REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INVESTIGATION OF THE EFFECTS OF CYCLOASTRAGENOL ON *ARABIDOPSIS THALIANA* BY HIGH THROUGHPUT SEQUENCING

Wissem MHIRI

DOCTOR OF PHILOSOPHY THESIS

Department of Chemistry

Chemistry Program

Advisor

Prof. Dr. Barbaros NALBANTOĞLU

Co-Advisor

Assoc. Prof. Dr. Özgür ÇAKIR

February, 2021

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INVESTIGATION OF THE EFFECTS OF CYCLOASTRAGENOL ON *ARABIDOPSIS THALIANA* BY HIGH THROUGHPUT SEQUENCING

A thesis submitted by Wissem MHIRI in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY** is approved by the committee on 24.02.2021 in Department of Chemistry, Chemistry (English) Program.

Prof. Dr. Barbaros NALBANTOĞLU	Assoc. Prof. Dr. Özgür ÇAKIR
Yıldız Technical University	Istanbul University
Advisor	Co-Advisor

Approved By the Examining Committee

Prof. Dr. Barbaros NALBANTOĞLU, Advisor _____

Yıldız Technical University

Prof. Dr. Lokman TORUN, Member _____

Yıldız Technical University

Assoc. Prof. Dr. Neslihan TURGUT KARA, Member _____

Istanbul University

Assist . Prof. Dr. Çiğdem BILEN, Member

Yıldız Technical University _____

Assoc. Prof. Dr. Zafer Ömer ÖZDEMIR, Member

University of Health Sciences _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Investigation of the Effects of Cycloastragenol on *Arabidopsis thaliana* by High Throughput Sequencing supervised by my supervisor, Prof. Dr. Barbaros NALBANTOĞLU. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Wissem MHIRI

Signature

This study was supported by the Research Funds of Yıldız Technical University (BAP project No: 3489)

Dedicated to my parents,

to my brothers,

to all my family and friends

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and my special thanks to my supervisor Prof. Dr Barbaros NALBANTOĞLU for his pure-hearted support, help and guidance during the last 5 years of my PhD journey. I would like to express my gratitude and thanks to my co-supervisor Assoc. Prof. Dr Özgür ÇAKIR for allowing me to work on this project, for his guidance and his kind advices through my PhD research.

I placed on record my sincere gratitude to Istanbul University for allowing me to work in their laboratories and using their facilities.

I acknowledge the financial support for my research study of Yıldız Technical University through the project BAP 3489.

I thank the graduate school of and all the teaching staff for their efforts, their commitment and their help during the first part of my PhD.

To my colleagues and friends and all members of GDO lab in the molecular biology and genetic department of Istanbul University; Assoc Pr. Dr Neslihan TURGUT-KAYA, Merve CEYLAN, Burcu KADIOĞLU, Seda YAŞAR, Yağmur Vecide YEŞILDIREK, Bekir ILGAR, Haluk ÇELIK, Burcu ARIKAN and Albert PREMKUMAR. Thank you for your help, for your companionship and for providing a pleasurable and friendly work atmosphere.

Many thanks go to the people I consider my family in Istanbul; Aymen HACHANA, Olfa BEN YAHIYA, Saeed ZINEDDINE, Hussam and Sara ALMASHHADANI, Mikela GJAPI, Müge ULAŞ, Sefa AKYÜREK, Sarvenaz NOUBAKHSH, Kübra AYDIN, Okan DERINALP, Ahmed MILHEM, Mohammad Fadheel HADDAD, Hind El-JANABI, Tuğba DAYIOĞLU, Fareed ASSAD and Doctor Ebru ERBAYAT; Thank you for the support, for the help, for all moment we shared, you are unforgetTable.

I would like to thank my best friends Rim JALLALI, Reem BEN BOUBAKER, Meher DERBEL and Nebila LECHIHEB for their support and their help.

To the best people in my life, I express my extreme gratitude for my parents Noureddine MHIRI and Afifa BOUZIRI-MHIRI for their unconditional support, for their help and their love. I would like to thank my brothers Aziz MHIRI, Ilyas MHIRI, Mohamad Karim MHIRI and his wife Safa DHAMEN-MHIRI for all what they have been doing for me.

Wissem MHIRI

TABLE OF CONTENTS

LIST OF SYMBOLS

atm	atmosphere
bp	Base pair
β	Beta
°C	Centigrade degree
C	Concentration
cm	centimeter
Gbp	Gega base pair
g	Gram
g	Gravity force (for centrifuge)
kb	Kilobasepair
L	Liter
m	Meter
mVA	Megavolt ampere
min	minute
µ	Micro
µl	Microliter
µm	Micrometer
µmol	Micromole

Mbp	Mega base pair
mg	Milligram
ml	Milliliter
M	Molar
ng	Nanogram
nm	Nanometer
rpm	Revolutions per minute
V	Volume

LIST OF ABBREVIATIONS

AMPK	Adenosine Monophosphate activated Protein Kinase
APS	Adenosine Phospho-Sulfate
AST	Astragolside
BGI	Beijing Genomics Institute
bHLH	basic Helix-Loop-Helix
BP	Biological Process
CAG	Cycloastragenol
CC	Cellular Component
ccDNA	closed circular DNA
cGMP	current Good Manufacturing Practice
CMV	Cytomegalovirus
DE	Differentially Expressed
DEFL	Defensin-like
DEGs	Differentially Expressed Genes
DEPC	Diethyl Pyrocarbonate
DESs	Differentially Expressed Small RNAs
DNA	Deoxyribonucleic Acid
DNBs	DNA nanoballs

ddNTPs	dideoxy Nucleotide triphosphates
DSG	Differentially Splicing Gene
EGFR	Epidermal Growth Factor Receptor
ERK	Extracellular Signal-Regulated Kinase
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase Million
FXR	Farnesoid
GATK	Genome Analysis Toolkit
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRAS	Generally Recognized As Safe
GRF	Growth Regulation Factor
GSL	Glucosinolate
HISAT	Hierarchical Indexing for Spliced Alignment of Transcripts
HGP	Human Genome project
HSPs	Heat Shock Proteins
JAK	Janus Kinase
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEK	Mitogen-Activated protein Kinase
MeS	Methylation Sequencing

MF	Molecular Function
miRNAs	Micro RNAs
MPSS	Massively Parallel Signature Sequencing
mRNA	messenger RNA
MS	Murashige and Skoog
MXE	Mutually Exclusive Exons
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PA	Phosphatidic acid
PAGE	Polyacrylamide Gel Electrophoresis
PC	Phosphoatidyl Choline
PE	Phosphoatidyl Ethanolamine
PIPmiR	Pipeline for the Identification of Plant miRNAs
piRNA	piwi-associated RNAs
PLD	Phospholipase D
PPi	PyroPhosphate
RCR	Rolling Circle Replication
RdDM	RNA-directed DNA methylation
Rfam	RNA families
RI	Retained Intron

RNA	Ribonucleic Acid
RNA pol	RNA polymerase
RNA-seq	RNA sequencing
SBS	Sequencing By Synthesis
SE	Skipped Exon
siRNAs	small interfering RNAs
SMS	Single Molecule Synthesis
SMRT	Single-Molecule Real-Time sequencing
sncRNAs	small non-coding RNAs
sRNAs	small RNAs
SRA	Sequence Read Archive
STAT	Signal Transducer and Activator of Transcription
SVM	Support Vector Machine
SNV	Single nucleotide variant
TEM	Transmission electron microscopy
TERT	Telomerase reverse transcriptase
TFs	Transcription Factors
TiRNAs	Transcription-initiation RNAs
TPM	Transcripts Per Kilobase Million
TRAP	Telomere Repeat Amplification Protocol

tRNA	Transfer RNA
VCF	Variant Call format
WES	Whole-Exome Sequencing
WGS	Whole Genome Sequencing
WTSS	Whole Transcriptome Shotgun Sequencing SMWT

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Investigation of The Effects of Cycloastragenol on *Arabidopsis Thaliana* by High-Throughput Sequencing

Wissem MHIRI

Department of Chemistry

Doctor of Philosophy Thesis

Advisor: Prof. Dr. Barbaros NALBANTOĞLU

Co-advisor: Assoc. Prof. Dr. Özgür ÇAKIR

Cycloastragenol (CAG) is a compound isolated from ‘Astragalus membranaceus’ which has been proven to significantly stimulate telomerase activity, cell proliferation and prevent some diseases in human. Similarly, plants endure different stress conditions leading to physiological, biochemical and molecular perturbations and affecting their growth and development. CAG would be for plants a potential molecule preventing and remedying these disorders. In this frame, our study, the first to investigate CAG in plants, aims to determine effects of cycloastragenol on different signaling mechanisms, to have an overview of transcriptional and post- transcriptional responses analysis through high throughput sequencing methodologies, in order to reveal CAG potential help to plants in surmounting different environmental stresses. RNA sequencing and small RNA sequencing strategies were employed in A. thaliana roots calli under CAG treatment to determine the transcriptional profiles, and to assess known and unknown small RNAs and evaluate their possible functions in the affected pathways. The generated data revealed the upregulation of different genes and stress responsive protein families such as cytochrome P450s transporters. Besides, several miRNAs and their targets associated to

metabolic and stress tolerance signaling pathways were found to be up-regulated in treated sample, like miR3434-5p and its target genes which interfere positively in plant responses to hyperosmotic stresses.

Keywords: High-throughput sequencing, cycloastragenol, RNA sequencing, small RNA sequencing, *A. thaliana*.

ÖZET

Sikloastagenol'ün *Arabidopsis Thaliana* Üzerindeki Etkilerinin Yüksek Verimli Dizileme İle Araştırılması

Wissem MHIRI

Kimya Bölümü

Doktora Tezi

Danışman: Prof. Dr. Barbaros NALBANTOĞLU

Eş-Danışman: Assoc. Prof. Dr. Özgür ÇAKIR

Sikloastragenol (CAG), telomeraz aktivitesini, hücre proliferasyonunu önemli ölçüde uyardığı ve insandaki bazı hastalıkları önlediği kanıtlanmış '*Astragalus membranaceus*' dan izole edilen bir moleküldür. Aynı şekilde, bitkiler fizyolojik, biyokimyasal ve moleküler pertürbasyonlara yol açan ve büyümelerini ve gelişimlerini etkileyen farklı stres koşullarına dayanırlar. CAG, bitkiler için bu bozuklukları engellemek ve iyileştirmek için anahtar bir molekül olacaktır. Bu çerçevede, bitkilerde CAG'yi ilk araştıran çalışmamız, bitkilerde farklı çevresel baskıların üstesinden gelmek için, CAG potansiyel yardımını ortaya çıkarmayi, farklı sinyalleşme mekanizmaları üzerindeki CAG etkilerini belirlemeyi, yüksek verimli sıralama metodolojileri aracılığıyla genel bir transkripsiyonel ve transkripsiyon sonrası yanıt analizine sahip olmayı amaçlamaktadır. RNA dizileme ve küçük RNA dizileme stratejilerini ve transkripsiyonel profillerini belirlemek için bilinen ve bilinmeyen küçük RNA'ları değerlendirmek ve etkilenen yollardaki olası işlevlerini değerlendirmek için CAG muamelesi altında *A. thaliana* kökleri kalluslari kullanılmıştır. Üretilen veriler, sitokrom P450s taşıyıcıları gibi farklı genlerin ve strese duyarlı protein

ailelerinin upregülasyonunu ortaya çıkarmıştır. Bununla birlikte, mir3434-5p ve hiperosmotik strese bitki tepkilerini olumlu yönde etkileyen hedef genleri gibi muamele edilen örneklerde, metabolik ve stres toleransı sinyal yolları ile ilişkili birkaç mirna ve hedeflerinin yukarı doğru düzenlendiği bulunmuştur.

Anahtar Kelimeler: Yüksek verimli dizileme, Sikloastragenol, RNA dizileme, küçük RNA dizileme, *A. thaliana*.

1
INTRODUCTION

1.1 Literature review

Plants have an outstanding place in human life, they shaped the world we are living in and are fundamentally responsible for creating human life conditions. About 450 million years ago, the earliest land plants showed up on Earth and extraordinarily changed the Earth's atmosphere by increasing oxygen amount level and reducing carbon dioxide amount, which allowed to other organisms to evolve and flourish, including our evolutionary ancestors.

Plants are able to harvest energy through photosynthesis and make it available for both human and animals. They are also producing food which is ultimately the main element connecting human to plants. However, they also have an important role to orchestrate in human health. Actually, archaeological records report that plants, particularly medicinal plants, have been used for healing earlier than at least 5000 years [1]. The strongest significant connection between human health and plants was revealed by Friedrich Bayer and his partners (1897) when they launched aspirin, the synthetic acetylsalicylic acid, which is a salicylic acid synthetic analogue and an active ingredient of willow bark, known from very ancient civilizations to be a remedy for headache and fever [2]. In more recent time, Galantamine, a chemical isolated from daffodil has been used to treat Alzheimer's disease [3].

In the same frame, cycloastragenol (CAG, MW: 490,72 g/mol) is a molecule derived from *Astragalus IV*, which is extracted from Chinese medicinal herbs and has been proved to stimulate telomerase activation which led to report anti-aging role of CAG. Moreover, several studies report that CAG, besides its anti-aging effect, has more than one pharmacological action such as anti-bacterial, anti-fibrosis and anti-inflammation. Others proved that CAG stimulate circulatory and urinary systems in humans and helps in healing burns and in preventing several

heart diseases [4], [5]. Cycloastragenol improves cardiac malfunction and remodeling through stimulation of myocardial cells autophagy [6]. CAG boosts telomerase activity and cell proliferation in humans and rats and reported in other study to reduce depression-like behavior in mice. Indeed, it was shown that a long-time exposure to chronic stress or depression is associated with rapid telomerase shortening, suggesting the existence of a relation between depression and telomere maintenance [7].

To highlight all these facts, TA-65 a diet supplement derived from CAG was launched [8]. This would push us to wonder if cycloastragenol which was proved to contribute to the improvement of human health like other plants-derived molecules, why it cannot help plants themselves? A full perception would give also some achievements to plants, which are similar to humans, subjected to different biotic and abiotic stresses, leading to several physiological, biochemical, molecular and gene expression variations that can be affected by CAG.

The present study is the primary research work investigating CAG effects in plants. Actually, CAG would be a key molecule to help plants to overcome the effects provoked by the different biotic and abiotic stresses that plants undergo, as well as it was proved on humans. Several available approaches can help to discern the involved pathways, determining the expression of concerned genes, developing then the perception of the affected mechanisms. Genome sequence of *Arabidopsis* availability and the evolution of the sequencing technologies favor genes determination and their analysis at the transcriptome level. Fred Sanger and Walter were the pioneers to determine base sequences in nucleic acids and since that, sequencing technologies knew revolutionary development and brought innovative accomplishments in the biological research field.

High-Throughput Sequencing (HTS) also called deep sequencing or Next Generation Sequencing (NGS), is a fundamentally different technology, comparing to Sanger sequencing, that contributed to distinguished discoveries and revealed impressive observations in relation with genome, transcriptome and epigenome. However, NGS technologies output a huge amount of data which does

not sometimes have a meaningful biological sense. In this frame, bioinformatics combines mathematics, computer science and molecular biology to generate useful tools that are used in analyze the obtained data from NGS. Next generation sequencing was used in the beginning for whole genome sequencing (WGS) [9], however these technologies have been developed for other applications such as whole transcriptome shotgun sequencing (WTSS) which is also known by RNA sequencing (RNA-seq)[10], whole-exome sequencing (WES) [11] or methylation sequencing (MeS)[12].

The evolution of next-generation sequencing (NGS) technologies has highlighted the relevance of sequencing-based approaches in gene expression profiling. The remarkable sensitivity and high-throughput feature made of the NGS the method of choice for gene expression analysis [13]. Used to identify the coding and non-coding transcriptional events or to select a subgroup of targeted RNA genes in a given sample [10], RNA sequencing gives an accurate and sensitive gene expression level measurements [14]. While small RNA sequencing is generally used to identify new small RNA and to give a quantitative profile of non-coding RNA expression [15] since their importance and multiple roles, particularly miRNAs and siRNAs, in gene regulation [16].

The present study represents the first research work evaluating cycloastragenol in plants. In literature, all studies have been evaluating this molecule on human cells and highlight its multiple effects particularly in human health. We applied the NGS Illumina/Solexa technology to Figure out the changes in transcriptome and miRNAome in *Arabidopsis thaliana*, by respectively RNA-seq and small RNA-seq, after cycloastragenol treatment in order to assess its effect on plants. In this work, I start by defining cyclostragenol and explaining its effect in humans and how it would be useful to be used with plants. Then, I describe the sequencing methodologies from the earliest to the newest ones with a focus on the next-generation sequencing and how it is employed in gene expression analysis. In the last chapter of the introduction, I illustrate gene expression and explain briefly its transcriptional and post-transcriptional regulation.

1.2 Objective of the thesis

As it has been proved in several studies [17], [5], [6], [7], cyclostragenol, derived from Chinese medicinal herb, improves the functioning of the immune system and protects from different diseases in human. Plants are like humans, submitted to different type of stresses which cause different physiological disorders and affect their immune system. Considering *Arabidopsis thaliana*, a model biological organism, the aim of the current study is to Figure out the effect of CAG on plants on the transcriptomic level, to determine the affected pathways and the involved genes that can be used later in other innovative breeding programs for agricultural and/or pharmaceutical purposes.

1.3 Hypothesis

CAG treatment affects specific pathways and leads to induce the transcription of some interested genes that are in relation with plant adaptation and/or overcoming some kind of stresses, like it was shown for humans. Hence, these genes and/or their products can be used to establish plant breeding program and/or develop novel plant-based pharmaceutical products.

2

CYCLOASTRAGENOL

2.1 Origin and effects

Cycloastragenol (CAG, $C_{30}H_{50}O_5$, 490.72 g/mol) is an aglycone derived from Astragaloside IV (AST IV), which is the principal saponin isolated from *Astragalus* genus (*Fabaceae*) roots, broadly known to be used in herbal Chinese medicine [18], [19]. CAG is a triterpenoid compound obtained from AST hydrolysis (Figure 2.1)[5].

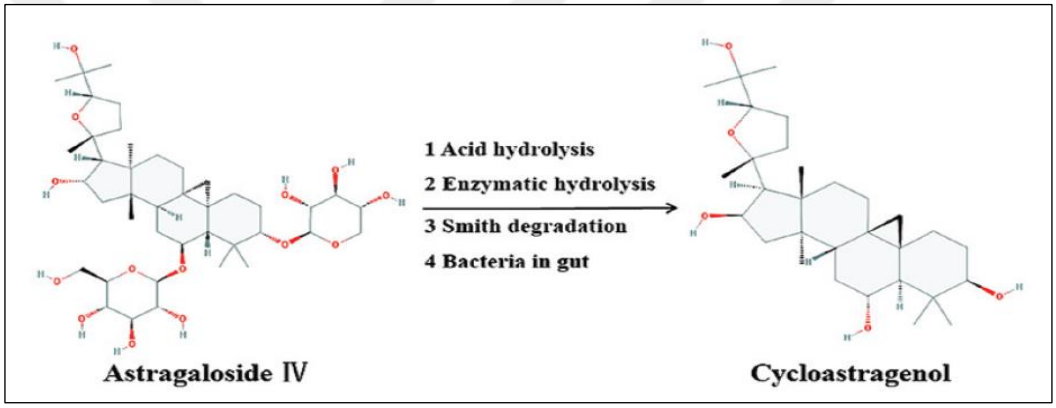


Figure 2. 1 Schematic hydrolysis reaction of cycloastragenol [5]

Actually, AST IV is the main active ingredient of *Radix Astragalis* (RA and Huang-qi in China) which is one of the most ancient and the most used Chinese medicine herbs derivative from dried roots of *Astragalus membranaceus* (*Fisch*.) Bge. var. *mongholicus* (Bge.) *Hsiao* or *Astragalus membranaceus* (*Fisch.*) Bge. RA has been used in traditional Chinese medicine, in several herbal preparations, for more than 2000 years, for its various effects. According to 'Chinese Materia Medica', it has been acknowledged to boost the immune functions, to function as diuretics, and to have anti-stress, anti-hypertensive, and considerable anti-bacterial properties in

humans, it also protective effects regarding the liver, the lung, cardiovascular and neural tissues and renal function[20], [21], [5].

A lot of researchers confirm that CAG is a triterpene and derivative aglycone of AST IV which is easily deglycosylated in the gastrointestinal tract [18], [22], [7], [23], [6], [5]. Indeed, AST has physiochemical characteristics that may contribute to its poor bioavailability such as a high molecular weight, high molecular flexibility, significant hydrogen-bonding capacity and low membrane permeability. However, AST goes through intestinal bacterial transformation giving result to CAG which can be immediately absorbed to get to the circulation system.

As a triterpene, several studies report the pharmacological effects of CAG (Figure 2.2) such as anti-viral, anti-inflammatory, anti-microbial, anti-hepatotoxic, anti-tumor, anti-leukemia, anti-nociceptive and immunomodulate activity [23]; For instance, CAG has a significant anti-inflammatory activity related to the decrease of the general intracellular CA^{2+} by inhibiting lymphocytes activation, proliferation and cytokines expression [24]. CAG was also reported to increase the anti-viral function of CD^{8+} lymphocytes from HIV-1-infected patients [25]. Anderson Cancer Center in Texas found out that AST and CAG enhance the immune system after the decline of mortality of to 50% to 10% of patients with breast cancer whose were given astragalus and their chances of survival are doubled [4]. Furthermore, CAG ameliorates cardiac dysfunction and remodeling through promoting autophagy in myocardial cells [6]. CAG inhibits the activation of stress-associated TXNIP/NLRP3 inflammasome in the endoplasmic reticulum, and reduces cell apoptosis ameliorating then the endothelial dysfunction [26].

Various research studies report that CAG have also an anti-aging action. Actually, cycloastragenol was reported to be a telomerase activator substance in the traditional Chinese medicine which helps in restoring telomeres after cell division [25], [8], [4]. Discovered by the Telomere Repeat Amplification Protocol (TRAP) screening, CAG delayed telomere shortening and increased then telomerase activity and cell proliferation in neonatal keratinocytes, it reduces depression-like

behavior in Mice showing that a continuous exposure to stress or depression is associated with telomeres curtailment, which led to suggest that depression and telomere maintenance are linked somehow [27], [7], [24].

CAG has a wide range of pharmacological effects, however, the fundamental details of the mechanisms for most of these effects still ambiguous.

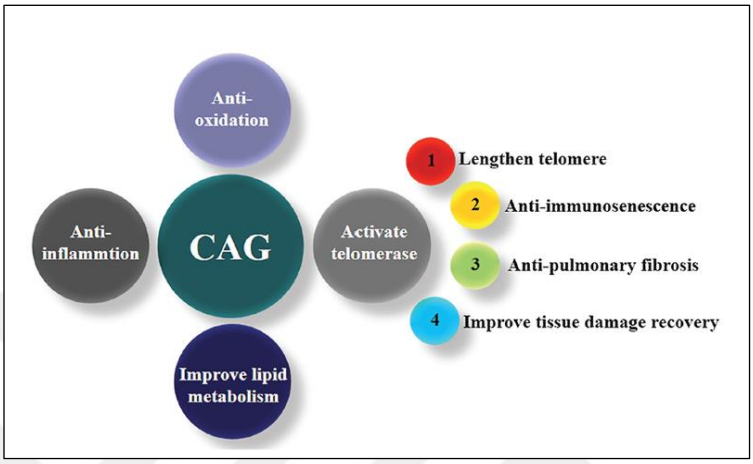


Figure 2. 2 Pharmacological effects of CAG [5]

The Figure below summarizes CAG-associated effects based on previous literature and represents a hypothetical mechanisms for CAG effects; Arrowheads mean trigger modifications while continuous lines represent the direct activation modifications according to the literature. Dashed lines indicate modifications by stimulation and grey dashed lines show the possibly activation modifications.

CAG can stimulate telomerase activation via pathway 1, 2, 3 and 4, and then enhance different effects. Over pathway 5, CAG stimulates in a direct way the FXR to better hepatitis and stimulates indirectly AMPK through pathway 6 to improve inflammation.

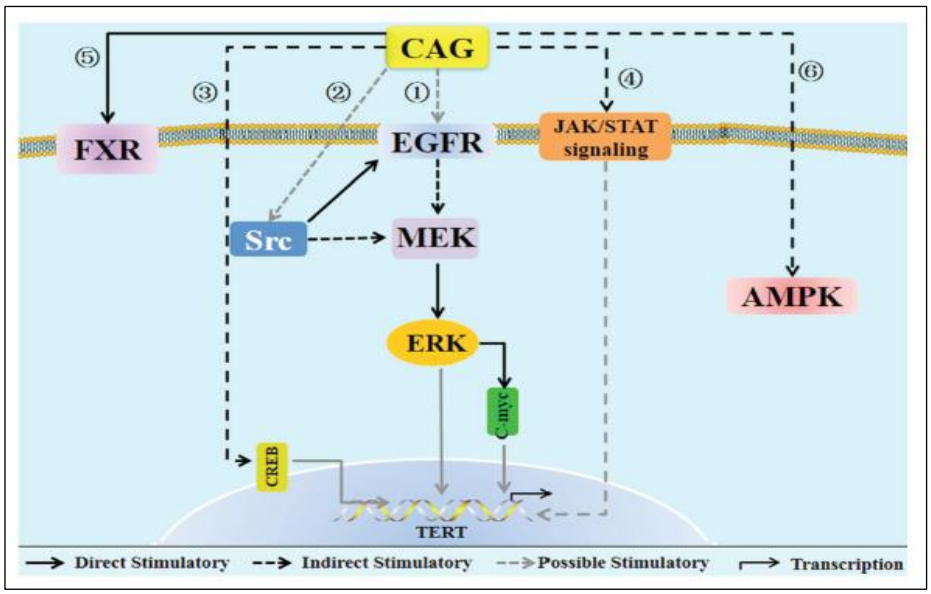


Figure 2. 3 Hypothetical mechanisms associated with CAG [5]

AMPK: adenosine monophosphate-activated protein kinase, **CAG**: cycloastragenol, **EGFR:** epidermal growth factor receptor, **ERK:** extracellular signal-regulated kinase, **FXR:** farnesoid, X receptor, **MEK:** mitogen-activated protein kinase, **JAK:** Janus kinase, **STAT:** signal transducer and transcription activator, **TERT:** telomerase reverse transcriptase.

2.2 From CAG to TA-65

Hong Kong University of Science and Technology analyzed several natural elements extracted from Astragalus genus in collaboration with the Geron Corporation and identified CAG as anti-aging compound which know to be the only one promoting telomerase activity [5]. Telomerase is a ribonucleoprotein complex, composed of a reverse transcriptase that elongates telomeres by repeat DNA sequences at the ends of chromosomes, preventing then their shortening [28], [29]. Several studies have been investigated telomere shortening in human cells and genetic using telomerase with mutation and animals with defected telomerase, that led to confirm the relation between cell aging and telomere loss [30], [31]. In the same frame, other researchers report telomere shortening in human as one of the factors risks for several health problems such as atherosclerosis, metabolic syndrome, cardiovascular disease, diabetes, hypertension, infections, fibrosis, Alzheimer disease, cancer, and overall mortality [8]. PattonProtocol-1, a commercial health care program, which was launched in

2007, introduced a CAG product-derivative telomerase activator; TA-65 (10-15 mg daily) as a food supplement. TA-65 formulation was manufactured under the regulation of current good manufacturing practice (cGMP) and produced as GRAS (Generally Recognized As Safe) to be sold as dietary supplement in the medical food by TA sciences company. Initial trials focused on the immune system and showed that low quantities of TA-65 activated, moderately, telomerase enzyme in human keratinocytes, fibroblasts and immune cells in culture [32]. These authors presented an evidence from an arbitrary, double blind, placebo-controlled study demonstrated that TA-65 as a dietary supplement lengthen telomeres and improve human health with no safety concerns after adding the product in the dietary of Cytomegalovirus (CMV) infected subjects.

3
SEQUENCING

3.1 Sequencing methods

Nucleic acids order in polynucleotide chains carries the hereditary information and the biochemical properties of life. Therefore, being able to determine sequences is very important in the biology research. Sequencing methodology aims to obtain the exacting order of a sequence of nucleotides in nucleic acids which help researchers to illuminate the genetic information from any biological system. DNA/RNA sequences deciphering is imperative for almost all branches of life sciences and its insight and cost-effectiveness have developed exponentially in the past decades, through several refinements of the first sequencing technique 'Sanger sequencing' to the invention of totally new technologies (Figure 3.1). The booming success of the Human Genome project (HGP) is widely the result of the early and continuous development sequencing methods. Over the last decade, and thanks to the automation and the refinement of the established sequencing methods, the HGP provoked a 100-fold reduction in sequencing costs [33]. Several academic and commercial efforts remain in progress to develop new ultra-low-cost sequencing technologies in order to reduce over and over the cost of sequencing and perfect their effectiveness.

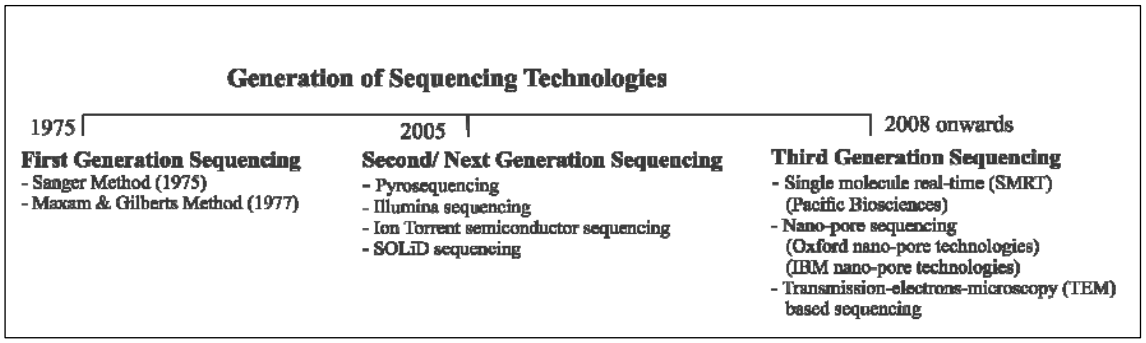


Figure 3. 1 Overview of sequencing generations [34]

3.1.1 First generation sequencing

The three-dimensional structure of DNA was discovered in 1953 based on crystallographic data produced by Rosalind Franklin and Maurice Wilkins and con-tributed to the theoretical DNA replication and encoding proteins in nucleic acids concepts [35]. However, reading or sequencing the DNA did not follow rapidly. Actually, development of some strategies was able to find out the sequence of protein chain earlier than the one of nucleic acids, which were much longer and made of fewer similar units, making hard to distinguish them. However, the discovery of RNase enzyme which is capable to cut RNA chain at specific sites, the ability to measure nucleotide composition but not the order and the selective ribonuclease treatments enhance the production of fully and partially degraded RNA fragments [36], [37]. It was also able to produce in 1965 the first whole nucleic acid sequence that of alanine tRNA from *Saccharomyces cerevisiae* [38]. Parallelly, a new technique based on detecting radio labelled partial-digestion fragments using two-dimensional fractionation was developed [39], helping to study the sequences of ribosomal and transfer RNA sequences [40]. Exploiting the same technique, first complete protein-coding gene sequence; the coat protein of bacteriophage MS2, was produced in 1972 followed by its whole genome in 1976. And around that time, several researchers were established to replace 2-D fractionation by electrophoresis separation by polyacrylamide (PPA) gels giving birth in mid-70 to the first two widely-known sequencing methods; Alan Coulson and Sanger's ‘plus and minus’ system[41], [42] and Allan Maxam and Walter Gilbert's chemical cleavage technique [43].

3.1.1.1 Maxam-Gilbert sequencing: Chemical cleavage sequencing

Developed by Allan Maxam and Walter Gilbert between 1976 and 1977, this methodology is based on DNA chemical modifications following by cleavage at specific bases (Figure 3.2). Restriction digestion of plasmid followed by electrophoretic separation are used to obtain the DNA fragment to be sequenced, which is isolated thanks to its labelling. Indeed, the fragment can be 5’ or 3’ 32P terminally alkaline phosphatase labelled and polynucleotide kinase labelled for 5’

or using terminal transferase labelling for 3'. Double stranded fragment is denatured and single strands are separated and isolated on PAA gel, which enable the sequencing of each strand [43].

This approach was not extensively used because of two major inconveniences; the use of dangerous chemicals such as hydrazine and the fact that abundant DNA amount is needed which leads to have several reactions and make the technique heavy. Still, this method could be used to sequence very short fragments [44].

3.1.1.2 Chain termination method: Sanger sequencing

Being more efficient and using fewer toxic chemicals and lower radioactivity amount than the chemical cleavage sequencing, Sanger sequencing technique is based on the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators [45]. Before establishing this sequencing method, almost the bacteriophage FX174 complete sequence genome was found out, using 'plus and minus' sequencing method that it was previously developed [42]. Shortly after, chain terminator sequencing was developed and then the FX174 complete sequence of 5368 bp long was determined using the same technique, describing the first sequenced DNA genome.

Unlike Maxam and Gilbert sequencing, Sanger technique requires DNA polymerase I, *Klenow* fragment as primer, single-stranded DNA template, labelled nucleotides by radioactivity or fluorescence and modified nucleotides for DNA strand elongation. DNA sample is divided in four sequencing reaction mixtures containing the previously mentioned requirements besides the standard deoxynucleotides (dATP, dGTP, dCTP, dTTP), three of them (example 32P-dATP, dCTP, dGTP) are in equimolar concentration while the fourth one (example dTTP) in lower concentration. Only one of the dideoxy nucleotides (ddATP, ddGTP, ddCTP, ddTTP) presenting the chain terminating nucleotides without 3'-OH group, to each reaction, is required for a phosphodiester bond between two nucleotides. Both dNTP and ddNTP enables different length chains synthesis. The four reaction, with 32P-ssDNA, are separated on the lanes next to each other of

the denaturing PAA gel followed by autoradiograph analysis for reading of the sequence (Figure 3.2) [44], [40].

Chain termination method has notably simplified DNA sequencing. Non-specific binding of the primer to the DNA affecting accurate read-out of the DNA sequence, and DNA secondary structures disturbing the fidelity of the sequence, present the limitation of sanger sequencing [45]. However, thanks to the simplicity and the reliability of this technique, it quickly became the most used one.

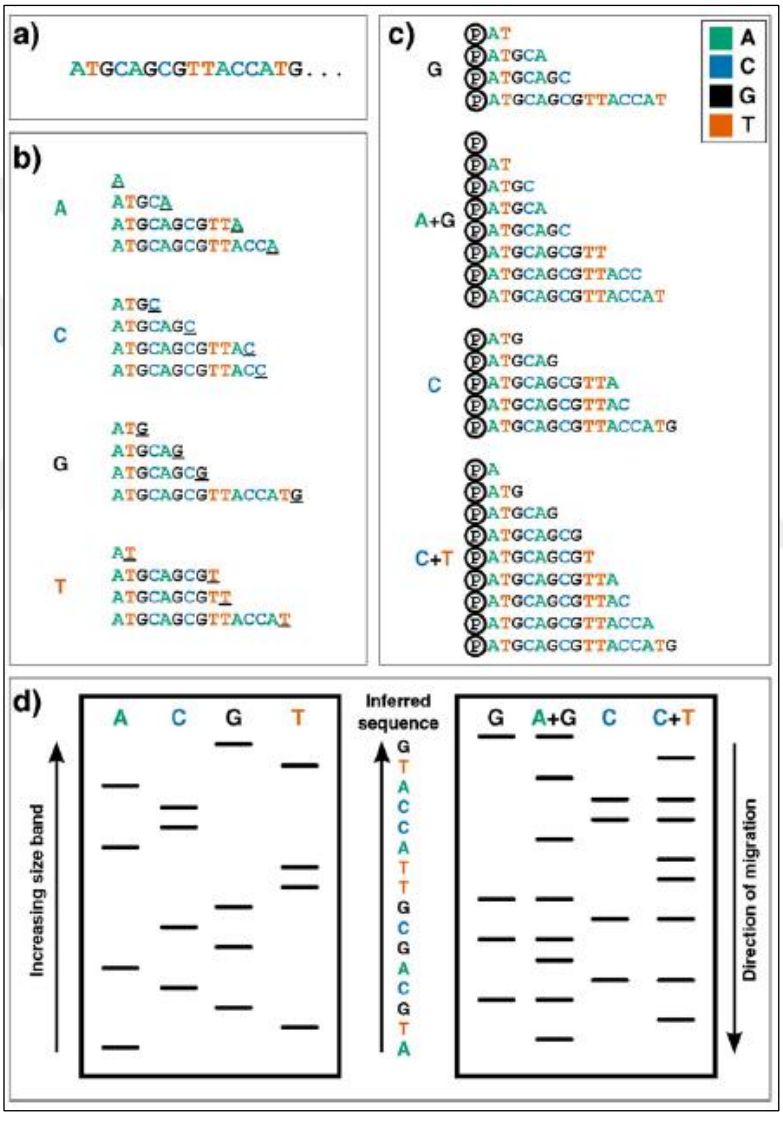


Figure 3. 2 First generation sequencing technologies [40]

a) DNA example to be sequenced **b)** Sanger sequencing: Chain termination sequencing **c)** Maxam-Gilbert sequencing: Chemical cleavage sequencing **d)** Fragments generated from each methodology.

A lot of enhancement have been done on Sanger sequencing like the detection through capillary based electrophoresis which lead to the development of the automated sequencing machines used in the sequencing of complex species genomes, such as the first bacterial genome *Haemophilus influenzae*, the first eukaryotic genome *Saccharomyces cerevisiae*, the first multicellular animal genome *Caenorhabditis elegans* and the first plant genome *Arabidopsis thaliana* [44], [40].

3.1.2 Second generation sequencing: Next generation sequencing

Because of the limitations in throughput, resolution, speed and scalability of first-generation Sanger sequencing approach, second-generation of sequencing techniques, called Next-Generation Sequencing (NGS), have been developed for cheaper and faster sequencing technology.

The main characteristic of this new sequencing generation is the parallelization of an important number of sequencing reactions where millions of DNA fragments are accurately sequenced from one single sample. Sequencing reactions of NGS are reduced in number and time compared to Sanger sequencing. The workflow of second generation sequencing technologies can be divided principally in library preparation, sequencing, imaging and data analysis [46].

It all started with the first NGS technology, the Massively Parallel Signature Sequencing (MPSS, Lynx Therapeutics Company). After MPSS, 454 sequencing was launched by Life Sciences, taken over later by Roche company to improve the technique and create platform. Initially 454 GS 20 were producing 20 Mbp and then upgraded during the last decade to be able to produce up to 1000 Mbp in 2014. In 2006, Solexa released Genome Analyzer, acquired by Illumina which has developed Mi-Seq and Hi-Seq platforms; the first platform is able to sequence up to 15 Gbp while the second one can sequence up to 600 Gbp. Later, Illumina developed Hi-seq 2500 system that is able to sequence a human genome in one day. Coming after, other technologies have been established like the SOLiD platform (Applied Biosystems), Ion Torrent PGM (Thermo Fisher Scientific), Complete Genomics (Beijing Genomics Institute (BGI)) and the Polonator

(Dover/Harvard). All these technologies can be summarized by continuous enzymatic manipulation and imaging-based data collection [45], [34], [47], [40].

All NGS platforms have three main steps (Figure 3.3):

1) Sample preparation for sequencing that involves the addition of the adapters at the end of the fragmented sample sequences. Preparation of DNA with universal nucleic acid terminals is considered as 'sequencing library'. Adapters are necessary to attach the library DNA fragments to a solid surface in order to describe the starting site of the sequencing reaction. 2) Sequencing library amplification, whether in solution in emulsion or in situ, in order to generate a cluster of copies. 3) Synthesizing or using fluorescent nucleotides for ligation during sequencing, [48].

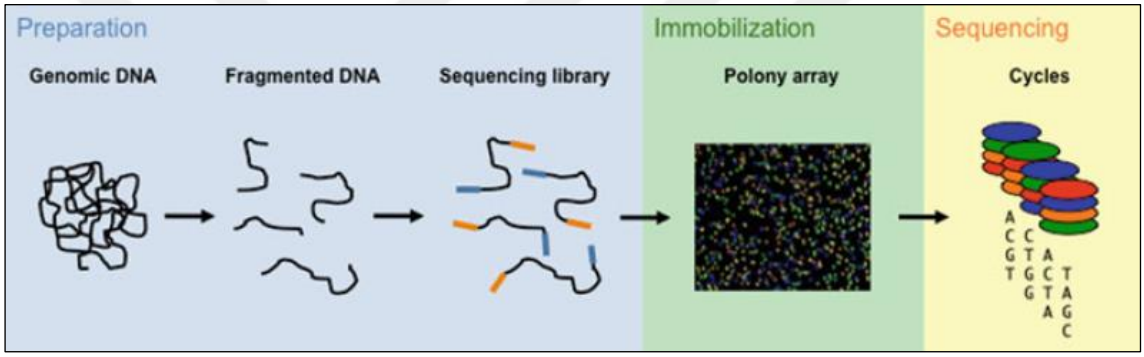


Figure 3. 3 High-throughput sequencing general workflow [48]

3.1.2.1 Lynx therapeutics massively parallel signature sequencing (MPSS)

Considered as the first 'Next generation sequencing' technology, MPSS was launched in 1990s by Lynx therapeutics company. This approach was based on bead using adapter ligation followed by adapter decoding. The sequence is read in increment of four nucleotides making it very susceptible to lose specific sequences. MPSS was a high throughput sequencing but its output data included hundreds of thousands of short sequences. This method was very used to sequence cDNA sequences for gene expression level measurements. Lynx therapeutics was taken by Solexa and then by Illumina [45].

3.1.2.2 Pyrosequencing/454 Technology

Launched by Life Sciences and developed later to 454 technology, pyrosequencing method lean on light generation detected by the added nucleotides simultaneously, while their incorporation by the polymerase and the incorporation of dNTP releases pyrophosphate (PPi). It is an enzymatic system-based methodology including 4 enzymatic systems (Figure 3.4); *Klenow* fragment of DNA Polymerase I, ATP sulfurylase, Luciferase and Apyrase. It also involved the enzyme substrates adenosine phosphosulfate (APS), D-luciferin and the sequencing template with an annealed primer to be used as starting material for the DNA polymerase. Actually, Pyrosequencing uses luciferase to generate light, PPi is converted to ATP with the help of ATP sulfurylase and adenosine 5'-phosphosulfate (APS). Luciferase-ATP moderates the conversion of luciferin to oxyluciferin generating then a light signal to detect the incorporated single nucleotides to the starting DNA sequence. The generated light amount is proportional to the number of molecules of ATP, themselves proportional to the released pyrophosphates number then to the incorporated nucleotides [44], [45],[49], [50]. Pyrosequencing was then combined with PCR emulsion to develop the 454 Life Sciences platform by Roche Applied Sciences. This technology is considered as miniaturized and massively parallelized pyrosequencing, using GS FLX flow cell which is introduced as 'PicoTiterPlate' and invented from the fused fiber-optic bundle.

