Studies on "COMPARATIVE GENOME ANALYSIS OF *Paenibacillus macerans* CMB402, CMB401 AND CMB393"

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

Master of Science

in

Plant Biotechnology

Submitted by

KHUSHBOO VERMA

Supervisor

DR. VISHAL SRIVASHTAV

Co-Supervisor

DR. ABHINAV SINGH

DEPARTMENT OF GENETICS AND PLANT BREEDING

INSTITUTE OF AGRICULTURAL SCIENCES

BANARAS HINDU UNIVERSITY

VARANASI – 221005

ID No. 19430PLB011 **2021** **Enrolment no. 419060**

BANARAS HINDU UNIVERSITY
कृषि विज्ञान संस्थान
INSTITUTE OF AGRICULTURAL SCIENCES

Dr. VISHAL SRIVASHTAV
Assistant Professor

Department of Genetics and Plant breeding
Institute of Agricultural Sciences
BHU, Varanasi -221005

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, B.H.U, Varanasi.

Dear Sir,

I have great pleasure in forwarding the thesis entitled **Studies on** "**Comparative genome analysis of *Paenibacillus macerans* CMB402, CMB401, CMB393**" submitted by **Ms. Khushboo Verma, I.D. No.19430PLB011,** in partial fulfillment of the requirements for the degree of **Master of Science** in **Plant Biotechnology**, from Department of Genetics and Plant breeding, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi.

I certify that the entire scheme of investigation reported herein, 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.

Thanking you,

Forwarded by: Your's faithfully

Prof. B. Sinha (Head)

Dr. Ravindra Prasad (Course Coordinator)

Dr. Vishal Srivashtav (Supervisor)

Studies on "COMPARATIVE GENOME ANALYSIS OF *Paenibacillus macerans* CMB402, CMB401 AND CMB393"

By

Khushboo Verma

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science (Agriculture) in

Plant Biotechnology

DEPARTMENT OF GENETICS AND PLANT BREEDING

INSTITUTE OF AGRICULTURAL SCIENCES

BANARAS HINDU UNIVERSITY

VARANASI - 221 005

ID. No.19430PLB011 **2021** **Enrolment No: 419060**

APPROVED BY ADVISORY COMMITTEE

Advisor	:	**Dr. Vishal Srivashtav** Assistant Professor (Plant Biotechnology) Department of Genetics and Plant breeding, Institute of Agricultural Sciences, Rajiv Gandhi South Campus, BHU, Barkachha, Mirzapur, India
Co-advisor	:	**Dr. Abhinav Singh** Assistant Professor, Agriculture Statistics, B.Sc. (Ag.) Course, Rajiv Gandhi South Campus, BHU, Barkachha, Mirzapur, India
Member	:	**Dr. Rajesh Kumar** Assistant Professor, Department of Genetics and Plant breeding, Institute of Agricultural Sciences, Rajiv Gandhi South Campus, BHU, Barkachha, Mirzapur, India
Member	:	**Dr. Ashok Kumar** Assistant Professor (Plant Biotechnology) Department of Genetics and Plant breeding, Institute of Agricultural Sciences, Rajiv Gandhi South Campus, BHU, Barkachha, Mirzapur, India
EXTERNAL EXAMINER	:	

DEDICATED TO MY MOM AND DAD

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the God, the Almighty, for his showers of blessings throughout my research work to complete the research successfully.

I am extremely grateful to Prof. B. Sinha, Head of the Department of Genetics and Plant Breeding for his inspiring guidance. I emphatically and gratefully acknowledge extend my loyal and venerable thanks to Dr. Ravindra Prasad for liberal approach as a co-ordinator of this degree programme.

Foremost, I would like to express my sincere gratitude to my advisor Dr. Vishal Srivashtav (Assistant Professor) Plant Biotechnology Laboratory, 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, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I emphatically and gratefully acknowledge extend my loyal and venerable thanks to my respected Co-advisor Dr. Abhinav Singh (Assistant Professor) their irreplaceable suggestions, guidance and kind attitude leading my path to achieve the destination during the entire move despite of his heavy schedule of work. Besides my advisor, I would like to thank the rest

of my advisory committee: Dr. Rajesh Kumar (Assistant Professor) and Dr. Ashok Kumar (Assistant Professor), for their encouragement, insightful comments, and hard questions.

My sincere thanks also goes to Dr. Roli Budhwar (Scientific Officer, Bionivid Technology Pvt. Ltd., Banglore), for offering me the online internship opportunities in their groups and leading me working on diverse exciting projects.

I would like to thank my family: my parents Dileep Kumar Verma and Mithlash Verma, for giving birth to me at the first place and supporting me spiritually throughout my life. I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I would like to thank my younger, little brother (Rudra Verma) for not disturbing me in my entire thesis. Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

I thank my fellow classmate Raksha Bhale, for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four months. Also I thank my friends: Muskaan Kaushal, N.Sai Prakash, Itishree Behera, Anamika Sharma.

Date :

Place : Varanasi KHUSHBOO VERMA

LIST OF CONTENTS

List of Abbreviations

DNA	Deoxyribonucleic Acid
NGS	Next Generation Sequencing
WGS	Whole Genome Sequencing
ENA	European Nucleotide Archive
NCBI	National Center for Biotechnology Information
BLAST	Basic Local Alignment Search Tool
BARRNAP	Basic Rapid Ribosomal RNA Predictor
RAST	Rapid Annotation using Subsystem Technology
PHASTER	PHAge Search Tool Enhanced Release
RGI	Resistance Gene Identifier
BRIG	BLAST Ring Image Generator
CARD	Comprehensive Antibiotic Resistance Database
ARO	Antibiotic Resistance Ontology
dNTP	deoxy Nucleoside Triphosphate
SNP	Single Nucleotide Polymorphism
et al.	et alia (and others)

List of Tables

List of Figures

INTRODUCTION

CHAPTER 1: INTRODUCTION

Genus Paenibacillus is a group of rod – shaped bacteria that make ATP either oxygen is present or not (Liu *et al.,* 2019). Paenibacillus name derived from Latin adverb paene which means almost; almost a Bacillus (Ash *et al.,* 1993). *Paenibacillus macerans* had flagella which are projecting in all directions and its size ranges from 0.6 – 3.5µm, while *P. thermophilus* and *P. macerans* both showed sequence similarity of 99.3% (Zhou *et al.,* 2012). Pseudomonas, Rhizobium and Bacillus genus are among the foremost powerful phosphate solubilizers (Rodriguez *et al.,* 1999). This genus mostly are plant growth promoting rhizobacteria which helps plant for providing resistance to particular diseases and in agricultural productivity (Seldin, 2011). However, *P. macerans* participate in fermentation of hexoses, cellulose etc. and showed high metabolic rates and also helps in production of fuels and chemicals (Gupta *et al.,* 2017). To ensure crop productivity, nitrogen availability in the soil is a major limiting factor for plant growth

Member of genus Paenibacillus are capable of nitrogen fixation for those plant species which are tolerant to heavy metals and grows in extreme environments (Navarro *et al.,* 2012). It is a nitrogen fixing bacteria but it also has a drawback that, by repeated sub culturing it loses its activity as compared to phosphate solubilizing fungi (Sharma *et al.,* 2013). Another study showed that it is paraphyletic, this genus currently comprises around 200 species and able to produced variety of biocidal substances and further nitrogen fixing ability determined by $15N_2$ fixing assays used to estimate nitrogenase activity (Grady *et al.,* 2016).

Many species of Paenibacillus which almost includes *P. azotofixans*, *P. polymyxa*, *P. macerans*, *P. odorifer*, *P. graminis*, *P. sabinae*, *P. zanthoxyli*, *P. peoriae*, *P. brasilensis* etc. showed the nitrogenase activity (Hong *et al.*, 2009; Jin *et al.*, 2011). Genome of this species has *nif* gene operon at almost 10.5kb region which also includes nine genes and that particular *nif* gene operon involves in nitrogen fixation (Xie *et al.*, 2016). It also involves in phosphate solubilization because of presence of *gcd* (glucose dehydrogenase) gene which involves in oxidation of glucose into gluconic acid (Li *et al.*, 2019) and iron acquisition (Wen *et al.*, 2011). *Paenibacillus macerans* ATCC8244 strain is predominantly involve in nitrogen fixation. (Daligault *et al.*, 2014). Many species of Paenibacillus which almost includes *P. mucilaginosus* (Hu *et al.*, 2006), *P. elgii* (Das *et al.*, 2010), *P. kribbensis* (Marra *et al.*, 2012), *P. xylanilyticus* (Pandya *et al.*, 2015), *P. peoriae* (Xie *et al.*, 2016), *P. polymyxa* and *P. macerans* (Wang *et al.*, 2012) showed the phosphate solubilization activity.

Few *P. macerans* show inhibitory effect against *R. solanacearum* strains and reduced the disease incidence (Li *et al.*, 2017). It has negative quality that it acts as opportunistic infectors of humans and causes spoilage of pasteurized dairy products and also helps to remove contaminants from the waste water (Grady *et al.*, 2016). *Paenibacillus macerans* causes allergies because it produces histamines (Jerez *et al.*, 1994). It creates intracranial infection which leads to periorbital puncture (Bert *et al.*, 1995). *P. macerans* were vancomycin resistant and not really show the sensitivity against erythromycin and further it shows 100% resistant against ampicillin with minimum inhibitory concentration 6 – 7.2 mg/L (Nieto *et al.*, 2017). It cause infection in humans and remain dormant until it revert into vegetative spores, infection requiring life-long antibiotic therapy (Szaniawski *et al.*, 2019).

Paenibacillus macerans involves in production of chitosanase by using chitinous materials which also plays important role in biomedical (Doan *et al.,* 2016). It acts as thermostable xylanase producing microbes which is used in paper and pulp industry (Dheeran *et al.,* 2012). It inhibit the growth of *Ralstonia solanacearum* because it acts as a biocontrol agent which involve competition for nutrients and suppress the growth of bacterial wilt and helps in sustainable management strategies for bacterial wilt of sweet peppers and other solanaceous crop (Mamphogoro *et al.,* 2020). It produced exopolysaccharides and biosurfactant which has potential role in cosmetics (Liang *et al.,* 2014). Wang *et al.,* 2012 showed that *Paenibacillus macerans* solubilize and form clear halozone when $Ca_3(PO_4)$ and $CaHPO_4$ used as phosphate source in National Botanical Research Institute's Phosphate growth medium but it was not able to show phosphate solubilization activity when $AlPO_4$ or $FePO_4$ used as phosphate source.

For identification of *Paenibacillus macerans*, *rpoB* gene which helps to discriminate nitrogen fixing genus Paenibacillus, used as an option to the 16SrRNA gene (Mota *et al.,* 2004) because different copies of 16SrRNA gene restricts its use in *Paenibacillus macerans* (Berge *et al.,* 2002). rpoB DNA sequences of this bacteria was CAGTCC and after alignment of rpoB DNA sequences, reverse primer rpoB PAEN is CTIAGI was obtained (Mota *et al.,* 2005). *P. macerans* 3CT49 that is ranging from 5.1 to 7.1 Mb with 49 to 53% GC contents isolated from cheese curds were sequenced by using MiSeq platform using a MiSeqV2 reagent kit (Olajide *et al.,* 2020).

High throughput sequencing that is Illumina sequencing involves in investigating bacteria for knowing about outbreak and spread of drug resistance (Loman *et al.,* 2012). Assembly is a step which is used to generate contigs, and for that assembler is used that is, SPAdes which

works for both single cell and multicell, it is an open source software (Bankevich *et al.*, 2012). RAST server used for annotating bacterial genomes which identifies protein encoding, helps to reconstruct the metabolic network and it is freely available to the community (Aziz *et al.*, 2008). Mauve is a contig ordering tool that orders or orients the contigs into scaffolds, use for microbial genome comparison because it aligns homologous regions even if genome undergone deletions or insertions (Rissman *et al.,* 2009).

For comparison and visualization of Prokaryote genomes, BRIG is used which can visualize or display the presence, absence or variation among different strains of bacteria and also display the custom graphs and annotation (Alikhan *et al.*, 2011). By using genomic comparison tool, not only analyzed IAA biosynthesis, nitrogen fixation and phosphate solubilization but also able to analyzed systemic resistance inducer production and reveals percentage identity between comparative genome (Li *et al.,* 2020). For identification of major biocontrol mechanisms and functional genes among related organisms, comparative genomics is recognized as an important tool (Helfrich *et al.,* 2014).

Currently, genomic sequences of *P. macerans* ATCC8244 (GenBank accession number – NZ_KN125580) have completed (Daligault *et al.,* 2014). To advance our understanding of the genome, used those strains of *P.macerans* which were not yet analyzed and their genomic characterization were not conducted yet.

The study **Studies on "Comparative Genome analysis of *Paenibacillus macerans* strains CMB402, CMB401 and CMB393"** was aimed to achieve the following objectives:

1. To study genome sequence and comparative analysis of three different strains of *Paenibacillus macerans*.

2. To identify plasmid, prophage and antibiotic resistance gene of *Paenibacillus macerans*.

LITERATURE

REVIEW

CHAPTER 2: LITERATURE REVIEW

This present study of **Studies on "Comparative genome analysis of *Paenibacillus macerans* CMB402, CMB401, CMB393"** was done through online mode at Rajiv Gandhi South Campus, Banaras Hindu University, Mirzapur, Uttar Pradesh.

2.1. Paenibacillus genus

Paenibacillus is a genus of bacteria that can produce ATP by aerobic respiration if oxygen is available, but may also ferment and generate endospores if oxygen is not available (Ash *et al.*, 1993). Paenibacillus has been found or isolated in a number of locales and samples, including Antarctic sediment (Montes *et al.,* 2004), boreal soil, rhizospheric soil, waste water, clinical samples, and others, and is hostile to the psychrophilic phytopathogenic fungus (Hoshino *et al.*, 2009). Few Paenibacillus species produce catalase and oxidase but are unable to digest casein, collagen, starch, or DNA, and *Paenibacillus anaericanus* is its closest phylogenetic cousin (Lee *et al.*, 2007). *Paenibacillus dendritiformis* has a colony diameter of 5cm and a C morphotype pattern (Ben – Jacob E *et al.*, 1997).

Agriculture uses *Paenibacillus macerans*, horticulture uses *Paenibacillus polymyxa*, industrial uses *Paenibacillus amylolyticus*, and medicinal uses *Paenibacillus peoriate* (Choi *et al.,* 2004). Paenibacillus sp. is a kind of bacteria which called as desulfurizing bacteria that makes 2-(2'-Hydroxyphenyl) benzene sufinate desulfinate (Konishi *et al.,* 2003). Various Paenibacillus species produces antimicrobial substances, various extracellular enzymes which affect fungi, soil bacteria and even anaerobic pathogens (Piuri *et al.,* 1998).

Paenibacillus which includes almost 74 and above species from which few are named in [Figure 2.1].

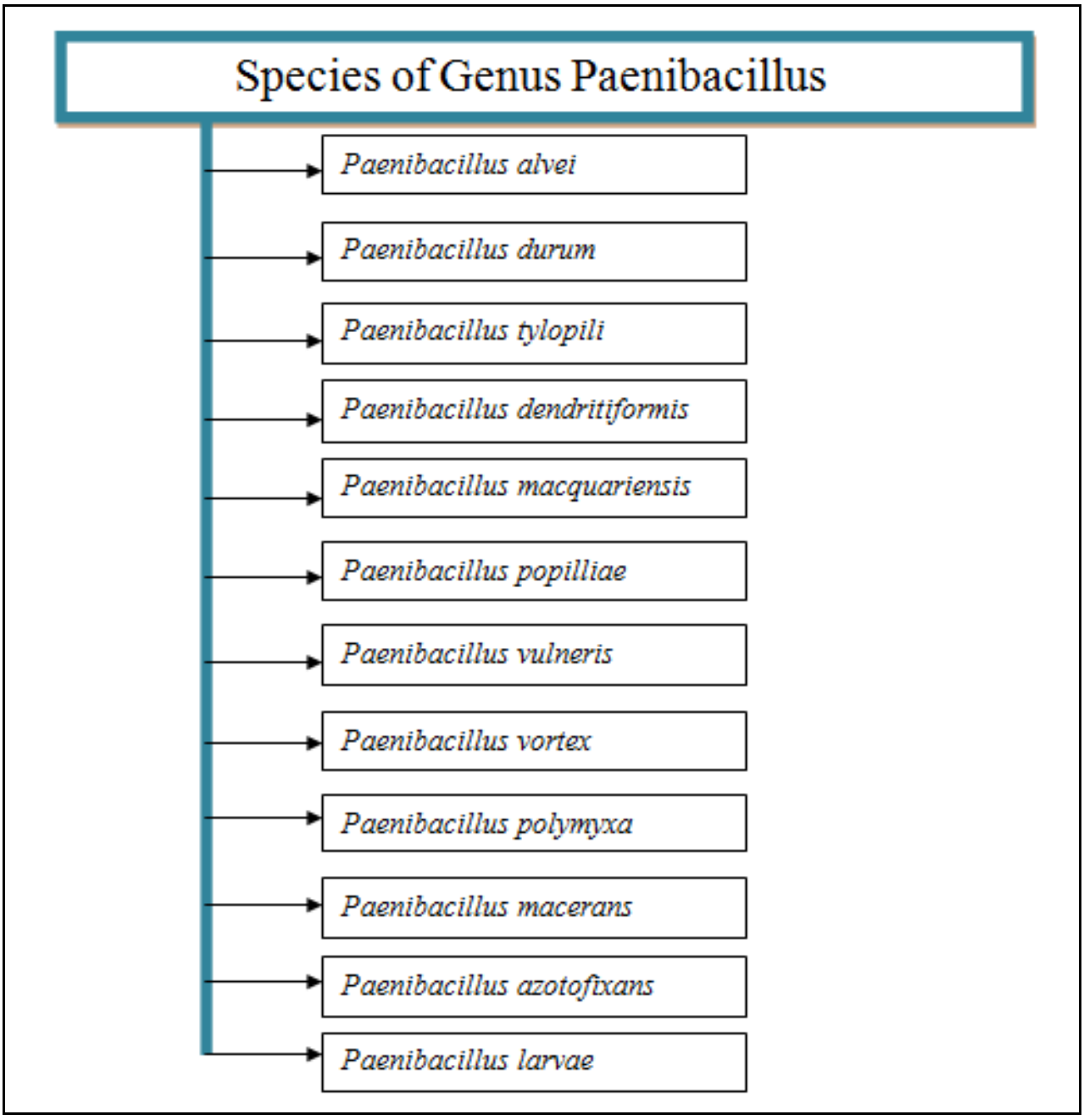


Figure 2.1. Species of Genus Paenibacillus

2.2 ***Paenibacillus macerans***

Schardinger (Austrian Biologist) found *Paenibacillus macerans* in 1905, according to Ash *et al.*, 1994 (Oren, 2019) [Table 2.1]. At pH levels ranging from 4.0 to 4.6, spore-forming organisms such as *Paenibacillus macerans* may survive and thrive (Ramesh, 2003). *Paenibacillus macerans* is a kind of bacteria that can fix atmospheric nitrogen and is also involved in fermentation. It may be found in soil.

Table 2.1. Taxonomic classification

Scientific classification of *Paenibacillus macerans* (Schardinger, 1905)	
Kingdom	Bacteria
Division	Firmicutes
Class	Bacilli
Order	Bacillales
Family	Paenibacillaeae
Genus	Paenibacillus Ash *et al.*, 1994
Species	*Paenibacillus macerans*

Paenibacillus macerans does not have capsule and have appendage that protrudes from the cell body which acts as flagella for movement, it maybe Gram positive, Gram variable or Gram negative and in bacilli shaped which has 0.7 – 2.5 µm in size and it also shows positive results for some biochemical tests for example, catalase, hydrolysis of starch, acid production from glucose, sucrose, lactose, fructose, glycerol, xylose, sorbitol, maltose and reduction of nitrate to nitrite but it also has a drawbacks because for few biochemical tests it shows negative

results for example, Indole production, H_2S production and casein decomposition (Ding *et al.*, 2005). It has 24 rRNA, 82 tRNA, 6234 protein, 6545 gene, 200 pseudogene, and 5 additional RNAs and is 7.39 Mb in size with 52.6 percent GC content (Daligault *et al.*, 2014 and Kobayashi *et al.*, 2019).

Menaquinone, discovered in Gram positive bacteria *Paenibacillus macerans*, is an example of an enzyme that catalyses the oxidation of succinate (Lancaster, 2018). It is involved in spoiling in less acid foods or fruit pulps with a pH of 3.7 to 4.5 (Silva *et al.*, 2014), as well as nitrogen fixation, phosphate solubilization, plant growth stimulation, iron acquisition, phytohormone synthesis, and functions as a biocontrol agent (Grady *et al.*, 2016). Paenibacillus species can generate immunological asexual spores that develop inside the cell of certain bacteria and are found all over the world (Bloemberg *et al.*, 2001).

2.3 *Paenibacillus macerans*, a nitrogen fixing bacteria

Agricultural output rises, but agricultural product traits diminish, as the fraction of light energy converted into chemical energy and leaf area growth is impacted by nitrogen element (Erisman *et al.*, 2008). Nitrogen is a necessary component of DNA, RNA, proteins, and other biological structures. This region of the soil is enriched with various compounds such as sugars, complex polysaccharides, amino acids, proteins, and other compounds that are directly influenced by root secretions and microorganisms. It is a zone of interactions between microorganisms (Garcia *et al.*, 2018) and their respective plant (Badri *et al.*, 2009).

The *Nif* gene genes for an enzyme that is responsible for nitrogen fixation in the atmosphere (Mus *et al.,* 2018). Nif regulon consists of seven operons and 17 *nif* genes, such as *nif A*, *D*, *K*, *F*, *H*, *U*, *W*, *Q*, *B*, *L*, *M*, *V*, *X*, *E*, *N*, *T*, *K*, *J*, and these genes conduct diverse tasks, such as *nif A* activating positive transcriptional activator, nif Q incorporating molybdenum into Fe-Mo cofactor, and so on (Lee *et al.*, 2001 and Khanal *et al.*, 2020) [Figure 2.2].

Paenibacillus macerans, *Paenibacillus sonchi*, and other Paenibacillus genus members are capable of nitrogen fixation. *Paenibacillus macerans* is a rhizobacterium that promotes plant development (Mohamed *et al.,* 2019). The *nif H* gene is a 323-bp segment that is used to identify nitrogen-fixing bacteria (Auman *et al.,* 2001). It demonstrates nitrogen fixation, which is supported by acetylene-reduction activity and a *nif H* fragment sequence (Achouak *et al.,* 1999).

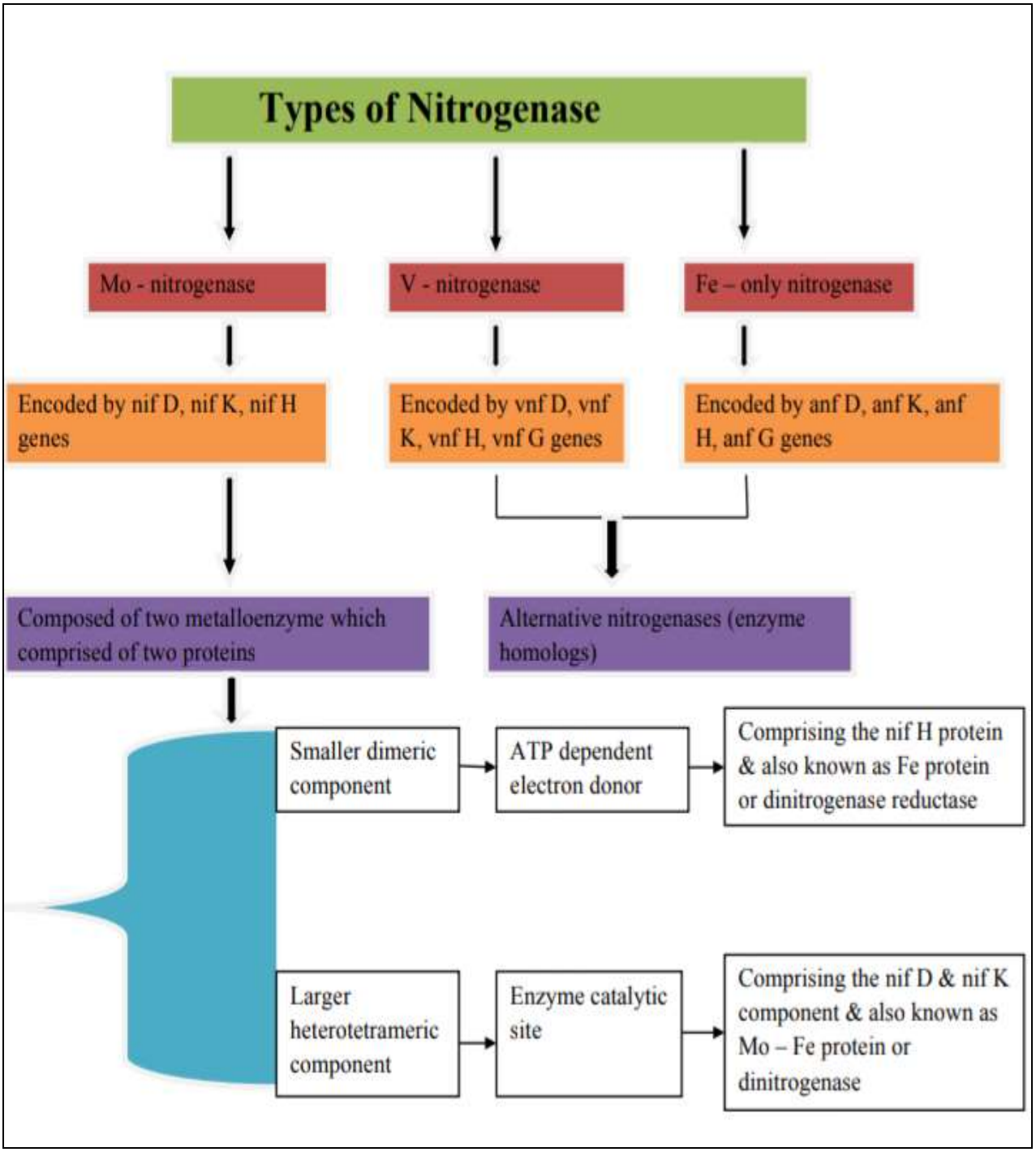


Figure 2.2. Types of nitrogenase

2.4 *Paenibacillus macerans* : Phosphate solubilization and what it produced ?

