

# **De novo transcriptomic profiling of sap of papaya plant under drought stress condition**



**THESIS SUBMITTED IN PARTIAL FULFILMENT OF  
THE REQUIREMENTS FOR THE AWARD OF  
DEGREE OF**

**Master of Science**  
**in**  
**Plant Biotechnology**

**Submitted by**

**Gargi Kashyap**

**Supervisor**

**Dr. Rajesh Kumar**

**DEPARTMENT OF GENETICS AND PLANT BREEDING  
INSTITUTE OF AGRICULTURAL SCIENCES  
BANARAS HINDU UNIVERSITY, VARANASI -221005  
INDIA**

Ref. No.....

Date.....

### **CERTIFICATE**

To,  
The Registrar (Academic)  
Banaras Hindu University,  
Varanasi-221005 (India)

Through: The Head, Department of Genetics and Plant Breeding, Institute of  
Agricultural Sciences, Banaras Hindu University, Varanasi.

Dear Sir,

I have great pleasure in forwarding the thesis entitled “**De Novo transcriptomic profiling of sap of papaya plant under drought stress condition**” submitted by **Ms. Gargi Kashyap (ID. No. 20430PLB016)** in partial fulfilment of the requirements for the award of the degree of “**Master of Science in Plant Biotechnology**”, from Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, Banaras Hindu University.

I certify that the entire scheme of investigation reported here in, was planned and carried out by the candidate under my guidance. To the best of my knowledge and belief, the data presented in the thesis are genuine and original. No part of the work has been submitted for any degree or distinction.

Thanking you,

**Forwarded by**

**Head**  
**Genetics and Plant Breeding**

**Course Coordinator**  
**Plant Biotechnology**

**Dr. Rajesh Kumar**  
**Advisor-Chairman**

# **De novo transcriptomic profiling of sap of papaya plant under drought stress condition**



THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT OF  
OR THE AWARD OF THE DEGREE OF

**Master of Science**

in

**Plant Biotechnology**

By

*Gargi Kashyap*

**ID.No.20430PLB016**

**2022**

**Enrolment No.435225**

**APPROVED BY ADVISORY COMMITTEE**

- 1. Dr. Rajesh Kumar (Supervisor)**  
Assistant Professor  
Dept. of Genetics and Plant Breeding  
Institute of Agricultural Sciences  
RGSC, Banaras Hindu University
- 2. Dr. Ashok Kumar (Member)**  
Assistant Professor  
Dept. of Genetics and Plant Breeding  
Institute of Agricultural Sciences  
RGSC, Banaras Hindu University
- 3. Dr. Vishal Srivastav (Member)**  
Assistant Professor  
Dept. of Genetics and Plant Breeding  
Institute of Agricultural Sciences  
RGSC, Banaras Hindu University
- 4. External**

# DECLARATION

I hereby declare that the dissertation entitled Studies on **“De Novo transcriptomic profiling of sap of papaya plant under drought stress condition”**. submitted to the Department of genetics and plant breeding, Banaras Hindu University, Varanasi for partial fulfillment of the requirement for the master’s degree in Plant Biotechnology is a bonafide record of research work carried out by me under the guidance and supervision of Dr. Rajesh kumar, Assistant Professor (Plant Biotechnology) Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, BHU, Varanasi, India and no part of this report has been presented by me for any degree.

**Date:**

**Place:** Varanasi

**(GARGI KASHYAP)**

# AKNOWLEDGEMENT

---

First and foremost, praise and appreciation to **God**, the Almighty, for his showers of blessings during my research effort, allowing me to successfully complete my research.

I bow my head and offer flowers at the pious feet of founder of this BHU Bharat Ratna **Mahamana Pandit Madan Mohan Malviya ji** for creating this unique university which is temple for thousands of students like me.

**Prof. B. Sinha**, Head of the Department of Genetics and Plant Breeding, has been an inspiration to me. First and foremost, I would like to express my heartfelt gratitude to my advisor **Dr. Rajesh Kumar** (Assistant Professor) Plant Biotechnology , Department of Genetics and Plant Breeding ,Institute of Agricultural Sciences, Banaras Hindu University, and also Chairman of my Advisory Committee for his enduring interest, continuous support of my M.Sc. study and research, for his patience, motivation, enthusiasm His advice was invaluable during the research and writing of this thesis.

Aside from my adviser, I'd like to thank the remainder of my advisory committee: **Dr. Vishal Srivastav** (Assistant Professor) and **Dr. Ashok Kumar** (Assistant Professor) for their support, insightful comments, and difficult questions.

My heartfelt gratitude also goes to **Dr. Mrunal Gokhale** (Scientific Officer, Bionivid Technology Pvt. Ltd., Bangalore) for providing me with online internship possibilities in their groups and guiding me through the process of working on different and intriguing projects.

I'd like to thank my parents, **Anupma and Harendra Kumar**, for giving me life in the first place and for their spiritual support throughout my life. I am grateful to my parents for their love, prayers, care, and sacrifices in teaching and preparing me for the future. I'd want to thank my younger, small brother (**Prakhar kashyap**) for not bothering me throughout my thesis.

Finally, I'd want to thank everyone who helped me complete the research work, whether directly or indirectly.

I would like to thanks my roommate and a lovely friend **Pragati Mardi** for being a pillar of strength through the thick and thins of this phase. I would also thanks to **Shrestha Yadav** for being incredible partner in this process . I would thanks **Jaya , Pranay , Jyotsna and Ishita, Anamika, Anshuli and Uyyala Sivakumar** for the support throughout.

I would specially thanks to **Neha kumari** for being the pillar of strength and motivation for every possible way and with whole zest . And infinite thank you to my all time favourite **Rashila Maharjan** for making this process smooth and helping me through my vulnerable times.

Lastly, I would like to thanks everybody collectively , even if not mentioned as it doesn't means lack of gratitude, just a human error.

Date

Place:

(GARGIKASHYAP)

## **LIST OF CONTENTS**

| <b>Serial no.</b> | <b>Particulars</b>            | <b>Page no.</b> |
|-------------------|-------------------------------|-----------------|
| <b>1.</b>         | <b>INTRODUCTION</b>           | 1-4             |
| <b>2.</b>         | <b>REVIEW OF LITERATURE</b>   | 5-15            |
| <b>3.</b>         | <b>MATERIALS AND METHODS</b>  | 16 - 24         |
| <b>4.</b>         | <b>RESULTS</b>                | 25-55           |
| <b>5.</b>         | <b>DISSCUSSION</b>            | 56-60           |
| <b>6.</b>         | <b>CONCLUSION AND SUMMARY</b> | 61-62           |
| <b>7.</b>         | <b>BIBLIOGRAPHY</b>           | i-iii           |

## **LIST OF ABBREVIATIONS**

|                               |                                                                    |
|-------------------------------|--------------------------------------------------------------------|
| RNA                           | Ribonucleic Acid                                                   |
| NGS                           | Next Generation Sequencing                                         |
| DNA                           | Deoxyribo nucleic Acid                                             |
| EMBL                          | European Molecular Biology Laboratory                              |
| NCBI                          | National Center for Biotechnology Information                      |
| BLAST                         | Basic Local Alignment Search Tool                                  |
| SAM                           | Sequence Alignment / Map                                           |
| BED                           | Brower Extensible Data                                             |
| BAM                           | Binary Alignment And Map                                           |
| DAVID                         | Database For Annotation , Visualisation , and Integrated Discovery |
| bioDB net                     | Biological Database And Network                                    |
| DEG                           | Differentially Expressed Genes                                     |
| PCR                           | Polymerase chain reaction                                          |
| <i>et al.</i>                 | et alia (and others)                                               |
| KEGG                          | Kyoto Encyclopedia of Gene amd Genomics                            |
| H <sub>2</sub> O <sub>2</sub> | Hydrogen peroxide                                                  |
| GO                            | Gene ontology                                                      |
| Mi RNA                        | MicroRNA                                                           |
| LEA                           | Late embryogenesis abundant                                        |
| SNP                           | Single nucleotide polymorphism                                     |
| Cov.                          | Coverage                                                           |

## **LIST OF TABLES**

| <b>S. No</b> | <b>Table No.</b> | <b>Particulars</b>                              | <b>Page no.</b> |
|--------------|------------------|-------------------------------------------------|-----------------|
| 1.           | 2.1              | Taxonomic classification                        | 3               |
| 2.           | 2.2              | Medicinal and therapeutic value of papaya       | 5               |
| 3.           | 2.3              | History of DNA Sequencing methods               | 8               |
| 4.           | 3.1              | Selected raw reads                              | 19              |
| 5.           | 3.2              | URLs of different database , software and tools | 24              |
| 6.           | 4.1              | Raw reads for analysis                          | 25              |
| 7.           | 4.2              | Basic statistics of reads after Fastqc          | 26              |
| 8.           | 4.3              | Depiction of sequencing data                    | 31              |
| 9.           | 4.4              | Data before doing NGSQC Toolkit                 | 32              |
| 10.          | 4.5              | Data after doing NGSQC Toolkit                  | 33              |
| 11.          | 4.6              | Trinity results                                 | 52              |
| 12.          | 4.7              | bioDBnet annotation                             | 54              |
| 13.          | 4.8              | Final ontology                                  | 55              |

## **LIST OF FIGURES**

| <b>SL. No</b> | <b>Table No.</b> | <b>Particulars</b>                                                                | <b>Page no.</b> |
|---------------|------------------|-----------------------------------------------------------------------------------|-----------------|
| 1.            | 2.1              | Various tools included in NGSQC Toolkit                                           | 11              |
| 2.            | 2.2              | Flow chart of genome assembly and genome annotation                               | 13              |
| 3.            | 3.1              | Work pipeline de novo Transcriptome analysis                                      | 19              |
| 4.            | 2.4              | Flow chart of genome assembly and genome annotation                               | 20              |
| 5.            | 4.1.1            | Per base sequence quality for SRR7145702_1.fastq.gz.                              | 27              |
| 6.            | 4.1.2            | Per base sequence quality for SRR7145702_2.fastq.gz                               | 27              |
| 7.            | 4.2.1            | Per base sequence quality for SRR7145708_1.fastq.gz                               | 28              |
| 8.            | 4.2.2            | Per base sequence quality for SRR7145708_2.fastq.gz                               | 28              |
| 9.            | 4.3.1            | Per sequence quality score for SRR7145702 (forward read)                          | 29              |
| 10.           | 4.3.2            | Per sequence quality score for SRR7145702 ( backward read)                        | 29              |
| 11.           | 4.4.1            | Per sequence quality score for SRR7145708 ( forward read)                         | 30              |
| 12.           | 4.4.2            | Per sequence quality score for SRR7145708 ( backward read)                        | 30              |
| 13.           | 4.5.1            | Per base average quality score for SRR7145702 ( forward read)                     | 34              |
| 14.           | 4.5.2            | Per base average quality score for SRR7145702 ( backward read)                    | 35              |
| 15.           | 4.6.1            | Per base average quality score for SRR7145708 ( forward read)                     | 36              |
| 16.           | 4.7.1            | Read count % per base for SRR7145708 Forward read before and after NGSQC toolkit  | 36              |
| 17.           | 4.7.2            | Read count % per base for SRR7145708 Forward read before and after NGSQC toolkit  | 37              |
| 18.           | 4.7.3            | Read count % per base for SRR7145708 Backward read before and after NGSQC toolkit | 38              |
| 19.           | 4.7.4            | Read count % per base for SRR7145708 Backward read before and after NGSQC toolkit | 38              |
| 20.           | 4.8.1            | Read count % per base for SRR7145702 forward read before and after NGSQC toolkit  | 39              |
| 22.           | 4.8.3            | Read count % per base for SRR7145702 backward read before and after NGSQC toolkit | 39              |
| 23.           | 4.8.4            | Read count % per base for SRR7145702 backward read before and after NGSQC toolkit | 40              |
| 24.           | 4.9.1            | Base composition % SRR7145702 Forward Read                                        | 40              |
| 25.           | 4.9.2            | Base composition % SRR7145702 Backward Read                                       | 41              |
| 26.           | 4.10.1           | Base composition % SRR7145708 Forward Read                                        | 42              |
| 27.           | 4.10.2           | Base composition % SRR7145708 Backward Read                                       | 43              |
| 28.           | 4.11.1           | % GC content SRR7145702 (FORWARD READ )                                           | 44              |

|     |        |                                                 |    |
|-----|--------|-------------------------------------------------|----|
| 29. | 4.11.2 | % GC content SRR7145702 (BACKWARD READ )        | 45 |
| 30. | 4.12.1 | % GC content SRR7145708 (FORWARD READ )         | 46 |
| 31. | 4.12.2 | % GC content SRR7145708 (BACKWARD READ )        | 47 |
| 32. | 4.13.1 | Qualitydistribution of SRR7145702 FORWARD READ  | 48 |
| 33. | 4.13.2 | Qualitydistribution of SRR7145702 BACKWARD READ | 49 |
| 34. | 4.14.1 | Qualitydistribution of SRR7145708 FORWARD READ  | 49 |
| 35. | 4.14.2 | Qualitydistribution of SRR7145708 BACKWARD READ | 50 |
| 36. | 4.15.1 | QC Summary of SRR7145702                        | 51 |
| 37. | 4.15.2 | QC Summary of SRR7145708                        | 51 |

## **CHAPTER 1: INTRODUCTION**

Papaya being tropical plant natively from tropical America from family Caricaceae. Being grown under equatorial conditions where rainfall is not limiting factor it is considered as drought susceptible crop. So not availability of ample amount of water may cause direct root injury, sap distortion and causing fatal conditions. Even cultural and agronomic measures adoptions help us overcome the drought stress conditions even in recent times but still it threatens to decrease the productivity of crop at agricultural level causing severe economical losses. Because of its high nutritional content, outstanding flavour, and several uses in the food business, papaya production has increased dramatically over the last few years. The papaya is now the third most popular tropical fruit in the world. Because it can be grown in such a wide range of settings, cultivated papaya has grown in popularity in tropical and subtropical countries. However, in order to reach great output, this crop requires continual irrigation (Fuchigami *et al.*, 1981). Commercial papaya genotypes may have 50% poorer yields in water limited conditions.

Wild native *C. papaya* populations can be found on the Yucatan Peninsula's undisturbed desert and semi-arid regions. In their natural habitat, these wild papayas have been exposed to a variety of environmental challenges, including drought and high temperatures. Wild papayas may have adapted to the harsh conditions of their environment (Hu H, Xiong *et al.*, 2014). The fruits of these wild papayas, however, are little and of poor quality, making them unmarketable. Commercial genotypes, on the other hand, were selected to provide large fruit yields, although they are vulnerable to abiotic conditions like drought and heat.

Under drought stress condition papaya being tropical plant starts showing adjustments in roots, sap and leaf like modification of roots architecture, modification of sodium, potassium content and increase of biosynthesis of suberin in organ etc to regulate water stress condition.

Amid drought stress conditions, many protein coding genes, which are being involved in regulating stress conditions are regulated by ABA dependent and ABA non dependent signaling pathways. These genes are classified as regulatory and functional genes (Laxmi *et al.*, 2015). Late embryogenesis abundant (LEA) proteins, heat shock proteins (HSP), reactive oxygen species (ROS), scavenging enzymes, and osmoprotectant synthesizing enzymes are only a few examples of functional genes. Transcription factors (TF), kinases, and phosphatases are examples of regulatory genes that affect the expression and/or activity of other genes.

For coordinated plant development and quick responses to unfavourable conditions, efficient communication of sensed external signals between distal tissues is critical. The vascular system controls the long-distance transport of a variety of molecules (including water, nutrients, and photo assimilates) between distal tissues (Wolf *et al.*, 2013). The mobility, distribution, and trafficking of these and other macromolecules, such as proteins and RNAs, is controlled by the phloem sap (Buhtz *et al.*, 2008). The phloem's molecular reaction to drought, on the other hand, has received less attention.

Plants' complicated molecular responses to drought and other types of stress have necessitated the use of omics methods to decipher their sophisticated mechanisms. The introduction of Next Generation Sequencing (NGS) technology, also known as High-Throughput Sequencing, has allowed researchers to examine numerous plant transcriptomes by sequencing their RNA (RNA-Seq) species, such as mRNA and microRNA (Giovannoni *et al.*, 2013). In reaction to dryness, this has resulted in large data sets containing hundreds or thousands of regulated genes. The major goal of any transcriptome investigation is to give an in-depth comparative examination of the geographic and temporal profile of expressed genes as well as the abundance of individual transcripts in different samples. The biological sample may be chosen to examine a certain stage or body portion

of an organism during its development or evolution, in response to a certain therapy or just to construct a transcriptome. An organism's atlas RNA-Seq is now the method of choice for Transcriptome studies: little amounts of input RNA/can be applied from single cells to the whole organism level; creates low levels of background signal and informs about transcript abundance; and allows investigation of gene expression in a species with or without a reference gene. The RNA-Seq technique and the different software tools used for de novo transcriptome assemblies have opened up a new avenue for investigating non-model species that do not have a sequenced reference genome. The transcriptome assembly from scratch may be used to align sequence reads from the same or a different experiment in order to discover differential gene expression and to investigate the hereditary variety. RNA-sequencing (RNA-seq) is a versatile tool with applications in a wide range of plant biology domain. This approach can be used to find expressed sequences in a particular tissue at any given time (Snyder *et al.*, 2009). It's been utilised to investigate tissue-specific transcripts in response to abiotic and biotic stressors. There have been several transcriptomic studies in papaya roots, flowers, fruit ripening, and a transcriptome analysis of leaves, sap, and roots of papaya plants under drought stress, but no reports on transcriptomic studies in response to drought in papayas have been published to date. It's important to compare transcriptome changes in wild and commercial papaya genotypes in response to WDS because understanding the mechanisms involved in this tropical crop's drought stress response is crucial. Similarly, finding differently expressed TF genes in papaya plants subjected to drought may shed light on the molecular mechanisms that confer drought tolerance in this species. We used Illumina RNA-Seq to examine the transcriptome of papaya leaves, sap, and roots under well-watered (control) conditions, as well as after 10 and 20 days of drought stress. We used reference and de novo assembly methodologies to identify tissue-specific collections of differentially expressed genes (DEGs).

Specific biological processes regulated among tissues under control and in response to drought were revealed by functional enrichment analysis of these sets of DEGs. Furthermore, we identified numerous abiotic stress-related TFs using co-expression network analysis, which may operate as putative regulatory hubs in sap and roots under drought stress. Our findings give us a better knowledge of the molecular responses of papaya plants to drought stress, as well as important baseline data for future genetic improvement. The discovery of differentially expressed TF genes in drought-stressed papaya plants may shed light on the molecular mechanisms that confer drought tolerance in this species.

The goal of this work was to uncover some of the physiological and molecular factors that allow papayas to withstand drought better than commercial papayas.

## **CHAPTER 2: LITERATURE REVIEW**

### **PAPAYA ORIGIN AND DISTRIBUTION**

The papaya tree (*C. papaya*) is a popular and economically significant tropical and subtropical fruit tree. The fruit is consumed as a fresh fruit, a vegetable, or a processed product. Worldwide over 6.8 million tonnes (Mt) of fruit were produced in 2004 on about 389,990 Ha (FAO 2004). Of this volume, 47% was produced in Central and South America (mainly in Brazil), 30% in Asia, and 20% in Africa (Fao *et al.*, 2004). The papaya industry in Brazil is one of the largest worldwide that continues to show rapid growth. (Sousa *et al.*, 2003) reported on a 151% increase in total area cultivated over the past decade (16,012 ha in 1990 to 40,202 ha in 2000) and a 164% increase in the quantity produced during the same period (642,581 to 1,693,779 fruits from 1990 to 2000). In 11 years, the volume exported increased 560% from 4,071 t to 22,804t in 2001 (Secexmdic *et al.*, 2002) and 38,760 t in 2005 (Fao *et al.*, 2005). Although papaya is mainly grown (>90%) and consumed in developing countries, it is fast becoming an important fruit internationally, both as a fresh fruit and as processed products

The classification of papaya has changed significantly throughout time. *C. papaya* was previously categorised as a member of the Passifloraceae, Cucurbitaceae, Bixaceae, and Papayaceae plant families. However, it is currently classified as part of the Caricaceae family, which includes 35 latex-producing species divided into four genera: *Carica*, *Cylico*, *Morpha*, *Jarilla*.

## CLASSIFICATION AND BOTANY

Table 2.1. Taxonomic classification (Ayurvedic pharmacopoeia of India, Govt of India)

| Scientific classification of <i>C. papaya</i> |                            |
|-----------------------------------------------|----------------------------|
| Kingdom                                       | PLANTAE                    |
| Division                                      | SPERMOPHYTA                |
| Class                                         | MAGNOLIOPSIDA              |
| Order                                         | BRASSICALES                |
| Family                                        | CARICACEAE                 |
| Genus                                         | <i>Carica</i>              |
| Species                                       | <i>Carica papaya</i> Linn. |

The papaya tree is a huge, single-stemmed herbaceous perennial tree that grows to a height of 20–30 feet. The leaves are huge (up to 212 feet broad), palmately lobed or deeply incised with whole margins, and petioles ranging from 1-3 feet long. The stems are hollow, light green to tan brown in hue, and have an 8-inch diameter with noticeable scars.