For sequencing, a DNA template library is developed by fragmentation by the way of nebulization or sonication. Fragments of certain hundred base pairs in length are end-repaired and ligated to adapter oligonucleotides. The library is denatured by dilution to a particular molecule concentration, and hybridized to separate beads containing complementary sequences to adapter oligonucleotides. DNA molecules are bound to the beads which are classified into water-in-oil micro-vesicles during emulsion PCR. Emulsion PCR is disturbed after amplification which enhances the enrichment of the clonally amplified DNA template that beads contain. Beads are removed from each other by dilution and accumulated into separated PicoTiterPlate wells and mixed with the enzymes. Plates are loaded into

GS FLX system to act as flow cell and pyrosequencing is carrying out by adding four dNTPs. Every incorporation of a nucleotide is a well that hold clonally amplified template producing a transmitted luminescence along the fiber optic plate. Every transmission is made on a charge-coupled camera and wells are imaged with every dNTP reagent flow, then analyzed and translated into a linear sequence output [50].

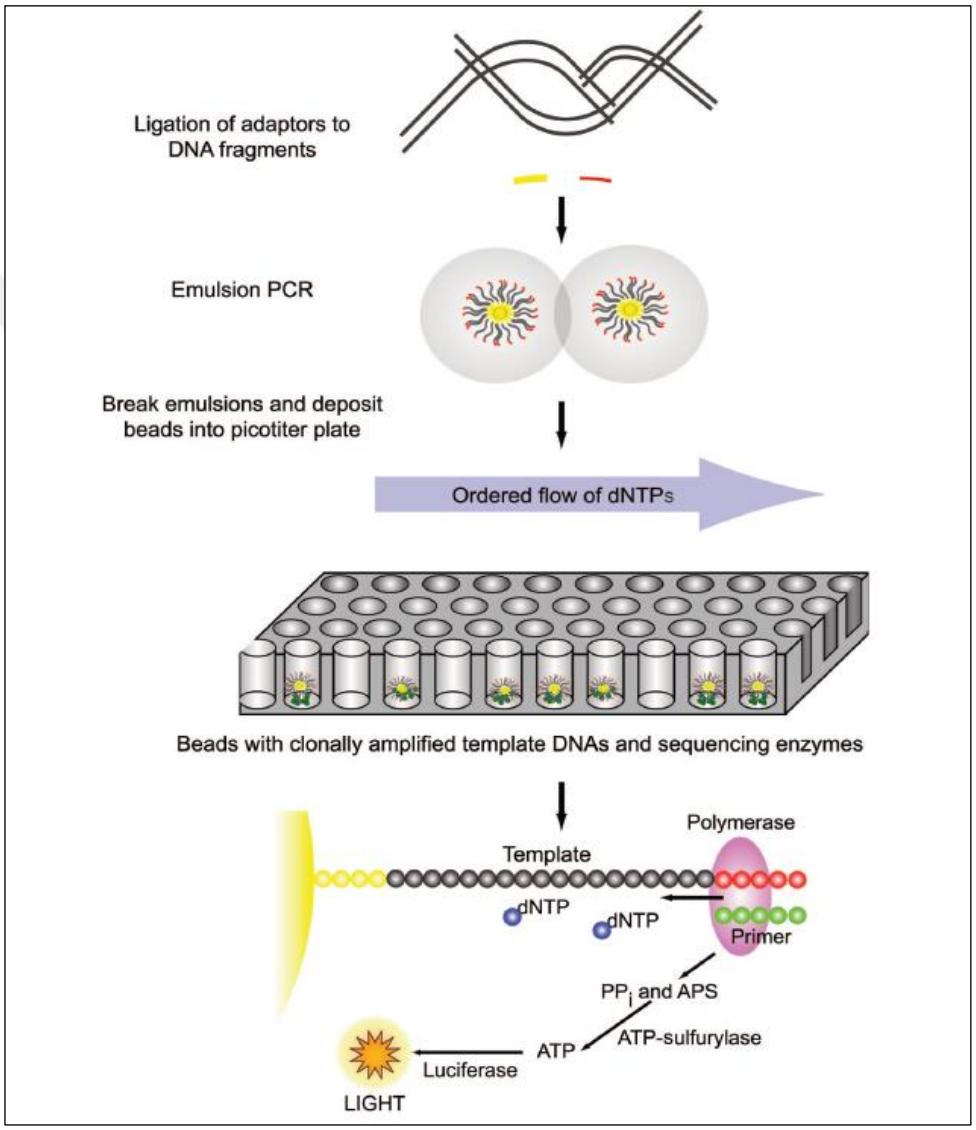


Figure 3. 4 Overall workflow of pyrosequencing technology [50]

3.1.2.3 Illumina/Solexa sequencing

Able to produce larger data volume, Illumina technology is the first 'short read' sequencing platform that is based on sequencing by synthesis and on dye

terminators, using bridge amplifications to generate clonally-enhanced template DNA (Figure 3.5). DNA template is divided into various of hundreds of base pairs lengths with end-repaired, generating termini of 5' phosphorylated blunt ends. A base is attached at 3' end of the phosphorylated blunt end of the DNA fragments with the help of *Klenow* fragment. DNA fragments were then prepared to be ligated to oligonucleotides adapters that have an overhang of T base at their 3' end, increasing then the efficiency of the ligation. Actually, two distinct set of adaptors are attached to fragments termini; one set is required to bind the flow cell and a second set is for sequencing. Templates are attached to the surface of a glass slide (the flow cell) thanks to the sequence of the adapters. The attached primers and adapter complementarity of the single-stranded template free-end contribute to the formation of *bridge amplifications*. These bridges present the annealing of adapters to the primers. The second strand is generated after the addition of PCR reagents. The denaturation step contributes to the bridge dissociation; multiple amplification cycles generate clonally amplified clusters which are denatured and washed leaving only the forward strands. Every sequencing cycle is initiated by the forward strand hybridization to a complementary primer to the adapter sequence, then followed by adding the polymerase. Reversible dye terminators which are colored with four different fluorescents, are incorporated to the complementary sequence of all strands in a cluster of clones. The excess reagents are washed and the fluorescence is then recorded. The fluorescent dye from incorporated nucleotides is removed by chemical treatment and a new sequencing cycle is then performed.

A single flow cell of the Illumina/Solexa platform consists of eight separated lanes. Each of them contains several million clusters. Then, the platform is able to sequence parallelly eight independent samples during one sequencing run. Illumina have strategies to sequence single-read and paired-read sequencing. Paired-end methodology generate sequence data from both ends of each template fragment. The whole process requires 2.5 days to generate 50 x 106 clusters per flow cell. However, the newest platform Genome Analyzer II, is able to perform higher cluster density. Illumina/Solexa system is joined to a software that

generates the first data including a base calling algorithm transforming image-based signals into nucleotides and the poor quality are removed by the same software. Supplementary bioinformatics analysis process is required to optimize the results.

This approach can widely read up to 36 bp, with an average raw error rate of 1-1.5%. However generated longer read are possible which can provoke a higher error rate [44], [47], [49]. In spite of its drawbacks; relatively high error percentage especially the error of substitution, which decreases the accuracy towards 3' ends and induces read mapping variation in case of long reads production, Illumina is the most used NGS technology for DNA and RNA analysis as well as for small RNA studies [46].

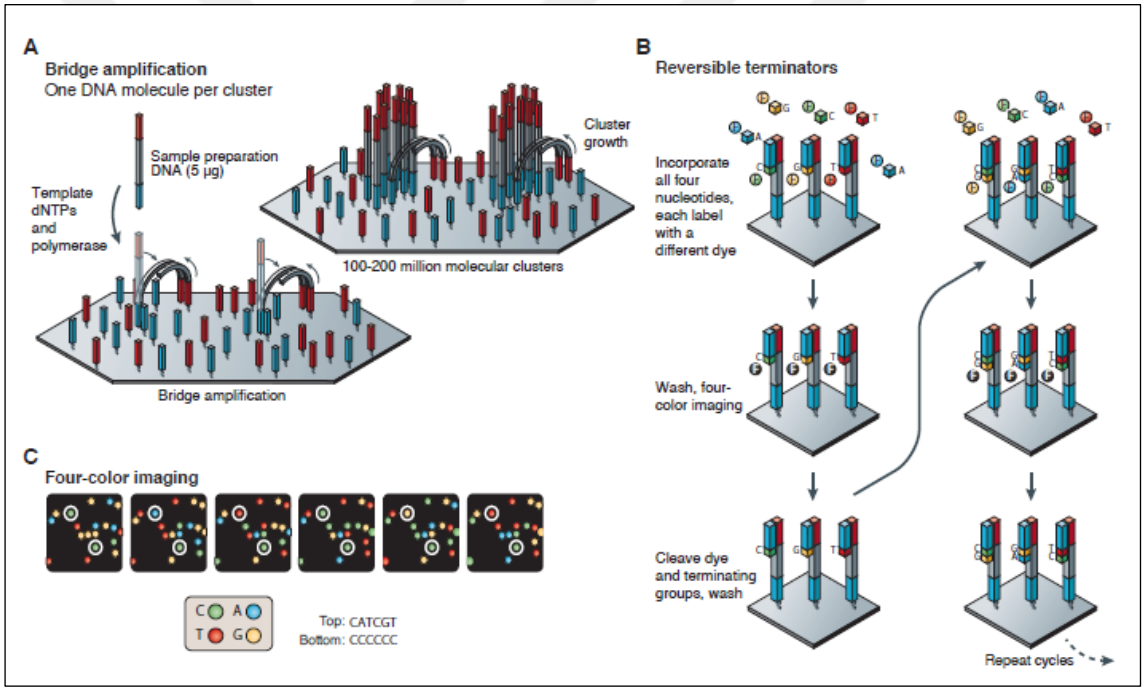


Figure 3. 5 Illumina/Solexa sequencing workflow (adapted from [46])

3.1.2.4 SoLiD sequencing

Launched in 2007, SOLiD ou Supported Oligonucleotides Ligation and Detection is one of the next generation sequencing technologies of short-read. It is based on using emulsion PCR to lock up the library of DNA to solid support and sequencing by ligation (Figure 3.6).

Sample prepration has similarities with 454 technology; adapters are ligated to DNA fragments which are attached to beads and clonally amplified by emulsion PCR. Beads are deposited on derivatized-glass flow and the sequencing is begun. The complementary primer is annealing to the adapter at its attachment to the template and nucleotides are translated by penetrating the beads in the mixtures of 5' octamer probes specified with fluorescent. Only two 3' known bases of the probes are used leading to a limited number of used fluorescent dyes, because of 16 possible combinations of the two bases. Probes are competing to anneal to the template sequences. A ligation step occurs and the unbound probes are washed. The record of the fluorescence signals is done before the cleavage of the ligated probes. A second pool of probes labelled with a second dye are added and the process is repeated. Every 4 cycles, two nucleotides are read then another cycle started after 5 bases upstream from the starting site. The first sequencing primer is discarded after 7 cycles of sequencing and the following primer started at n-1 site. A 6-days run generated around 35 bases read length sequence [44], [48], [50].

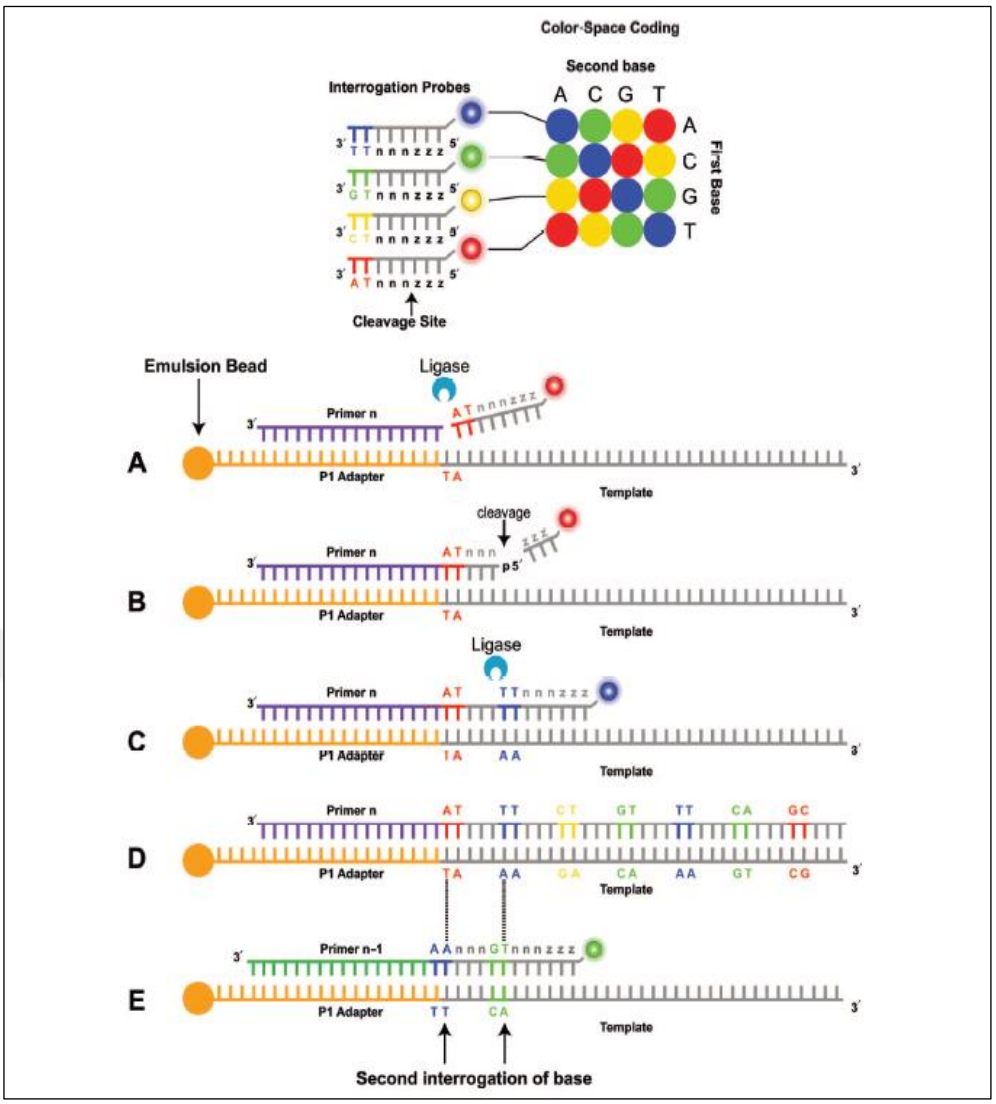


Figure 3. 6 SOLiD sequencing by ligation workflow [50]

Other NGS methodologies have appeared and have been used successfully like DNA nanoball sequencing and Ion torrent technology, however, in the recent last years, Illumina sequencing was the most successful to near-monopoly which led to consider its greatest contribution to the second-generation sequencers [40].

In the present study, Illumina sequencing, SEQ-500 platform, was used for RNA and small RNA sequencing.

3.1.3 Third generation sequencing (TGS)

A comparison between first, second and then third generation technologies lead to conclude that TGS were developed to lower the price of sequencing and to

facilitate its process. In order to reduce the reagent cost number, sample preparation and amplification by sequencing single molecule were avoided. Besides, third generation techniques use non-optical systems to improve the detection system of incorporation events (Table 3.1).

Third generation sequencing methods can be divided into three main categories: 1) Sequencing By Synthesis (SBS) technologies when a single molecule of DNA is synthesized by single molecules of DNA polymerase 2) Nanopore-sequencing technologies when single molecules of DNA are directed to a nanopore or put proximity to a nanopore, and bases are individually detected as they penetrate the nanopore 3) Sequencing technologies based on advanced microscopy techniques for a direct-imaging of separated DNA molecules (Figure 3.7) [44], [51].

Here, we are giving and example from each type of TGS.

3.1.3.1 Single molecule real time sequencing

The single molecule real-time sequencing (SMRT) was developed by Pacific Biosciences. It is the first single-molecule synthesis (SMS) among the third-generation sequencing technologies. Based on SBS, it leads to a direct observation of single DNA polymerase molecule, since it synthesizes a DNA strand and highlights the rapidity and the specific process of the enzyme to accentuate on many of the limitations of SGS. To realize the SMS, a chip with wells in the zeptoliter (10 to 21 liters) was developed. DNA polymerase molecule is immobilized on each well bottom. Once single strand DNA template is attached to the polymerase, labelled dNTPs are added and the detection system detects only the fluorescent nucleotides which were incorporated in the growing DNA strand. The attached phosphate group to the fluorescent dye causes the release of fluorescent signal from the incorporated nucleotide and signal detection is then stopped. The incorporations during second strand synthesis are measured in real time. The simultaneous addition of all four nucleotides and real-time measurement increase the speed of sequencing, compared to other sequencing technologies (Figure 3.7A) [44].

3.1.3.2 Direct imaging of DNA using electron microscopy

This technology has adopted Richard Feynman theory, when he reported that the easiest way to study nucleic acids and proteins is when you directly stare at them.

Halcyon Molecular, genome sequencing platform, has used transmission electron microscopy (TEM) to develop SMS methodology to screen and observe atoms and identifying then DNA nucleotides. Several supporting technologies were developed to improve TEM-based DNA sequencing such using functionalized needles to connect DNA molecules to a substrate for the direct imaging (Figure 3.7B)[51].

3.1.3.3 Oxford Nanopore sequencing

Nanopore sequencing depends on the passage of a DNA molecule through a pore and on the detection of bases by their effect on an optical signal or electrical conduction [48]. Nanopores are pores made of proteins such as biological channels [44]. Nanopore technology allows to detect directly nucleotides without active DNA synthesis. Actually, single-stranded DNA crosses the protein nanopore that is fixed on an electrically resistant polymer membrane. A voltage across this membrane generates ionic conduction changes that are detected and shifted by nucleotides filling the pore in real-time when the DNA molecule passes through (Figure 3.7C).

This technology has its unique advantages that does not exist in other techniques; actually, since no amplification and no ligation occur during nanopore sequencing and the input molecule is directly detected, this technology does not require enzymes from one side and does not have a length DNA limit to be sequenced from another side. However, the challenge is in library preparation which needs intact extremely high molecular weight DNA into the flow cell of the sequencer. To detect nucleotides, nanopore technology does not require imaging devices. Therefore, the weight of the device, fitting with the palm of one hand, is only 90 g and its price is much less expensive compared to other massively parallel sequencers. The main drawback of this technology is the high error rate which varies from 5-20% [44], [52].

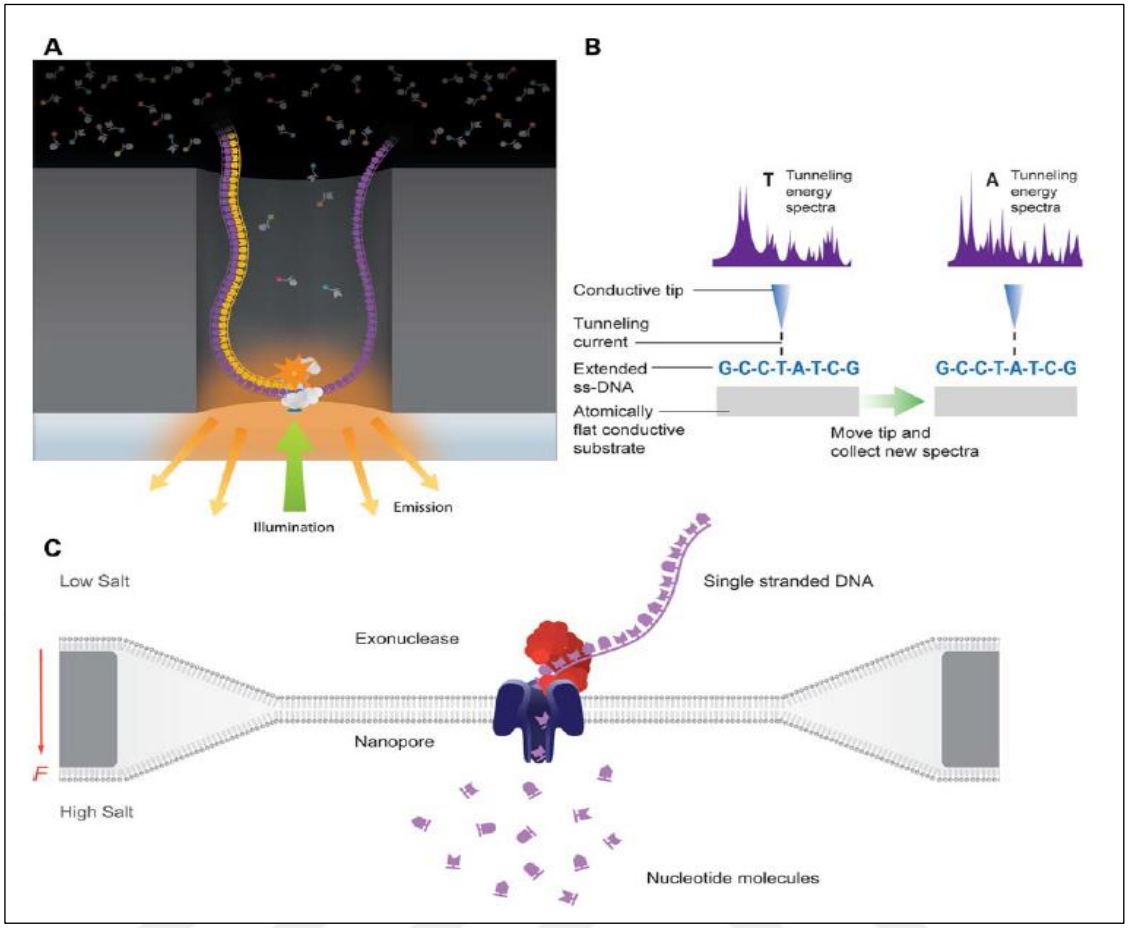


Figure 3. 7 Third generation sequencing workflow [51]

A: DNA synthesize by single-molecule real-time sequencing technology **B:** sequence DNA by direct inspection using electron microscopy **C:** Nanopore technology: measuring of the translocation of the cleaved nucleotides from a molecule of DNA over a pore, stimulated by differential ion concentrations force through the membrane.

Table 3.1 Comparison of the three generation of sequencing [51]

	First generation	**Second generation**	**Third generation**
Principal of the technology	Separation of labeled fragments based on their size	Wash-and-scan sequencing by synthesis	Single molecule real time sequencing
Resolution	Averaged across many copies of the DNA molecule	Averaged across many copies of the DNA molecule	Single DNA molecule
Raw read accuracy	High	High	Lower
Current read length	Moderate (800-1000 bp)	Short (shorter than Sanger)	> 1000 bp
Current cost	High cost per base, Low cost per run	Low cost per base, High cost per run	Low cost per base, High cost per run
RNA-sequencing methodology	cDNA sequencing	cDNA sequencing	Direct RNA sequencing
Duration	Hours	Days	< 1 day
Sample preparation	PCR amplification is not required	PCR amplification is required	Diverse

3.2 Overview of NGS applications

Over the past 20 years, NGS technologies have been used to answer biological questions in different species, including human. Different NGS approaches contributed to several revolutionary discoveries and brought innovations in the genomic research by remedying to issues related to genome, transcriptome and epigenome. The ability to deciphered and read any kind of DNA sequence is one of the distinguished advantages of NGS sequencing.

Actually, there are more than 200 applications of NGS as whole genome sequencing, chromatin immunoprecipitation sequencing, gene expression profile or small RNA sequencing besides several novel applications, which are being often published (Figure 3.8)[53].

Here we discuss the three most common applications of NGS; DNA sequencing, RNA sequencing and small RNA sequencing and we present the guidelines for data processing pipeline.

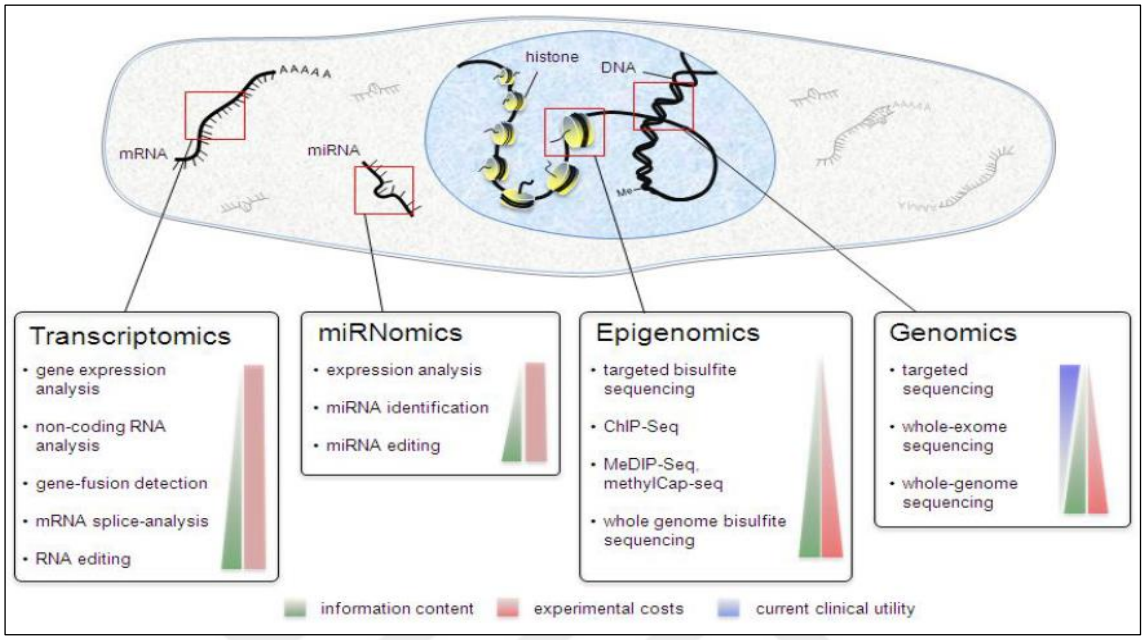


Figure 3. 8 NGS applications: Different methodologies for transcriptomic, miRNomic, epigenomic and genomic studies [54]

3.2.1 DNA sequencing

DNA sequencing is one of the first applications of NGS. It can be applied to a whole genome in order to get information about any component of the genome. Or else, on shorter sequence DNA fragment, an amplicon or on different loci especially in evolutionary studies, to focus on a specific set of genes or genomic regions [55].

The generated data is used as following:

- **Processing procedures:** Each run of NGS generates raw reads which are translated to FASTQ format files required to the following analysis. Raw images are used for clusters detecting, determination of their positions and intensity and evaluating each cluster's noise. Base-calling is the following step in which base reads sequence of each cluster are identified and filtered to remove low quality reads.
- **Reads alignment:** Also called reference-based assembly [56].

Read alignment tools are used to align huge number of reads based on a reference genome. MAQ and BFST software are used to map reads to a reference sequence [57][58]. Novoalign, a package of algorithms, is used to optimize the alignment by querying short reads with low error rate and long reads with high error rate.

- ***de novo* assembly:** *de novo* approaches are organizing short reads into contigs and aligning them into scaffolds to reassemble the original genomic DNA for other species. *de novo* assembly confronts length of short reads compared to repeats in the genome which can be overcome using algorithm extensions [59].
- **Single nucleotides:** After read assembling, pipeline analysis is done using single nucleotide variant (SNV) to identify genetic variant. Different bioinformatic tools are used like 'GTAK' process to distinguish segregating variations, 'VarScan' platform that employs statistical tests to identify SNVs and indels (insertions/deletions) and the probabilistic model 'JointSNVMix' to estimate the differences probability between different genotypes [59].
- **Structural variation detection:** Oppositely to SNVs which imply small genetic changes, structural variation includes large DNA variation (length from 1 to 3 Mb). These variations involve indels, copy-number variants, inversions and translocations [60]. They are detected by bioinformatics software such BreakDancer which provides a genome-wide screening of multiple samples and libraries and PEMer that is able to detect, simulate and annotate structural variations [59].

3.2.2 RNA sequencing: Transcriptome sequencing

High-throughput sequencing has been also employed to study the gene expression and, in this case only the transcriptome is sequenced, or to study differential gene expression and compare expression profiles by sequencing two or more groups of samples, in order to identify responsible genes for the expressed phenotype.

RNA-seq can target protein-coding RNA and non-coding RNA [10]. RNA-seq is also used to study epigenetics and identify post-transcriptional modifications [61].

RNA-seq basic applications imply transcriptome information and gene annotation including list of genes, their compositions, and their positions in the reference genome [62].

3.2.2.1 RNA sequencing general workflow

As we mentioned previously, a lot of technologies are available to perform high-throughput sequencing, however, the most frequently used one is Illumina's platform, the technology we used in the present study.

RNA-seq is typically initiated by the purification of RNAs group from a pool of total RNA. As examples, selection of polyadenylated molecules enrich mRNAs while RNAs spectrum of is enriched by rRNAs removal [63], [64].

Most of the sequencing platform require short length molecules to be sequenced. That is why and after purification, RNA is fragmented into small pieces by hydrolysis or nebulization to create a library template. Otherwise, the fragmentation can be carried out on cDNA using DNase I treatment or sonication. Most of the platforms allow the sequencing of DNA molecules, then cDNA double stranded molecules are synthesized by reverse transcriptase using RNA molecules as templates. The reaction requires primers that can be short sequences of oligo-dTs, complimentary to RNA poly(A) tail, or sequences of random 6 bases. The second type of primers is preferred because they are able to incorporate random sites throughout the length of the RNA molecule [65]. Double-stranded cDNAs, under the action of specific enzymes, engender blunt ends, both ligated to adapters (Figure 3.9)[66]. The fragmentation step generates molecules of different length and the molecules of the desired length is purified by gel extraction. Annealed primers to the adapters of the purified cDNA molecules are used then for a PCR reaction. Before sequencing, denaturation of double-stranded molecules into single-stranded ones occurred bypassing the flow cell with immobilized adapters-complementary oligo sequences.

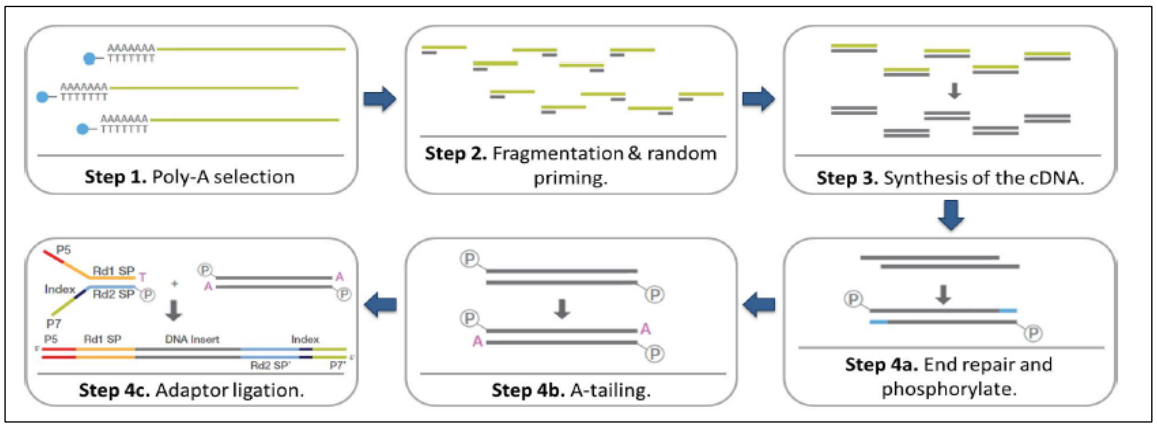


Figure 3. 9 Creation of library template steps [66]

Bridged amplified single-stranded DNAs create clonal clusters of identical molecules inside cell flow, serving for templates to create the complementary stands (Figure 3.10) [67]. Sequencing primers hybridize with every molecule of the clonal clusters, then DNA templates are reversely complemented used fluorescent-labelled nucleotides. This process is called sequencing by synthesis (Figure 3.11). Every incorporated nucleotide enhances a laser excitement which lead to a fluorescence corresponding to the last added base. When the fluorescent died, the terminal group is eliminated from the novel nucleotides and the operation is repeated for the required times (from 30 up to 200 times). A sequence of image is the out-put of this process; One image sequence for each added nucleotide giving the fluorescence of each cluster, and each light color represent a base type.

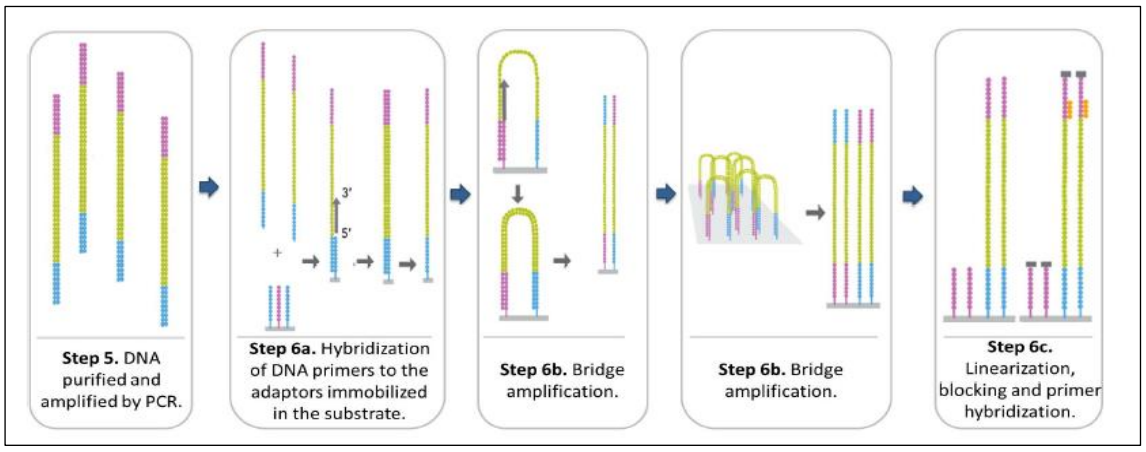


Figure 3. 10 Cluster generation steps [67]

Then, the combination of the sequence of images gives the sequence of the nucleotide in each cluster. The information is afterwards saved in FSTQ file for each cluster, called reads. Each read contains a unique ID read identification, raw nucleotides sequence and every base quality value in the read, called Phred quality score Q (Q= -10log10(P), (P = probability of the base call to be incorrect)[68], [69].

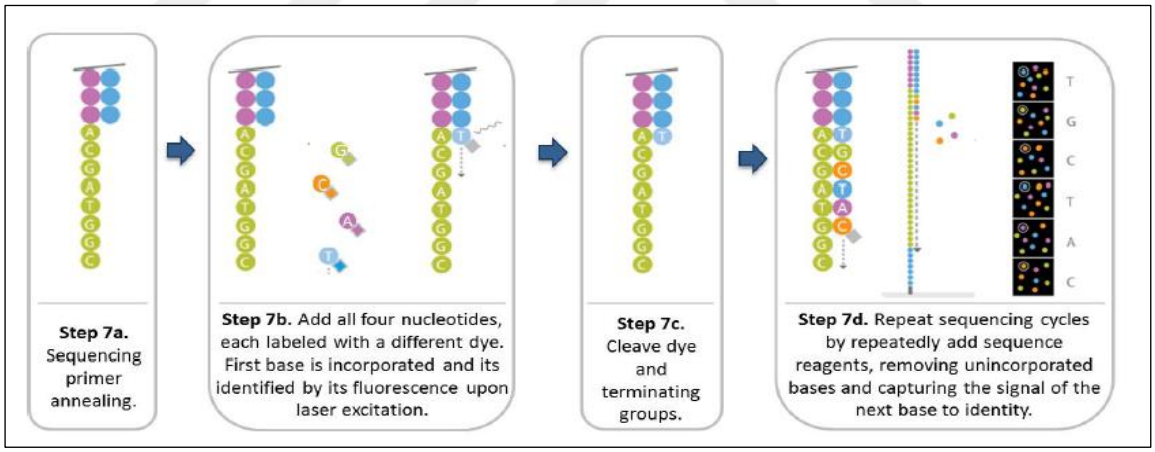


Figure 3. 11 Sequencing by synthesis process [68]

3.2.2.2 Reads mapping strategies

RNA sequencing has several objectives, but most of the studies aim to estimate the expression of a specific region of the genome, that might be genes, exons, splice junction, isoforms or even new transcribed regions.

After identifying the features presented in the sequencing library, reads mapping are employed for *de novo* transcriptome assembly, variant calling and transcriptomic epigenetics and analysis of expression profiling [59].

- ***de novo* transcriptome assembly:** By exploiting reads, *de novo* assembly aims to determine a set of contigs defined as longest contiguously expressed regions, using algorithmic strategies such as pree-fix tree, overlap-layout-consencus and the most common one, Brujin graph which has been used by several assembling programs. The annotation of the long-assembled RNA contigs is done to similar species in order to investigate the genome by the mean of Basic Local Alignment Search Tool (BLAST), and can be considered as a reference for additional abundance profiling [59], [70].
- **Expression profiling analysis:** Main applications of sequencing of RNA are gene expression levels profiling and identifying differentially expressed transcripts.

  Expression profiling analysis includes mapping reads procedures based on reference, determining reads for each transcript and some statistical tests for differentially expressed ones. Short reads are aligned oppositely to sequences of a reference in FASTA files like annotated genome sequences from NCBI or RNA transcripts de novo assembled from new species. Some alignment computer software that has been developed, like Bowtie and Bowtie2 are able to find longer alignments. TopHat adds to Bowtie2 the ability to integrate the known splice junctions and to identify novel ones. Transcripts raw expression levels are presented by read counts and used for differentially expressed statistics tests such as DeSeq. Differentially expressed transcripts are analyzed for gene ontology (GO) and pathways enrichment (KEGG) [59], [70].
- **Variant calling and transcriptomic epigenetics:** RNA sequencing is an effective way to identify coding variants. Actually, certain studies succeeded to provide variants in vertebrates from RNA-seq data. SAM tools determines calling variants by alignment to reference sequences [71]. The

mutation mapping analysis pipeline for pooled RNA-seq (MMAPPR) have been proposed recently [72]. Some new research areas use RNA-seq to transcriptomic epigenetics studies like transcription start site-associated RNAs [73], promoter-associated RNAs [74], transcription-initiation RNAs (tiRNAs). However, they require more bioinformatic and biostatistical input [59]

3.2.3 Small RNA sequencing

Small RNA sequencing can be considered as a special protocol for deep sequencing of small non coding RNAs [75].

Because of their crucial role in gene expression post-transcriptional regulation, studying small RNAs (miRNAs, siRNAs and piRNAs) and especially sequencing has become a way for sRNA discovery and sRNA profiling.

3.2.3.1 Small RNA seq general workflow

Because of the existence of different classes of small non-coding RNA, specific small RNA-seq protocols exist. The workflow differs based on the aim and on the sequencing platform. However main steps are generally followed as it mentions in the Figure 3.12.

Raw cDNA reads libraries are obtained after the sequencing process. Sequencing reads with adapters are removed with the help of a toolkit provided by the sequencing platform like Flicker of Illumina or a third-party toolkit like FASTX toolkit. Reads are then filtered and low-quality reads are castoff using specific tools, as the FASTX toolkit FASTQ quality filter, or NGS QC toolkit. The filtered reads must be verified by alignment to a genome database reference. Generally, Bowtie and Bowtie2 are used for short read alignment since their application give optimized and memory-efficient algorithm, and produce a lot of built-in-indexes for the reference of the genome database [76].

In sRNA-seq, some databases are frequently used; Rfam, the open access database where data about non-coding RNAs, like snoRNA, rRNA and tRNA, is found [77]

and helps to identify known sRNA reads, and miRbase database is also often used since it has sequences and annotations of all miRNAs of all species [78].

3.2.3.2 Read mapping applications

- **Small RNA prediction:** Identification of novel sRNAs was easier by NGS technology. Recently, several algorithms have been developed to predict miRNAs, which are applied in tools like miRTRAP [79], miRExplorer [80] and miRDeep [81]. Most of these tools were developed for animal species however miRDeep-P was proposed for plants [82]. Prediction tools continue to make progress and there is no most-preferred tool, every tool can be used based on the aim of the study [59].
- **miRNAs characterization:** miRNAs profiling is another major sRNA sequencing application. miRNAs are important gene regulators and have large range of functionalities such as marker for tissues or stage of cell development or diseases [83]. It was also reported by [84] that miRNAs have been used for drug development. The sRNA-seq is widely used for miRNAs characterization, by providing high precision to distinguish miRNAs with closed sequences like isomiRs and by identifying novel miRNAs at the same time [59].

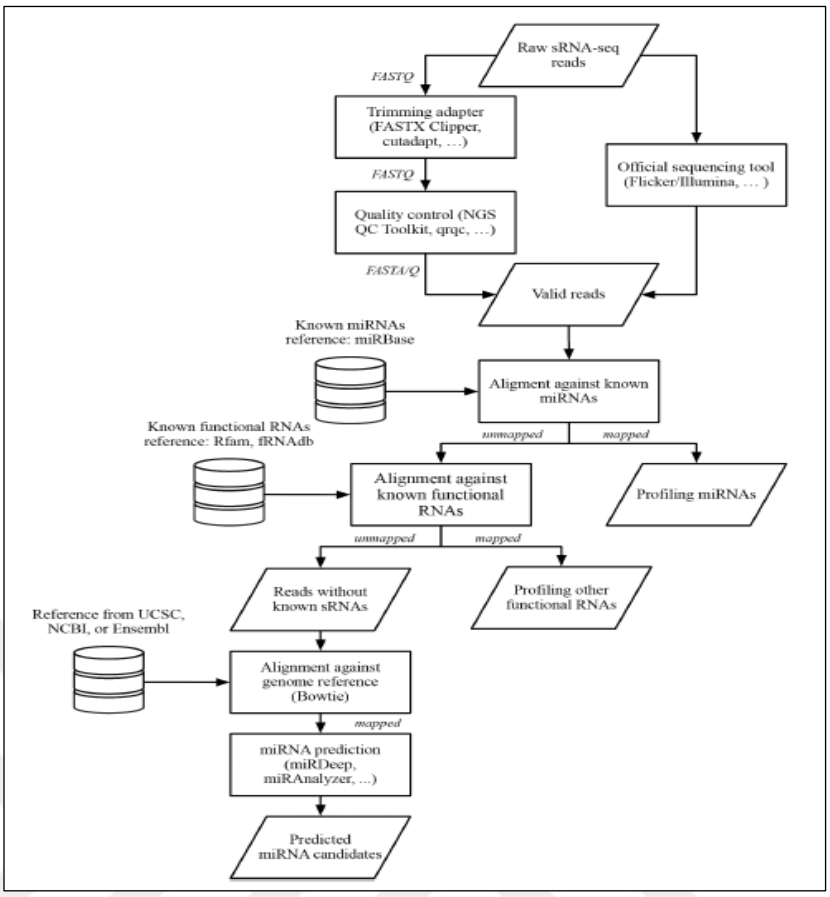


Figure 3. 12 Small RNA sequencing workflow [59]

High throughput sequencing or next generation sequencing have been applied for different purposes among which gene expression analysis.

Gene expression analysis allows the comprehension of the transcriptomic behaviors of any biological system. It can be done for several aims such as identifying differentially expressed genes between tissues or different experimental conditions like the first part of the present study [85], to analyze heterogenous diseases like cancer [86], to diagnose some diseases or even to found out new drugs [87]. Gene expression analysis can also be employed to produce a database that gives information about living mechanisms. Expression of genes of interest can be analyzed by genome-wide approach such as RNA-seq.