Paenibacillus macerans and other bacterial strains are important in the phosphorus cycle because they convert mineral phosphorus or organic phosphorous into an accessible form that is dependent on the culture's nutritional, physiological, and growth conditions, and phosphatases is an enzyme that aids in the mineralization of organic phosphorous into organic acids (Behera *et al.,* 2014). When bacterial samples were collected from non-arbuscular mycorrhiza or bulk soil, it exhibited greater phosphate solubilization (Wang *et al.,* 2012).

Paenibacillus mucilaginosus, *Paenibacillus macerans*, *Paenibacillus skribbensis*, and *Paenibacillus elgii* were among the Paenibacillus species that demonstrated phosphate solubilization, while *Paenibacillus mucilaginosus*, *Paenibacillus macerans* showed the potassium solubilization (Hu *et al.,* 2006). The activity of phosphate solubilizing bacteria was stimulated by a lower quantity of exogenous soluble phosphate, and vice versa (Patel *et al.,* 2008). Genes involved in phosphate solubilization are *pst* (Pi – specific transporter), *phoA* (alkaline phosphatase), *phyC* (phytase), *ushA* (nucleotidase) and *glpQ* (glycerophosphoryldiester phosphodiesterase) (Pragai *et al.,* 2004).

Paenibacillus macerans produce enzymes that act on alpha – glucans belonging to CGTase, phytohormones, glucanases, cellulose and proteases that are involved in destruction of eukaryotic cell walls, antimicrobial includes peptides, enzymes and volatile organic compounds [Table 2.2].

Table 2.2. Products produced by *Paenibacillus macerans*

Phosphate Solubilizing Bacteria	Enzymes / Hormones that are produced	References
Paenibacillus macerans	Enzymes that act on alpha – glucans belonging to CGTase	Taniguchi *et al.*, 2009
	Cyclodextringylcosyltransferase	Fravel, 2008
	Production of phytohormones that is indole – 3 – acetic acid	Patten *et al.*, 2013
	Hydrolysis of chitin by chitinase	Grady *et al.*, 2016
	Glucanases, cellulose & proteases that are involved in destruction of eukaryotic cell walls	Naing *et al.*, 2014
	Antimicrobial includes peptides, enzymes and volatile organic compounds	Grady *et al.*, 2016

2.5 Role of Bioinformatics to know about microorganism

The American Academy of Microbiology in 2009 stated that the use of computer examination to studies the composition, structure and interactions of cellular molecules, known as Bioinformatics. Whole Genome Sequencing has the ability to collate and look into the genetic sequences of bacteria, archaea etc. and tell about the processes microbes carry out.

From wet lab, there are lots of data available which are not analysed or examined yet and to full fill this goal, research by using bioinformatics has to be done which is further classified into three approaches that is examine based upon the available wet-lab data results, use of mathematical modeling, integrated approach that combine search techniques with mathematical approach and bioinformatics research used to know about particular genome function, metabolic pathways and gene expression of particular microorganisms and to know about pathogenicity, antimicrobial genes and comparative analysis between different strain of particular species and many more (Bansal, 2005). Applications of Bioinformatics are – Whole Genome Sequencing, Annotation, Alignment, Phylogenetics or evolutionary relationships, protein structure prediction and in Homology modeling (Varli H *et al.*, 2014).

DNA sequencing is any chemical, catalyst or technological procedure for decisive the linear order of ester bases in deoxyribonucleic acid. Sanger sequencing by replicative synthesis within the presence of dideoxy ester chain slayer monomers has currently given thanks to next generation sequencing by short parallel read technologies. The term typically applies to the complete sequence determination pipeline as well as post-sequencing software package analysis.

Many scientists were involved in developing chain termination technique sequencing, cappillary DNA sequencing, and next generation sequencing, including Richard Holley, who made the first attempt to sequence the nucleic acid [Table 2.3].

Table 2.3. History of DNA Sequencing method

Name of Scientist / Biotechnology company	Year	Achievements
Richard Holley	1964	First attempt to sequence the nucleic acid
Fredrick Sanger	1977	Chain termination method of sequencing
Maxam and Walter Gilbert	1977	Genome of bacteriophage X174 was sequenced using 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

2.5.1. Next Generation Sequencing

Next Generation Sequencing divides millions of fragments of DNA in parallel fashion and further reads are evaluated by numerous computational method (Behjati *et al.,* 2013). Following sequencing platforms which comes under next generation sequencing:

2.5.1.1. 454 Pyrosequencing

454 Pyrosequencing was called after the company 454 Life Sciences, which invented it. Pyrosequencing employs polymerase enzyme to produce complementary strands and dideoxynucleotides to inhibit chain amplification. Primer, which is ssDNA, is used to form the complementary strand by adding dNTPs, and when the right dNTPs are added, pyrophosphate is released, which is then converted to ATP. Adenosine triphosphate (ATP) transforms luciferin to oxyluciferin, which creates light (Gharizadeh *et al.,* 2003).

2.5.1.2. 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 bases added in a row might be tricky (Rothberg *et al.,* 2011).

2.5.1.3. Sequencing by ligation (SOLiD)

SOLiD is a deoxyribonucleic acid ligase-based sequencing technique. Desoxyribonucleic acid ligase is a widely utilised catalyst in biotechnology because of its capacity to ligate double-stranded deoxyribonucleic acid strands (Ho *et al.*, 2011).

2.5.1.4. Illumina sequencing

There are four fundamental phases in the Illumina sequencing process: Sample preparation, cluster formation, sequencing, and knowledge analysis. Adaptors are added to the ends of deoxyribonucleic acid in all preparation techniques. Additional features, such as sequencing binding sites, indices, and sections complementary to the flow cell oligos, are added by decreased cycle amplification. Illumina's HiSeq and MiSeq systems are used for sequencing (www.illumina.com/technology/next-generation-sequencing.html).

2.6 Genome assembly and annotation

Since the invention of the Sanger sequencing process, experts all over the world have focused their efforts on making advancements in the sector in order to provide the most cutting-edge technology. The development of next-generation sequencing (NGS) is a game-changing breakthrough that promises to lead to significant improvements in our understanding of how nucleic acids work. Illumina sequencing is a type of high throughput sequencing that is used to investigate microorganisms in order to learn about drug resistance outbreaks and dissemination (Loman *et al.*, 2012). Different sequencing platforms are used to generate reads. To examine the raw reads, a fastqc report was generated to determine the quality of the reads; it requires less

memory to run and allows for more flexible HTML report display (Brandine *et al.*, 2019). Next-generation sequencing technology generated a large amount of data, which could be long or short reads depending on the sequencing technique utilized. Because read mistakes, low quality reads, and primer / adaptor contamination are relatively common in the data, the NGS QC Toolkit is utilized for filtering and trimming the huge quantity of data created and employed in Quality check and filtration is necessary. It may accept sequencing data in a variety of formats as input and execute quality checks with default/user-defined parameters while it also provide QC reports for unfiltered (input) and filtered (output) data, as well as filtered HQ data files, are created in a variety of formats (Patel *et al.*, 2012) (Figure 2.3).

Sequencing readings are converted into contiguous sequences using sequence similarity during genome assembly (Clum, 2018), and gaps between contigs are filled in by sequential joining to construct scaffolds. The graph methods utilised by de novo assembly include overlap – layout consensus, which shows an overlap graph, and De Bruijn graph approach, which shows a compact representation (Miller *et al.*, 2010). VELVET or SPADes are two genome assembly techniques that are employed.

The ribosomal RNA (rRNA) genes have a wide range of applications in bioinformatics due to their highly conserved sequences and prevalence across all genomes. RNAmmer and Barrnap are two methods for predicting rRNA genes that both use a Hidden Markov Model methodology (HMM) (Seemann, 2013).

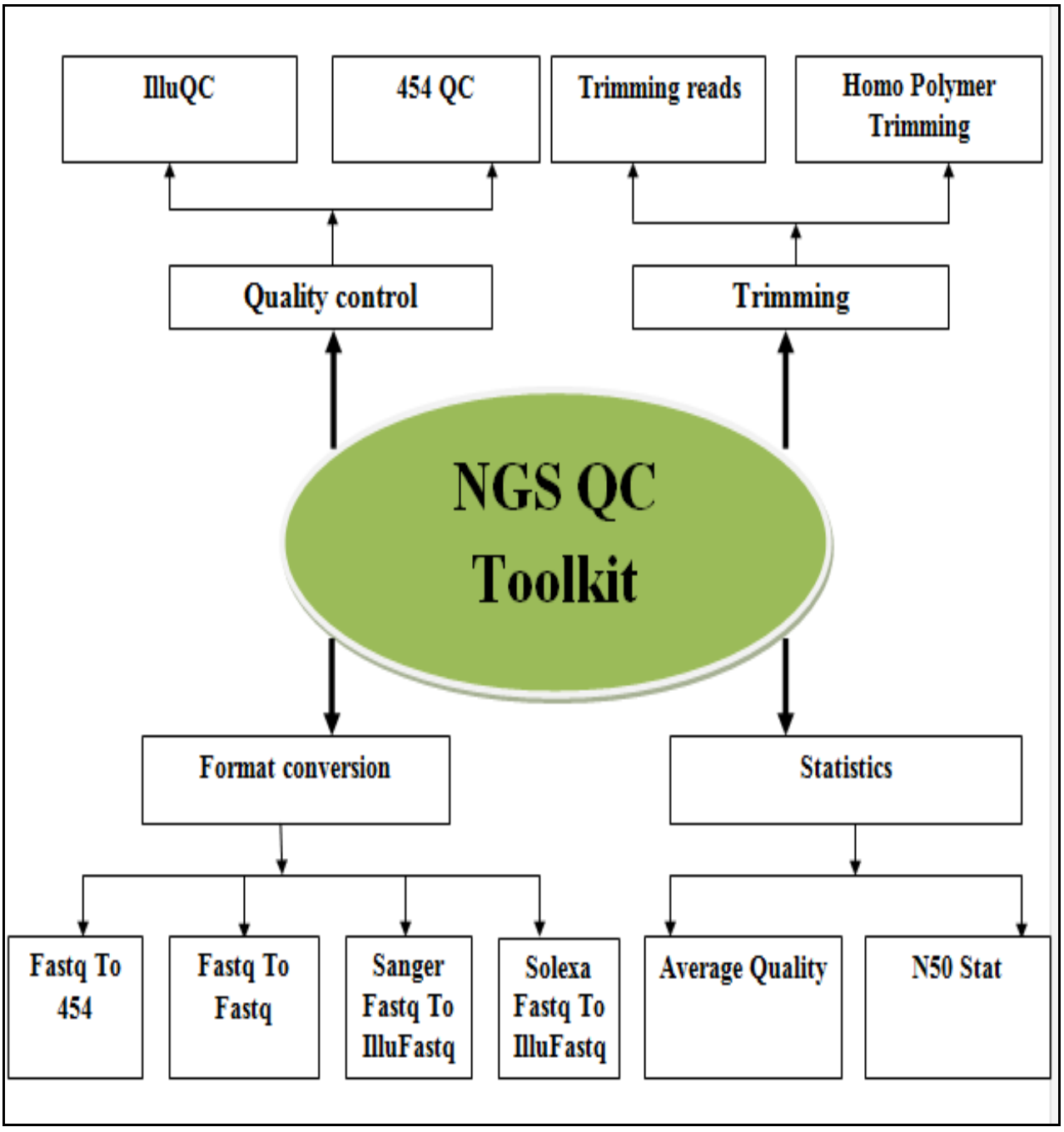


Figure 2.3 Various tools included in NGS QC Toolkit

In comparative, functional, and structural genomics, completing an organism's genome sequence is a critical endeavour. However, this is still a difficult problem to solve, both computationally and experimentally. De novo assembly genome scaffolding is frequently the initial step in most genome finishing workflows (Bosi *et al.*, 2015). Wet laboratory procedures, in silico approaches, or a combination of both are used to finish a genomic scaffold. The paired-read data from the sequencing stage could be used as an example of a computational method. The presence of paired reads in separate contigs can be utilised to determine the order and distance between these contigs probabilistically. MEDUSA scaffolded the same datasets in less than 10 minutes on average (Barton *et al.*, 2012). Contiguator: a bacterial genomes finishing tool for structural insights on draft genomes (Glardini *et al.*, 2011), Ragout-a reference-assisted assembly tool for bacterial genomes are just a few of the various ways for mapping (and later scaffolding) the generated draft contigs (Kolmogorov *et al.*, 2014).

BASys (Bacterial Annotation System) is a web service that permits for the automatic annotation of bacterial genomic (chromosomal and plasmid) sequences. It receives raw DNA sequence data as well as an optional list of gene identification information, and outputs comprehensive textual annotation and hyperlinked images, including gene/protein name, GO function, COG function, possible paralogues and orthologues, subcellular localization, signal peptides, transmembrane regions, secondary structure, 3D structure, reactions, and pathways (Gary *et al.*, 2005). Magnifying Genomes (MaGe) is a bacterial genome annotation system that incorporates a web interface for finishing the genome annotation tasks (Vallenet *et al.*, 2006).

KAAS (KEGG Automatic Annotation Server) is a programme that uses a fast approach to assign K values to genes in the genome, allowing KEGG pathways to be reconstructed (Moriya *et al.,* 2007). Genome annotation is a process that identifies the relevant features associated with specific genome sequences. A variety of tools are available for annotation, some of which are online web servers and others are command line based tools, such as RAST (Rapid Annotation Using SubsystemTechnology) (Aziz *et al.,* 2008), DIYA (Stewart *et al.,* 2009) [Figure 2.4].

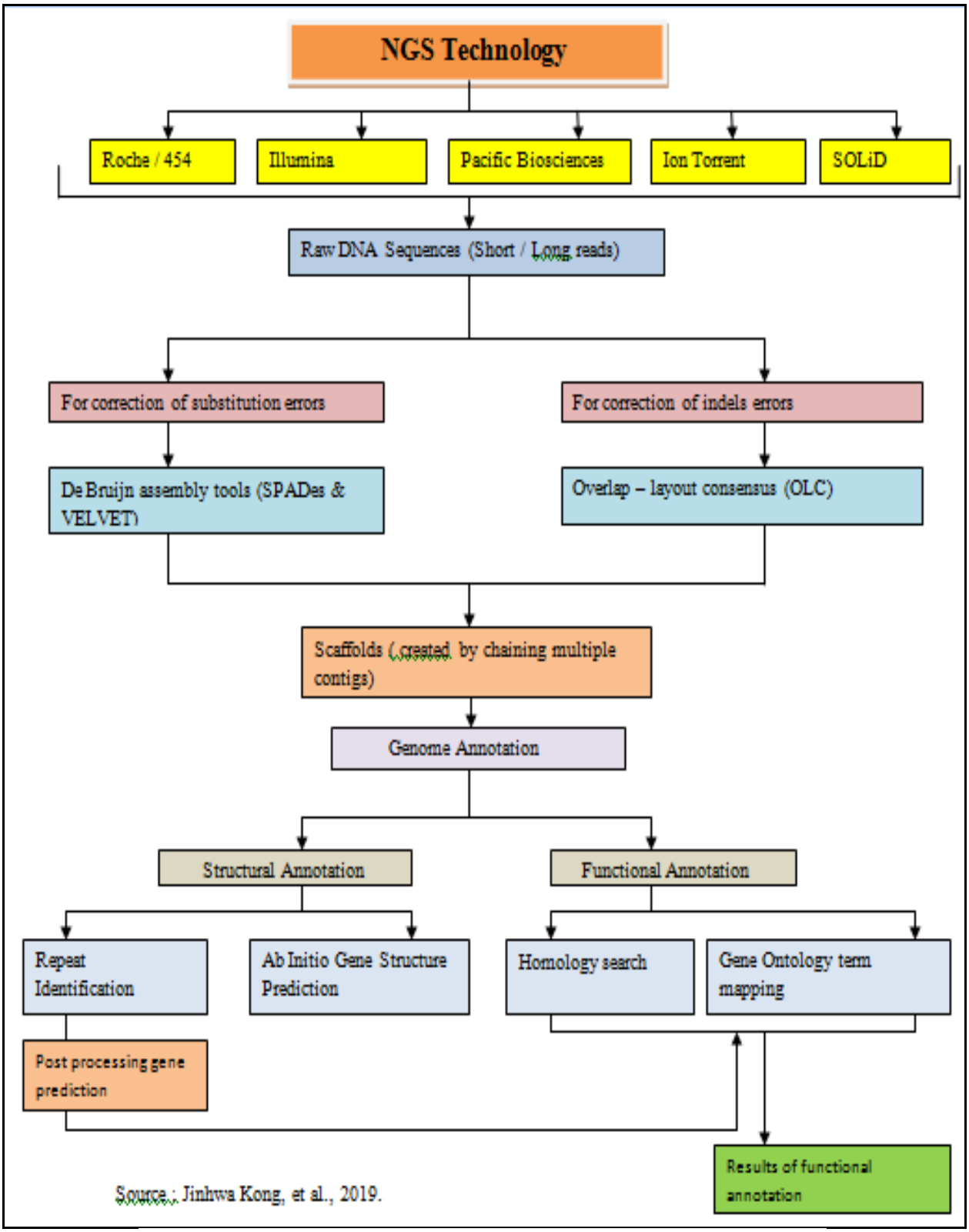


Figure 2.4 Workflow of genome assembly and genome annotation

2.7 Post annotation: Comparative analysis

Linear or circular representations are used in genome visualisation approaches. Genome search by DNA hybridization is included in linear representations such as those generated by Artemis Comparison Tool (ACT), which allows an interactive visualisation of comparisons between complete genome sequences and associated annotations (Carver *et al.,* 2005), Mauve (Darling *et al.,* 2004), and Genomorama (probe binding and PCR amplification). In displaying insertions and deletions between genomic sequences, efficient multi-scale display and manipulation of multiple genomes, support for many genome file types, and the ability to search for and retrieve data from the National Center for Biotechnology Information (NCBI) Entrez server (Gans *et al.,* 2007) have advantages, and certain programmes, such as Mauve and ACT, can show genome rearrangements. Mauve is software that can be used to identify and align conserved genomic DNA in the presence of horizontal transfer and rearrangements. Progressive Mauve generates numerous genomic alignments based on positional homology. Glocal genome alignments, in which all copies of a repeating gene family may become aligned to one other, are very different from these alignments. Downstream alignment tasks such as phylogenetic inference of nucleotide substitution, phylogenetic inference of gene gain and loss (Didelot *et al.,* 2008), phylogenetic inference of rearrangement and even inference of homologous recombination-induced lateral gene transfer (Darling *et al.,* 2008) are made easier with positional homology alignment (Didelot *et al.,* 2007). However, applying these tools to summarise huge datasets is problematic. Microbial Genome Viewer (Kerkhovan *et al.,* 2004) and Genome Projector (Arakawa *et al.,* 2009) are two programmes that generate circular figures. They are meant to annotate a single chromosome and do not support whole genome comparative data. BRIG (BLAST Ring Image Generator) is a cross-platform desktop tool that allows users to

quickly visualise BLAST comparisons to one or more central reference sequences utilising complete, draft, or unassembled genome data (Alikhan *et al.*, 2011).

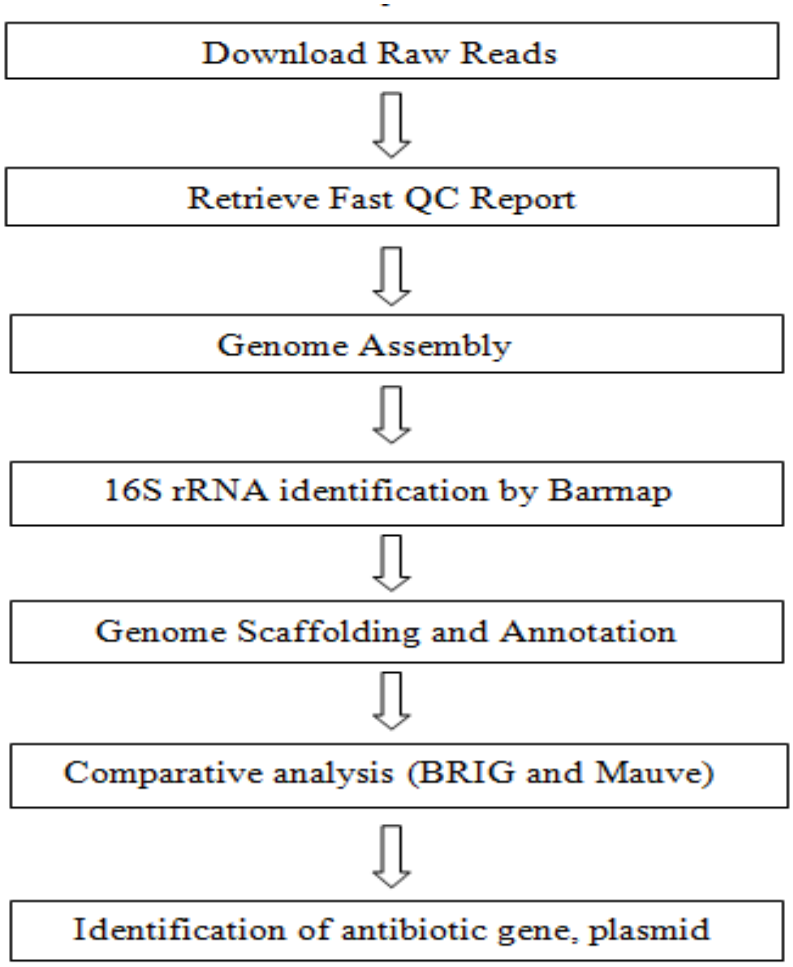
Unlike BRIG, comparable programmes limit the amount of genome comparisons that may be displayed on a single image and do not allow you to combine several sequences into a single lane. For most sequencing investigations, comparing to other genomes or sequences is a vital step. In many situations, users are also interested in trying to find specific genes with recognised roles, such as virulence genes or medication resistance determinants. Most users will need to be able to visualise these comparisons in order to aid understanding and interpretation of the data, as well as to obtain figures for conveying conclusions.

MATERIALS AND METHODS

CHAPTER 3: MATERIALS AND METHODS

This present study of **Studies on "Comparative genome analysis of *Paenibacillus macerans* CMB402, CMB401, CMB393"** was done through online mode at Rajiv Gandhi South Campus, Banaras Hindu University, Mirzapur, Uttar Pradesh.

3.1 Schematic workflow of this study



3.2 Bacterial Strain

Raw data for selected bacterial strain (*Paenibacillus macerans* CMB402, CMB401 and CMB393) were downloaded from European Nucleotide Archive site as shown in Table 3.1. This site is a sequence repository which includes three main databases that is Sequence Read Archive, Trace Archive and EMBL – Bank and also it promotes the use of nucleotide sequencing as an experimental research (Leinonen *et al.*, 2011).

Table 3.1: Selected Strain for genomic analysis and further downstream analysis.

S.No.	Strain	Sample	Run	Experiment	Read
1	CMB393	Sample has been submitted on 2021-01-04; *Paenibacillus macerans*	Illumina MiSeq sequencing; Bovine Rumen Microbiome isolates	Illumina MiSeq paired end sequencing; Bovine Rumen Microbiome isolates	SRR11410553_1.fastq SRR11410553_2.fastq
2	CMB401	Sample has been submitted on 2021-01-04; *Paenibacillus macerans*	Illumina MiSeq equencing; Bovine Rumen Microbiome isolates	Illumina MiSeq paired end sequencing; Bovine Rumen Microbiome isolates	SRR11410549_1.fastq SRR11410549_2.fastq
3	CMB402	Sample has been submitted on 2021-01-04; *Paenibacillus macerans*	Illumina MiSeq sequencing; Bovine Rumen Microbiome isolates	Illumina MiSeq paired end sequencing; Bovine Rumen Microbiome isolates	SRR11410548_1.fastq SRR11410548_2.fastq

3.3 Bioinformatics Tool for Quality Check

(i) Fast Quality Check

Fast QC accustomed fathom internal control on data coming back from high throughput sequencing technology. It gives a preliminary report into data quality. Prior to scanning the sequence to draw biological interpretation you want to forever perform some straightforward control checks to verify that the data information looks wise and there do not seem to be any problems or biases in your data that might have a bearing on but you will usefully use it. Many sequencers will produce a QC report as the location of their analysis pipelines (Andrews, 2010).

3.4 Bioinformatics Tool for Quality Control

(i) NGS QC Toolkit: Next Generation Sequencing Quality Control Toolkit

Raw reads are converted into filtered sequences using NGS QC Toolkit. NGS QC Toolkit used for control and filtering of next-generation sequence data and it generates elaborate finally end up within the form of tables and graphs at the aspect of filtering of the data to return up with high-quality sequence data and further it incorporates a spread of modules that facilitate varied totally different pre-processing like format conversion and trimming and statistics. After trimming the reads, low-quality reads area unit removed and reads containing primer/adaptor contamination area unit cut as per given criteria. Finally, High Quality (HQ) reads and QC statistics area unit generated within the output folder (Patel *et al.*, 2012).

3.5 Bioinformatics Tool for Genome Assembly

(i) SPAdes: Genome Assembler

After getting filtered files, genome assembly has to be done by using the SPAdes tool. It includes four stages that are: assembly graph construction; k-bimer adjustment using joint analysis of distance histograms and strategies among the assembly graph; constructs the paired assembly graph; contig construction (Bankevich *et al.*, 2012).

(ii) BARRNAP: BAsic Rapid Ribosomal RNA Predictor

Barrnap tool is employed to predicts the placement of rRNA genes in genomes whereas this tool supports bacteria (5S, 23S, 16S), archaea (5S, 5.8S, 23S, 16S), eukaryotes (5S, 5.8S, 28S, 18S). The largest sequenced 16S rRNA that is obtained from Barrnap is employed in BLAST (Seemann, 2013).

(iii) BLAST: Basic Local Alignment Search Tool

It is used to seek out the nearest reference genome. It is used for searching the sequences and for aligning them. It defines the connection between novel sequenced DNA and already submitted or identified sequences. Blast compare the query sequence to every sequence in a large database of sequences. It judges the degree of similarity and then withdraw all sequences that share small region of similarity to the query sequence (Boratyn *et al.*, 2013).

(iv) MEDUSA: Genome Finishing Tool

MeDuSa is employed for generating scaffolds from contigs. It is a multi draft primarily based scaffolder and exploits data obtained from a collection of genomes from connected organisms to work out the right order and orientation of the contigs (Bosi *et al.,* 2015). During this server, comparison ordination is needed that may be the nearest reference sequence that is obtained by victimization BLAST.

