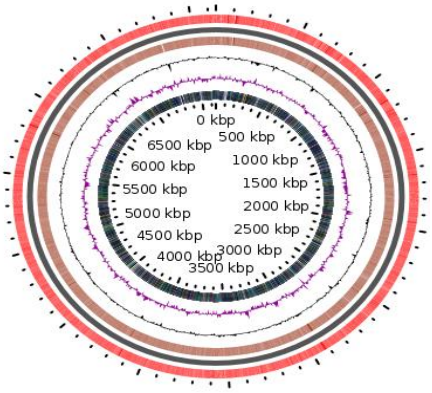
Chapter 6

Whole-genome sequencing of *Streptomyces* sp. S-9 and Transcriptome analysis of Pigeon pea in control and wilt condition



Chapter 6

6 Whole-genome sequencing of *Streptomyces* sp. S-9 and Transcriptome analysis of Pigeon pea under control and wilt condition

6.1 Introduction

Streptomyces is the source of the vast majority of the chemicals of microbial origin that have antibacterial, anticancer, or immunosuppressive properties that have been found to date (Braña, et al. 2015). It has been hypothesized that these bacteria could create a great deal more metabolites than those that have been uncovered to this point. In addition to their medical use, *Streptomyces* bacteria have important ecological and environmental implications (Sayed et al. 2020). Because of their capacity to break down a wide variety of organic molecules, these bacteria are often regarded as being among the most important participants in the process of biomass breakdown. The rise of dangerous diseases that are resistant to several drugs is a problem that the world is now experiencing. It's important to find metabolically active species capable of providing novel secondary metabolites. The ecological and physiological effects offered by secondary metabolites produced by the bacteria are significant. Their contribution is especially important in severe environments, where bacteria have adapted to live and multiply (Makhalanyane et al. 2015; Lo Giudice et al. 2015). There is a possibility that many species living in harsh conditions might be new taxa. These bacteria could also be a useful resource for discovering new bioactive chemicals and enzymes with potential commercial uses. The genus *Streptomyces*, which accounts for fifty % of the entire population of soil actinobacteria, is the most prevalent of all actinobacterial species (Parte et al. 2020).

It is widely recognized for generating a significant quantity of different bioactive metabolites. *Streptomyces* was responsible for the discovery of about 75% of all of the antibiotics that were reported (Olanrewaju and Babalola 2019).

Numerous strains of *Streptomyces* are regarded to be biocontrol agents due to the fact that they efficiently colonize the rhizosphere of a wide variety of plant species, one of which is rice, generate a broad spectrum of antimicrobials, and are able to survive in harsh environments (Qin et al. 2011; Kinkel et al. 2012). This finding suggests that chromosomal linearity is likely

widespread among streptomycetes (Huang et al. 1998). The majority of the *Streptomyces* chromosomal DNA molecules have a length of around 8 megabases, and the 5' end is thought to include terminal-inverted repeats as well as covalently associated terminal proteins. When compared to other well-known microorganisms like *Escherichia coli* and *Bacillus subtilis*, this size is enormous for a bacterium to have. The G + C concentration of streptomycetes is greater than that of almost every other kind of creature by more than 70 % age points. Therefore, the chromosome of *Streptomyces* is one of a kind, both in terms of its structure and its size.

Here, we discuss the genome-wide study structure and sequencing features of *Streptomyces* species. The aim of the study is to see the novel genes study using Genome annotation and assembly of those genes which are related to secondary metabolism. We place great emphasis on the description of this microorganism's secondary metabolite production. In addition, whole-genome sequencing and bioinformatics tools were used to analyze the genomic DNA and protein sequences of *Streptomyces* strain from Pigeon pea.

Multilocus sequence typing (MLST) is an unambiguous procedure for characterizing isolates of bacterial species using the sequence of internal fragments of sixhousekeeping genes.

In a number of crop species, genomic approaches have been effective at overcoming production barriers (Varshney et al., 2013; Kole et al., 2015). Given the draught genome sequence for pigeon pea (Varshney et al., 2012) and a variety of genomic resources, it is now possible to conduct genomics-assisted breeding (GAB). The genome sequence, molecular markers, genetic maps, quantitative trait loci (QTLs), and transcriptome assembly are among these genomic resources (Pazhamala et al., 2015). Additionally, in pigeon pea, expression studies have been carried out using transcriptome sequencing and quantitative real-time PCR to understand the plant's response to biotic stresses (fusarium wilt and sterility mosaic disease; Singh et al., 2016) as well as abiotic stresses (drought and salinity); Sinha et al., 2015) (Pazhamala et al., 2016).