Actually, in case the genes of interest are not known RNA-seq is used for profiling cell transcriptome which can be very helpful to provide the connection between the encoded information and the presented phenotype [10]. Moreover, small

RNAs, particularly micro RNAs, have an essential role in studying gene expression. That is why with the progress of high throughput sequencing technologies and the availability of some species genome sequence such *A. thaliana*, studying gene expression remain possible also by identifying and discovering miRNAs and their target genes.

4

GENE EXPRESSION

4.1 Concept of gene expression

In 1958, Francis Crick presented molecular biology central dogma. He stated that the central dogma 'deals with the detailed residue-by-residue transfer of sequential information' and explained notably that the information carried by nucleic acids might be conserved or converted to proteins. However, their conversion is irreversible (Figure 4.1). Actually, the mechanism during which a specific strand of deoxyribonucleic acid (DNA) is reproduced into a ribonucleic acid (RNA) that itself will be the template to synthesize a functional gene product, is called gene expression [88].

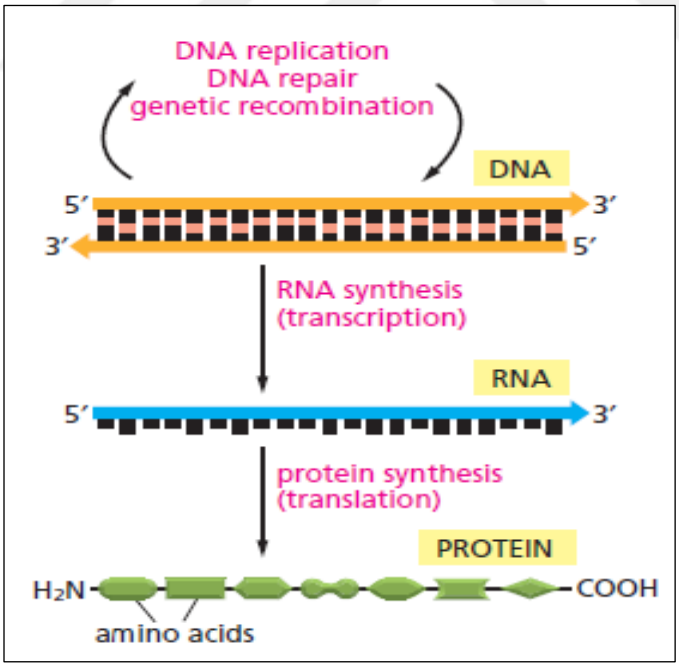


Figure 4. 1 Central dogma of molecular biology [89]

Gene expression process starts with the transcription that occurred in the cell nucleus. It represents deoxyribonucleic acid (DNA) transformation into a ribonucleic acid (RNA) with the help of RNA polymerase (RNA pol) enzyme by catalyzing a phosphodiester bonds to associate the nucleotides together and shape RNA molecule backbone.

In eukaryotic cells, mainly three types of RNA polymerase exist; RNAs pol I catalyze the transcription of ribosomal RNAs, RNAs pol II catalyze principally the transcription of protein-coding genes (mRNA), and RNAs pol III that catalyze transfer RNA (tRNA) transcription. All these RNA polymerases ensure the synthesis of small RNAs with important catalytic and structural roles in the cell. Gene transcription consists of three main steps: Initiation, elongation and termination.

Initiation step starts with the binding of RNA pol to the 5' end region of the DNA strand where the promoter is. RNA pol needs the help of transcription factors (TFs) which must be gathered around the promotor with RNA pol before the transcription. TFs are released from the DNA molecule after initiation. The double-stranded DNA molecule is unwinded and RNA pol scan the strand of the template and adds nucleotides of the elongated RNA molecule.

The termination step occurs when RNA polymerase arrives at a termination site and separated from the DNA and RNA is then cleaved and come out of the transcriptional complex. The produced RNA molecule is processed depending on its type during the transcription. Polyadenylation is a crucial processing feature for the transcripts which will be a messenger RNA (mRNA), also other processing steps are required to make from the newly synthesized RNA molecule a functional one. Actually, in eukaryotic cells including plant cells, mRNA is represented by altering genes coding parts which are the exons and gene non-coding portions which are named the introns. A splicing process detached the introns from the mRNA inducing exons to be stick together. The mature RNA is transferred from the nucleus to the cytosol, where mRNA is subjected to translational process and transformed into a protein; ribosomes use a mRNA molecule as a template, to

match each codon composed from three nucleotides on the mRNA with an anti-codon of three complimentary nucleotides sequence on a tRNA molecule. The tRNA plays the role of a translator of RNA nucleotides sequence into amino acids in proteins. The mRNA moves through ribosome that links every amino acid by peptide bond to the end of a growing polypeptide chain. The completion of the translation occurs when the mRNA comes to a termination codon. The protein is then released with the mRNA from the ribosome that is divided into two separated subunits.

Concisely, gene expression is the mediation translating the genetic information (genotype) to a visible manifestation (phenotype) through gene transcription and mRNA processing (Figure 4.2) [89].

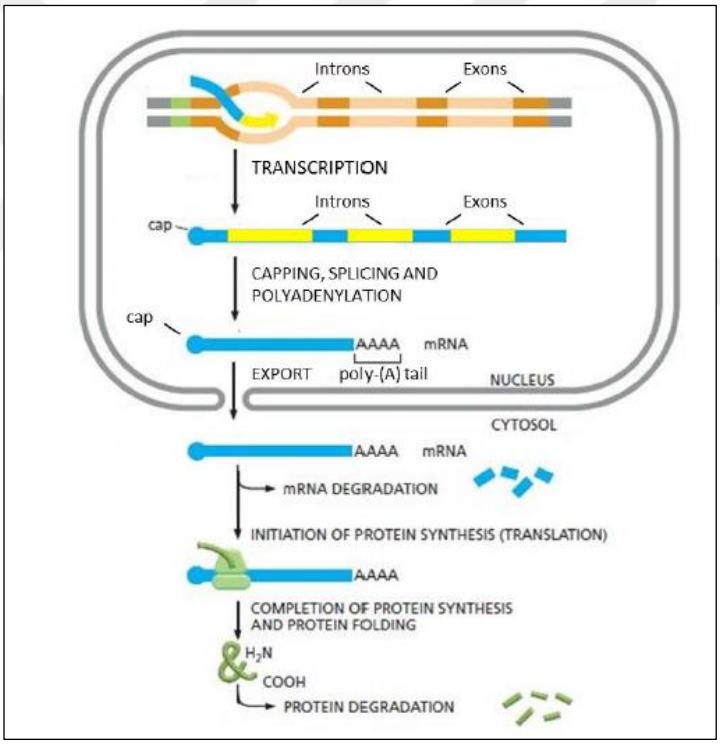


Figure 4. 2 Summary of gene expression process (adapted from [89])

4.2 Regulation of gene expression

Genes encode for proteins that guide the structural and functional properties of the cell. All steps in the process of the information translation from DNA to RNA

to protein, contribute to the cell functioning self-regulation. This system enables cells to face the variations of the environment and keep their specific expression patterns. Cell regulatory system ensures:

- Controlling the time and the rate of the transcription of a gene;
- Controlling the processing of the transcript;
- Controlling the rate of transcript degradation;
- Selecting the exported mRNA from the nucleus to the cytosol;
- Selecting the translated mRNAs in ribosomes;
- Controlling activated and inactivated proteins after synthesis;
- Selecting degrading proteins (Figure 4.3) [89]

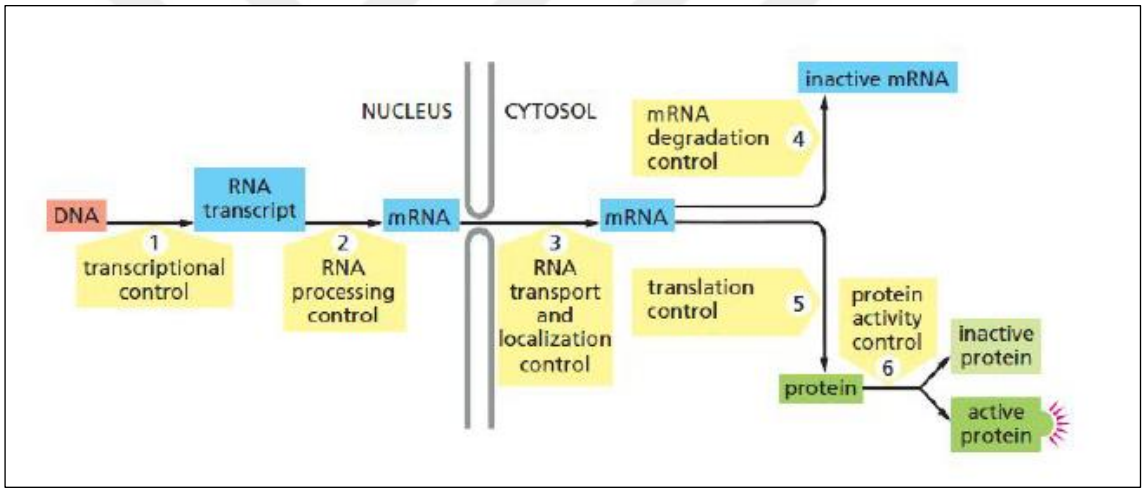


Figure 4. 3 Gene regulation ways in Eukaryotic cell [89]

4.2.1 Transcriptional regulation

Transcriptional regulation plays a crucial role in controlling the gene expression. It represents the first step in the gene expression process. The regulation is performed at the promoter level (Figure 4.4).

TFs are associated with the promoter, influence the binding of RNA polymerase and consequently regulate the initiation of the transcription. Indeed, transcription activity of basically all genes are controlled by regulatory DNAs regions, called enhancers or silencers, that increase or decrease the transcription activity by

switching, respectively, on or off the gene. Specific sequences of TFs are bound to these regulatory regions, assembled in the promoter region and interact with other proteins called co-factors. Gene transcription regulation rate is accomplished by helping or inhibiting aggregation of the TFs and RNA polymerase at the promoter [89], [90], [91]. Moreover, the DNA chain needs to be accessible so that TFs bind the regulatory regions, and the RNA pol can start its polymerization reaction. Then the transcriptional activity is affected by the DNA packaging level.

DNA is packed with histones to form the chromatin. Chromatin has the nucleosome as the fundamental subunit where the transcription might be prevented if the gene promoter region is blocked. Therefore, RNA pol II activity rate decreased until it stopped at the proximal position of the promoter. Depending on the transcription and elongation factors type, the elongation phase starts [92], [93].

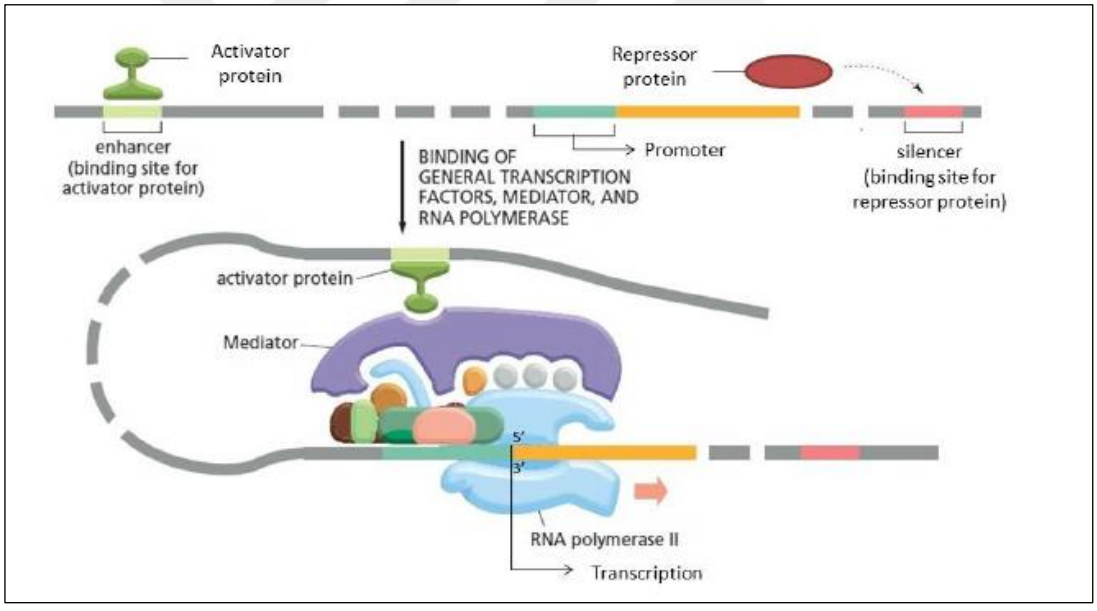


Figure 4. 4 Schematic presentation of the regulatory regions controlling transcription initiation in eukaryotic gene (adapted from [89])

4.2.2 Post-transcriptional regulation

Post-transcriptional regulations of gene expression take place at the levels of pre-messenger RNA (mRNA) processing, mRNA stability, and mRNA translation. They are defined as all regulated steps from mRNA molecule to the synthesis of protein

which are; mRNA processing and stability, compartmentalization and translational regulation (Figure 4.5) [94].

The polyadenylation process protects the transcripts from degradation and helps in their exportation to the cytosol. Similarly, the capping process, which consists of adding a modified guanine nucleotide cap to the 5' end of pre-mRNA molecule, is important to get new transcripts out of the nucleus of the cell. Hence, the two processes are crucial for the mRNA molecule stability. In addition, the process of splicing participates to produce proteins from the same transcripts source gene. Gene parts inclusion or exclusion allow a single coding gene to produce different transcripts and increase functional gene products number. Mature mRNAs resulting from this process are called isoforms.

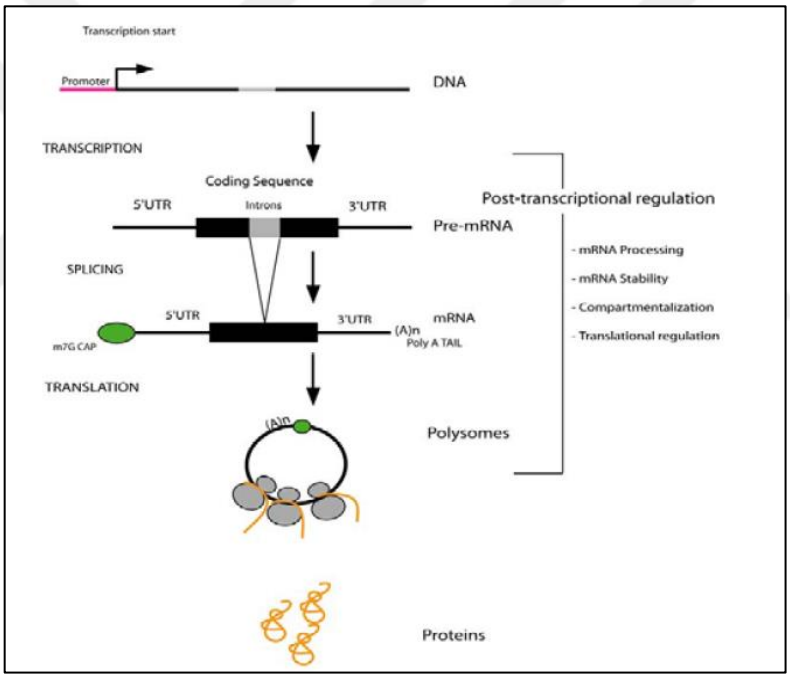


Figure 4. 5 Gene expression post-transcriptional regulations [94]

The mRNA degradation can be considered as a defense mechanism from different RNAs such as viruses, but it is also a way to control the stabilization and the translations of mRNAs. Micro RNAs (miRNAs), small interfering RNAs (siRNAs) and piwi-associated RNAs (piRNAs) are three classes of small regulatory RNAs (20 - 30 nucleotides) which currently appeared to be eminent mRNAs regulators [95].

The small RNAs three species are different in their biogenesis, act in different biological pathways and have different regulation ways to their targets. However, they have the same functional form by being part of protein RNA complexes associated to Argonaute protein family members. The protein RNA complexes target RNAs that can be completely or partially complementary to the small RNAs and the RNAs translation is then repressed or degradation of targets themselves occurred [96]. Certain mechanisms inhibiting protein degradation, translation initiation or elongation, or early termination have been proposed [97]. For instance, the mechanism of silencing by degradation. Actually, in plants, a miRNA can be totally complementary to its target(s), which lead to a transcript endonuclease cleavage by Argonaute protein followed by its degradation. The degradation mechanism is characterized by the development of the RISC complex which is composed of a small interfering RNAs (siRNA) associated with a group of proteins. RISC complex inspects the cytoplasm in order to find complementary mRNAs to the siRNA they carry. A molecule of mRNA is destroyed in case of carrying a sequence, complementary to the siRNA incorporated RISC complex [89]. Based on computational predictions and miRNA targets genome-wide screens, the majority of mammalian transcriptomes (up to 50% of human protein-coding genes) are exposed to the regulation by up to 500 genomes encoded miRNAs, despite the fact that the proportions of degradation or translational silencing regulation remain imprecise [97]. However, siRNAs can be derived from external or internal sources, and will always be complementary in total to their target(s) causing a cleavage. The last discovered and most unknown class of the regulatory sRNAs are piRNAs. Their function seems to be mainly the oppression through transposable elements that modify chromatin. In addition to the transposable elements silencing, several piRNAs were reported to target protein-coding genes causing their degradation. However, this type of regulation is still not elucidated and restricted to a few cases that have so far been observed in vivo [98].

5

MATERIALS AND METHODS

5.1 Materials

5.1.1 Chemicals

Chemicals and kits (Table 5.1) used in the present study are listed below

Table 5.1 Chemicals, solutions, and kits used

CAS Number	Company	Product Code	Name
9002-18-0	Sigma	A1296	Agar
9012-36-6	Sigma	A9539	Agarose
64-19-7	Merck	1000632500	Acetic Acid
115-39-9	Sigma	B0126	Bromothymol Blue
67-66-3	ISOLAB	910.033	Chloroform
78574-94-4	Bio-norm	-------	Cycloastragenol
1609-47-8	Sigma	D5758	Diethyl Pyrocarbonate (DEPC)
94-75-7	Sigma	D70724	2,4-Dichlorophenoxyacetic acid
64-17-5	ISOLAB	920.026	Ethanol Absolute
1239-45-8	SIGMA	E7637	Ethidium Bromide
60-00-4	SIGMA	E9884	Ethylenediaminetetra-acetic acid (EDTA)
43-688-14	Applied Biosystem (Thermofisher scientific)	007-43367	High-Capacity cDNA Reverse Transcription Kit
67-63-0	Sigma	24137-M	Isopropanol or Isopropyl Ethanol

Table 5.1 Chemicals, solutions, and kits used (cont.)

0118-CC-342	Hibrigen	MG-LDR-100	Ladder (100 bp DNA Ladder)
NA.72	Sigma	M3900	Murashige and Skoog Vitamin Solution
MFCD00240976	Sigma	M5519	Murashige and Skoog Basal Salt Mixture
1310-73-2	Merck	1.06462.1000	Sodium Hydroxide
57-50-1	Sigma	M84097	Sucrose
0920-UB-961	Hibrigen	MG-SYBR-01	2x Syber Green qPCR Master Mix
43-665-96	Applied Biosystem (Thermofisher Scientific)	-------	TaqMan™ MicroRNA Reverse Transcription Kit
77-86-1	Sigma	648310-M	Tris Base
15-596-026	Invitrogen	15596026	Trizol®
121B-BM-519	Hibrigen	MG-KTAG-01	Taq DNA Polymerase

5.1.2 Devices

Devices used in the present study are listed below (Table 5.2):

Table 5.2 Experimental devices

Name	**Company**	**Series/Code**
Analytical balance	SHIMADZU	TX4202L
Autoclave	Nüve	OT 012
Automatic pipettes	Eppendorf	2.5μl/20μl/200μl/1000μl

Table 5.3 Experimental devices (cont.)

Growth chamber	Panasonic	Fitotron
Cabinet (Hood)	Telstar	Bio II Advance / Tel_9009
Electrophoresis power source	VWR	300 V
Gel imaging system	SYNGENE	SYNGENE-G: BOX-F3
Vortex	Labnet	C1301P-230V
Incubator/ Thermo Shaker	BIOSAN	TS-100
Ice maker	----------	IMS-50
Microcentrifuge	Hettich	Micro120
Centrifuge	NÜVE	NF 800R
ScanSpeed mini	Labnet International	SN 1204
Nanodrop spectrophotometer	Thermo Scientific	Nanodrop 2000
pH- meter	Mettler Toledo	FiveEasy
Thermocycler / PCR	Bio-Rad	T100™
Thermocycler / qRT - PCR	Bio-Rad	CFx96 Touch
Water purification system	Millipore	Direct Q3
Hot air oven	WiseVen®	WOF-155
Hotplate stirrer	WiseStir®	MSH-20A
Freezer 2	Arçelik	----------
Freezer 1	Regal	----------
Deep freezer (-86°C)	HAIER	DW-86L388
Microwave	Arçelik	MD 574 S

5.1.3 Plant material

Wild-type seeds (Colombia Col-0) from the model plant *Arabidopsis thaliana* were used in the present work. They were provided from Dr Ralph Stracke (Bielefeld University, Biotechnology Center, Genome Research Department).

5.2 Experimental methods

5.2.1 Plant tissue culture and cycloastragenol application

5.2.1.1 Media preparation

Murashige and Skoog (MS) medium [99] was used for *A. thaliana* tissue culture studies. To prepare 500 ml of MS medium, 2.2 grams of Murashige and Skoog Basal Salt Mixture and 15 g of sucrose as a carbon source were dissolved in 400 mL of distilled water. The pH was adjusted using 5M NaOH to 5.7, then 4.5 grams of agar were added and the solution volume was adjusted up to 500 mL by adding 100 mL of distilled water. Medium was sterilized for 15 min under a pressure of 1.2 atm and a temperature of 121°C. After cooling, 0.5 mL of MS vitamin solution was added and the medium was poured into 25mL/9 cm of diameter Petri dishes.

5.2.1.2 Surface sterilization of *A. thaliana* seeds

Around 5 mg of *Arabidopsis* seeds were put in 1.5 mL column tube and then sterilized as following:

First wash: for three times, 1ml of 70% Ethanol was added to the tube, and then removed by centrifugation for 2 min at 600 rpm using ScanSpeed mini.

Second wash: 1ml of 100% Ethanol was added and then removed by centrifugation for 2 min at 600 rpm using ScanSpeed mini.

Seeds were after dried for 20-30 min in laminar air flow in sterile conditions under the cabin.

5.2.1.3 Cultivation of *A. thaliana* plants

Sterile seeds were sown in 9 diameter Petri dishes containing around 30 mL of MS medium and incubated in plant growth cabin for 10-12 days under 1400 lux fluorescent light, altering between 16h light and 8h dark at a same temperature of 25°C until seedlings grew-up. Seedlings were then transferred into test tubes containing MS medium for 21 days more. Grown plants were used in callus formation.

5.2.1.4 Preparation of cycloastragenol solutions

The commercial cycloastragenol product TA-65 (Table 5.1) was dissolved; a stock solution of 10 mM was prepared and stored at -20°C. Then, besides control concentration (0 μM), 1 μM and 10 μM solutions were prepared from the stock solution and applied to *Arabidopsis* calli.

5.2.2 Callus culture

Roots were used as explants for callus culture. Excised from plants and surface sterilized, explants were afterwards aseptically cut into small segments (few millimeters in size). Roots were then transferred on MS medium containing 1mg of 2,4-dichlorophenoxyacetic acid (2,4-D).

Each Petri dish was containing around 10 explants of 1-2 cm in length. The obtained calli were transferred into fresh medium containing 1mg of 1-2, 4-D and cycloastragenol (CAG) in three biological repetitions; three different concentrations: Control (0 μM), 1 μM, and 10 μM. Each concentration was applied in two series, subdivided into three technical repetitions; six months, nine months and eleven months old calli. Calli of each Petri dish have been sub-cultured and weighed every three weeks.

5.2.3 Growth Index measurement

Based on cycloastragenol concentration, Petri dishes were divided into three biological repetitions; Control (0 μM), 1 μM and 10 μM as experimental groups. Every group was carried out in two series of three replicates, and the callus weight

was measured before and after treatment in order to calculate Weight Index based on the below equation [100];

$$\% \text{ Growth } = \frac{\text{Wf} - \text{Wi}}{\text{Wf}} \times 100 \qquad (\mathbf{5.1})$$

W_f: Final Weight

W_i: Initial Weight

5.2.4 RNA isolation and quality determination

5.2.4.1RNA Isolation

Total RNAs were isolated from calli tissue in order to analyze CAG effects and changes occurring on the gene expression level. RNAs were extracted using Trizol reagent. Around 100 mg of each callus were crushed in a cold mortar after adding liquid Nitrogen until obtaining of a powder. One mL of Trizol was added to the powdered samples and transferred to a 1.5 mL microcentrifuge tubes. Samples were then incubated for 5 min at room temperature.

After incubation, 0.2 mL of chloroform was added to each sample. Microcentrifuge tubes were gently inverted and mixed for 15 sec. Tubes were incubated at room temperature for 5 min and then centrifuged at 4°C, 12000 x g' for 20 min. Supernatant of each sample was afterwards withdrawn and transferred into a new 1.5mL microcentrifuge tube to which 500 μL of isopropanol was added. Tubes were gently mixed and incubated for 10 min at room temperature. They were then centrifuged at 4°C for 10 min at 12000 xg'. After centrifugation, RNA was observed as a white pellet at the bottom of the tube. The supernatant was discarded and 1 mL of 75% ethanol was added to each pellet in order to wash it by centrifugation for 5 min at 7500 xg' at 4°C.

The ethanol was discarded and the RNA pellet was left to dry around 10-15 min at room temperature. Fifty milliliters of DEPC-dH_2O were added to the dried RNA pellets.

Samples were incubated on the hotplate stirrer at 37°C for 20-25 min to allow RNAs to be dissolved.

5.2.4.2 Determination of RNA concentration

After the dissolution of RNA samples in DEPC-dH_2O, purity and concentration (µg/ml) of each sample were determined by measuring the absorbance values at 260 and 280 nm wavelengths according to the equation below.

Measurements were performed using Nanodrop spectrophotometer.

$$\text{RNA (µg/ml)} = \text{A260} \times \text{dilution coefficient} \times 40 \qquad (\mathbf{5.2})$$

5.2.4.3 Gel Electrophoresis

Agarose gel electrophoresis technique aims to determine the quality of the extracted RNAs, while the Nanodrop spectrophotometer determines their concentrations. A 0.8% agarose gel was run to check RNAs integrity and visualize the ribosomal bands to control total RNA quality. The gel was prepared by dissolving 8 g of agarose in 100 mL of 1x TAE buffer in microwave for approximately 2 min. After cooling, 3 μL of ethidium bromide was added to the gel which was then poured into a gel dock with a comb and allowed to solidify for around 20 min at room temperature.

Samples were loaded on the separated wells after mixing 5 μL of each RNA sample with 1 μL of loading buffer. The gel was run for 20 min at 90 V, and then exposed to UV light to be visualized on the gel imaging system.

5.2.5 High Throughput Sequencing (NGS)

After checking RNA samples quality, based on the growth index results and on [101], a study during which the same calli samples were used, 1 μM was chosen to be the optimum CAG concentration to affect our plant samples. Two samples of RNA, Control and 1 μM of CAG, were sent to Beijing Genomics Institute (BGI, China) for RNA-sequencing and small RNA sequencing. Each sample was a pool of nine RNAs from the two different concentrations of 9 months old calli.

5.2.5.1 Transcriptomic analysis: RNA-sequencing

- **Transcriptome sequencing study process**

This part is divided into a molecular experiment workflow and a bioinformatics workflow. Actually, total RNA Quality control (RNA concentration, RIN value, 28S/18S and the fragment length distribution) was done using Agilent 2100 Bioanalyzer (Agilent RNA 6000 Nano Kit NanoDrop™) was used to identify the purity of the RNA samples.

The molecular experiment workflow, corresponding to mRNA fragmentation, cDNA libraries construction and sequencing, was performed as it is shown in the Figure 5.1; Poly(A)-containing mRNA molecules were purified employing poly(T) oligo-attached magnetic beads and the fragmentation of the purified mRNAs into small pieces was performed under high temperature using divalent cations. The obtained RNAs were used as template to synthesize cDNA using DNA polymerase I and Rnase H. The cDNA fragments have an additional single 'A' base and subsequent ligation of the adapter. The products were purified and enriched with PCR amplification and after quantification by Qubit, they were pooled together to make a closed circular DNA circle (ccDNA circle) giving the final library.

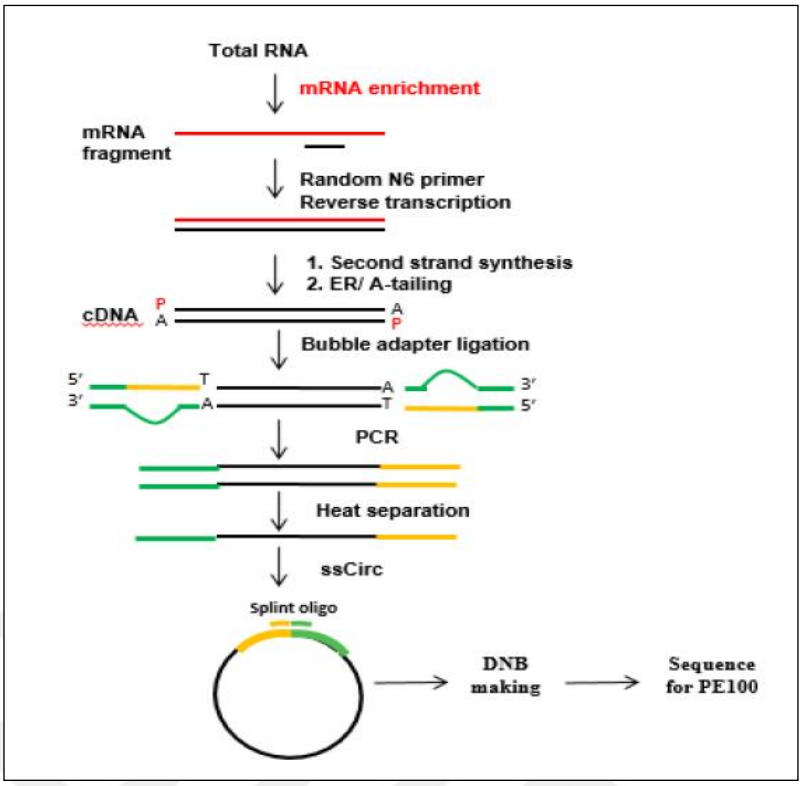


Figure 5. 1 Transcriptome experimental workflow

The ssDNA circles were used to generate DNA nanoballs (DNBs) by Rolling Circle Replication (RCR) in order to have intense fluorescent signals during sequencing process. The DNA nanoball-based nanoarrays and stepwise sequencing were then combined using Combinational Probe-Anchor Synthesis Sequencing Method and the DNBs were loaded into the patterned nanoarrays. The 100 bp pair-ended reads were read through on the BGISEQ-500 platform for the following data analysis study.

The bioinformatics workflow is summarizing in the Figure 5.2. Consequently, the low-quality reads were firstly filtered then clean reads were mapped onto reference genome followed with novel gene prediction SNP and INDEL calling and gene-splicing detection. Differentially expressed genes (DEGs) between samples were finally identified and clustering analysis in addition to the functional annotations were done.

- **Sequence reads filtering**

More than 20% with lower than 10 quality bases, the reads with adaptors and those with unknown bases (N bases more than 5%) are considered low-quality reads and filtered using the software **SOAPnuke** (version: 1.5.2; parameters: -l 15 –q 0.2 –n 0.05) to clean the data. The data was then stored in FASTQ format.

Figure 5. 2 Transcriptome resequencing analysis pipeline

- **Genome mapping**

Clean reads were mapped onto reference genome using **HISAT** (Hierarchical Indexing for Spliced Alignment of Transcripts) software (Version: v2.0.4 and Parameters: --phred64 –sensitive –no-discordant –no-mixed –I 1 –X 1000), which is a fast, sensitive and a high accuracy one. The mapping was done as it is shown in the Figure 5.3.

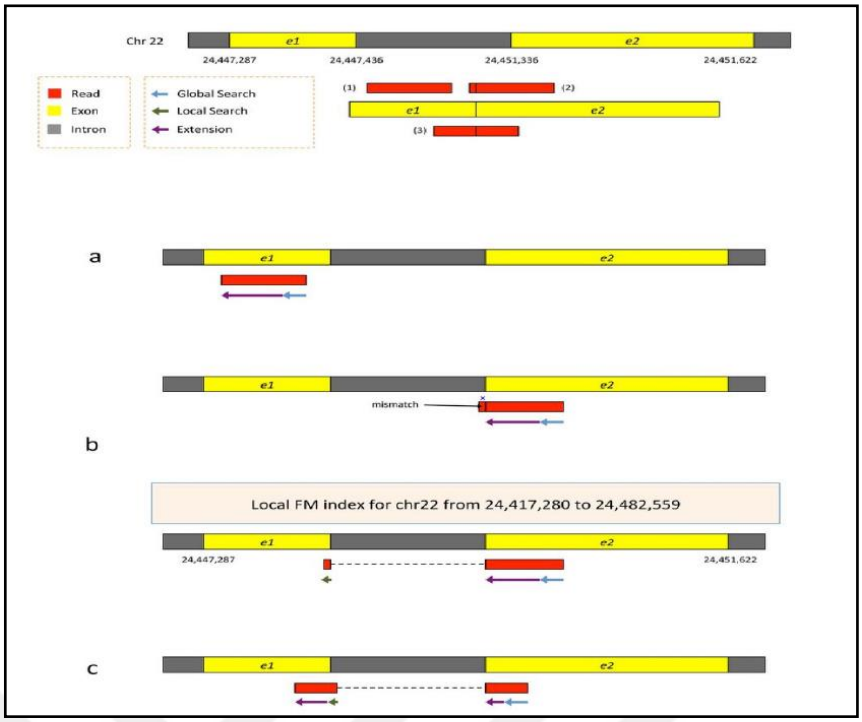


Figure 5. 3 HISAT mapping presentation

- **Novel transcript prediction**

StringTie (Version: 1.0.4 and Parameters: -f 0.3 –j 3 –c 5 –g 100 –s 10000 –p 8), fast and accurate software for transcriptome assembly, was used to recreate the transcripts. **Cuffcompare** (Cufflinks tools) (Version: 2.2.1 and Parameters: -p 12) was employed to compare the transcripts to the reference annotation which is presented by *A. thaliana* known genes, as it explained in Figure 5.4, then 'u', 'i', 'o', 'j' class code types, were selected, as novel transcripts (Table 5.3).

CPC software (Version: v0.9-r2; Parameters: Default) predicts the coding of new transcripts, assembled with the reference transcripts and the complete reference is used to a downstream analysis.

Table 5.4 Explanation of class codes

Class Code	Explanation
u	Unknown, Intergenic transcript
i	A transfrag falling entirely within a reference intron.
o	Generic exonic overlap with a reference transcript
j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript.

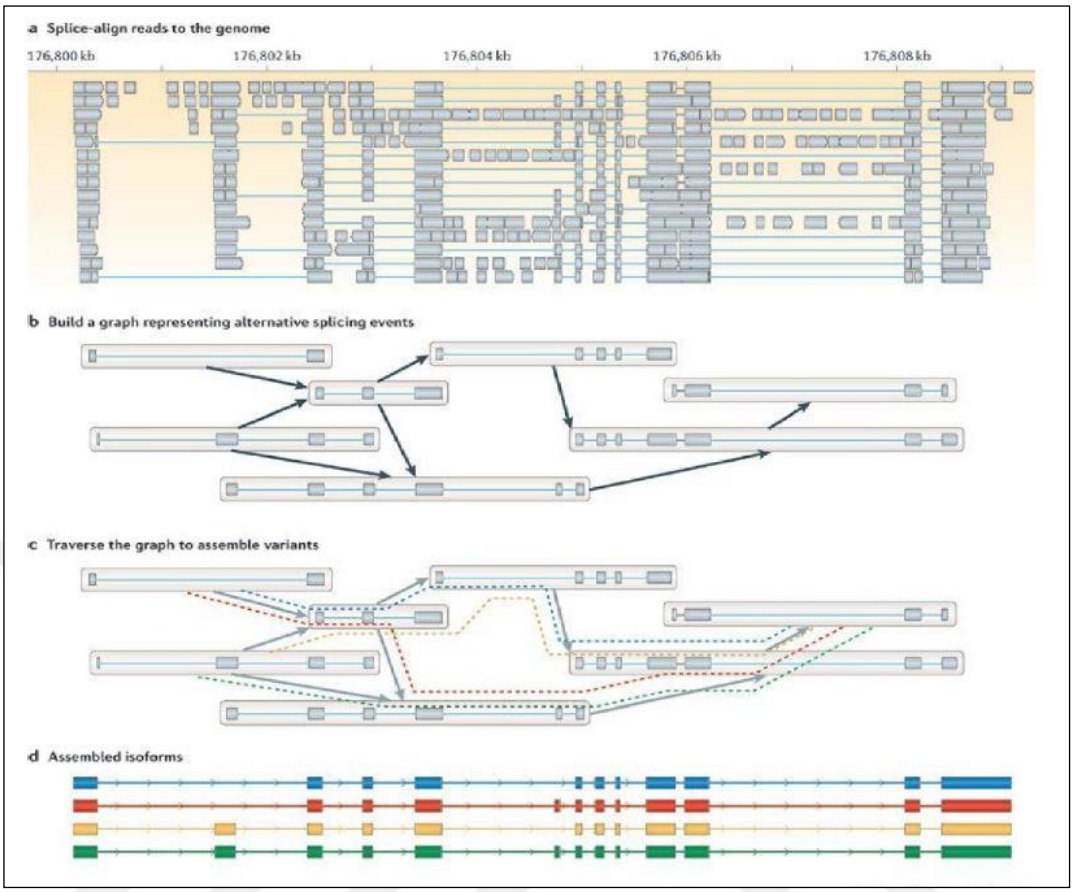


Figure 5. 4 Transcriptome assembly based on reference

- **SNP and INDEL detection**

GATK (The Genome Analysis Toolkit) was used on genome mapping result to identify SNP and INDEL for the two samples. The final SNP and INDEL in Variant Call format (VCF) was obtained after filtering out the unreliable sites.

- **Differentially splicing gene detection**

The software **RMATS** (Version: v3.2.5; Parameters: -analysis U –t paired –a 8) was used to detect differentially splicing gene, indicating the differential isoform relative abundance between control and 1 CAG samples. It is a computational tool to detect differential alternative splicing events from RNA-Seq data, it calculates the inclusion and skipping isoform as it is shown in Figure 5.5.

The statistical model of **MATS** calculates the *P-value* and false discovery rate (FDR); Gene with FDR ≤ 0.05 was defined as significant differentially splicing gene (DSG).

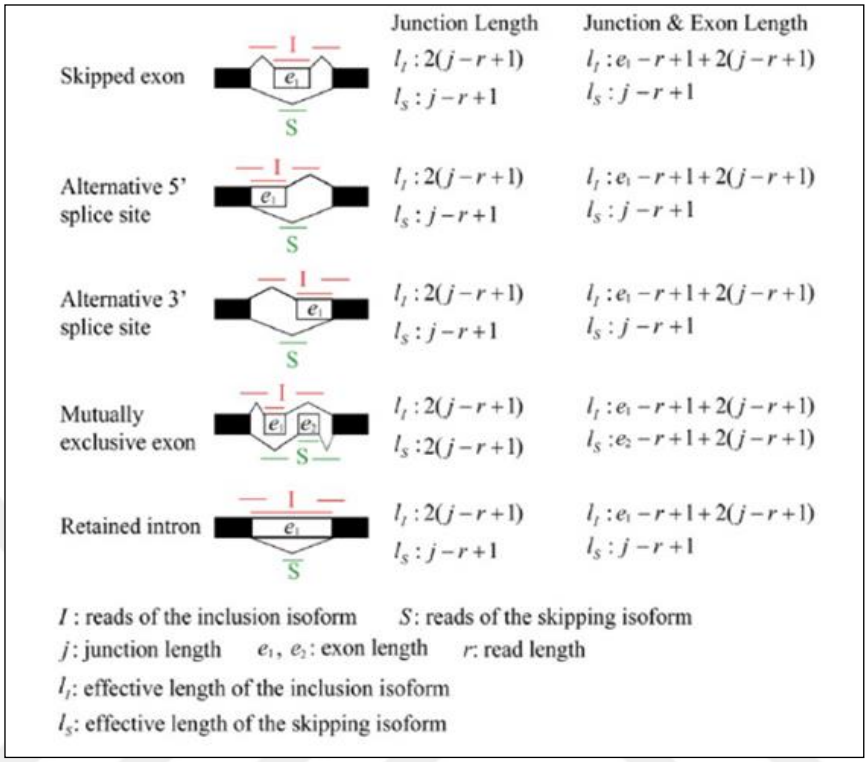


Figure 5. 5 Relative abundance calculation of differential isoforms.

- **Gene expression analysis**

Clean reads were mapped to a reference by the mean of **Bowtie2** (Version: v2.2.5 and parameters: -q --phred64 --sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min L, 0, -0.1 -I 1 -X 1000 --nomixed--no-discordant -p 1 -k 200).

Gene expression level was calculated using **RSEM** (Version: v1.2.12 and parameters: default), which is a software package used to estimate gene and isoform expression levels from an RNA sequencing data.

Pearson correlation between the samples was determined using **cor** and diagrams were drawn with **ggplot2,** both are **R** functions.

- **DEGs detection and hierarchical clustering analysis of DEGs**

DEGs have been detected with **PossionDis** (Parameters: Fold Change ≥ 2.00 and FDR ≤ 0.001) based on the poisson distribution and performed as described by [102].

Hierarchical clustering for DEGs was performed using **pheatmap**, a function of **R**. For cluster more than two groups, the intersection and union DEGs between them were done.

- **Gene Ontology analysis**

Gene Ontology (GO) annotation represents a link between a gene product type and a molecular function, biological process, or cellular component type.

The GO functional enrichment was performed using **phyper**, a function of **R**, and then False Discovery Rate (FDR) was calculated for each *p-value*. Generally, the terms which have 0.01 as a maximum FDR value, are defined as significantly enriched. DEGs were then classified according to the GO annotation result.

- **Pathway analysis (KEGG) of DEGs**

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource that integrates genomic, chemical and systemic functional information.

Pathway functional enrichment was carried out using **phyper**, a function of **R**, and then False Discovery Rate (FDR) was calculated for each *p-value*. Actually, the terms which have 0.01 as a maximum FDR value, are defined as significantly enriched. DEGs were organized based to the official classification of KEGG.

- **Transcription factor prediction of DEG**

The **getorf** software (Version: EMBOSS: 6.5.7.0 and parameters: -minsize 150) was used to find ORF of each DEG. We aligned ORF to transcription factor (TF) domains from **PlntfDB** (Version: v23.0 and Website: http://plntfdb.bio.uni-potsdam.de/v3.0/) using **hmmsearch** (Version: v3.0 and parameters: default).