The fruits are large oval fruits that are sometimes referred to as pepo-like berries because they have a central seed cavity that resembles that of a melon. Fruits are borne on the main stem axillary, generally single but occasionally in tiny groups. Fruits range in size from 0.5 to 20 pounds and start out green before turning yellow or crimson orange when ripe. At maturity, the flesh is yellow-orange to salmon (pinkish orange). The edible component of the seed is located around the huge central seed cavity. Individual fruits might take anywhere from 5 to 9 months to ripen, depending on the cultivator and temperature. In 6 to 12 months, the plants begin to bear fruit.

Papaya plants are either dioecious or hermaphroditic, meaning they produce blooms that are either male, female, or bisexual (hermaphroditic). Papayas are frequently referred to be

"trioecious," which means that different plants produce male, female, or bisexual blooms. Female and bisexual blooms are waxy, ivory white, and borne in leaf axils along the main stem on small penducles. Flowers are solitary or in tiny cymes of three.

### **CHEMICAL COMPOSITION OF PAPAYA**

Chemical composition of various parts of *C. papaya* (Jean Bruneton et al., 1999)

- 1. Juice** - N-butyric, n-hexanoic and n-octanoic acids, lipids; myristic, palmitic, stearic, linoleic, linolenic acids-vaccenic acid and oleic acid.
- 2. Fruit-** protein, fat, fibre, carbohydrates, minerals, calcium, phosphorus, iron, vitamin C, thiamine, riboflavin, niacin, and caroxene, amino acid, citric acids and molic acid (green fruits), volatile compounds : linalol, benzyliothiocynate, cis and trans 2, 6-dimethyl-3,6 expoxy-7 octen-2-ol. Alkaloid,  $\alpha$ ; carpaine, benzyl- $\beta$ -d glucoside, 2-phenylethyl- $\beta$ -D-glucoside, 4-hydroxyl -phenyl-2 ethyl-B-D glucoside and four isomeric malonated benzyl- $\beta$ -D glucosides.
- 3. Seed** - Fatty acids, crude proteins, crude fibre, papaya oil, carpaine, benzyliothiocynate, benzylglucosinolate, glucotropacolin, benzylthiourea, hentriacontane,  $\beta$ -sistosterol, caricin and an enzyme nyrosin.
- 4. Latex-** proteolytic enzymes, papain and chemopapain, glutamine cyclotransferase, chymopapainA, B and C, peptidase A and B and lysozymes.
- 5. Bark-**  $\beta$ -sitosterol, glucose, fructose, sucrose, galactose and xylitol.
- 6. Leaves-** Alkaloids carpain, pseudocarpain and dehydrocarpaine I and II, choline, carposide, vitamin C and E.

**Table 2.2 MEDICINAL AND THEAURPATIC VALUES OF PAPAYA**

| <b><u>PART</u></b>          | <b><u>MEDICINAL USE</u></b>                           |
|-----------------------------|-------------------------------------------------------|
| <b><u>Seed</u></b>          | Carminative, Vermifuge                                |
| <b><u>Unripe fruits</u></b> | Antibacterial and antinflammatory to spleen and liver |
| <b><u>Ripe fruits</u></b>   | Digestive and tackles obesity                         |
| <b><u>Latex</u></b>         | Cures whooping cough                                  |
| <b><u>Stem bark</u></b>     | Cures jaundice                                        |
| <b><u>Leaves</u></b>        | Cures asthma                                          |
| <b><u>Root</u></b>          | Diuretic                                              |
| <b><u>Seed juice</u></b>    | Cures bleeding piles                                  |

**OTHER APPLICATIONS**

- a) Colon cancer:- Papaya fibre binds cancer-causing chemicals in the colon, keeping them away from healthy colon cells. These nutrients work together to protect colon cells from DNA damage caused by freeradicals.
- b) Anti-inflammatory effects:- Protein enzymes like papain and chymopapain, as well as antioxidant elements like vitamin C, E, and betacarotene contained in papaya, help to lower the severity of illnesses including asthma, osteoarthritis, and rheumatoid arthritis.
- c) Rheumatoid arthritis:- Vitamin C-rich foods like papaya protect individuals from inflammatory polyarthritis, a kind of rheumatoid arthritis that affects two or more joint.

- d) Anti-sickling activity:- Recent research indicates that papaya has anti-sickling properties.
- e) Men who consume lycopene-rich foods are less likely to develop prostate cancer. papaya, tomatoes, apricots, and other fruits and vegetables. Pink grapefruit, watermelon, and guava all had an 82 percent lower sugar content. Those who have prostate cancer are more likely to have it than those who don't. consuming the fewest meals high in lycopene.
- f) Anticoagulant action of papain extract injection in a dog. Prothrombin and coagulation are increased thrice. In addition, it is claimed that the enzyme removes necrotic tissues from the body. Chronic wounds, burns, and ulcers are all examples of chronic wounds. Papain is also a type of enzyme. Commercial significance in the brewing sector textile industry and the food industry. (Harish *et al* ., 2013).

## **EVOLVING ROLE OF BIOINFORMATICS IN REASEARCH**

Bioinformatics, as defined by the American Academy of Microbiology in 2009, is the use of computer analysis to study the composition, structure, and interactions of cellular molecules. Whole Genome Sequencing can compile and analyse the genetic sequences of bacteria, archaea, and other microorganisms to reveal information about their processes.

There is a lot of data available from the wet lab that hasn't been analysed or examined yet, and in order to achieve this goal, bioinformatics research must be conducted, which can be divided into three approaches: examine based on the available wet-lab data results, use of mathematical modelling, integrated approach that combines search techniques with mathematical approach, and bioinformatics research used to learn about a specific genome function or metabolic pathway and gene expression of specific microbes, as well as pathogenicity, antibiotic genes, and comparative examination of multiple strains of a specific species, among other things (Bansal, 2005). Whole genome sequencing, annotation,

alignment, phylogenetics or evolutionary links, protein structure prediction, and homology modelling are all examples of bioinformatics applications (VarliH *et al.*, 2014).

Any chemical, catalytic, or technical process for determining the linear order of ester bases in deoxyribonucleic acid is known as DNA sequencing. Sanger sequencing using replicative synthesis in the presence of dideoxy ester chain slayer monomers has now been replaced by next-generation sequencing using short parallel read technology. The word is most commonly used to describe the entire sequence determination pipeline as well as post-sequencing software package analysis. Many scientists contributed to the development of chain termination technique sequencing, capillary DNA sequencing, and next generation sequencing, including Richard Holley, who attempted to sequence the nucleic acid for the first time.

**Table 2.3 History of DNA Sequencing method**

| <b>Name of Scientist<br/>/Biotechnology<br/>company</b> | <b>Year</b> | <b>Achievements</b>                                                             |
|---------------------------------------------------------|-------------|---------------------------------------------------------------------------------|
| Richard Holley                                          | 1964        | First attempt to sequence the nucleic acid                                      |
| Fredrick Sanger                                         | 1977        | Chain termination method of sequencing                                          |
| Maxam and Walter<br>Gilbert                             | 1977        | Genome of bacteriophage X174 was sequenced using<br>Chemical degradation method |
| Lorey and Smith                                         | 1986        | First semi-automated DNA method                                                 |
| Applied Biosystem                                       | 1987        | Fully automated machine-controlled DNA sequencing method                        |
| Applied Biosystem                                       | 1996        | Capillary DNA Sequencing                                                        |
| Solexa/Illumina                                         | 2005        | Next Generation Sequencing                                                      |

Next-generation sequencing separates millions of DNA fragments in parallel, and subsequent reads are analysed using a variety of computer methods (Behjati et al.,2013).

Following are the platforms that fall within the category of next-generation sequencing:

#### **454 Pyrosequencing**

454 Pyrosequencing was named after the firm that pioneered it, 454 Life Sciences. To prevent chain amplification, pyrosequencing uses the polymerase enzyme, which produces complementary strands and dioxynucleotides. The complementary strand is formed by adding dNTPs to a primer, which is ssDNA, and when the proper dNTPs are added, pyrophosphate is released, which is then converted to ATP. Luciferin is converted to oxyluciferin by adenosine triphosphate (ATP), which produces light (Gharizadeh *et al.*,2003).

### **Ion torrent semiconductor sequencing**

Ion torrent sequencing employs a "sequencing by synthesis" method, in which a new DNA strand is produced one base at a time, complementary to the target strand. The DNA library fragment is flooded successively with each nucleoside triphosphate during read generation using emulsion PCR(dNTP). If the dNTP is complementary to the nucleotide on the target strand, it is integrated into the new strand. A hydrogen ion is released each time a nucleotide is successfully inserted, and it is measured by the sequencer's pH sensor. Unfortunately, counting the number of similar base added in a row might be difficult (Rothberg *et al.*,2011).

### **Sequencing by ligation (SOLiD)**

SOLiD is a deoxyribonucleic acid ligase-based deoxyribonucleic acid sequencing method. Because of its ability to ligate double-stranded deoxyribonucleic acid strands, deoxyribonucleic acid ligase is an extensively used catalyst in biotechnology (Ho *et al.*,2011)

### **Illumina sequencing**

Sample preparation, cluster creation, sequencing, and knowledge analysis are the four fundamental aspects of the Illumina sequencing process. In all preparation processes, adaptors are introduced to the ends of deoxyribonucleic acid. Reduced cycle amplification adds extra features such as sequencing binding sites, indices, and sections that are complementary to the flow cell oligos. Sequencing is done with Illumina's HiSeq and MiSeq systems ([www.illumina.com/technology/next-generation-sequencing.html](http://www.illumina.com/technology/next-generation-sequencing.html))

## **2.6 GENOME ASSEMBLY AND ANNOTATION**

Experts all over the world have focused their efforts on making advancements in the sector since the invention of the Sanger sequencing process in order to provide the most cutting-edge technology.

Next-generation sequencing (NGS) is a game-changing technology that promises to significantly increase our understanding of how nucleic acids work. Illumina sequencing is a sort of high-throughput sequencing that is used to study microorganisms to learn about medication resistance outbreaks and spread (Loman et al., 2012). Reads are generated using a variety of sequencing platforms. A fastqreport was prepared to determine the quality of the rawreads; it just takes a fewminutes to complete.

It takes less memory to run and enables for more flexible HTML report presentation (Brandine et al., 2019). Depending on the sequencing, next-generation sequencing technique produces a vast amount of data, which could be long or short reads. Depending on the sequencing technique used, next-generation sequencing technology generated a vast amount of data, which could be long or shortreads. Because read errors, low-quality reads, and primer/adaptor contamination are relatively common in the data, the NGS QC Toolkit is used for filtering and trimming the massive amount of data generated and used in Qualitycheck. It may receive sequencing data in a number of formats and perform quality checks using default or user-defined parameters, as well as generate QC reports. For unfiltered (input) and filtered (output) data, as well as filtered HQdata files, a variety of formats are used (Patel et al., 2012). During genome assembly (Clum, 2018), sequencing data are transformed into contiguous sequences using sequence similarity, and gaps between contigs are filled in by

sequential joining to create scaffolds. De novo assembly employs graph methods such as overlap – layout consensus, which displays an overlap graph, and De Bruijn graph approach, which displays a compact representation (Miller et al., 2010). Two genome assembling approaches are used: VELVET and SPADes.

Because of their highly conserved sequences and prevalence across all genomes, ribosomal RNA (rRNA) genes offer a wide range of uses in bioinformatics. RNAmmer and Barrnap are two approaches for predicting rRNA genes that both employ the HMM methodology (Seemann, 2013).

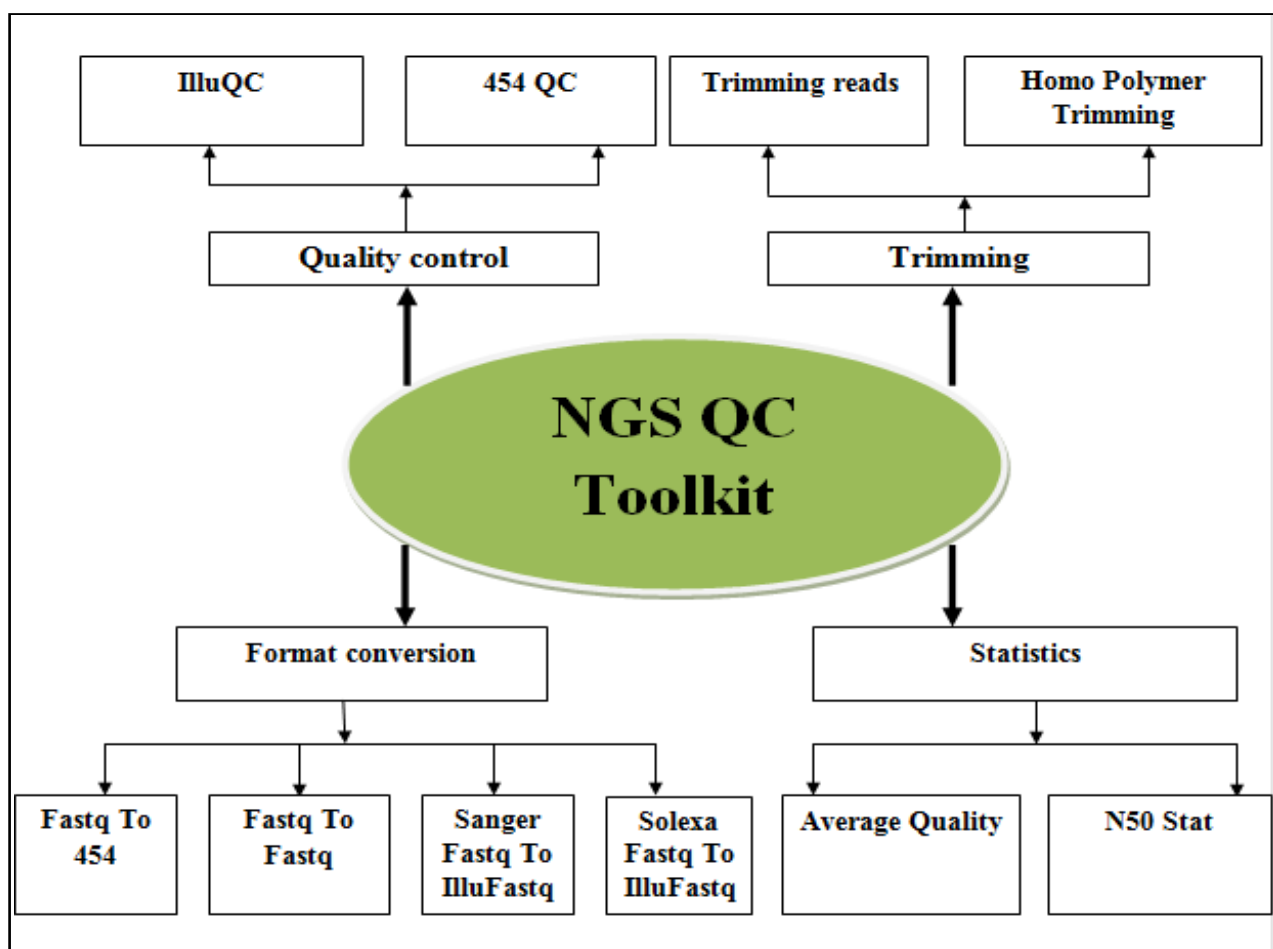


**Fig 2.1 :** Various tools included in NGSQC Toolkit

Completing an organism's genome sequence is a vital endeavor in comparative, functional, and structural genomics. However, both computationally and experimentally, this is still a challenging challenge to tackle. In most genome finishing workflows, de novo assembly genomic scaffolding is typically the first step (Bosi *et al.*, 2015). To complete a genomic scaffold, wetlaboratory procedures, in silico approaches, or a combination of both are used. The sequencing stage's paired-read data could be used as an example of a computational technique. The presence of paired reads in various contigs can be used to probabilistically identify the order and distance between them. The same datasets were scaffolded by MEDUSA in less than 10 minutes on average (Barton *et al.*, 2012). Contiguator, a bacterial genome finishing tool for structural insights on draft genomes (Gardini *et al.*, 2011), and Ragout, a reference-assisted assembly tool for bacterial genomes, are just a couple of the different approaches for mapping (and later structuring) the generated draft contigs (Kolmogorov *et al.*, 2014).

BASys (Bacterial Annotation System) is a web service that automatically annotates bacterial genomic (chromosomal and plasmid) sequences. It takes raw DNA sequence data and an optional list of gene identification information and generates a comprehensive textual annotation and hyperlinked images that include gene/protein name, GO function, COG function, possible paralogues and orthologues, subcellular localization, signalpeptides, transmembrane regions, secondary structure, 3D structure, reactions, and pathways (Gary *et al.*, 2005). MaGe (Magnifying Genomes) is a bacterial genome annotation system with a web interface for completing genome annotation jobs (Vallenet *et al.*, 2006).

KAAS (KEGG Automatic Annotation Server) is a programme that assigns K values to genes in the genome quickly, allowing for the reconstruction of KEGG pathways (Moriya *et al.*, 2007).

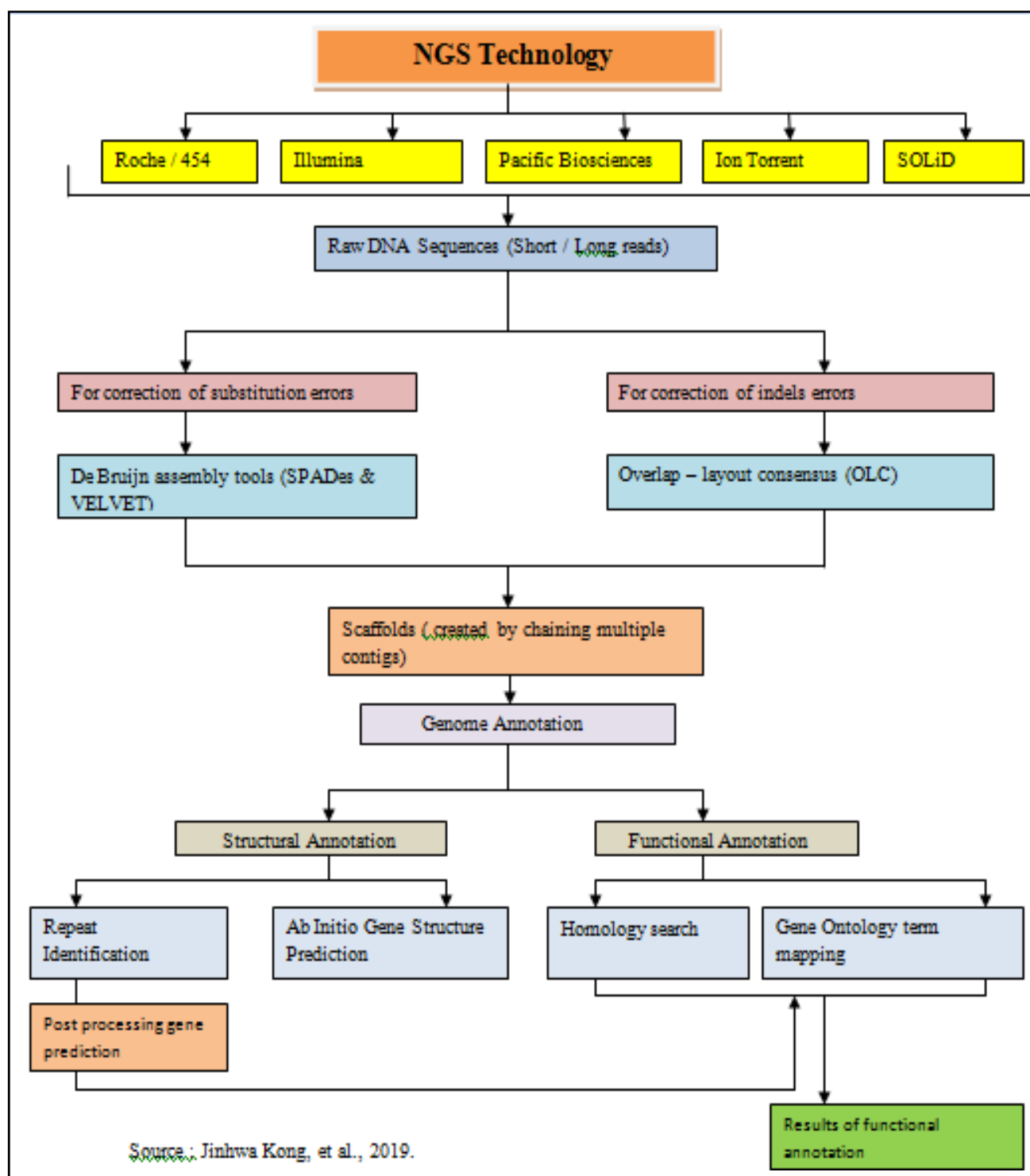


**Fig 2.2 :** Flow chart of genome assembly and genome annotation

## **2.7 Post annotation: Comparative analysis**

In genome visualization methods, linear or circular representations are used. Artemis Comparison Tool (ACT), which offers an interactive visualisation of comparisons between whole genome sequences and associated annotations (Carver *et al.*, 2005), Mauve (Darling *et al.*, 2004), and Genomorama (Darling *et al.*, 2004), are examples of genome search by DNA hybridization (probe binding and PCR amplification). The ability to search for and retrieve data from the National Center for Biotechnology Information (NCBI) Entrez server (Gans *et al.*, 2007) has advantages in displaying insertions and deletions between genomic sequences, and certain programmes, such as Mauve and ACT, may show genome rearrangements. Mauve is a piece of software that can be used to find and align conserved genomic DNA in the face of horizontal transfers and rearrangements. Progressive Mauve builds a large number of genomic alignments using positional homology.