The transcriptome is the whole complement of ribonucleic acid (RNA) transcripts in a cell, which includes both coding (1–4 % - messenger) and non-coding (>95 percent -ribosomal, transfer, small nuclear, small interfering, micro, and long-non-coding) RNAs (Berg et al., 2007; Mattick & Makunin, 2006). Transcriptomics consists of comprehensive transcript cataloguing,

dynamic transcript profiling, and study of gene expression regulatory networks. It is possible to analyse the transcriptome using a variety of methods, including cDNA-AFLP, sequence tag-based technologies such as Expression Sequence Tags (EST), Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS), Open Reading Frame EST, and digital expression analysis with next-generation sequencer (RNAseq). Plant genome assemblies housed at the National Center for Biotechnology Information (NCBI) are in the hundreds (Bolger et al., 2017). The genome of the tomato contains 296,963 ESTs and 18,346 unigenes, while the genome of the potato contains 237,320 and 18,825, respectively. RNA sequencing is an efficient next-generation sequencing technique (Wang et al., 2009).

RNA sequencing also allows for gene expression investigation without prior information (de novo) of the transcriptome. qPCR and microarrays, on the other hand, rely on past knowledge and cannot uncover unique information on a wide scale. RNA-Seq provides a unique mix of coverage of the whole transcriptome, sensitivity, and discovery potential. RNA-Seq can be utilised to investigate functional pathways, isoforms, and resistance. Gene RNA polymerase activity is controlled by transcription factors. Gene expression is a very dynamic process that permits a tissue or organism to develop stress resistance by expressing specific genes. The transcripts correspond to the gene in a complimentary manner. The transcription factors govern the activity of RNA polymerase. mRNA is translated into proteins, and the differential expression of genes is recorded. PCR is used to quantify gene expression in real-time (RT-PCR). RT-PCR identifies the intensity and pattern of an illness. Multiplex RT-PCR employs fluorochromes for gene identification, whereas transcriptome analysis begins with RNase- and temperature-sensitive RNA extraction. The KEGG (Kyoto Encyclopedia of Genes and Genomes) is a reference database for gene-related information and biological pathways (Kanehisa et al., 2019).

The database enhances knowledge about conserved genes, evolved genes, and genomes in many species. KO (KEGGOrthology) represents biological processes with conserved characteristics. Since the publication of KEGG in 1995, databases pertaining to biological pathways, genes, chemicals, and enzymes have been improved. KEGG now maintains 18 databases, including KEGGBRITE, MODULE, GLYCAN, and REACTION, among others.

MicroRNAs (miRNAs) are a family of 20-24 nucleotide-long non-coding RNAs generated from single-stranded RNA precursors that can form stem-loop structures. They regulate gene expression post-transcriptionally through translational repression or target degradation, resulting in gene silencing (Jones-Rhoades et al., 2006). The status of target mRNA is determined by the degree of complementarity between the miRNA and target mRNA (Bentwich, 2005). MiRNAs were first identified in the soil nematode *Caenorhabditis elegans* in 1993, and they were also found in plants in 2002 (Reinhart et al., 2002). According to current knowledge, miRNAs are engaged in several cellular, biochemical, and metabolic processes, such as defining cell fate and differentiation, organ development, phase change regulation, disease and environmental stress response, auxin signalling, and reproductive development (Sun, 2012).

In 2012, the genome of Pigeonpea was sequenced to expedite the application of genomics to crop enhancement, which resulted in a massive data explosion (Varshney et al., 2012). As many biological processes depend on the temporal and spatial control of genes present in every organism, the transcriptome data obtained from pigeonpea led to high-throughput gene expression research (Dubey et al., 2011; Kudapa et al., 2012).

The present study focuses on identifying and characterizing Pigeonpea miRNAs using bioinformatics approaches and understanding their role in wilt tolerance. The conserved miRNA genes were identified from the draft whole-genome sequence of Pigeon pea hosted at NCBI (http://www.ncbi.nlm.nih.gov/genome/?term=cajanus+cajan) using bioinformatics approaches. Moreover, we attempted to establish the role of candidate miRNAs in wilt stress tolerance by analysing various physiological parameters and gene expression analysis of miRNAs and corresponding target transcripts.

6.2 Materials and methods

6.2.1 Genomic DNA (gDNA) extraction

A total of 50 millilitres (mL) of liquid culture medium was inoculated with streptomycetes and then cultured in an orbital shaker at a temperature of thirty degrees Celsius. After harvesting 25 mL of cultured cells while they were in the exponential growth phase, the cells were washed twice with 10 mM EDTA, and then they were subjected to a 45-minute lysozyme treatment at 37

degrees Celsius. gDNA was isolated by employing QiAmp, which is the DNA Purification Kit offered by Qiagen (USA). Electrophoresis on an agarose gel containing 1 % and a Nanodrop 1000 instrument were used to analyse gDNA samples (Thermo Scientific, USA). As a method of measuring DNA purity, OD260/OD280 nm and OD260/OD230 (>2.0) were utilised.

6.2.2 Whole-genome sequencing of S-9

Streptomyces genome was sequenced using Illumina Hiseq. Nextera XT sample preparation kit (Illumina, San Diego, CA, USA) was used to prepare Illumina sequencing library from genomic DNA. After fragmented DNA samples were cleaned and end-repaired, adaptor ligation and bead-based size selection were performed. The library size was evaluated using an Agilent 2200 Bioanalyzer. Illumina's library was sequenced using paired-end sequencing kits on their in-house Illumina Hiseq. The base calling software trimmed Illumina adapters.