- **Plant disease resistant gene prediction**

BLAST (Version: v2.5.0; Parameters (Running): -evalue 1e-5 -outfmt 6 -max_target_seqs 1; Parameters (Selecting): query coverage >= 50%, identity >= 40%) and **DIAMOND** (Version: v0.8.31; Parameters (Running): --evalue 1e-5 --outfmt 6 --max-target-seqs 1 --more-sensitive; Parameters (Selecting): query coverage ≥ 50%, identity ≥ 40%) were used to map DEGs to the **PRGdb** database in order to detect plant disease resistant genes based on the query coverage and identity requirement.

5.2.5.2 Small RNA sequencing

- **Small RNA sequencing process**

Sequencing process is divided into an experimental pipeline and a bioinformatics one. Every step of the experimental pipeline, from sampling to sequencing, influences the data quality and quantity which affects directly the bioinformatics analysis results.

The experimental pipeline is described in Figure 5.6. Actually, total RNAs were run into polyacrylamide gel electrophoresis (PAGE), in order to select small RNAs segments (18-30 nt). The 3' ends of the selected segments were bound to 3-blocked single stranded DNA adapters using 5-adenylated linkage and a reverse primer annealing reaction was occurred: RT primer was added to the solution and cross-linked to the 3' adapter of RNAs. The 5' ends of the segments produced after the annealing reaction were bound to a 5' adapter. cDNA strand was then synthesized and amplified using a high-ping polymerase and enriched with both 3' and 5' adapter. Library fragments (PCR products of 100～120 bp) were selected by PAGE to eliminate primer, dimer and other byproducts. Separation of the obtained fragments and the cyclization of single-strand DNA have occurred under high temperature. ssDNA circle was used to develop DNA nanoballs (DNBs).

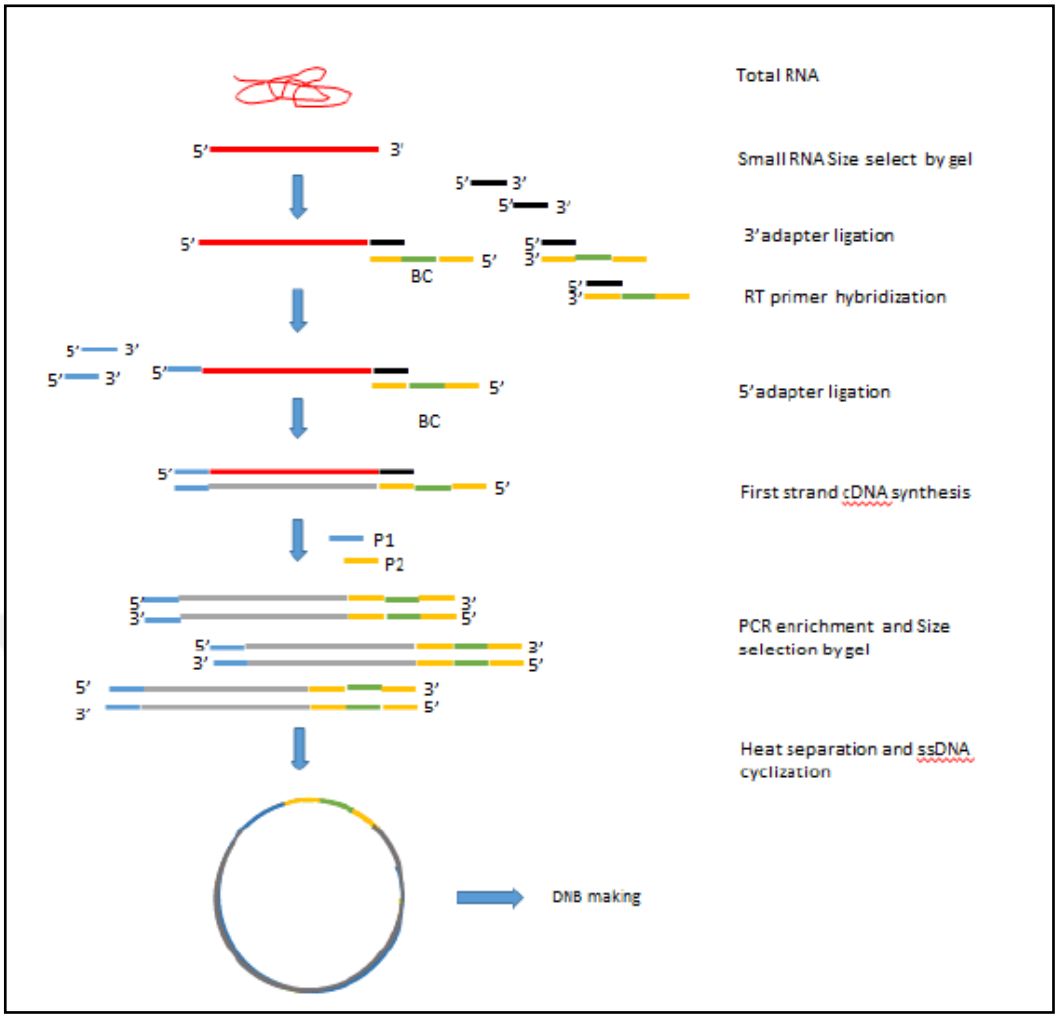


Figure 5. 6 Small RNA-seq experimental process

Bioinformatics analysis pipeline for small RNA sequencing (Figure 5.7) consists in; Eliminate low-quality reads, adaptors and other contaminants to get clean reads, then summarize their length distribution by identifying common and specific sequences between samples. Clean tags were annotated in different categories followed by the prediction of the novel miRNAs and the function annotation of the known miRNAs.

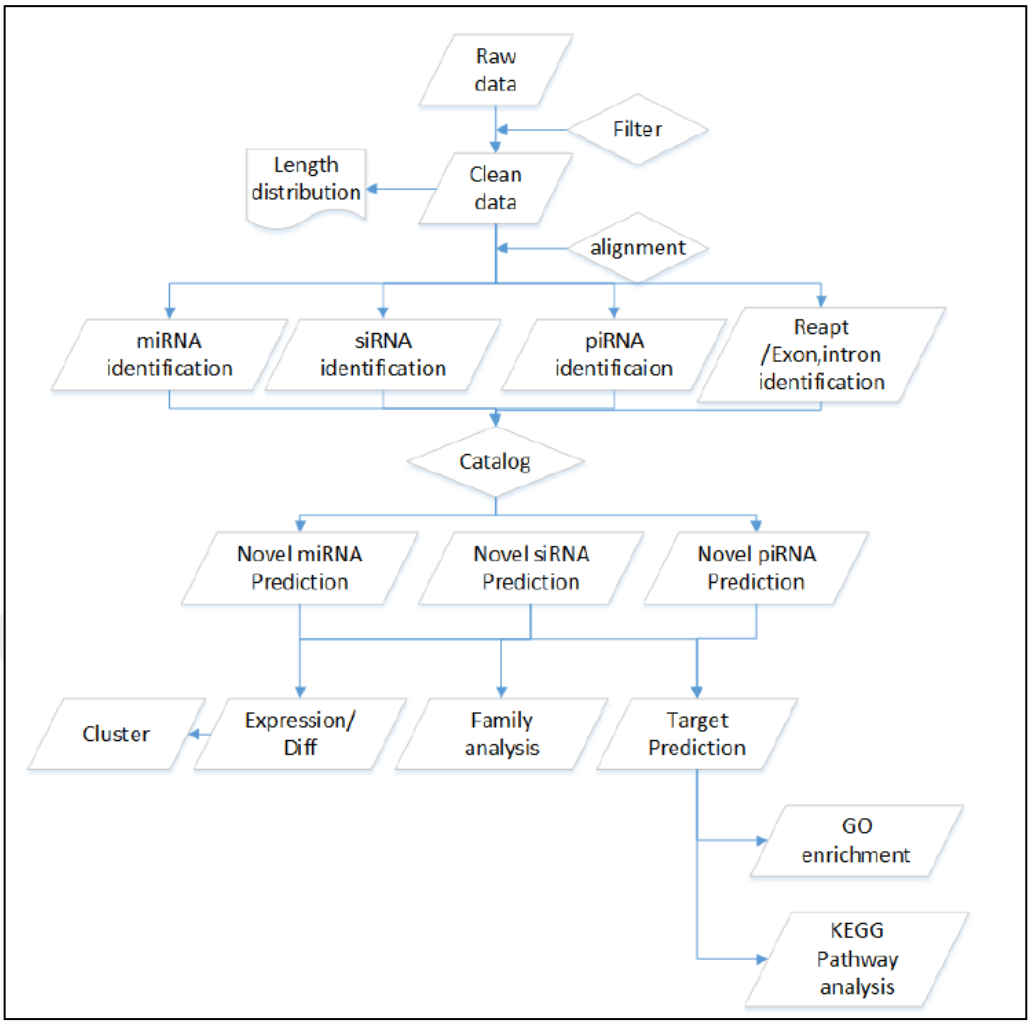


Figure 5. 7 Bioinformatics analysis pipeline for small-RNA

- **Data filtering**

All contaminant tags were removed following this order;

- Low quality tags
- Tags with 5' primer contaminants
- Tags without 3' primer
- Tags without insertion
- Poly-Adenylated tags
- Short tags (shorter than 18 nt)

Clean tags length distribution was afterward summarized. Length distribution of small RNA is between 18nt and 30nt; miRNA is normally 21nt or 22nt, siRNA is

24nt, and piRNA is 30nt. The remaining tags after filtering are called 'Clean tags' and stored in **FASTQ** format [69].

- **Reads mapping**

Bowtie2(Version2.3.5; http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) was used to map clean reads to the reference genome (*Arabidopsis thaliana* genome)[76].

- **sRNA classification**

In the annotation information of different RNAs, some small RNA tags can be mapped to more than one category. To make sure that every small RNA is mapped to only one category, the sequence tags were compared with the sequences of noncoding RNAs (rRNA, tRNA, snRNA, snoRNA) available in **Rfam** database [103] (https://www.sanger.ac.uk/science/search-tools-and-software) and the GenBank noncoding RNA database (https://www.ncbi.nlm.nih.gov/) to classify degradation fragments of non-coding RNA. Filtered tags were aligned against the sRNA database **miRbase** (http://www.mirbase.org/) to identify based on the utterly matched sequences known and conserved miRNAs.

To make every unique small RNA mapped to only one annotation, the small RNAs were annotated with the priority rule: miRNA > piRNA > snoRNA > Rfam: known miRNA > repeat > exon > intron.

- **sRNA prediction**

PIPmiR (Pipeline for the Identification of Plant miRNAs) is used to predict novel miRNA [104] by exploring the characteristic hairpin structure of miRNA precursor (Figure 5.8). **Piano** program (http://ento.njau.edu.cn/Piano.html) developed by [105] is used to predict piRNAs. This algorithm is based on the Support Vector Machine (SVM) algorithm and transposon interaction information.

Small interfering RNAs (siRNAs) are 22-24 nt double-stranded RNAs, each strand is 2 nt longer than the other. Because of this structural feature, tags were aligned to each other to find sRNAs meeting these criteria. These resultant tags are most likely to be siRNA candidates.

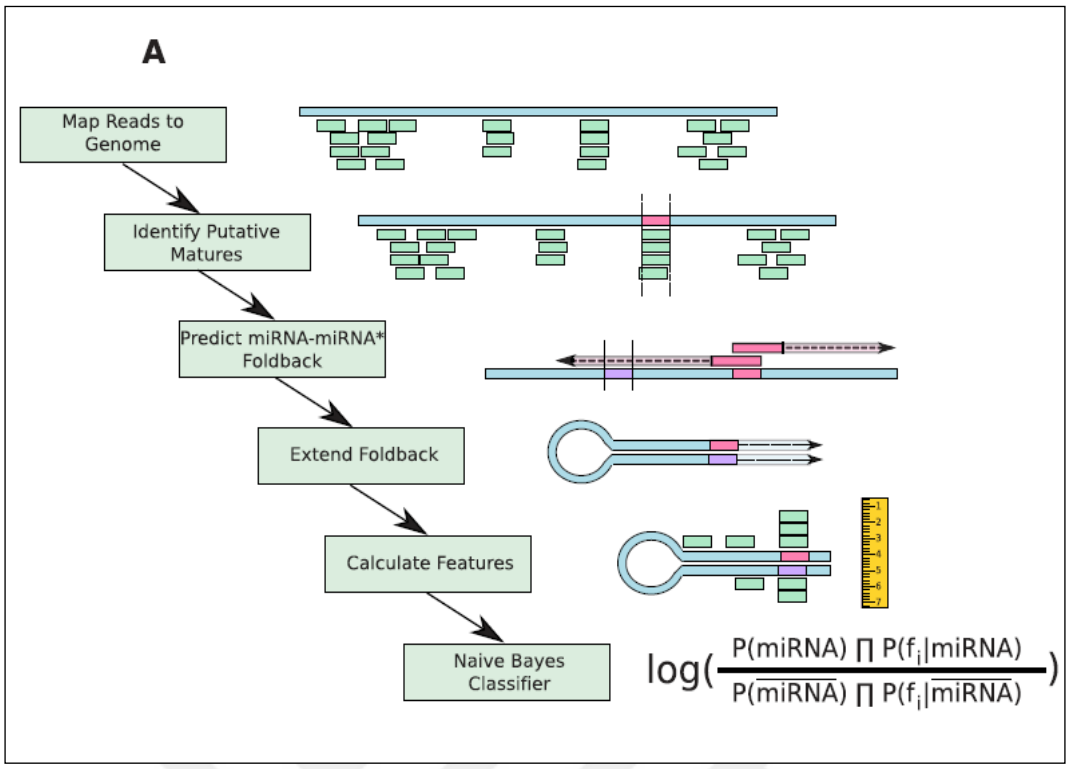


Figure 5. 8 A schematic depiction of PIPmiR pipeline steps [104]

- **sRNA expression**

The method of Transcripts Per Kilobase Million (TPM) was applied to measure small RNA expression level. TPM method removes the sequencing discrepancy influence on level of small RNA expression calculation.

Gene expression difference between samples is compared using the measured gene expression level.

TPM is calculated following the above formula:

$$TPM = \frac{C * 10^6}{N} \qquad (5.3)$$

- **Target prediction**

Two types of software were used are:

psRobot (http://omicslab.genetics.ac.cn/psRobot/; -gl 17-p8-gn 1) [106] **TargetFinder** (http://www.bioit.org.cn/ao/targetfinder.htm; parameters: -c4;)[107] used to find more accurate specific plants targets.

- **Screening DESs (DEGs)**

RNA sequencing could be designed as an arbitrary sampling process, in which every read is modeled independently and uniformly from every possible nucleotide in the sample [108]. Under this assumption, the number of reads generating from a gene (or transcript isoform) follows a binomial distribution and could be approximated by a Poisson distribution. **DEGseq** software (http://bioinfo.au.tsinghua.edu.cn/software/degseq/) was used using this statistical assumption and based on the MA-plot, to determine gene expression difference. The hypothesis test of H_0: $p_1=p_2$ versus H_1: $p_1\neq p_2$ is attributed to each gene on the MA-plot. *P-values* could be assigned according to the conditional normal distribution. They were calculated for each gene and adjusted *to Q-values* for multiple testing corrections. To improve accuracy of DEGs result, a gene is defined as a differentially expressed when reads number fold change is ≥ 2 and *Q-value* ≤ 0.001.

- **Hierarchical clustering analysis**

pheatmap function was used to perform the hierarchical clustering for differentially expressed miRNA.

- **Gene Ontology enrichment analysis**

GO enrichment analysis identifies all significantly enriched GO-terms in a list of DESs targets. It determines the corresponding genes to specific biological functions. GO analysis was performed by first, mapping all genes to GO-terms in the database (http://geneontology.org/), which calculates the gene numbers for every term. Then the test was used to find significantly enriched GO-terms in the input gene list. It is based on 'GO: TermFinder' (https://www.yeastgenome.org/help/analyze/go-term-finder). The *P*-value was corrected by using the Bonferroni method [109]; a corrected *P-value* ≤ 0.05 is taken as a threshold. GO terms fulfilling this condition were defined as significantly enriched GO terms.

- **Pathway enrichment analysis: KEGG**

KEGG analysis identifies the significantly enriched pathways in DESs target genes comparing to the whole genome background. The *P-value* is corrected by using the Bonferroni method [109], a corrected *P-value* ≤ 0.05 was taken as a threshold. KEGG terms conform with this condition are defined as significantly enriched KEGG pathways.

5.2.6 Real time quantitative PCR analysis: Sequencing validation

Two pools of RNAs, control and 1 μM CAG were sent for sequencing. Each of it contained three samples. In order to confirm RNA-seq and small RNA-seq results, the same RNAs samples from control and 1 μM concentration were used, in triplicates, in the qRT-PCR analysis.

5.2.6.1 RNA seq results validation

- **Synthesis of cDNA**

cDNA was synthesized from RNA samples that we used for sequencing, using High-Capacity cDNA Reverse Transcription Kit from Applied Biosystem™, composed of 10x RT Buffer (1.0 mL), 10xRT Random Primers (1.0 mL), 25xdNTP Mix (100 mM, 1.0 mL), and MultiScribe™ Reverse Transcriptase (50 U/μL, 1.0 mL). A mix of 20 μL per tube was prepared as it is described on the Table 5.4 below;

Table 5.5 cDNA synthesis mixture components

Component	Volume/tube (μL)
10x RT Buffer	2
10xRT Random Primers	2
25xdNTP Mix	0.8
MultiScribe™ Reverse Transcriptase	1
dH_2O	9.86
RNA	4.34

Volume of our RNA samples was calculated based on their concentration. The RNA sample should have a concentration of 1000 ng. The PCR program was as follows; 10 min at 25°C, 120 min at 37°C then 5 min at 85°C. Gel electrophoresis was run to check the amplification efficiency of our cDNAs.

- **Primer's design and gradient PCR**

RNA-Seq generated a DEGs list from which ten genes were chosen randomly in order to validate the obtained results. Ten primer sets (Table 5.5) were designed with Primer3 Input software (version 0.4.0, https://bioinfo.ut.ee/primer3-0.4.0/).

Table 5.6 Primers pairs list used in RNA seq results validation

Genes	Product size (bp)	Sequence
AT5G47560	168	F 5′-CTTGTTATTGGGAGCTGGATTC-3′ R 5′- GCGTTGTTTGAAGTGAACTCTG-3′
AT1G05560	148	F 5′- TTCCCTCTGTTCATCTCTGGAT-3′ R 5′-GTGTTGGAAGGTGAGAGGAAAG-3′
AT2G43535	126	F 5′- CTTCGCCATCTTTATCATCCTC-3′ R 5′-AATATCTTAGGCGCACAGAAGC-3′
AT5G38200	151	F 5′- GCTTGTGGTGGGACTCTTTATC-3′ R 5′-ACCAGGAATGAAGTGGACTGTT-3′
AT3G48360	195	F 5′-CATCTTTTAGCGTTGTCTCACG-3′ R 5′-TTCAGTCTGCTCAACGGTCTTA-3′
AT5G19500	142	F 5′- GAAGGAGACTTGCCGAGAGTAA-3′ R 5′- GAGGATCGACCATCTTCTCAAC-3′
AT2G25810	157	F 5′- CTATTGTGGATCCGAAGAAAGG-3′ R GTCCAGTTTCCAGAGACCAAAG-3′
AT3G53980	189	F 5′- CTCTTCAGTGCCTCAAATTTCC-3′ R 5′- GTTCCGATGAGGAGAATCAGAC-3′
AT1G21270	163	F 5′- TTTCACAACCATCCTACTGTGC-3′ R 5′-TGCAAGTCTTGTCTCCGATAGA-3′
AT4G27030	148	F 5′- CCAACCAAGTACCCTCTACGAC-3′ R 5′- GATTCACAGTGCGTTGCTCTAC-3′
AT3G18780 (Actine ACT2)	313	F 5′-TGCTGACCGTATGAGCAAA-3′ R 5′-CTCCGATCCAGACACTGTA-3′

To identify their annealing temperature, the synthesized cDNAs were used to perform a gradient PCR between 50°C and 65°C, using DNA Taq polymerase. A mixture of 25 μl by tube was prepared as it is described on the Table 5.6.

Table 5.7 Gradient PCR mixture components

Component	**Volume/tube (μL)**
Buffer 10x	2.5
dNTP Mix (10 mM)	0.5
MgCl2 (50 mM)	2
DNA Taq Pol (5U/μL)	0.25
Primers (F+R)	0.5 + 0.5
cDNA	1
dH2O	17.75

The gel electrophoresis was run after gradient PCRs in order to identify the annealing temperature of each primer before running the qRT-PCR.

- **Real Time qRT-PCR analysis**

After identification of the annealing temperature for each gene, qRT-PCR analysis was performed in triplicate for each control and 1μM sample. Each reaction mixture was prepared with 10 µL of 1×SYBR Green Supermix, 2 µL of cDNA, and 1 µL from primer, forward and reverse, at 10 µmol/L of concentration.

The program of qRT-PCR we ran was as following: 5 min of denaturation at a temperature of 95°C, then 45 amplification cycles of 15 s each at 95°C, and according to the used primer, a 60 s cycle at the temperature determined by gradient PCR followed by another one for 30 s at 72°C. Amplification specificity was checked by running a melting curve with an increase of 0.5°C per cycle that last for 5 s, from 65 up to 95°C. Gene expression in all samples we used in qRT-PCR was normalized using the expression of *A. thaliana* actin 2 (Table 5.5). The ΔΔ cycle threshold method [110] was used to calculate the quantitative variation between samples.

5.2.6.2 Small RNA seq results validation

- **Stem-loop RT primers design and pool preparation**

Stem-loop primers (Figure 5.9) of twenty microRNAs selected (Table 5.7) were designed with the help of miRNA primer designer software developed by Fuliang Xie from East Carolina University and provided by Macrogen Inc. (Ankara, Turkey).

Table 5.8 Stem-Loop primers designed for qRT-PCR

miRNA	**Primers sequences (5'-3')**
miR160a-5p,miR160b, miR160c-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTGGCAT **F**: GCGGCGGTGCCTGGCTCCCTG
miR158a-3p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTGCTTT **F**: GCGGCGGTCCCAAATGTAGAC
miR319a, miR319b	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACAGGGAG **F**: GCGGCGGTTGGACTGAAGGGAG
miR845a	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACCATCAA **F**: GCGGCGGCGGCTCTGATACC
miR158b	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTGCTTT **F**: GCGGCGGCCCCAAATGTAGAC
miR169d, miR169e, miR169f-5p, miR169g-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACCGGCAA **F**: GCGGCGGTGAGCCAAGGATGAC
miR866-3p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTCTTCA **F**: GCGGCGGACAAAATCCGTCTTTG
miR172c, miR172d-3p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACCTGCAG **F**: GCGGCGGAGAATCTTGATGATG
miR2112-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACACATTG **F**: GCGGCGGCGCAAATGCGGATATC
miR398a-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTGTGTT **F**: GCGGCGGAAGGAGTGGCATGTG
miR157a-5p,miR157b-5p, miR157c-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGTGCTC **F**: GCGGCGGTTGACAGAAGATAGAG
miR408-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACCATGCT **F**: GCGGCGGACAGGGAACAAGCAG

Table 5.7 Stem-Loop primers designed for qRT-PCR (cont.)

miR827	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACAGTTTG **F**: GCGGCGGTTAGATGACCATCAAC
miR396b-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACAAGTTC **F**: GCGGCGGTTCCACAGCTTTCTTG
miR399f	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACCCGGGC **F**: GCGGCGGTTCCACAGCTTTCTTG
miR5650	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTGTATC **F**: GCGGCGGTTGTTTTGGATCTTAG
miR863-3p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACATTATG **F**: GCGGCGGTTGAGAGCAACAAGAC
miR399e	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACCGACGC **F**: GCGGCGGTGCCAAAGGAGATTTG
miR837-5p	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTGAAAC **F**: GCGGCGGATCAGTTTCTTGTTC
miR8173	**SL**:GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCAC TGGATACGACTCCCAC **F**: GCGGCGGATGTGCTGATTCGAG
Universal reverse Primer	GTGCAGGGTCCGA GGT

SL: Stem-Loop, **F:** Forward Primer

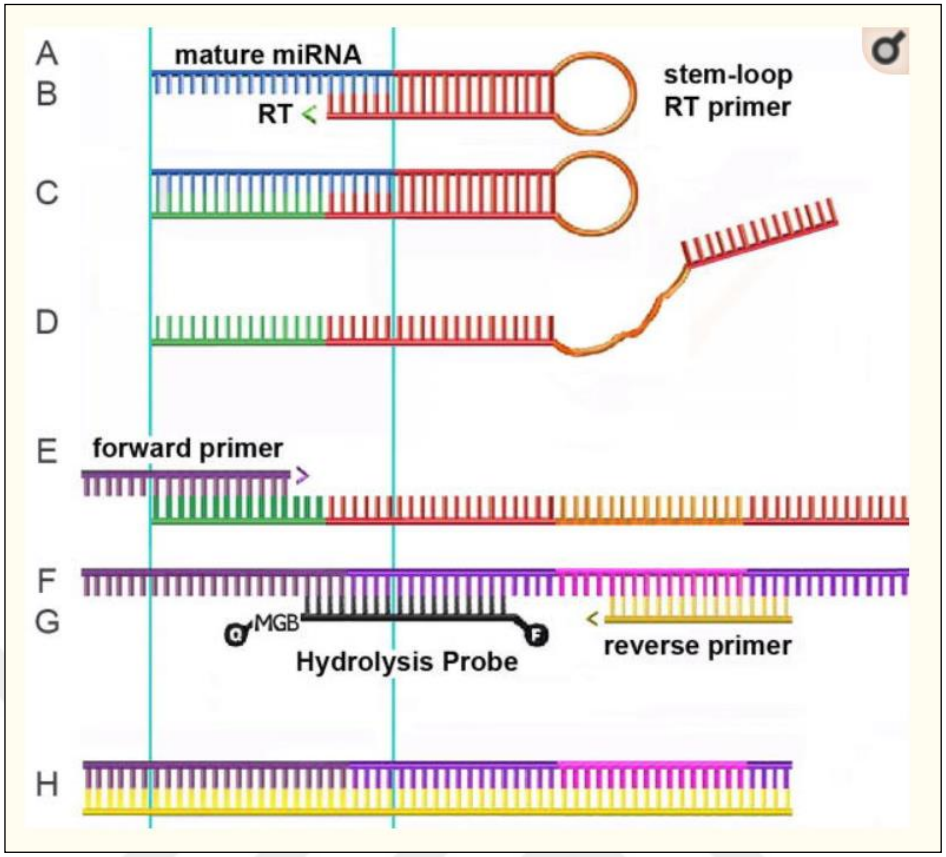


Figure 5. 9 Nucleic acid reagent used in stem-loop primers production

Stem-Loop primers were divided in 4 pools in order to have primer pools of 10 mM. Each pool contains 10 μL of 5 stem-loop primer, 20 μL of oligo (dt) 18 primer and 30 μL of dH_2O to have in total 100 μL of pool primers.

- **cDNA synthesis**

cDNA was synthesized using TaqMan™ MicroRNA Reverse Transcription Kit from Applied Biosystems, composed of 100mM dNTPs (with dTTP), MultiScribe™ Reverse Transcriptase (50 U/μL), 10X Reverse Transcription Buffer and RNase Inhibitor (20 U/μL). A mix of 15 μL per tube was prepared using Stem-loop RT primer pool, as described in the Table below (Table 5.8).

Table 5.9 cDNA synthesis mixture components for miRNA

Component	**Volume/tube (μL)**
Stem-loop RT primer pool (10mM)	6
10x Reverse Transcription Buffer	1.5
DNTPs (100mM)	0.3
MultiScribe™ Reverse Transcriptase	3
RNase Inhibitor	0.2
ddH_2O	3
RNA	1

The reactions were then placed in a thermal cycler under the program described in Table 5.9. After the PCR reaction, 80 μl of dH_2O is added to each tube.

Table 5.10 PCR steps of cDNA synthesis

Step	**Temperature**	**Time**
Reverse transcription	16°C	30 min
	42°C	30 min
Stop reaction	85°C	5 min
Hold	12°C	∞

- **Real Time qRT-PCR analysis**

Every reaction mixture contained 10 μL of 1 × SYBR Green Supermix (Hibrigen, Turkey), 2 μL of cDNA, 1 μL of forward primer (corresponding to the Stem-loop primer) and 1 μL of reverse primer (Universal reverse primer, Table 5.7) (10 μmol/L). The program of qRT-PCR we set is as following: 5 min of denaturation at a temperature of 95°C, then 45 amplification cycles of 15 s each at 95°C, and according to the used primer, a 60 s cycle at the temperature determined by gradient PCR followed by another one for 30 s at 72 °C. A melting curve was run to check the amplification specificity, by increasing 0.5°C per cycle that last for 5 s, from 65 up to 95°C. Gene expression in all samples we used in qRT-PCR, was normalized using the expression of *A. thaliana* actin (Table 5.5) by running actin-specific primer. Fold change results of qRT-PCR were determined using ΔΔ cycle threshold method [110].

6
RESULTS AND DISCUSSION

6.1 Results

Arabidopsis thaliana Col-0 calli were grown on medium with and without cycloastragenol during eleven months. We had three groups of calli considered as three technical repetitions; six months old calli, 9 months old calli and 11 months old calli. During each repetition, three different concentrations, presenting three biological repetitions were applied: 0 μM (Control), 1 μM and 10 μM. The calli have been continuously sub-cultured every three weeks. Our experiment aims to understand CAG effect on *A. thaliana* calli specifically and on plants generally. During the first part of our study, Growth Index have been calculated by measuring the weight of our calli before and after CAG treatment, then the transcriptome and miRNAome of gene expression pattern were analyzed by NGS after RNA isolation from the chosen treated calli.

6.1.1 *A. thaliana* plant growth and callus formation

Arabidopsis thaliana seeds which were sown in Petri dishes containing MS medium were allowed to germinate in the plant growth chamber. Seven-days old seedlings obtained as a result of 90% of germinated seeds (Figure 6.1 A) were transferred into test tubes that contain MS medium for extra 21 days. Roots of the obtained plants were used as explants for calli formation (Figure 6.1 B).

Explants were excised in 2 cm pieces and transferred into MS medium containing 1 mg / L 2,4-D to induce callus formation.

Calli formation started to be observed after approximately 4 weeks. They have been sub-cultured every three weeks (Figure 6.2).

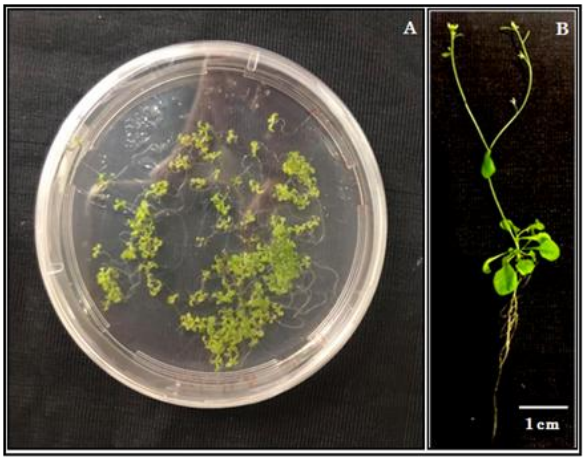


Figure 6.1 (A) *A. thaliana* seedlings (B) 30 days old *Arabidopsis* plant of which roots were used as explants

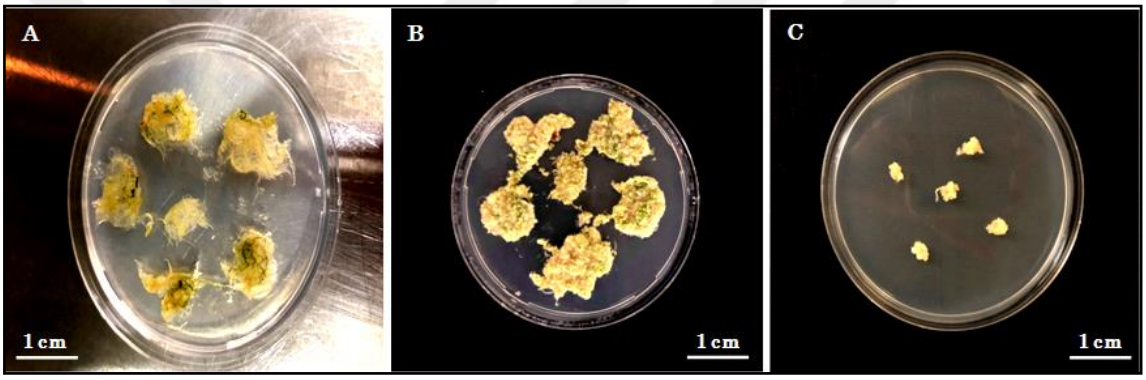
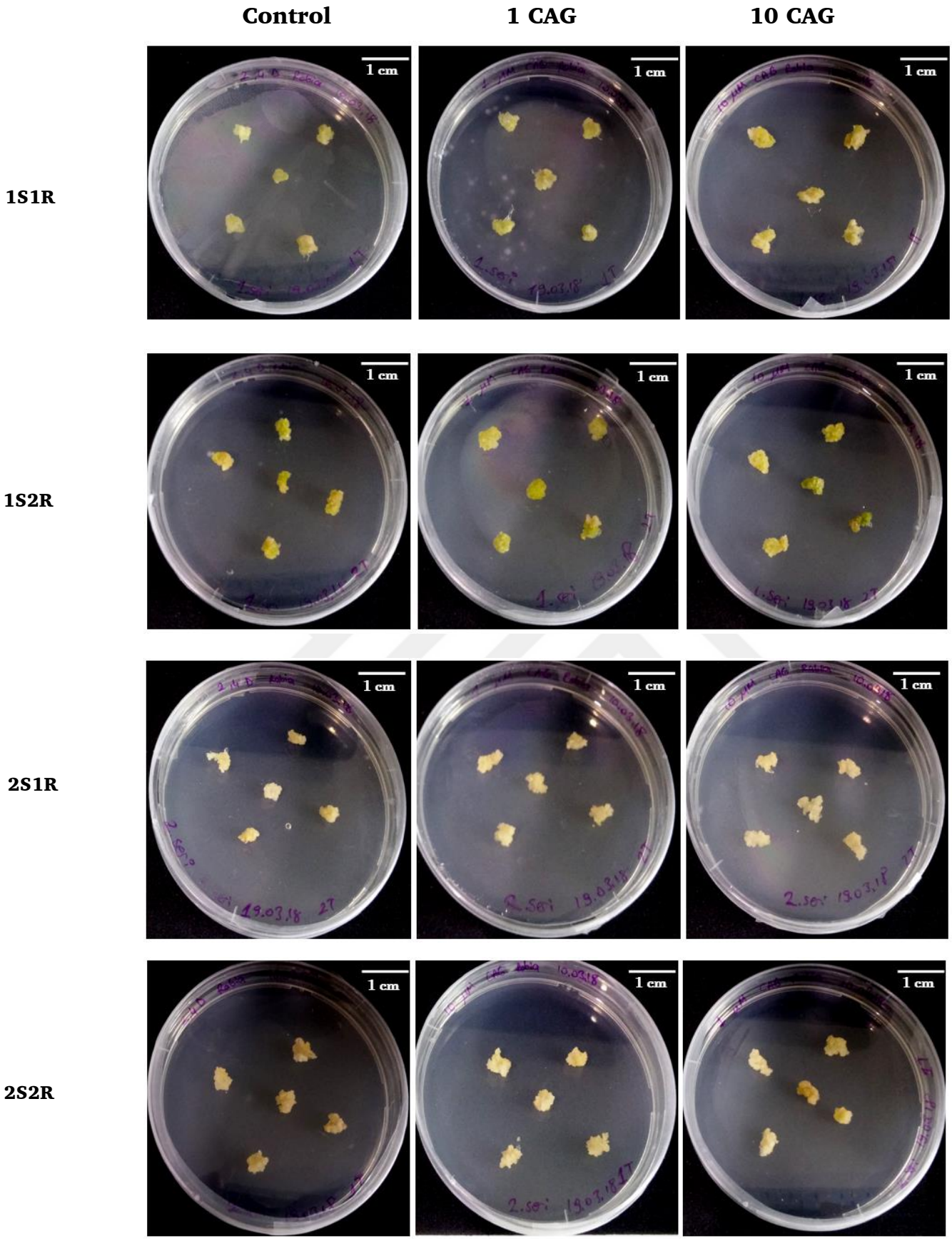


Figure 6. 2 Callus development. (A) Callus in first development stage (first week); (B): Callus after 4 weeks of development; (C): sub-cultured callus

6.1.2 Growth index (GI) measurement

A control concentration (0 μM), 1 μM and 10 μM of CAG were applied on our calli to identify the optimum CAG concentration affecting plants. CAG was applied in two series for each concentration, and each series presented three repetitions. Calli weight was measured before and after treatment to calculate growth index. Morphologically, a remarkable difference was noticed on some calli. Figures 6.3 and 6.4 show the 9 months old calli before and after CAG treatment, their weights are presented in the Table 6.1.



S : Serie
R : Repetition

Figure 6. 3 9 months-old calli before CAG treatment

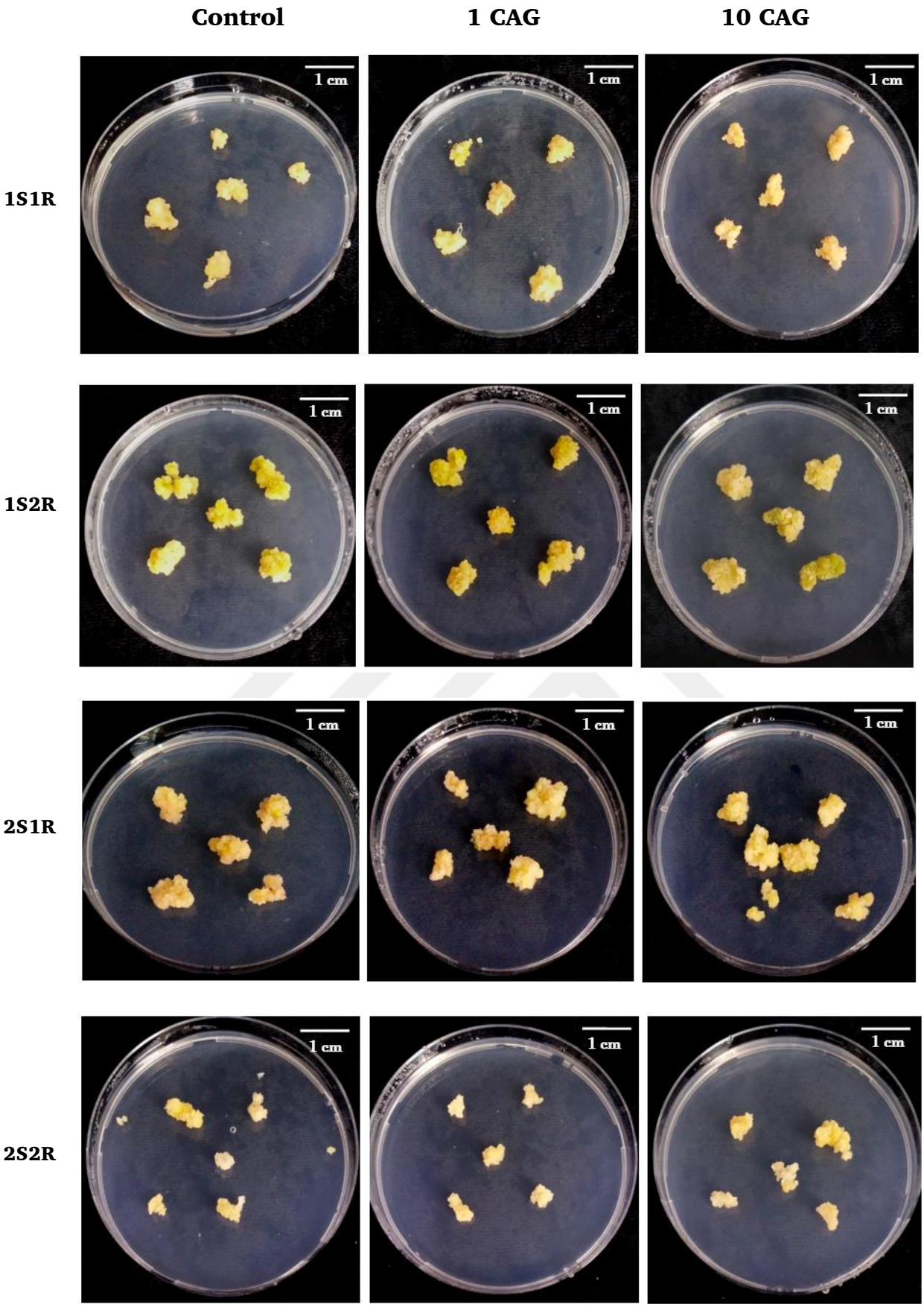


Figure 6. 4 9 months-old calli after CAG treatment

S : Serie
R : Repetition

Table 6.1 9 months old calli weight and growth index measurement

	Initial weight	**Final weight**	**Growth Index**
Control			
A1 (1S-1R)	0,37	0,77	1,08
D1 (1S-2R)	0,47	1,75	2,72
G1 (1S-3R)	0,54	1,3	1,40
M1 (2S-1R)	0,27	1,44	4,33
P1 (2S-2R)	0,25	1,62	5,48
T1 (2S-3R)	0,27	1,58	4,85
Average			3,31
1 μM CAG			
B1 (1S-1R)	0,48	0,85	0,77
E1 (1S-2R)	0,59	1,51	1,56
H1 (1S-3R)	0,48	1,52	2,16
N1 (2S-1R)	0,26	1,75	5,73
R1 (2S-2R)	0,24	1,75	6,29
U1(2S-3R)	0,27	1,11	3,11
Average			3,27
10 μM CAG			
C1 (1S-1R)	0,59	0,7	0,18
F1 (1S-2R)	0,57	1,95	2,42
I1 (1S-3R)	0,44	1,43	2,25
O1(2S-1R)	0,24	1,62	5,75
S1 (2S-2R)	0,25	2,01	7,04
V1(2S-3R)	0,25	0,75	2
Average			3,27

Statistically, no significant difference was determined.

Same measurements were done on 6 months old and 11 months old calli. After calculation, growth index average was 3.180 ±1.73-fold change for control concentration, 2.659 ± 2.1-fold change for 1 μM CAG and for 10 μM CAG, 2,941 ± 1.97-fold change was observed. Student *t-test* was done and showed no significant change observed in the growth rate of our calli after cycloastragenol application (Figure 6.5).

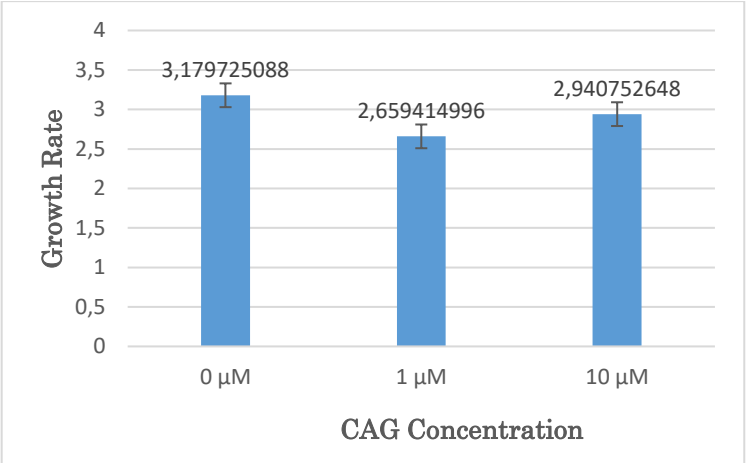


Figure 6. 5 Growth index results after student test for 9 months old calli

6.1.3 RNA extraction and concentration measurement

Total RNAs (Figure 6.6 A) were extracted from calli tissues which were developed under the three different concentrations. The integrity of all RNAs (the presence of three RNA specific bands) was checked by agarose gel electrophoresis (Figure 6.6 B).