These alignments are significantly distinct from global genome alignments, in which all copies of a repeating gene family may become aligned to one another. Phylogenetic inference of nucleotide substitution, phylogenetic inference of gene gain and loss (Didelot *et al.*, 2008), phylogenetic inference of rearrangement, and even phylogenetic inference of homologous recombination-induced lateral gene transfer (Darling *et al.*, 2008) are all made easier with positional homology alignment (Didelot *et al.*, 2007). However, using these tools to summarise large datasets is difficult. Two programmes that generate circular figures are Microbial Genome Viewer (Kerkhovan *et al.*, 2004) and Genome Projector (Arakawa *et al.*, 2009). They're only good for annotating a single chromosome and don't facilitate full genome comparisons. BRIG (BLAST Ring Image Generator) is a cross-platform desktop tool that allows users to use entire, draft, or unassembled genome data to quickly visualise BLAST comparisons to one or more central reference sequences.

(Alikhan *et al.*, 2011).

Comparable programmes, unlike BRIG, limit the number of genome comparisons that may be displayed on a single image and don't allow you to merge several sequences into a single lane. Comparing to other genomes or sequences is an important step in most sequencing investigations. In many cases, users are also looking for specific genes with well-defined activities, such as virulence genes or medicine resistance determinants. The majority of users will require the ability to visualise these comparisons in order to improve understanding and interpretation of the data.

## **CHAPTER 3: MATERIALS AND METHODS**

This reasearch study of “**De novo transcriptome profiling of plant *Carica papaya* sap in drought stress condition**” was done through online mode at Rajiv Gandhi South Campus ,Banaras Hindu University, Mirzapur, Uttar Pradesh.

### **STEPS FOR DE NOVO TRANSCRITOME ANALYSIS :**

#### **1. Generation or downloading of raw reads for NGS in fastq format .**

The National Centre for Biotechnology Information provided raw reads for the start of the transcriptome analysis procedure (NCBI). NCBI acts as an international resource for the scientific community by providing open source databases and software tools for evaluating biological data and doing computational biology research. The reads for sap are SRR7145702 AND SRR7145708 of Project number: PRJNA47602.

#### **2. OC analysis and OC filtration is been done.**

Qc analysis with FASTQC to get idea about quality of data downloaded, NGSQC to quality filter the data . steps and commands on PUTTY

##### **a. Downloading Miniconda**

```
$ wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh
```

##### **b. Installation of Miniconda**

```
$ sh Miniconda3-latest-Linux-x86_64.sh
```

```
-----exit the terminal and relogin-----
```

##### **c. Create an environment for QC**

```
$ conda create --name qc
```

- d. Activate the environment

```
$ conda activate qc
```

- e. Install fastqc

```
$ conda install -c bioconda fastqc
```

- f. Create an output folder

```
$ mkdir qc_out
```

- g. Run Fastqc

```
$fastqc -o qc_out/ SRR1039508_1.fastq SRR1039508_2.fastq
```

### 3. Use of trinity

Assembly of all the samples to get transcriptome N50 analysis using statistics program of NGSQC Toolkit.

### 4. Redudancy and clustering removal of transcriptome

Confirmation with N50 analysis.

### 5. Alignment and validation is to be checked

Use of bowtie tool to check the alignment and validation of individual sample against indexed reference transcriptome.

### 6. Use of SAM tools and BED tools

By using these tools generation of bed files and collating the validation of sorted BAM files is to be done.

### 7. Use of Transimprove

All the files generated in step 6. Has to be used as input , and generation of coverage files is to be done.

### 8. Extraction of all the transcript Ids of all the sample of true records.

9. Extraction of sequences for transcript Ids for true records using FastQC Records.

10. Downloaded the relevant databases from Ensembl Plants.

11. Indexed databases using BLAST. Commands be like:

- a. Download and install miniconda and activate it.
- b. Create an env db search and activate it.
- c. Install blast from bioconda  
\$ conda install -c bioconda blast.
- d. Prepare the database from search  
\$ makeblastdb -dbtype nucl -in
- e. Use following output format for -outfmt "6 qacc sseqid sacc slen qlen qstart qend sstart send  
eval score qframe" with blastx query

12. BLASTn search against indexed databases with extracted sequences in step 10.

13. BLASTn search results were to be filtered for the criteria of query coverage greater than 60%.

14. Using BioDBnet and DAVID Annotation resource annotation of unique peptide Ids is to be done.

15. Differential expressed gene (DEG) Analysis is to be done using R platform to get

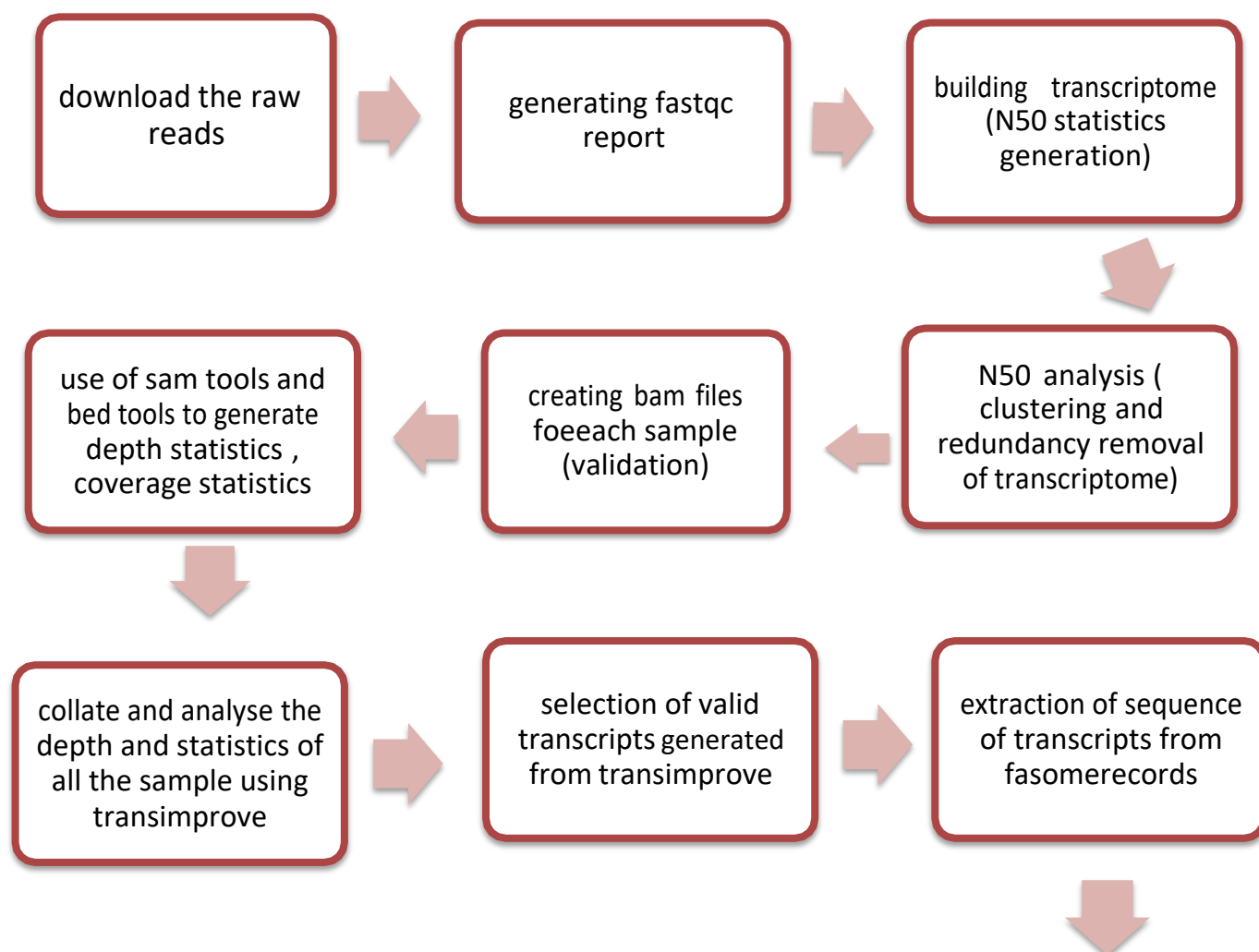
16. Merging of cov. Records for samples and filtration to be done to remove false records.

17. Differential expressed gene

- a. Input preparation : read counts for control and treated coverage output fromtransimprove
- b. Metadata of read counts : .cov. files

**Table 3.1 RAW DATA READS TO BE OBTAINED OF SAMPLE**

| <b><u>SAMPLE NAME</u></b> | <b><u>FORWARD RAW READ</u></b> | <b><u>REVERSE RAW READ</u></b> |
|---------------------------|--------------------------------|--------------------------------|
| <b><u>SRR7145702</u></b>  | SRR7145702_ 1.fastq.gz         | SRR7145702_ 2.fastq.gz         |
| <b><u>SRR7145708</u></b>  | SRR7145708 _1.fastq.gz         | SRR7145708_2.fastq.gz          |

**WORK PIPELINE OF DE NOVO TRANSCRIPTOME PROFILING**

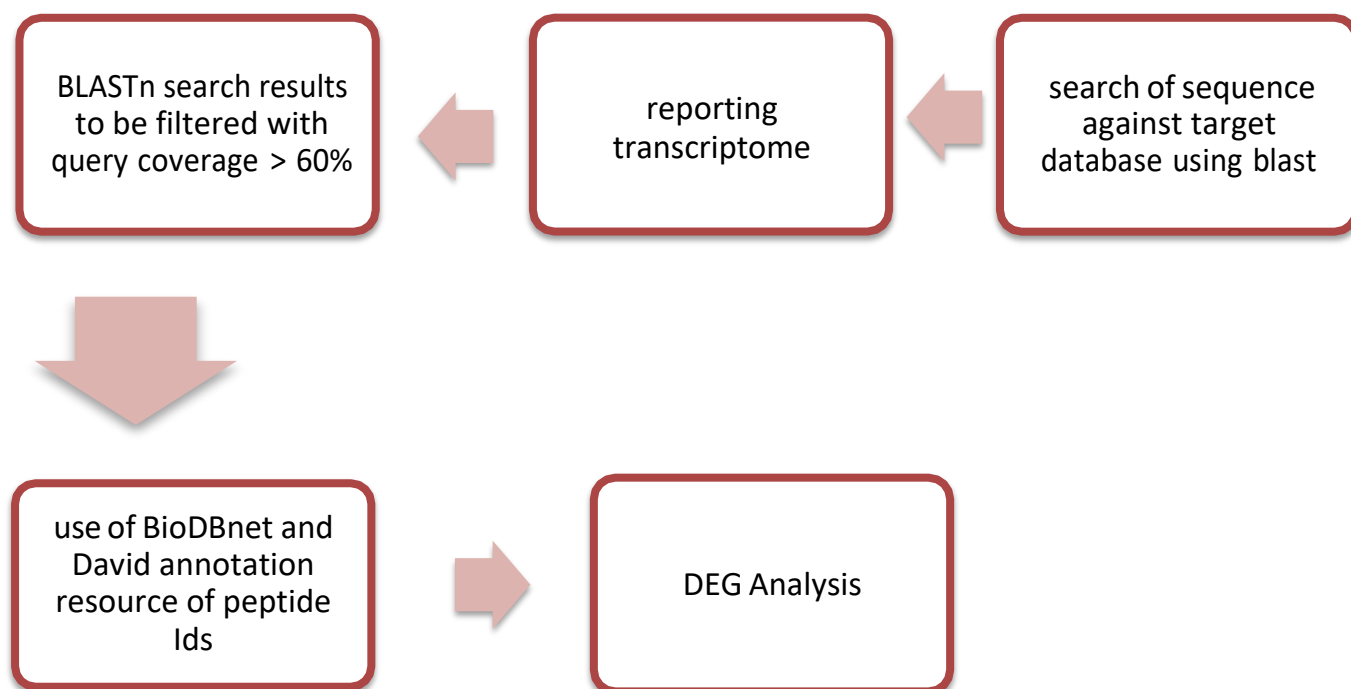


**Fig. 3.1:** Work Pipeline

## **BIOINFORMATICS TOOLS USED FOR THESE PROCESS**

### **1. FAST QUAITY CHECK (FASTQC)**

Fast QC has been accustomed to fathoming internal control over data returned from high throughput sequencing technology. It generates a preliminary data quality report. Prior to scanning the sequence for biological interpretation, you should always do some basic checks to ensure that the data information is correct and that there are no flaws or biases in the data that could affect how you use it. As the location of their analysis pipelines, many sequencers will create a QC report. (Andrews, 2010).

### **2. NGSQC Toolkit: Next Generation Sequencing Quality Control Toolkit**

NGS QC Toolkit is used to convert raw readings into filtered sequences. NGS QC Toolkit is a tool for controlling and filtering next-generation sequence data. It generates detailed reports in the

form of tables and graphs when the data is filtered, resulting in high-quality sequence data. It also includes a number of modules that allow for a variety of different pre-processing such as format conversion, trimming, and statistics. Low-quality reads are deleted after trimming, and reads containing primer/adaptor contamination are clipped according to the criterion. Finally, within the output folder, High Quality (HQ) readings and QC statistics are generated (Pateletal., 2012).

### 3. SEQUENCE ALIGNMENT / MAP (SAM Tools)

The SAM (Sequence Alignment/Map) format is a general format for storing nucleotide sequence alignments of huge sizes. SAM aspires to be a format that satisfies the following criteria: Is versatile enough to contain all alignment data generated by various alignment programmes; is simple enough to be generated by alignment programmes or converted from current alignment formats; Is little in terms of file size; Allows most alignment operations to be performed on a stream rather than loading the entire alignment into memory; allows the file to be indexed by genomic position to quickly get all reads aligning to a certain locus. SAM Tools offers a number of utilities for handling SAM alignments, including sorting, merging, indexing, and producing per-position alignments. (Li H A *et al.* , 2011)

### 4. BROWSER EXTENSIBLE DATA (BED TOOLS)

The BEDTools make it easy to compare big datasets of genomic characteristics in a quick and flexible way. The BEDtools software help with tasks like detecting feature overlaps and computing coverage in genomics. Its name comes from a historical purpose, however it can now process the most common feature file formats, such as BED, GFF, VCF, and SAM. (Giardine B, *et al.*, 2005)

## 5. **TRINITY**

Trinity, created at the Broad Institute and the Hebrew University of Jerusalem, is a unique method for de novo transcriptome reconstruction from RNA-seq data that is both efficient and reliable. Trinity is made up of three separate software modules: Inchworm, Chrysalis, and Butterfly, which are used in order to handle massive amounts of RNA-seq readings. Trinity divides the sequence data into multiple de Bruijn graphs, each representing the transcriptional complexity at a single gene or locus, and then processes each graph separately to extract full-length splicing isoforms and separate transcripts produced from paralogous genes.( Grabherr MGet *al.*, 2011)

## 6. **BLAST**

BLAST (basic local alignment search tool) is a bioinformatics technique and programme for comparing primarybiological sequence information, such as protein amino-acid sequences or DNA and/or RNA nucleotide sequences. A BLAST search allows a researcher to compare a query protein or nucleotide sequence to a library or database of sequences and find database sequences that are similar to the query sequence beyond a particular threshold( <https://blast.ncbi.nlm.nih.gov/Blast.cgi>)

## 7. **bioDBnet tool**

The biological DataBase Network is an application that connects a variety of biological databases such as Gene, UniProt, Ensembl, GO, Affy, and RefSeq. Data is downloaded from a variety of public sources to generate the databases. The Advanced Biomedical Computing Center formats and maintains them in a relational framework (ABCC). Notes on what bioDBnet currently has to offer :

- a. All conversions from one database identifier to another are handled by db2db.You can use dbWalk to walk through your own bioDBnet path.
- b. For a provided identifier, dbReport generates a report with all available information.
- c. dbFind assesses the kind of identifiers and changes them all to the database identifier type of

yourchoice.

- d. dbOrtho translates one organism's identifiers into homolog identifiers for another organism.
- e. For various biological identifiers, dbAnnot supports numerous sorts of annotations.
- f. dbOrg makes organism-wide transformations available as downloaded files.
- g. In addition to the core capabilities, bioDBnet provides analysis tools.
- h. orgTaxon is a tool for determining an organism's taxon ID.
- i. To see the parents of any go accession, use goTree.
- j. chrView is a chromosomal data visualisation programme.
- k. bioTaxon is a tool for locating genes.

## 8. David annotation tool

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) provides investigators with a complete collection of functional annotation tools to help them comprehend the biological significance of enormous lists of genes. The DAVID Knowledgebase, which is based on the DAVID Gene concept and pulls together numerous sources of functional annotations, powers these products. DAVID tools can do the following for any given gene list:

- a. Identify biological topics that have been enriched, particularly GO keywords.
- b. Discover functionally similar gene groupings that have been enriched.
- c. Duplicate annotation terms should be grouped together.
- d. Genes can be seen on BioCarta and KEGG pathway maps.
- e. On a 2-D perspective, display associated many-genes-to-many-terms.
- f. Look for functionally similar genes that aren't on the list.
- g. Compile a list of proteins that interact.
- h. Gene names can be looked up in bulk.
- i. Connect the dots between genes and diseases.
- j. The functional domains and motifs of proteins are highlighted.

k. Redirect to literatures that are connected.

l. Gene IDs can be converted.

**Table 3.2 : URLS OF DIFFERENT TOOLS , SOFTWARE AND DATABASE**

| <b>TOOLS / SOFTWARE /<br/>DATABASE</b> | <b>URLS</b>                                                                                                                         |
|----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| NCBI                                   | <a href="https://www.ncbi.nlm.nih.gov/">https://www.ncbi.nlm.nih.gov/</a>                                                           |
| NGSQCToolkit                           | <a href="http://www.nipgr.res.in/ngsqctoolkit.html">http://www.nipgr.res.in/ngsqctoolkit.html</a>                                   |
| TRINITY                                | <a href="https://github.com/trinityrnaseq/trinityrnaseq/">https://github.com/trinityrnaseq/trinityrnaseq/</a>                       |
| FastQC                                 | <a href="https://www.bioinformatics.babraham.ac.uk/projects/fastqc/">https://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a> |
| SAM TOOLS                              | <a href="http://samtools.sourceforge.net/">http://samtools.sourceforge.net/</a>                                                     |
| BED TOOLS                              | <a href="https://bedtools.readthedocs.io/">https://bedtools.readthedocs.io/</a>                                                     |
| DAVID ANNOTATION TOOL                  | <a href="https://david.ncifcrf.gov/">https://david.ncifcrf.gov/</a>                                                                 |
| bioDBnet tool                          | <a href="https://biodbnet-abcc.ncifcrf.gov/">https://biodbnet-abcc.ncifcrf.gov/</a>                                                 |

## **CHAPTER 4: RESULTS**

The results of this whole “**De novo transcriptome profiling of plant Papaya plant sap in drought stress condition**” which includes quality control of data, validation of transcriptome and annotation is cited here.

### **4.1 SAMPLE SELECTION :**

The Sample of Papaya plant sap selected from NCBI database under project number PRJNA470602.