3.6 Bioinformatics server for downstream analysis

(i) RAST: Genome Annotation

It is a tool that helps in the identification of genes and also tells about their function. For annotation of sequence, RAST (Rapid Annotation Subsystem Technology) is utilized. It may be a fully automated service for step-up complete or nearly complete organism and archaeal genomes. It provides top of the range ordination annotations for these genomes across the complete phylogenetic tree. Users of the provision upload a genome as a set of contigs in FASTA format, and they receive ingress to an annotated genome in an environment that supports collation with an unification of hundreds of subsisting genomes. It makes a SEED-quality annotation out there as a service with 48 hours. The SEED environment and SEED information structures unit accustomed to reckon the machine-controlled annotations; however, knowledge is not extra into the SEED automatically. Once the annotation is completed, genomes are downloaded in an extremely variety of formats or viewed online (Aziz *et al.,* 2008).

(ii) PHASTER: PHAge Search Tool Enhanced Release

It is used for the rapid identification and annotation of prophage (A bacteriophage genome that has been inserted and integrated into the circular bacterial DNA chromosome or that exists as an extrachromosomal plasmid) sequences within bacterial genomes and plasmids. Phaster has three input options, upload a GenBank formatted file or nucleotide sequence file, fasta format. Download the complete set of results in a zipped folder or view them in three files (A summary file, detailed file, and an interactive genome viewer) through which the results can be analyzed (Arndt *et al.*, 2016; Zhou *et al.*, 2011).

(iii) Plasmid Finder

It is a web tool for *in silico* discernment and characterization of whole plasmid sequence data. Plasmid Finder can discern an extensive variety of plasmids that are frequently related with antimicrobial resistance. It is based on a curated database of plasmids replicons intended for the identification of plasmids in whole genome sequences originating from Paenibacillus species using direct high throughput raw reads, assembled contigs. Upon sequence submission, a percent identity of 100%, 95%, or on down to 50% can be selected (Caratfoli *et al.*, 2014).

(iv) CARD (Comprehensive Antibiotic Resistance Database):

The CARD ordered by the Antibiotic Resistance Ontology (ARO) and Antimicrobial Resistance (AMR) factor discernment models that may be a precisely curated variety of peer-reviewed resistance factor and related antibiotics. The CARD comprise tools for inspection of molecular sequences and also the Resistance Gene Identifier (RGI) code for prediction of genes that confirm resistance to antibiotics that support the similarity and SNP models. CARD

information and ontologies will be downloaded in an exceeding variety of formats (McArthur *et al.,* 2013).

3.7 Bioinformatics Tool for Comparative Analysis

(i) MAUVE: For Contigs Alignment

Mauve could be a software that tries to align sequences in different species that evolve from common ancesteral sequences and xenologous regions of sequences that have gone through each local and large-scale change. As results of recombination can cause genomic sequences rearrangements, orthologous regions of one ordering sequence might even be reordered or inverted relative to special genomic sequences. All over the alignment strategy, Mauve point out preserved segments that seem to be internally free from genome sequences rearrangements. Such region's area units brought up as Locally Collinear Blocks (LCBs) (Darling *et al.*, 2010).

(ii) BRIG: BLAST Ring Image Generator

Visualizing an organism order as a circular image has become a robust means that of displaying informative comparisons of one sequence to a variety of others. BRIG will generate pictures that show multiple organism order comparisons, while not associate arbitrary limits on the number of genomes compared. The output image shows the similarity between a central reference sequence and different sequences as a collection of coaxial rings, wherever BLAST matches area unit colored on a wage schedule indicating an outlined share identity. Pictures can even embrace draft order assembly information to point out browse coverage, assembly breakpoints and folded repeats. BRIG is quickly accessible to any user; because it assumes no

specialist process information and can perform all needed file parsing and BLAST comparisons mechanically. It's a graphical interface programmed on the Swing framework, that takes the user piecemeal through the generation of a circular image. The settings accustomed to generate a specific image are saved for re-use with completely different order information, or the complete session is bundled and saved for later. The image is generated in JPEG, PNG, SVG, SVGZ format (Alikhan *et al.*, 2011).

3.8 Bioinformatics software, tools and their web addresses for downloading purposes

Software programs that are designed for extracting useful information from the sequences and their URL for downloading all above software, database and tools, refer Table no. 3.2.

Table 3.2: URL of different Databases, Software and Tools

Databases / Software / Tools	**URL**
European Nucleotide Archive	http://www.ebi.ac.uk/ena
NGS QC Toolkit	http://www.nipgr.res.in/ngsqctoolkit.html
SPAdes	http://bioinf.spbau.ru/spades
Barrnap	https://github.com/tseemann/barrnap
Medusa	http://combo.dbe.unifi.it/medusa
RAST	https://rast.nmpdr.org
PHASTER	https://phaster.ca
PLASMID FINDER	https://cge.cbs.dtu.dk
CARD	https://card.mcmaster.ca
Mauve	http://darlinglab.org/mauve/mauve.html
BRIG	http://sourceforge.net/projects/brig/
BLAST	http://blast.ncbi.nlm.nih.gov
Fast QC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

RESULTS

CHAPTER 4: RESULTS

The results of study entitled **Studies on "Comparative genome analysis of *Paenibacillus macerans* CMB402, CMB401 and CMB393"** such as quality control, validation of assembly data, prophage and plasmid identification, antibiotic resistance gene identification and comparative genome analysis of three different strains of *Paenibacillus macerans* are discussed below.

4.1. Sample Collection

Raw reads of three different strains of *Paenibacillus macerans* were downloaded from European Nucleotide Archive site [Table 4.1].

Table 4.1. Raw Reads for genomic analysis

Strains of *Paenibacillus macerans*	**CMB402**	**CMB401**	**CMB393**
Forward raw reads	SRR11410548_1.fastq	SRR11410549_1.fastq	SRR11410553_1.fastq
Reverse raw reads	SRR11410548_2.fastq	SRR11410549_2.fastq	SRR11410553_2.fastq

4.2. Fast Quality Check Report

Fast QC was used to check the quality of reads of three different strains of *Paenibacillus macerans*. Total sequences are different in reads of three different strains. There are 1196964 sequences present in CMB402, 1228980 sequences present in CMB401 and 1410585 sequences present in CMB393. Percentage of GC content is almost similar in all three strains that is 53-54% [Table 4.2].

Table 4.2. Basic Statistics of the reads after Fast QC

Measure	CMB402		CMB401		CMB393	
File name	SRR11410548_1 .fastq.gz	SRR11410548_2 .fastq.gz	SRR11410549_1 .fastq.gz	SRR11410549_2 .fastq.gz	SRR11410553_1 .fastq.gz	SRR11410553_2 .fastq.gz
File type	Conventional base calls	Conventional base calls	Conventional base calls	Conventional base calls	Conventional base calls	Conventional base calls
Encoding	Sanger / Illumina 1.9	Sanger / Illumina 1.9	Sanger / Illumina 1.9	Sanger / Illumina 1.9	Sanger / Illumina 1.9	Sanger / Illumina 1.9
Total Sequences	1196964	1196964	1228980	1228980	1410585	1410585
Sequence length	35-301	35-301	35-301	35-301	35-301	35-301
%GC	53	54	53	54	53	53

4.3. NGS QC Report

CMB402 strain showed higher cut – off quality score that is 38 than other two strains that is 30. CMB401 and CMB393 strain showed 99.9% inferred base call accuracy and probability of incorrect base call was 1 in 1000 [Table 4.3].

Table 4.3. Characterization of sequencing data

Strain	CMB402	CMB401	CMB393
Library File	Paired End	Paired End	Paired End
Input File	SRR11410548_1.fastq SRR11410548_2.fastq	SRR11410549_1.fastq SRR11410549_2.fastq	SRR11410553_1.fastq SRR11410553_2.fastq
Primer / Adaptor Library	Paired End DNA Library	Paired End DNA Library	Paired End DNA Library
Cut-off read length for HQ	70%	80%	80%
Cut-off Quality score	38	30	30
Only Statistics	Off	Off	Off
Number of CPUs	2	2	2

4.4.1. Before using NGS QC Toolkit

The reverse reads of CMB393 strain show higher total number of non-ATGC bases that are 111237 because of that it show higher polymorphism and forward reads of CMB401 strain show lesser total number of non-ATGC bases that are 33302 because of that it show lower polymorphism. The percentage of high quality reads are higher in genome sequence of CMB402 strain that is 57.93 and percentage of high quality reads are lesser in genome sequence of CMB393 strain that is 29.82 [Table 4.4].

Table 4.4. QC Statistics Before Using NGS QC Toolkit

Strain	**CMB402**		**CMB401**		**CMB393**	
File Name	SRR11410548 _1.fastq	SRR11410548 _2.fastq	SRR11410549 _1.fastq	SRR11410549 _2.fastq	SRR11410553 _1.fastq	SRR11410553 _2.fastq
Percentage of HQ Reads	57.93%	57.93%	29.92%	29.92%	29.82%	29.82%
Percentage of HQ Bases in HQ Reads	97.09%	83.24%	97.51%	89.24%	97.45%	89.48%
Average Read Length	269.85	270.69	275.02	276.76	275.53	276.46
Percentage of Reads with Non-ATGC Bases	0.23%	0.47%	0.17%	0.40%	0.23%	0.20%
Non-ATGC Bases	55879	62521	33302	39564	56668	111237

4.4.2. After using NGS QC Toolkit

After filtration and trimming, the reverse reads of CMB402 strain show higher total number of non-ATGC bases that are 2198 because of that it show higher polymorphism and reverse reads of CMB393 strain show lesser total number of non-ATGC bases that are 124 because of that it show lower polymorphism [Table 4.5].

Table 4.5. Detailed QC Statistics after Using NGS QC Toolkit

Strain	**CMB402**		**CMB401**		**CMB393**	
File Name	SRR11410548_1.fastq_filtered	SRR11410548_2.fastq_filtered	SRR11410549_1.fastq_filtered	SRR11410549_2.fastq_filtered	SRR11410553_1.fastq_filtered	SRR11410553_2.fastq_filtered
Total Number of Reads	693386	693386	157	497	420573	420573
Total Number of Reads with Non-ATGC Bases	347	1213	0.04%	0.14%	118	38
Total Number of HQ Bases	168400746	144670756	81146630	74307457	92264592	84769828
Non-ATGC Bases	487	2198	237	714	135	124

4.5. Quality control data from NGSQC Toolkit

4.5.1. Per base average quality score

4.5.1.1. CMB402 strain

The average quality score is the chance that the given base is called incorrectly by the sequencer. While increasing base position, the average quality score of raw reads and filtered reads decreases, and because of that it limits the length of high quality reads. From base position 1, the average quality score of raw reads and filtered reads of forward reads are 34 but at base position 5, its average quality score increases to 38 and then decreases [Figure 4.1.1]. For reverse reads, from base position 1for filtered reads show a higher average quality score that is 34 than raw reads that is 33 [Figure 4.1.2].

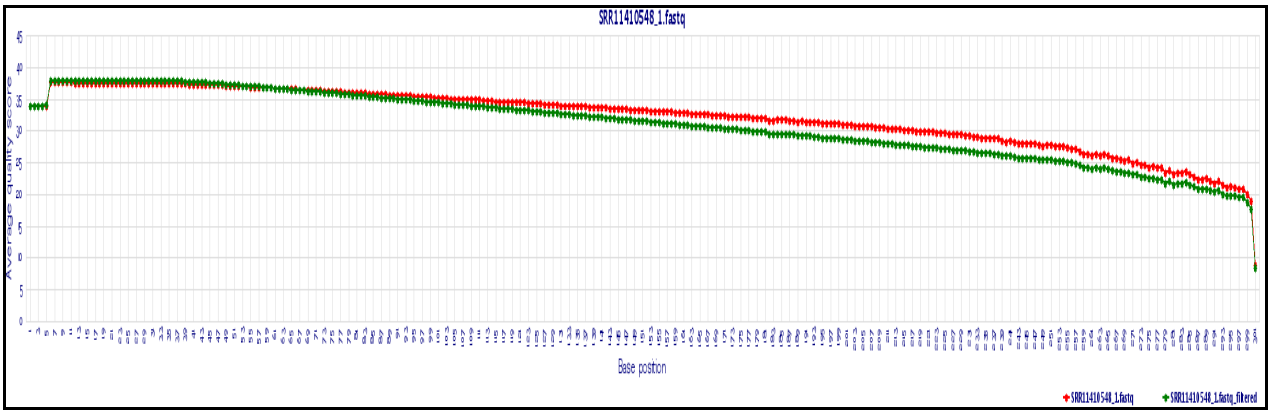


Figure 4.1.1. Per base average quality scores for input file, SRR11410548_1.fastq before and after QC of CMB402 strain.

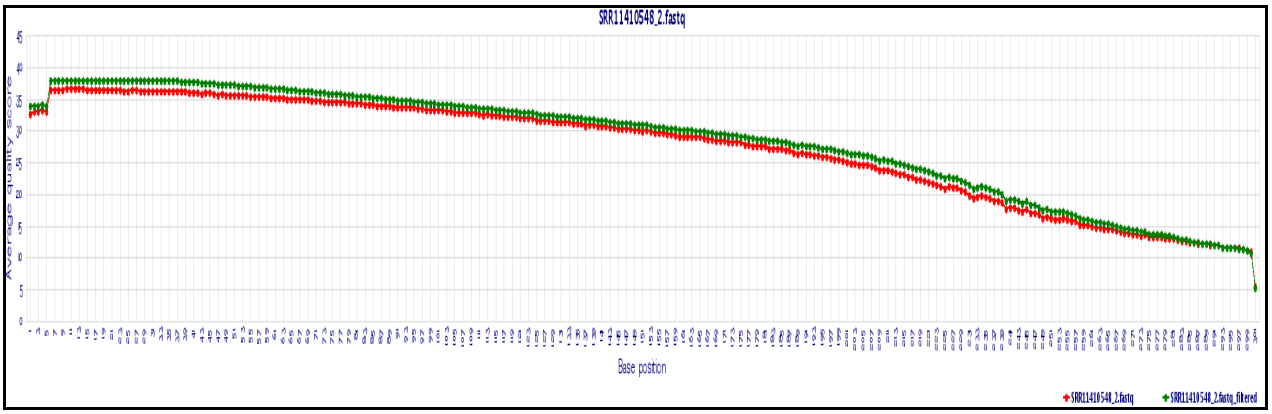


Figure 4.1.2. Per base average quality scores for input file, SRR11410548_2.fastq, before and after QC of CMB402 strain.

4.5.1.2. CMB401 strain

The average quality score relative to the base position of raw reads and filtered reads of forward and reverse reads while from base position 1, the average quality score of raw reads and filtered reads of forward reads are 34 but at base position 5, its average quality score increases to 38 and then decreases [Figure 4.2.1 and 4.2.2].

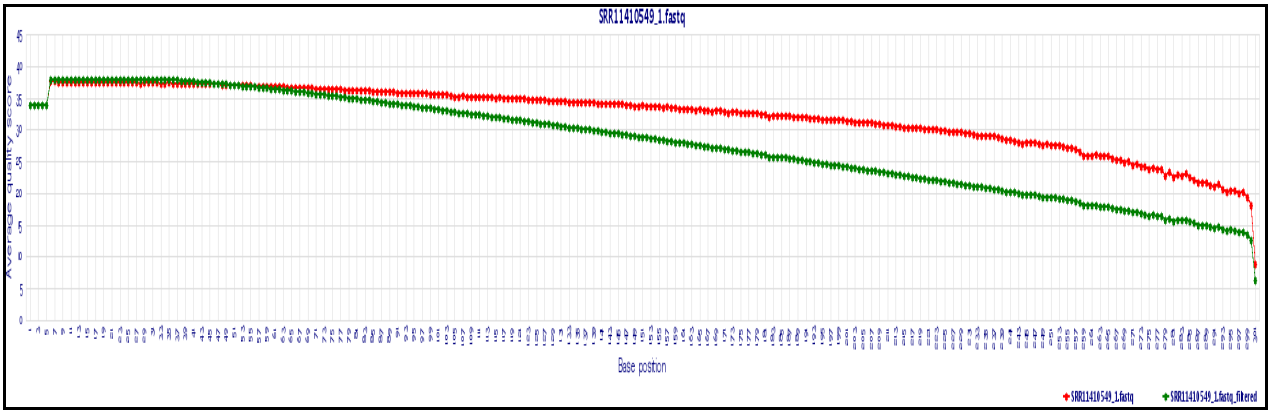


Figure 4.2.1. Per base average quality scores for input file, SRR11410549_1.fastq before and after QC of CMB401 strain.

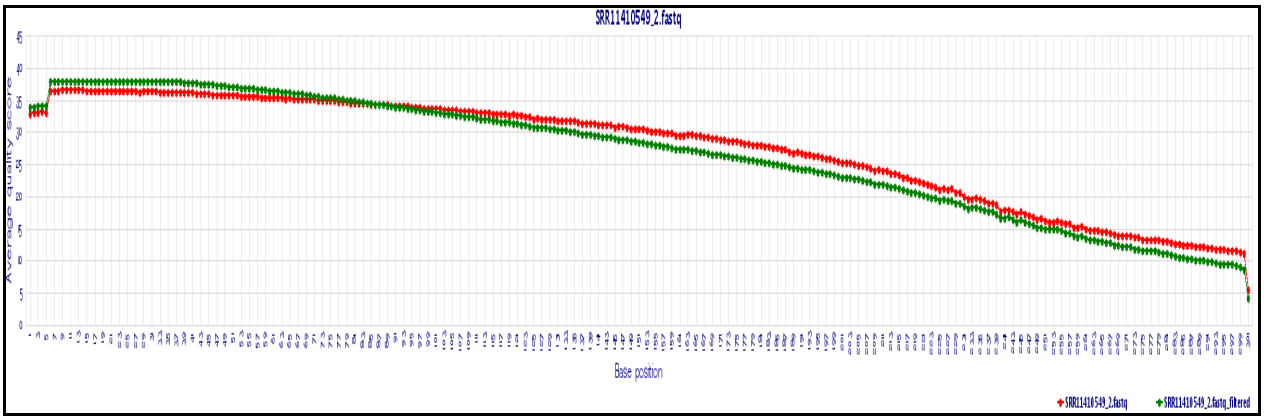


Figure 4.2.2. Per base average quality scores for input file, SRR11410549_2.fastq before and after QC of CMB401 strain.

4.5.1.3. CMB393 strain

At base position 1, raw and filtered reads both show the same average quality score that is 34 in forward reads but in reverse reads both show different average quality score that for filtered reads it is 34 and for raw reads it is 33 [Figure 4.3.1 and 4.3.2].

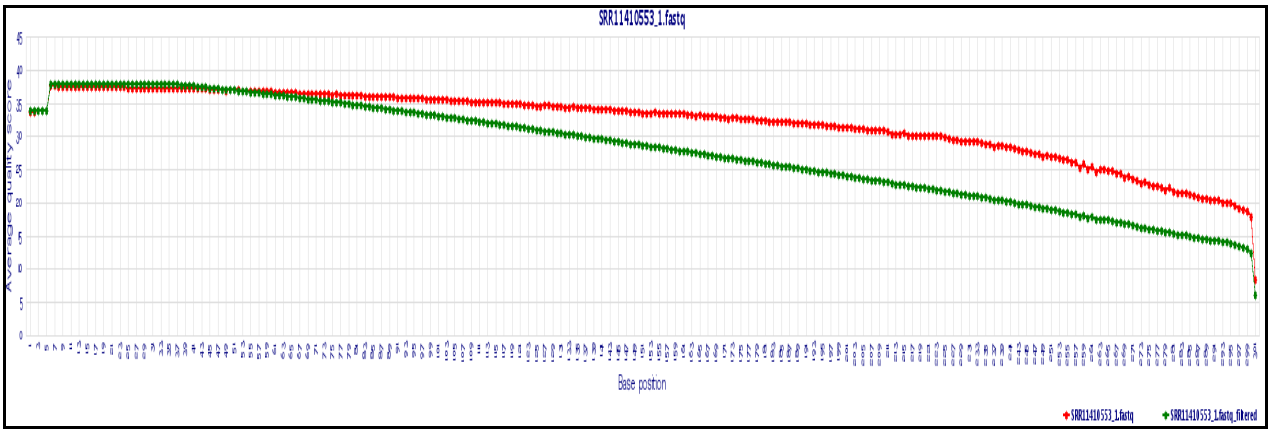


Figure 4.3.1. Per base average quality scores for input file, SRR11410553_1.fastq before and after QC of CMB393 strain.

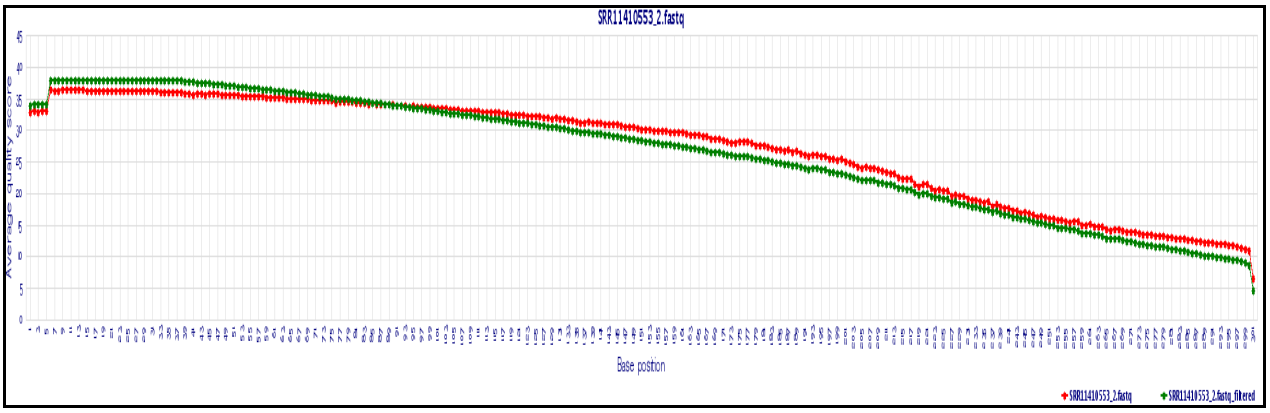


Figure 4.3.2. Per base average quality scores for input file, SRR11410553_2.fastq before and after QC of CMB393 strain.

4.5.2. Read count (%) per base for different quality ranges

4.5.2.1. CMB402 strain

At base position 1, read count percentage is nearly 100 for quality score between 31 – 40 for forward read [Figure 4.4.1]. At base position 1, read count percentage is approximately 99 for quality score between 31 – 40 for reverse read [Figure 4.4.2].

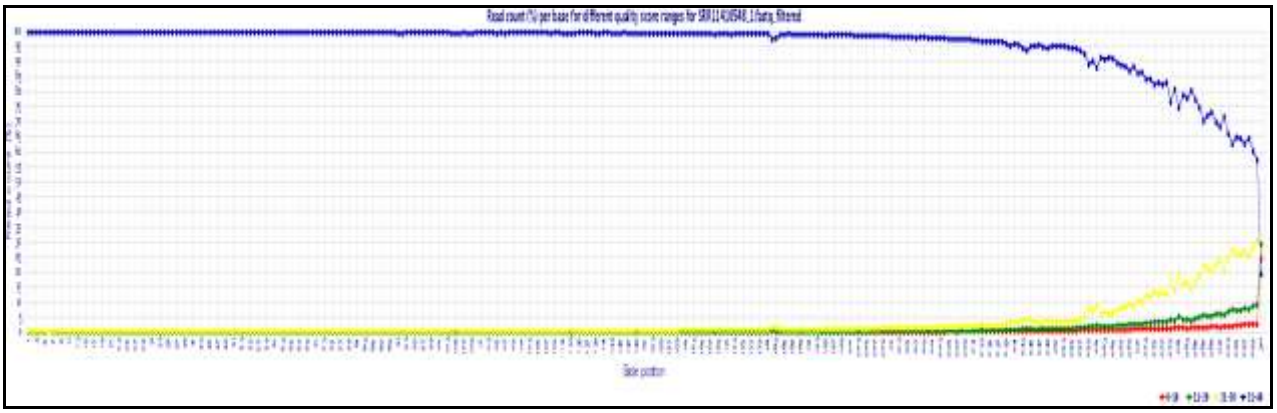


Figure 4.4.1. Read count(%) per base for SRR11410548_1.fastq, before and after QC of CMB402 strain.

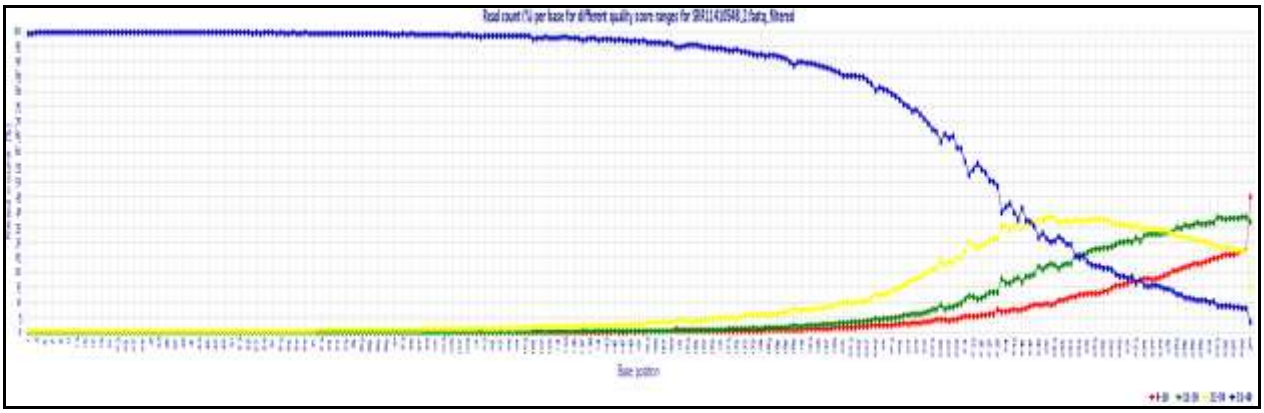


Figure 4.4.2. Read count(%) per base for SRR11410548_2.fastq, before and after QC of CMB402 strain.

4.5.2.2. CMB401 strain

At base position 1, read count percentage is approximately 99 for quality score between 31 – 40 for forward read [Figure 4.5.1]. At base position 1, read count percentage is approximately 100 for quality score between 31-40 for reverse read [Figure 4.5.2].