6.2.3 Quality control and assembly of the S-9 genome

FASTQ files must undergo quality control and preprocess in order to provide clean data for subsequent analysis. We utilized fastp, a very quick FASTQ preprocessor with practical quality control and data-filtering capabilities (Chen et al. 2018). With just one scan of the FASTQ data, it can complete processes like quality control, adapter cutting, quality filtering, per-read quality pruning, and many more. This utility was created in C^{++} and supports many threads. A sequence assembler's objective is to create lengthy contiguous sequences (contigs) from these reads. Contigs may then be arranged and positioned in respect to one another to create scaffolds. Using the SPAdes (Bankevich et al. 2012) assembler with the default parameters, the high-quality (HQ), filtered reads from each of the seven samples were individually assembled.

Using the FastQ C (v. 0 11.5) tool, the quality of each raw file produced after sequencing was examined. Using the FASTX toolkit, (accessible at http://hannonlab.cshl.edu/fastx_toolkit/), ambiguous bases and low-quality readings were diverted. High-quality readings were assembled from scratch using the SPAdes (v. 3.11.1) program. Utilizing the QUAST (v. 5.0.2) tool, the constructed sequence's quality was evaluated. For the gene annotation, the Prokka tool (version 1.13.30, accessible at: https://github.com/tseemann/prokka) was utilized. The Centre for Genomic Epidemiology (CGE) toolbox's MLST (v.2.0.4, accessible at:

https://cge.cbs.dtu.dk/services/MLST/) was used to calculate multilocus sequence typing (MLST) for an isolate based on six housekeeping genes.

6.2.4 Genome Annotation and Assembly

Prokka 1.11 was used for genome annotation (https://github.com/tseemann/prokka) (Seemann, 2014) and RAST Rapid Annotation using Subsystem Technology (https://rast.nmpdr.org/) (Aziz et al., 2008). The RAST server classified the bacterial genome's subsystem characteristics, which mostly concerned amino acids, cofactors, vitamins, prosthetic groups, etc. Unassigned proteins were sent to the nr database for information on their alleged functions. The MEGA7 performed the phylogenetic analysis using a 1000 bootstrap value. The EggNOG orthologous database used HMMER profiling to analyze Gene Ontology and KEGG pathways. The genome's secondary metabolites gene clusters were identified using the AntiSMASH platform.

6.2.5 Antibiotic Resistance

ResFinder and CARD were used to find the antibiotic-resistant gene in the genome. Open Reading Frame (ORF) prediction uses Prodigal, homolog discovery uses DIAMOND, and resistance profiling uses CARD-curated bitscore cut-offs. Screening the assembled genome's predicted protein sequence for antibiotic resistance genes using RGI 5.1.1 and CARD 3.1.0.

6.2.6 COG Analysis

For the purpose of clusters of orthologous groups (COG) annotation, functional categories of coding sequences (CDSs) were identified using WebMGA by utilizing the RPSBLAST software (with an applied threshold of 1e5).

6.2.7 Secondary Metabolite

Bacterial secondary metabolism produces bioactive molecules with medicinal potential. It contains biosynthetic pathways of antibiotics, cholesterol-lowering medicines, and anti-tumor medications. These chemicals could be drugs. The genes that code for the secondary metabolite biosynthesis pathway are often grouped on the chromosome. Many species have a "secondary metabolite biosynthesis gene cluster." These 24 genomic architecturesmake it easy to locate

secondary metabolite manufacturing routes by identifying gene clusters. antiSMASH screens secondary metabolites. AntiSMASH accurately identifies gene clusters that code for secondary metabolites of all known chemical groups because it uses profile hidden Markov models of genes (Blin et al. 2017). These models target certain gene clusters.

6.2.8 Transcriptome analysis and miRNA study

6.2.9 Plant material used

The Pigeon pea (BDN 2) seeds were obtained from Model Farm, AAU, Vadodara. These seeds were germinated on Petri dish followed by sowing in pots, and were kept in a greenhouse. After four weeks of sowing, pots were divided into three groups and subjected to different levels of fungal stress to standardize wilt imposition.

6.2.10 RNA Extraction and quality check

Extracted RNA quantity was checked on Qubit 4.0 fluorometer (Thermofisher #Q33238) using RNA HS assay kit (Thermofisher #Q32851) following manufacturer's protocol. To measure the purity of the extraction, we also measured the concentration of RNA on Nanodrop 1000. The quality of the quantified RNA was confirmed on 1.5 % agarose gel. In brief, 30 ng of the RNA was mixed with 2 ul of 6x Loading dye (Invitrogen) and subjected to electrophoresis at 120 volts for 30 mins.