**Table 4.1 RAW READS FOR ANALYSIS :**

| <b><u>SAMPLE NAME</u></b> | <b><u>FORWARD RAW<br/>READ</u></b> | <b><u>REVERSE RAW<br/>READ</u></b> |
|---------------------------|------------------------------------|------------------------------------|
| <b><u>SRR7145702</u></b>  | SRR7145702_1.fastq.gz              | SRR7145702_2.fastq.gz              |
| <b><u>SRR7145708</u></b>  | SRR7145708_1.fastq.gz              | SRR7145708_2.fastq.gz              |

## 4.2 FAST QUALITY CHECK REPORT

Table 4.2: BASIC STATISTICS OF READS AFTER FASTQC

| Measure         | SRR7145702                |                           | SRR7145708                |                           |
|-----------------|---------------------------|---------------------------|---------------------------|---------------------------|
| File name       | SRR7145702_1<br>.fastq.gz | SRR7145702_2<br>.fastq.gz | SRR7145708_1<br>.fastq.gz | SRR7145708_2<br>.fastq.gz |
| File type       | ConventionalBase<br>calls | ConventionalBase<br>calls | ConventionalBase<br>calls | ConventionalBase<br>calls |
| Encoding        | Sanger/ Illumina1.9       | Sanger/ Illumina1.9       | Sanger/ Illumina1.9       | Sanger/ Illumina<br>1.9   |
| Total Sequences | 39861192                  | 39861192                  | 37912534                  | 37912534                  |
| Sequence length | 35 - 76                   | 35 - 76                   | 35 - 76                   | 35 - 76                   |
| %GC             | 44                        | 44                        | 43                        | 43                        |

## 4.2.1 PER BASE SEQUENCE QUALITY FOR BOTH THE SAMPLE READS (FORWARD AND REVERSE)

### 1. SRR7145702\_1.fastq.gz

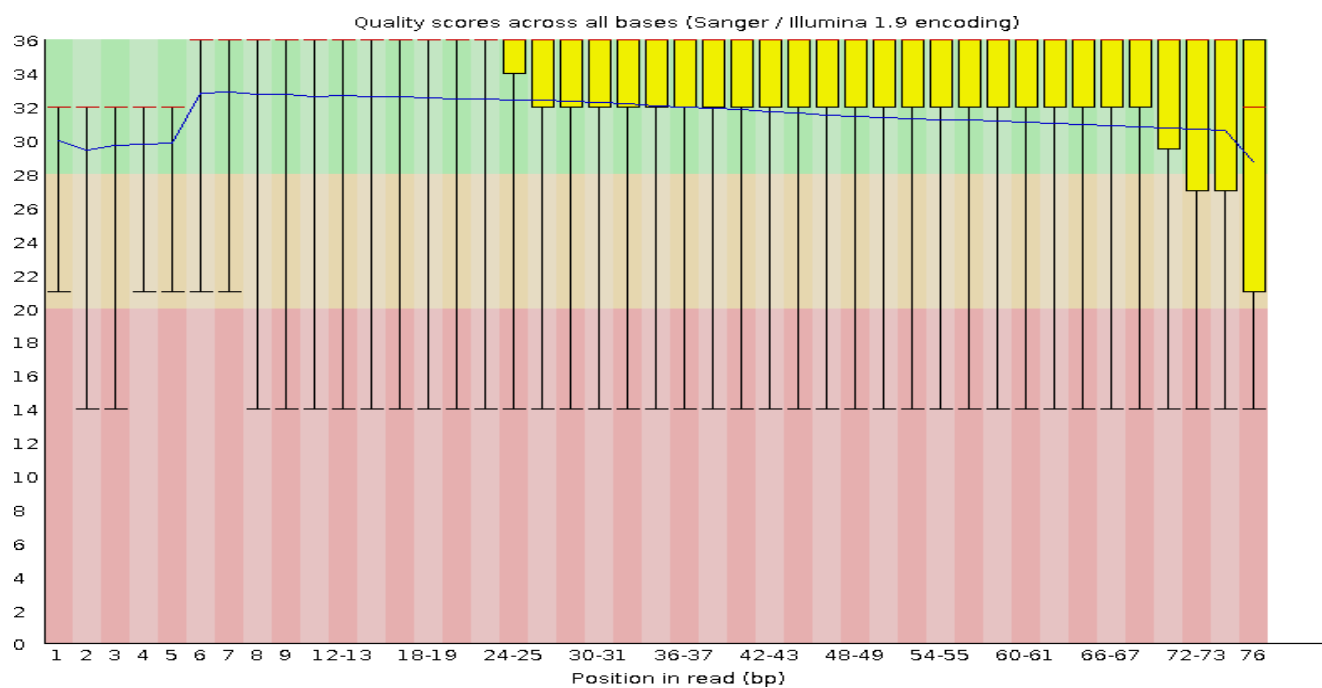


**Fig. 4.1.1:** per base sequence quality for SRR7145702\_1.fastq.gz

### 2. SRR7145702\_2.fastq.gz

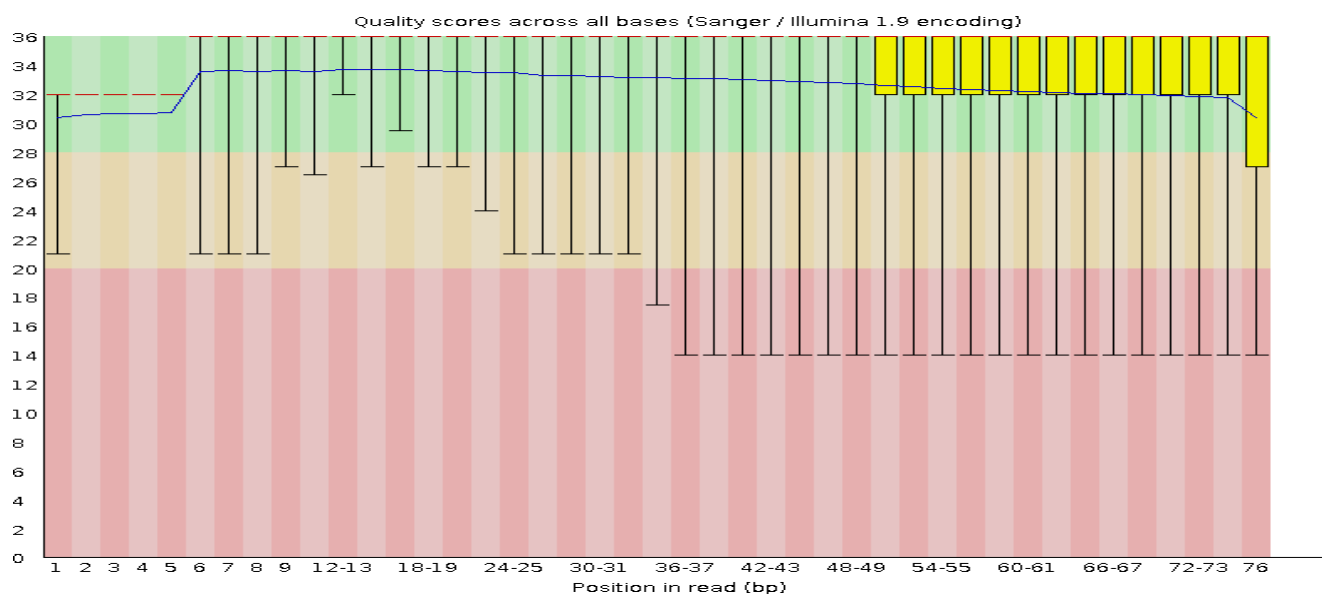


**Fig. 4.1.2:** per base sequence quality for SRR7145702\_2.fastq.gz

### 3. SRR7145708\_1.fastq.gz

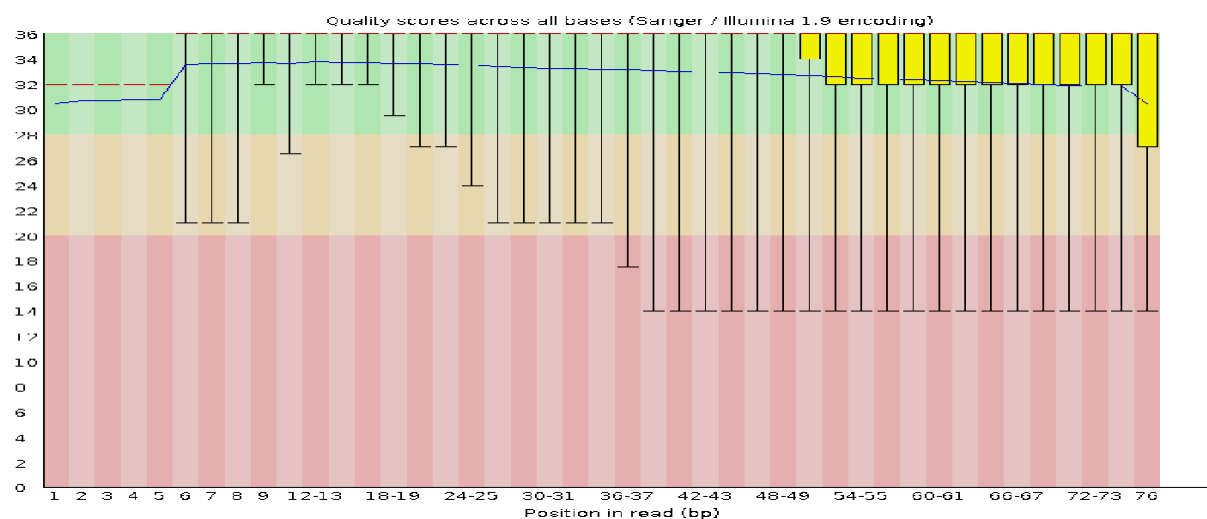


**Fig. 4.2.1:** per base sequence quality for SRR7145708\_1.fastq.gz

### 4. SRR7145708\_2.fastq.gz

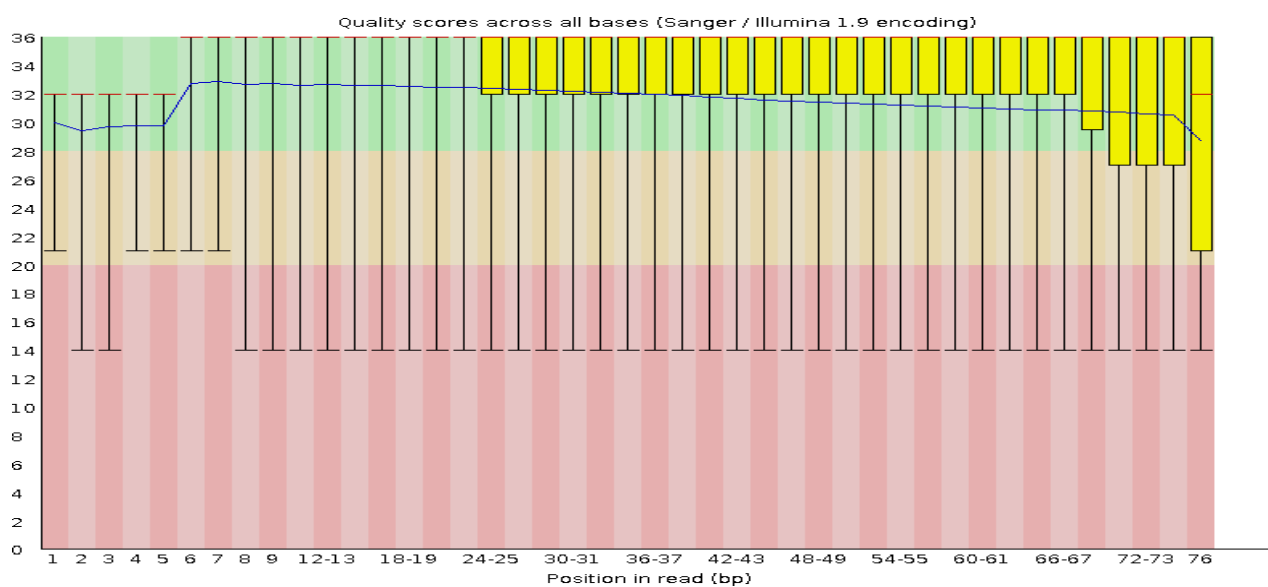


**Fig.4.2.2 :** per base sequence quality for SRR7145708\_2.fastq.gz

## 2.2 PER SEQUENCE QUALITY SCORE

### A. SRR7145702 ( FORWARD AND REVERSE READ)

For forward read

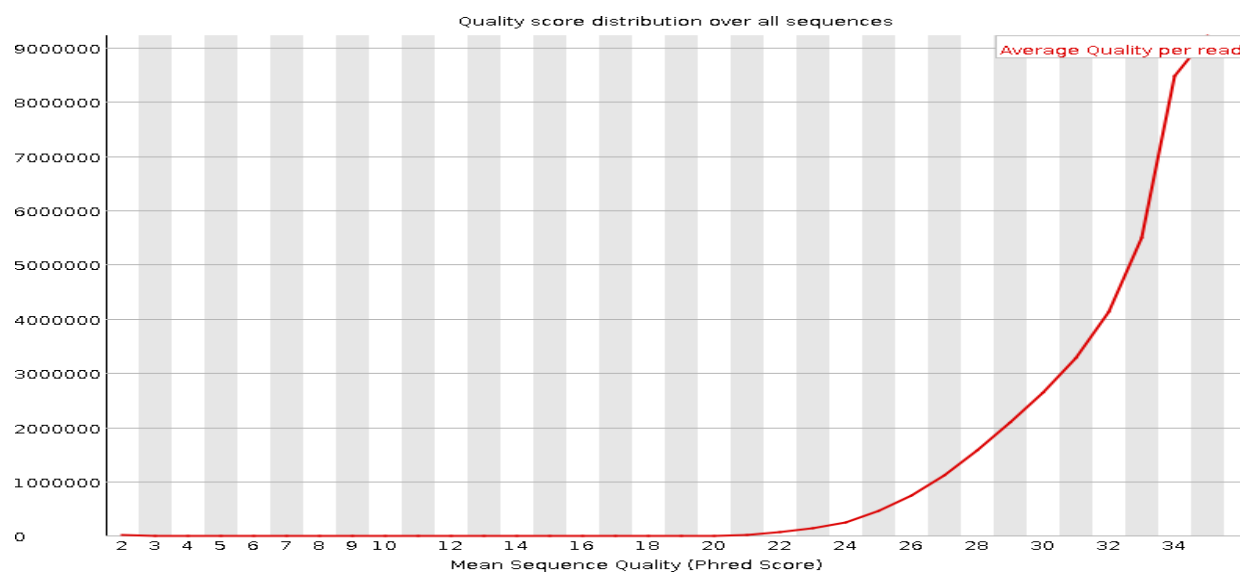


**Fig. 4.3.1:** per sequence quality score for SRR7145702 ( forward read)

For reverse read

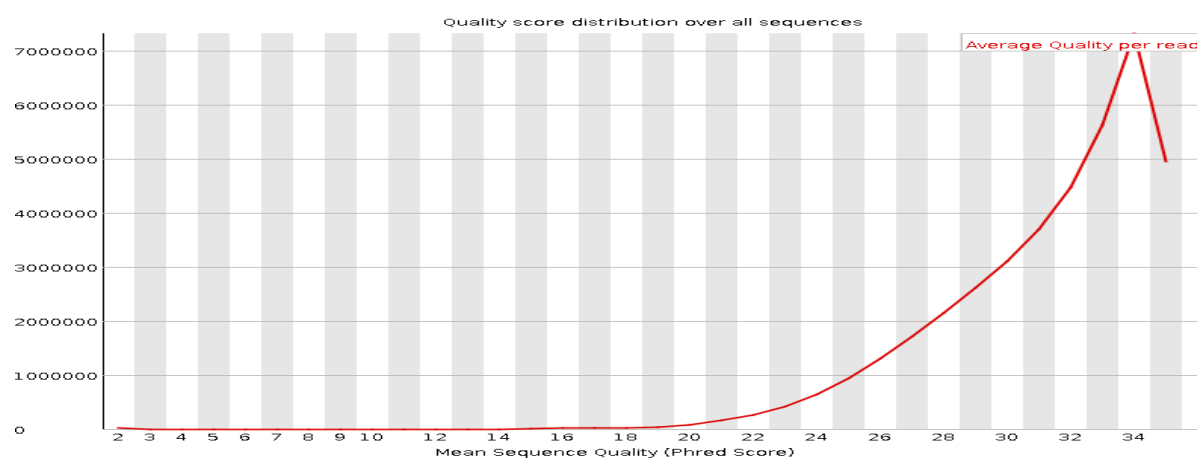


**Fig 4.3.2:** per sequence quality score for SRR7145702 ( backward read)

## **B) FOR SRR7145708 ( FORWARD AND BACKWARD READS)**

### **Forward read**

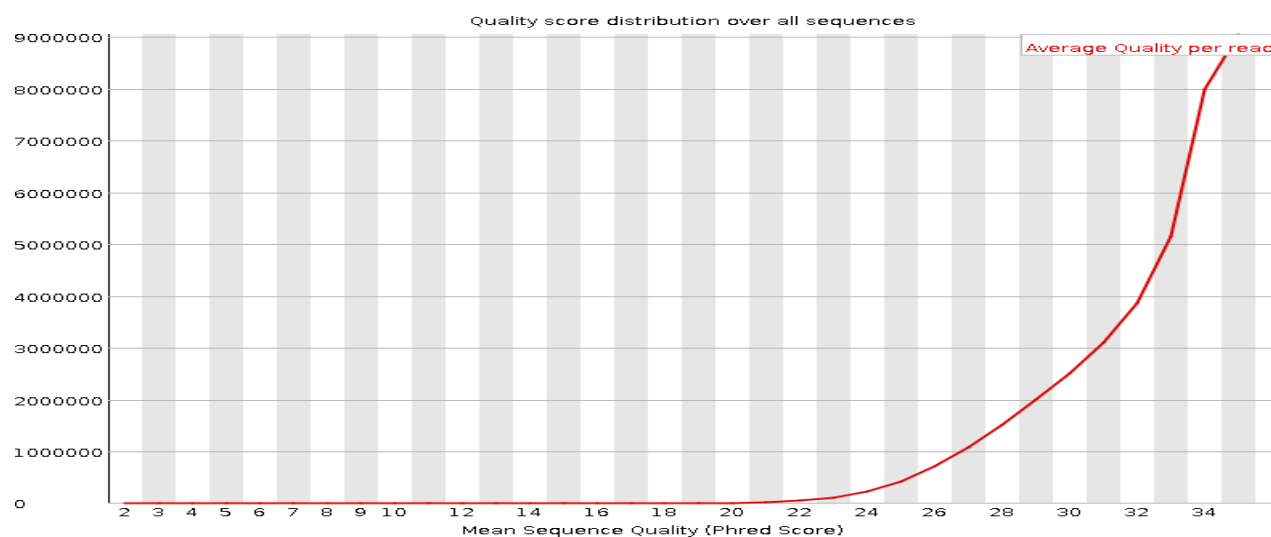


**Fig .4.4.1:** per sequence quality score for SRR7145708 ( forward read)

### **Backward read**

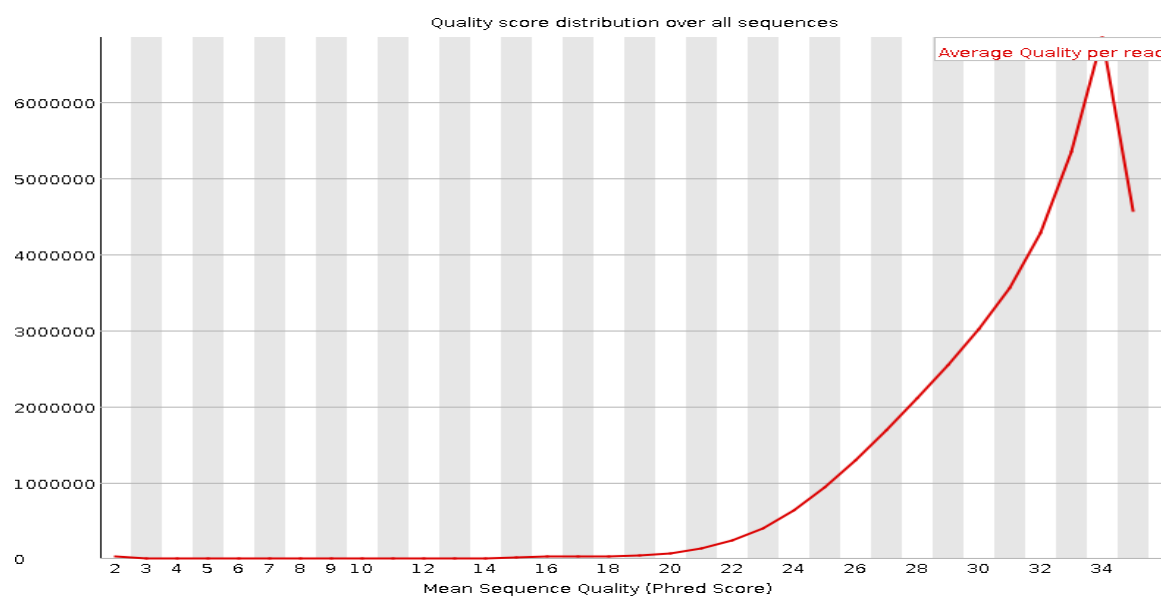


**Fig .4.4.2:** per sequence quality score for SRR7145708 ( backward read)

## **NGSCQC REPORT**

Both the sample SRR7145702 and SRR7145708 have shown same cutoff quality score of 20 and cutoff read length forHQ 70%.

**Table 4.3: DEPICTION OF SEQUENCING DATA**

| <b>SAMPLE</b>                     | <b>SRR7145702</b>                        | <b>SRR7145708</b>                        |
|-----------------------------------|------------------------------------------|------------------------------------------|
| <b>Library File</b>               | Paired End                               | Paired End                               |
| <b>Input Fil</b>                  | SRR7145702_1.fastq<br>SRR7145702_2.fastq | SRR7145708_1.fastq<br>SRR7145708_2.fastq |
| <b>Primer/Adaptor Library</b>     | Paired End DNA<br>Library                | Paired End DNA Library                   |
| <b>Cut-off read length for HQ</b> | 70%                                      | 70%                                      |
| <b>Cut-off Quality score</b>      | 20                                       | 20                                       |
| <b>Only Statistics</b>            | Off                                      | Off                                      |
| <b>Number of CPUs</b>             | 3                                        | 3                                        |

### **Ahead of using NGSCOC TOOL KIT :**

The reverse reads of both the sap sample SRR7145702 and SRR7145708 show higher number of non ATGC bases which shows higher polymorphism which is 1593926 and 1525591 respectively. Number of primer contaminated for forward read of SRR7145708 is 0 comparing all the reads of both samples. Other details are summarized in table below:

**Table 4.4: Data before using NGSQC TOOLKIT**

| Sample                                        | SRR7145702             |                        | SRR7145708             |                        |
|-----------------------------------------------|------------------------|------------------------|------------------------|------------------------|
| <b>File Name</b>                              | SRR7145702_1.<br>fastq | SRR7145702_2.fa<br>stq | SRR7145708_1.fa<br>stq | SRR7145708_2.fa<br>stq |
| <b>Percentage of HQ Reads</b>                 | 86.66%                 | 86.66%                 | 86.54%                 | 86.54%                 |
| <b>Percentage of HQ BasesIn HQ Reads</b>      | 93.05%                 | 90. 27%                | 93.12%                 | 90.10%                 |
| <b>Average Read length</b>                    | 75. 50                 | 75. 50                 | 75.49                  | 75.49                  |
| <b>Percentage of Reads withNon-ATGC Bases</b> | 0.03%                  | 0.13%                  | 0.03%                  | 0.13%                  |
| <b>Non-ATGC Bases</b>                         | 22085                  | 1593926                | 21119                  | 1525591                |

### Afterwards using NGSOC TOOL KIT

Even after filtration and trimming te reverse reads of both the samples shows higher number of non ATGC bases which shows higher polymorphism but number of non ATGC bases have been decreased from unfiltered reads. other details comprises of :

**Table 4.5: Data after using NGSQC TOOLKIT**

| <b>Sample</b>                                     | <b>SRR7145702</b>                   |                                     | <b>SRR7145708</b>                   |                                     |
|---------------------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| <b>File Name</b>                                  | SRR7145702_1<br>.<br>fastq_filtered | SRR7145702_2<br>.<br>fastq_filtered | SRR7145708_1<br>.<br>fastq_filtered | SRR7145708_2<br>.<br>fastq_filtered |
| <b>Total Number of Reads</b>                      | 34514063                            | 34514063                            | 32782812                            | 32782812                            |
| <b>Total Number of Reads with Non- ATGC Bases</b> | 6579                                | 18933                               | 6362                                | 18101                               |
| <b>Total Number ofHQ Bases</b>                    | 2424782943                          | 2352361308                          | 2304646086                          | 2229793153                          |
| <b>Non-ATGC Bases</b>                             | 11674                               | 84565                               | 11337                               | s79739                              |

## QUALITY CONTROL DATA FROM NGSQC TOOL KIT :

### Per base average quality score:

#### SRR7145702

The average quality score is the likelihood that the sequencer would call the provided base erroneously. The average quality score of raw reads and filtered reads drops as base position increases, limiting the duration of high-quality reads. In forward read , from base position 1 to 5 the average quality score remains 30 then increasing sharply and gradually decreasing till 73 base position and then sharp turn of 20 quality score for base position 76 . likewise for reverse reads the quality score remains constant 30 till base position 5 and increasing sharply till 34 at position 7 then decreasing gradually to 19 at base position 76, these trends are for both before and after filtration.