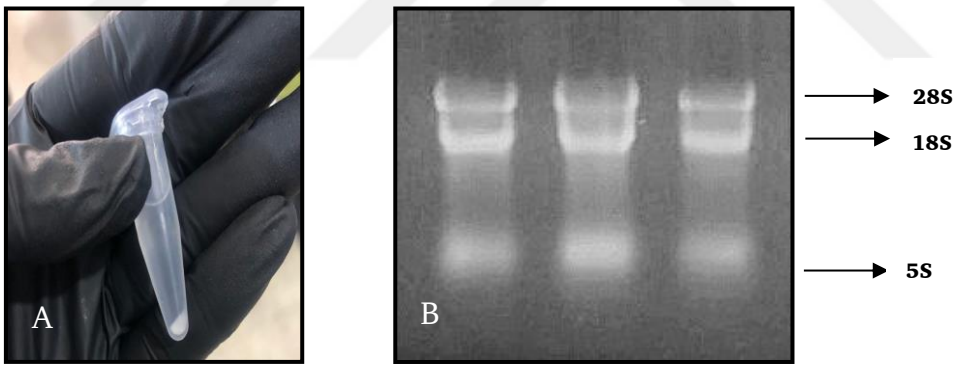


Figure 6. 6 (A) RNA pellet after ethanol wash (B) RNAs isolated from callus visualized on agarose gel electrophoresis

Concentration of RNA samples, dissolved in DEPC-dH_2O, was measured by determining the absorbance values at 260 and 280 nm wavelengths using Nanodrop spectrophotometer. Table 6.2 presents the measured concentration of RNAs extracted from 9 months old calli.

Table 6.2 Concentration of RNAs extracted from 9 months-old calli

Callus age	Group	Sample	RNA Concentration (μg/ml)
9	1S1R Control	D	2182.5
9	1S1R 1 CAG	E	2061,5
9	1S1R 10 CAG	F	5872,8
9	1S2R Control	G	6800
9	1S2R 1 CAG	H	2749
9	1S2R 10 CAG	I	5705
9	2S1R Control	M	3890,6
9	2S1R 1 CAG	N	3842,9
9	2S1R 10 CAG	O	2168,9
9	2S2R Control	T	1496,4
9	2S2R 1 CAG	U	6177,2
9	2S2R 10 CAG	V	7033,3

The same calli were used parallelly in [101] study, and found out that 1 μM of CAG caused an increase in the transcription levels of various genes that act as positive regulators on the telomere/telomerase system. Besides, the initial results showed that 1 μM of CAG induces higher GI of our calli than the control one. Despite a non-significant increase in GI, preliminary results encourage us to investigate on the CAG-induced physiological responses. Then, two pools of control of 1 μM CAG RNA samples were chosen to be sent to BGI for RNA and small RNA sequencing.

6.1.4 High-throughput sequencing

After checking RNA samples quality, and according to the growth index results and on [101], 1 μM was chosen to be the potential CAG concentration affecting plants.

Two pools of RNA, one control and one 1 μM of CAG, were sent to Beijing Genomics Institute (BGI, China) for RNA-sequencing and small RNA sequencing.

Each pool is a mix of three RNAs from the two different concentrations of 9 months old calli.

6.1.4.1 Transcriptomic analysis: RNA-sequencing

In order to study the molecular and the physiological changes and the transcriptional responses induced by cycloastragenol, two cDNA libraries, corresponding to treated and control samples, were established using total RNA extracted from calli of wild type *A. thaliana*, and sequenced.

The sequenced dataset was delivered to the scientific public through the two databases; Sequence Read Archive (SRA) under the accession number of SRP285089 and Gene Expression Omnibus (GEO) with GSE158409 accession number. The two databases belong to the National Center for Biotechnology Information (NCBI).

Deep sequencing of the two libraries in the BGISEQ-500 platform generated about 5.94 Gb per sample. The average rate of the genome mapping was around 96% while the average rate of gene mapping was higher than 91%. A total of 22.593 genes were expressed among which 22.593 are known genes and none of them are novel genes. In addition, 4.843 novel transcripts were identified among which 4.347 of are splicing event for known genes, 145 of them are novel coding transcripts without any known features, and the remaining 351 transcripts, are long non-coding RNA [85].

6.1.4.1.1 Sequencing reads filtering

Transcriptome libraries sequencing produced a total of 133.586.630 raw reads, divided into 63.509.590 reads for control sample and 70.077.040 reads for CAG-treated sample (Table 6.3).

To certify the validity of our analysis, initial data was filtered to predicate read quality. The original data performance of each sample is shown in Figure 6.7. Reads which had adapter, those with low base quality (≤ Q20%), sequences with less than 5% of N and rRNA sequences, were removed to keep for, respectively, control and 1 CAG libraries 56.215.604 and 62.620.022 clean reads; around 48%

and 52%. Q20 reads values were over than 96% in the two samples (Table 6.3) producing 11.88 GB of generated clean bases, representing 5.6 GB and 6.2 GB for, respectively, control and treated sample.

Table 6.3 Filtering of raw reads obtained through high throughput sequencing of RNA-Seq libraries

Sample	Total raw reads	Total clean reads	Total clean bases	Clean reads Q20	N reads ratio	Clean reads ratio
Control	63.509.590	56.215.604	5.621.560.400	96.4%	0.0%	88.51
1CAG	70.077.040	62.620.022	6.262.002.200	96.55%	0.0%	89.36

Sample: Sample Name
Total Raw Reads (Mb): The reads amount before filtering, Unit: Mb
Total Clean Reads (Mb): The reads amount after filtering, Unit: Mb
Total Clean Bases (Gb): The total base amount after filtering, Unit: Gb
Clean Reads Q20 (%): The Q20 value for the clean reads
N reads ratio: The total amount of reads which contain more than 5% unknown N base
Clean Reads Ratio (%): The ratio of the amount of clean reads

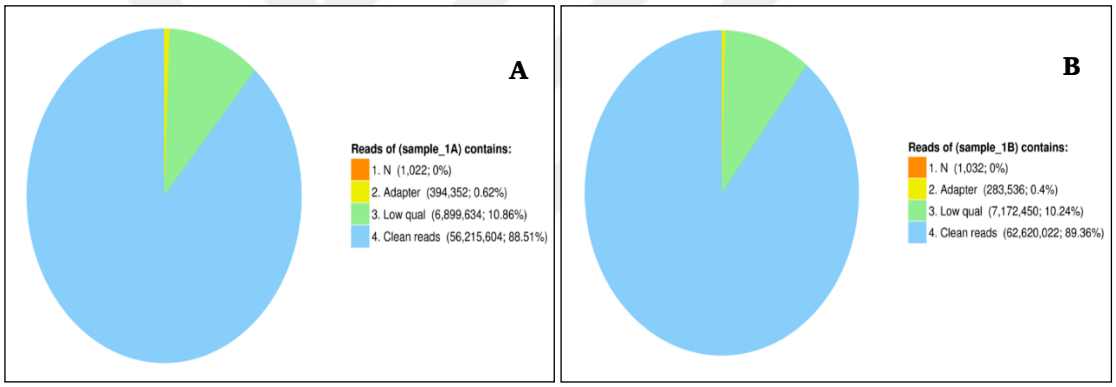


Figure 6. 7 Classification of raw reads in (A) Control sample and (B) 1μM CAG sample

6.1.4.1.2 Gene mapping

Using **HISAT** software, clean reads were mapped to the reference genome. As it shown in Table 6.4, genome mapping rate had an average of 96.72%, corresponding to 96.54% and 96.89% for non-treated and treated samples respectively while the average of gene mapping was 91.94%.

Mapped clean reads and measured gene expression through **RSEM** software generated 43.541 as total gene number corresponding to 21.639 and 21.902

genes, represented by 33.972 and 34.193 transcripts in control and 1 CAG sample respectively (Table 6.4).

Table 6.4 Summary of genome mapping

Sample	Total Clean Reads	Total Mapping Ratio	Total Gene Number	Total Transcript Number
Control	56.215.604	96.54%	21.639	33.972
1 μM	62.620.022	96.89%	21.902	34.193

6.1.4.1.3 Novel transcript prediction

Genome mapping is followed by the reconstruct of transcripts using **StringTie** software. **Cuffcompare** used the information of genome annotation to identify novel transcripts and with the help of **CPC** software, transcripts coding ability was identified. As it mentioned in Table 6.5, 4.843 novel transcripts were identified.

Table 6.5 Summary of novel transcripts

Total Novel Transcript	Coding Transcript	Non-coding Transcript	Novel isoform	Novel gene
4.843	4.492	351	4.347	145

6.1.4.1.4 SNP and INDL detection

SNP and INDEL variant for each sample were determined after genome mapping using **GATK**, and the final results were stored in VCF format. The Table 6.6 presents the different SNP variant types while Figure 6.8 summarizes SNP results.

Table 6.6 SNP variant types

Sample	A-G	C-T	Transition	A-C	A-T	C-G	G-T	Transversion	Total
Control	529	331	860	307	348	64	100	819	1679
1 μM	507	320	827	280	343	64	99	786	1613

Sample: Sample name, A-G: The amount of A-G variant type, C-T: The amount of C-T variant type,
Transition: The amount of A-G and C-T variant type
A-C: The amount of A-C variant type, A-T: The amount of A-T variant type, C-G: The amount of C-G variant type, G-T: The amount of G-T variant type
Transversion: The amount of A-C, A-T, C-G and G-T variant type, Total: The amount of all variant type

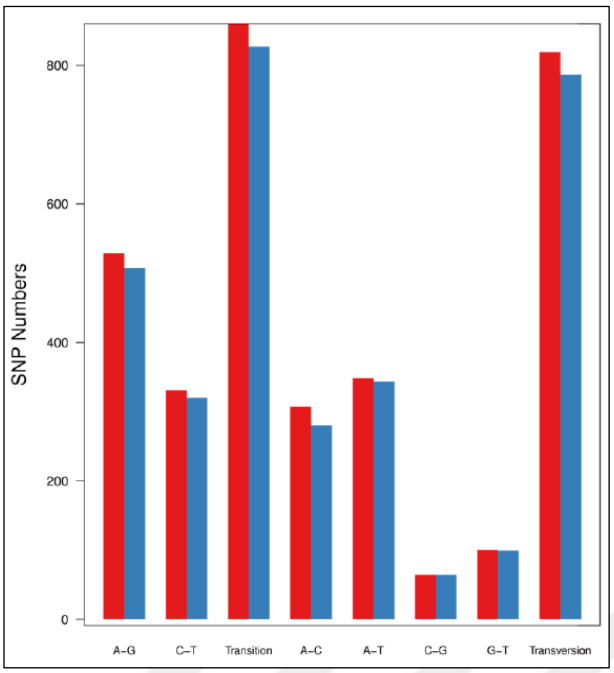


Figure 6. 8 SNP variant type distribution

X axis represents the type of SNP. Y axis represents the number of SNP.

Location of SNP and INDEL is statistically explained below in Figure 6.9 A and B.

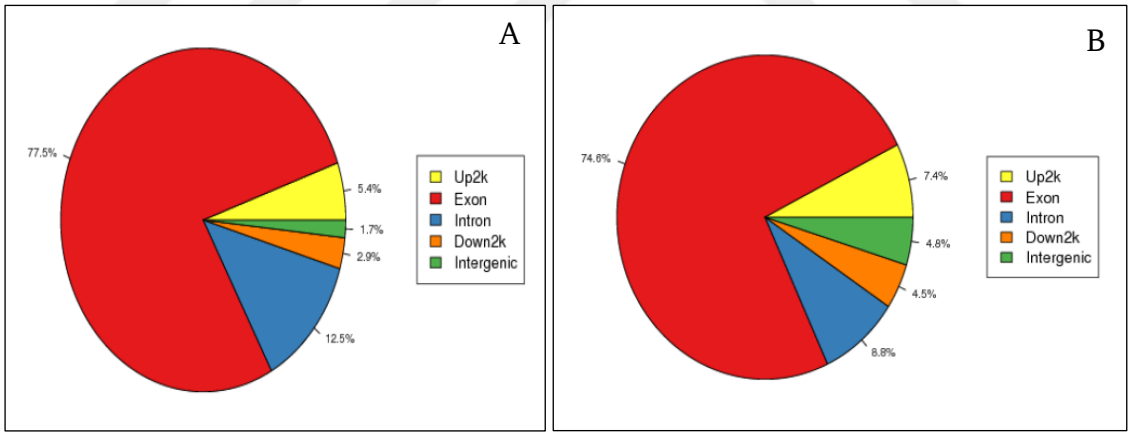


Figure 6. 9 Distribution of (A) SNP location and (B) INDEL location

6.1.4.1.5 Differentially splicing gene detection

Genome mapping is also followed by detecting differentially splicing genes between control and treated sample using **rMAT**.

DSGs are regulated by alternative splicing (AS), which allows the production of different isoforms from one gene only. Changes in relative abundance of isoforms,

regardless of the expression change, indicate a splicing-related mechanism. Five types of AS events including, Skipped Exon (SE), Alternative 5' Splicing Site (A5SS), Alternative 3' Splicing Site (A3SS), Mutually Exclusive Exons (MXE) and Retained Intron (RI), were detected (Figure 6.10).

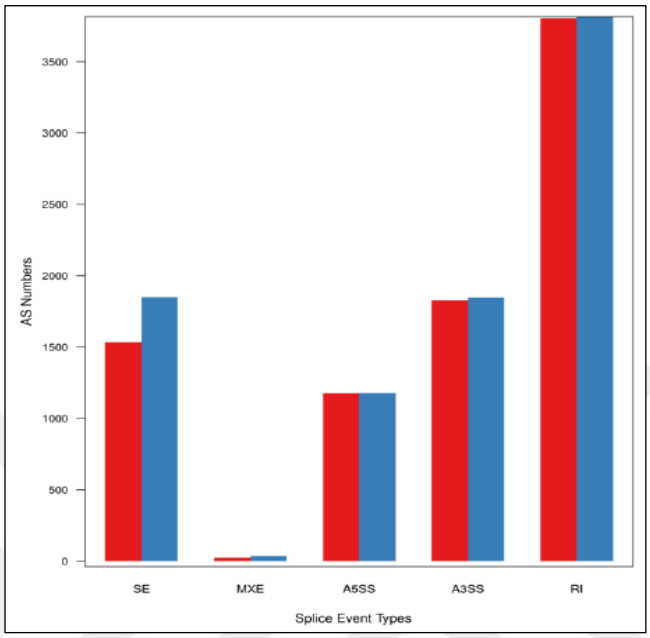


Figure 6. 10 Statistic of splicing

X axis means the type of splicing. Y axis means the amount. Different columns represent different splicing events.

6.1.4.1.6 Gene expression analysis

- **Gene mapping and expression**

After novel transcript detection, novel coding transcripts with reference transcripts were merged to get complete reference. The mapped clean reads and the measured gene expression by **RSEM** software gave totally of 43.541 as gene numbers corresponding to about 21.639 and 21.902 genes, and 33.972 and 34.193 transcripts, for control and CAG-treated sample respectively (Table 6.4).

Tables 6.7 and 6.8 present some genes from the results of gene expression in the two samples.

Table 6.7 Example of expressed genes of control sample

Gene_id	Transcript_id(s)
AT1G01010	AT1G01010.1
AT1G01020	AT1G01020.1, AT1G01020.2, AT1G01020.3, AT1G01020.4, AT1G01020.5, AT1G01020.6
AT1G01030	AT1G01030.1, AT1G01030.2
AT1G01040	AT1G01040.1, AT1G01040.2
AT1G01050	AT1G01050.1, AT1G01050.2
AT1G01060	AT1G01060.1, AT1G01060.2, AT1G01060.3, AT1G01060.4, AT1G01060.5, AT1G01060.6, AT1G01060.7, AT1G01060.8
AT1G01070	AT1G01070.1, AT1G01070.2
AT1G01080	AT1G01080.1, AT1G01080.2, AT1G01080.3
AT1G01090	AT1G01090.1

Table 6.8 Example of expressed genes of treated sample

Gene_id	Transcript_id (s)
AT1G02065	AT1G02065.1, AT1G02065.2
AT1G02090	AT1G02090.1, AT1G02090.2, AT1G02090.3
AT1G02140	AT1G02140.1
AT1G02280	AT1G02280.1, AT1G02280.2
AT1G02300	AT1G01050.1, AT1G01050.2
AT1G02305	AT1G02305.1
AT1G03120	AT1G03120.1, AT1G03120.2
AT1G03687	AT1G03687.1, AT1G03687.2
AT1G03920	AT1G03920.1, AT1G03920.2, AT1G03920.3

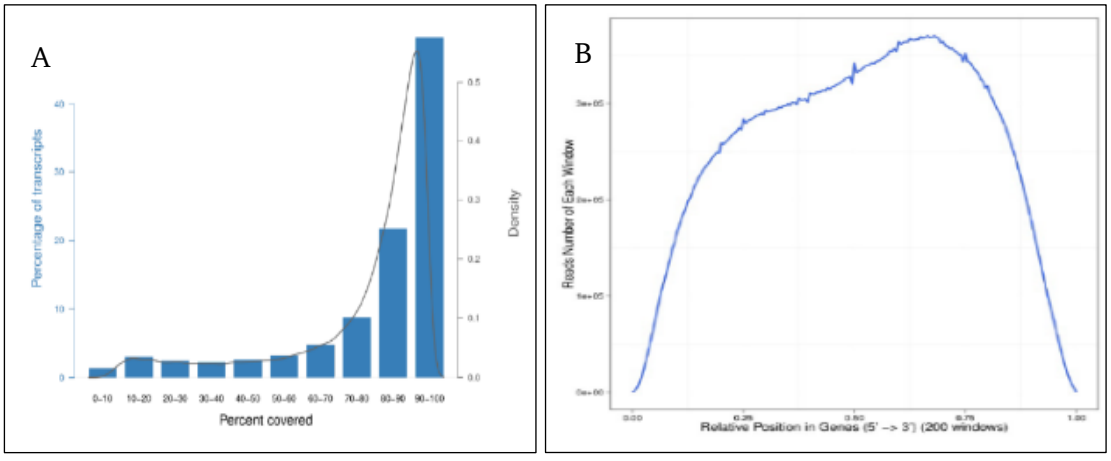


Figure 6. 11 (A) Reads coverage on transcripts **(B)** Reads distribution on transcripts

Reads coverage and reads distribution were calculated for all detected transcripts and presented in Figure 6.11 A and B. To display expressed genes between the two samples, we used Venn diagram as it presented in Figure 6.12.

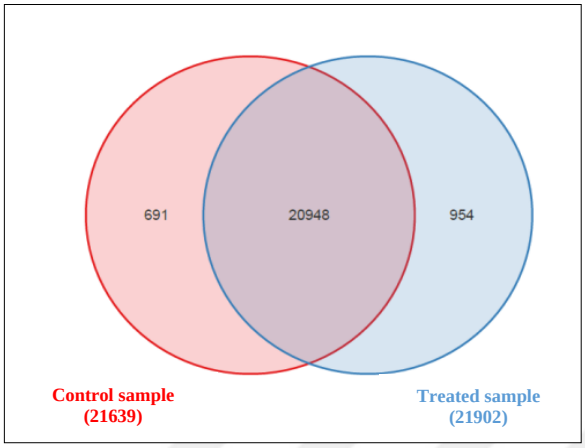


Figure 6. 12 Venn diagram analysis

- **Correlation between samples**

Pearson correlation was calculated between the two samples as it shown in Figure 6.13.

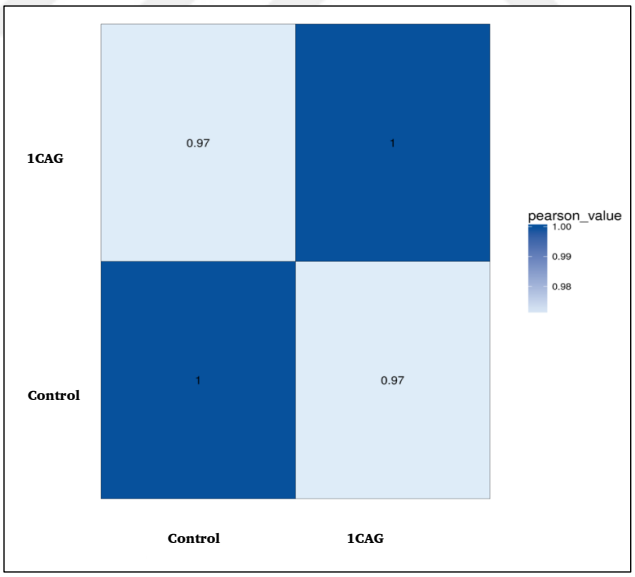


Figure 6. 13 Heatmap of Pearson correlation between control and treated samples

Both X and Y axis represent each sample. Coloring indicate Pearson correlation (high: blue; low: white).

- **Distribution of gene expression**

Based on the expression information, we preformed box plot to show the distribution of the gene expression level of each sample, and we observed the dispersion of the distribution, as shown in Figure 6.14 A.

The density map shows the change of gene abundance and reflect the concentration of gene expression in the sample interval (Figure 6.14 B).

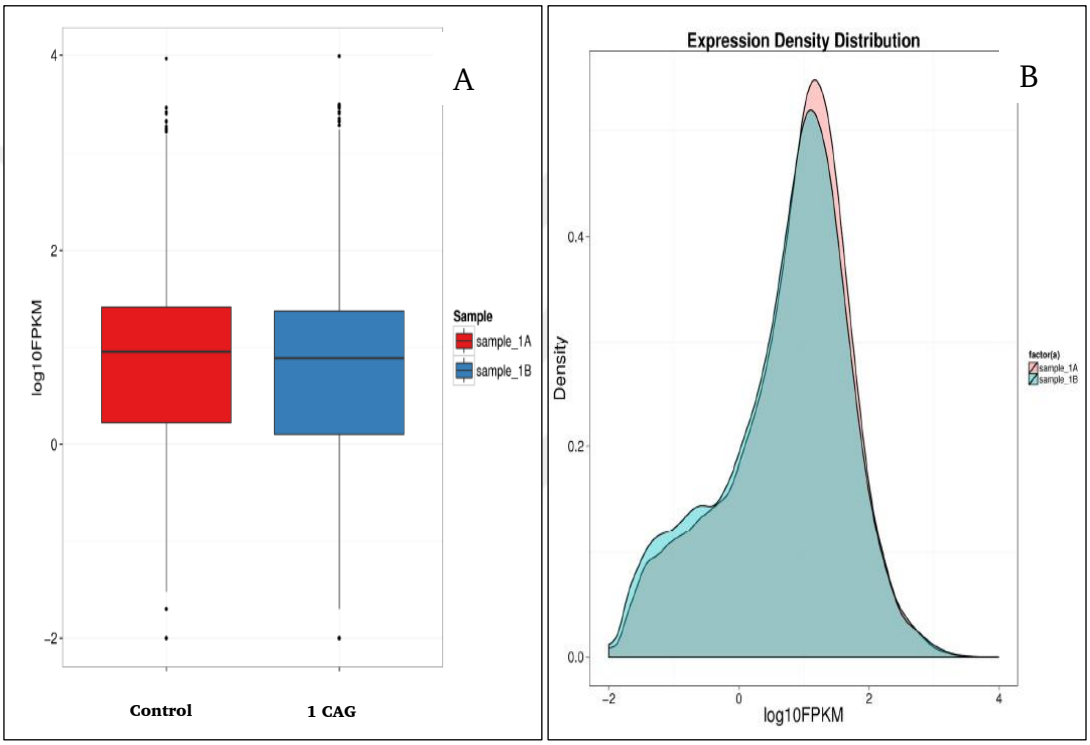


Figure 6. 14 Distribution of gene expression **(A)** Gene expression box-plot **(B)** Gene expression density map

(A) X axis represents the sample name. Y axis represents the log10FPKM value (B) X axis represents the log10FPKM value. Y axis represents the gene density

The gene amount was calculated according to three values of Fragments Per Kilobase Million (FPKM): FPKM ≤ 1, FPKM 1~10 and FPKM ≥ 10. Around 10.000 genes having FPKM value ≥ 10 were determined in treated and non-treated samples (Figure 6.15).

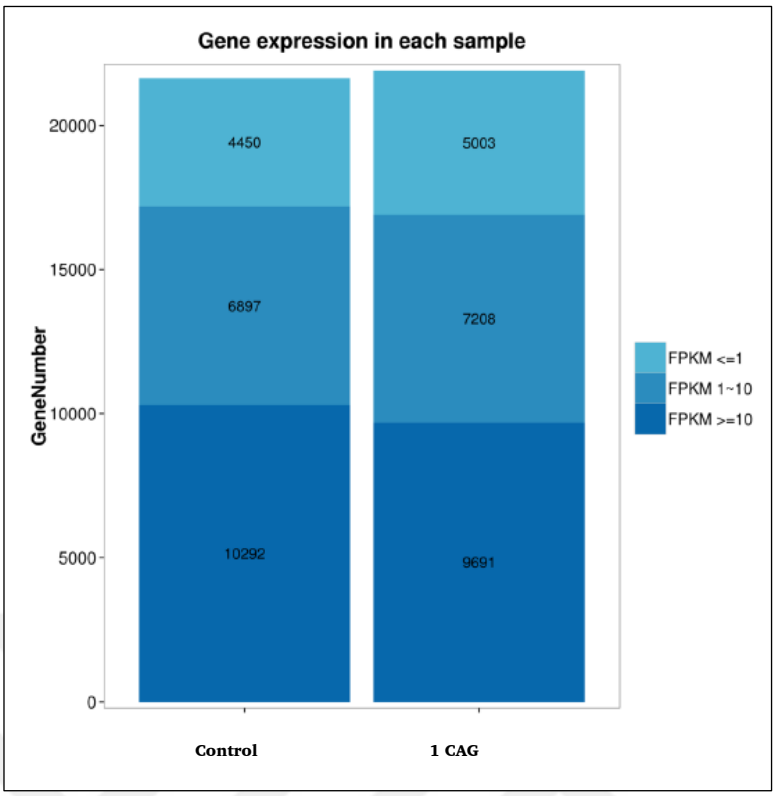


Figure 6. 15 Gene expression distribution

X axis corresponds to the sample name. Y axis corresponds to the gene amount. The dark color = high expression level (FPKM value >= 10, light color = low expression level (FPKM value <= 1)

6.1.4.1.7 Differentially Expressed Genes (DEGs) detection

DEGs between control and treated samples were identified according to the gene expression level. **PossionDis** algorithm was used to detect the DEGs. Some DE genes are presented in the Table 6.9. Totally, 22.593 DEGs were identified and then filtered up to 1045; 184 genes with an FDR≤ 0.001 and log2FoldChange ≥ 1, were remarkably up-regulated by cycloastragenol, while 861 genes with an FDR ≤ 0.001 and log2FoldChange ≤ -1 were reported to be significantly down-regulated as it shown in Figures 6.16 A and B. Based on their size, the filtered DEGs were distributed into 5 groups; 30 unigenes shorter than 500 pb, 191 which the size is longer than 500 and shorter than 1000 pb, 607 unigenes which are between 1000 and 2000 pb and 217unigenes longer than 2000 pb.

Table 6.9 Examples of differentially expressed genes

GeneID	Length	Control Expression	1CAG-Expression	log2FoldChange (1CAG/Control)	FDR	Up/Down-Regulation (1CAG/Control)	P-value	Symbol
AT5G05250	1500	3.02	41.16	3.76862252756889	0	Up	0	AT5G05250// AT5G05250
AT4G16370	2698	14.19	95.02	2.74335661735029	0	Up	0	OPT3
AT1G16030	2570	70.27	142.67	1.02170119966516	0	Up	0	Hsp70b
AT5G19600	2359	29.37	0.33	-6.4757334309664	0	Down	0	SULTR3%3B5
AT5G23010	1729	44.12	0.56	-6.29986215353378	0	Down	0	MAM1
AT1G62360	1704	57.98	2.37	-4.61259636916314	0	Down	0	STM
AT3G05930	926	2.50	1.04	-1.26534456652099	0.000400952171264368	Down	0.0001019018	GLP8
AT5G25160	1228	1.31	0.42	-1.64110557875869	0.00040736753794542	Down	0.0001037306	ZFP3
AT2G03500	2100	0.80	0.27	-1.56704059272389	0.000413499736619229	Down	0.0001054568	AT2G03500// AT2G03500

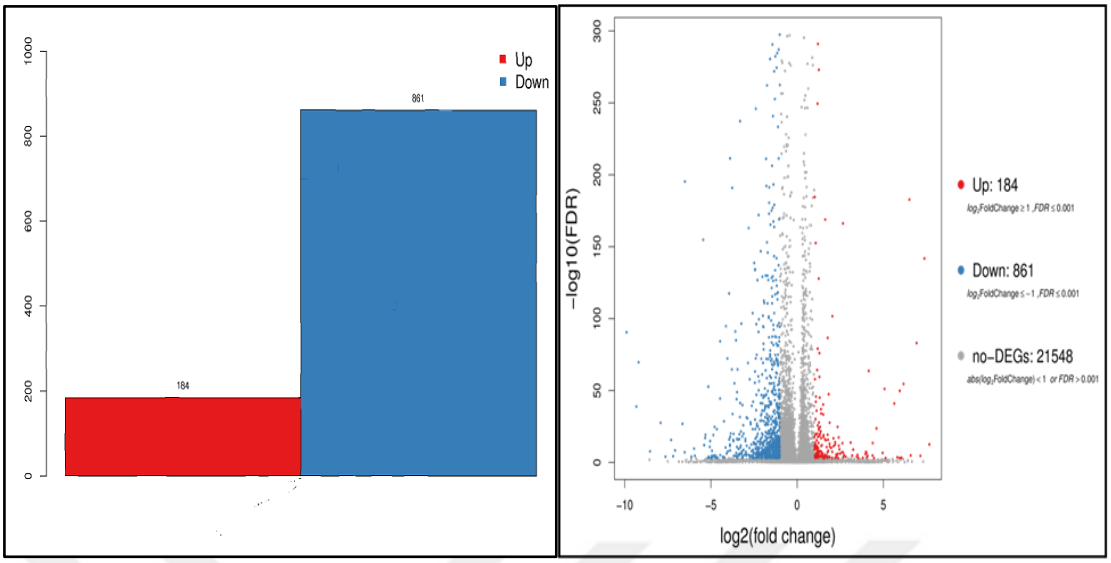


Figure 6. 16 Presentation of DEGs (Left) Summary of DEGs (Right) Volcano plot of DEGs

6.1.4.1.8 Gene Ontology analysis of differentially expressed genes

DEGs were analyzed in Gene Ontology (GO) classification and functional enrichment with the use of **AmigoGO** database.

GO was studied to check the mechanisms to which these belong to and to classify them regarding to their biological process (BP), cellular component (CC) and molecular function (MF). According to GO analysis, DE genes were extremely enriched for different GO processes and function categories. All DE genes were associated to one or more different GO according to the category function. However, some DEGs seem to not be annotated to any GO term. The identified DE genes were associated with 1754 GO terms and organized in 46 functional subcategories. Most of the GO annotations were part of the biological processes with 1139 gene ontology terms and 672 genes (64% from total DEGs), then 427 GO terms with 707 genes (67% from total DEGs) belong to the molecular function and the rest 188 GO terms, with 72% from total DEGs corresponding to 757 genes, were associated with the cellular components.

Twelve subcategories were the most abundant among all GO terms subcategories. They were, based on the percentage of the annotated DEGs: cell (GO:0005623), metabolic process (GO:0008152), cell part (GO: 0032990), cellular process (GO:0009987), catalytic activity (GO:0003824), binding (GO:0005488), membrane (GO:0016020), membrane part (GO:0048501), biological regulation (GO:0065007), response to the stimulus (G0:0050896), regulation of biological process (GO:0050789) and organelle (GO:0043226). The GO classification of up and down-regulated genes showed that all of them were distributed over the three ontologies. However, the majority belonged to cell and cell part GO terms. Figure 6.17 represent the most enriched GOs and the number of their corresponding DE genes. 'Cellular process', 'metabolic process' and 'response to stimulus' ontologies were the most enriched three gene ontologies of the biological process category with, respectively, 466, 448, and 297 unigenes. In the same category, with respectively, 5, 3, and 2 unigenes, the 'rhythmic process', 'locomotion', and 'cell proliferation' were the less enriched ontologies. 'Cell', 'cell part' and 'organelle' are the most three annotated terms of the cellular component category with number of unigenes, respectively, 585, 581 and 384. However, 'extracellular region part' with 9 unigenes, 'nucleoid' with 3 and 'supramolecular complex' with one gene were the three less enriched terms. For the molecular function category, 'signal transducer activity', 'molecular transducer activity', in addition to 'nutrient reservoir activity' were found to be the least three annotated categories having respectively 13, 8 and 7 genes. However, 'binding', 'catalytic activity', together with 'transcription regulator activity' were the three most annotated GO terms with accordingly 422, 401 and 108 unigenes (Figure 16.18).

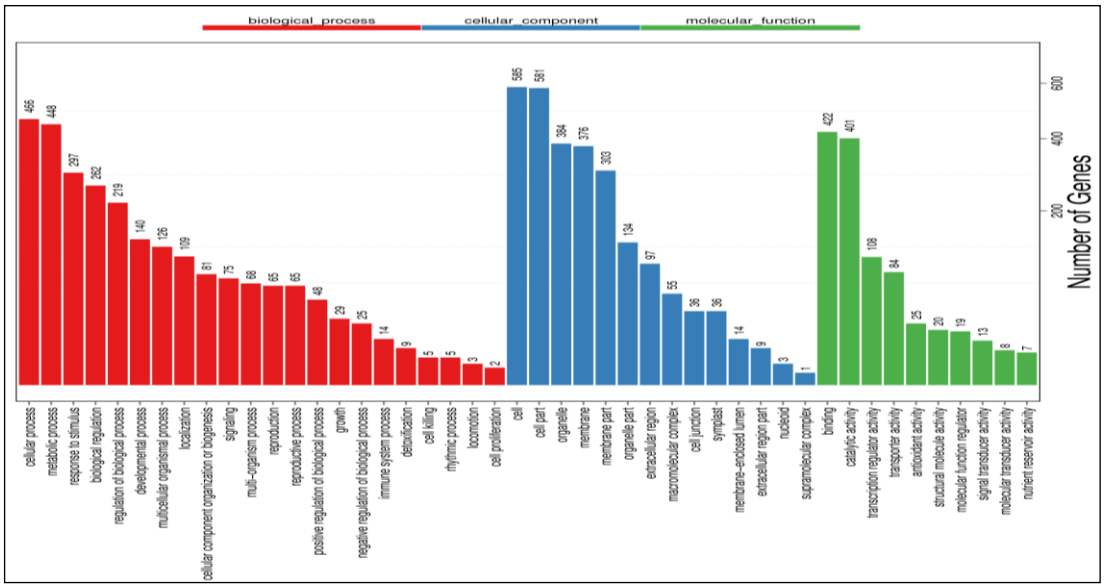


Figure 6. 17 Gene ontology classification

X axis represents number of DEG. Y axis represents GO terms

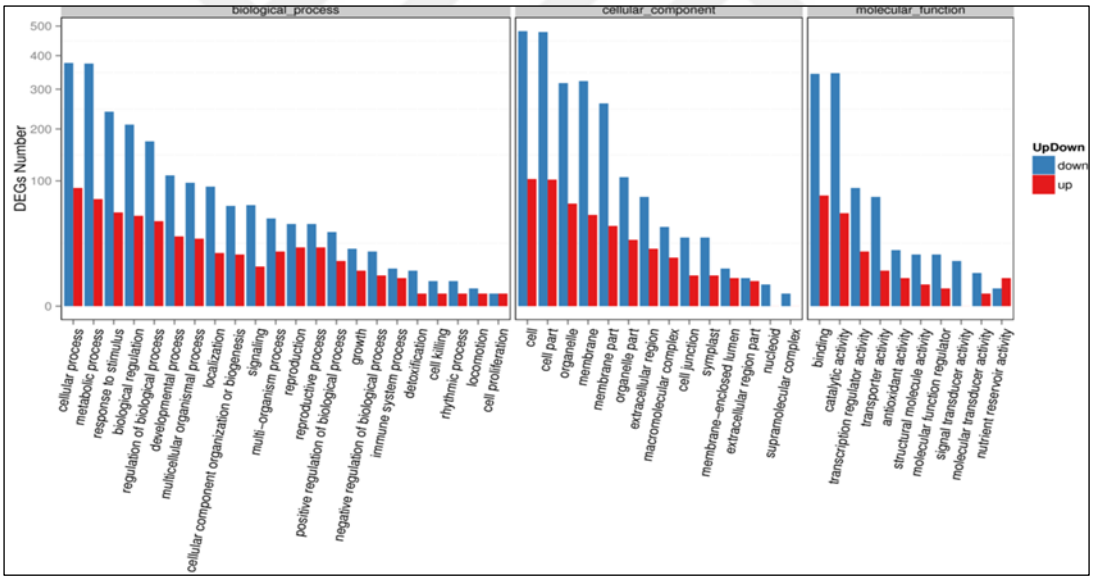


Figure 6. 18 GO classification of up-regulated and down-regulated genes

X axis: GO term. Y axis: the amount of up/down-regulated genes

6.1.4.1.9 KEGG analysis of differentially expressed genes

Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was carried out and displayed 784 assigned unigenes among 20.101 annotated ones to KEGG pathways.

The KEGG pathways are determined according to the ratio of the enriched genes to a pathway and to *p-values* and *q-values* which represent enrichment degree.

The enriched pathways of the present work with the most significant *p-values* are: 'Glucosinolate metabolism' with a *p-value* of 1.743362e-06, 'Riboflavin metabolism' with 0.0002860938, '2-Oxocarboxylic acid metabolism' and its *p-value* is 0.0007007765 and 'Phenylpropanoid biosynthesis' pathway with 0.00130124 (Table 6.10).

Differentially expressed genes were studied using KEGG database to be classified according to the involved pathways and their functions. The analysis revealed that they are assigned to 118 pathways and classified in 5 biological function; Genetic information processing, Cellular processes, Environmental information processing, Organism systems and Metabolism (Figure 6.19 A). Most of these pathways are associated with the metabolism category by 548 DE unigenes. Genetic information processing is the second category with 127 unigenes and with 65 genes the environmental information processing is the third one.

The quotient of foreground value (DEGs number) and the background value (total gene amount), called 'Rich Factor', highlighted three functional pathways as the most enriched ones; plant hormone signal transduction, biosynthesis of secondary metabolites and metabolic pathways (Figure 6.19 B). These results agree with the determined enriched pathways according to the implicated DEGs (Table 6.5). Actually, 'Metabolic pathways' (23.34%, ko01100), 'Biosynthesis of secondary metabolites' (13.52%, ko01110) and 'Plant hormone signal transduction' (4.97%, ko04075, Figure 6.20) are the most abundant KEGG pathways in the present study.

Table 6.10 The most enriched pathway functional results

Pathway	DEGs with pathway annotation (784)	All genes with pathway annotation (20101)	P-value	Q-value	Pathway ID
Glucosinolate metabolism	10 (1.28%)	25 (0.12%)	1.477425e-08	1.743362e-06	ko00966
Riboflavin metabolism	7 (0.89%)	34 (0.17%)	0.0002860938	1.687953e-02	ko00740
2-Oxocarboxylic acid metabolism	13 (1.66%)	118 (0.59%)	0.0007007765	1.687953e-02	ko01210
Phenylpropanoid biosynthesis	30 (3.83%)	424 (2.11%)	0.00130124	3.838658e-02	ko00940
Biosynthesis of secondary metabolites	106 (13.52%)	2078 (10.34%)	0.002334762	5.510038e-02	ko01110
Metabolic pathways	183 (23.34%)	3987 (19.83%)	0.007615617	1.011071e-01	ko01100
Circadian rhythm - plant	17 (2.17%)	238 (1.18%)	0.01232007	1.346494e-01	ko04712
Plant hormone signal transduction	39 (4.97%)	691 (3.44%)	0.01369316	1.346494e-01	ko04075
Protein processing in endoplasmic reticulum’	24 (3.06%)	462 (2.3%)	0.09507395	3.739575e-01	ko04141
MAPK signaling pathway-plant	24 (3.06%)	603 (3%)	0.4892315	9.284486e-01	ko04016
Starch and sucrose metabolism	22 (2.81%)	568 (2.83%)	0.5446477	9.284486e-01	ko00500

DEGs pathway enrichment analysis showed detailed information, representing a map for every pathway and its up and down-regulated genes. The pathway functional enrichment results are presented in Figure 6.19 C.

In the plant hormone signal transduction pathway, genes coding SAUR-like auxin-responsive protein family, EXS (ERD1/XPR1/SYG1) family protein, squamosa promoter binding protein-like 4, basic helix-loop-helix (bHLH) DNA-binding superfamily protein and response to ABA and SALT1 were induced by by cycloastragenol.

Other pathways, like ‘Phenylpropanoid biosynthesis’ enriched by 3.83% (ko00940, Figure 6.20), ‘MAPK signaling pathway-plant’ with an enrichment of 3.06%,

(ko04016, Figure 6.21), ‘Protein processing in endoplasmic reticulum’ (3.06%, ko04141) and ’Starch and sucrose metabolism’ (2.81%, ko00500) were annotated to a significant DEGs number. Actually, genes related to phenylpropanoid biosynthesis are implicated in phenolic acids synthesis and known to be one of the most significant secondary metabolisms in plants. Same for the genes associated with the metabolism of starch and sucrose, they play a role in carbohydrate metabolism and proved to play a role in the formation of cell wall [111], [112].

A high number of genes are annotated to ‘Glucosinolate metabolism’ pathway (1.28%, ko00966) which is associated with metabolism processing and considered as an essential pathway of plant secondary metabolism. Several researches work reported the implication of glucosinolates and glucosinolate-derived metabolites in plant metabolism, growth, and their defense mechanisms.