6.2.11 cDNA preparation

To prepare the cDNA, we evaluated the concentration of the RNA using a qubit fluorometer. The obtained results were provided in the following table. We used 50ng of the total RNA to convert the RNA into cDNA. Conversion of the RNA to cDNA was performed using superscript III Reverse transcriptase (ThermoFisher #18080093) by following the manufacturer's protocol.Primers were ordered through Sigma, and the provided sequence for the primers is presented in Table 6.1

Table 6.1: Primer Sequence

Primer serial number	Primer name	Primer Sequences
1	Cajanus-GR466360-FW	ACCTGTTTTGTTCGCCTTGT
2	Cajanus-GR466360-Rev	CGGGATCGTAGTGAAAATGGT
3	Fusarium-GR464381-Fw	CCAAGGGAAAAACTTGGTGGC
4	Fusarium-GR464381-Rev	AAGGAAAAACCCTCTCCCCG
5	Fusarium-GR4652931-Fw	GCGCAAAACGGACACAATCC
6	Fusarium-GR465293-Rev	AAAGATGAGCCGGAGAACGG

6.2.12 qRT-PCR

qRTPCR was performed using Cobas480 (Roche) using Sybergreen dye (KCQS00). The reaction setup was done as manufacturer's protocol. The optimum Tm was identified at 60C. The final obtained Ct values are provided in the below table. The result cannot be analyzed as the internal control genes are not amplified.

6.2.13 Novel Gene Prediction

Mapping information from all samples is combined and placed as input into the regular Cufflinks assembler. The assembled transfrags are then compared to the reference transcripts to determine if they are sufficiently different to be considered novel. In brief, in this process we can (1) identify novel genes, (2) identify novel exons of known genes, and (3) optimize the start and end information of known transcripts. The outputs are provided as GTF files; more information about GTF format is available at (http://mblab.wustl.edu/GTF22.html).

6.2.14Functional Analysis

Through the enrichment analysis of the differential expressed genes, we can find out which biological functions or pathways are significantly associated with differential expressed genes. The clusterProfiler (Yu G, 2012) software was used for enrichment analysis, including GO Enrichment, DO Enrichment, KEGG and Reactome database Enrichment.

6.2.15 GO Enrichment Analysis

GO is the abbreviation of Gene Ontology (http://www.geneontology.org/), which is a major bioinformatics classification system to unify the presentation of gene properties across all species. It includes three main branches: cellular component, molecular function and biological process. GO terms with padj< 0.05 are significant enrichment.

6.2.16 KEGG Enrichment Analysis

The interactions of multiple genes may be involved in certain biological functions. KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.jp/) is a collection of manually curated databases containing resources on genomic, biological-pathway and disease information (Kanehisa,2008). Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed genes, comparing the whole genome background. KEGG terms with padj< 0.05 are significant enrichment

6.3 Results

6.3.1 Whole genome sequencing

6.3.2 Phylogenetic analysis and General Genome features of S-9

The anticipated 16S rRNA sequence was matched with the 16S rRNA database at the NCBI, and the top 50 aligned hits were selected to create a tree. Multiple alignments were carried out using conserved regions after concatenating all the sequences. Phylogeny was created using MAFFT version 7's online tool (Fig. 2). MEGA7 built the tree based on the Genera+l Time Reversible model and the Maximum Likelihood approach. The S-9 strain was grouped with other members of the *Streptomyces* genus. Table 6.2 provides a list of S-9's general characteristics. Using the Illumina Hiseq technology, the entire genomes of S-9 were sequenced, and Busco verified the gene completeness (Seppey et al. 2019). The GC content of the genome, which has a total size of 0.97 Mb, is 72.60 %. A 1total of 214 genes are expected to make up S-9's genome, with 77 rRNA and 9 tRNA genes, respectively, making up the full set (Fig 3). The whole genome of *Streptomyces* sp. S-9 was submitted to NCBI with a Bioproject ID PRJNA695540 and Biosample ID SAMN17616131.