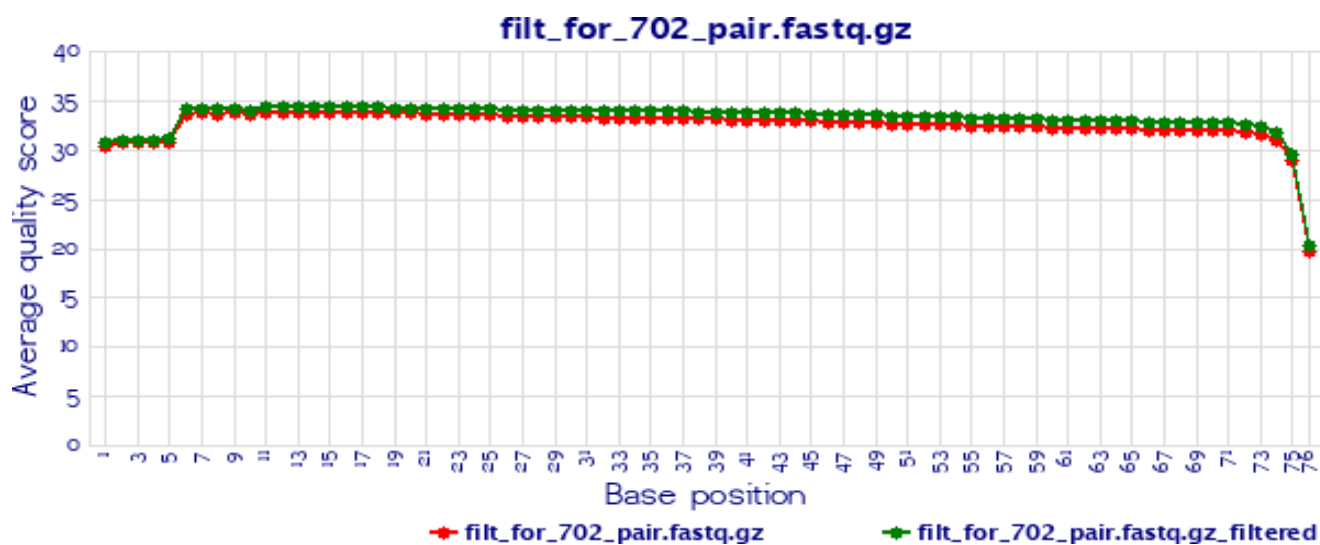


**Fig .4.5.1:** per base average quality score for SRR7145702 ( forward read)

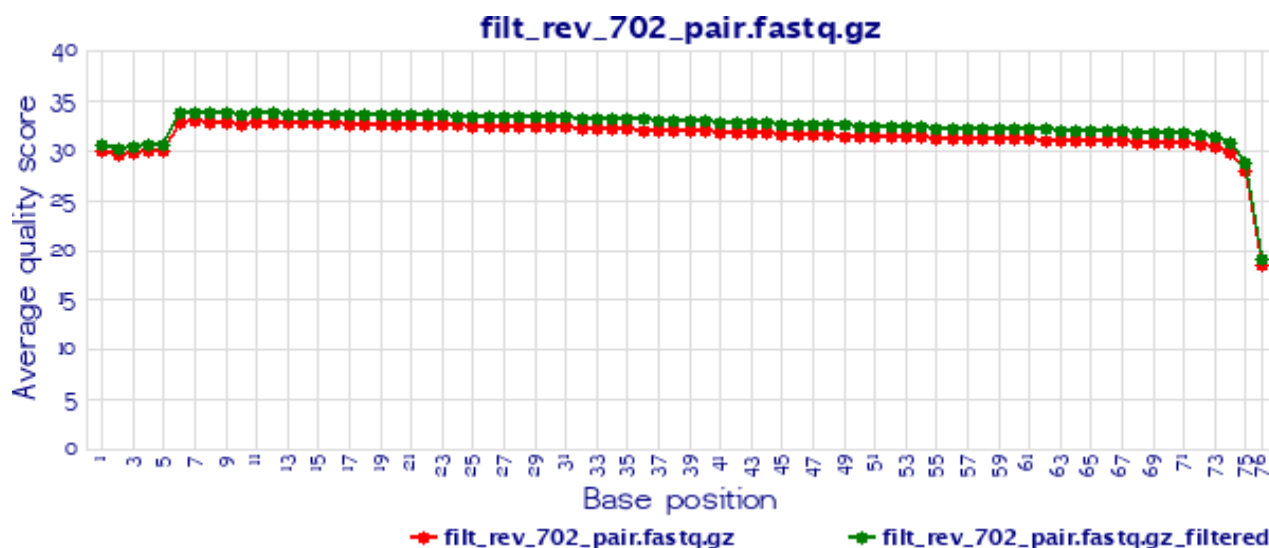


**Fig. 4.5.2** :per base average quality score for SRR7145702 ( backward read)

### SRR7145708

For sample SRR7145708 the quality score for forward read is showing same trend as sample 702 like at at base position 1 to 5 quality score 31 and then increasing to 34 sharply at base position 7 then gradually decreasing after being steady to 20 at base position 76. And in reverse read from 1 to 5 the quality score being 30 increasing to 34 atbase position 7 then being constant and slightly changing till 71 base postion then sharp decrease to 19 for base position 76 and these trends are for before and after filtration.

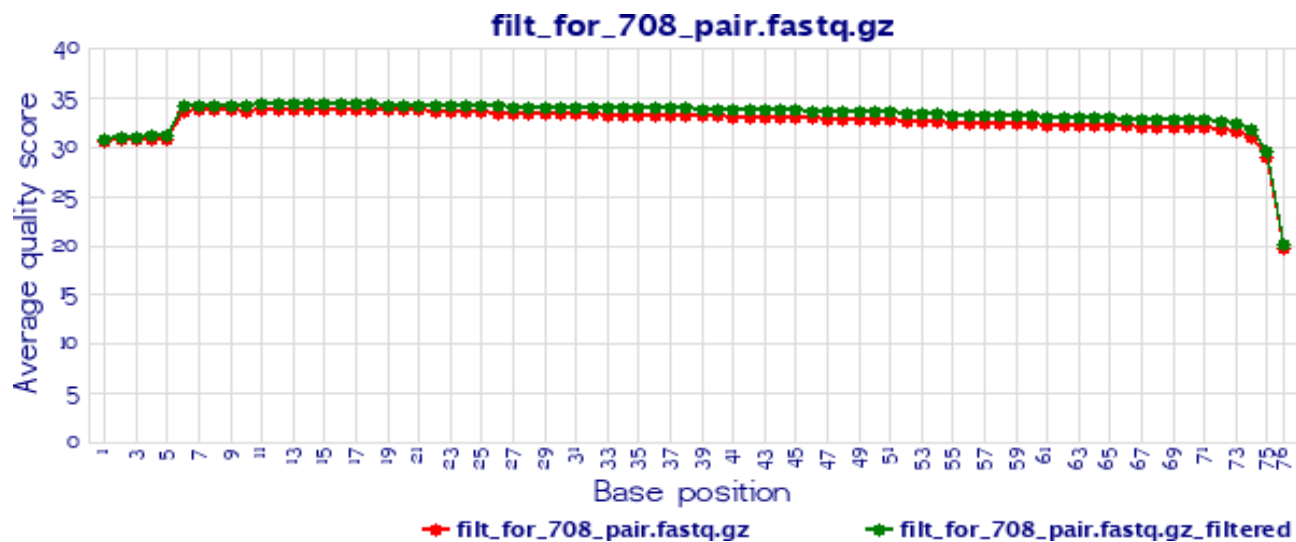


**Fig .4.6.2:** per base average quality score for SRR7145708 ( forward read)

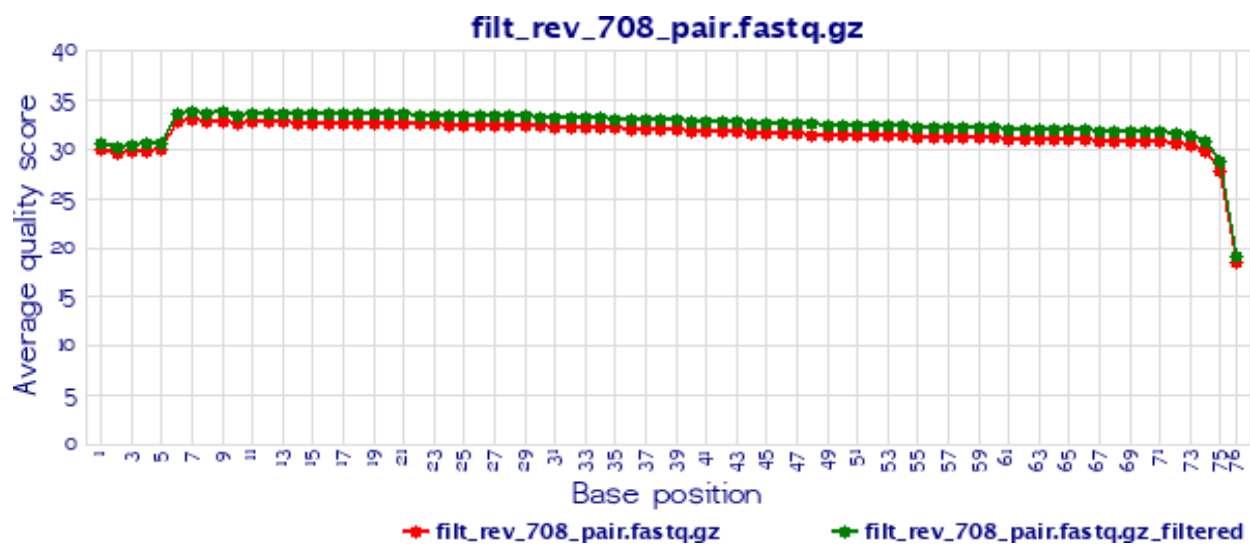


**Fig. 4.6.2** per base average quality score for SRR7145708 ( backward read)

## READ COUNT % FOR DIFFERENT SAMPLES AT DIFFERENT QUALITY RANGE

### SRR7145708

For the base position 1 for quality score 31 to 40 forward read is 90 before qc and after qc for same condition qualityscore be like slightly 91. For the base position 1 the quality score 31 to 40 for reverse read for sample is 88 before qcand 91 after qc for same reverse read . other range data is depicted in figure 4.7.1

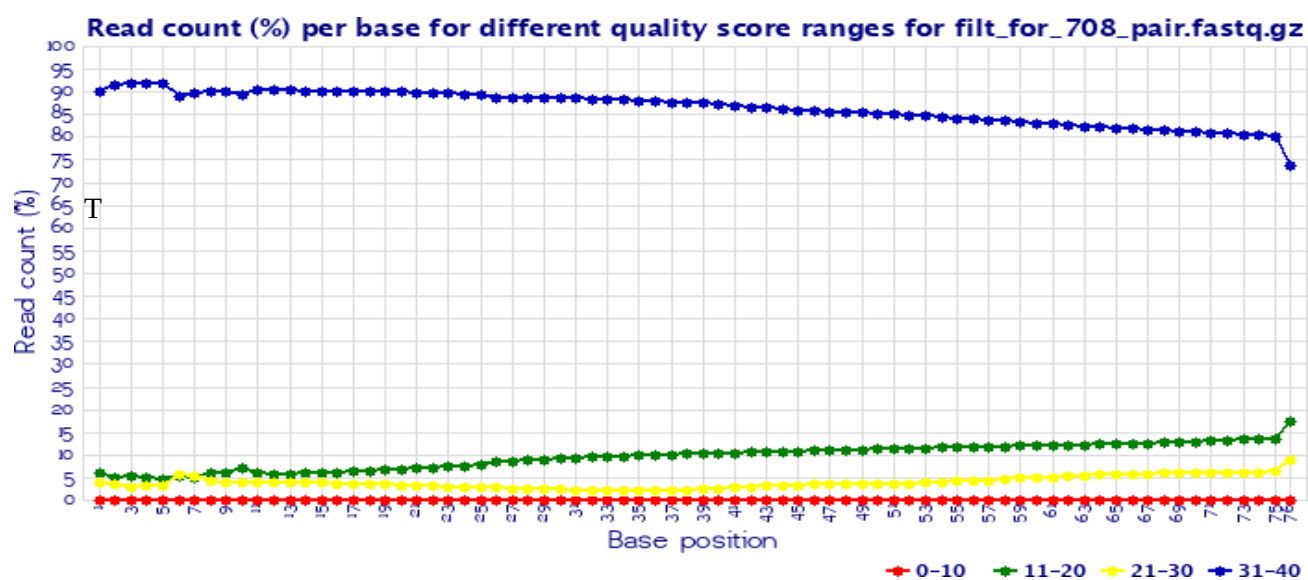
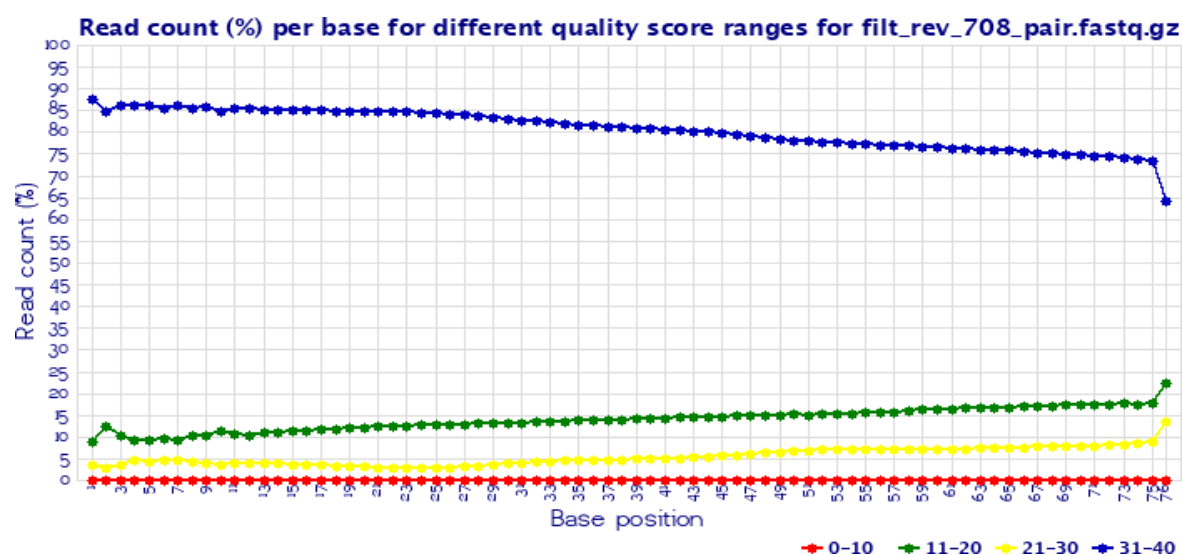
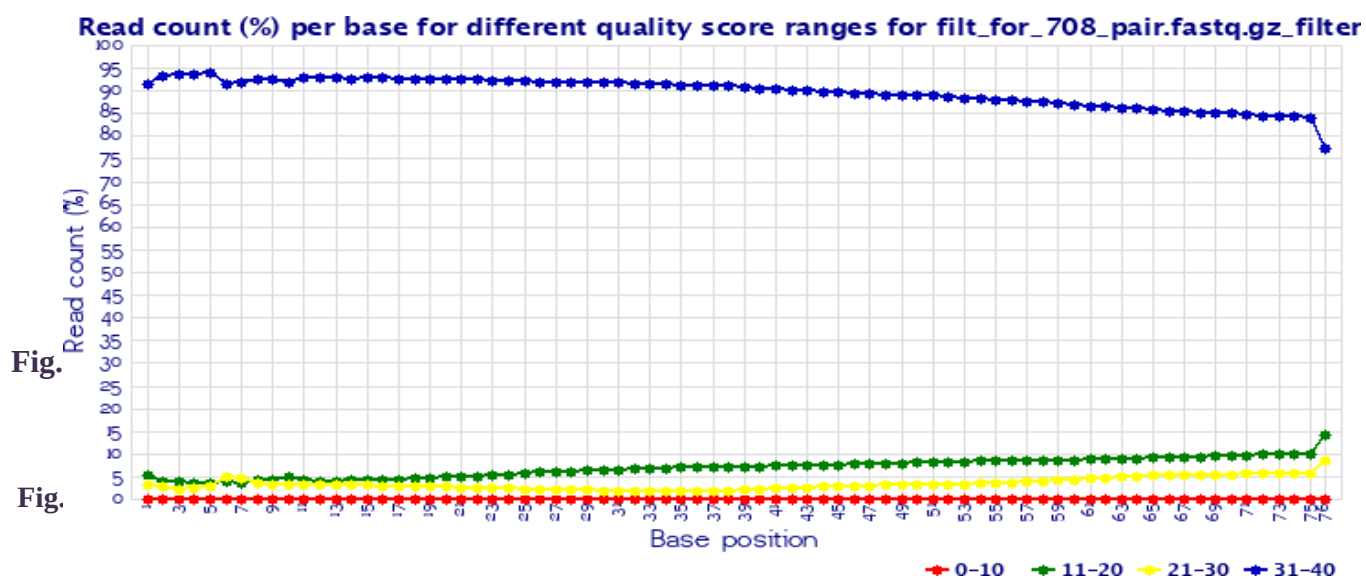


**Fig.4.7.1:** Read count % per base for SRR7145708 Forward read before and after NGSQC toolkit



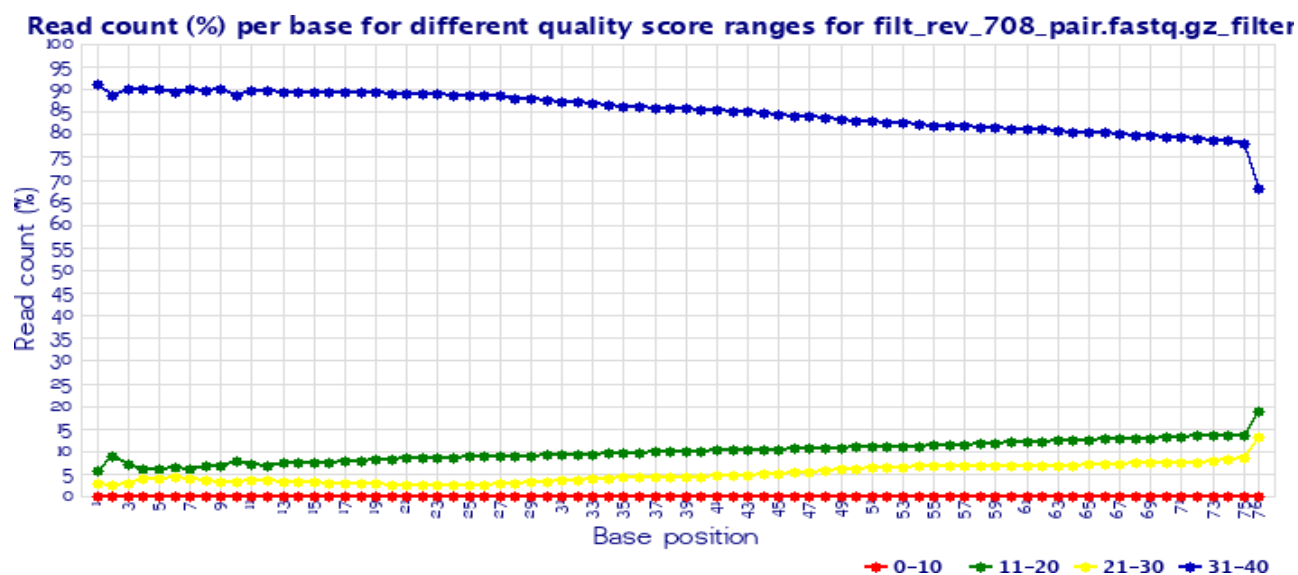


Fig.4.7.4 : Read count % per base for SRR7145708 Backward read before and after NGSQC