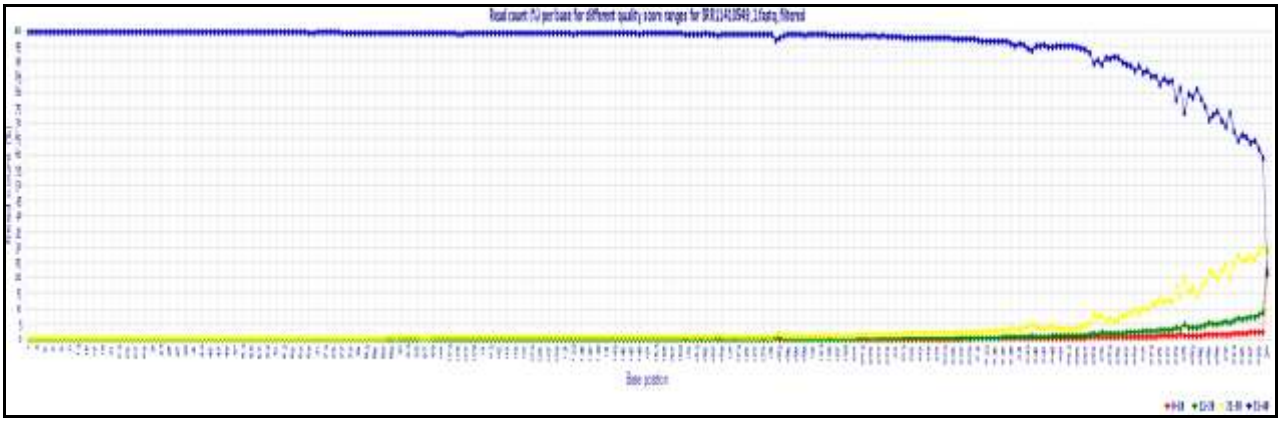


Figure 4.5.1. Read count(%) per base for SRR11410549_1.fastq, before and after QC of CMB401 strain.

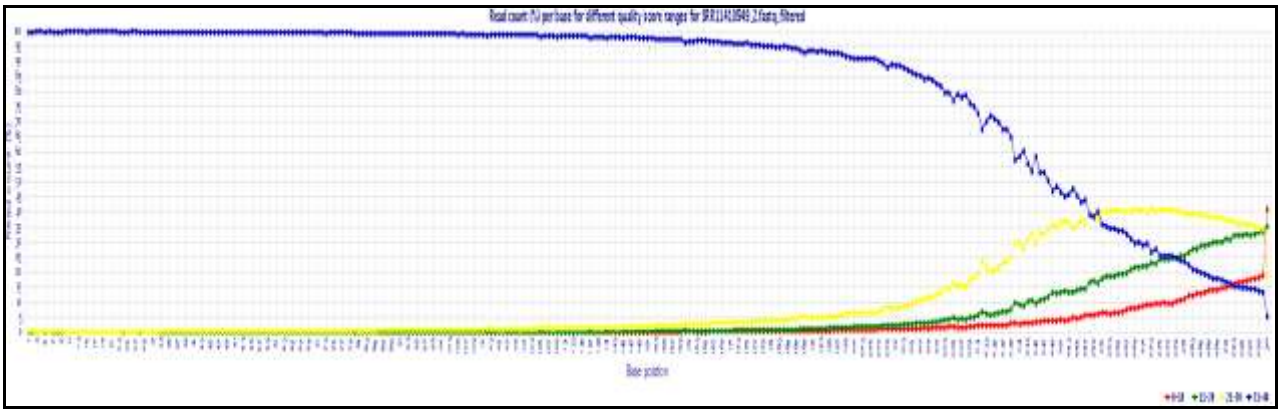


Figure 4.5.2. Read count(%) per base for SRR11410549_2.fastq, before and after QC of CMB401 strain.

4.5.2.3. CMB393 strain

At base position 1, read count percentage is approximately 99 for quality score between 31 – 40 for forward read [Figure 4.6.1]. At base position 1, read count percentage is approximately 100 for quality score between 31 – 40 for reverse read [Figure 4.6.2].

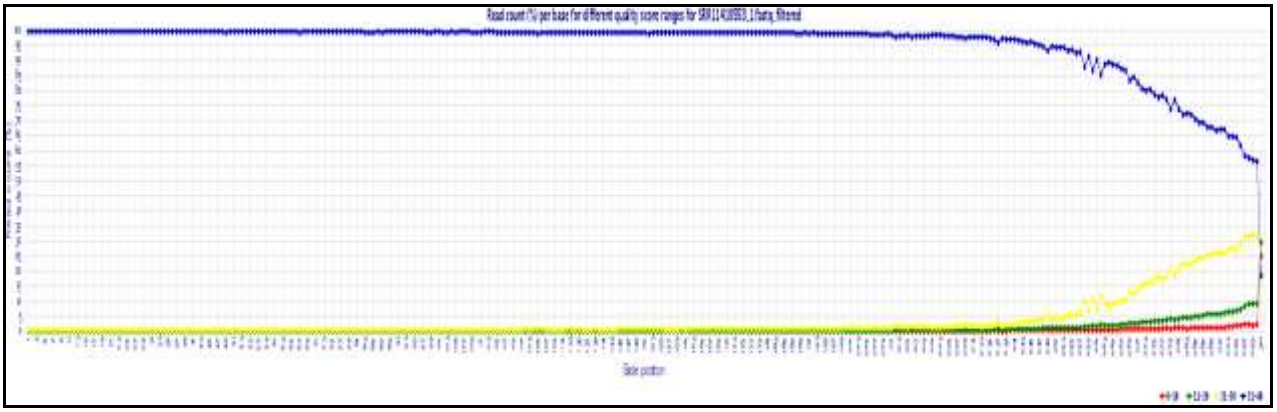


Figure 4.6.1. Read count(%) per base for SRR11410553_1.fastq, before and after QC of CMB393 strain.

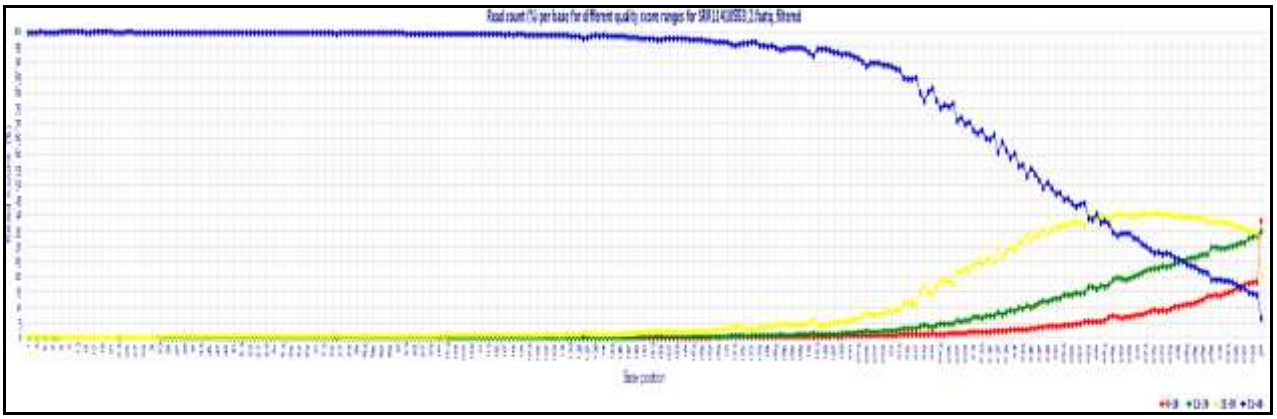


Figure 4.6.2. Read count(%) per base for SRR11410553_2.fastq, before and after QC of CMB393 strain.

4.5.3. Base composition

4.5.3.1. CMB402 strain

In forward and reverse both reads, after filtration or trimming base composition will change. Adenine composition is 23.34%, Thymine is 23.20%, Guanine is 26.86%, Cytosine is 26.58% and non – ATGC is 0.02% in raw reads but these percentages vary in filtered reads which represent a green bar in the graph for forward reads but this composition may vary for reverse reads [Figure 4.7.1 and 4.7.2].

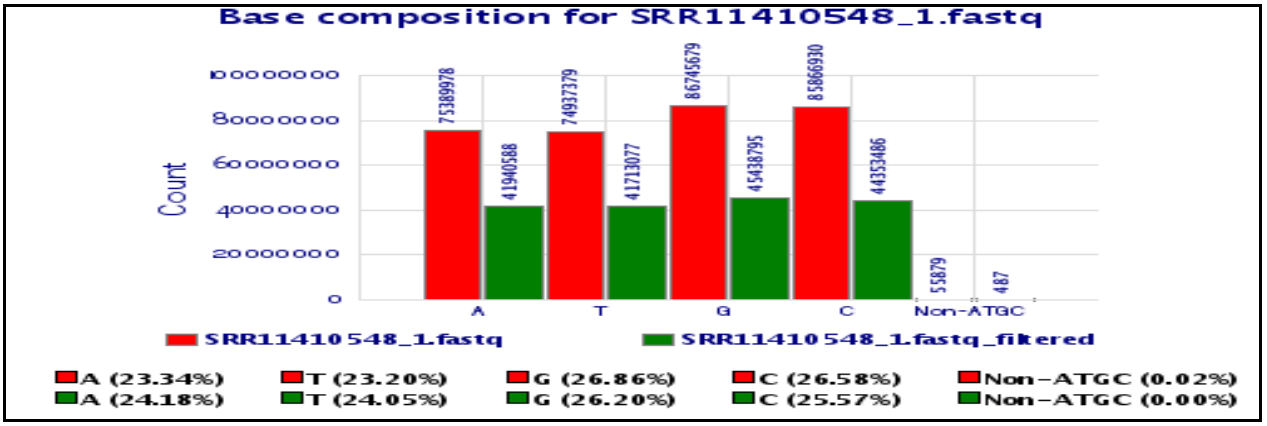


Figure 4.7.1. Base composition for input file, SRR11410548_1.fastq, before and after QC of CMB402 strain.

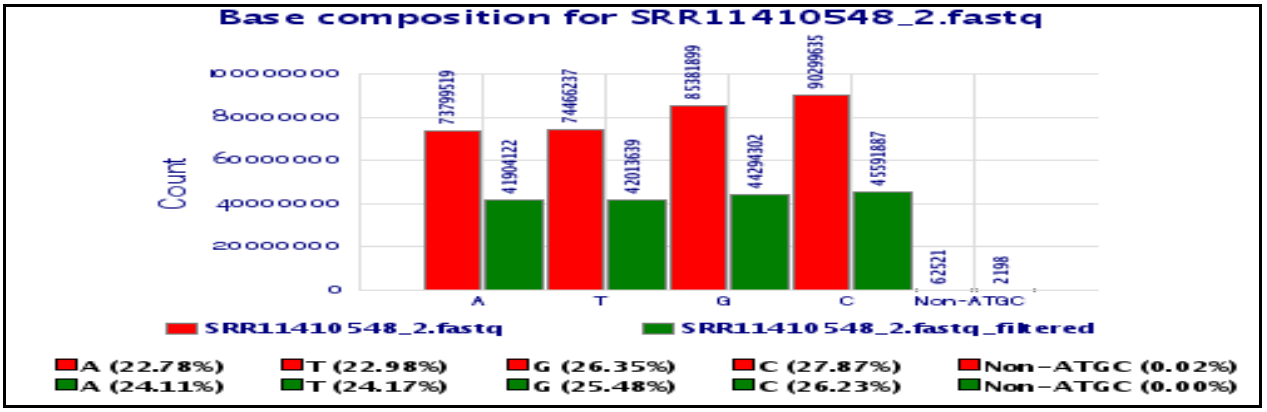


Figure 4.7.2. Base composition for input file, SRR11410548_2.fastq, before and after QC of

4.5.3.2. CMB401 strain

Percentage of four different base composition of raw reads and filtered reads for SRR11410549 raw data are, Adenine composition is 23.35%, Thymine is 23.18%, Guanine is 26.89%, Cytosine is 26.57% and non – ATGC is 0.01% in raw reads but these percentages vary in filtered reads that is Adenine composition is 25.26%, Thymine is 25.13%, Guanine is 25.22%, Cytosine is 24.39% and non – ATGC is 0.00% [Figure 4.8.1]. For reverse reads the adenine composition is 22.66%, thymine is 22.95%, guanine is 26.33%, cytosine is 28.05% and non – ATGC is 0.01% in raw reads but these percentages vary in filtered reads that is Adenine composition is 25.36%, Thymine is 25.42%, Guanine is 24.16%, Cytosine is 25.06% and non – ATGC is 0.00% [Figure 4.8.2].

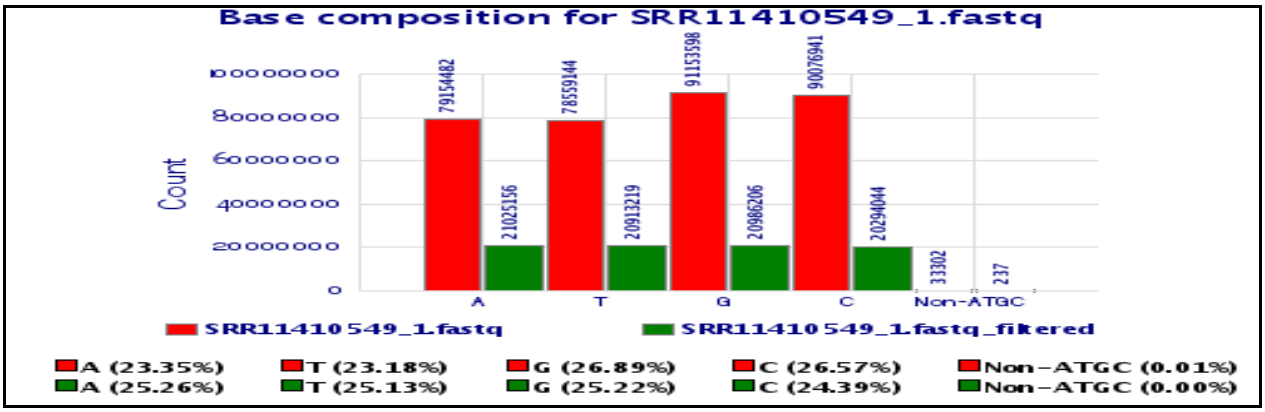


Figure 4.8.1. Base composition for input file, SRR11410549_1.fastq, before and after QC of CMB401 strain.

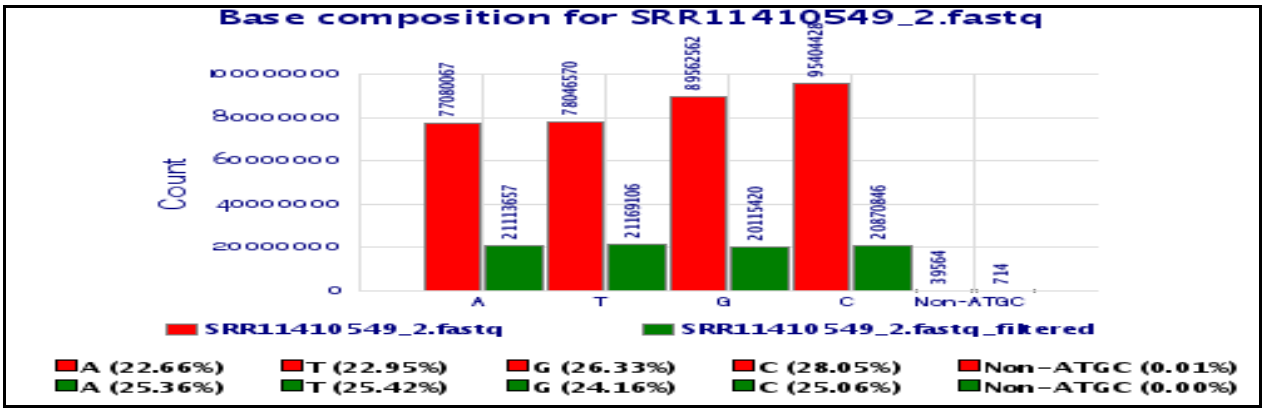


Figure 4.8.2. Base composition for input file, SRR11410549_2.fastq, before and after QC of CMB401 strain.

4.5.3.3. CMB393 strain

Percentage of four different base compositions (A,T,G,C) of raw reads and filtered reads for SRR11410553 raw data but both forward and reverse reads shows different composition of Non - ATGC of raw reads are, Adenine composition is 23.47%, Thymine is 23.34%, Guanine is 26.70%, Cytosine is 26.47% and non – ATGC is 0.01% in raw reads but these percentages vary in filtered reads that is Adenine composition is 25.40%, Thymine is 25.16%, Guanine is 25.10%, Cytosine is 23.34% and non – ATGC is 0.00% [Figure 4.9.1]. For reverse reads the adenine composition is 23.01%, thymine is 23.08%, guanine is 26.31%, cytosine is 27.58% and non – ATGC is 0.03% in raw reads but these percentages vary in filtered reads that is Adenine composition is 25.40%, Thymine is 25.56%, Guanine is 24.11%, Cytosine is 24.93% and non – ATGC is 0.00% [Figure 4.9.2].

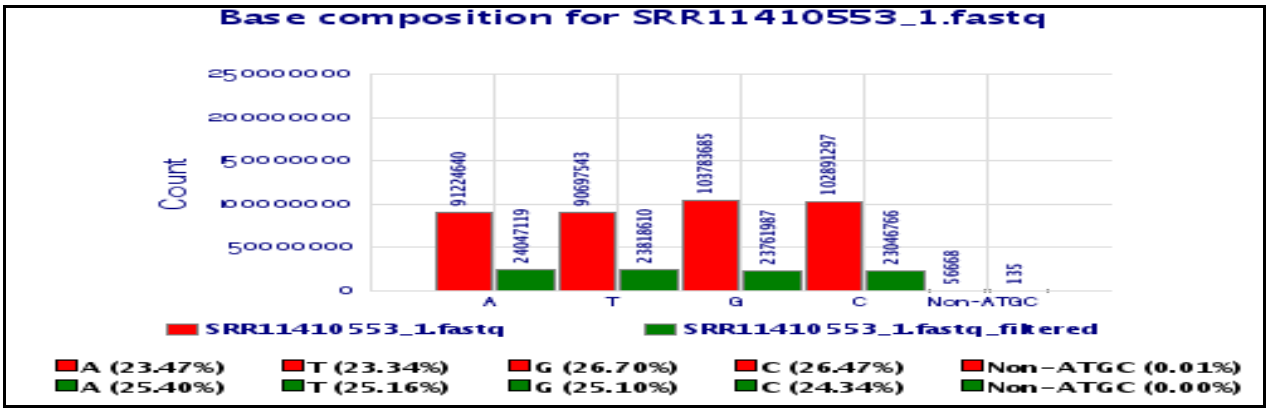


Figure 4.9.1. Base composition for input file, SRR11410553_1.fastq, before and after QC of CMB393 strain.

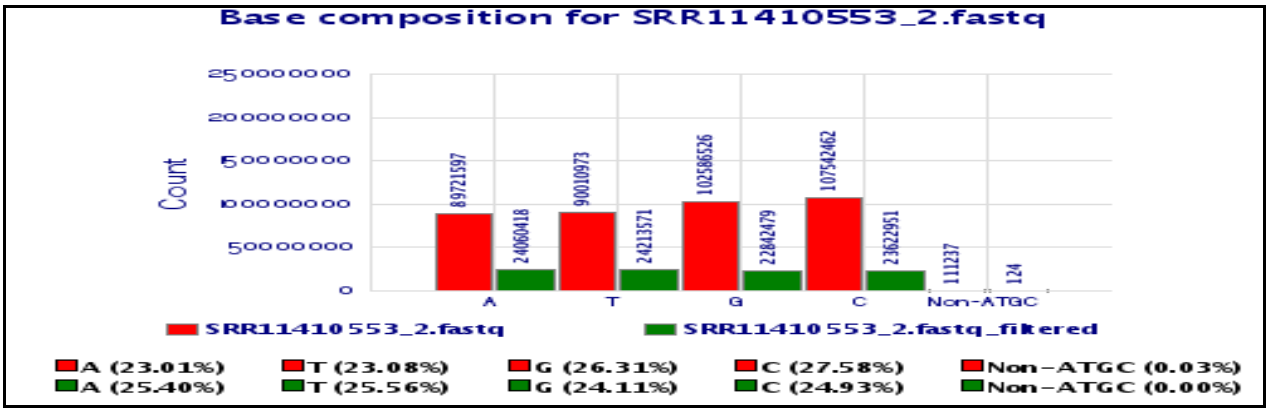


Figure 4.9.2. Base composition for input file, SRR11410553_2.fastq, before and after QC of CMB393 strain.

4.5.4. <u>GC content distribution</u>

4.5.4.1. CMB402 strain

Graph shows the GC content distribution in raw reads and filtered reads. For CMB402 strain, GC content distribution is almost the same in filtered reads that is 50 – 55% but varies in raw reads for both forward and reverse reads [Figure 4.10.1 and 4.10.2].

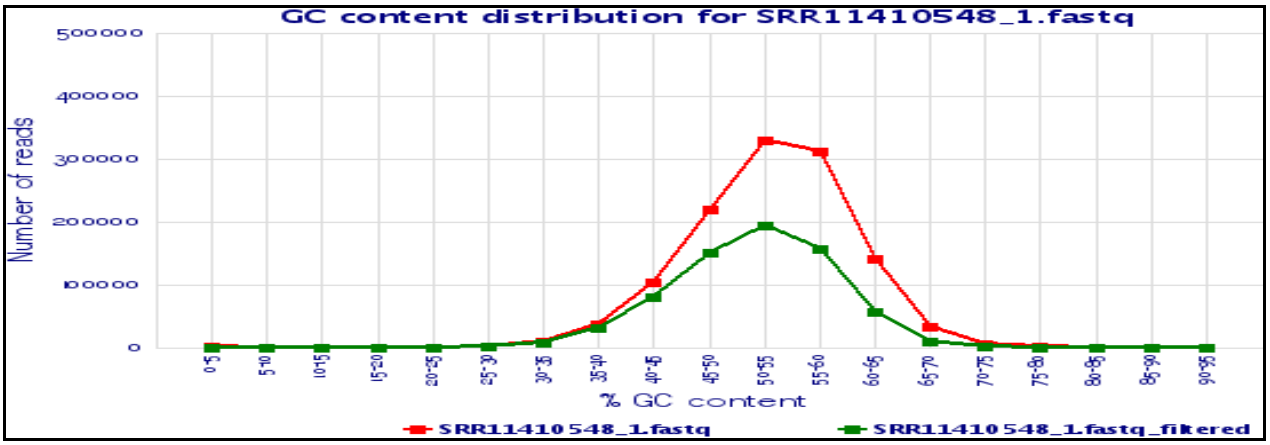


Figure 4.10.1. GC content distribution for input file, SRR11410548_1.fastq, before and after QC of CMB402 strain.

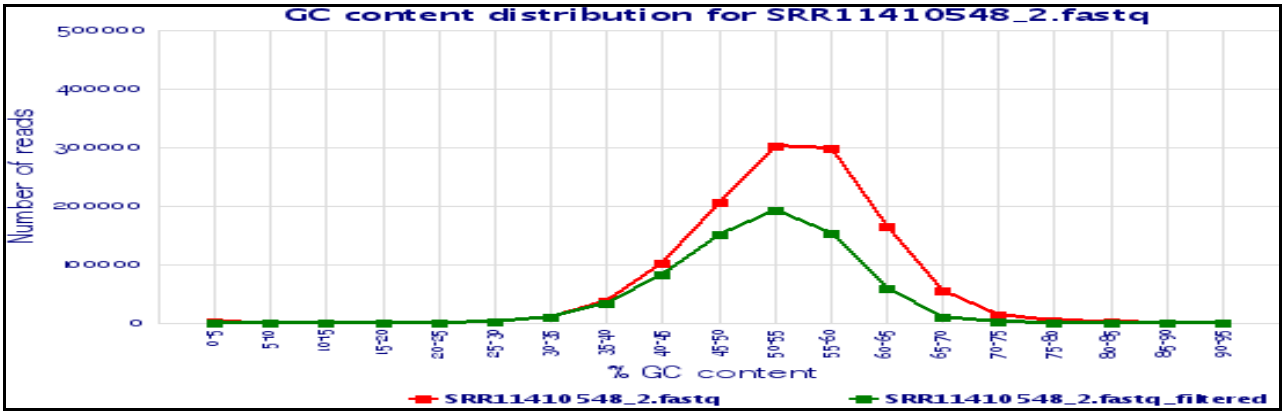


Figure 4.10.2. GC content distribution for input file, SRR11410548_2.fastq, before and after QC of CMB402 strain.

4.5.4.2. CMB401 strain

For CMB401 strain, for about 10,000 reads the percentage of GC content distribution is 45 – 55% for both forward and reverse reads of filtered reads but it varies for raw reads [Figure 4.11.1 and 4.11.2].

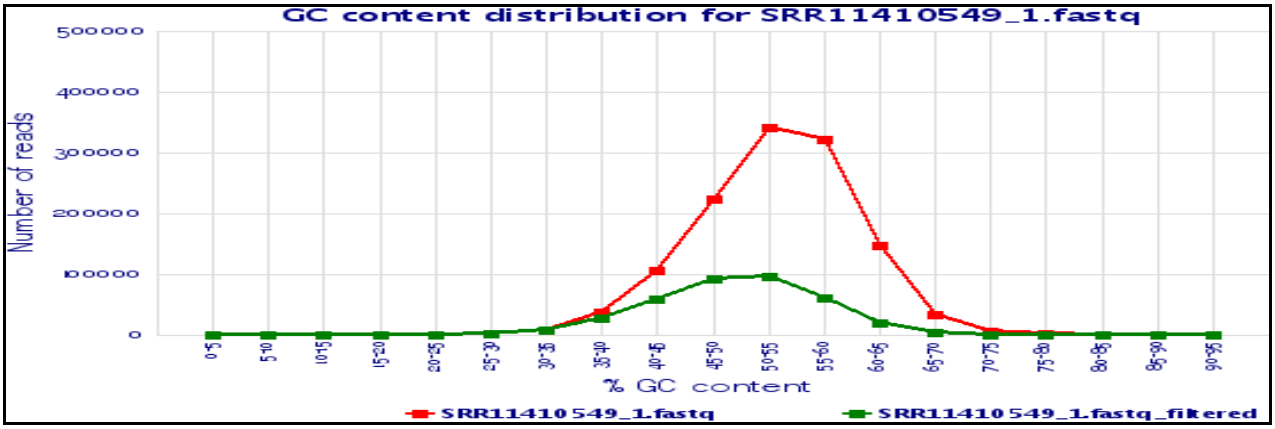


Figure 4.11.1. GC content distribution for input file, SRR11410549_1.fastq, before and after QC of CMB401 strain.