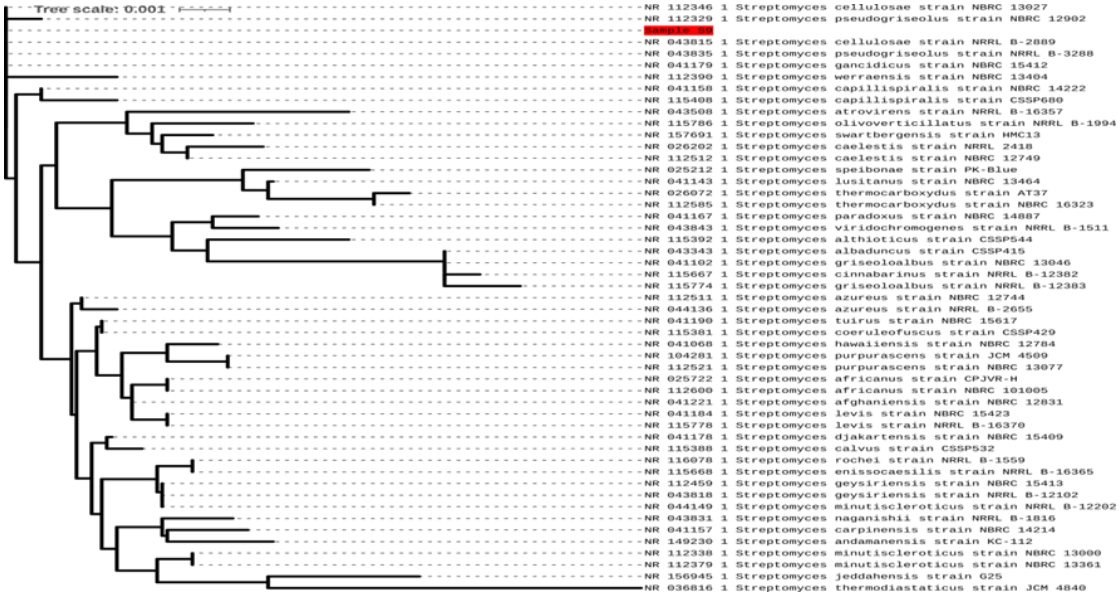


Fig 6.1: Phylogenetic analysis 16S rRNA sequence proclaimed that S-9 belonged to the genus *Streptomyces* and species *pseudogriseolus*. Scale bar shows 0.05 substitutions per site for bootstrap values**.**

6.3.3 Genome Annotation

Prokka version 1.11 was used to annotate the S-9 genome, finding 6872 protein-coding genes (CDSs), 77 tRNA, and 9 rRNA genes shown in Table 6.1. Plasmids weren't discovered. The closest relatives based on the highest similarity rate to the *Streptomyces* species were identified using the BlastN run of the entire S-9 genome. *Streptomyces pseudogriseolus* 100 %, which had a 99 % identity to S-9, was the closest strain.The house keeping genes of *Streptomyces* sp. (atpD, gyrB, recA, rpoB, trpB) were used for determining the MLST of *Streptomyces* sp. (S-9) that is ST93. However the housekeeping genes in MLST profile with *Streptomyces* sp. match with ST 93 shown in Figure 6.2.

Table 6.2: General genome features of *Streptomyces* sp. S-9

Genome size (base pairs)	142393
GC content(%)	72.60%
tRNA genes	77
rRNA genes	09

Total genes	6872

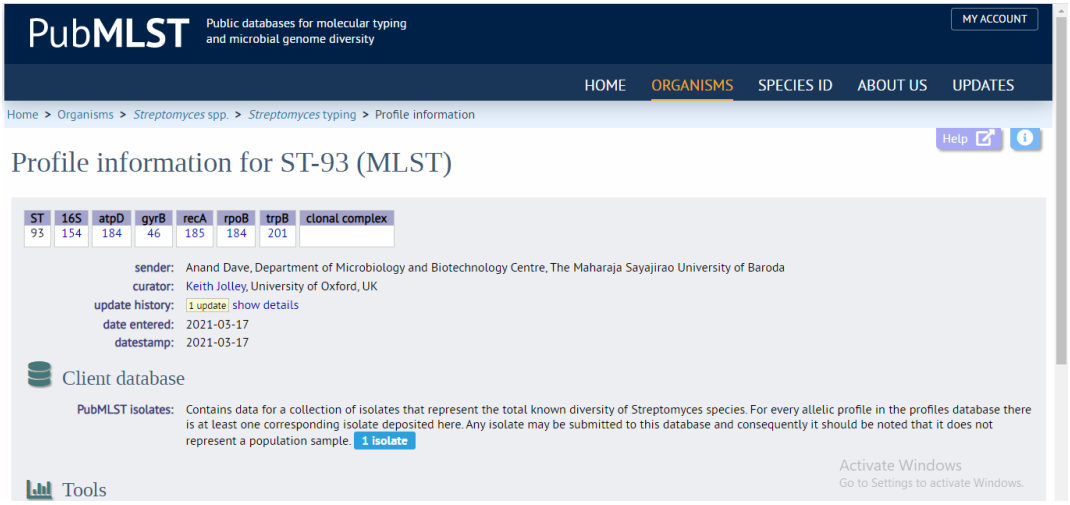


Figure 6.2: PubMLST ST Number

6.3.4 Average nucleotide identity (ANI)

ANI was calculated using the nearest reference genome for the sample's assembled genome using ChunLab's online Average Nucleotide Identity (ANI) calculator (https://www.ezbiocloud.net/tools/ani). ANI is used to compare prokaryotic genome sequences when classifying and identifying bacteria. An ANI above 95% indicates two genomes are the same species. The ANI and BlastN results were used to strengthen taxonomy identification, resulting in an ANI match of over 83%. Sample S-9 uses *Streptomyces azureus* ATCC 14921 because *Streptomyces pseudogriseolus* assembly is unavailable. That is the reason for such a low OrthoANIu value (89%).

6.3.5 Average Amino-acid Identity (AAI)

Average Amino-acid Identity (AAI) was carried out for the S-9 species in order to achieve a better resolution in illuminating taxonomic structure beyond the species rank (Fig 6.3). Since DNA-DNA reassociation values, which are the traditional method for identifying species in prokaryotes, and the rate of genome mutation exhibit a strong correlation, average amino-acid identity (AAI) represents a very reliable indicator of the genetic and evolutionary relatedness

between two strains. The outcome amply demonstrated the S-9's resemblance to *Streptomyces* species (Fig 6.3).