### SRR7145702

For the base position 1 for quality score 31 to 40 forward read is 90 before qc and after qc for same condition qualityscore be like slightly 91. For the base position 1 the quality score 31 to 40 for reverse read for sample is 87 before qc and 92 after qc for same reverse read . other range data is depicted in figure 4.8.1

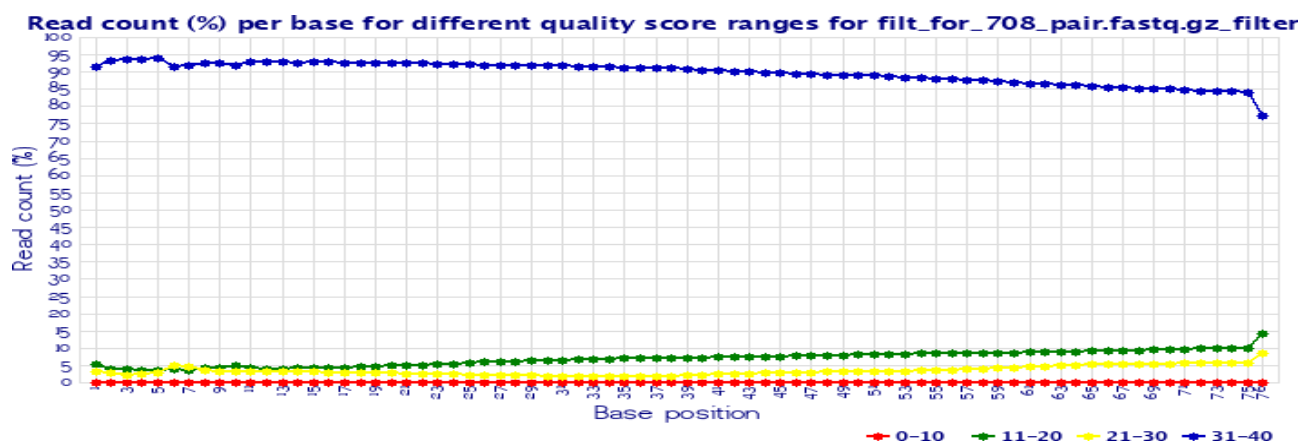


Fig 4.8.1 : Read count % per base for SRR7145702 Forward read before and after NGSQC

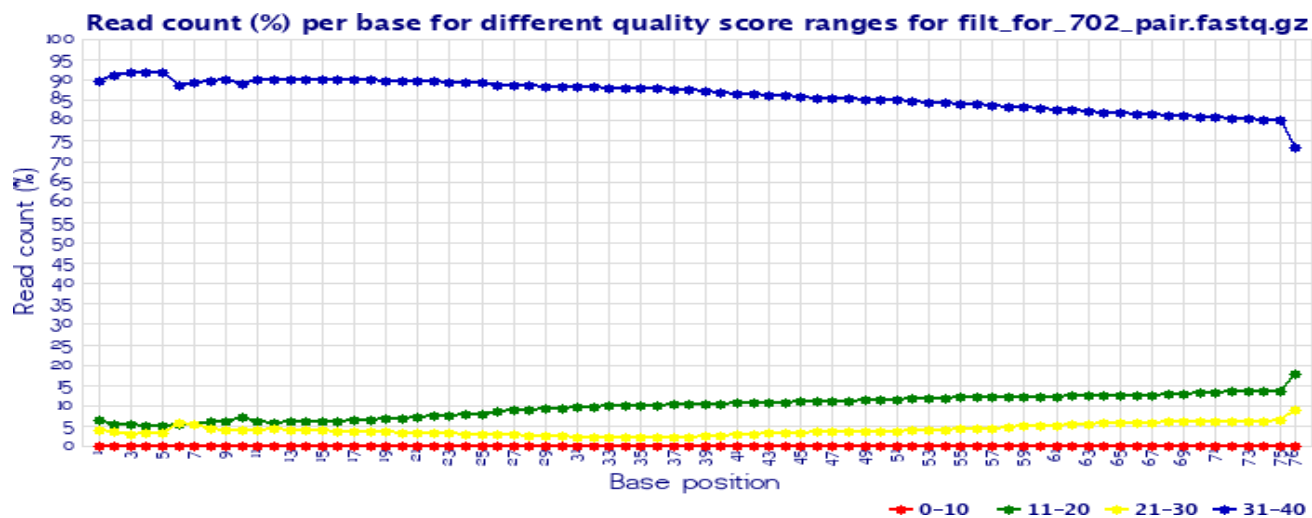


Fig.:4.8.2 : Read count % per base for SRR7145702 Forward read before and after NGSQC

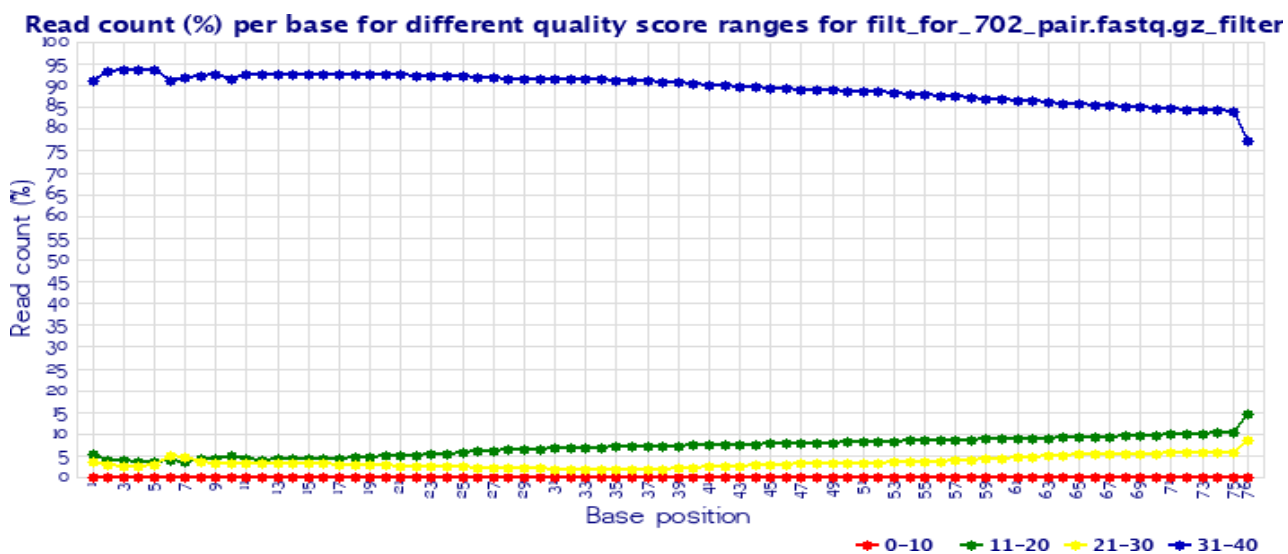


Fig. 4.8.3 Read count % per base for SRR7145702 Backward read before and after NGSQC

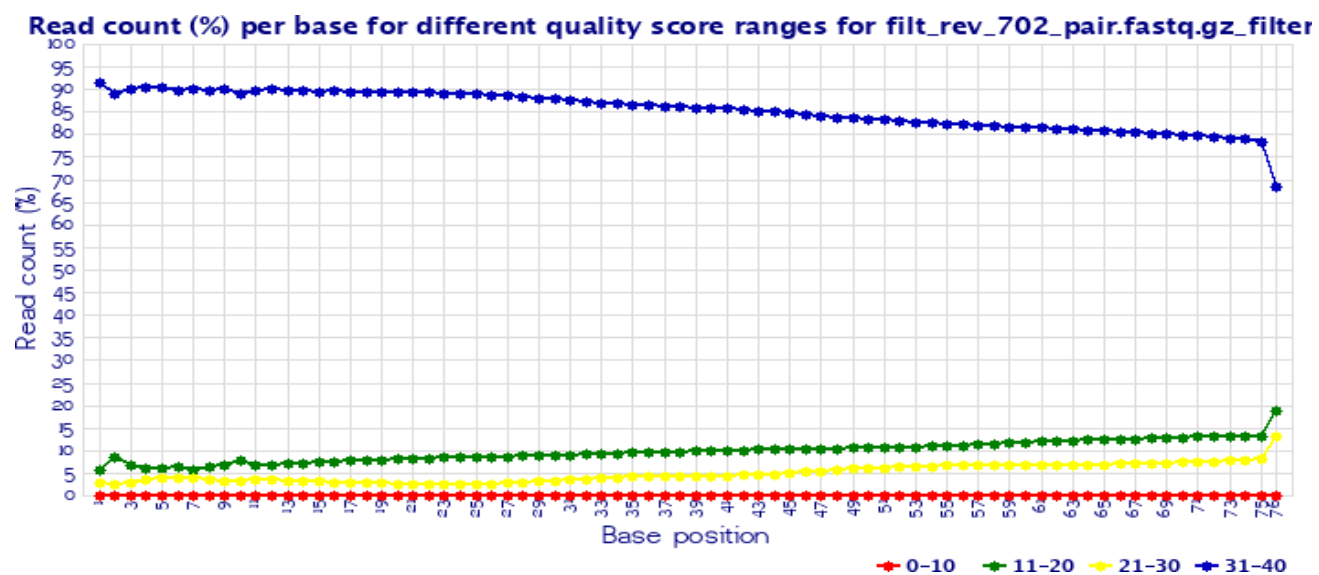


Fig .4.8.4 : Read count % per base for SRR7145702 Backward read before and after NGSQC

### 4.3.2.1 BASE COMPOSITION

#### SRR7145702

After filtration or trimming base composition will alter in both forward and reverse readings. In raw reads, Adenine makes up 27.95 percent, Thymine makes up 27.97 percent, Guanine makes up 21.59 percent, Cytosine makes up 22.49 percent, and non-ATGC makes up 0.00 percent, but these percentages change in filtered reads, which are represented by a green bar in the graph for forward reads but for reverse reads the adenine is 28.00 percent, Thymine is 27.81 percent, guanine is 21.61 percent, cytosine is 22.53 percent and non ATGC is 0.05 percent in raw reverse reads and for filtered data the percentage change is depicted by green bar. Fig 4.9.1 and 4.9.2

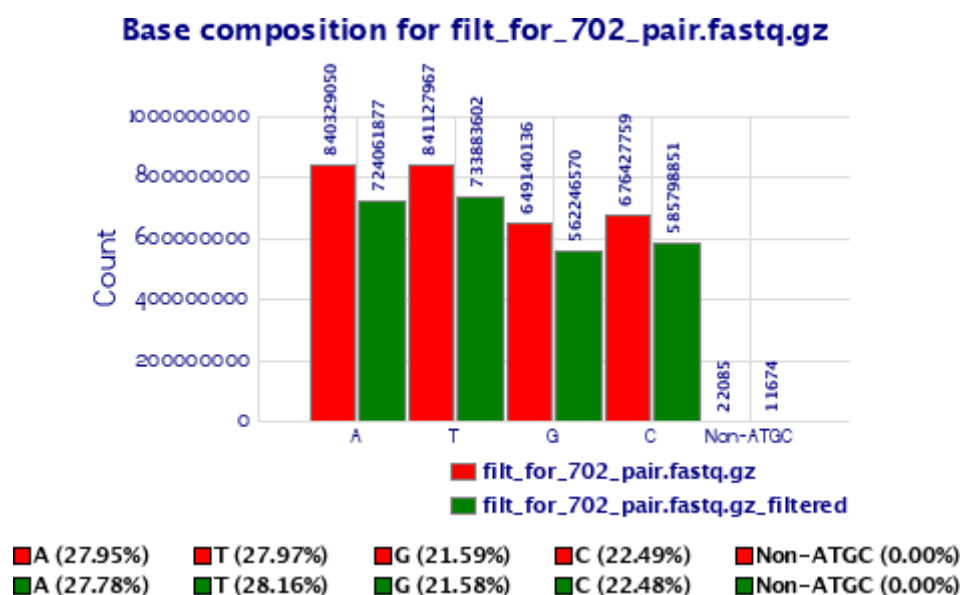


**Fig .4.9.1:** Base composition % SRR7145702 Forward Read

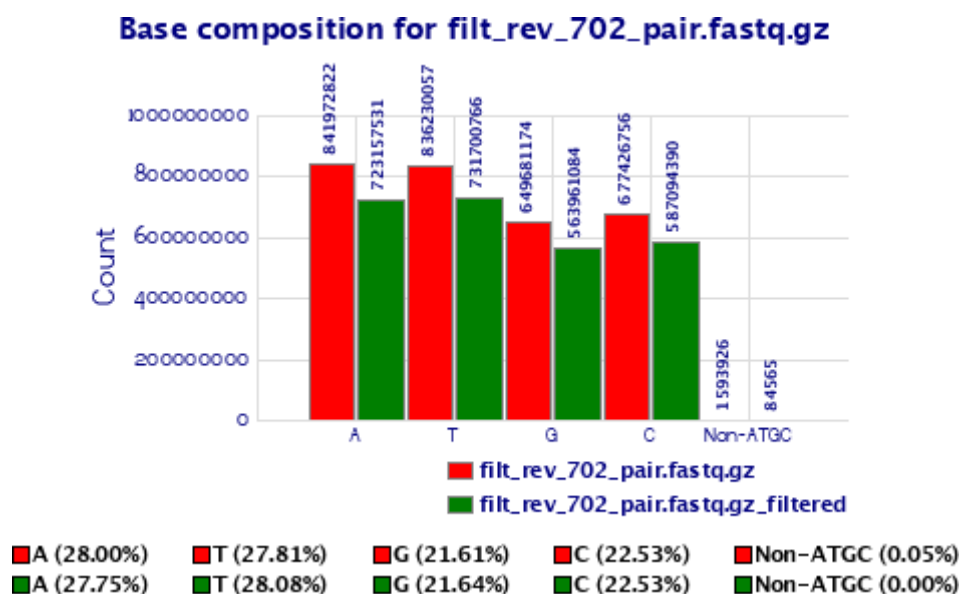
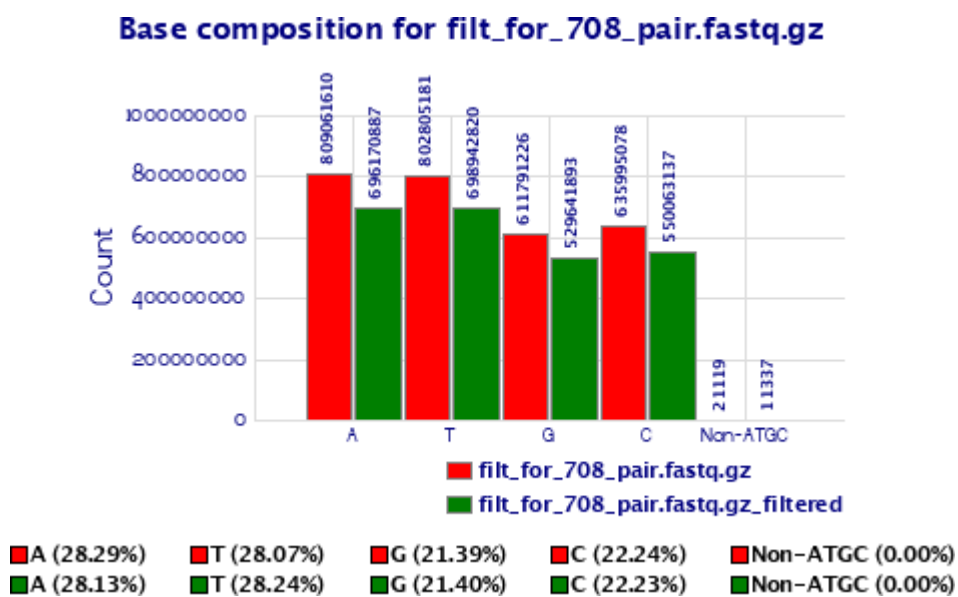


**Fig .4.9.2:** Base composition % SRR7145702 Backward Read

### SRR7145708

After filtration or trimming base composition will alter in both forward and reverse readings. In raw reads, Adenine makes up 28.30 percent, Thymine makes up 27.92 percent, Guanine makes up 21.47 percent, Cytosine makes up 22.26 percent, and non-ATGC makes up 0.00 percent, but these percentages change in filtered reads, which are represented by a green bar in the graph for forward reads but for reverse reads the adenine is 28.30 percent , Thymine is 27.92 percent , guanine is 21.47 percent , cytosine is 22.26 percent and non ATGC is 0.05 percent in raw reverse reads and for filtered data the percentage change is

depicted by green bar.



. Fig .4.10.1: Base composition % SRR7145708 Forward Read

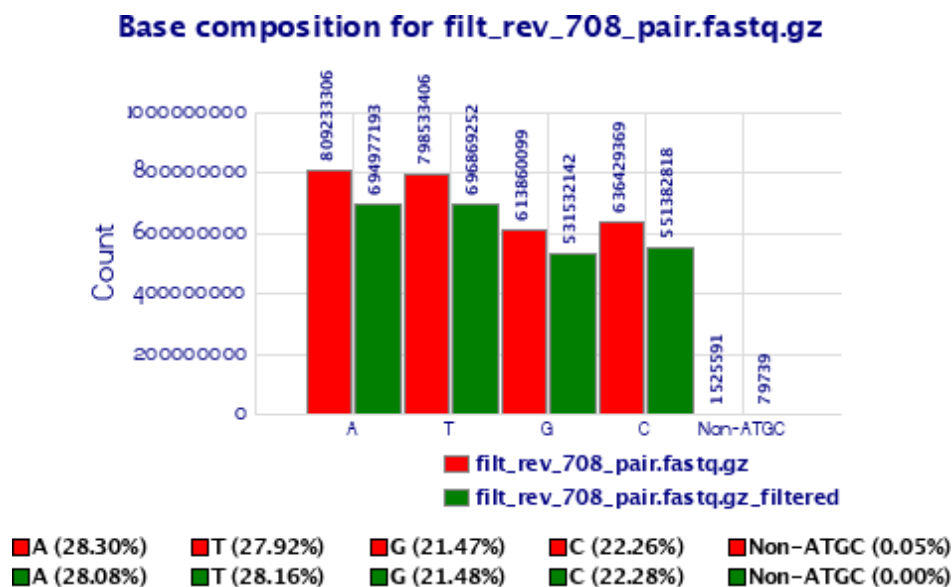


**Fig .4.10.2:** Base composition % SRR7145708 Backwar

## GC CONTENT

### SRR7145702

Graph shows the GC content distribution in raw reads and filtered reads. For sample 702, GC content distribution is almost the same in filtered reads that is 40 – 45 % in both forward and reverse reads at slightly same amount of reads. Its there peak of gccontent at that position . other conyents are deoicted in graph below. Fig 4.11.1 and 4.11.2.