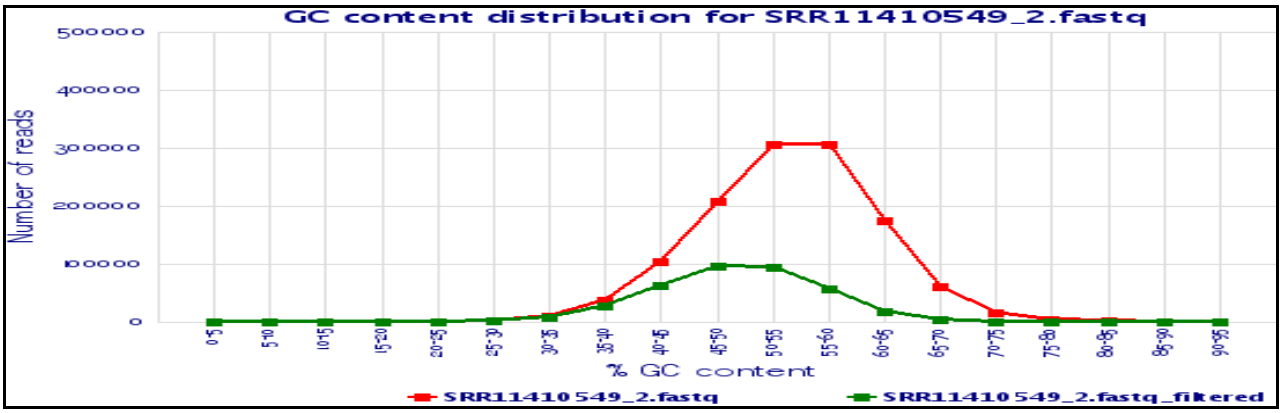


Figure 4.11.2. GC content distribution for input file, SRR11410549_2.fastq, before and after QC of CMB401 strain.

4.5.4.3. CMB393 strain

For CMB393 strain, the GC content distribution is 45 – 55% for both forward and reverse reads of filtered reads at or above 10000 reads [Figure 4.12.1 and 4.12.2].

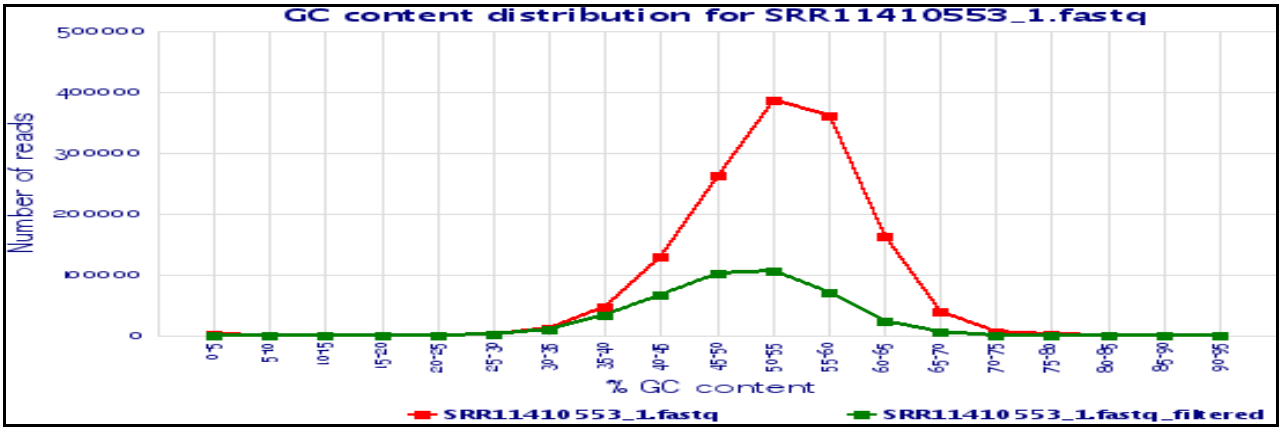


Figure 4.12.1. GC content distribution for input file, SRR11410553_1.fastq, before and after Q C of CMB393 strain.

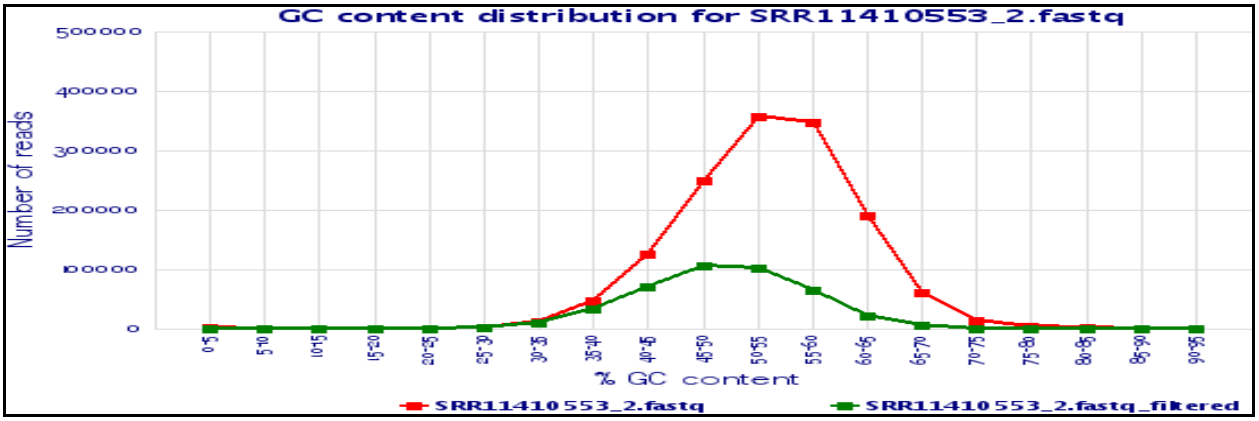


Figure 4.12.2. GC content distribution for input file, SRR11410553_2.fastq, before and after QC of CMB393 strain.

4.5.5. Quality distribution

4.5.5.1. CMB402 strain

Average phred quality score is 38 for an average 300000 reads for forward reads and for reverse reads it is 38 for an average above 100000 reads [Figure 4.13.1 and 4.13.2].

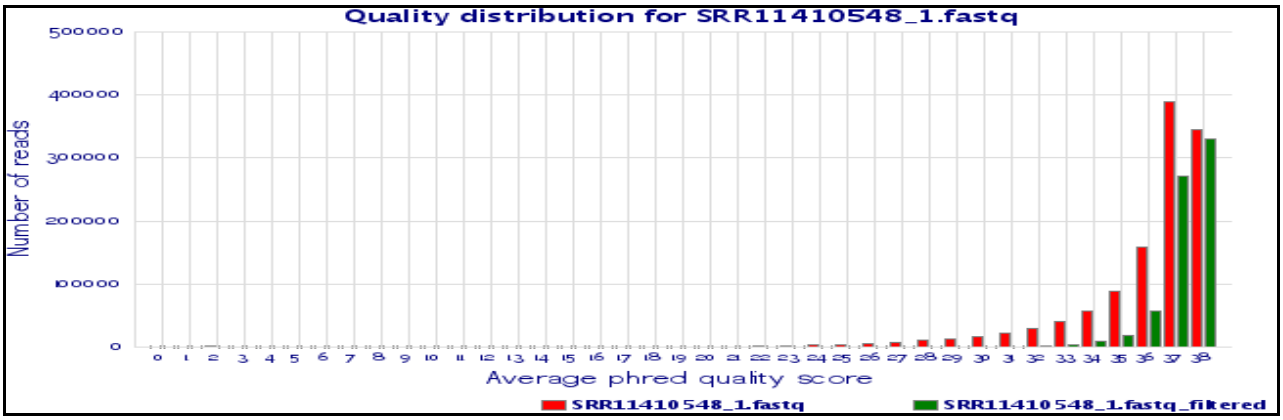


Figure 4.13.1. Quality distribution for input file, SRR11410548_1.fastq, before and after QC of CMB402 strain.

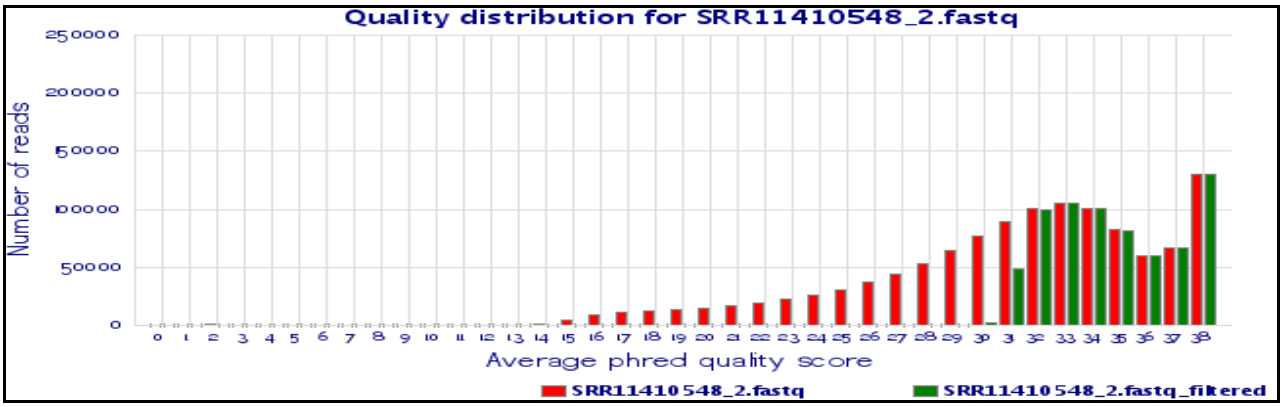


Figure 4.13.2. Quality distribution for input file, SRR11410548_2.fastq , before and after QC of CMB402 strain.

4.5.5.2. CMB401 strain

For CMB 401 strain, its forward reads consist 38 average phred quality score for less than 200000 reads and reverse reads consists 38 quality score for less than 100000 reads [Figure 4.14.1 and 4.14.2].

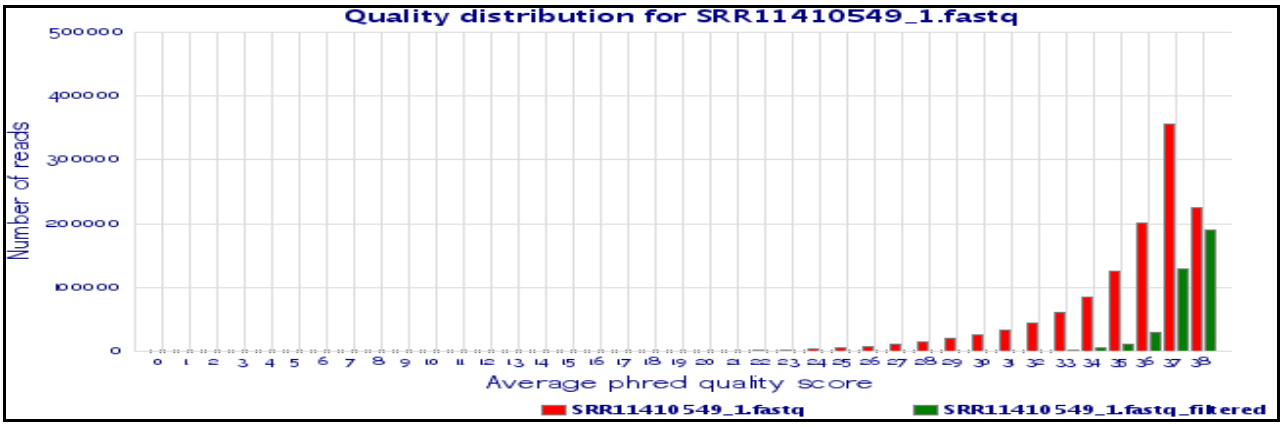


Figure 4.14.1. Quality distribution for input file, SRR11410549_1.fastq , before and after QC of CMB401 strain.

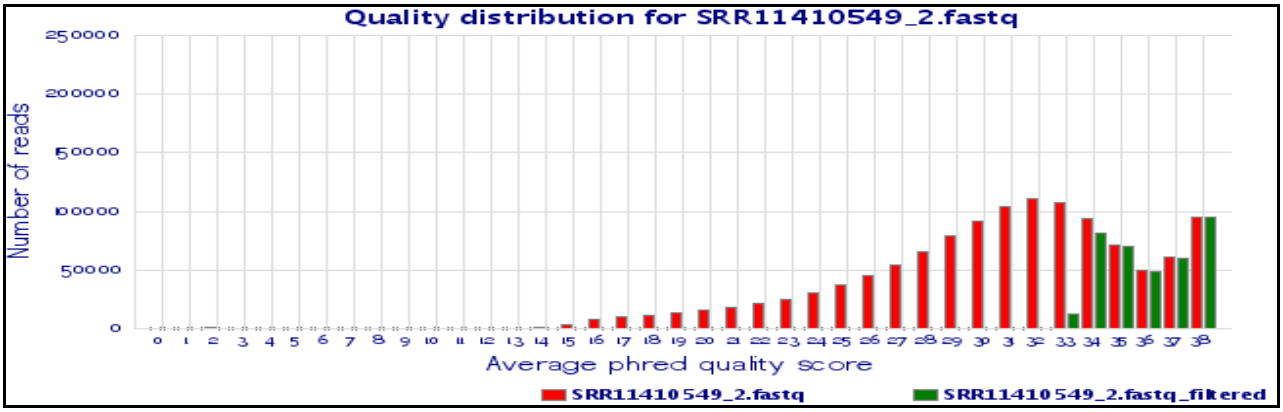


Figure 4.14.2. Quality distribution for input file, SRR11410549_2.fastq , before and after QC of CMB401 strain.

4.5.5.3. CMB393 strain

For CMB393 strain, its forward reads show 38 quality scores for above 200000 reads and reverse reads show 38 quality scores for above 100000 reads [Figure 4.15.1 and 4.15.2].

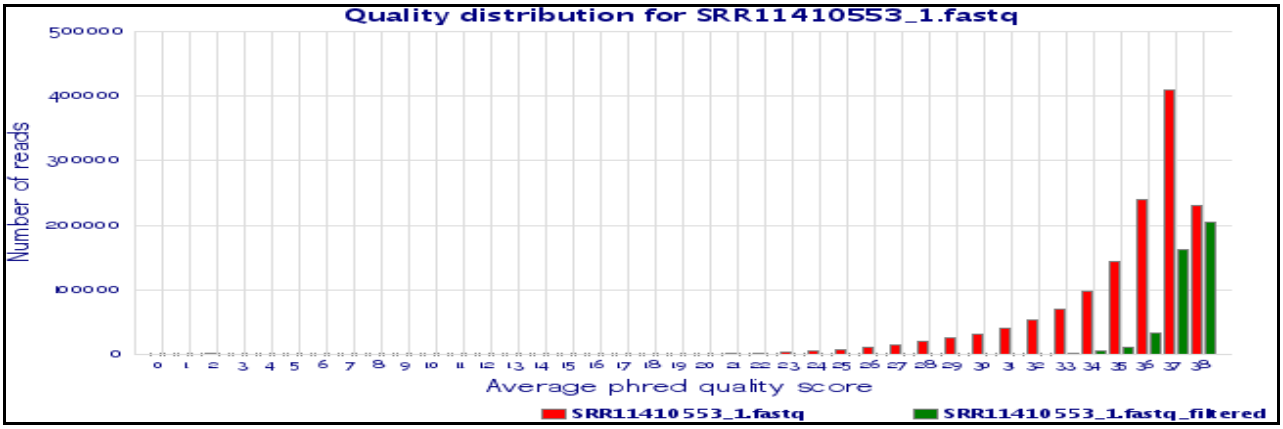


Figure 4.15.1. Quality distribution for input file, SRR11410553_1.fastq , before and after QC of CMB393 strain.

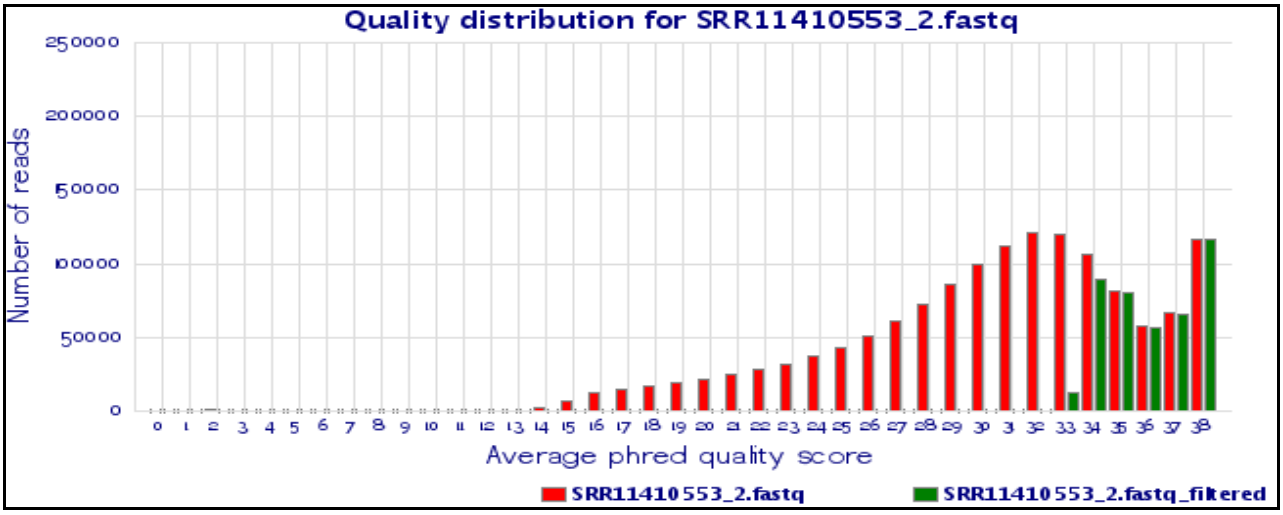


Figure 4.15.2. Quality distribution for input file, SRR11410553_2.fastq , before and after QC of CMB393 strain.

4.5.6. Summary of QC

4.5.6.1. CMB402 strain

Reads of CMB402 strain shows, 57.93% of high quality filtered reads which is indicated by green color, 42.07% of low quality reads which is indicated by red color and there is no primer/adaptor contaminated reads for both forward and reverse reads [Figure 4.16].

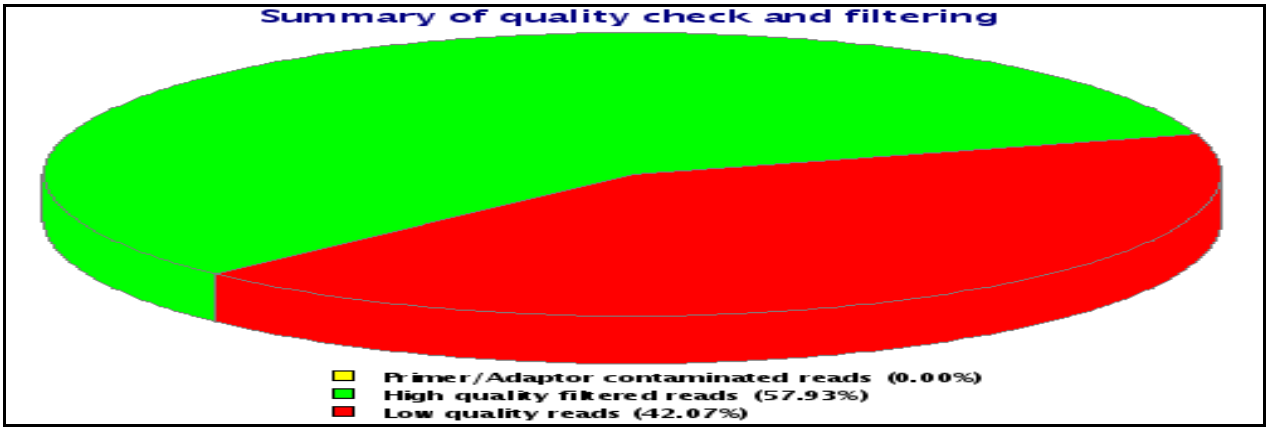


Figure 4.16. The summary of quality check for both SRR11410548_1.fastq and SRR11410548_2.fastq of CMB402 strain.

4.5.6.2. CMB401 strain

Reads of CMB401 strain show 29.92% of high quality filtered reads which is indicated by green color, 70.08% of low quality reads which is indicated by red color for both forward and reverse reads [Figure 4.17].

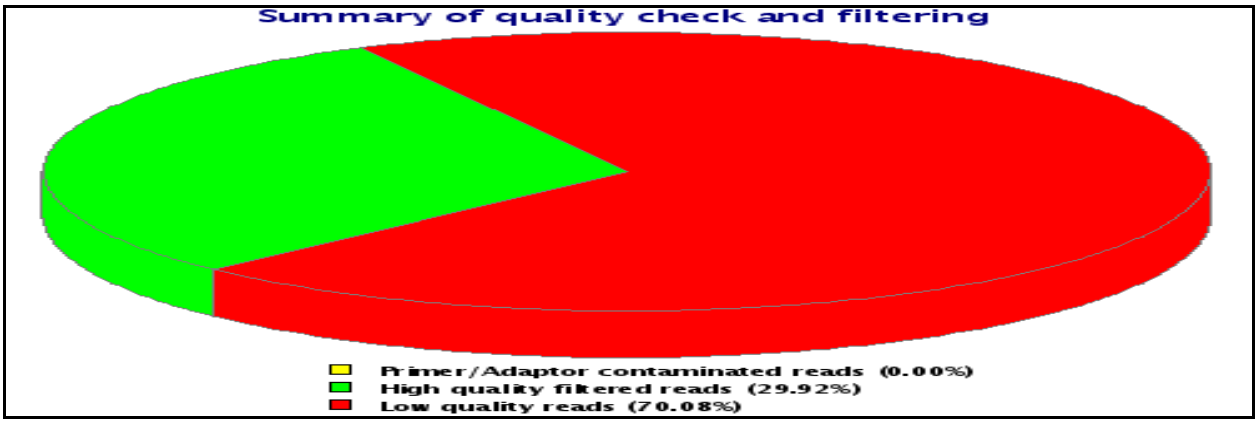


Figure 4.17. Above figure depicts the summary of quality check and filtering for both SRR11410549_1.fastq and SRR11410549_2.fastq of CMB401 strain.

4.5.6.3. CMB393 strain

Reads of CMB393 shows 29.82% of high quality filtered reads which is indicated by green color, 70.18% of low quality reads which is indicated by red color for both forward and reverse reads [Figure 4.18].

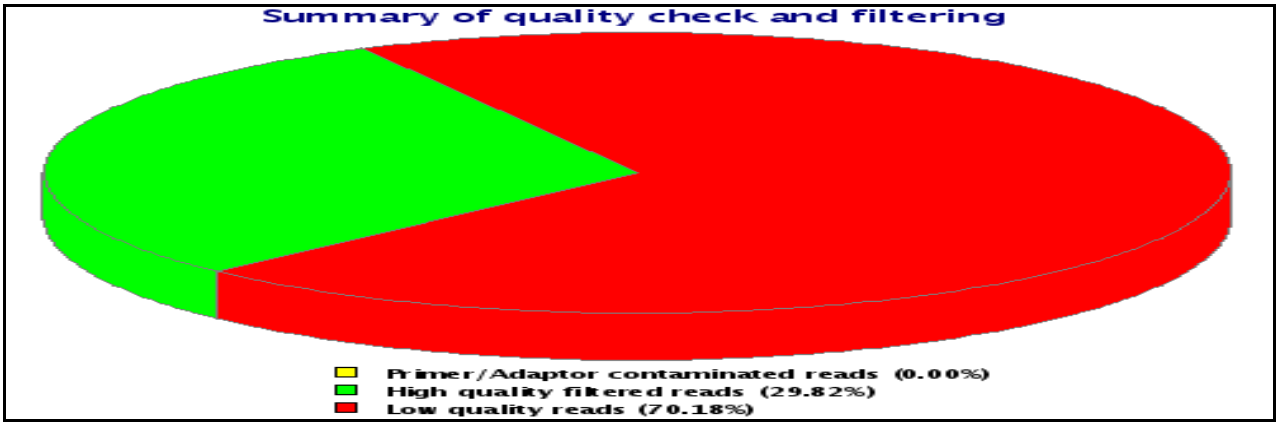


Figure 4.18. Above figure depicts the summary of quality check and filtering for both SRR11410553_1.fastq and SRR11410553_2.fastq of CMB393 strain.

4.5.7. Validation data of assembled genome

Validation data of assembled genome tells about their base composition (A, T, G, C), N25, N50, N75, N90, N95 length values of three different strains of the same species. N50 is the minimum contig length to cover 50% of the genome and it describes the quality of a draft assembly. CMB402 strain had high quality draft assembly as compare to CMB401, CMB393 strain because it has high N50 value that is 159232 [Table 4.6].

Table 4.6. Values for assembly validation

Strain	**CMB402**	**CMB401**	**CMB393**
Total Sequences	196	213	263
Total Bases	7142881	7144711	7319099
Min sequence length	128	128	128
Max sequence length	374120	696080	437958
Average sequence length	36443.27	33543.24	27829.27
Median sequence length	443.50	452.00	537.00
N25 length	259788	254705	263558
N50 length	159232	150671	144293
N75 length	97075	76365	88076
N90 length	48238	43489	38419
N95 length	28397	28421	23276
As	23.35%	23.51%	23.57%
Ts	23.77%	23.62%	23.67%
Gs	25.77%	26.35%	26.14%
Cs	27.11%	26.51%	26.62%
(A+T)s	47.12%	47.13%	47.25%
(G+C)s	52.88%	52.87%	52.75%

4.6 Genome annotation

4.6.1. CMB402 strain

Subsystem coverage which depicts the percentage of features those are in subsystem. 21% feature counts are present in subsystem that includes 137 subsystem feature counts related to cofactors, vitamins, prosthetic groups, pigments. It also includes 50 subsystem feature counts related to phosphorus metabolism, 31 subsystem feature counts related to nitrogen metabolism and 64 subsystem feature counts related to virulence, disease and defence [Figure 4.19] while its annotated gene [Table 4.7].