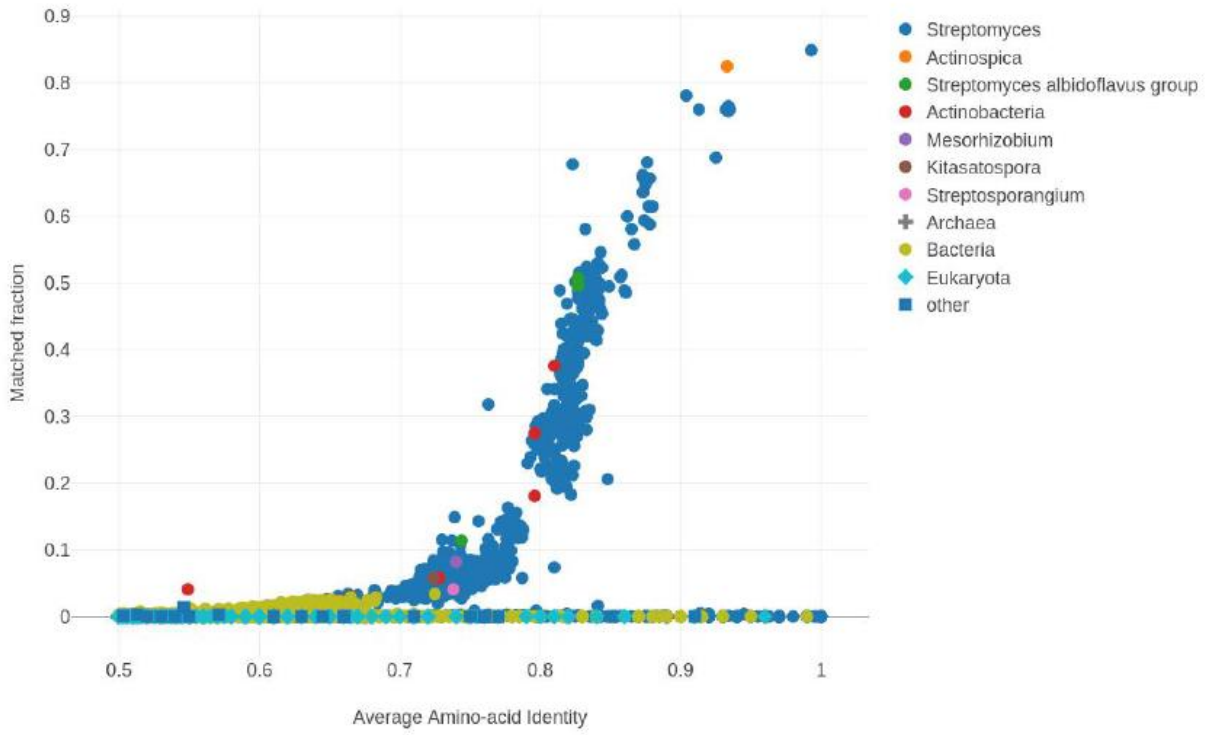


Fig 6.3: Dot plot showed the similarity index of S-9 species

Genomic map comparison (Fig 6.4) using CGViewer webserver clearly indicates Circular genome of *Streptomyces* sp. S-9. The outer circle indicates the nucleotide base positions, the next inner circle represents the forward cds, the pink circle represents the reverse cds, the blue lines, and green lines indicate the tRNA and rRNA positions respectively, and the second circle from the inside represents the GC skew and the innermost circle represents the GC plot.