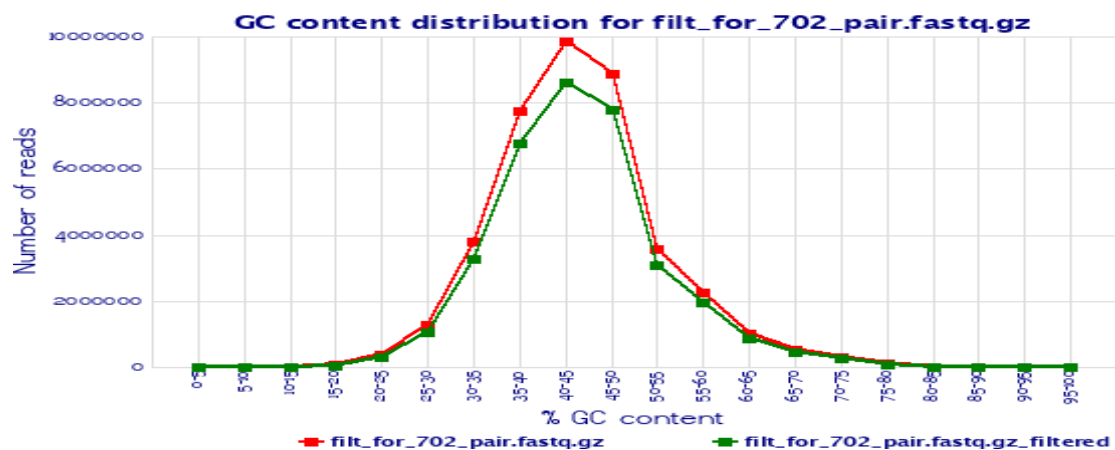


Fig 4.11.1: % GC content SRR7145702 (FORWARD READ )

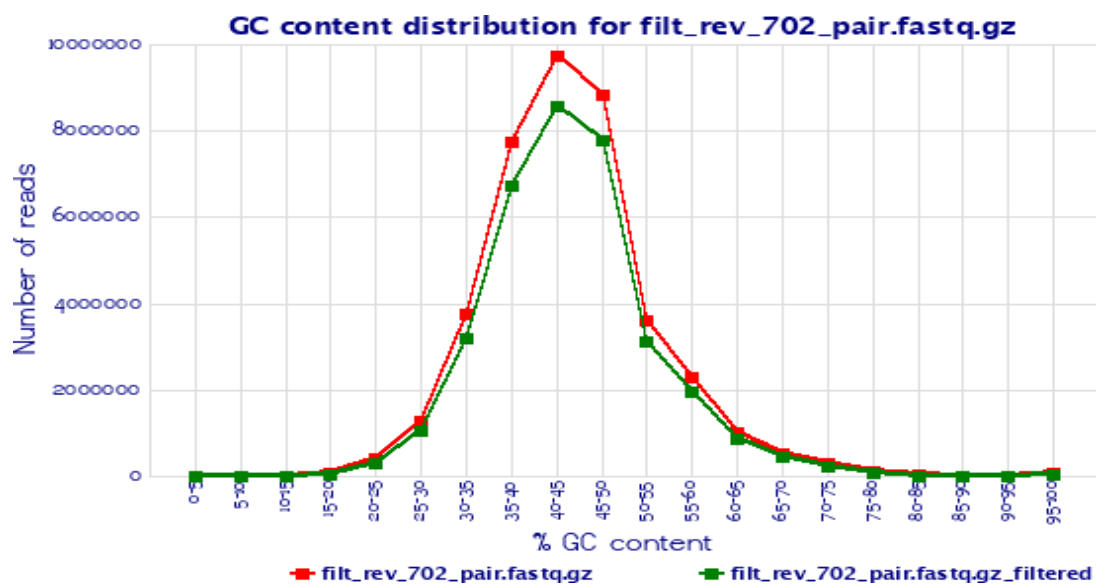


Fig 4.11.2: % GC Content SRR7145702 (BACKWARD READ)

### SRR7145708

Graph shows the GC content distribution in raw reads and filtered reads. For sample 708, GC content distribution is almost the same in filtered reads that is 40 – 45 % in both forward and reverse reads at slightly same amount of reads. Its there peak of gccontent at that position . other contents are depicted in graph below Fig 4.12.1 and 4.12.2

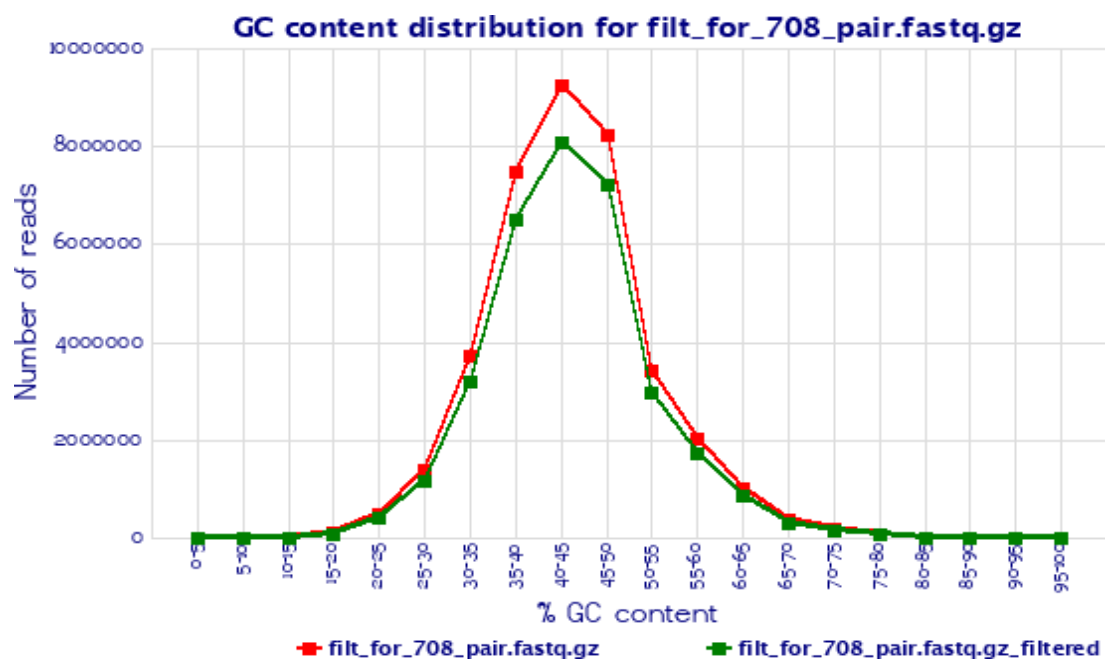


**Fig.4.12.1:** % GC Content SRR7145708 Forward read

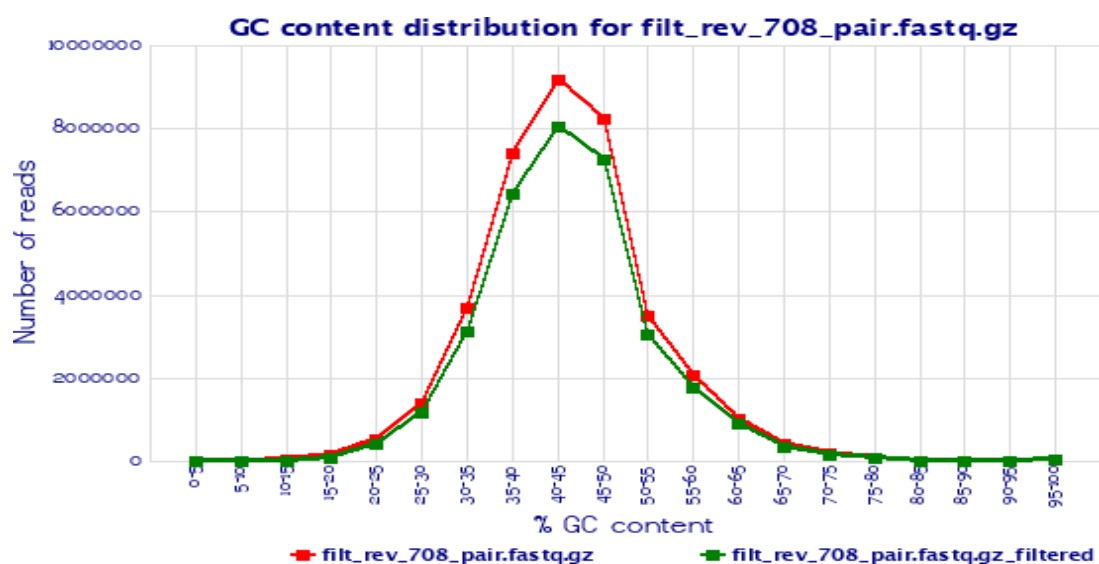


**Fig.4.12.2:** % GC Content SRR7145708 Backward read

## QUALITY DISTRIBUTION

### SRR7145702

The highest phred quality score for before and after qc for forward read is 36 at 3000000 average reads and for reversereads the average phred quality score is 36 before and after qc is 1000000 average reads. Fig .4.13.1 and Fig.4.13.2

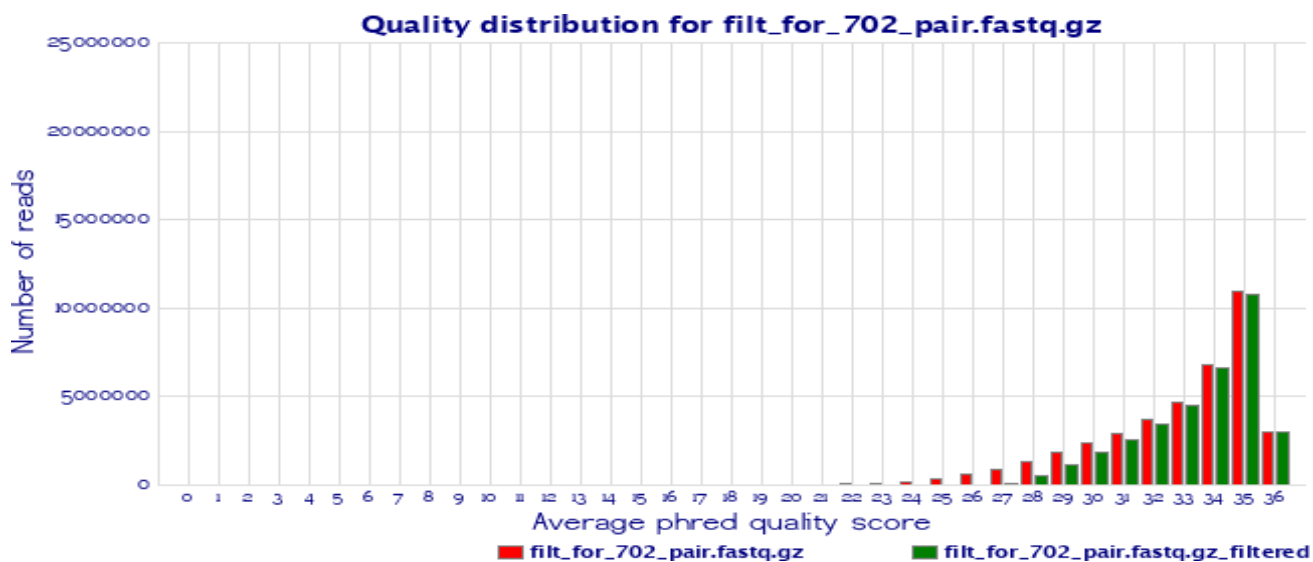


**Fig . 4.13.1:** Quality distribution of SRR7145702 FORWARD READ

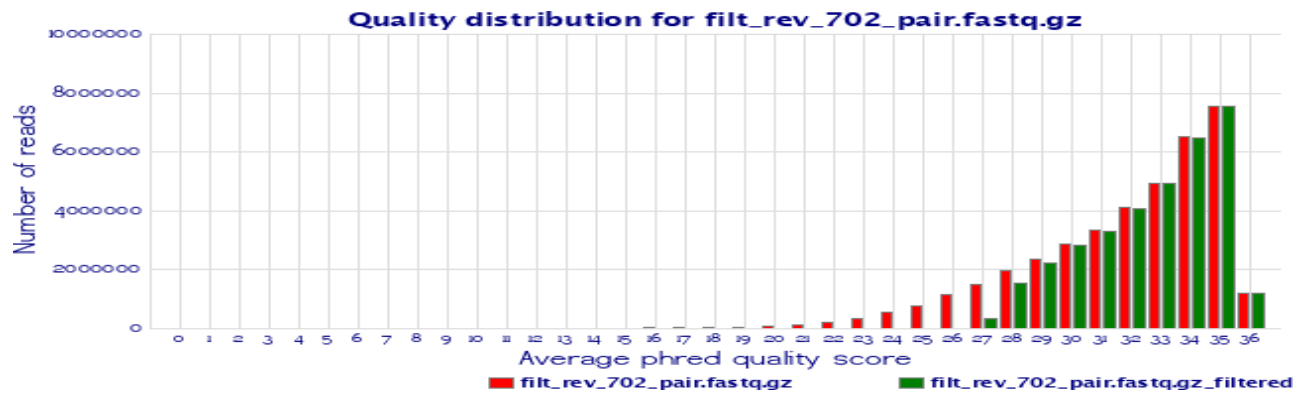


**Fig . 4.13.2. :** Quality distribution of SRR7145702 BACKWARD READ

### SRR7145708

The highest phred quality score for before and after qc for forward read is 36 at 3000000 average reads and for reverse reads the average phred quality score is 36 before and after qc is 1000000 average reads. Fig. 4.14.1 and Fig. 4.14.2

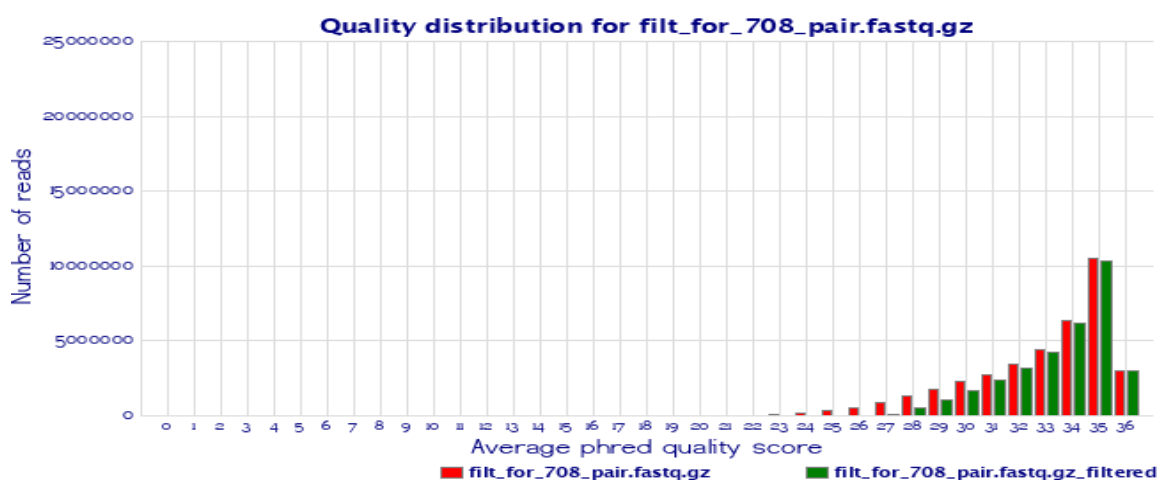


Fig . 4.14.1: Quality distribution of SRR7145708 FORWARD READ

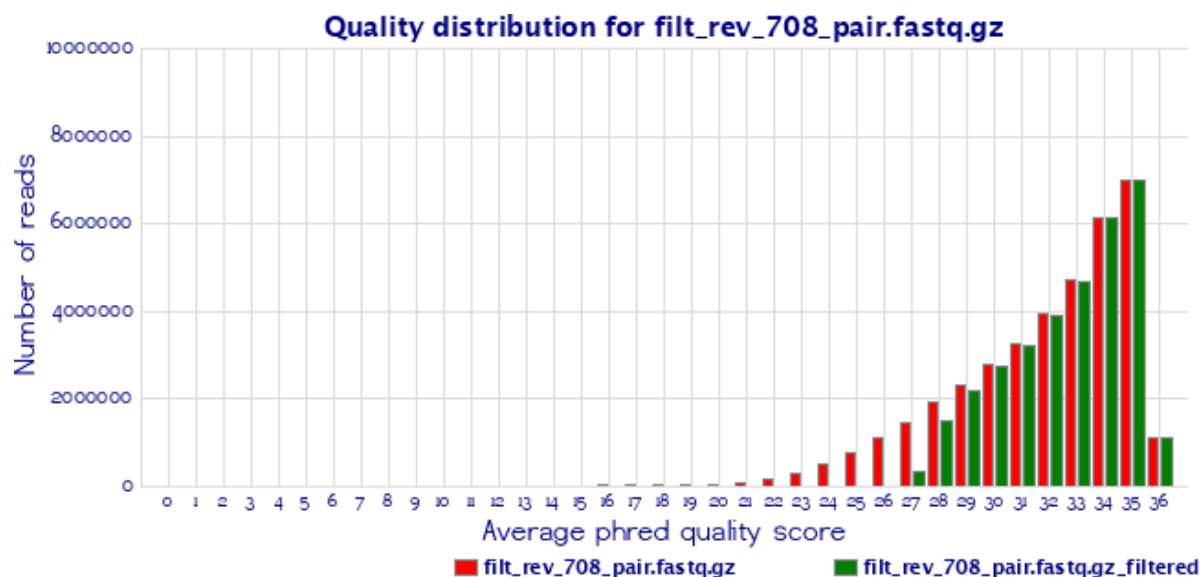


Fig . 4.14.2: Quality distribution of SRR7145708 BACKWARD READ

## SUMMARY OF QC ANALYSIS

### A) SRR7145702

Reads from the sample SRR7145702 demonstrates that there are 86.66 percent of high quality filtered reads (green colour), 13.34 percent of low quality readings (red colour), and no primer/adaptor contaminated reads for both forward and reverse reads.

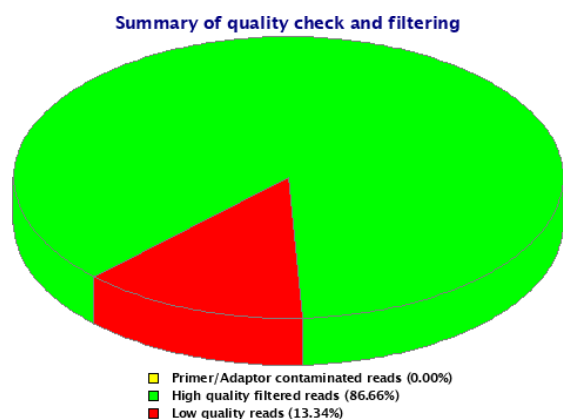


**Fig.4.15.1:** Summary of QC Analysis SRR7145702

### SRR7145708

Reads from sample SRR7145708 demonstrates that there are 86.54 % high quality filtered reads ( green colour). 13.46% low quality and no primer contaminated reads for both forward and reverse reads .

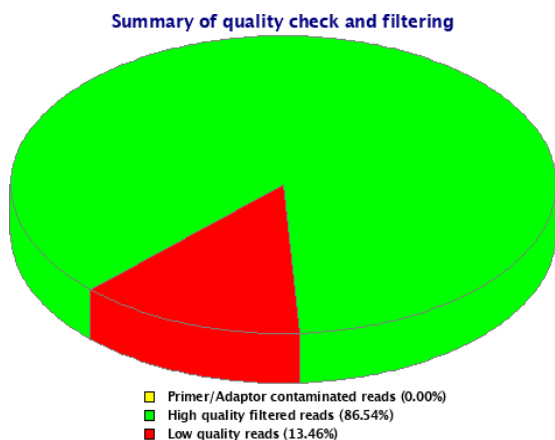


**Fig.4.15.2:** Summary of QC Analysis SRR7145708

## **VALIDATION FOR BUILDED TRANSCRIPTOME**

The assembled genome's validation data reveals the base composition (A,T,G, C), as well as the N25, N50, N75, N90, and N95 length values of three separate strains of the same species. The minimal contig length required to cover 50% of the genome is known as N50, and it indicates the quality of a prototype assembly. Trinity is to be run to obtain the data to check validation of transcriptome.

**Table 4.6 : TRINITY Results**

| <b>Parameter</b>        | <b>Trinity results</b> |
|-------------------------|------------------------|
| Total Sequences         | 42775                  |
| Total Bases             | 54523907               |
| Min sequence length     | 201                    |
| Max sequence length     | 15507                  |
| Average sequence length | 1274.67                |
| Median sequence length  | 917.00                 |
| N25 length              | 3062                   |
| N50 length              | 2035                   |
| N75 length              | 1241                   |
| N90 length              | 580                    |
| N95 length              | 344                    |
| As                      | 29.23%                 |
| Ts                      | 29.19%                 |
| Gs                      | 20.98%                 |
| Cs                      | 20.60%                 |

|        |        |
|--------|--------|
| (A+T)s | 58.42% |
| (G+C)s | 41.58% |

## ANNOTATION

We went for two types of DAVID and bioDBnet annotation to identify the role and biological operations of transcriptome and proteins they cipher. *Arabdiopsis thaliana* is been taken as model plants.and there are results we obtained from both the processes.

### David annotation

With model plant *Arabdiopsis thaliana* the following results were obtained which shows path ways obtained with particular nucleotide ids , gene name and log 2 fold change . log 2 fold change depicts logratio of transcripts expression value in two different conditions.