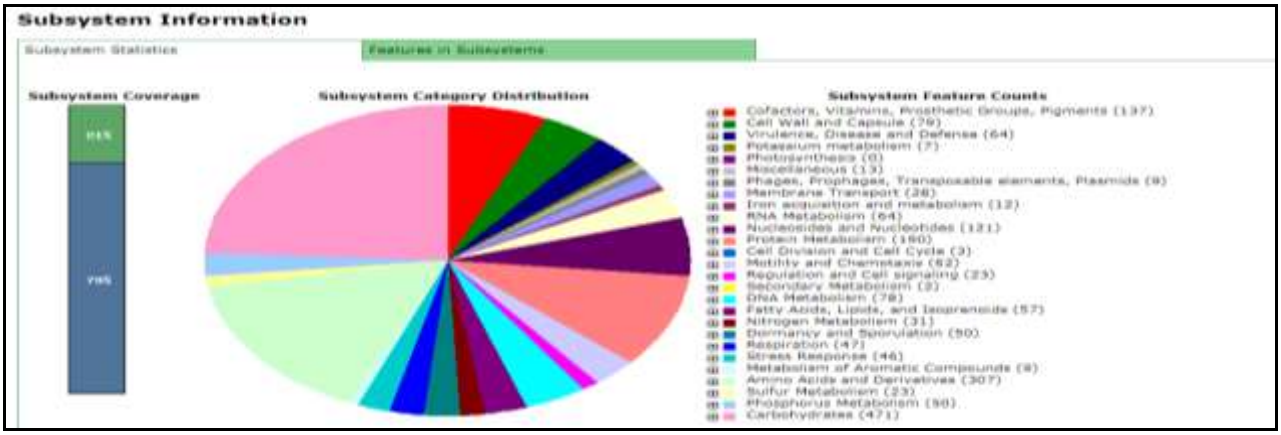
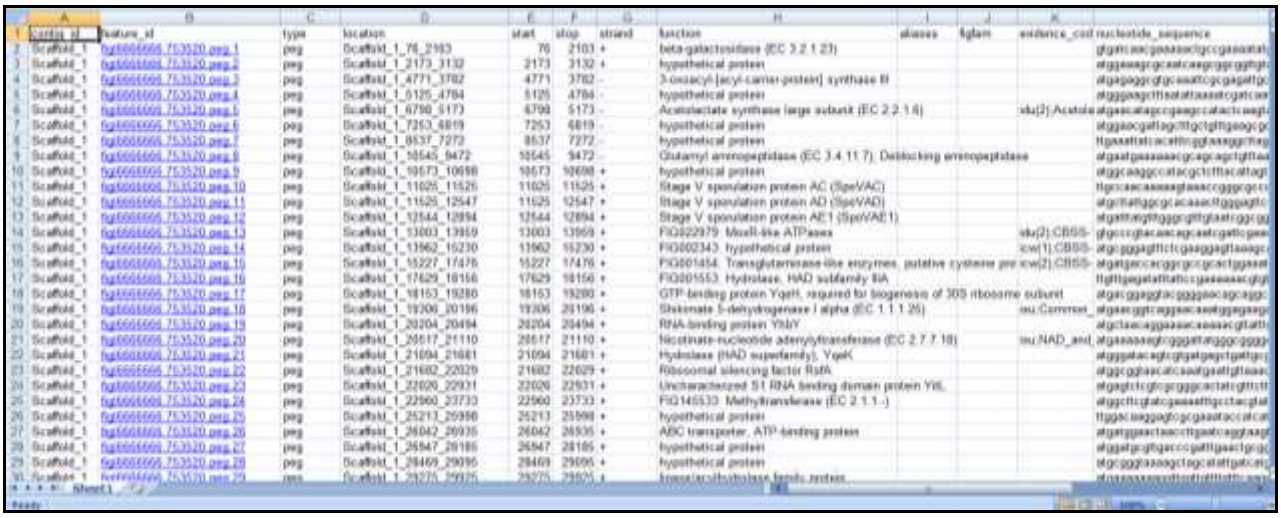


Figure 4.19. Subsystem category distribution of strain CMB402

Table 4.7. Annotation table of strain CMB402



4.6.2. CMB401 strain

21% feature counts are present in subsystem that includes 136 subsystem feature counts related to cofactors, vitamins, prosthetic groups, pigments. It also includes 47 subsystem feature counts related to phosphorus metabolism, 31 subsystem feature counts related to nitrogen metabolism and 66 subsystem feature counts related to virulence, disease and defence [Figure 4.20] while its annotated gene [Table 4.8].

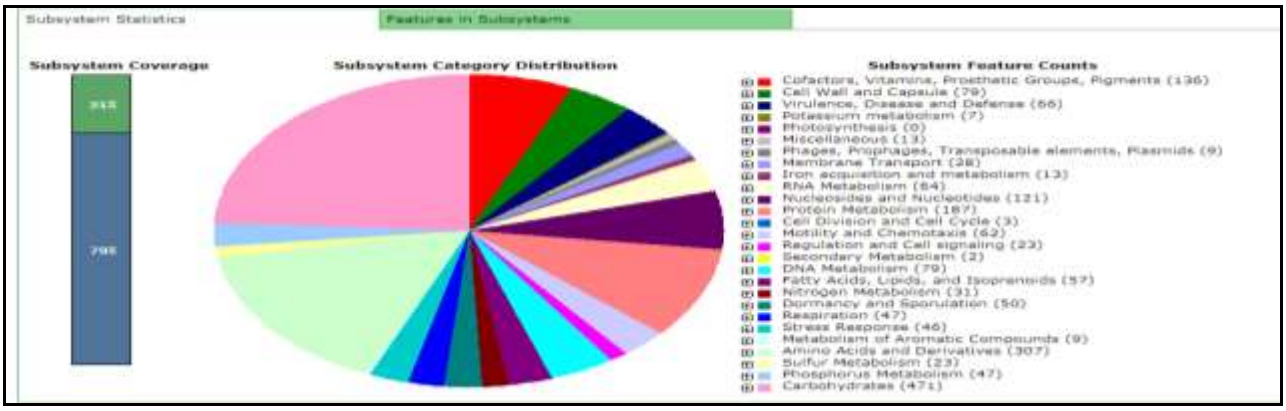
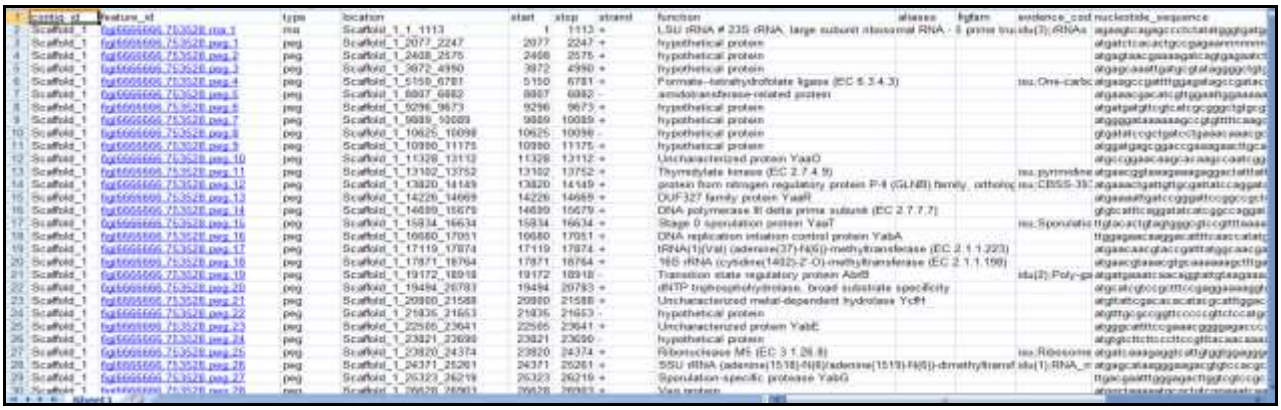


Figure 4.20. Subsystem category distribution of strain CMB401

Table 4.8. Annotation table of strain CMB401



4.6.3. CMB393 strain

21% feature counts are present in subsystem that includes 147 subsystem feature counts related to cofactors, vitamins, prosthetic groups, pigments. It also includes 56 subsystem feature counts related to phosphorus metabolism, 32 subsystem feature counts related to nitrogen metabolism and 62 subsystem feature counts related to virulence, disease and defense [Figure 4.21] while its annotated gene [Table 4.9].

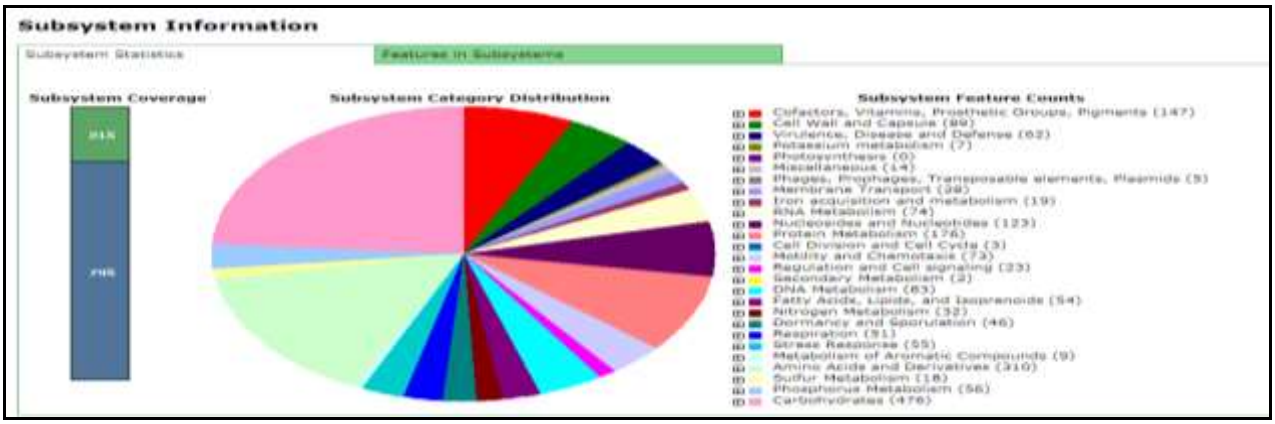
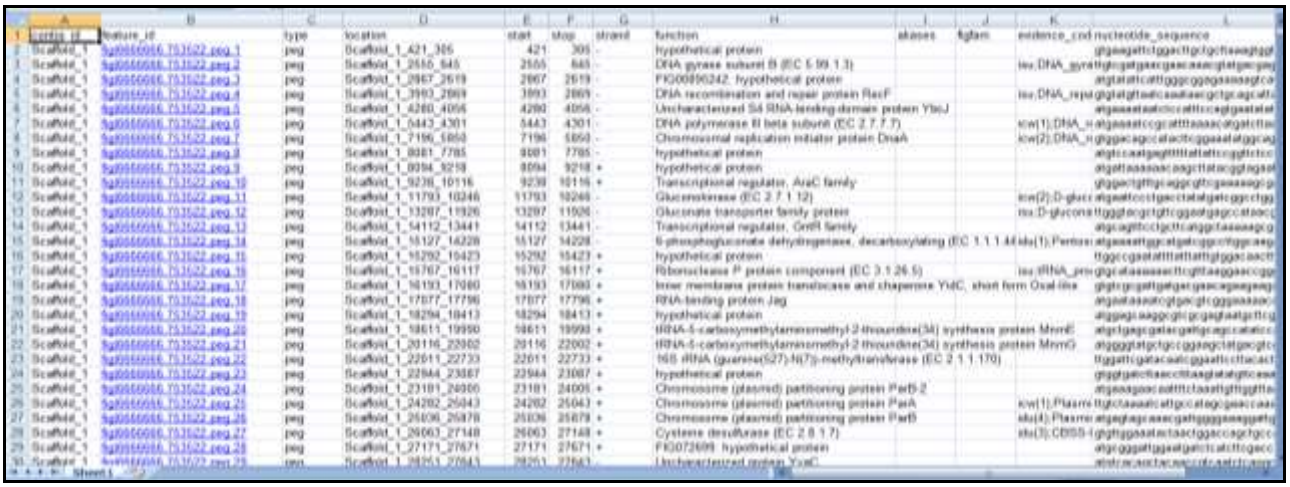


Figure 4.21. Subsystem category distribution of strain CMB393

Table 4.9. Annotation table of strain CMB393



4.7. Antibiotic Resistance Gene

4.7.1. CMB402 strain

In CMB402, three antibiotic resistance genes are found which are *vanI* (vancomycin resistance gene), *LimA 23S ribosomal methyltransferase* and *fexA* (florfenicol – chloramphenicol resistance gene) also present. Alteration in the target sites of antibiotics that is *vanI* is a common mechanism of resistance which consists of 69.1% identity of matching region. *LimA 23S ribosomal methyltransferase* consist of 84.97% identity of matching region. Antibiotic efflux is a key mechanism of resistance in Gram negative bacteria which is involved by *fexA* antibiotic resistance gene and it shows 68.58% identity of matching region [Figure 4.22].

Figure 4.22. Antibiotic resistance genes within strain CMB402

4.7.2. CMB401 strain

In CMB401, three antibiotic resistance genes are found which are *fexA* (florfenicol – chloramphenicol resistance gene), *LimA 23S ribosomal methyltransferase* and *vanI* (vancomycin resistance gene) also present. Alteration in the target sites of antibiotics that is *vanI* is a common mechanism of resistance which consists of 69.1% identity of matching region. Antibiotic efflux is a key mechanism of resistance in Gram negative bacteria which is involved by *fexA* antibiotic resistance gene and it shows 68.58% identity of matching region. *LimA 23S ribosomal methyltransferase* consist of 84.97% identity of matching region [Figure 4.23].

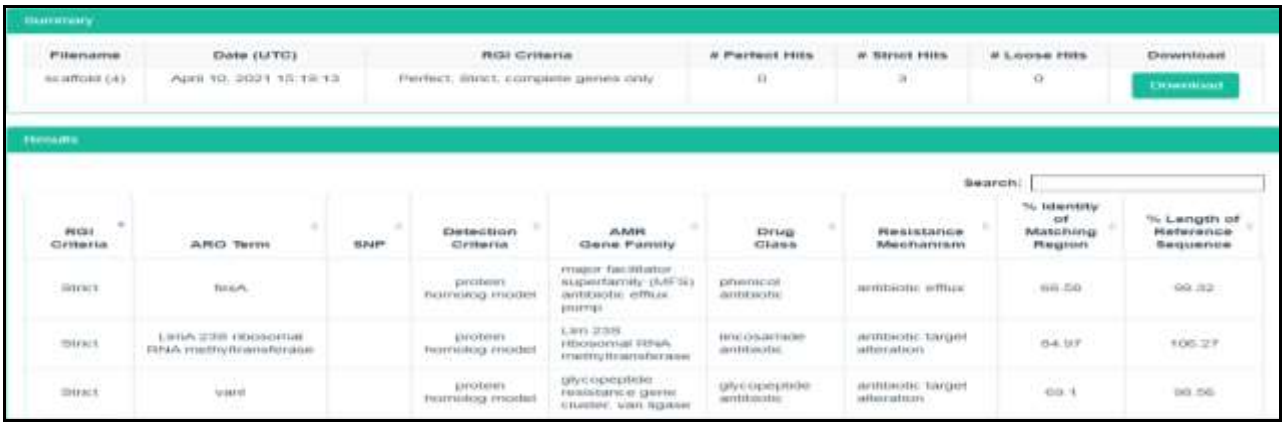


Figure 4.23. Antibiotic resistance genes within strain CMB401

4.7.3. CMB393

In CMB393, two antibiotic resistance genes are found which are *LimA 23S ribosomal methyltransferase* and *vanl* (vancomycin resistance gene). Both antibiotic resistance gene shows antibiotic target alteration mechanism. *LimA 23S ribosomal methyltransferase* resistance gene shows 86.01% identity of matching region and *vanI* shows 68.22% identity of matching region [Figure 4.24].

Figure 4.24. Antibiotic resistance genes within strain CMB393

4.8. <u>Prophage identification</u>

4.8.1. CMB402 strain

In CMB402, 6 prophage regions are found. From these six regions four regions are incomplete and their score is less than 70. Two regions are intact that is region 3 and region 5 and their score is 150. Prophage carrying population might in fact benefit from the release of free phages, because they may infect and kill any competing phage susceptible cells so CMB402 strain is more advantageous than the CMB401 and CMB393 strains [Figure 4.25.1 and 4.25.2].

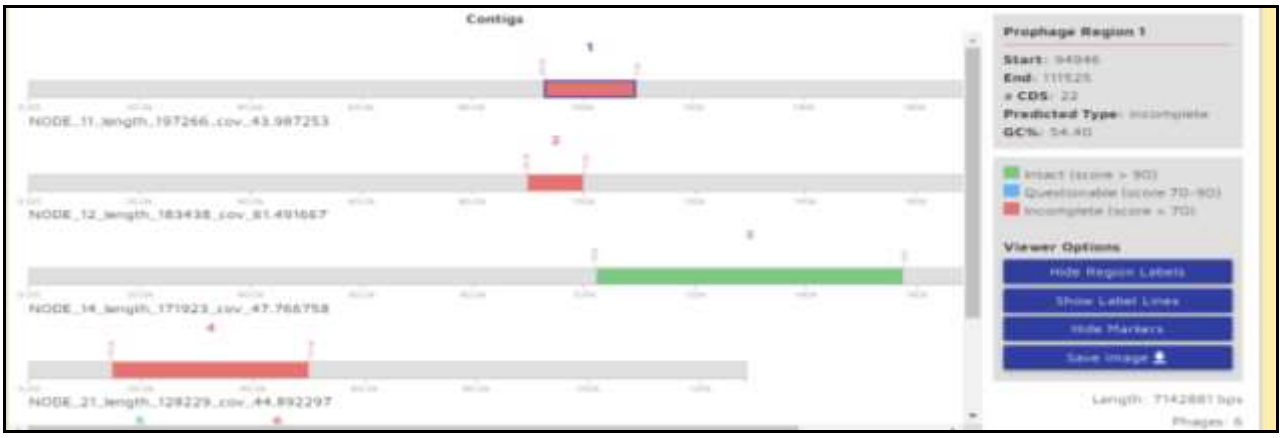


Figure 4.25.1. Phages associated with strain CMB402

Figure 4.25.2. Phages associated with strain CMB402

4.8.2. CMB401 strain

Prophage carrying population might in fact benefit from the release of free phages, because they may infect and kill any competing phage susceptible cells. In CMB401, 5 prophages are found. From these five prophages, three regions are incomplete and two regions are intact. The two regions are intact that is region 1 and region 5 and their score is 150. [Figure 4.26.1 and 4.26.2].

Figure 4.26.1. Phages associated with strain CMB401

Figure 4.26.2. Phages associated with strain CMB401

4.8.3. CMB393 strain

In CMB393, three prophages are found. These prophages are actually incomplete and their score is 20 [Figure 4.27].

Figure 4.27. Phages associated with strain CMB393

4.9. Plasmid finder

There is no plasmid present in any of the three strains of the same bacteria [Figure 4.28.1 – 4.28.3].

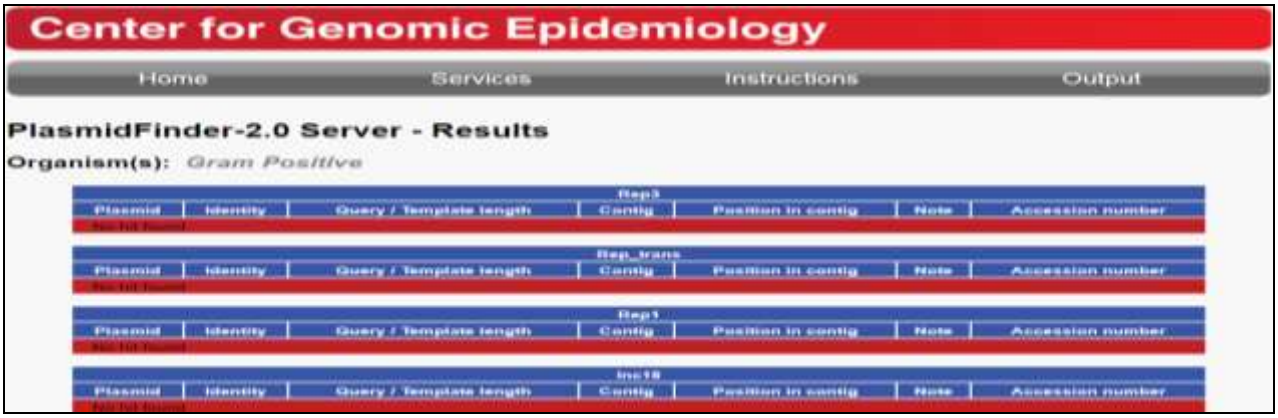


Figure 4.28.1. Plasmid associated with strain CMB402

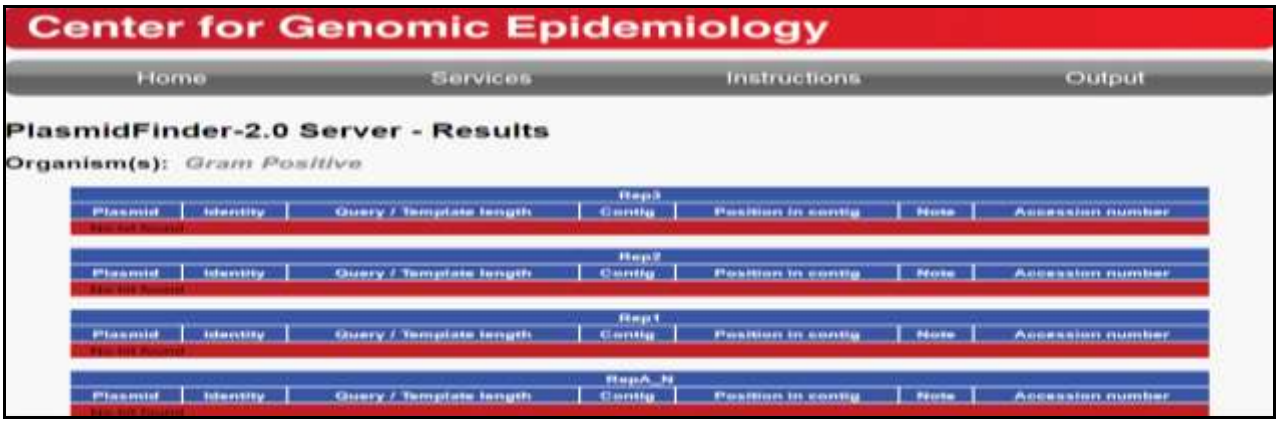


Figure 4.28.2. Plasmid associated with strain CMB401

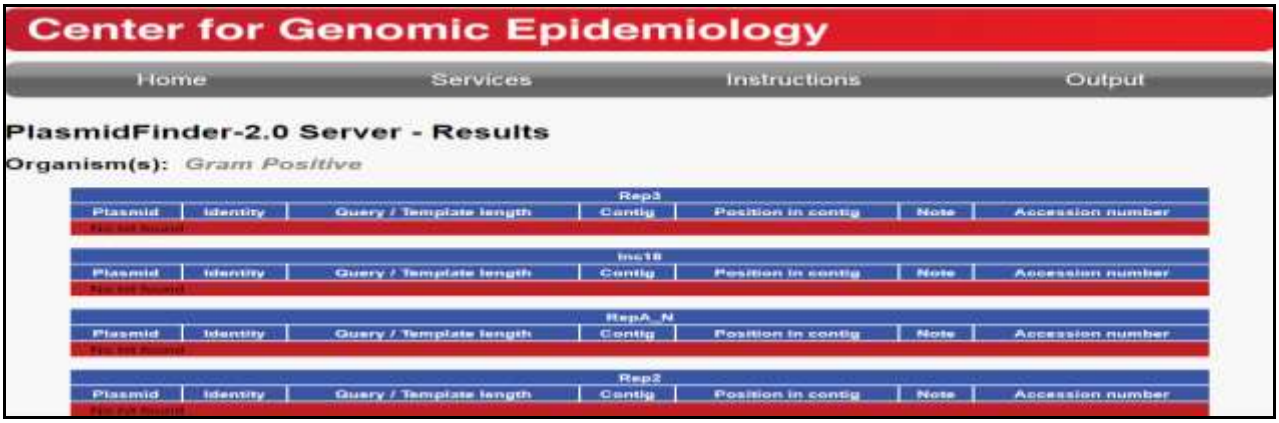


Figure 4.28.3. Plasmid associated with strain CMB393

4.10. Visualization of comparative genome analysis

(A) Comparative genome visualization through circular representation

BRIG : Blast Ring Image Generator

BRIG software was used for BLAST comparison of three unpublished genome of *Paenibacillus macerans* against the simulated draft genome. This image show similarity between central reference sequence (contigs.fasta) which indicate as red colour and other sequence as a set of concentric rings (genome sequence of CMB402, CMB401 and CMB393). Starting from the innermost circle going outwards: major tick (500 kb) and minor tick (100 kb) measurements of the contigs.fasta that is genome sequence of 3CT49 strain which is used as reference sequence; BLAST comparisons of CMB402 genome against the 3CT49 genome (green ring); BLAST comparisons of CMB401 genome against the 3CT49 genome (blue ring); BLAST comparisons of CMB393 genome against the 3CT49 genome (ocher ring); BLAST matches are coloured on a sliding scale indicating a defined percentage identity. Colored ring shows 100% identity with reference draft genome and gap (white color in the ring of genome sequence of CMB402, CMB401 and CMB393) in the ring represent less than 50% identity [Figure 4.29].