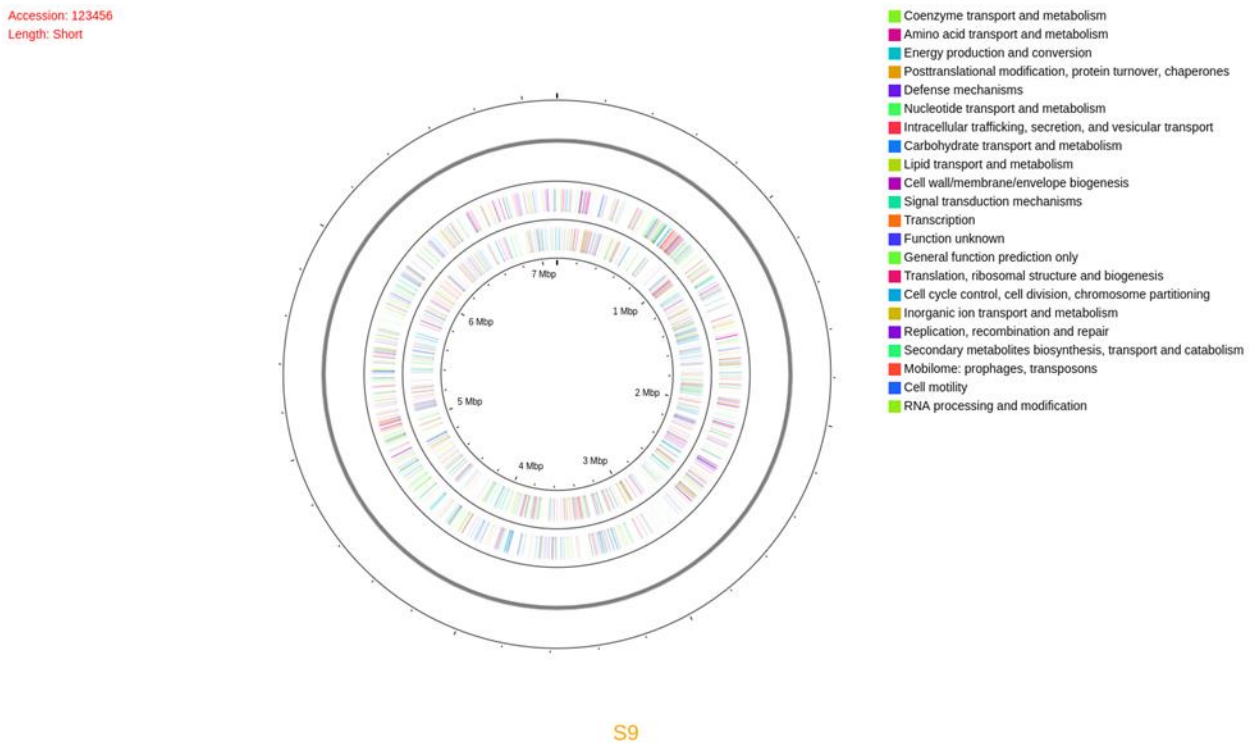


Fig.6.4: Genomic map of *Streptomyces* sp. S-9 as a reference sequence blasted with circular genome

6.3.6 Subsystem features of S-9

On the S-9 genome, subsystem feature analyses were carried out. Utilizing the RAST annotation server (Rapid Annotation using Subsystem Technology, http://rast.nmpdr.org/), the Draft genome was submitted to gene prediction and annotation. Annotation parameters used included Genetic code = 11, E Value cut-off for selection of opinneMetricdCDSs = 1e-20. Complete or almost complete bacterial and archaeal genomes can be annotated using the completely automated service known as RAST (Rapid Annotation using Subsystem Technology). The entire phylogenetic tree's worth of these genomes' high-quality genomic annotations are provided.Subsystem features revealed the distribution of genes that play essential roles in a variety of metabolic pathways as well as salt stress resistance. The functions of the genes were cataloged into different functional classes (Fig. 6.5). According to the findings, the highest numbers of genes (310) were involved in the metabolism of carbohydrates, while the lowest

numbers of genes (07) were involved in secondary metabolism. There were no genes discovered that play a part in the process of photosynthesis. On the other hand, the genome of S-9 displayed genes that are responsible for the stress response (66) as well as dormancy and sporulation (17). In addition, the S-9 genome accounted for genes associated with the promotion of plant growth, such as those involved in the metabolism of nitrogen (N), phosphorus (P), and potassium (K), as well as iron acquisition and metabolism (Fig.6.5).

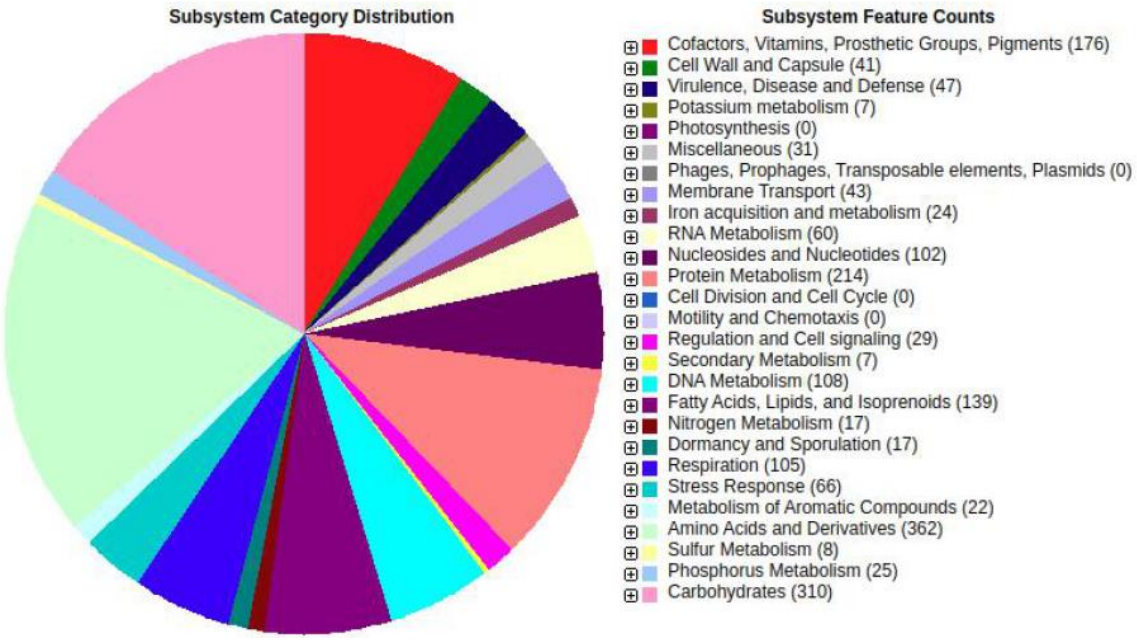


Fig. 6.5: Distribution of the genes of S-9 under various subsystem features of RAST

6.3.7 Antibiotic Resistance Gene Prediction Analysis

The Comprehensive Antibiotic Resistance Database (CARD; http://arpcard.mcmaster.ca/) was used to identify the antibiotic resistance gene in the assembled genome. S-9 species contains the macrolide antibiotic resistance gene with gimA family macrolide glucosyltransferase AMR gene family Table 6.3.