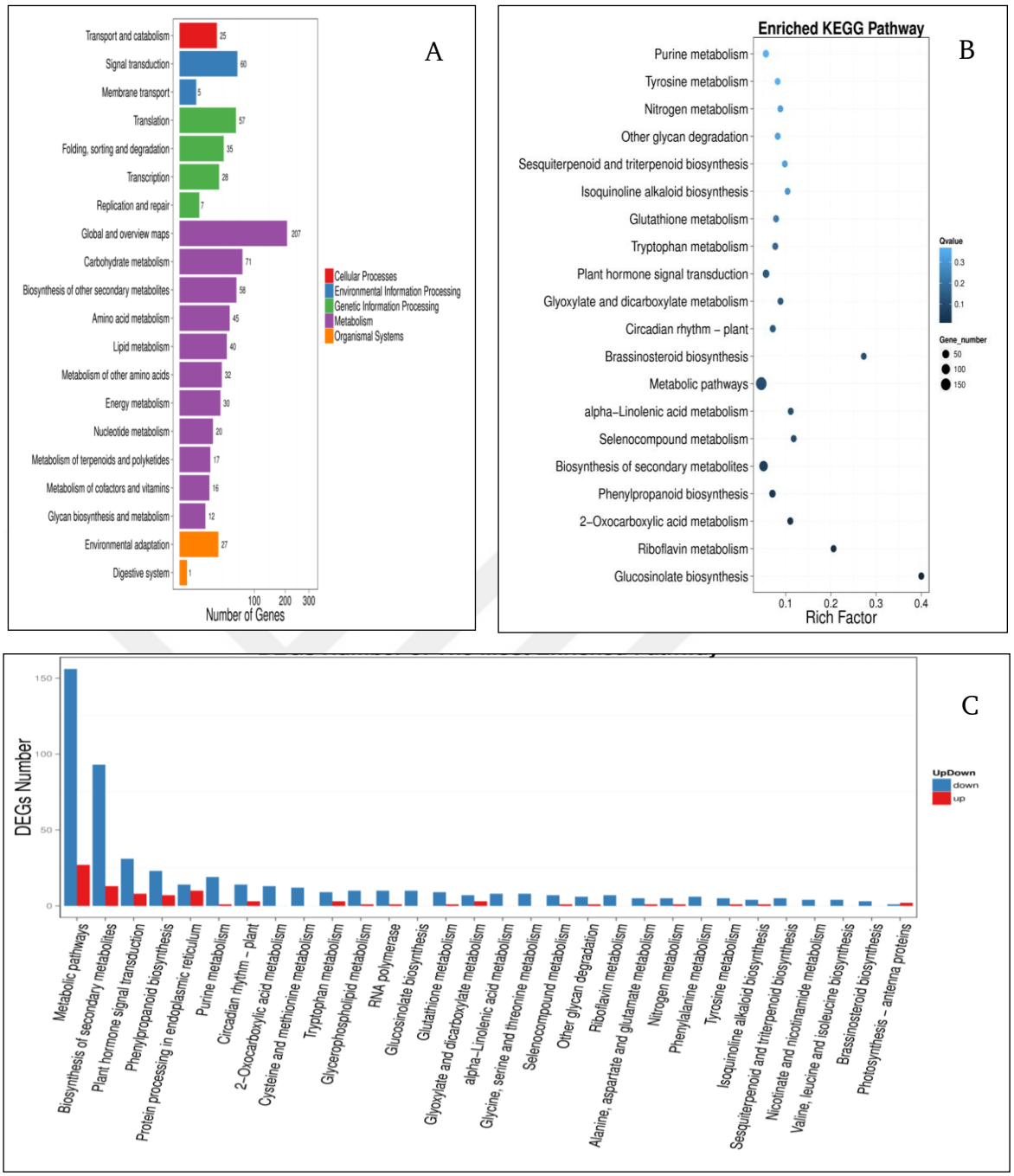


Figure 6. 19 KEGG pathway classification and functional enrichment

(A) The pathway classification of DEGs results. X axis represents number of DEG. Y axis represents functional classification of KEGG. **(B)** The pathway functional enrichment results. X axis: enrichment factor. Y axis: pathway name. the q-value is indicated by color (high: white, low: blue), the lower q-value corresponds to the more significant enrichment. Point size indicates DEG number (The bigger dots refer to larger amount). **(C)** The pathway functional enrichment result for up/down regulation genes. X axis: the terms of Pathway. Y axis: the number of up/down regulation genes.

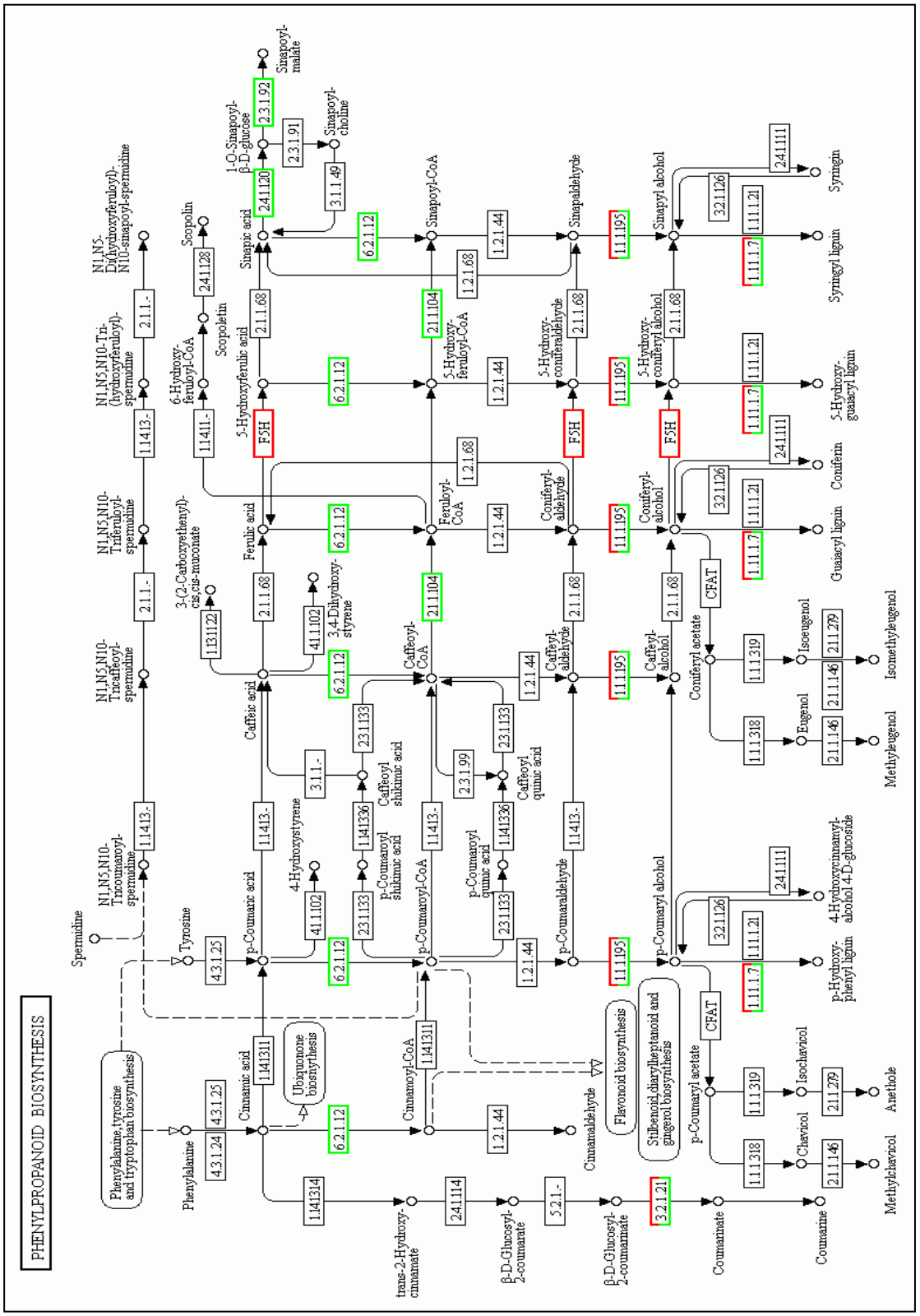


Figure 6. 20 Phenylpropanoid biosynthesis pathway

Red boxes and green boxes represent up-regulated and down-regulated genes, respectively under CAG treatment

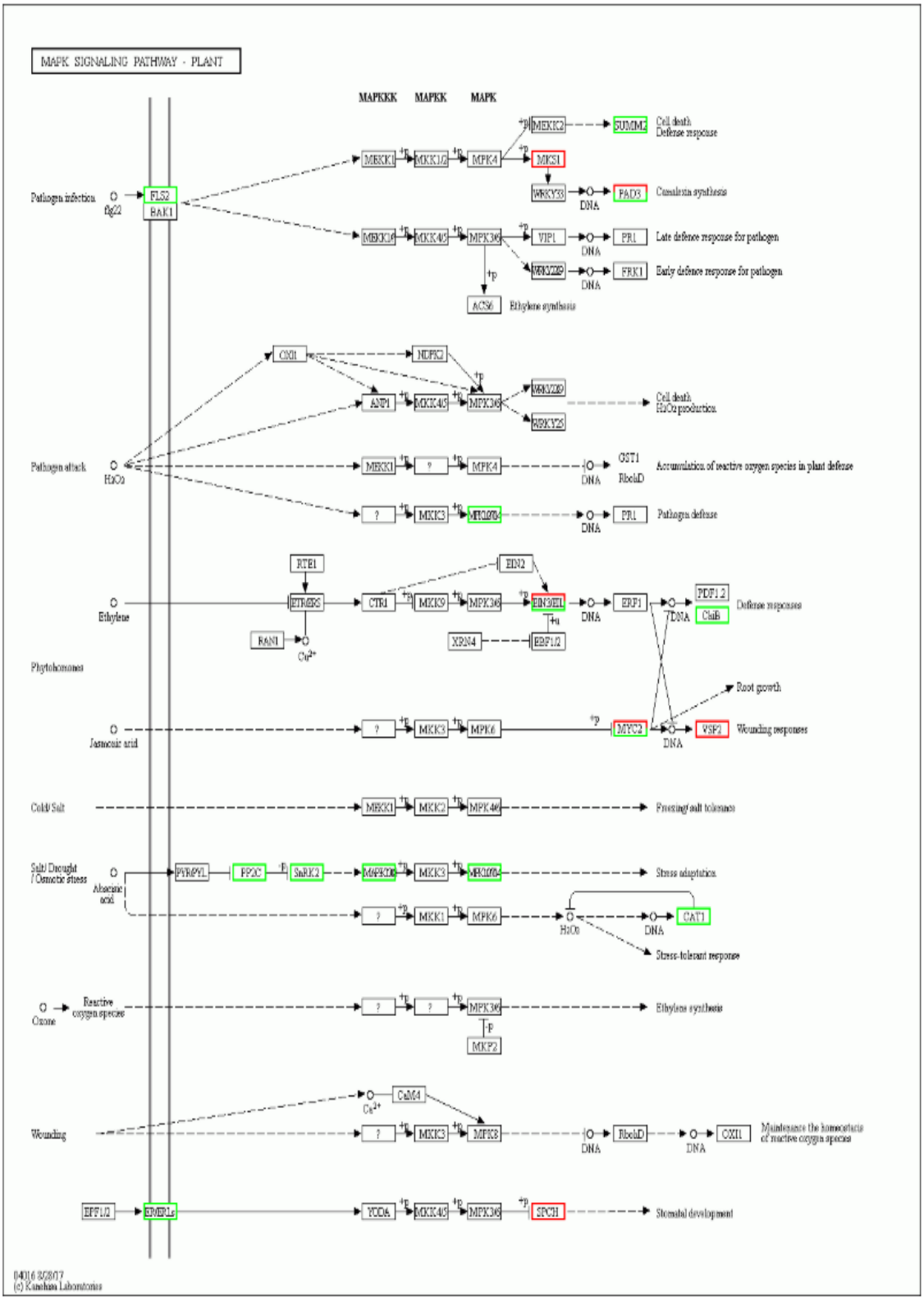


Figure 6. 21 MAPK signaling-plant pathway

Red boxes and green boxes represent up-regulated and down-regulated genes, respectively under CAG treatment.

6.1.4.1.10 Transcription factors prediction

Totally, 2000 transcription factors were determined in *Arabidopsis thaliana* calli treated with CAG, classified in 60 different TF common families (Figure 6.22). TFs that were expressed differentially were counted to 240 (Table 6.11).

The largest group was the MYB family with 276 TFs presenting around 13.8% from the total TFs number, followed by MYB-related family with 229 TFs (11.45%), the AP2_EREBP family (146, 7.3%) and the bHLH family with 141, around 7.05% from the total identified TFs.

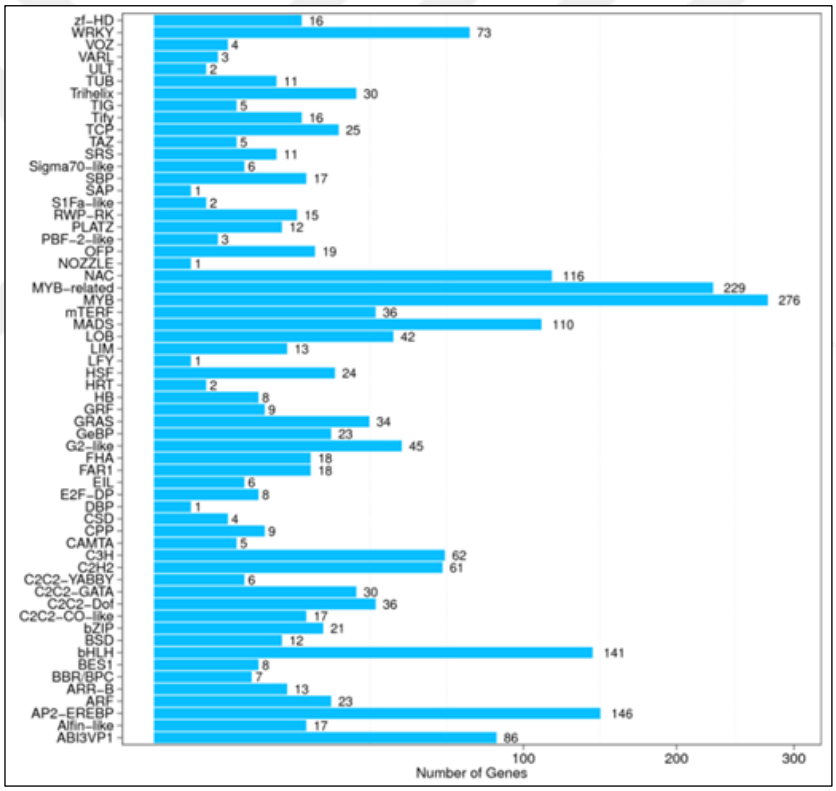


Figure 6. 22 Gene classification of TF families

Table 6.11 Examples of differentially expressed TFs

Gene ID	Control-Expression	1CAG-Expression	log2FoldChange (Control/1CAG)	P-value	TF family
AT3G56980	0.10	6.18	5.94953493301701	6.0398e-52	bHLH
AT2G41240	0.11	5.46	5.63332552228256	4.85362e-43	bHLH
AT2G41240	0.11	5.46	5.63332552228256	4.85362e-43	bHLH
AT1G53160	0.01	0.35	5.12928301694497	0.000267744	SBP
AT1G53160	0.01	0.35	5.12928301694497	0.000267744	SBP
AT1G53160	0.01	0.35	5.12928301694497	0.000267744	SBP
AT3G56970	0.23	7.70	5.06515267952525	2.50786e-53	bHLH
AT1G26870	0.08	0.73	3.18982455888002	3.69168e-06	NAC
AT1G62085	0.10	0.66	2.72246602447109	1.592166e-06	mTERF
AT1G62085	0.10	0.66	2.72246602447109	1.592166e-06	mTERF
AT1G67260	0.39	2.41	2.62748711736771	4.22712e-15	TCP
AT1G67260	0.39	2.41	2.62748711736771	4.22712e-15	TCP

6.1.4.2 miRNA analysis: SmallRNA sequencing

Small RNA sequencing was performed to study the expression profiles of microRNAs, increase the understanding of how CAG would regulate cells and identify the differential expression of small RNA particularly microRNAs caused by cycloatragenol.

In this frame, two cDNA libraries were prepared using total RNA isolated from our treated and non-treated calli and sequenced.

The high throughput sequencing of the two libraries using BGISEQ-500 platform produced a total of 69.835.028 raw tags and detected a number of known and novel small non-coding RNAs.

6.1.4.2.1 Data statistics

The sequencing generated totally 96.835.028 raw reads corresponding to 35.061.175 and 34.773.853 for control and CAG-treated sample, respectively. Before carrying the data analysis, the generated data was cleaned and filtered from all contaminant tags to produce, respectively, 32.895.000 and 32.191.888 clean reads for non-treated and treated samples (Table 6.12). The distribution of

base quality on clean tags is presented in Figure 6.23. Length distribution analysis allowed to see the composition of a sRNA sample, as it shown in Figure 6.24. Generally, the length of small RNA is found to be between 18 nt and 30 nt with a majority of 21 to 24 nt sequences.

Table 6.12 Summary of the sequencing data in the two samples

Samples	Raw tag count	Clean tag count	Percentage %
Control	35.061.175	32.895.000	93.82
1 CAG	34.773.853	32.191.888	92.57

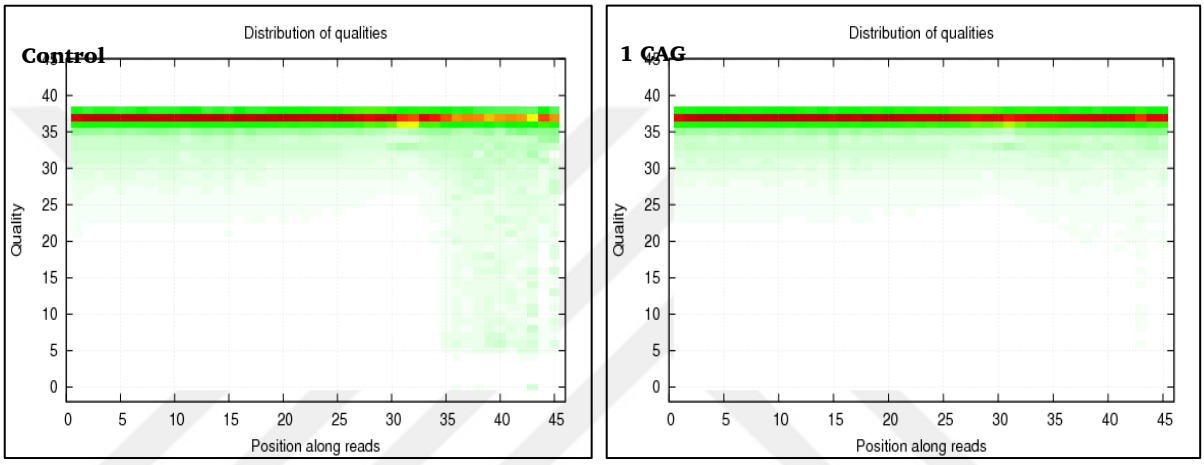


Figure 6. 23 Distribution of base quality on clean tags

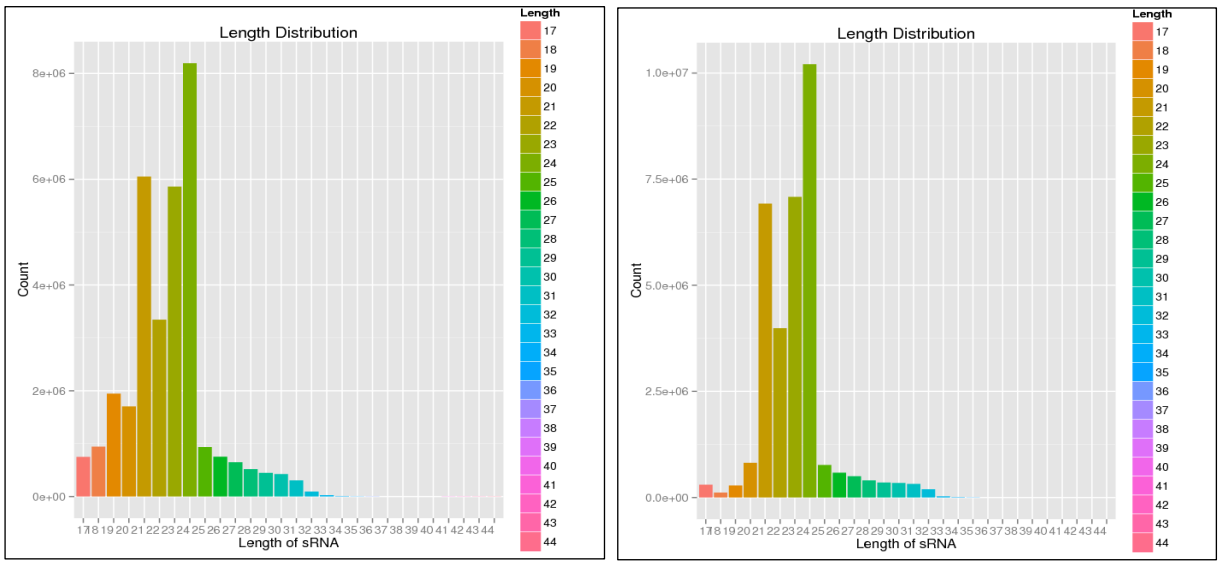


Figure 6. 24 Length distribution of sRNAs in control sample (left) and treated one (right)

Table 6.13 summarizes the detected sncRNAs for each sample. Actually, 1117 micro RNAs were detected; 266 known miRNAs and 298 novel ones in control sample against 261 known and 292 novel miRNAs in CAG-treated sample. In addition, 4648 new small interference RNAs were identified while no piRNA or known siRNA were detected.

Table 6.13 Summary of detected sncRNAs for control and treated sample

Samples	Known miRNA count	Novel miRNA count	Known piRNA count	Novel piRNA count	Known siRNA count	Novel siRNA count
Control	266	298	0	0	0	3180
1 CAG	262	292	0	0	0	1486

6.1.4.2.2 sRNA annotation

Filtered tags were mapped to the used sRNA database **miRbase**. Respectively, for control and treated sample, 96% and 95% of clean tags were mapped corresponding to 31.783.363 and 30.588.444 mapped tags were determined (Table 6.14).

Table 6.14 Statistics of tags alignment to reference genome

Samples	Clean tag count	Mapped tag	Percentage %
Control	32.895.000	31.783.363	96.62
1 CAG	32.191.888	30.588.444	95.02

To make every unique small RNA mapped to only one annotation, we followed the following priority rule: miRNA > piRNA > snoRNA > Rfam > other sRNA. Proportions of all obtained sRNAs are shown in Figure 6.25.

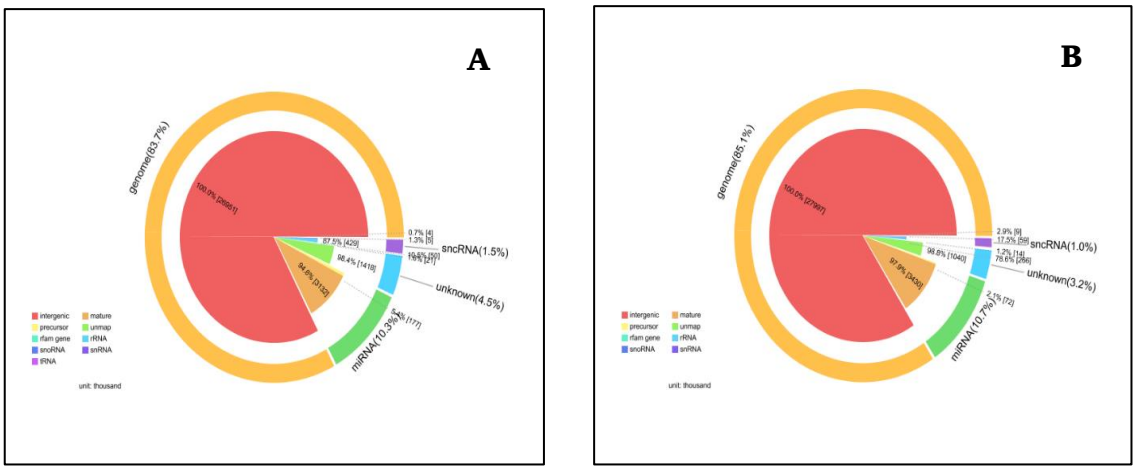


Figure 6. 25 The proportion of sncRNAs in control (**A**) and in treated (**B**) samples

6.1.4.2.3 sRNA prediction

After sRNA annotation, the unknown tags were used to predict novel sRNA based on their architectural features. An example for the stem loop structure of precursors of predicted miRNA is shown in Figure 6.26. The structure of each candidate miRNA was stored in precursor.tar.gz.

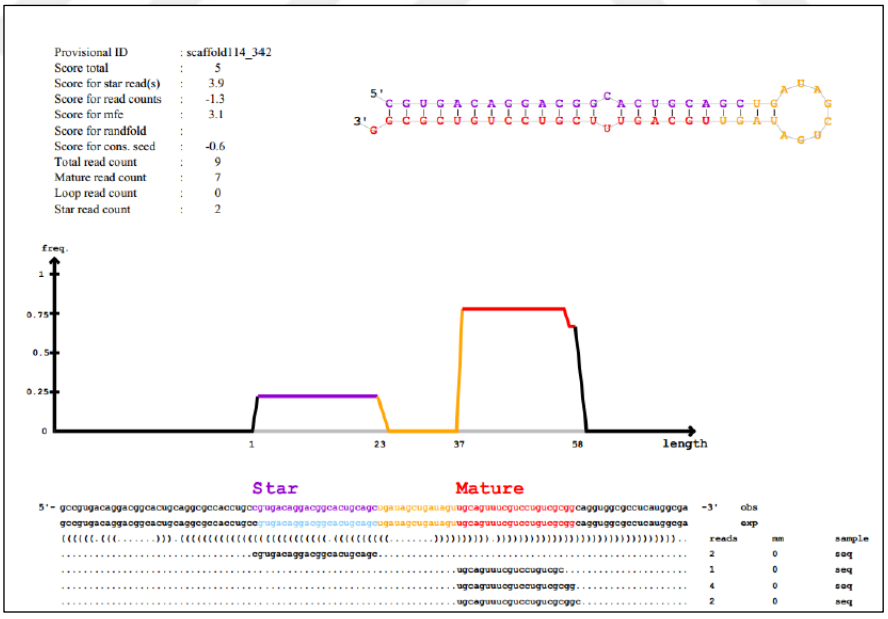


Figure 6. 26 The stem loop structure of precursors of predicted miRNA

Novel miRNAs were predicted as the Table 6.15 shows some of them and the details of their sequences, while the Table 6.16 give some of the predicted novel siRNAs.

Table 6.15 Characteristic of predicted novel microRNAs samples

miRNA id	Chr	Strand	Seq(mature)	Seq(star)	Start	End
novel_ mir1	1	_	TTTTCTTAATAGACTTA ATGGGCTACAGAG	TGTATAAATAGTTTTTTATTT TATTTTT	3091 74	3093 49
novel_ mir2	1	+	TGGACATGGTCTTTAT TGGGCAT	TTTTTCGGACCATTGTCCAT TA	4319 02	4321 35
novel_ mir3	1	+	ATCGTTGATTCGGCCA AATCGACTCACCAT	GGTGAGTCGATTTGGCCGAG TCAACGATA	5127 79	5130 17
novel_ mir4	1	+	CCGTATGGACAGTTTT TTTTATGGGCCAA	TGCCCAATAAAGACCATGTC CATATGGGCA	1430 268	1430 421

Chr: chromosome
Seq:sequence

Table 6.16 Characteristic of predicted novel small interference RNAs samples

miRNA id	Guide Sequence	Passenger Sequence	Tag info
novel_sir1	CTTTTGATCGATCTATCTAAA	AAACTAGCTAGATAGATTTTT	tag9514
novel_sir2	AAAGATTTGGTACGTCGTTGA	AAACCCATGATCCGGCGTTAA	tag65344
novel_sir3	CCTTTGGGTACTAGGCCGCAA	AAACCCATGATCCGGCGTTAA	tag286664
novel_sir4	GGTTTTTTTTAACAACGGCAA	AAAAAAAATTGTTGCCGTTGG	tag119731

Unknown sncRNAs were also detected (Table 6.17).

Table 6.17 Characteristic of predicted unknown sncRNAs samples

miRNA id	Ch	Start	End	Control sample	1CAG sample	Sequence
Tag1	2	2606385	2606408	11	0	TCATCGTTGGGT ATGGTCTTTGGT
Tag2	1	257723	257746	77	68	AAAAAACCCTCA ACTTTGGCCAGA
Tag3	3	7232052	7232074	12	0	TGGCGTGTCTTT GAAGGGAAATA
Tag4	1	9852790	9852813	18	0	GATTAGAACTGT CAATTGGGTGGA

6.1.4.2.4 sRNA Expression

Small RNA expression level was measured in Transcripts Per Million (TPM). The Table 6.18 below shows the expression of some sRNAs of control and treated samples.

Table 6.18 Expression of some miRNAs & siRNAs in control and treated samples

Control				**1 CAG**			
sRNA id	**Count**	**Type**	**TPM**	**sRNA id**	**Count**	**Type**	**TPM**
novel_mir1	6200	miRNA	180.88	novel_mir1	9640	miRNA	296.84
novel_mir2	1355	miRNA	39.53	novel_mir2	743	miRNA	22.88
novel_mir3	4994	miRNA	145.69	novel_mir3	3244	miRNA	99.89
novel_sir1	18	siRNA	0.53	novel_sir4	98	siRNA	3.02
novel_sir3	41	siRNA	1.20	novel_sir5	81	siRNA	2.49
novel_sir4	179	siRNA	5.22	novel_sir19	148	siRNA	4.56

- **Expression analysis of known miRNAs**

A total of 273 known micro RNAs were detected; 261 known miRNAs, belonging to 86 known miRNA families, in control sample and 259 known miRNAs, that belong to 83 known miRNA families, in CAG-treated sample. A total of 80 miRNAs families, corresponding to 247 common miRNAs, were commonly expressed in treated and non-treated samples (Figure 6.27a). However, 4 miRNA families were specifically expressed in the control sample; MIR5020, MIR4245, MIR847 and MIR869, while 3 miRNA families; MIR854 MIR8167 MIR4240, were specific to the treated sample. Actually, around 92% and 94% of the totally expressed miRNAs in, respectively, control and CAG-treated sample were found to be common. 14 miRNAs were uniquely expressed in control sample; miR847, miR416, miR5630b, miR156j, miR164c-3p, miR407, miR8173, miR156f-3p, miR4245, miR869.2, miR5020c, miR5639-3p, miR399e and miR5661 when 12 miRNAs were expressed uniquely in the treated one; miR5665, miR4240, miR2936, miR865-3p, miR5636, miR1888a, miR5023, miR5666, miR5014b, miR2112-5p, miR831-5p and miR5632-3p.

- **Expression analysis of novel miRNAs**

Among the totally 298 novels expressed miRNAs, 292 miRNAs were expressed in both samples, 6 were specifically expressed in control sample which are; mirn123, mirn127, mirn122, mirn143, mirn58 and mirn169 while no novel miRNA was found to be unique to CAG-treated one (Figure 27b). The majority of these novel miRNAs had mature sequences varying between 26 and 28 nt. They were found to be distributed in the five chromosomes of Arabidopsis thaliana, but most of them were located in chromosome 1, 3 and 5.

6.1.4.2.5 miRNA target prediction

The two software **psRobot** and **TargetFinder** were used to identify target genes of miRNAs and extract their intersection or union. Target result are shows in Table 6.19.

Table 6.19 Statistics of target genes

MiRNA count	Target count	Target Finder target	psRobot
455	8550	3015	7538

Referring to the genome of *A. thaliana*, a total of 10.928 target genes, filtered up to 6160 were identified. Among identified filtered targets, 5803 genes were targets of known miRNAs, 357 by novel miRNAs and 51 genes were commonly targeted by both types of miRNAs (Figure 27c). Some of the genes were targeted by more than two know and/or novel miRNAs.

Actually, applying the appropriate filters like MFE on the intersection targets was set for the further analysis. The targets genes result before and after filtering are shown in Figure 6.28 and details of some targets are presented in Table 6.20.

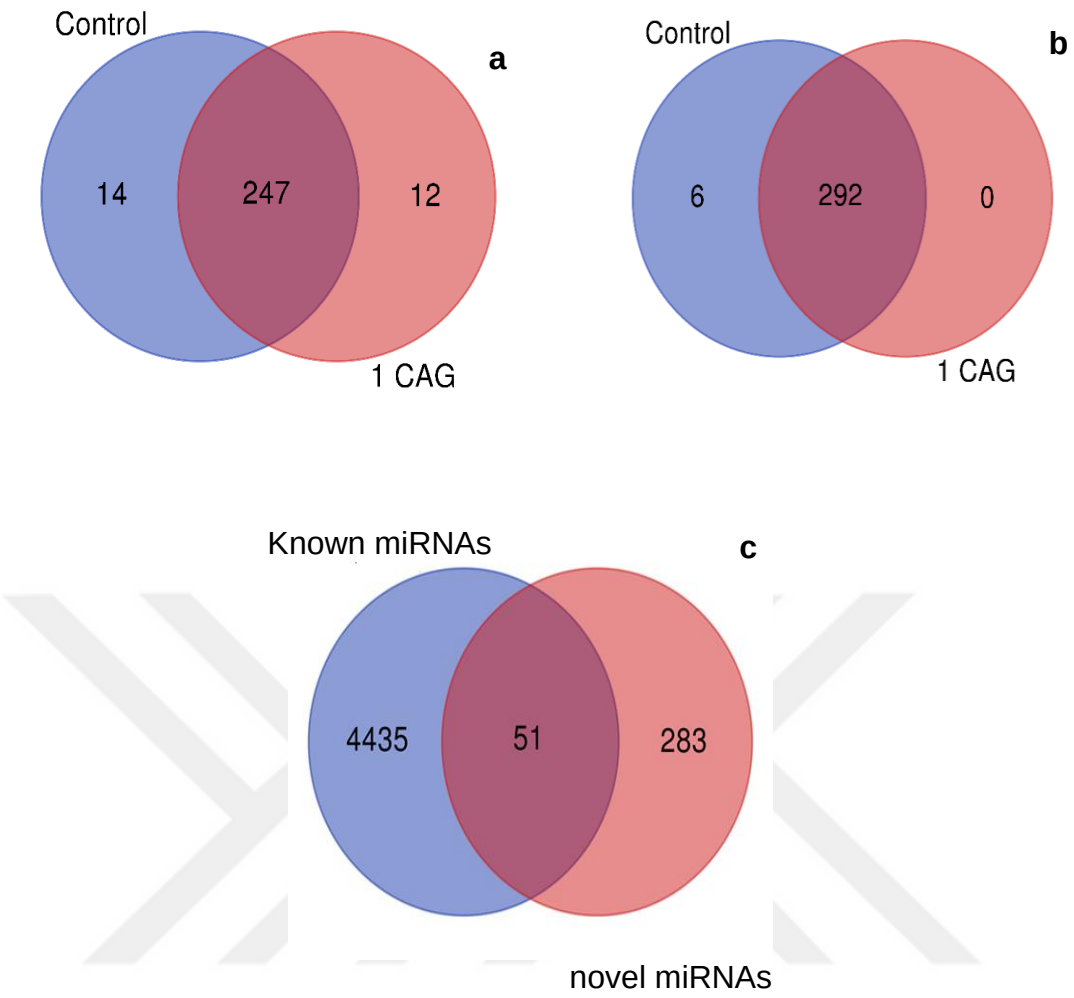


Figure 6. 27 Common and unique (**a**) known miRNAs (**b**) novel miRNAs and (**c**) target genes between control and CAG-treated libraries

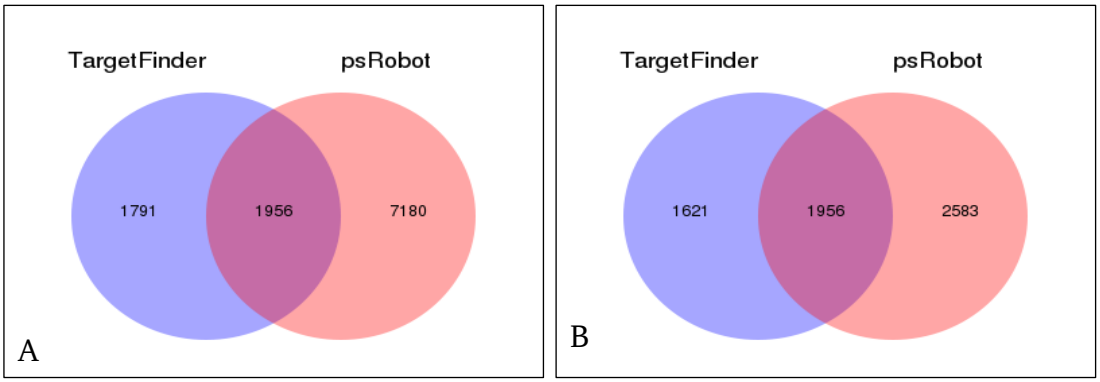


Figure 6. 28 Venn statistics of (A) target predictors and (B) filtered targets

Table 6. 20 Examples of some target genes from filtered results

miRNA id	Target id	TargetFinder score	psRobot score	NR evalue
miR10515	AT2G20610.1	3	1.5	3.6e-73
miR833b	AT4G15090.1	NA	2.5	9.7e-26
miR835-3p	AT4G19600.1	4	2.5	3.7e-43
miR156a-3p	AT3G23930.1	NA	2.5	1.6e-10
miR156a-5p	AT3G51910.1	4	NA	2.4e-22
miR156g	AT1G25500.2	NA	2.5	2.8e-42
miR156h	AT1G09170.6	4	2.5	1.7e-72
miR837-5p	AT1G52590.1	4	2.5	1.2e-09
miR838	AT1G67230.1	NA	1.8	2.0e-47
miR841a-3p	AT5G28320.1	4	2.2	9.6e-09
miR842	AT2G37340.1	3	2.0	1.2e-32
miR870-3p	AT5G20360.4	NA	2.5	3.5e-15
mirn230	AT1G06220.1	3.5	NA	0.0e+00
mirn136	AT5G26710.1	4	NA	6.4e-185

6.1.4.2.6 Screening differentially expressed sRNAs

DEGseq software was used to found out the differentially expressed small RNAs between our two samples that will be used for further analysis.

- **Differentially expressed siRNAs**

The Figure 6.28 below shows the number of up and down regulated DE siRNAs in each pairwise.

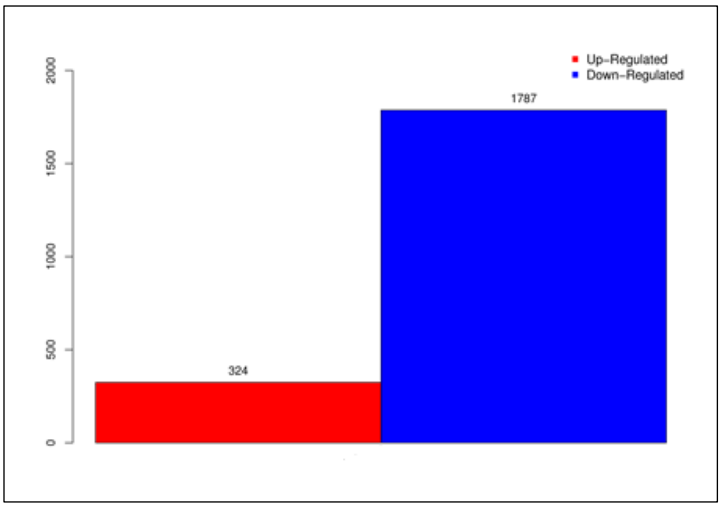


Figure 6. 29 Differentially expressed siRNA

- **Differentially expressed miRNAs**

The number of DE miRNAs is summarized in the Figure 6.30. The total number of differentially expressed miRNAs is 119, among which 81 are found to be down-regulated and 38 were up-regulated.

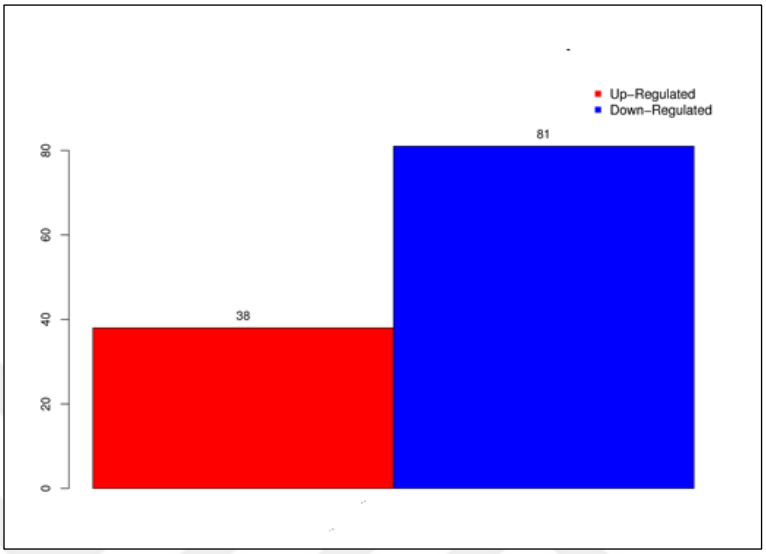


Figure 6. 30 Differentially expressed miRNAs

➢ Differentially expressed known miRNAs

Among the total known miRNA number, 59 ones were found to be differentially expressed between the two samples (Table 21). Actually, microRNAs with expression level changes ≥ 1.0-fold change were considered as differentially expressed while miRNAs that showed variation in expression level ≤ 1.0-fold change were identified in baseline-expression zone. DE miRNAs were classified into 22 up-regulated and 37 down-regulated ones. The microRNA miR2112-3p were the most significantly up-regulated known miRNA by 5.88-fold change while the miRNAs (miR172c, miR172d-3p) from MIR172 family were the less up-regulated ones with a fold change of 1.02. miR2112-3p was followed by (miR394a, miR394b-5p) from MIR394 family, miR2112-5p, (miR160a-5p, miR160b, miR160c-5p) from MIR160 family, miR401, miR398a-5p, miR845a and miR5654-5p were found to be significantly up-regulated by a fold change of, respectively, 4.75, 4.53, 4.39, 4.21, 3.67, 3;09 and 3.07. the most notably down-regulated known miRNA is miR399d by 6.84-fold change, followed by

(miR2111a-5p, miR2111b-5p) from MIR2111 family, miR156f-3p, miR399a, miR399e, miR399c-5p and miR2111a-3p by, respectively, 6.42, 5.31, 4.89, 4.69, 4.44 and 4.10.