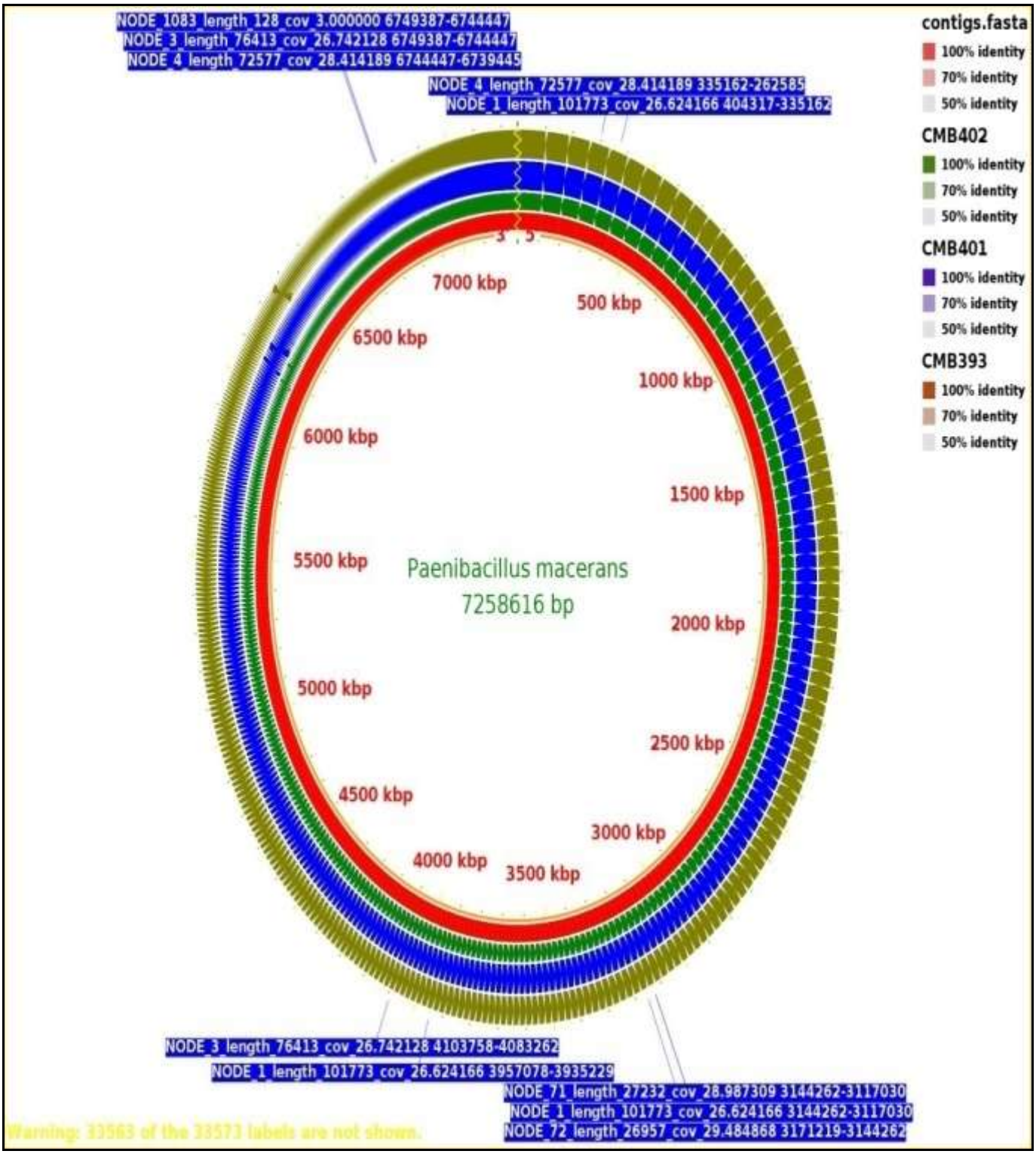


Figure 4.29. Comparative genome analysis using BRIG.

(B) Comparative genome visualization through linear representation

Mauve

Mauve software was used for alignment of *P.macerans* 3CT49 (used as a nearest reference sequence), *P.macerans* CMB402, *P.macerans* CMB401 and *P.macerans* CMB393. Notice how inverted regions in the *P.macerans* CMB401 and *P.macerans* CMB393 are clearly depicted as blocks below a genome's center line. These four genome were aligned with Progressive Mauve using default parameters, as shown in below figure. The colored blocks in the first genome are connected by lines to similarly colored blocks in the second, third and fourth genomes. These lines indicate which regions in each genome are homologous. The region of sequence covered by colored block is entirely collinear and homologous among the genomes. The boundaries of colored blocks usually indicate the breakpoints of genome rearrangement, unless sequence has been gained or lost in the breakpoint region [Figure 4.30]. The image show linear comparison between genome sequence of CMB402 (Scaffold(2).fasta), CMB401 (Scaffold(3).fasta), CMB393 (Scaffold(4).fasta) and 3CT49 (Scaffold(1).fasta) [Table 4.10].

Table 4.10. Sequence name represent the strain

Strain	Sequence name
CMB402	Scaffold(2).fasta
CMB401	Scaffold(3).fasta
CMB393	Scaffold(4).fasta
3CT49	Scaffold(1).fasta

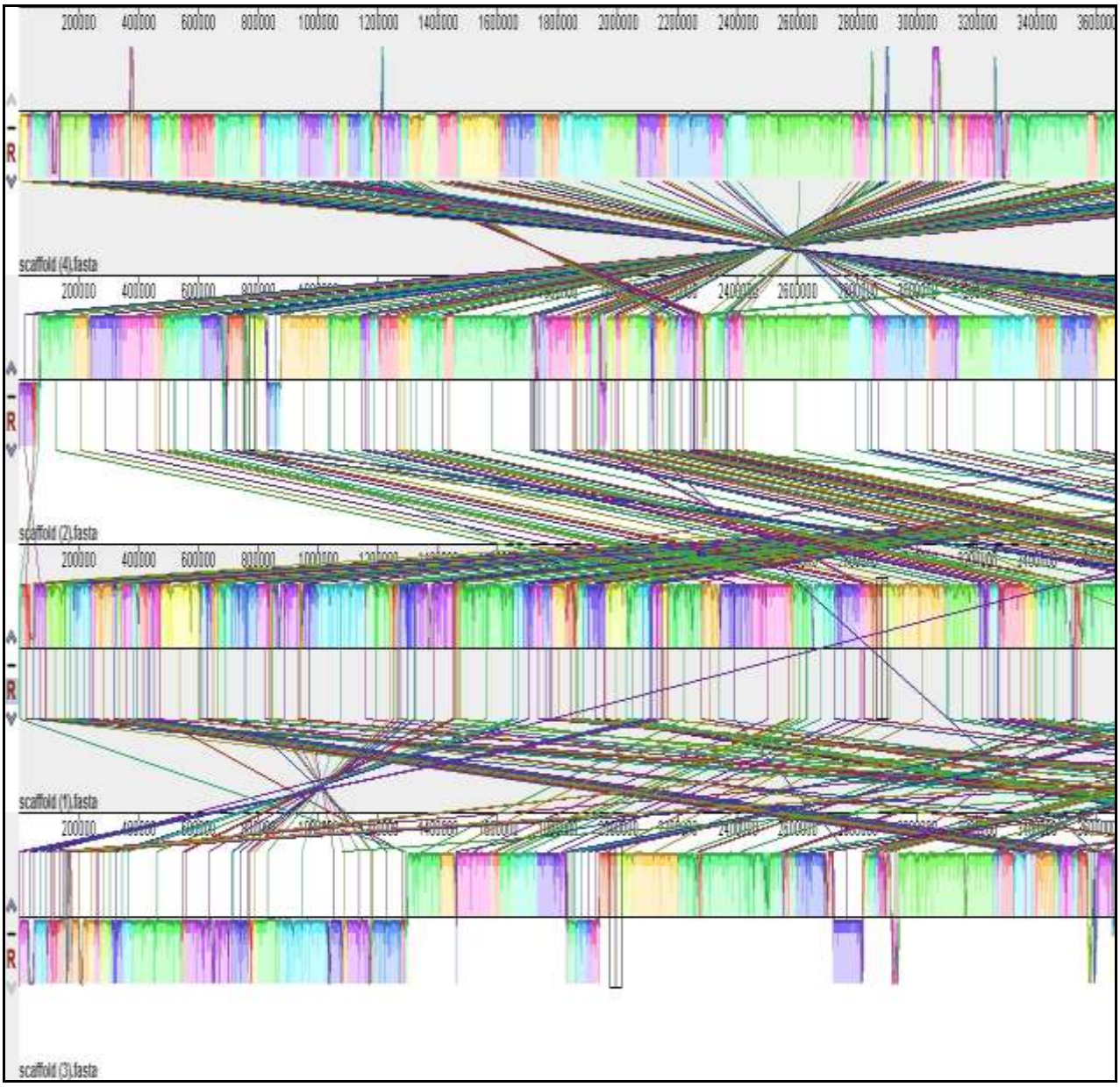


Figure 4.30. Alignment of genome sequence of CMB402, CMB401 and CMB393 strain using MAUVE

4.11. Annotated genes

4.11.1 Annotated genes involved in Phosphate and Nitrogen metabolism

In CMB402, CMB401 and CMB393 strain, genes are annotated which involve in phosphate metabolism for example, *SphR*, *PhoU*, *PhoP*, *PhoH* and *SphS* etc. The *phoA* gene encode for alkaline phosphatase. Many protein encode genes are present in these strain that are phosphate regulon sensor protein *PhoR*, phosphate regulon sensor protein *PhoB*, Alkaline phosphatase synthesis transcriptional regulatory protein *PhoP* and Phosphate transport system regulatory protein *PhoU* [Table 4.11].

In CMB402, CMB401 and CMB393 strain, genes are annotated which involve in nitrogen metabolism for example, Nitrite transporter from formate/ nitrite family, *NirC*, Nitrate / nitrite sensor protein Respiratory nitrate reductase which are involved in nitrate / nitrite ammonification. Glutamine synthetase type I, Glutamate synthase (NADPH) small chain, Ammonium transporter, Glutamate synthase (NADPH) large chain are involved in ammonia assimilation. Nitrous oxide reductase maturation protein *NosD*, Nitrous oxide reductase maturation transmembrane protein *NosY*, Nitrous oxide reductase maturation protein outer membrane lipoprotein *NosL* acts as denitrifying reductase gene clusters [Table 4.12].

Table 4.11. Genes involved in Phosphorus metabolism

Category	Sub - category	Gene/ Enzyme
Phosphorus metabolism	Phosphate metabolism	Phosphate regulon sensor protein *PhoR* (*SphS*)
		Phosphate transport regulator (distant homolog of *PhoU*)
		Inorganic pyrophosphate *PpaX*
		Phosphate starvation inducible protein *PhoH*, predicted ATPase
		Alkaline phosphatase
		Predicted ATPase related to phosphate starvation – inducible protein *PhoH*
		Probable low affinity inorganic phosphate transporter
		Alkaline phosphatase synthesis transcriptional regulatory protein
		Phosphate transport system regulatory protein *PhoU*
		Phosphate regulon transcriptional regulatory protein *PhoB* (*SphR*)

Table 4.12. Genes involved in Nitrogen metabolism

Category	Subsystem	Gene / Enzyme
Nitrogen metabolism	Nitrate and nitrite ammonification	Respiratory nitrate reductase alpha chain
		Nitrite transporter from formate / nitrite family
		Nitrite transporter *NirC*
		Nitrite reductase probable (NaD(P)H) large subunit
		Nitrite reductase probable (NaD(P)H) small subunit
	Ammonia assimilation	Glutamate synthase (NADPH) small chain
		Glutamate synthase (NADPH) large chain
	Denitrifying reductase gene clusters	Nitrous oxide reductase maturation protein *NosD*
		Nitrous oxide reductase maturation transmembrane protein *NosY*
		Nitrous oxide reductase maturation protein outer membrane lipoprotein *NosL*

4.11.2 Annotated genes involved in DNA metabolism and Virulence, Disease and Defence

CRISPR sequences found in the genomes of three different strain of *Paenibacillus macerans* for example, CRISPR associated RAMP Cmr1 – 4, CRISPR associated protein Csd1 family, CRISPR associated protein Cas2 and CRISPR associated protein Cas1[Table 4.13].

From CMB402, CMB401 and CMB393 strain, genes are annotated which involve in virulence disease and defence for example, Macrolide specific efflux protein *MacA*, Acriflavin resistance protein and Multi antimicrobial extrusion protein (Na(+) / drug antibiotic family of MDR efflux pumps) categorized under multidrug resistance efflux pumps. Chromate transport protein *ChrA*, cadmium efflux system accessory protein, copper translocating p-type ATPase, copper – zinc – cadmium and transcriptional regulator *Mer R* family involve in resistance to toxic compounds. Ribosome protection type tetracycline resistance group 2, translation elongation factor G and fosfomycin resistance protein *FosB* involve in resistance to antibiotics [Table 4.14].

Table 4.13. Genes involved in DNA metabolism

Category	Sub category	Gene / Enzyme
DNA metabolism	CRISP Cmr Cluster	CRISPR – associated RAMP Cmr3
		CRISPR – associated RAMP Cmr4
		CRISPR – associated RAMP Cmr1
		CRISPR – associated RAMP Cmr2
		CRISPR – associated RAMP Cmr6
	CRISPRs	CRISPR – associated protein, Csd1 family
	CRISPRs	CRISPR – associated protein Cas2
		CRISPR – associated protein Cas1

Table 4.14. Genes involved in Virulence, Disease and Defence

Category	Subsystem	Gene / Enzyme
Virulence, Disease and Defence	Multidrug resistance efflux pumps	Macrolide specific efflux protein *MacA*
		Acriflavin resistance protein
		Multi antimicrobial extrusion protein (Na(+) / drug antibiotic family of MDR efflux pumps
	Resistane to chromium compounds	Chromate transport protein *ChrA*
	Cadmium resistance	Cadmium efflux system accessory protein
	Fosfomycin resistance	Fosfomycin resistance protein *FosB*
	Copper homeostasis	Copper translocating P-type ATPase
	Copper – Zinc – Cadmium resistance	Copper – Zinc – Cadmium resistance protein
		Transcriptional regulator, *MerR* family

DISCUSSION

CHAPTER 5: DISCUSSION

The *Paenibacillus macerans* represent the phosphate solubilizing and nitrogen fixing microorganism. This microorganism was used for the study entitiled **Studies on "Comparative genome analysis of *Paenibacillus macerans* CMB402, CMB401 and CMB393."** In the present study, we firstly reported the detailed study of *Paenibacillus macerans* CMB402, CMB401 and CMB393 genome sequences and predicted several phosphate solubilization, nitrogen metabolism, antibiotic resistance and metal resistant genes.

Genome Assembly and Annotation

Genome assembly refers to the process of placing nucleotide sequence into the right order (Foxman, 2012). One of the primary issues that have been discovered when utilising short-read sequencing methods is the huge number of contigs following assembly (Smits, 2019). The quality of de novo genome assembly may be evaluated using a variety of factors, including the number of contigs and scaffolds available and their sizes, the proportion of reads that can be assembled, and the contig N50 value, which is a frequently used to measure and evaluate the quality of assembly (Choudhuri, 2014).

In this study, N50 value of contigs of *Paenibacillus macerans* CMB402, CMB401 and CMB393 are 159232, 150671 and 144293. *Paenibacillus macerans* CMB402 strain had high quality draft assembly as compare to CMB401and CMB393 strain because it has high N50 value i.e., 159232. GC percentage of *Paenibacillus macerans* CMB402, CMB401 and CMB393 are 52.88%, 52.87% and 52.75% and their genome size are 7142881 bp, 7144711 bp and 7319099

bp, respectively. The regions with 50–60% GC content obtaining the foremost coverage and regions with high (70–80%) or low (30–40%) GC content receiving considerably less coverage (Gnirke *et al.,* 2009). *Paenibacillus macerans* ATCC strain has 53% of GC content and it's genome size is 7331450 bp which is almost similar to this study (Daligault et al, 2014). *Paenibacillus macerans* CMB402 has higher GC percentage that is 52.88 which showed higher coverage and better data quality.

The RAST annotation engine was created to annotate bacterial and archaeal genomes, and it works by providing a standard software workflow for detecting genomic characteristics (such as protein-encoding genes and RNA) and annotating their activities (Brettin *et al.,* 2015). Furthermore, only 25% of total sequences fulfill the N50 contig scaffolding criteria, which measures how much of the genome is assembled into bigger contigs (Schatz *et al.*, 2010 and Lewin *et al.*, 2018). The number of contigs decreased by increasing the coverage and N50 value (Peng *et al.*, 2014). The contigs were further assembled to scaffolds, in which the N50 value increased to 278921 bp of *Paenibacillus macerans* CMB402, 227313 bp of *Paenibacillus macerans* CMB401 and 228196 bp of *Paenibacillus macerans* CMB393. *Paenibacillus macerans* 3CT49 has 108 contigs, 6340 coding sequences and 120 RNAs (Olajide *et al.,* 2020). However, in *Paenibacillus macerans* CMB402, number of contigs, coding sequences and RNAs are 69, 6985 and 99 respectively. In *Paenibacillus macerans* CMB401, number of contigs, coding sequences and RNAs are 79, 7013 and 97 respectively. In *Paenibacillus macerans* CMB393, number of contigs, coding sequences and RNAs are 210, 7161 and 96 respectively. Based on genome assembly and structural genome annotation, *Paenibacillus macerans* CMB402 considered to be the best strain because it has highest N50 value of contigs and scaffolds, lower number of contigs and higher percentage of GC content as compared to

Paenibacillus macerans CMB401, CMB393 and 3CT49 strain from which it conclude that if number of contigs are lower, than N50 stats value will increase.

Annotated genes

The total number of genes in *Paenibacillus macerans* CMB402, CMB401 and CMB393 are 7285, 7311 and 7798 from which 5946 hypothetical genes in CMB402, 5947 hypothetical genes in CMB401 and 5942 hypothetical genes in CMB393. *PhoR, PhoU, PpaX, PhoH, glpQ* and *PhoB* genes are annotated which involves in phosphate solubilization in all three strains of *Paenibacillus macerans* i.e., CMB402, CMB401 and CMB393. The total number of phosphate metabolism genes present in *Paenibacillus macerans* CMB402, CMB401 and CMB393 are 33, 30 and 35, respectively. According to Beneit *et al.,* 2015, some bacteria present (Pseudomonas, Bacillus) regulatory mechanism of inorganic phosphate that involves two regulatory systems called as pho regulon. The PhoP-PhoR involve in regulation of Pho regulon (Shi *et al.,* 1999). Phosphate deprivation increased the expression of genes involved in phosphate absorption and organic phosphate solubilization, such as *pst* (Pi-specific transporter), *phoA* (alkaline phosphatase), *glpQ* (glycerophosphoryldiester phosphodiesterase), and *ushA* (nucleotidase) (Ishige *et al.,* 2003 and Pragai *et al.,* 2004). All the three strains (*Paenibacillus macerans* CMB402, CMB401 and CMB393) involve in phosphate solubilization.

Paenibacillus macerans has also found to accommodate metal resistance genes that are *ChrA* and *MerR* which provide cadmium, zinc and copper resistance. According to Khan *et al.,* 2012, Paenibacillus species has also found to promote growth of plants in heavy metal polluted sites. The total number of nitrogen metabolism genes present in *Paenibacillus macerans* CMB402, CMB401 and CMB393 are 31, 31 and 32, respectively. *NosD*, *NirC*, *glnA*, *NosY* and *NosL* genes are involved in nitrogen metabolism. All the three strains of *Paenibacillus macerans*

accommodate these genes involve in nitrogen metabolism. According to Xie *et al.*, 2014, some Paenibacillus species participate in nitrogen metabolism. *glnA* gene encoding the glutamine synthetase, *gdhA* gene encoding the glutamate dehydrogenase which play important role in ammonia assimilation in *Helicobacter pylori* (Mobley *et al.*, 2001).

In *Paenibacillus macerans* CMB402 and CMB401 strain accommodate 10 CRISPRs array and in CMB393 strain CRISPRs array present are 17. However, in *Paenibacillus larvae,* DSM25430 strain accommodate 7 CRISPRs array (Stamereilers *et al.,* 2021). Based on CRISPR array accumulation, *Paenibacillus macerans* CMB393 is the best strain because it has higher number of CRISPR array as compared to Paenibacillus CMB402, CMB401 and *Paenibacillus larvae* DSM25430 strain. CRISPR (clustered regularly interspaced short palindromic repeats) is a defensive system in prokaryotic cells that develops resistance to foreign genetic material, such as that found in plasmids or phages, by allowing them to identify and destroy the virus's DNA (Mohamadi *et al.,* 2020). The Paenibacillus strain sp. E222 genome is 7.5 Mb in size, with a G+C content of 46% and 1 CRISPR/Cas System (Bastias, *et al.*, 2020).

Antibiotic resistance gene

The Comprehensive Antibiotic Resistance Database's Resistance Gene Identifier (RGI) was used to search for putative antimicrobial resistance genes in genomic sequences (McArthur *et al.,* 2013). The *vanI* and *LimA 23S ribosomal methyltransferase* antibiotic resistance genes are identified in *Paenibacillus macerans* CMB402, CMB401 and CMB393 strains in this study. But *fexA* antibiotic resistance gene identified only in *Paenibacillus macerans* CMB402 and CMB401 strains. According to Pasari *et al.,* 2019, Fosfomycin, vancomycin, tetracycline and many more antibiotic resistance genes are identified in *Paenibacillus polymyxa* A18. In staphylococci, the *fexA* gene codes for the florfenicol efflux protein (Kehrenberg *et al.,* 2004).

Prophage

A prophage is a bacteriophage genome that has been introduced and integrated into the circular bacterial DNA chromosome or lives as a plasmid outside of the chromosome. Prophages, which may fill up to 20% of bacterial chromosomes and so give novel roles to their hosts, are a key source of new genes for bacteria (Wang *et al.*, 2016).

Prophage carrying population might in fact benefit from the release of free phages, because they may infect and kill any competing phage susceptible cells so *Paenibacillus macerans* CMB402 and CMB401 strain is more advantageous than the CMB393 strain. In *Paenibacillus macerans* CMB402 strain accommodate six prophages from which two are intact means not damaged or impaired in any way. In *Paenibacillus macerans* CMB401 strain accommodate five prophages from which two are intact. In *Paenibacillus macerans* CMB393 strain accommodate three prophages and all three prophages are incomplete. Similarly, in *Paenibacillus larvae* DSM25430 strain accommodate 12 prophages from which two are intact (Stamereilers *et al.*, 2021). Because intact prophages are likely to destroy the cell when the lytic cycle is induced, vigorous selection for mutations that cause prophage inactivation should be made (Bobay *et al.*, 2014).

Visualization of comparative genome analysis

BLAST search of nucleotide sequence between strain 3CT49 of *Paenibacillus macerans* and other three strains that are *Paenibacillus macerans* CMB402, CMB401 and CMB393 showed that 3CT49 strain has highest similarity with CMB402 (Figure 10(A)) using the BRIG program. Paenibacillus species A2 showed similarity with *Paenibacillus elgii* B69 (Zhang *et al.*, 2016).

Mauve aligns genomes in the same way regardless of input order by detecting multi-MUMs in subsets of the genomes (Darling *et al.*, 2004). The complete genome sequences of the sequenced strains were compared using Mauve to determine the evolutionary gap between them and various *Paenibacillus polymyxa* strains (Li *et al.*, 2020). *Paenibacillus macerans* CMB401 and CMB393 genomes have undergone more genome rearrangements and showed more deletions and inversions than CMB402 which further helps in evolutionary study.

All of the foregoing discoveries might lead to a better knowledge of *Paenibacillus macerans* CMB402, CMB401, and CMB393's genomic architecture, showing significant promise for this bacterium's use in agricultural and medicinal research.

SUMMARY AND CONCLUSION

CHAPTER 6: SUMMARY AND CONCLUSION

Paenibacillus macerans is a bacterium that aids in agricultural purposes through phosphate solubilization and nitrogen fixation. The first genome sequence of *Paenibacillus macerans* CMB402, CMB401, and CMB393 strains were acquired from European Nucleotide Archive for the paper **Studies on "Comparative genome analysis of *Paenibacillus macerans* CMB402, CMB401, and CMB393."**

The quality check was performed using the fastqc programme, which provided an html report that displayed the quality score and percentage of the GC content of readings from three distinct strains of same species: *Paenibacillus macerans* CMB402, CMB401, and CMB393. For the sequences of three strains of *Paenibacillus macerans*, quality control was performed using the NGS QC Toolkit, which was able to provide filtered reads in html report form, indicating how much better quality reads were present in certain sequences. The CMB402 strain of *Paenibacillus macerans* had a greater number of high-quality bases. Because *Paenibacillus macerans* CMB402 had a high N50 value of 159232, the assembled genome validation data showed that it had a high quality draft assembly.

RAST software was used to annotate the genome of *Paenibacillus macerans*. A total of 7285 genes were annotated in the CMB402 strain, 7311 genes in the CMB401 strain, and 7798 genes in the CMB393 strain. Genes involved in phosphate metabolism (*SphR*, *PhoU*, *PhoP*, *PhoH*, and *SphS*), nitrogen metabolism (*NosD*, *NirC*, *NosY*, *NosL*), virulence, and defence metabolism were identified in these strains. CARD (Comprehensive Antibiotic Resistance

Database) was utilized to find antibiotic resistance genes. *Paenibacillus macerans* CMB402 and CMB401 had three antibiotic resistance genes (*vanI, LimA 23S ribosomal methyltransferase*, and *fexA*), whereas CMB393 had just two (*limA 23S ribosomal methyltransferase* and *vanI*). The discharge of free phages may actually benefit prophage-carrying populations. The phaster database was used to find or investigate prophage. *Paenibacillus macerans* CMB402 had a larger number of prophage (6 in total) as comparison to *Paenibacillus macerans* CMB401 (5 in total) and *Paenibacillus macerans* CMB393 (3 in total). Plasmid Finder was used to identify plasmids. This database revealed that none of the three strains of *Paenibacillus macerans* had plasmids.

Two software programmes were utilised in the comparative analysis study: BRIG (Blast ring image generator) and Mauve. These three strains have 100 percent identity against the genome's reference sequence (*Paenibacillus macerans* 3CT49) at maximal places or bases, but less than 50 percent identity at a few base locations. *Paenibacillus macerans* CMB402 had a greater level of similarity to the reference genome than Paenibacillus CMB401 and CMB393 strains. Mauve might be a programme that attempts to align sequences in different species that evolved from a common ancestral sequences and xenologous sections of sequences that have undergone local and large-scale changes. The genomic rearrangement in *Paenibacillus macerans* CMB402 was less inverted and had fewer breakpoints.

Conclusion

Finally, this research concluded that *Paenibacillus macerans* CMB402 has an excellent strain as comparison to *Paenibacillus macerans* CMB401 and CMB393 because it accommodate genes related to phosphate solubilization (*PhoR*, *PhoB*, *PhoU*, *glpQ*, *PhoH*), nitrogen metabolism (*NirC*, *NosD*, *NosY*, *NosL*, *glnA*), metal resistant (*ChrA*, *MerR*) and antibiotics resistant (*fexA*, *vanI*, and *LimA 23S ribosomal methyl transferase*), which are involved in illness and defense mechanisms, according to the findings. It has a higher GC percentage (52.88%), high quality draft assembly (N50 Stats is 159232) and 2 intact prophages. *Paenibacillus macerans* CMB402 genome had undergone less genome rearrangements and showed less deletions and inversions than other strains. But based on CRISPR array, *Paenibacillus macerans* CMB393 has an excellent strain because it has 17 CRISPR array encoded genes which involve in defensive system. *Paenibacillus macerans* CMB402 might play a significant role in the future in agricultural area for phosphate solubilization, nitrogen fixation and also used for agricultural land which is saturated with toxic heavy metals, their genes will be manipulated for enhancement of these activities and in biomedical field for the use of antibiotic resistant genes.

REFERENCES

REFERENCES

Achouak, W., Normand, P., & Heulin, T. (1999). Comparative phylogeny of rrs & nif H genes in the Bacillaceae. *International Journal Systematic Bacteriology*, **49**(3), 961-967.