### DAVID ANNOTATION TABLE REPRESNTATION

| ID     | Gene Name       | Species     | GOTERM     | KEGG_PATHWAY |
|--------|-----------------|-------------|------------|--------------|
| 820102 | Heat shock      | Arabidopsis | GO:000641  | GO:000033    |
| 824798 | Melibiose       | Arabidopsis | GO:000591  | GO:000045    |
| 822036 | Serine          | Arabidopsis | GO:0016020 | GO:0016021   |
| 839174 | Zinc-binding    | Arabidopsis | GO:000691  | GO:0016628   |
| 820288 | fatty acid      | Arabidopsis | GO:000661  | GO:0016717   |
| 831020 | heat shock      | Arabidopsis | GO:000641  | GO:000033    |
| 830139 | myo-inositol    | Arabidopsis | GO:000601  | GO:000562    |
| 816757 | myo-inositol    | Arabidopsis | GO:000601  | GO:000562    |
| 816394 | phytochrome     | Arabidopsis | GO:000631  | GO:000561    |
| 833406 | purple acid     | Arabidopsis | GO:004274  | GO:000033    |
| 837716 | starch synthase | Arabidopsis | GO:001001  | GO:000095    |
| 842930 | transmembrane   | Arabidopsis | GO:000641  | GO:000551    |

**Table 4.7 bioDBnet ANNOTATION**

| TrID                      | nucleotide id | log2fold change | Pathway                                                                                                               |
|---------------------------|---------------|-----------------|-----------------------------------------------------------------------------------------------------------------------|
| TRINITY_DN1138_8_c0_g1_i1 | AT2G18790.1   | 5.414141414     | Circadian rhythm – plant                                                                                              |
| TRINITY_DN1642_9_c0_g1_i1 | AT3G56310.1   | 4.489594356     | Glycosphingolipid biosynthesis - globo series; Galactose metabolism; Glycerolipid metabolism; Sphingolipid metabolism |
| TRINITY_DN2215_7_c0_g3_i1 | AT2G22240.3   | 0.24849742      | Inositol phosphate metabolism; Metabolic pathways; Biosynthesis of antibiotics                                        |

## FINAL ONTOLOGY REPRESENTATION

The screenshot shows a Microsoft Excel spreadsheet titled "gargi arab DEG final (1) - Microsoft Excel". The spreadsheet contains a table with columns labeled A through O. The first column (A) is labeled "TrID" and contains identifiers like "TRINITY\_DN11388\_c0\_g1\_i1". Columns B through H contain numerical data: "baseMean", "baseMean", "baseMean", "foldChange", "log2FoldChange", "p-value", and "padj". Column I is labeled "GOTERM BP DIRECT" and contains GO terms like "GO:0006325\*chromatin". Column J is labeled "GOTERM GOTERM\_MF\_DIRECT" and contains GO terms like "GO:0005524\*ATP binding". The table is sorted by the "log2FoldChange" column in descending order. The bottom of the spreadsheet shows the Windows taskbar with the search bar and system clock.

| TrID                     | baseMean | baseMean | baseMean | foldChange | log2FoldChange | p-value  | padj     | GOTERM BP DIRECT                                    | GOTERM GOTERM_MF_DIRECT |
|--------------------------|----------|----------|----------|------------|----------------|----------|----------|-----------------------------------------------------|-------------------------|
| TRINITY_DN11388_c0_g1_i1 | 299.3419 | 93.3381  | 505.3456 | 5.414141   | 2.436733       | 0.002959 | 0.200065 | GO:0006325*chromatin                                | GO:0005524*ATP binding  |
| TRINITY_DN15216_c0_g1_i1 | 788.4241 | 241.8305 | 1335.018 | 5.520468   | 2.464791       | 0.000359 | 0.049465 | GO:0042744*hydrogen peroxide catabolic process      | GO:0005524*ATP binding  |
| TRINITY_DN15216_c0_g1_i1 | 788.4241 | 241.8305 | 1335.018 | 5.520468   | 2.464791       | 0.000359 | 0.049465 | GO:0051174*regulation of gene expression            | GO:0005524*ATP binding  |
| TRINITY_DN16017_c0_g1_i1 | 63.63961 | 8.485281 | 118.7939 | 14         | 3.807355       | 0.004606 | 0.263872 |                                                     |                         |
| TRINITY_DN16429_c0_g1_i1 | 1834.117 | 668.2159 | 3000.018 | 4.489594   | 2.166585       | 0.001462 | 0.131326 | GO:0005975*carbohydrate metabolic process           | GO:0005524*ATP binding  |
| TRINITY_DN186_c0_g1_i1   | 868.5628 | 207.8894 | 1529.236 | 7.356009   | 2.878923       | 3.66E-05 | 0.010098 | GO:0006979*response to stress                       | GO:0005524*ATP binding  |
| TRINITY_DN1885_c0_g1_i1  | 496.8604 | 186.6762 | 807.0445 | 4.323232   | 2.11211        | 0.00378  | 0.238281 | GO:0010021*amylopectin catabolic process            | GO:0005524*ATP binding  |
| TRINITY_DN21394_c0_g1_i1 | 61.16474 | 116.6726 | 5.656854 | 0.048485   | -4.36632       | 0.001884 | 0.157226 | GO:0006629*lipid metabolic process                  | GO:0005524*ATP binding  |
| TRINITY_DN21394_c0_g1_i1 | 61.16474 | 116.6726 | 5.656854 | 0.048485   | -4.36632       | 0.001884 | 0.157226 | GO:0006636*unsaturated fatty acid metabolic process | GO:0005524*ATP binding  |
| TRINITY_DN21667_c0_g1_i1 | 930.3168 | 1624.931 | 235.7023 | 0.145054   | -2.78534       | 5.60E-05 | 0.012305 | GO:0006457*protein folding                          | GO:0005524*ATP binding  |
| TRINITY_DN21667_c0_g1_i1 | 930.3168 | 1624.931 | 235.7023 | 0.145054   | -2.78534       | 5.60E-05 | 0.012305 |                                                     |                         |
| TRINITY_DN22157_c0_g1_i1 | 2766.32  | 4431.438 | 1101.201 | 0.248497   | -2.0087        | 0.004353 | 0.255921 | GO:0006021*inositol metabolic process               | GO:0005524*ATP binding  |
| TRINITY_DN22157_c0_g2_i1 | 2766.32  | 4431.438 | 1101.201 | 0.248497   | -2.0087        | 0.004353 | 0.255921 | GO:0006021*inositol metabolic process               | GO:0005524*ATP binding  |
| TRINITY_DN22157_c0_g3_i1 | 2766.32  | 4431.438 | 1101.201 | 0.248497   | -2.0087        | 0.004353 | 0.255921 | GO:0006021*inositol metabolic process               | GO:0005524*ATP binding  |
| TRINITY_DN22157_c0_g4_i1 | 2766.32  | 4431.438 | 1101.201 | 0.248497   | -2.0087        | 0.004353 | 0.255921 | GO:0006021*inositol metabolic process               | GO:0005524*ATP binding  |
| TRINITY_DN3100_c0_g1_i1  | 496.8604 | 169.7056 | 824.0151 | 4.855556   | 2.279636       | 0.001899 | 0.157501 |                                                     |                         |
| TRINITY_DN3100_c0_g2_i1  | 496.8604 | 169.7056 | 824.0151 | 4.855556   | 2.279636       | 0.001899 | 0.157501 |                                                     |                         |
| TRINITY_DN3100_c0_g1_i1  | 632.2713 | 129.4005 | 1135.142 | 8.772313   | 3.132957       | 1.84E-05 | 0.006741 | GO:0006468*protein phosphorylation                  | GO:0005524*ATP binding  |
| TRINITY_DN4364_c0_g2_i1  | 607.7583 | 129.4005 | 1086.116 | 8.393443   | 3.069263       | 2.87E-05 | 0.00912  | GO:0006468*protein phosphorylation                  | GO:0005524*ATP binding  |
| TRINITY_DN8603_c0_g1_i1  | 485.7824 | 140.0071 | 831.5576 | 5.939394   | 2.570316       | 0.00057  | 0.067941 |                                                     |                         |
| TRINITY_DN9939_c0_g1_i1  | 989.9495 | 1697.056 | 282.8427 | 0.166667   | -2.58496       | 0.000156 | 0.026553 |                                                     |                         |

**Table 4.8 Final representation of gene name obtained.**

| <u>sseqid</u>      | <u>TrID</u>                     | <u>Gene Name</u>                                                                                 |
|--------------------|---------------------------------|--------------------------------------------------------------------------------------------------|
| <u>AT2G18790.1</u> | <u>TRINITY_DN11388_c0_g1_i1</u> | <u>phytochrome B(PHYB)</u>                                                                       |
| <u>AT5G34850.1</u> | <u>TRINITY_DN15216_c0_g1_i1</u> | <u>purple acid phosphatase</u><br><u>26(PAP26)</u>                                               |
| <u>AT3G56310.1</u> | <u>TRINITY_DN16429_c0_g1_i1</u> | <u>Melibiose</u>                                                                                 |
| <u>AT1G26320.1</u> | <u>TRINITY_DN186_c0_g1_i1</u>   | <u>Zinc-binding dehydrogenase</u><br><u>family</u>                                               |
| <u>AT1G11720.2</u> | <u>TRINITY_DN1885_c0_g1_i1</u>  | <u>starch synthase 3(SS3)</u>                                                                    |
| <u>AT3G11170.1</u> | <u>TRINITY_DN21394_c0_g1_i1</u> | <u>fatty acid desaturase 7(FAD7)</u>                                                             |
| <u>AT5G02500.1</u> | <u>TRINITY_DN21667_c0_g1_i1</u> | <u>Heat shock protein 70</u>                                                                     |
| <u>AT4G39800.1</u> | <u>TRINITY_DN22157_c0_g1_i1</u> | <u>myo-inositol-1-phosphate</u><br><u>synthase 1</u>                                             |
| <u>AT1G66150.1</u> | <u>TRINITY_DN4364_c0_g1_i1</u>  | <u>transmembrane kinase 1</u>                                                                    |
| <u>AT3G24460.2</u> | <u>TRINITY_DN8603_c0_g1_i1</u>  | <u>Serine-domain containing</u><br><u>serine and sphingolipid</u><br><u>biosynthesis protein</u> |

## **CHAPTER 5: DISCUSSION**

*C. papaya* being the tropical plant and highly susceptible to drought conditions shows the irrigation is highly recommended. In our study of denovo transcriptome profiling of *Carica papaya* Linnsap under drought stress condition. We obtained following results :

### **5.1 TRANSCRIPTOME ASSEMBLY AND ANNOTATION**

The process of putting nucleotide sequences in the proper order is known as transcriptome assembly. The enormous number of contigs that follow assembly is one of the main problems that have been identified when using short-read sequencing techniques. The quantity and size of contigs and scaffolds that are available, the percentage of assembled reads, and the contig N50 value, which is widely used to measure and evaluate the quality of assembly, are all possible ways to assess the quality of de novo genome assembly (Choudhuri, 2014).

The N 50 value for the samples obtained is 2035 . The data shows the contigs have high (A+T )% which is 58.42% and remaining ( G+C)% which is 41.58%. which denotes low coverage as G+ C percentage is directly associated with the coverage.

For annotation process we went through different DAVID and bioDBnet annotation tools which provides the biological significance of enlisted genes . The results provides us name of gene name with its associated TrIDs with its pathways .

By using them the we obtained many genes involved in regulation of papaya in drought stress condition .the genes are:

### **A) PHYTOCHROME B (PHYB)**

PHYB gene having photoreceptors in the red and far red that control de-etiolation. two interchangeable variants exist: Pr and Pfr (active). involved in both the reaction to avoid shade and the light-promotion of seed germination. encourages seedling etiolation both with and without phytochrome A. Under far- red light, overexpression causes etiolation. after being exposed to far red light, accumulates in the nucleus. The phytochrome B molecule's dark reversion is affected by the phosphorylation status of the Ser-86 residue. It is related to stress condition in papaya . (Li H, *et al .*, 2022)

### **B) PURPLE ACID PHOSPHATASE 26(PAP26)**

PAP family of gene are potential candidates for improving P usage efficiency. In order to give background information for this family, we discovered PAP homologs in the genomes of 10 vegetable species, using *Arabidopsis thaliana* and *Amborella trichopoda* as references. Phosphate is an element that increases drought tolerating capacity so this gene is directly helps the plant to regulate conditions . PAP homologs with known enzyme structures and expression profiles could therefore be used as targets for plant breeding and/or genetic engineering initiatives to enhance P uptake and utilisation.

### **C) MELIBIASE FAMILY PROTEIN(AT3G56310)**

This gene codes and convert melibiose in glucoe and galactose , and causes cell expansion in different parts of plant to facilitate during drought stress condition.

### **D) ZINC-BINDING DEHYDROGENASE FAMILY PROTEIN(AT1G26320)**

Plant growth, development, and stress tolerance are regulated by zinc-finger proteins, a superfamily of proteins with a characteristic structural domain that coordinates a zinc ion and

binds nucleic acids. Zinc- finger proteins are involved in plant growth and resistance to abiotic challenges such salt, oxidative stress, cold temperatures, drought, and flooding. Zinc-binding alcohol dehydrogenases are NAD(P)-dependent oxidoreductases that catalyse the reversible oxidation of alcohols to aldehydes or ketones bytransferring hydride ions from alcohols to NAD+.

#### **E) STARCH SYNTHASE 3 (SS3)**

Contains a starch synthase that not only aids in the manufacture of starch but also performs a detrimentalregulating role in the biosynthesis of transient starch. It appears that the protein has a starch-binding domain (SBD).

#### **F) FATTY ACID DESATURASE 7 (FAD7)**

A key precursor for the synthesis of jasmonic acid (JA), a phytohormone that can affect plant growth, development, and primary metabolism, is provided by fatty acid desaturase 7 (FAD7), which also produces polyunsaturated fatty acids that encourage the desaturation of chloroplast membranes. Drought Stress Signaling via Jasmonic Acid Drought stress or a lack of water reduces turgor pressure, raises ion toxicity, and limits photosynthesis. JA signalling pathways have been linked to drought stress reduction in various studies. So connectively it controls the condition under drought stress in plants.

#### **G) HEAT SHOCK PROTEIN 70 (HSP 70) FAMILY PROTEIN(AT3G09440)**

Heat shock protein 70 (Hsp70) is a molecular chaperone that is activated by stress. Hsp70 performs this function by binding to its protein substrates and stabilising them against denaturation or aggregation until conditions improve. Hsp70 assists in the folding of newly generated proteins, the sub cellular transport of proteins and vesicles, the formation and dissociation of complexes, and the destruction of undesired proteins, in addition to its roles

during a stress response.

#### **H) MYO-INOSITOL-1-PHOSPHATE SYNTHASE 2(MIPS2)**

MIPS catalyses the rate-limiting step in the production of myo-inositol, which is a key molecule in the cell. MIPS genes are found in plants and encode substantially comparable enzymes. MIPS helps in sustaining ascorbic acid, phosphatidylinositol, and ceramide levels that control growth, development, and cell death.

#### **I) TRANSMEMBRANE KINASE 1(TMK1)**

The auxin-signaling proteins transmembrane kinase (TMK) interact with plasma membrane H<sup>+</sup>-ATPases, causing their phosphorylation and driving cell-wall acidification and hypocotyl cell elongation in Arabidopsis. Within seconds, auxin produced contacts between TMKs and H<sup>+</sup>-ATPases in the plasma membrane, as well as TMK-dependent phosphorylation of the H<sup>+</sup>-ATPases' penultimate threonine residue. TMKs directly phosphorylate plasma membrane H<sup>+</sup>-ATPase and are essential for auxin-induced H<sup>+</sup>-ATPase activation, apoplastic acidification, and cell growth, according to genetic, biochemical, and molecular data. Thus, the findings show that auxin and plasma membrane H<sup>+</sup>-ATPase activation play an important role in regulating apoplastic pH fluctuations and cell growth via TMK-based cell surface auxin signaling and sustaining drought condition.

#### **J) SERINC-DOMAIN CONTAINING SERINE AND SPHINGOLIPID BIOSYNTHESIS PROTEIN(AT3G24460)**

Serinc1-5, which integrates the polar amino acid serine into membranes and aids in the synthesis of two serine-derived lipids, phosphatidylserine and sphingolipids. Serinc is a distinct protein family with no amino acid homology to other proteins but is highly conserved

among eukaryotes. Sphingolipids provide membrane stiffness, aid in the trafficking of secretory proteins, and begin programmed cell death during stress condition. These listed upregulated and downregulated gene information denotes the way to regulate the conditions under drought stress. we talked about 9 gene expression in which 6 up regulated and 3 down regulated genes.

## **CHAPTER 7 : CONCLUSION**

After going through whole procedure of “**de novo transcriptome profiling of *C. papaya* sap under drought stress condition**” which was done at Banaras Hindu University , Rajiv Gandhi South Campus , Mirzapur , Uttar Pradesh . We went to all the steps involved in this process to get anticipated set of results . We started from downloading the reads from reliable and authentic source NCBI to proceeding to all the steps involved in de novo transcriptome profiling . Sample reads underwent Fastqc , NGS Toolkit , TRINITY, SAM tools , BED tools , Transimprove, BLASTn search and finally landing for annotation through DAVID and bioDBnet annotation . These process lastly concludes us the results for results of transcriptome profiling of Papaya sap under drought stress condition.

We have came across 9 genes listed as Phytochrome B, Purple acid phosphatase 26, Melibiase family protein , Zinc binding dehydrogenase family, Strach synthase 3, Fatty acid destaurase 7, Heat shock protein 70, Myo inisitol 1 phosphate synthase 2, Transmembrane kinase1 and Serinc-domain containing serine and sphingolipid biosynthesis protein which are being annotated in the process. Among them 6 of them are found to be upregulated genes and 3 of them are to be found downregulated . Being upregulated genes means these genes enhance the bulk of cellular factor in response of external flash whereas down regulated gene expression means decrease of bulk of cellular factor in response of external factor.

Phytochrome B gene is upregulated as its is found that it enhance plant chance to sustain drought tolerance by increasing ABA vulnerability , by sending signals to stress defensive proteins. Purple acid phosphatase 26 is upregulated as it increase phosphate uptake in plant which increases maximal quantam efficacy and balancing water content in plant parts , ultimately leading to drought tolerance. Melibase family of protein also upregulated alpha galactose activity as increasing raffinose synthesis which decreasing osmotic stress condition as by flow mechanism still remains uninsightful. Zinc binding dehydrogenase family being upregulated

helps in plants growth and catalyses photochemical aldehydic and ketonic reactions in plants and regulates ABA transduction signal.

Starch synthase 3 gene which increases starch accumulation and synthesis which is upregulation activity to shock up drought activity. FAD7 is found to be down regulated gene expression it reduces the level of linoleic acid as linoleic acid increases drought susceptibility in crops. HSP 70 downregulated protein shows that it dissociates protein complexes to reduce programmed cell deaths stabilizing under drought tolerance. MIPS 2 is also found to be downregulated in gene expression alleviating ROS scavenging causing biosynthesis of secondary metabolites and effecting drought stress tolerance. TMK1 involved in auxin signaling upregulated gene expression by controlling apoplastic  $p^h$  fluctuations contributing in drought stress control. Serinc-domain containing serine and sphingolipid biosynthesis protein lastly being upregulated in gene expression which is highly conserved and helps in protein trafficking and causing membrane stiffness to contribute to drought tolerance. With this disclosure we conclude that 6 genes are upregulated and 3 are down regulated .

# **BIBLIOGRAPHY**

Anna, I.R., Bob, M., Bryan, S.B., Aaron, E.D., Jeremy, D.G., & Nicole, T.P. (2009). Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, 25(16),2071-2073.

Bankevich, A., Nurk, S., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assesmbly algorithm and its applications to single cells sequencing. *Journal of Computational Biology*, 19(5),455–477.

Barton, M. D., & Barton, H. A. (2012). Scaffolder-software for manual genome scaffolding. *Source code for biology and medicine*, 7(1), 1-6.

Brainerd, K. E. & Fuchigami, L. ( 1981). Acclimatization of aseptically cultured apple plants to low relative humidity. *American Society for Horticulture Science*, 515–518.

Fiancette, R., Finlay, C.M., Willis, C., Bevington, S.L., Soley, J., Baker, S.M., Andrews, S., Hepworth, M.R., Withers, D.R. (2021). Reciprocal transcription factor networks govern tissue-resident ILC3 subset function and identity. *Nature Immunology*, 22(10):1245-1255.

Gans, J. D., & Wolinsky, M. (2007). Genomorama: genome visualization and analysis. *Biomedcentral bioinformatics*, 8(1), 1-5.

Giardine, B. (2005). A platform for interactive large-scale genome analysis. *Genome Resolution*, 15:1451–145

Guo, T., Luna, A., Rajapakse, V. N., Koh, C. C., Wu, Z., Liu, W., ... & Aebersold, R. (2019). Quantitative proteome landscape of the NCI-60 cancer cell lines. *Iscience*, 21, 664-680.

Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). Genome Project Data Processing Subgroup The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9.

Hu, H., Xiong, C. (2014). Genetic engineering and breeding of drought-resistant crops. *Review of Plant Biology*,

65:715–741.

Huang, D. W., Sherman, B. T., & Lempicki, H. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37, 1–13.

Kehr, J. & Buhtz, S. (2008). A Long distance transport and movement of RNA through the phloem. *Journal of Experimental Botany*, 59, 85–92.

Madroñer, L. (2018). Transcriptome analysis provides insights into the delayed sticky disease symptoms in *Carica papaya*. *Plant Cell Reports*, 37, 967–980 .

Mahouachi, J., Arbona, V. & Cadenas, G. (2007). Hormonal changes in papaya seedlings subjected to progressive water stress and re-watering. *Plant Growth Regulation*, 53, 43–51.

Mahouachi, J., Socorro, A. R. & Talon, S. (2006). Responses of Papaya Seedlings (*Carica papaya* L.) to Water Stress and Re-Hydration Growth, Photosynthesis and Mineral Nutrient Imbalance. *Plant Soil*, 281, 137–146.

Martin, L.B., Fei. Z., Giovannoni, J.J., Rose. H. (2013). Catalyzing plant science research with RNA-seq. *Frontiers of Plant Science*, 4:66.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for Next Generation Sequencing Data. *Genomics*, 95(6), 315-327.

Morf, J, Wingett, S.W., Farabella, I., Cairns, J, Furlan-Magaril, M, Jiménez-García, L.F., Liu, X, Craig, F.F, Walker, S., Segonds-Pichon, A., Andrews, S., Marti-Renom, M.A., Fraser, P. (2019). RNA proximity sequencing reveals the spatial organization of the transcriptome in the nucleus. *Nature Biotechnology*, 37(7):793-802.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS—an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35,W182-5.

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, 7(2), e30619.

Peng, Y., Lai, Z., Lane, T., Nageswara-Rao, M., Okada, M., Jasieniuk, M., ... & Stewart Jr, C. N. (2014). De

novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. *Plant Physiology*, 166(3), 1241-1254.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., ... & Fraser, P. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, 42(1), 53-61.

Singh, D. & Laxmi, H.S (2015). Transcriptional regulation of drought response: a tortuous network of transcriptional factors. *Frontiers in Plant Science*, 6, 895.

Spiegelman, Z., Golan, G. & Wolf, K. (2013). Don't kill the messenger: Long-distance trafficking of mRNA molecules. *Plant Science*, 213, 1–8.

Wang, Z., Gerstein, M., & Snyder, D. (2009) M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63.