Table 6.21 Differentially expressed known miRNAs after CAG treatment

miRNA id	**Expression**		**Fold Change** (log2Ratio(1CAG/Control))
	Control	**1 CAG**	
miR160a-5p, miR160b, miR160c-5p	306,44	5871,43	4,40
miR158a-3p	880,93	3531,24	2,14
miR157a-5p, miR157b-5p, miR157c-5p	1854,41	532,89	-1,66
miR319a, miR319b	113,40	711,98	2,79
miR845a	38,22	297,24	3,09
miR408-5p	668,37	259,09	-1,23
miR158b	353,53	666,04	1,05
miR827	273,74	76,46	-1,70
miR396b-5p	225,81	63,34	-1,70
miR399f	88,02	6,28	-3,67
miR394a, miR394b-5p	1,55	37,94	4,75
miR399b, miR399c-3p	95,28	24,30	-1,84
miR399d	35,04	0,28	-6,85
miR774a	21,09	67,44	1,81
miR399c-5p	30,28	1,26	-4,45
miR390b-3p	39,09	82,22	1,21
miR399a	11,09	0,34	-4,90
miR163	58,73	25,99	-1,04
miR160c-3p	25,12	8,96	-1,35
miR397a	11,90	2,46	-2,14
miR2111a-5p, miR2111b-5p	5,81	0,06	-6,42
miR156a-3p	8,96	1,51	-2,43
miR159b-5p	5,13	12,99	1,47
miR778	5,57	0,86	-2,56
miR831-3p	4,64	0,59	-2,85
miR5654-5p	0,50	3,82	3,08
miR8183	15,14	6,28	-1,13
miR166e-5p	8,26	2,40	-1,65
miR845b	0,73	3,66	2,46
miR5663-3p	7,00	2,34	-1,45
miR2112-3p	0,03	1,57	5,89
miR393a-5p, miR393b-5p	0,93	3,20	1,91
miR826a	6,51	2,59	-1,20
miR830-5p	8,49	3,85	-1,01
miR5024-5p	4,17	1,35	-1,49
miR5663-5p	1,08	3,08	1,65

Table 6.21 Differentially expressed known miRNAs after CAG treatment (cont.)

miR396b-3p	2,07	0,37	-2,35
miR169f-3p	0,96	2,37	1,44
miR839-5p	1,20	2,68	1,30
miR156f-3p	0,67	0,00	-5,31
miR171c-5p	1,02	0,15	-2,59
miR395a, miR395d, miR395e	0,44	1,39	1,80
miR2111a-3p	0,58	0,03	-4,11
miR3434-5p	0,55	1,48	1,55
miR833a-3p	2,74	1,17	-1,09
miR1886.1	0,93	0,18	-2,20
miR5650	0,67	0,09	-2,73
miR863-3p	0,67	0,09	-2,73
miR399e	0,44	0,00	-4,69
miR169d, miR169e, miR169f-5p, miR169g-5p	0,26	0,89	1,90
miR837-5p	1,78	0,71	-1,19
miR866-3p	0,09	0,52	2,72
miR5659	1,23	0,40	-1,48
miR172c, miR172d-3p	0,96	1,79	1,03
miR2112-5p	0,00	0,31	4,54
miR398a-5p	0,03	0,34	3,67
miR8173	0,29	0,00	-4,11
miR401	0,00	0,25	4,21
miR171a-3p	0,99	0,37	-1,29

- Differentially expressed novel miRNAs

Among the total novel identified miRNAs, 60 were observed to be differentially expressed between the two samples; 14 were up-regulated under CAG while 46 were down-regulated (Table 22). The most significantly up-regulated miRNAs were miRn14, miRn216, miRn231 and miRn275 with a fold change of, respectively, 1.74, 1.57, 1.51 and 1.41, and the most significantly down-regulated ones are miRn123 with 10.06-fold change, miRn58, miRn127and miRn169 with 9.42, 9.33 and 8.27, respectively.

Table 6. 22 Differentially expressed novel miRNAs after CAG treatment

miRNA id	Expression		Fold Change (log2Ratio(1CAG/Control))
	Control	1 CAG	
miRn183	222,39	449,54	1,15
miRn48	32,67	0,43	-6,11
miRn9	60,39	130,44	1,25
miRn59	36,41	2,46	-3,75
miRn27	24,94	1,94	-3,55
miRn122	33,14	0,00	-10,94
miRn62	27,10	2,86	-3,11
miRn267	22,23	1,32	-3,93
miRn171	99,98	41,38	-1,14
miRn42	40,96	8,84	-2,08
miRn298	57,41	17,18	-1,61
miRn208	93,97	38,95	-1,14
miRn117	105,08	45,97	-1,06
miRn102	37,31	77,75	1,19
miRn162	25,53	3,48	-2,74
miRn163	84,02	36,09	-1,08
miRn225	68,88	26,82	-1,23
miRn19	18,93	1,85	-3,22
miRn45	15,08	0,77	-4,16
miRn25	20,86	3,02	-2,65
miRn88	55,78	21,00	-1,27
miRn123	18,18	0,00	-10,07
miRn26	44,23	14,93	-1,43
miRn135	21,01	45,85	1,26
miRn20	46,36	17,21	-1,29
miRn224	16,77	38,80	1,34
miRn46	45,19	17,43	-1,24
miRn58	11,61	0,00	-9,42
miRn14	7,00	21,43	1,75
miRn206	13,36	31,01	1,35
miRn272	34,13	12,44	-1,32
miRn127	10,94	0,00	-9,34
miRn205	38,33	15,18	-1,20
miRn181	15,75	33,29	1,21
miRn13	24,42	44,50	1,00
miRn142	47,35	21,52	-1,00
miRn168	27,07	9,33	-1,40
miRn133	25,09	8,28	-1,46
miRn261	24,21	8,56	-1,36
miRn268	13,01	2,74	-2,11

Table 6.22 Differentially expressed novel miRNAs after CAG treatment (cont.)

miRn275	8,49	20,63	1,42
miRn16	14,53	3,54	-1,90
miRn12	23,37	8,65	-1,30
miRn182	29,29	12,66	-1,08
miRn216	5,34	14,47	1,57
miRn253	14,76	4,46	-1,59
miRn214	8,05	1,35	-2,44
miRn169	5,25	0,00	-8,28
miRn145	12,72	3,70	-1,65
miRn43	16,37	6,07	-1,30
miRn231	4,41	11,49	1,52
miRn193	19,08	7,94	-1,13
miRn139	15,26	5,88	-1,24
miRn195	6,51	1,45	-2,03
miRn295	4,06	9,67	1,39
miRn54	8,23	15,00	1,00
miRn136	13,77	6,19	-1,02
miRn143	2,01	0,00	-6,90
miRn256	7,53	3,02	-1,18
miRn277	7,99	3,33	-1,13

6.1.4.2.7 Differentially expressed small RNAs (DESs) target prediction

The DESs target were identified using **DEGseq** software. Totally, 961 targets were observed to be differentially expressed between our two samples and some of them were functionally described.

6.1.4.2.8 Gene Ontology analysis

Gene Ontology (GO) enrichment analysis was performed for screened DESs target genes with the help of **AmigoGO2** database.

GO Analysis was carried out to determine the mechanisms that these genes might belong to and to divide total target genes into three main groups; molecular function (MF), cellular component (CC) and biological process (BP).

GO analysis organizes the target genes under many GOs processes and function categories; among the 961 target genes, 286 genes were determined to be

associated to one or more GO terms. totally, 1031 GO terms were associated with the differentially expressed target genes and classified in 34 annotated functional subcategories. Biological processes have the majority of the GO annotations; 640 GO terms (152 genes, 15.82% of all DEGs). Molecular functions include 222 GO terms (258 genes, 26.85% of all DEGs) while GO terms of cellular components are 170 (178, 18.52% of all DEGs).

The most abundant GO terms present the highest DE target genes. Eight subcategories were identified as the most abundant ones; cellular process (GO:0009987), single organism process (GO:0044236), cell (GO: 0005623), cell part (GO:0032990), organelle (GO:0043226), binding (GO:0005488) and catalytic activity (GO:0003824).

The most enriched GOs of DE genes are shown in Figure 16.31. In biological process category, cellular process, single-organism process, and metabolic process were the top three gene ontology terms with, respectively, 133, 123, and 98, of differentially expressed target genes. In the same category, positive regulation of biological process, developmental process, and multi-organism process represented the less three enriched GO terms with, respectively, 3, 6, and 7 target genes. Cell, cell part, and organelle terms were the most three enriched three GO terms, of the cellular component category, with 140, 138 and 107 target genes respectively. The last three categories were virion, virion part, and extracellular region with respectively 1, 1 and 2 targets. The molecular function category had catalytic activity with 177, binding with 157, and transporter activity with 21 target genes as the top three GO terms while structural molecule activity, enzyme regulator activity and molecular transducer activity with, respectively, 3, 4 and 6 targets were the less enriched GO terms.

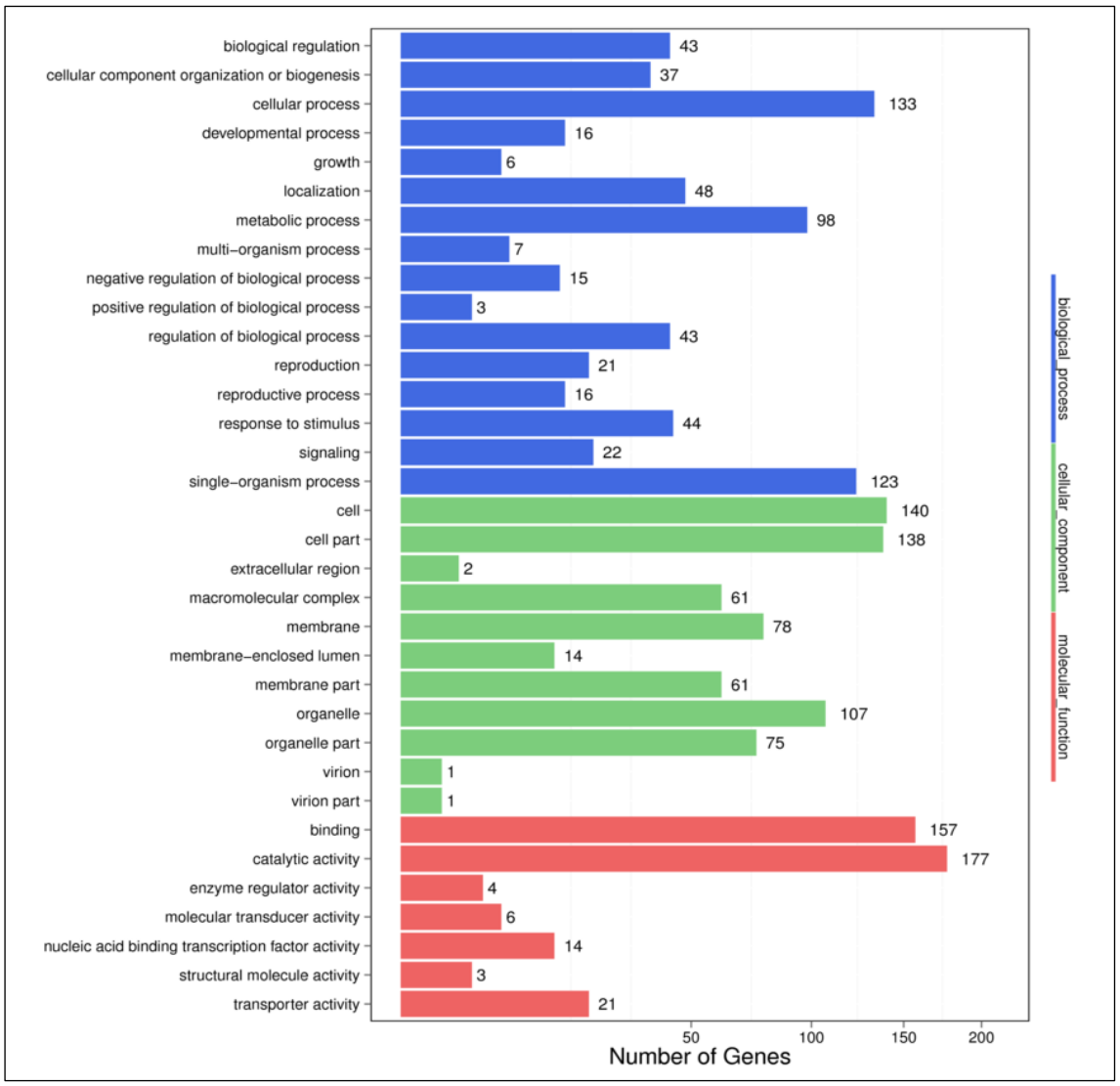


Figure 6. 31 GO functional classification

6.1.4.2.9 KEGG analysis

Kyoto Encyclopedia of Genes and Genomes or KEGG analysis was performed on the DE target genes and 532 targets were assigned to 63 KEGG pathways and arranged into 6 biological functions; Environmental information processing, Cellular processes, Genetic information processing, Metabolism, Organismal Systems and Human Diseases, which are categorized into 20 biological subfunctions (Figure 32). Based on *P-values* and *Q-values* that show the enrichment degree (Figure 33). Indeed, the more significantly DE target genes enriched the pathway the smaller *p-value* and *q-value* should be. The five most significantly enriched pathways are 'Ribosome biogenesis in eukaryotes',

'Monobactam biosynthesis', 'Sulfur metabolism', 'Selenocompound metabolism', and 'Cicardian rhythm-plant' with a *p-value* of, 7.849401e-08, 9.303572e-05, 0.001413715, 0.005986308 and 0.00643223, respectively (Table 6.23).

Table 6. 23 The most enriched pathways after small RNA-seq

Pathway	**DEGs with pathway annotation (532)**	**All genes with pathway annotation (24615)**	***P-value***	***Q-value***	**Pathway ID**
Ribosome biogenesis in eukaryotes	23 (4.32%)	286 (1.16%)	7.849401e-08	4.945123e-06	ko03008
Monobactam biosynthesis	4 (0.75%)	12 (0.05%)	9.303572e-05	2.930625e-03	ko00261
Sulfur metabolism	6 (1.13%)	57 (0.23%)	0.001413715	2.968802e-02	ko00920
Selenocompound metabolism	4 (0.75%)	34 (0.14%)	0.005986308	8.104610e-02	ko00450
Circadian rhythm - plant	3 (0.56%)	18 (0.07%)	0.00643223	8.104610e-02	ko04712
Biosynthesis of secondary metabolites	27 (5.08%)	1734 (7.04%)	0.974437	9.999877e-01	ko01110
Metabolic pathways	59 (11.09%)	4052 (16.46%)	0.9998409	9.999877e-01	ko01100
RNA degradation	12 (2.26%)	374 (1.52%)	0.1139787	5.717938e-01	ko03018
Amino sugar and nucleotide sugar metabolism	15 (2.82%)	499 (2.03%)	0.1260904	5.717938e-01	ko00520
Endocytosis	20 (3.76%)	839 (3.41%)	0.3586703	9.601941e-01	ko04144

The majority of the pathways are associated to metabolism processing with 166 differentially expressed target genes. Genetic Information Processing is the second with 102 DE targets and Cellular processes is the third category with 31 differentially expressed target genes.

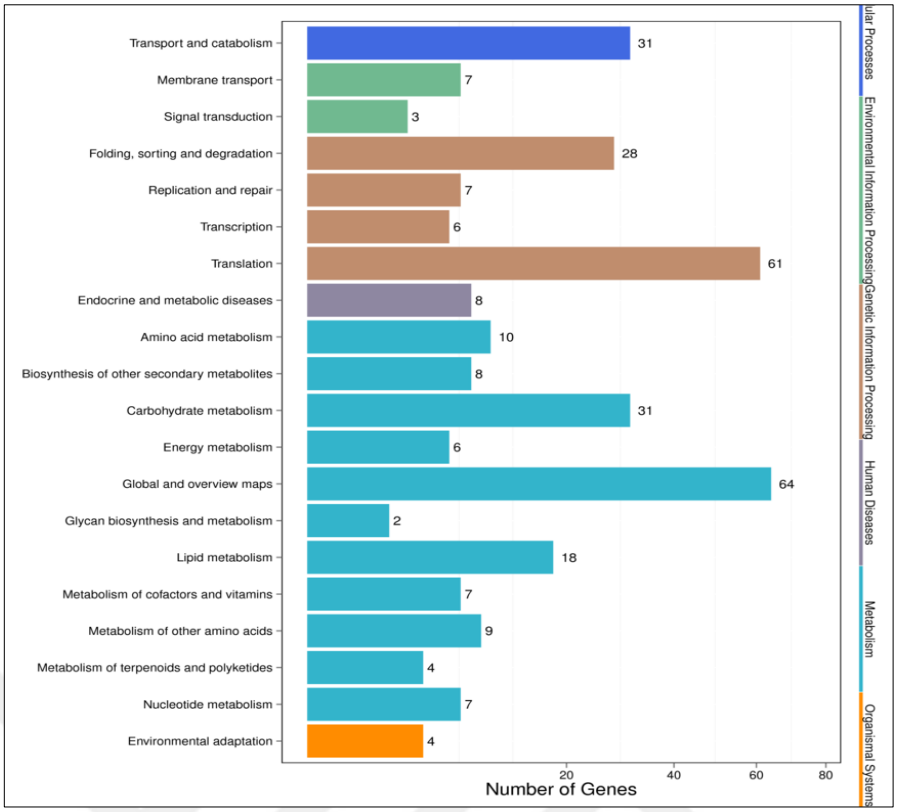


Figure 6. 32 KEGG classification

The Rich Factor, presenting the foreground quotient value (DEGs number) and background value (total Gene amount) defined ribosome biogenesis in eukaryotes, RNA transport, and amino sugar and nucleotide sugar metabolism as the most enriched functional pathways (Figure 6.33).

The enrichment analysis of the targets gives accurate information about their corresponding miRNAs, illustrating the map of every pathway with the associated up and down-regulated target genes. According to the implicated DEGs number, the most abundant KEGG pathways in the present work are ‘Metabolic pathways’ (11.09%, ko01100, no map in kegg), ‘Biosynthesis of secondary metabolites’ (5.08%, ko01110, No map in Kegg) and ‘Ribosome Biogenesis in eukaryotes (4.32%, ko03008, Figure 6.34).

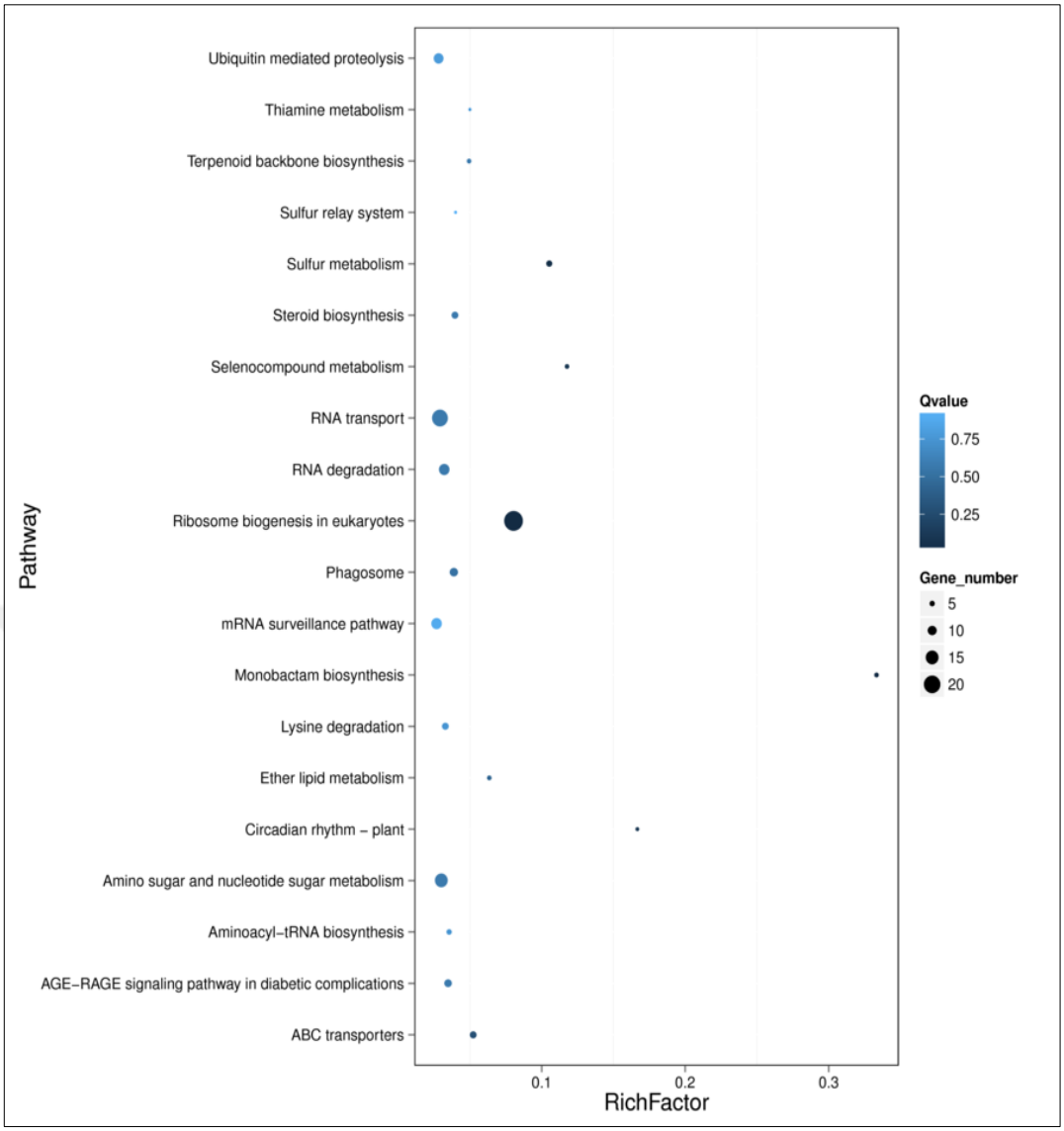
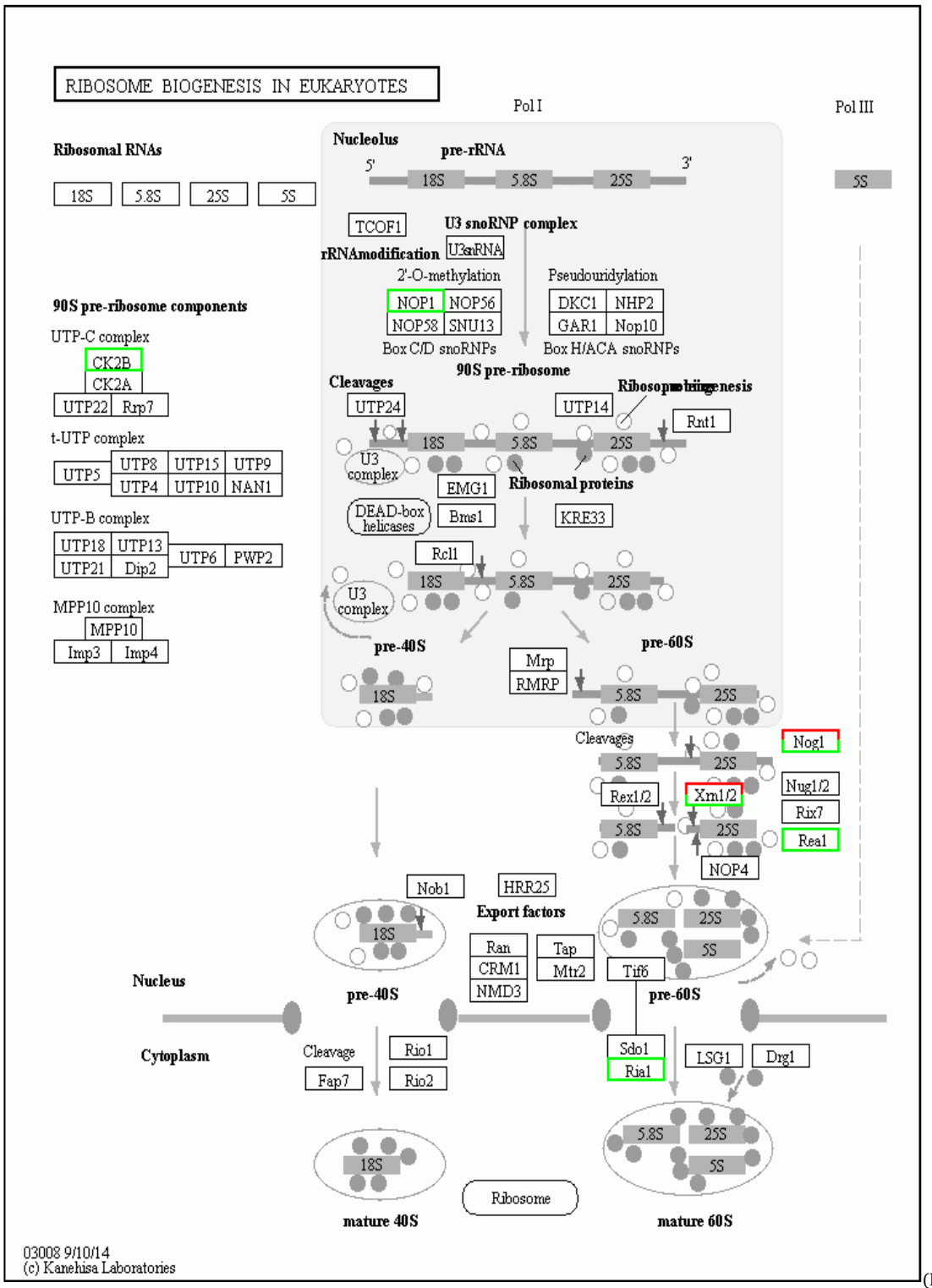


Figure 6. 33 Statistics of pathway enrichment

In the monobactam biosynthesis, AT4G14680 gene encoding *A. thaliana* ATP-sulfurylases which is the first enzyme of sulfate assimilation that catalyzes the formation of adenosine-5'-phosphosulfate from ATP and sulfate is stimulated by CAG. The AT4G14680 gene is involved in cellular response to sulfate starvation, hydrogen sulfide biosynthetic process, protein trimerization and sulfate assimilation and expressed in several plant structures. Some pathways, such as 'RNA transport' (3.57%, ko03013, Figure 6.35), 'Amino sugar and nucleotide sugar metabolism' (2.82%, ko00520), and 'Endocytosis' (3.76%,

ko04144, Figure 6.36) have a significant portion of DE targets with pathway annotation which mean that they were remarkably affected by CAG.

The genes; AT3G12550.1, AT3G12550.3, AT3G12550.5, AT3G12550.2, AT3G12550.4 related to RNA transport pathway, belong to the subgroup of SGS3-like proteins that act redundantly in RNA-directed DNA methylation (RdDM) pathway were up-regulated under CAG treatment [113]. In the endocytosis pathway, up-regulated target genes, encoding for C2-PLD subfamily (Phospholipase D Alpha); AT5G25370.1, AT5G25370.2, AT5G25370.3, and AT5G25370.4, wereidentified. PLD proteins are an enzyme family hydrolyzing phospholipid of the membrane, like phosphoatidyl choline (PC) and phosphoatidyl ethanolamine (PE), producing hence phosphatidic acid (PA) and a free-head alcohol [114], and play a crucial role in responding to several biotic and abiotic stresses. The gene family regulate plant responses to salt and drought stresses by interfering in the mediation of abscisic acid signaling in order to control stomata closure which is a common adaptation response of plants in stress conditions [115].



(Red boxes and green boxes represent up-regulated and down-regulated genes, respectively)

Figure 6. 34 Ribosome biogenesis pathway in *Arabidopsis* treated calli

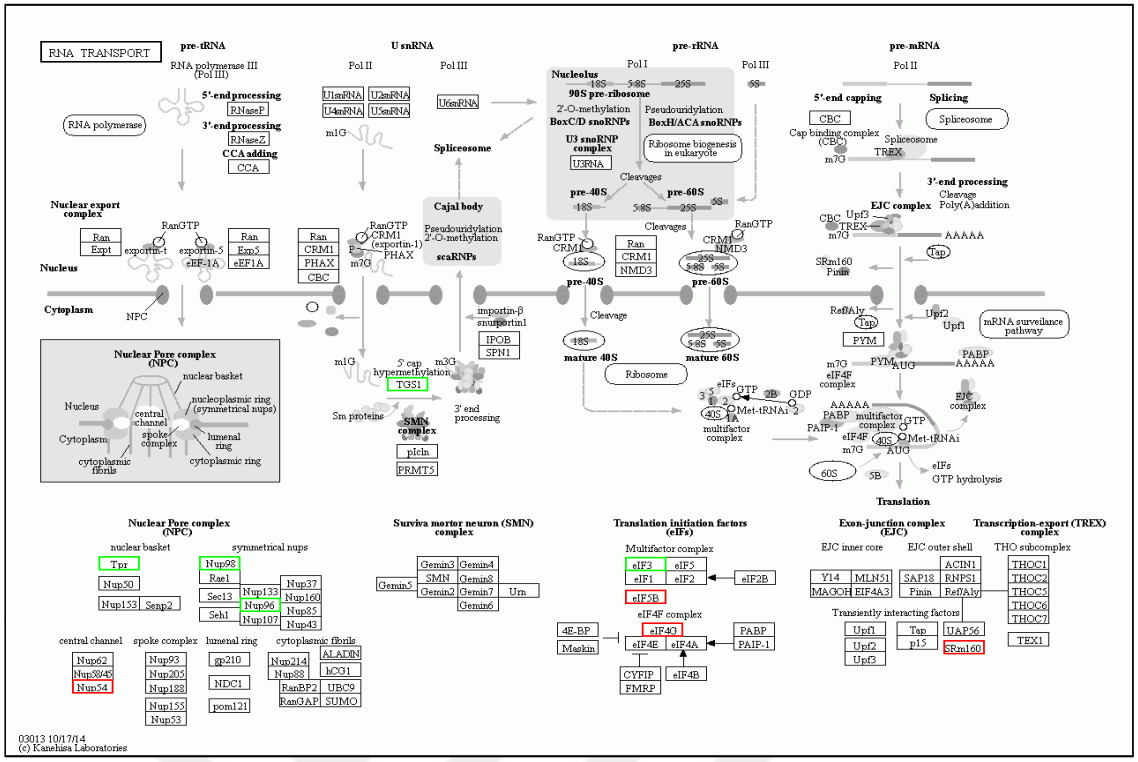


Figure 6. 35 RNA transport pathway in *Arabidopsis* treated calli

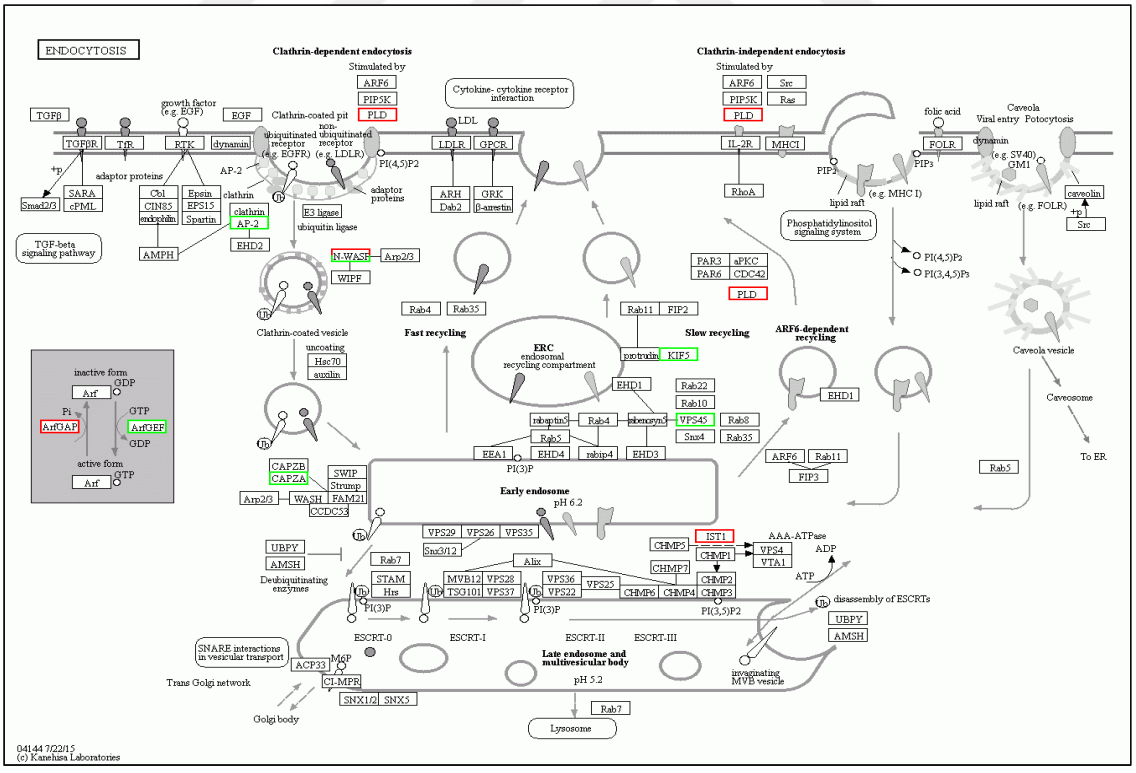


Figure 6. 36 Endocytosis pathway in *Arabidopsis* treated calli

(Red boxes and green boxes represent up-regulated and down-regulated genes, respectively)

6.1.5 Real-time quantitative PCR analysis: Sequencing validation

In order to check and validate our HTS results, the RNAs samples from control and 1 μM concentration which were sent for sequencing, were used in the qRT-PCR analysis.

6.1.1 RNA-seq validation

After cDNA synthesis, and checking the annealing temperature by gradient PCR, qRT-PCR was performed, following the previously mentioned program, to seven randomly genes using the three control and three 1CAG samples in triplicate. The results obtained by qRT-PCR were used to calculate Fold-change for each primer using ΔΔ cycle threshold method. Real-time quantitative PCR analysis exhibited that the relative expression patterns of the chosen genes were consistent with RNA-Seq data with a correlation coefficient of 0.95 between qRT-PCR and RNASeq. The results demonstrated that our RNA-Seq data are reliable (Figure 6. 37).

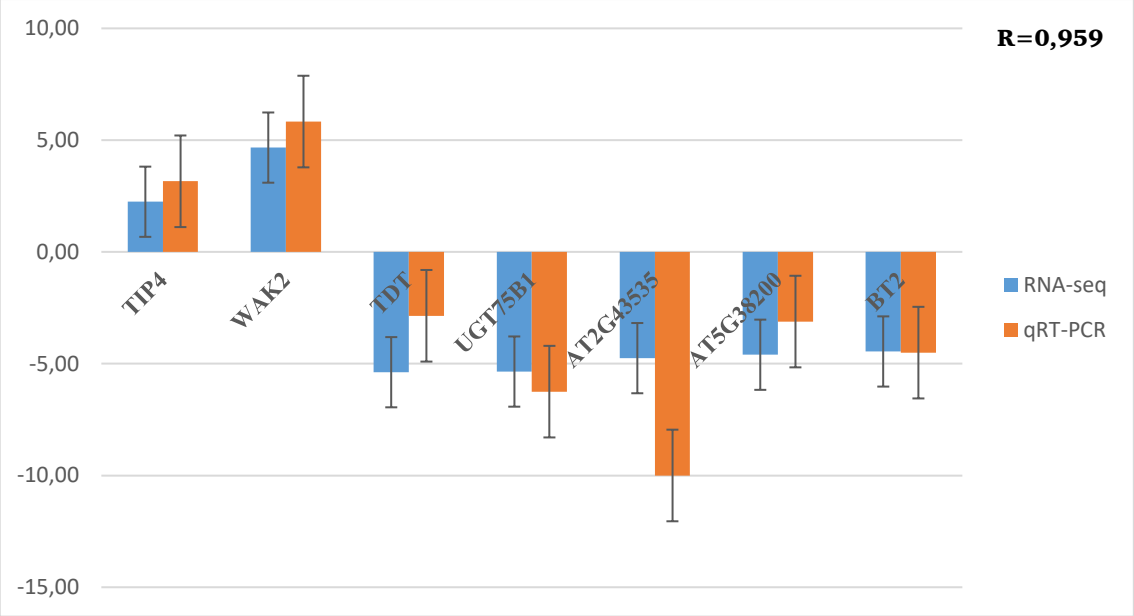


Figure 6. 37 qRT-PCR validation of RNAseq results

6.1.2 SmallRNA-seq validation

After cDNA synthesis, qRT-PCR was performed to 9 randomly chosen miRNAs, following the same program, for the three control and three 1 CAG samples in triplicate.

The results obtained by qRT-PCR were used to calculate Fold-change for each miRNA following ΔΔ cycle threshold method and the analysis exhibited that the relative expression patterns of the chosen miRNAs were consistent with small RNA-Seq data with a correlation coefficient of 0.938 between qRT-PCR and small RNASeq (Figure 6. 38).

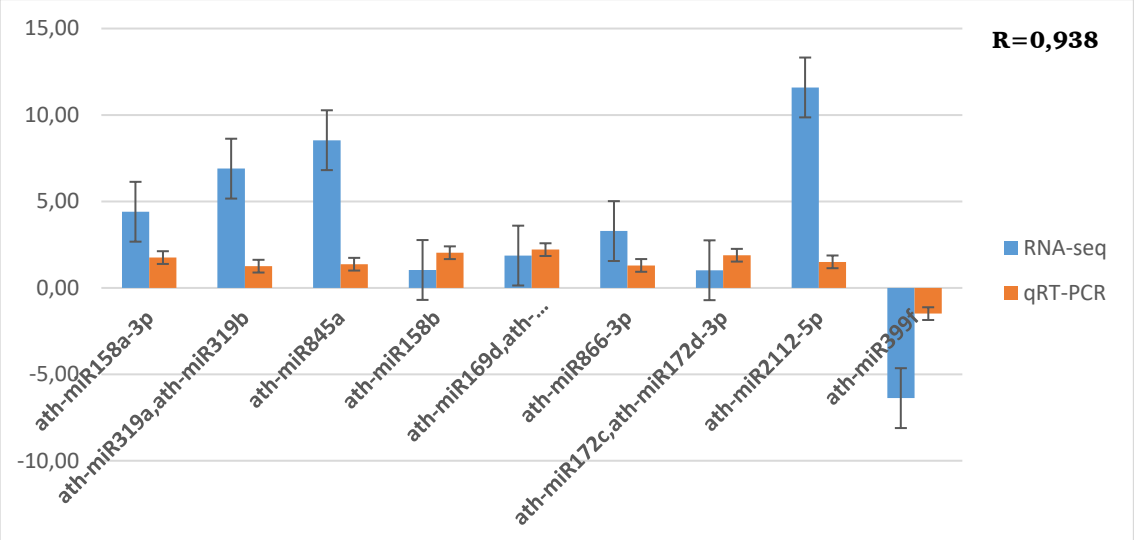


Figure 6. 38 qRT-PCR validation of SmallRNAseq results

6.2 Discussion

6.2.1 Discussion of the RNA-sequencing results

Technologies of NGS are effective ways to highlight known genes, identify new ones and explain their involvement in the concerned biochemical and physiological pathways.

Our study is the first to evaluate cycloastragenol on plants under common and/or specific conditions. RNA sequencing methodology was chosen as it is an efficient mean to achieve any specie or cell transcriptome data in specific conditions.

Indeed, the present research work presents a broad-spectrum analysis of transcriptomic responses in *A. thaliana* calli to cycloastragenol, while small RNA-seq was a way to study the expression profile of sncRNA, particularly microRNAs, in order to understand the cell regulation under CAG treatment and to identify CAG responsive miRNAs expressed in the model plant *Arabidopsis thaliana*.

After RNA sequencing, GO classification revealed that all genes, including the differentially expressed ones, were categorized into three principal groups; Biological Process (BP), Cellular Component (CC) and Molecular function (MF), classified in 46 functional subcategories. Preliminary results showed that treating calli with CAG did not induce a remarkable effect on callus growth index, however 1045 genes (DEGs) between the studied samples were expressed differentially. DEGs data authenticity was validated by qRT-PCR.

Up and down-regulated differentially expressed genes number of the most enriched gene ontology terms were practically the same. All subcategories have both types of genes expecting nucleoid (GO: 0009295) and signal transducer activity (GO: 0039562) that have only down-regulated ones. The DEGs are involved in all categories, however the most of them are down-regulated.

A high number of the differentially expressed genes were attributed to KEGG pathways and are particularly involved in metabolism, replication and repair, transcription, hormone signal transduction pathways and environmental adaptation.

Some pathways have a high number of annotated genes however few of them were expressed differentially. Indeed, 10 DEGs were associated to plant-pathogen interaction pathway among which only one was up-regulated while a total of 578 genes were associated with this pathway. Moreover, 691 genes were found to be associated with the plant-hormone signal transduction pathway and only 35 genes of them were reported to be expressed differentially. Transcriptional responses of adding CAG to *A. thaliana* calli has led this study to various signaling pathways.

KEGG study showed glucosinolate (GSL) biosynthesis as the most significant pathway with 10 DEGs; AT3G19710 (BCAT4), AT5G23020 (MAM3), AT5G23010

(MAM1), AT1G16410 (CYP79F1), AT4G13770 (CYP83A1), AT1G16400 (CYP79F2), AT3G58990 (IPMI1), AT2G22330 (CYP79B3), AT5G14200 (IMD1) and AT2G43100 (IPMI2). Implicated in glucosinolate synthesis (GSL), these genes are specifically involved in the deamination reaction of methionine giving 2-oxomethylthiodecanoic acid and then 8-methylthio-octyl-glucosinolte. GSLs are important defense elements in plants against biotic stresses, their hydrolysis products are toxic for the different species that attack plants. CAG treatment down-regulated all these genes. However, other authors reported that they had different regulation types in different conditions. Actually, CYP79F1, BCAT4, MAM1 and MAM3 in cabbage head were reported up-regulated under external selenium and sulfur treatment in different concentrations and these results support the idea that S and Se metabolisms and GSLs biosynthesis are interacted, such glucoraphanin the main cruciferous plants aliphatic GSL, and sulforaphane the anticancer compound precursor. But the same concentrations did not affect the other members of the same gene family which are UGT74B1 and UGT74C1 [116]. Se amino acids incorporated in proteins affects antagonistically GLSs biosynthesis causing then their decrease [117]. Then the effects of CAG on the regulation of GSLs biosynthesis genes would probably be induced by antagonist effects of the 1μM CAG concentration and a lower concentration would probably have different effect.

Plant-Hormone signaling transduction was among determined pathways the most enriched with 36 DEGs presenting around 4.97% from all DEGs among which 2 genes from the basic helix-loop-helix (bHLH) family; AtbHLH38 (AT3G56970) and AtbHLH39 (AT5G04150). The gene family and their homologs, the subgroup Ib BHLH genes presented by AtbHLH101 (AT3G56980) and AtbHLH100 (AT2G41240), are regulated in roots and leaves under several nutritional changes and in split-root situations related to iron insufficiency [118].

Known to regulate the transcription factors in iron deficiency, AtbHLH39, AtbHLH38, AtbHLH101 and AtbHLH100 are among the enriched gene groups, compared with the control sample. They are up-regulated by respectively, 1.2, 2.8, 7.9, and 1.8 of fold-change. They were also shown to be highly up-regulated in

conditions of Fe deficiency in roots [119]. AtbHLH101 was found to be up-regulated under the action of the allelochemical L-Dopa, which decreases root growth of a various species [120]. CAG can be involved in iron regulation or other mechanism associated with roots growth. Part of the heat shock proteins group and implicated in endoplasmic reticulum pathway processing, HSP90.1 (AT5G52640), HSP101 (AT1G74310), HSP17.6A (AT5G12030), Hsp70b (AT1G16030) and HSP17.4 (AT3G46230) were up-regulated by cycloastragenol. Heat shock proteins (HSPs) are induced by high temperature and inhibit the assembly of proteins sensitive to heat. Moreover, they have a crucial function in some plants adaptation features to abiotic and biotic stresses and stimulate their immunity by accumulating PR proteins in several stress conditions[121][122]. HSP101 and HSP17.6A have been proved to be up-regulated under Ca^{2+} excess in *A. thaliana*. Nevertheless, transcript abundance was not significantly increased [123]. CAG stimulation of HSPs expression reflect the fact that cycloastragenol can stimulate in an indirect way the stress tolerance mechanism.