Alikhan, N.F., Petty, N.K., Ben Zakour, N.L., & Beatson, S.A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, **12**, 1471-2164.

Andrews, S. (2010). Fast QC : a quality control tool for high throughout sequence data. *Babraham Bioinformatics*.

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.

Arakawa, K., Tamaki, S., Kono, N., Kido, N., Ikegami, K., Ogawa, R., & Tomita, M. (2009). Genome Projector – zoomable genome maps with multiple views. *BMC Genomics*, **10**, 31.

Arndt, D., Grant, J., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, **44**, 16-21.

Ash, C., Priest, F., & Collins, M.D. (1993). Molecular identification of rRNA group 3 bacilli (Ash, Farrow, Wallbanks & Collins) using a PCR Probe test. *Antonie Van Leeuwenhoek*, **64**, 253-260

Auman, A.J., Speake, C.C., & Lidstrom, M.E. (2001). Nif H sequences & nitrogen fixation in type I and type II methanotrophs. *Applied and Environmental Microbiology*, **67**(9), 4009-4016.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., & Kubal, M. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Badri, D.V., Weir, T.L., Lelie, D.V., & Vivanco, J.M. (2009). Rhizosphere chemical dialogues: plant : microbe interactions. *Current Opinion in Biotechnology*, **20**(6), 642-650.

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 assembly algorithm and its applications to single-cell sequencing. *J*ournal *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**, 4.

Bastias, D.A., Jauzegui, R., Applegate, E.R., Altermann, E., Card, S.D., & Johnson, L.J. (2020). Complete genome sequence of Paenibacillus sp. strain E222, a bacterial symbiont of an Epichloe Fungal Endophyte of Ryegrass. *Microbiology Resource Announcements*, **9**, e00786-20.

Behera, B.C., Singdevsachan, S.K., Mishra, R.R., Dutta, S.K., & Thatoi, H.N. (2014). Diversity, mechanism & biotechnology of Phosphate Solubilizing Microorganismsin mangrove – A review. *Biocatalysis and Agricultural Biotechnology*, **3**(2).

Ben – Jacob, E., & Cohen, I. (1997). Cooperative formation of bacterial patterns. In Bacteria as multicelluar organisms edited by shaprio JA, Dworkin M. *New York : Oxford University Press*, 394 -416.

Berge, O., Guinebretie`re, M.H., Achouak, W., Normand, P., & Heulin, T. (2002). *Paenibacillus graminis* sp. nov. and *Paenibacillus odorifer* sp. nov., isolated from plant roots, soil and food. *International Journal of Systematic Evolutionary Microbiology*, **52**(2), 607-616.

Bert, F., Ouahes, O., & Zechovsky, N.L. (1995). Brain abscess due to Bacillus macerans following a penetrating periorbital injury. *Journal of Clinical Microbiology*, **33**(7), 1950–1953.

Bobay, L.M., Touchon, M., & Rocha, E.P.C. (2014). Pervasive domestication of defective prophages by bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **111** (33), 12127-12132.

Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Thomason Iii, J.A., Stevens, R., Vonstein, V., Wattam, A.R., & Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Science Reports*, **5**, 8365.

Bloemberg, G.V., & Lugtenberg, B.J. (2001). Molecular basis of plant growth promotion & biocontrol by rhizobacteria. *Current Opinion in Plant Biology*, **4**(4), 343-350.

Boratyn, G.Z.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., Mc Ginnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST – a more efficient report with usability improvements. *Nucleic Acids Research*, **41**, W29-W33.

Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M.F., Lió, P., Crescenzi, P., Fani,R., & Fondi, M. (2015). MeDuSa: a multi-draft based scaffolder. *Bioinformatics*, **31**(15), 2443-2451.

Canova, M.J., Pascale, G.D., Ejim, L., Kalam, L., King, A.M., Kotena, K., Morar, M., Mulvey, M.R., O' Brien, J.S., Pawlowski, A.C., Piddock, L.J.V., Spanogiannopoulos, P., Sutherland, A.D., Tang, I., Taylor, P.L., Thaker, M., Wang, W., Yan, M., Yu, T., & Wright, G.D. (2013). The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, **57**(7), 3348 – 3357.

Caratfoli, A., Zankari, E., Fernandez, A.G., Larsen, M.V., Lund, O., Villa, L., Aarestrup, F.M., & Hasman, H., (2014). In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents & Chemotherapy*. **58**(7), 3895 – 3903.

Choudhuri, S., (2014). Chapter – 7 Additional Bioinformatic Analyses involving nucleic acid sequences. *Bioinformatics for Beginners*, 157-181.

Choi, K.K., Park, C.W., Kim, S.Y., Lyoo, W.S., Lee, S.H., & Lee, J.W. (2004). Polyvinyl alcohol degradation by Microbacterium barkeri KCCM 10507 & Paenibacillus amylolyticus KCCM 10508 in dyeing wastewater. *Journal of Microbiology and Biotechnology* , **14** (5),1009-1013.

Daligault, H.E., Davenport, K.W., Minogue, T.D., Lilly, K.A.B., Broomall, S.M., Bruce, D.C., Chain, P.S., Coyne, S. R., Frey, K. G., Gibbons, H. S., Jaissle, J., Korolena, G. I., Ladver, J. T., Lo, C-C., Munk, C., Palacios, G. F., Redden, C. L., Rosenzweig, C. N., Scholz, M. B., & Johnson, S.L. (2014). Twenty Whole Genome Bacillus sp. Assemblies. *Genome Announcements*, 2(5), e00158-14.

Darling, A.E., Miklos, I., & Ragan, M.A. (2008). Dynamics of Genome Rearrangement in Bacterial populations. *PLoS Genetics*, **4**(7), e1000128.

Darling, A.C.E., Mau, B., Blattner, F.R., & Perna, N.T. (2004). Mauve - Multiple Alignment of Conserved Genomic Sequence with Rearrangements. *Genome Reseach*, **14**, 1394-1403.

Das, S.N., Dutta, S., Kondreddy, A., Chilukoti, N., Pullabhotla, S.V.S.R.N., Vadla- mudi, S., et al. (2010). Plant growth-promoting chitinolytic Paenibacillus elgii responds positively to tobacco root exudates. *Journal of Plant Growth Regulation*, **29**, 409–418.

da Mota, F.F., Gome, E.A., Paiva, E., Rosad, A.S., & Seldin, L. (2004). Use of rpoB gene analysis for identification of nitrogen-fixing Paenibacillus species as an alternative to the 16SrRNA gene. *Letters in Applied Microbiology*, **39**(1), 34-40.

da Mota,F.F., Gomes, E.A., Paiva, E., & Seldin, L. (2005). Assessment of the diversity of Paenibacillus species in environmental samples by a novel rpoB – based PCR – DGGE method. *FEMS Microbiology Ecology*, **53**(2), 317-328.

Dheeran, P., Nandhagopal, N., Kumar, S., Jaiswal, Y.K., & Adhikari, D.K. (2012). A novel thermostable xylanase of *Paenibacillus macerans* IIPSP3 isolated from the termite gut. *Journal of Industrial Microbiology and Biotechnology*, **39**(6), 851–860.

Ding, Y., Wang, J., Liu, Y., & Chen,S. (2005). Isolation & identification of nitrogen fixing bacilli from plant rhizospheres in Beijing region. *Journal in Applied Microbiology*, **99**(5), 1271-1281.

Didelot, X., & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Comparative study*, **175**(3), 1251-1266.

Doan, C.T., Tran, T.N., Nguyen, V.B., Nguyen, A.D., & Wang, S.L. (2018). Reclamation of Marine chitinous materials for chitosanase production via microbial conversion by *P.macerans*. *Marine Drugs*, **16**(11), 429.

Erisman, J.W., Sutton, M.A., Galloway, J., Klimont, Z., & Winiwarter, W. (2008). How a century of ammonia synthesis changed the world. *Nature Geoscience*, 1, 636–639.

Fang, X.M., Li, Y.S., Nie, J., Wang, C., Huang, K.H., Zhang, Y.L., She, H.Z., Liu, X.B., Ruan, R.W., Yuan, X., & Yi, Z. (2018). Effects of nitrogen fertilizer and planting density on the leaf photosynthetic characteristics, agronomic traits and grain yield in common buckwheat (Fagopyrum esculentum M.). *Field Crops Research*, 219, 160-168.

Foxman, B., (2012). Chapter – 5 A Primer of Molecular Biology. *Molecular Tools and Infectious Disease Epidemiology*, 53-78.

Gans, J.D., & Wolinsky, M. (2007). Genomorama – genome visualization and analysis. *BMC Bioinformatics*, **8**, 204.

Gary, H., Domselaar, V., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A.C., Dong, X., Lu, P., Szafron, D., Greiner, R., & Wishart, D.S. (2005). BASys – a web server forautomated bacterial genome annotation. *Nucleic Acids Research*, **33**, W455-W459.

Garcia, J., & Kniffin, J.K. (2018). Microbial Group Dynamics in Plant Rhizospheres and their implications on Nutrient Cycling. *Frontiers in Microbiology*, 9, 1516.

Galardini, M., Biondi, E.G., Bazzicalupo, M., & Mengoni, A. (2011). CONFIGuator – bacterial genomes finishing tool for structural insights on draft genomes. *Source Code of Biology and Medicine*, **6**(11), 1751.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, B. D., Lander, E. S., & Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**(2), 182-189.

Grady, E.N., Donald, J.M., Liu, L., Richman, A., & Yuan, Z.C. (2016). Current knowledge & perspectives of Paenibacillus : a review. *Microbial Cell Factories*, **15**(1), 203.

Gupta, A., Murarka, A., Campbell, P., & Gonzalez, R. (2017). Anaerobic Fermentation of Glycerol in *P. macerans* : Metabolic Pathways & Environmental Determinants. *Applied and Environmental Microbiology*, **75**(18), 5871–5883.

Helfrich, E.J.N., Reiter, S., & Piel, J. (2014). Recent advances in genome based polyketide discovery. *Current Opinion Biotechnology*, **29**, 107-15.

Hong, Y.Y., Ma, Y.C., Zhou, Y.G., Gao, F., Liu, H.C., & Chen, S.F. (2009). *Paenibacillus sonchi* sp. Nov., a nitrogen – fixing species isolated from the rhizosphere of Sonchus oleraceus. *International Journal of Systematic and Evolutionary Microbiology*, **59**(11), 2656-2661.

Hoshino, T., Nakabayashi, T., Hirota, K., Matsuno, T., Koiwa, R., Fujiu, S., Saito, I., Tkachenko, O.B., Matsuyama, H., & Yumota, I. (2009). Paenibacillus macquariensis subsp. Defensor subsp. nov., isolated from boreal soil. *International Journal of Systematic and Evoutionary Microbiology*, **59**(Pt 8), 2074-2079.

Hu, X., Chen, J., & Guo, J. (2006). Two phosphate & potassium solubilizing bacteria isolated from Tianmu Mountain, Zhejiang, China. *World Journal of Microbiology and Biotechnology*, **22**, 983-990.

Ishige, T., Krause, M., Bott, M., Wendisch, V.F., Sahm H. (2003). The phosphate starvation stimulon of Corynebacterium glutamicum determined by DNA microarray analyses. *Journal of Bacteriology*, **185** (16), 4519-4529.

Jian bo Xie, J., Shi, H., Du, Z., Wang, T., Liu, X., & Chen, S. (2016). Comparative genomic and functional analysis reveal conservation of plant growth promoting traits in *Paenibacillus polymyxa* and its closely related species. *Scientific Reports*, **6**, 21329.

Jin, H.J. & Lv, J., & Chen, S.F. (2011). *Paenibacillus sophorae* sp. Nov., a nitrogen – fixing species isolated from the rhizosphere of Sophora japonica. *International Journal of Systematic and Evolutionary Microbiology*, **61**(4), 767-771.

Junqueira, D.M., Braun, R.L., & Verli, H. (2014). Alinhamentos. In: Bioinformatica da biologia a flexibilidade molecular (Verli H, ed.). SBBq, Sao Parlo, 38-61.

Kehrenberg, C., & Schwarz, S. (2004). fexA, a Novel Staphylococcus lentus gene encoding resistance to florfenicol and chloramphenicol. *Antimicrobial Agents Chemotherapy*, **48**(2), 615-618.

Khan, N., Mishra, A., Chauhan, P. S., Sharma, Y. K., & Nautiyal, C. S. (2012). Paenibacillus lentimorbus enhance growth of chickpea (Cicer arietinum L.) in chromium - amended soil. *Antonie Leeuwenhoek*, **101**(2), 453-459.

Kobayashi, H., Tanizawa, Y., Sakamoto, M., Nakamura, Y., Ohkuma, M., & Tohno, M. (2019). Reclassification of Paenibacillus thermophilus Zhou et al. 2013 as a later heterotypic synonym of Paenibacillus macerans (Schardinger 1905) Ash et al.1994. **62**(2), 417 – 421.

Kolmogorov, M., Raney, B., Paten, B., & Phem, S. (2014). Ragout – a reference assisted assembly tool for bacterial genomes. *Bioinformatics*, **30**(12), i302-i309.

Kong, J., Huh, S., Won, J.I., Yoon, J., Kim, B., & Kim, K. (2019). GAAP : A Gene Assembly + Annotation Pipeline. *Hindawi BioMed Research International*, **12**.

Konishi, J., & Maruhashi, K. (2003). 2-(2'-Hydroxyphenyl) benzene sulfinate desulfinase from the thermophilic desulfurizing bacterium Paenibacillus sp. strain AII – 2: purification and characterization. *Applied in Microbiology and Biotechnology*, **62**(4), 356-361.

Lancaster, C.R.D., in Reference Module in Biomedical Sciences (2018). Respiratory Chain Complex II & Succinate : Quinone Oxidoreductases. *Encyclopedia of Biological Chemistry*, 2004, 681-687.

Lee, M., Ten, L.N., Baek, S.H., Im, W.T., Aslam, Z., & Lee, S.T. (2007). Paenibacillus ginsengisoli sp. nov., a novel bacterium isolated from soil of a ginseng field in Pocheon Province, South Korea. *Antonie Van Leeuwenhoek*, **91**(2), 127-135.

Lee, S., Reth, A., Meletzus, D., Sevilla, M., & Kennedy, C. (2001). Characterization of a Major Cluster of nif, fix, and Associated Genes in a Sugarcane Endophyte, Acetobacter diazotrophicus. *Journal of Bacteriology*, **182**(24), 7088-7091.

Lewin, H.A., Robinson, G.E., Kress, W.J., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(17), 4325-4333.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive: International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, **39**, D19-D21.

Patel, R.K., & Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLOS ONE*, **7**(2), e30619.

Li, B., Su, T., Yu, R., Tao, Z., Wu, Z., Algam, S.A.E., Xie, G., Wang, Y., & Sun, G., (2017). Inhibitory activity of Paenibacillus macerans and Paenibacillus polymyxa against Ralstonia solanacearum. *Advanced Journal of Microbiology Research*, **12**(2), 001-007.

Li, J.Y., Gao, T.T., & Wang, Q. (2020). Comparative & functional Analyses of two sequenced P. polymyxa genomes provides insights into their potential genes related to plant growth promoting features & biocontrol mechanisms. *Frontiers in Genetics*, **11**, 564939.

Liang, T.W., Wu, C.C., Chang, W.T., Chen, Y.C., Wang, C.L., & Wang, I.L. (2014). Exopolysaccharides and antimicrobial biosurfactants produced by *P.macerans* TKU029. *Applied Biochemistry and Biotechnology*, **172**(2), 933-950.

Liu, X., Li, Q., Li, Y., Guan, G., & Chen, S. (2019). Paenibacillus strains with nitrogen fixation & multiple beneficial properties for promoting plant growth. *Peer Journal*, **7**, e7445.

Loman, N.J., Constantinidou, C., Chan, J.Z., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R., & Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, **10**, 599–606.

Li, Y., Zhang, J., Gong, Z., Xu, W., & Mou, Z. (2019). *Gcd* Gene Diversity of Quinoprotein Glucose Dehydrogenase in the Sediment of Sancha Lake and Its Response to the Environment. *International Journal of Enivronmental Research and Public Health*, **16**(1), 1.

Mamphogoro, T.P., Babalola, O.O., & Aiyevoro, O.A. (2020). Sustainable management strategies for bacterial wilt of sweet peppers (*Capsicum annum*) & other Solanaceous crops. *Journal of Applied Microbiology*, **129**(3), 496-508.

Marra, L.M., Sousa Soares, C.R.F., de Oliveira, S.M., Ferreira, P.A.A., Soares, B.L., de Carvalho, R.F., et al (2012). Biological nitrogen fixation and phosphate solubilization by bacteria isolated from tropical soils. *Plant Soil*, **357**, 289–307.

McArthur, A.G., Waglechner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar,K., Darling, A.E., Mau, B., & Perna, N.T. (2010). ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLOS One*, **5**, e11147.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents Chemotherapy*, **57**, 3348–3357.

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

Mohamed, I., Eid, K.E., Abbas, M.H.H., Salem, A.A., Ahmed, N., Ali, M., Shah, G.M., & Fang, C. (2019). Use of PGPR & mycorrhizae to improve the growth & nutrient utilization of common bean in a soil infected with white rot fungi. *Ecotoxicology and Environmental Safety*, **171**, 539-548.

Montes, M.J., Mercade, E., Bozal, N., & Guinea, J. (2004). Paenibacillus antarcticus sp. nov., a novel pyschrotolerant organism from the Antarctic environment. *International Journal of Systematic and Evolutionary Microbiology*, **54**(Pt 5), 1521-1526.

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.

Mohamadi, S., Bostanabad, S.Z., & Mirnejad, R. (2020). CRISPR Arrays: A Review on Its Mechanism. *Journal of Applied Biotechnology Reports*, **7**(2), 81-86.

Mobley, H.L.T., Mendz, G.L., & Hazell, S.L., (2001). Helicobacter pylori: Physiology and Genetics. *ASM Press*.

Mus, F., Alleman, A.B., Pence, N., Seefeldt, L.C., & Peters, J.W. (2018). Exploring the alternatives of biological nitrogen fixation. *Metallomics*, **10**(4), 523–538.

Navarro, Y.E., Hernandez, N.E., Morales, M.J., Jan, J.J., Martinez, R.E., Hernandez, R.C., & Rodriguez (2012). Isolation and characterization of nitrogen fixing heterotrophic bacteria from the rhizosphere of pioneer plants growing on mine tailings. *Applied Soil Ecology*, **62**, 52-60.

Oren, A., & Garrity, G.M. (2019). Notification that new names of prokaryotes, new combinations, and new taxonomic opinions have appeared in volume 69, part 2 of the IJSEM. *International Journal of Systematic Evolutionary and Microbiology*, **69**, 1251–1252.

Olajide, A.M., Chen, S., & Pointe, G. L. (2020). Draft Genome Sequences of Five Paenibacillus Species of Dairy Origin. *Microbiology Resource Announcements*, **9**, e00971-20.

Overmars, L., Kerkhoven, R., Siezen, R.J., & Francke, C. (2013). MGcV – the microbial genomic context viewer for comparative genome analysis. *BMC Genomics*, **14**, 209.

Pasari, N., Gupta, M., Eqbal, D., & Yazdani, S.S. (2019). Genome analysis of Paenibacillus polymyxa A18 gives insights into the features associated with its adaptation to the termite gut environment. *Scientific Reports*.

Pandya, M., Rajput, M., & Rajkumar, S. (2015). Exploring plant growth promoting potential of non rhizobial root nodules endophytes of Vigna radiata. *Microbiology*, **84**, 80–89.

Patel, D.K., Archana, G., & Kumar, G.N. (2008). Variation in the nature of organic acid secretion & mineral phosphate solubilization by Citrobacter sp. DHRSS in the presence of different sugars. *Current Microbiology*, **56**(2), 168-174.

Passera A, Marcolungo L, Casati P, Brasca M, Quaglino F, Cantaloni C, et al. (2018) Hybrid genome assembly and annotation of Paenibacillus pasadenensis strain R16 reveals insights on endophytic life style and antifungal activity. *PLoS ONE*, **13**(1), e0189993.

Patten, C.L., Blakney, A.J.C., & Coulson, T.J.D. (2013). Activity, Distribution & Function of Indole – 3 – acetic acid biosynthetic pathways in bacteria. *Critical Reviews in Microbiology*, **39**(4), 395-415.

Peng, Y., Lai, Z., Lane, T., Rao, M.N., Okada, M., Jasieniuk, M., Geen, H.O., Kim, R.W., Sammons, R.D., Rieseberg, L.H., & Stewart, C.N.Jr. (2014). De Novo Genome Assembly of the Economically Important Weed Horseweed Using Integrated Data from Multiple Sequencing Platforms. *Plant Physiology*, **166**, 1241–1254.

Piuri, M., Rivas, C.S., & Ruzal, S.M. (1998). A novel antimicrobial activity of a Paenibacillus polymyxa strain isolated from regional fermented sausages. *Letters in Applied Microbiology*, **27**(1), 9-13.

Pragai, Z., Allenby, N.E., O'Connor, N., Dubrac, S., Rapoport, G., Msadek, T., & Harwood, C.R. (2004). Transcriptional regulation of the phoPR operon in Bacillus subtilis. *Journal of Bacteriology*, **186**, 1182-1190.

Ramesh, M.N. (2003) in Encyclopedia of Food Sciences & Nutrition. Sterilization of Foods.

REGNUM PROKARYOTE. ABIS Encyclopedia. Tgw 1916.net. Retrieved 19 Dec. 2014.

Rodríguez, H., & Fraga, R. (1999). Phosphate solubilizing bacteria and their role in plant growth promotion**.** *Biotechnology Advances*, **17**(4-5), 319-339.

Rodriguez – Jerez, J.J., Giaccone, V., Colavita, G., & Parisi, E. (1994). Bacillus macerans – a new potent histamine producing microorganism isolated from Italian cheese. *Food Microbiology*, **11**, 409-415.

Santos-Beneit, F. (2015). The Pho regulon : a huge regulatory network in bacteria. *Frontiers in Microbiology*, **6**, 402.

Saez – Nieto, J.A., Medina – Pascual, M.J., Carrasco, G., Garrido, N., Fernandez – Torres, M.A., Villalon, P., & Valdezate, S. (2017). Paenibacillus spp. Isolated from humans & environmental samples in Spain : detection of 11 new species. *New Microbes New Infections*, **19**, 19-27.

Schatz, M.C., Delcher, A.L., & Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome Research, **20**(9), 1165-1173.

Seemann T. (2013). Barrnap 0.7: rapid ribosomal RNA prediction.

Seldin, L., (2011). Paenibacillus, nitrogen fixation & soil fertility. In: Logrn NA, De Vos P, eds. Endospore–forming Soil Bacteria. *Soil Biology*, **27**, 287-307.

Sharma, S.B., Sayyed, R.Z., Trivedi, M.H., & Gobi, T.A. (2013). Phoshphate Solubilizing Microbes : sustainable approach for managing Phosphorus deficiency in agricultural soils. *SpringerPlus*, **2**, 587.

Shi, L., & Hulett, F. M. (1998). The cytoplasmic kinase domain of PhoR is sufficient for the low phosphate inducible expression of Pho regulon genes in Bacillus subtilis. *Molecular Microbiology*, **131**, 211-222.

Shtratnikovaa, V.Y., Rudenskaya, Y.A., Gerasimov, E.S., Schelkunovc, M.I., Logachevaa, M.D., & Kolesnikov, A.A., (2020). Complete genome assembly data of paenibacillus sp. RUD330, a hypothetical symbiont of euglena gracilis. *Europe PMC*, **32**, 106070.

Silva, F.V.M., Gibbs, P.A., Nunez, H., Almonacid, S., R. Simpson in Encyclopedia of Food Microbiology (second edition) (2014). Thermal Processes/ Pasteurization.

Smits, T.H.M. (2019). The importance of genome quality to microbial comparative genomics. *BMC Genomics*, **20**, 662.

Stamereilers, C., Wong, S., & Tsourkas, P.K. (2021). Characterization of CRISPR Spacer and Protospacer Sequences in Paenibacillus larvae and Its Bacteriophages. . *Viruses* , **13**, 459.

Stewart, A.C., Osborne, B., & Read, T.D. (2009). DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics*, 25, 962-963.

Szaniawski, M.A., & Spivak, A.M. (2019). Recurrent Paenibacillus infection. *Oxford Medical Case Reports*, **5**, 034.

Taniguchi, H., & Honnda, Y. (2009). Amylases. *Encyclopedia of Microbiology* (Third Edition), 159-173.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., & Medigue, C. (2006). MaGe – a microbial genome annotation system supported by synteny results. *Nucleic Acids Research*, **34**(1), 53-65.

Wang, Y., Shi, Y., Li, B., Shan, C., Ibrahim, M., Jabeen, A., Xie, G., & Sun, G. (2012). Phosphate solubilization of P. polymyxa & P.macerans from mucorrhizal & non – mycorrhizal cucumber plant. *African Journal of Microbiology Research*, **6**(21).

Wang, X., & Wood, T. K. (2016). Cryptic prophages as targets for drug development. *Drug Resistance Updates*, **27**, 30–38.

Wen, Y., Wu, X., Teng, Y., Qian, C., Zhan, Z., Zhao, Y., et al., (2011). Identification & analysis of the gene cluster involved in biosynthesis of paenibactin, a catecholate siderophore produced by *P.elgii* B69. *Environmental Microbiology*, **13**, 2726-2737.

Xie, J.B., Du, Z., Bai, L., Tian, C., Zhang, Y., Xie, J.Y., Wang, T., Liu, X., Chen, X., Cheng, Q., Chen, S., & Li, J. (2014). Comparative genomic analysis of N2- fixing and non-N2 fixing Paenibacillus spp.: organization, evolution and expression of the nitrogen fixation genes. *PLoS Genetics*, **10**, e1004231.

Zheng, B., Zhang, F., Dong, H., Chai, L., Shu, F., Yi, S., Wang, Z., Cui, Q., Dong, H., Zang, Z., Hou, D., Yang, J., & She, Y. (2016). Draft genome sequence of Paenibacillus sp. strain A2. *Standards in Genomic Sciences*, **11**, 9.

Zhou, Y., Gao, S., Wei, D.Q., Yang, L.L., Huang, X., He, J., Zhang, Y.J., Tang, S.K., & Li, W.J. (2012). Paenibacillus thermophilous sp. Nov., a novel bacterium isolated from a sediment of hot spring in Fujian province, China. *Antonie Van Leeuwenhoek*, **102**(4), 601-609.

Zhou., Liang,Y., Lynch,K.H., Dennis,J.J., & Wishart,D.S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Research*, **39**, W347–W352.