Riboflavin metabolism was affected significantly by cycloastragenol. Actually, the 7 differentially expressed genes AT1G14700 (PAP3), AT2G01880 (PAP7), AT4G29690, AT4G29700, AT2G38740, AT2G01890 (PAP8) and AT1G25230 (OPT3) associated with this pathway were observed to be down-regulated. OPT3, an oligopeptide transporter playing a role in metal homeostasis particularly iron homeostasis, has been proved to be up-regulated by cadmium (Cd). This suggests a potential implication of this gene in the organization of iron distribution and its transfer [124]. However, OPT3 encodes Fe transporter that works in stem phloem and it is involved in Cd tolerance [125] which suggest a potential role of CAG in Cd tolerance mechanism.

A group of 4 unknown proteins was found to be highly up-regulated by CAG. They are; AT1G47395 and At1G47400, presented like hypothetical proteins, AT1G13609 which belongs to Defensin-like family protein (DEFL) and At2G30766; a member of cytochrome P450 family. AT1G13609 and At2G30766 were also shown to be up-regulated under lack of Fe in leaves [100]. Fe is associated to photosystems and important for photosynthesis suggesting that the

two genes work differently with photosynthesis and CAG have an antagonist action in this mechanism.

6.2.2 Discussion of small RNA sequencing results

MicroRNAs play an important role in gene expression regulation at post-transcriptional level. Small RNA sequencing is used in order to identify micro RNAs in genome plant.

This technique enables several researchers to identify and characterize miRNAs involved in different signaling pathways. To date, several miRNAs have been characterized and revealed to be functionally involved in various mechanisms such phosphorus and nitrogen nutrient signaling pathways [126], biotic and abiotic stresses [127] and plant development stages [128].

The purpose of the present study is to determine *Arabidopsis thaliana* responsive microRNAs to cycloastragenol molecule. In this frame, small RNA sequencing was performed using control and 1 μM CAG treated calli. Small RNA sequencing generated 119 differentially expressed miRNAs and 961 corresponding target genes. Authenticity of our data was validated by qRT-PCR analysis. Based on GO classification, target genes were grouped in three principal categories; Molecular function (MF), Cellular Component (CC), and Biological Process (BP) to have totally 1031 GO terms divided in 34 annotated functional subcategories. In the first part of the present study, CAG treatment had no significant effect on calli growth index. However, 961 target genes were differentially expressed between control and treated samples among which 286 genes were determined to be associated with one or more GO terms implicated in all subcategories. The three categories are presented by up and down-regulated miRNAs and their corresponding targets.

KEGG analysis showed that the majority of the targets were assigned to KEGG pathways and the study of their functions and the pathways to which they are associated showed that they are mainly involved in metabolism pathways, genetic information processing and cellular processes pathways. A high number of target genes were shown to be associated with some pathways, however few of them are

differentially expressed. In instance, 12 DE target genes were annotated to RNA degradation while totally 374 genes were associated to the same pathway, among which 6 genes were up-regulated. Totally, 534 genes were reported to be assigned to the spliceosome pathway, among which only 4 genes were found to be differentially expressed in the treated sample. CAG treatment significantly affected the expression of miRNAs. In total, 273 known and 298 novel miRNAs were identified (Table 6.21).

The most significant expression difference was occurred for miR2112-3p by 5.88-fold up regulation, (miR394a, miR394b-5p) and miR2112-5p were regulated by 4.75 and 4.53 respectively. Among the downregulated miRNAs, the most downregulated one is miR399d by -6.84. Despite the fact that miR2112-3p was the most regulated miRNA, its target AT4G23160.1 is not annotated. It is an organ-specifically and constitutively expressed microRNA in the siliques of *Arabidopsis* according to [129].

Regulation of gene expression of *A. thaliana* calli under CAG treatment was highlighted by several signaling pathways. According to KEGG analysis, the most significantly enriched pathway is 'ribosome biogenesis in eukaryotes' with 23 DE target genes among which, AT3G10300.1, AT3G10300.2, AT3G10300.3, AT3G10300.4, AT3G10300.5, AT3G10300.6, representing calcium-binding EF-hand protein family, targets of miR159b-5p, are involved in the regulation of ribosome biogenesis and RNA degradation pathways and were found to be up-regulated under CAG treatment. Actually, the Ca^{2+} messenger is generally occurred in plants and the EF-hand protein family play a crucial role as transducers mediating Ca^{2+} action in different hormonal and environmental signals including biotic and abiotic stresses [130], which mean that CAG can enhance Ca^{2+} action in different signaling pathways. Mir159b-5p was also found to be involved in metabolic processes pathway under arsenic contamination. Endocytosis pathway was annotated by 839 genes among which only 20 genes were differentially expressed. AT5G25370.1, AT5G25370.2, AT5G25370.3, AT5G25370.4, were up-regulated after cycloastragenol treatment. They represent C2-PLD subfamily coding for PHOSPHOLIPASE D ALPHA 3 which is an enzyme

responsible of hydrolyzing glycerol-phospholipids at the terminal phosphodiesteric bond to generate phosphatidic acids (PA). The enzyme interferes positively in plant responses to hyperosmotic stresses, a high PLDalpha3 expression enhances root growth, flowering and stress prevention [131]. The obtained data showed that CAG up-regulated the microRNA associated to these target genes of miR3434-5p, which can lead to deduce that CAG would also enhance flowering and root growth and avoid stresses. The same microRNA was found by [132] to be downregulated in response to high GSH level during a pathogen attack specifically and biotic stresses generally.

According to our data results, CAG upregulated some miRNAs and downregulated others, and among the upregulated ones, miR319 (a and b) which is involved in the regulation of the transcription factors of TCP family [133] showed a significant increase with 2.87-fold change. This microRNA was reported by [134] to be upregulated under cold stress in sugarcane. Mir158a, which plays an important role in the glucosilation by targeting PPR and fucosyltransferase gene families [135], and miR395, which was reported to be essential in the regulation of sulfate assimilation [136], were upregulated by 2.13-fold and 1.79-fold, respectively, under CAG treatment. It was shown up that their expression was also increased under, respectively, phosphore deprivation in tomato [137] and Cu-induced oxidative stress in *A. thaliana* [138], which prove that the two miRNAs are involved in nutrient signalling pathways. In contrast miR157, an important regulator of transition from juvenile to adult phase in plants and involved in shoot development and leaf and flower initiation, was found to be downregulated under CAG effect by -1.66-fold. However, [139] showed that miR157, is upregulated in the first step of *Arabidopsis* seed germination corresponding to water uptake under cold temperature; cold imbibition. Another microRNA was significantly downregulated by cycloastragenol, miR396b, with -1.69-fold. This miRNA which is known to target growth regulation factor (GRF) genes, which encode putative transcription factors involved in plant leaf growth, was found to be upregulated during Mn toxicity and attenuate then the expression of several transcription factors [137]. MiR399 genes; miR399a, miR399b, miR399c, miR399d, miR399e,

miR399f were proved to encode a phosphate starvation-responsive and involved in orthophosphate (Pi) deficiency signaling pathway targeting PHOSPHOTE2 (PHO2) gene which encodes for an enzyme modulate phosphate uptake and root-to-shoot allocation [140]. These microRNAs were downregulated by CAG. One of the most significantly affected by cycloastrogenol is miR399f with –3.67-fold. However, [139] have studied miR399 family in *Arabidopsis thaliana* under cold imbibition conditions and found out that low temperature upregulated miR399a-f.

6.3 Conclusion

The present work presents the first study of the effects of cycloastragenol in *A. thaliana* and in plants generally. All studies that have been done before, investigated about CAG effects in human and revealed its ability to overcome several health problems particularly its anti-aging effect by stimulating telomerase activity. High throughput sequencing technology was adopted in order to have an overall view of the molecular, biochemical and physiological impacts and to highlight all the involved genes and pathways after cycloastragenol treatment.

High throughput sequencing of coding and non-coding RNA (RNA-seq and small RNA-seq respectively) allows, at relatively low cost, to measure simultaneously the sequence and the expression of RNAs at whole cell level. However, the analysis of the generated data requires new bioinformatic approaches. In this study, RNA sequencing analysis identified more than 40.000 genes among which around 22.000 were differentially expressed between non-treated and CAG-treated sample. Most of the differentially expressed genes belong to physiological processes, signaling/metabolic pathways and regulatory networks such as BCAT4 and CYP79F1 associated with GSLs synthesis which play an important role in stress tolerance in plants. Various transcription factors were also determined and different protein kinases were revealed to be upregulated by CAG. Small RNA sequencing aimed to the identification of small non coding RNAs particularly microRNAs. Actually, more than 5.700 from the revealed sncRNAs, among which

more than 100 miRNAs were differentially expressed between the two samples and targeting more than 900 genes. Most of the up regulated microRNAs and their target genes are associated with regulatory pathways like the C2-PLD subfamily coding for PHOSPHOLIPASE D ALPHA 3 which play a role in hydrolyzing the glycerol-phospholipids of the phosphodiester bond.

RNA sequencing and small RNA sequencing highlighted, in our study, several pathways enriched with up and down-regulated genes, among which, metabolic pathways, biosynthesis of secondary metabolites and circadian rhythm-plant were observed to be three of the most enriched pathways in both analyses. The different determined pathways and the involved differentially expressed genes would be helpful to develop our knowledge about how CAG affect plants and would be exploited to create breeding programs overcoming biotic and/or abiotic stresses in order to improve some crop species and/or to develop CAG-derived products like pesticides and/or fertilizers.

REFERENCES

[1] I. Raskin *et al.*, “Plants and human health in the twenty-first century,” *Trends Biotechnol.*, vol. 20, no. 12, pp. 522–531, 2002, doi: 10.1016/S0167-7799(02)02080-2.

[2] V. Fuster and J. M. Sweeny, “Aspirin: A historical and contemporary therapeutic overview,” *Circulation*, vol. 123, no. 7, pp. 768–778, 2011, doi: 10.1161/CIRCULATIONAHA.110.963843.

[3] D. Prvulovic, H. Hampel, and J. Pantel, “Galantamine for Alzheimer’s disease,” *Expert Opin. Drug Metab. Toxicol.*, vol. 6, no. 3, pp. 345–354, 2010, doi: 10.1517/17425251003592137.

[4] P. Kumar Gupta, S. Pulapalli, and F. Correspondence Pranay Kumar Gupta, “Research and Reviews: Journal of Microbiology and Biotechnology Cycloastragenol(Telomeres Activator) and its relation with Cancer: A Brief Review,” vol. 4.

[5] Y. Yu, L. Zhou, Y. Yang, and Y. Liu, “Cycloastragenol: An exciting novel candidate for age-associated diseases (Review),” *Exp. Ther. Med.*, vol. 16, no. 3, pp. 2175–2182, 2018, doi: 10.3892/etm.2018.6501.

[6] J. Wang, M. L. Wu, S. P. Cao, H. Cai, Z. M. Zhao, and Y. H. Song, “Cycloastragenol ameliorates experimental heart damage in rats by promoting myocardial autophagy via inhibition of AKT1-RPS6KB1 signaling,” *Biomed. Pharmacother.*, vol. 107, no. July, pp. 1074–1081, 2018, doi: 10.1016/j.biopha.2018.08.016.

[7] F. C. F. Ip *et al.*, “Cycloastragenol is a potent telomerase activator in neuronal cells: Implications for depression management,” *NeuroSignals*, vol. 22, no. 1, pp. 52–63, May 2014, doi: 10.1159/000365290.

[8] C. B. Harley *et al.*, “A natural product telomerase activator as part of a health maintenance program,” *Rejuvenation Res.*, vol. 14, no. 1, pp. 45–56,

2011, doi: 10.1089/rej.2010.1085.

[9] H. Y. K. Lam *et al.*, “Performance comparison of whole-genome sequencing platforms,” *Nat. Biotechnol.*, vol. 30, no. 1, pp. 78–82, 2012, doi: 10.1038/nbt.2065.

[10] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics in Western Equatoria State,” *Nat. Rev. Genet.*, vol. 10, no. 1, p. 57, 2009.

[11] B. Rabbani, M. Tekin, and N. Mahdieh, “The promise of whole-exome sequencing in medical genetics,” *J. Hum. Genet.*, vol. 59, no. 1, pp. 5–15, 2014, doi: 10.1038/jhg.2013.114.

[12] M. Pelizzola and J. R. Ecker, “The DNA methylome,” *FEBS Lett.*, vol. 585, no. 13, pp. 1994–2000, 2011, doi: 10.1016/j.febslet.2010.10.061.

[13] F. Finotello and B. Di Camillo, “Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis,” *Brief. Funct. Genomics*, vol. 14, no. 2, pp. 130–142, 2015, doi: 10.1093/bfgp/elu035.

[14] J. K. Kulski, “Next-Generation Sequencing — An Overview of the History, Tools, and ‘Omic’ Applications,” *Next Gener. Seq. - Adv. Appl. Challenges*, no. January, pp. 2–60, 2016, doi: 10.5772/61964.

[15] L. C. Hsieh *et al.*, “Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing,” *Plant Physiol.*, vol. 151, no. 4, pp. 2120–2132, 2009, doi: 10.1104/pp.109.147280.

[16] A. Pauli, J. L. Rinn, and A. F. Schier, “Non coding RNAs regulation in embryogenesis,” *Nat Rev Genet.*, vol. 12, no. 2, pp. 136–149, 2011, doi: 10.1038/nrg2904.Competing.

[17] M. Gu *et al.*, “Cycloastragenol improves hepatic steatosis by activating farnesoid X receptor signalling,” *Pharmacol. Res.*, vol. 121, pp. 22–32, Jul. 2017, doi: 10.1016/j.phrs.2017.04.021.

[18] J. Zhu *et al.*, “In vitro intestinal absorption and first-pass intestinal and

hepatic metabolism of cycloastragenol, a potent small molecule telomerase activator," *Drug Metab. Pharmacokinet.*, vol. 25, no. 5, pp. 477–486, 2010, doi: 10.2133/dmpk.DMPK-10-RG-037.

[19] L. Y. Yung *et al.*, "Astragaloside IV and cycloastragenol stimulate the phosphorylation of extracellular signal-regulated protein kinase in multiple cell types," *Planta Med.*, vol. 78, no. 2, pp. 115–121, 2012, doi: 10.1055/s-0031-1280346.

[20] Y. Zu *et al.*, "Determination and quantification of astragalosides in Radix Astragali and its medicinal products using LC-MS," *J. Sep. Sci.*, vol. 32, no. 4, pp. 517–525, 2009, doi: 10.1002/jssc.200800499.

[21] Y. Liu, J. Liu, Y. Wang, A. Abozeid, and Z. H. Tang, "Simultaneous determination of six active metabolites in Astragalus mongholicus (Fisch.) Bge. under salt stress by ultra-pressure liquid chromatography with tandem mass spectrometry," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-2638-y.

[22] N. A. Kovalenko, D. D. Zhdanov, and T. F. Kovalenko, "Possibilities and effects of telomerase activation," *Mol. Biol.*, vol. 47, no. 4, pp. 476–487, 2013, doi: 10.1134/S0026893313040079.

[23] R. Ran *et al.*, "Evaluation and comparison of the inhibition effect of astragaloside IV and aglycone cycloastragenol on various UDP-glucuronosyltransferase (UGT) isoforms," *Molecules*, vol. 21, no. 12, Dec. 2016, doi: 10.3390/molecules21121616.

[24] C. Sun *et al.*, "Cycloastragenol mediates activation and proliferation suppression in concanavalin A-induced mouse lymphocyte pan-activation model," *Immunopharmacol. Immunotoxicol.*, vol. 39, no. 3, pp. 131–139, May 2017, doi: 10.1080/08923973.2017.1300170.

[25] S. R. Fauce *et al.*, " Telomerase-Based Pharmacologic Enhancement of Antiviral Function of Human CD8 + T Lymphocytes ," *J. Immunol.*, vol. 181, no. 10, pp. 7400–7406, 2008, doi: 10.4049/jimmunol.181.10.7400.

[26] Y. Zhao *et al.*, “Astragaloside IV and cycloastragenol are equally effective in inhibition of endoplasmic reticulum stress-associated TXNIP/NLRP3 inflammasome activation in the endothelium,” *J. Ethnopharmacol.*, vol. 169, pp. 210–218, 2015, doi: 10.1016/j.jep.2015.04.030.

[27] S. P. H. Alexander *et al.*, “The concise guide to PHARMACOLOGY 2013/14: Transporters,” *Br. J. Pharmacol.*, vol. 170, no. 8, pp. 1706–1796, 2013, doi: 10.1111/bph.12450.

[28] M. I. Zvereva, D. M. Shcherbakova, and O. A. Dontsova, “Telomerase: Structure, functions, and activity regulation,” *Biochem.*, vol. 75, no. 13, pp. 1563–1583, 2010, doi: 10.1134/S0006297910130055.

[29] M. Korandová and R. Č. Frydrychová, “Activity of telomerase and telomeric length in Apis mellifera,” *Chromosoma*, vol. 125, no. 3, pp. 405–411, 2016, doi: 10.1007/s00412-015-0547-4.

[30] C. Harley B., A. Futcher B., and C. Greider W., “Telomeres shorten during ageing of human fibroblasts,” *Nature*, vol. 345. pp. 458–460, 1990.

[31] I. Flores, R. Benetti, and M. A. Blasco, “Telomerase regulation and stem cell behaviour,” *Curr. Opin. Cell Biol.*, vol. 18, no. 3, pp. 254–260, 2006, doi: 10.1016/j.ceb.2006.03.003.

[32] L. Salvador, G. Singaravelu, C. B. Harley, P. Flom, A. Suram, and J. M. Raffaele, “A Natural Product Telomerase Activator Lengthens Telomeres in Humans: A Randomized, Double Blind, and Placebo Controlled Study,” *Rejuvenation Res.*, vol. 19, no. 6, pp. 478–484, 2016, doi: 10.1089/rej.2015.1793.

[33] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, “Advanced sequencing technologies: Methods and goals,” *Nat. Rev. Genet.*, vol. 5, no. 5, pp. 335–344, 2004, doi: 10.1038/nrg1325.

[34] K. Raza and S. Ahmad, “Recent Advancement in Next- Generation Sequencing Techniques,” *bioRxiv*, vol. 13, p. 341, 2017.

[35] F. Crick and J. Watson, “© 1953 Nature Publishing Group,” 1953.

[36] H. Jaffee, F. G. Bordwell, P. J. Barton, and C. Cooper, “+o. 10,” vol. 886, no. 4, 1961.

[37] R. W. Holley, J. T. Madison, and A. Zamir, “A new method for sequence determination of large oligonucleotides,” *Biochem. Biophys. Res. Commun.*, vol. 17, no. 4, pp. 389–394, 1964, doi: 10.1016/0006-291X(64)90017-8.

[38] R. W. Holley, J. Apgar, G. A. Everett, and J. T. Madison, “Structure of a Ribonucleic Acid Marquisee , Susan H . Merrill , John Robert Penswick and Ada Zamir Published by : American Association for the Advancement of Science STable URL : http://www.jstor.org/sTable/1715055 REFERENCES Linked references are availab,” vol. 147, no. 3664, pp. 1462–1465, 2017.

[39] F. Sanger, G. G. Brownlee, and B. G. Barrell, “A two-dimensional fractionation procedure for radioactive nucleotides,” *J. Mol. Biol.*, vol. 13, no. 2, pp. 373–398, 1965, doi: 10.1016/S0022-2836(65)80104-8.

[40] J. M. Heather and B. Chain, “The sequence of sequencers: The history of sequencing DNA,” *Genomics*, vol. 107, no. 1, pp. 1–8, 2016, doi: 10.1016/j.ygeno.2015.11.003.

[41] F. Sanger and A. R. Coulson, “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase,” *J. Mol. Biol.*, vol. 94, no. 3, pp. 441–448, 1975, doi: 10.1016/0022-2836(75)90213-2.

[42] P. M. S. & M. S. F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, “Nucleotide sequence of bacteriophage φX174 DNA,” 1977.

[43] A. M. Maxam and W. Gilbert, “A new method for sequencing DNA. 1977.,” *Biotechnology*, vol. 24, no. 2, pp. 99–103, 1992.

[44] M. Gužvić, “The history of DNA sequencing,” *J. Med. Biochem.*, vol. 32, no. 4, pp. 301–312, 2013, doi: 10.2478/jomb-2014-0004.

[45] A. Munshi, *DNA SEQUENCING – METHODS AND Edited by Anjana Munshi*.

2012.

[46] M. L. Metzker, “Sequencing technologies the next generation,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010, doi: 10.1038/nrg2626.

[47] J. Shendure and H. Ji, “Next-generation DNA sequencing,” *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1135–1145, 2008, doi: 10.1038/nbt1486.

[48] S. Myllykangas, J. Buenrostro, and H. P. Ji, “Bioinformatics for High Throughput Sequencing,” *Bioinforma. High Throughput Seq.*, pp. 11–25, 2012, doi: 10.1007/978-1-4614-0782-9.

[49] A. Ahmadian, M. Ehn, and S. Hober, “Pyrosequencing: History, biochemistry and future,” *Clin. Chim. Acta*, vol. 363, no. 1–2, pp. 83–94, 2006, doi: 10.1016/j.cccn.2005.04.038.

[50] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi, “Next-generation sequencing:from basic research to diagnostics,” *Clin. Chem.*, vol. 55, no. 4, pp. 641–658, 2009, doi: 10.1373/clinchem.2008.112789.

[51] E. E. Schadt, S. Turner, and A. Kasarskis, “A window into third-generation sequencing,” *Hum. Mol. Genet.*, vol. 19, no. R2, pp. 227–240, 2010, doi: 10.1093/hmg/ddq416.

[52] N. Kono and K. Arakawa, “Nanopore sequencing: Review of potential applications in functional genomics,” *Dev. Growth Differ.*, vol. 61, no. 5, pp. 316–326, 2019, doi: 10.1111/dgd.12608.

[53] “index @ enseqlopedia.com.” [Online]. Available: http://enseqlopedia.com/.

[54] K. S. Frese, H. A. Katus, and B. Meder, “Next-generation sequencing: From understanding biology to personalized medicine,” *Biology (Basel).*, vol. 2, no. 1, pp. 378–398, 2013, doi: 10.3390/biology2010378.

[55] S. M. Bybee *et al.*, “Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics,” *Genome Biol. Evol.*, vol. 3, no. 1, pp. 1312–1323, 2011, doi: 10.1093/gbe/evr106.

[56] J. Zhang, R. Chiodini, A. Badr, and G. Zhang, “The impact of next-generation sequencing on genomics,” *J. Genet. Genomics*, vol. 38, no. 3, pp. 95–109, 2011, doi: 10.1016/j.jgg.2011.02.003.

[57] H. Li, J. Ruan, and R. Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores,” *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, 2008, doi: 10.1101/gr.078212.108.

[58] N. Homer, B. Merriman, and S. F. Nelson, “BFAST: An alignment tool for large scale genome resequencing,” *PLoS One*, vol. 4, no. 11, 2009, doi: 10.1371/journal.pone.0007767.

[59] C. Y. Lee *et al.*, “Common applications of next-generation sequencing technologies in genomic research,” *Transl. Cancer Res.*, vol. 2, no. 1, pp. 33–45, 2013, doi: 10.3978/j.issn.2218-676X.2013.02.09.

[60] L. Feuk, A. R. Carson, and S. W. Scherer, “Structural variation in the human genome,” *Nat. Rev. Genet.*, vol. 7, no. 2, pp. 85–97, 2006, doi: 10.1038/nrg1767.

[61] M. O’Connell, “RNA modification and the epitranscriptome; the next frontier,” *Rna*, vol. 21, no. 4, pp. 703–704, 2015, doi: 10.1261/rna.050260.115.

[62] J. A. Martin and Z. Wang, “Next-generation transcriptome assembly,” *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 671–682, 2011, doi: 10.1038/nrg3068.

[63] M. A. Tariq, H. J. Kim, O. Jejelowo, and N. Pourmand, “Whole-transcriptome RNAseq analysis from minute amount of total RNA,” *Nucleic Acids Res.*, vol. 39, no. 18, pp. 1–10, 2011, doi: 10.1093/nar/gkr547.

[64] R. Sooknanan, J. Pease, and K. Doyle, “Novel methods for rRNA removal and directional, ligation-free RNA-seq library preparation,” *Nat. Methods*, vol. 7, no. 10, pp. i–ii, 2010, doi: 10.1038/nmeth.f.313.

[65] J. Z. Levin *et al.*, “Comprehensive comparative analysis of strand-specific RNA sequencing methods,” *Nat. Methods*, vol. 7, no. 9, pp. 709–715, 2010,

doi: 10.1038/nmeth.1491.

[66] Illumina Inc., “TruSeq™ RNA and DNA Library Preparation Kits v2 - Pub. No. 770-2009-039 Current as of 17 November 2014,” pp. 2–5, 2011, [Online]. Available:https://www.illumina.com/documents/products/datasheets/datasheet_truseq_sample_prep_kits.pdf.

[67] Illumina, “The Best Next-Gen Sequencing Workflow Just Got Better Breakthrough System for Cluster Generation,” no. Figure 3, 2011, [Online]. Available:https://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf.

[68] Illumina Inc, “Illumina Sequencing Technology - YouTube,” *Oct. 23*, 2013, [Online]. Available:https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf%0Ahttps://www.youtube.com/watch?v=womKfikWlxM.

[69] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants,” *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, 2009, doi: 10.1093/nar/gkp1137.

[70] B. Wajid and E. Serpedin, “Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers,” *Genomics, Proteomics Bioinforma.*, vol. 10, no. 2, pp. 58–73, 2012, doi: 10.1016/j.gpb.2012.05.006.

[71] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009, doi: 10.1093/bioinformatics/btp352.

[72] J. T. Hill, B. L. Demarest, B. W. Bisgrove, B. Gorsi, Y. C. Su, and H. J. Yost, “MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq,” *Genome Res.*, vol. 23, no. 4, pp. 687–697, 2013, doi:

10.1101/gr.146936.112.

[73] A. C. Seila *et al.*, "Divergent transcription from active promoters," *Science (80-.).*, vol. 322, no. 5909, pp. 1849–1851, 2008, doi: 10.1126/science.1162253.

[74] K. Fejes-toth *et al.*, "Nihms127518," vol. 457, no. 7232, pp. 1–12, 2009, doi: 10.1038/nature07759.Post-transcriptional.

[75] F. Ozsolak and P. M. Milos, "RNA sequencing: Advances, challenges and opportunities," *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, 2011, doi: 10.1038/nrg2934.

[76] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, 2009, doi: 10.1186/gb-2009-10-3-r25.

[77] S. W. Burge *et al.*, "Rfam 11.0: 10 years of RNA families," *Nucleic Acids Res.*, vol. 41, no. D1, pp. 226–232, 2013, doi: 10.1093/nar/gks1005.

[78] C. A. Raabe, T. H. Tang, J. Brosius, and T. S. Rozhdestvensky, "Biases in small RNA deep sequencing data," *Nucleic Acids Res.*, vol. 42, no. 3, pp. 1414–1426, 2014, doi: 10.1093/nar/gkt1021.

[79] D. Hendrix, M. Levine, and W. Shi, "Open Access METHOD miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data miRTRAP A novel method for prediction of miRs from deep sequencing data," *Genome Biol.*, vol. 11, p. 39, 2010, [Online]. Available: http://genomebiology.com/2010/11/4/R39.

[80] D.-G. Guan, J.-Y. Liao, Z.-H. Qu, Y. Zhang, and L.-H. Qu, "mirExplorer: Detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features," *RNA Biol.*, vol. 8, no. 5, pp. 922–934, 2011, doi: 10.4161/rna.8.5.16026.

[81] A. Mathelier, A. Carbone, and I. Hofacker, “MIReNA: Finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data,” *Bioinformatics*, vol. 27, no. 13, pp. 2226–2234, 2011, doi: 10.1093/bioinformatics/btq329.

[82] X. Yang and L. Li, “miRDeep-P: A computational tool for analyzing the microRNA transcriptome in plants,” *Bioinformatics*, vol. 27, no. 18, pp. 2614–2615, 2011, doi: 10.1093/bioinformatics/btr430.

[83] C. C. Pritchard, H. H. Cheng, and M. Tewari, “MicroRNA profiling: Approaches and considerations,” *Nat. Rev. Genet.*, vol. 13, no. 5, pp. 358–369, 2012, doi: 10.1038/nrg3198.

[84] J. E. Eipper-Mains, B. A. Eipper, and R. E. Mains, “Global approaches to the role of miRNAs in drug-induced changes in gene expression,” *Front. Genet.*, vol. 3, no. JUN, pp. 1–17, 2012, doi: 10.3389/fgene.2012.00109.

[85] W. Mhiri, M. Ceylan, N. Turgut-Kara, B. Nalbantoğlu, and Ö. Çakır, “Transcriptomic analysis reveals responses to Cycloastragenol in Arabidopsis thaliana,” *PLoS One*, vol. 15, no. 12 December, pp. 1–19, 2020, doi: 10.1371/journal.pone.0242986.

[86] A. Bhattacharjee *et al.*, “Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 24, pp. 13790–13795, 2001, doi: 10.1073/pnas.191502998.

[87] P. H. Johnson *et al.*, “Multiplex gene expression analysis for high-throughput drug discovery: Screening and analysis of compounds affecting genes overexpressed in cancer cells,” *Mol. Cancer Ther.*, vol. 1, no. 14, pp. 1293–1304, 2002.

[88] S. I. Rasool and C. De Bergh, “© 1970 Nature Publishing Group,” *Nat. Publ. Gr.*, vol. 228, pp. 726–734, 1970, [Online]. Available: http://www.mendeley.com/research/discreteness-conductance-chnge-n-bimolecular-lipid-membrane-presence-certin-antibiotics/.

[89] S. R. Goodman and W. E. Zimmer, *Cytoskeleton*. 2007.

[90] C. K. Sen and L. Packer, “Antioxidant and redox regulation of gene transcription,” *FASEB J.*, vol. 10, no. 7, pp. 709–720, 1996, doi: 10.1096/fasebj.10.7.8635688.

[91] A. Kreimer and I. Pe’er, “Variants in exons and in transcription factors affect gene expression in trans,” *Genome Biol.*, vol. 14, no. 7, 2013, doi: 10.1186/gb-2013-14-7-r71.

[92] A. Dvir, R. C. Conaway, and J. W. Conaway, “A role for TFIIH in controlling the activity of early RNA polymerase II elongation complexes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 17, pp. 9006–9010, 1997, doi: 10.1073/pnas.94.17.9006.

[93] Q. Zhou, T. Li, and D. H. Price, “RNA polymerase II elongation control,” *Annu. Rev. Biochem.*, vol. 81, pp. 119–143, 2012, doi: 10.1146/annurev-biochem-052610-095910.

[94] M. Floris, H. Mahgoub, E. Lanet, C. Robaglia, and B. Menand, “Post-transcriptional regulation of gene expression in plants during abiotic stress,” *Int. J. Mol. Sci.*, vol. 10, no. 7, pp. 3168–3185, 2009, doi: 10.3390/ijms10073168.

[95] M. Ghildiyal and P. D. Zamore, “Small silencing RNAs: An expanding universe,” *Nat. Rev. Genet.*, vol. 10, no. 2, pp. 94–108, 2009, doi: 10.1038/nrg2504.

[96] M. A. Valencia-Sanchez, J. Liu, G. J. Hannon, and R. Parker, “Control of translation and mRNA degradation by miRNAs and siRNAs,” *Genes Dev.*, vol. 20, no. 5, pp. 515–524, 2006, doi: 10.1101/gad.1399806.

[97] E. Huntzinger and E. Izaurralde, “Gene silencing by microRNAs: Contributions of translational repression and mRNA decay,” *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 99–110, 2011, doi: 10.1038/nrg2936.

[98] M. C. Siomi, K. Sato, D. Pezic, and A. A. Aravin, “PIWI-interacting small

RNAs: The vanguard of genome defence," *Nat. Rev. Mol. Cell Biol.*, vol. 12, no. 4, pp. 246–258, 2011, doi: 10.1038/nrm3089.

[99] T. Murasnige, "A Revised Medium for Rapid Growth and Bio Agsays with Tohaoco Tissue Cultures," vol. 15, 1962.

[100] E. G. Mendonça, L. V. Paiva, and V. C. Stein, "Growth Curve and Development of the Internal Calli Structure of Eucalyptus camaldulensis Dehn," vol. 55, no. December, pp. 887–896, 2012.

[101] İ. Ünİversİtesİ, T. Telomeraz, and Sï. Üzerİne, "FEN BİLİMLERİ ENSTİTÜSÜ," 2019.

[102] S. Audic and J. M. Claverie, "The significance of digital gene expression profiles," *Genome Res.*, vol. 7, no. 10, pp. 986–995, 1997, doi: 10.1101/gr.7.10.986.

[103] E. P. Nawrocki *et al.*, "Rfam 12.0: Updates to the RNA families database," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D130–D137, 2015, doi: 10.1093/nar/gku1063.

[104] N. W. Breakfield *et al.*, "High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis," *Genome Res.*, vol. 22, no. 1, pp. 163–176, 2012, doi: 10.1101/gr.123547.111.

[105] K. Wang *et al.*, "Prediction of piRNAs using transposon interaction and a support vector machine," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–8, 2014, doi: 10.1186/s12859-014-0419-6.

[106] H. J. Wu, Y. K. Ma, T. Chen, M. Wang, and X. J. Wang, "PsRobot: A web-based plant small RNA meta-analysis toolbox," *Nucleic Acids Res.*, vol. 40, no. W1, pp. 22–28, 2012, doi: 10.1093/nar/gks554.

[107] B. Xiaochen and S. Wang, "TargetFinder: A software for antisense oligonucleotide target site selection based on MAST and secondary structures of target mRNA," *Bioinformatics*, vol. 21, no. 8, pp. 1401–1402,

2005, doi: 10.1093/bioinformatics/bti211.

[108] H. Jiang and W. H. Wong, “Statistical inferences for isoform expression in RNA-Seq,” vol. 25, no. 8, pp. 1026–1032, 2009, doi: 10.1093/bioinformatics/btp113.

[109] H. Abdi, “The Bonferonni and Šidák Corrections for Multiple Comparisons,” *Encycl. Meas. Stat.*, pp. 103–107, 2007, doi: 10.4135/9781412952644.

[110] K. J. Livak and T. D. Schmittgen, “Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method,” *Methods*, vol. 25, no. 4, pp. 402–408, 2001, doi: 10.1006/meth.2001.1262.

[111] X. Shen *et al.*, “Transcriptomic profiling revealed an important role of cell wall remodeling and ethylene signaling pathway during salt acclimation in Arabidopsis,” *Plant Mol. Biol.*, vol. 86, no. 3, pp. 303–317, 2014, doi: 10.1007/s11103-014-0230-9.

[112] C. Fellenberg, C. Böttcher, and T. Vogt, “Phenylpropanoid polyamine conjugate biosynthesis in Arabidopsis thaliana flower buds,” *Phytochemistry*, vol. 70, no. 11–12, pp. 1392–1400, 2009, doi: 10.1016/j.phytochem.2009.08.010.

[113] M. Xie, G. Ren, P. Costa-Nunes, O. Pontes, and B. Yu, “A subgroup of SGS3-like proteins act redundantly in RNA-directed DNA methylation,” *Nucleic Acids Res.*, vol. 40, no. 10, pp. 4422–4431, 2012, doi: 10.1093/nar/gks034.

[114] X. Yuan, Z. Wang, J. Huang, H. Xuan, and Z. Gao, “Phospholipidase dδ negatively regulates the function of resistance to pseudomonas syringae pv. Maculicola 1 (RPM1),” *Front. Plant Sci.*, vol. 9, no. January, pp. 1–9, 2019, doi: 10.3389/fpls.2018.01991.

[115] B. O. R. Bargmann and T. Munnik, “The role of phospholipase D in plant stress responses,” *Curr. Opin. Plant Biol.*, vol. 9, no. 5, pp. 515–522, 2006, doi: 10.1016/j.pbi.2006.07.011.

[116] J. Wang *et al.*, “Effects of Sulfur and Selenium on Glucosinolate

Biosynthesis in Cabbage," *Plant Mol. Biol. Report.*, vol. 38, no. 1, pp. 62–74, 2020, doi: 10.1007/s11105-019-01178-x.

[117] T. C. Barickman, D. A. Kopsell, and C. E. Sams, "Selenium In fl uences Glucosinolate and Isothiocyanates and Increases Sulfur Uptake in Arabidopsis thaliana and Rapid-Cycling Brassica oleracea," 2013.

[118] H. W. Marco, K. Marc, H. Bäumlein, B. Weisshaar, and P. Bauer, "Iron de W ciency-mediated stress regulation of four subgroup Ib BHLH genes in Arabidopsis thaliana," pp. 897–908, 2007, doi: 10.1007/s00425-007-0535-x.

[119] Y. Yuan *et al.*, "FIT interacts with AtbHLH38 and AtbHLH39 in regulating iron uptake gene expression for iron homeostasis in Arabidopsis," pp. 385–397, 2008, doi: 10.1038/cr.2008.26.

[120] A. Golisz, M. Sugano, and S. Hiradate, "Microarray analysis of Arabidopsis plants in response to allelochemical L -DOPA," pp. 231–240, 2011, doi: 10.1007/s00425-010-1294-7.

[121] S. Haq *et al.*, "Heat Shock Proteins : Dynamic Biomolecules to Counter Plant Biotic and Abiotic Stresses," pp. 1–31.

[122] F. McLoughlin *et al.*, "Class I and II small heat shock proteins together with HSP101 protect protein translation factors during heat stress," *Plant Physiol.*, vol. 172, no. 2, pp. 1221–1236, 2016, doi: 10.1104/pp.16.00536.

[123] M. Weber, A. Trampczynska, and S. Clemens, "Comparative transcriptome analysis of toxic metal responses in Arabidopsis thaliana and the Cd 2 + -hypertolerant facultative metallophyte Arabidopsis halleri," pp. 950–963, 2006, doi: 10.1111/j.1365-3040.2005.01479.x.

[124] H. Pauliina *et al.*, "Transcriptional effects of cadmium on iron homeostasis differ in calamine accessions of Noccaea caerulescens," doi: 10.1111/tpj.14121.

[125] Z. Zhai *et al.*, "OPT3 Is a Phloem-Speci fi c Iron Transporter That Is Essential

for Systemic Iron Signaling and Redistribution of Iron and Cadmium in Arabidopsis," vol. 26, no. May, pp. 2249–2264, 2014, doi: 10.1105/tpc.114.123737.

[126] B. D. Pant *et al.*, "Identification of nutrient-responsive Arabidopsis and rapeseed microRNAs by comprehensive real-time polymerase chain reaction profiling and small RNA sequencing," *Plant Physiol.*, vol. 150, no. 3, pp. 1541–1555, 2009, doi: 10.1104/pp.109.139139.

[127] B. A. Akpinar, M. Kantar, and H. Budak, "Root precursors of microRNAs in wild emmer and modern wheats show major differences in response to drought stress," *Funct. Integr. Genomics*, vol. 15, no. 5, pp. 587–598, 2015, doi: 10.1007/s10142-015-0453-0.

[128] T. Hewezi, T. R. Maier, D. Nettleton, and T. J. Baum, "The arabidopsis microrna396-GRF1/GRF3 regulatory module acts as a developmental regulator in the reprogramming of root cells during cyst nematode infection," *Plant Physiol.*, vol. 159, no. 1, pp. 321–335, 2012, doi: 10.1104/pp.112.193649.

[129] C. Shao, X. Ma, M. Chen, and Y. Meng, "Characterization of expression patterns of small RNAs among various organs in Arabidopsis and rice based on 454 platform-generated high-throughput sequencing data," *Plant Omics*, vol. 5, no. 3, pp. 298–304, 2012.

[130] I. S. Day, V. S. Reddy, G. Shad Ali, and A. S. Reddy, "Analysis of EF-hand-containing proteins in Arabidopsis.," *Genome Biol.*, vol. 3, no. 10, pp. 1–24, 2002, doi: 10.1186/gb-2002-3-10-research0056.

[131] Y. Hong, X. Pan, R. Welti, and X. Wang, "Phospholipase Dα3 is involved in the hyperosmotic response in Arabidopsis," *Plant Cell*, vol. 20, no. 3, pp. 803–816, 2008, doi: 10.1105/tpc.107.056390.

[132] R. Datta, P. Boro, K. Mandal, A. Sultana, and S. Chattopadhyay, "Glutathione imparts stress tolerance against Alternaria brassicicola infection via miRNA mediated gene regulation," pp. 1–13, 2019, doi:

10.21203/rs.2.12645/v1.

[133] C. Schommer, E. G. Bresso, S. V. Spinelli, and J. F. Palatnik, “Role of MicroRNA miR319 in Plant Development,” pp. 29–47, 2012, doi: 10.1007/978-3-642-27384-1_2.

[134] F. Thiebaut *et al.*, “Computational identification and analysis of novel sugarcane microRNAs,” *BMC Genomics*, vol. 13, no. 1, 2012, doi: 10.1186/1471-2164-13-290.

[135] G. Liang, H. He, and D. Yu, “Identification of Nitrogen Starvation-Responsive MicroRNAs in Arabidopsis thaliana,” *PLoS One*, vol. 7, no. 11, 2012, doi: 10.1371/journal.pone.0048951.

[136] C. A. Matthewman, C. G. Kawashima, D. Húska, T. Csorba, T. Dalmay, and S. Kopriva, “MiR395 is a general component of the sulfate assimilation regulatory network in Arabidopsis,” *FEBS Lett.*, vol. 586, no. 19, pp. 3242–3248, 2012, doi: 10.1016/j.febslet.2012.06.044.

[137] S. Paul, S. K. Datta, and K. Datta, “miRNA regulation of nutrient homeostasis in plants,” *Front. Plant Sci.*, vol. 6, no. APR, pp. 1–11, 2015, doi: 10.3389/fpls.2015.00232.

[138] A. Noman and M. Aqeel, “miRNA-based heavy metal homeostasis and plant growth,” *Environ. Sci. Pollut. Res.*, vol. 24, no. 11, pp. 10068–10082, 2017, doi: 10.1007/s11356-017-8593-5.

[139] S. Sarkar Das *et al.*, “Expression dynamics of miRNAs and their targets in seed germination conditions reveals miRNA-ta-siRNA crosstalk as regulator of seed germination,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018, doi: 10.1038/s41598-017-18823-8.

[140] H. F. Kuo and T. J. Chiou, “The role of microRNAs in phosphorus deficiency signaling,” *Plant Physiol.*, vol. 156, no. 3, pp. 1016–1024, 2011, doi: 10.1104/pp.111.175265.

Publications from the Thesis

Papers

Mhiri, M. Ceylan, N. Turgut-Kara, B. Nalbantoğlu, and Ö. Çakır, "Transcriptomic analysis reveals responses to Cycloastragenol in *Arabidopsis thaliana*," PLoS One, vol. 15, no. 12 December, pp. 1–19, 2020, https://doi.org/10.1371/journal.pone.0242986.

Mhiri, M. N. Turgut-Kara, B. Nalbantoğlu, and Ö. Çakır, Analysis of microRNAs in response to cycloastragenol treatment by small RNA sequencing in *Arabidopsis thaliana*. Under submission.

Conference oral presentations

W. Mhiri, M. R. Ceylan, N. Turgut-Kara, B. Nalbantoğlu and Ö. Çakır, 2020. Transcriptomic analysis of *Arabidopsis thaliana* plants treated with Cycloastragenol. International Eurasian Conference on BioTechnology and BioChemistry (BioTechBioChem 2020), Ankara/Turkey.

Projects

Research Fund of Yildiz Technical University (BAP Project N°: 3489)