Table 6.3: Antibiotic Resistance Gene Prediction Analysis in S-9.

Sample	Drug Class	Resistance Mechanism	AMR Gene Family
S-9	Macrolide antibiotic	Antibiotic inactivation	gimA Family macrolide glycosyltransferases

6.3.8 Cluster of orthologous groups (COG) annotation

For the purpose of clusters of orthologous groups (COG) annotation, functional categories of coding sequences (CDSs) were identified using WebMGA by utilising the RPSBLAST software (with an applied threshold of 1e5). According to the findings of the COG analysis, the greatest proportion of genes does not have an assigned function (1978 genes). Fig. 6.6 shows that the number of genes involved in transcription (K), carbohydrate transport and metabolism (G), and amino acid transport and metabolism (E) is greater than that of the genes involved in any other function-related processes.

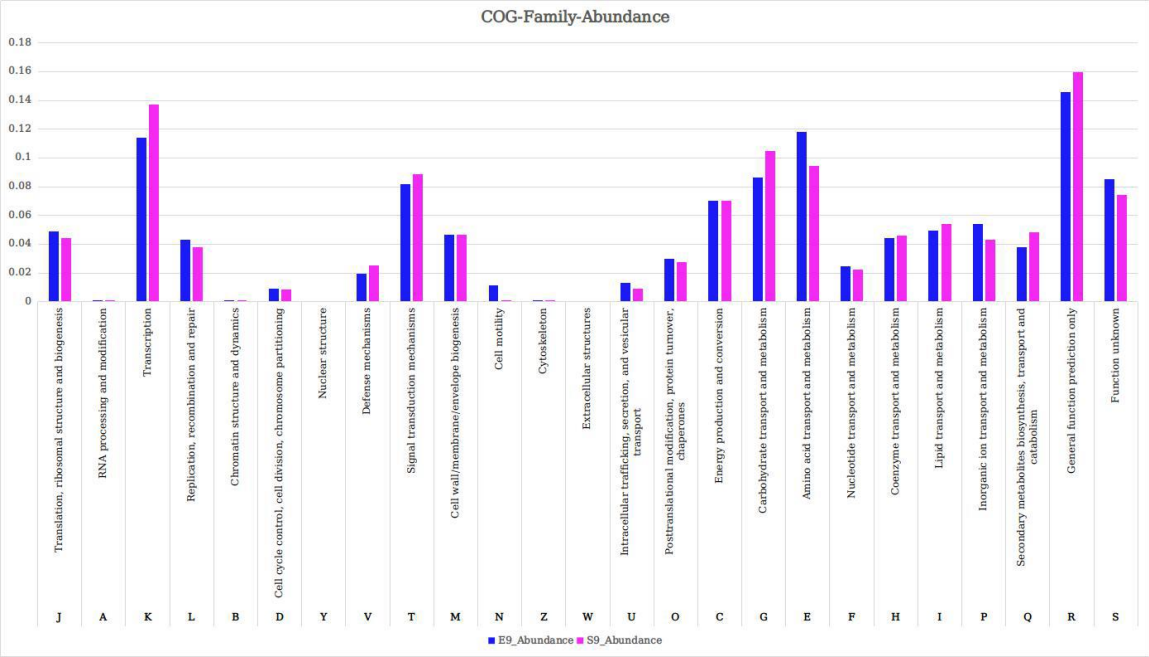


Fig 6.6: Cluster of Orthologous Groups (COG) database annotation of S-9 and relative abundance of proteins (%) in the two genomes

6.3.9 Secondary Metabolite Profiling and biosynthetic gene clusters

AntiSMASH 4.1.0 and BAGEL4 were used to check S-9 genomes for BGCs (Van Heel et al., 2018). AntiSMASH 4.1.0 identifies bacterial gene clusters using HMM and rules-based identification (BGCs). These BGCs encode polyketides, non-ribosomal peptides, terpenes, aminoglycosides, and RiPPs from bacterial genomes. BAGEL4 uses HMM to find RiPP-encoding genes. It's not dependent on the genome's ORF calls, thus it can more correctly detect the small precursor peptides in clusters of RiPP-expressing genes.

AntiSMASH predicted S-9 contained 5 metabolite clusters (Table 6.4). Most of these results were terpene BGCs, which resembled natural compounds (less than 92 %). These routes may encode new natural compounds or those without BGCs. S-9 strains share Lanthipeptide's non-ribosomal peptide synthesis pathway. Terpene, lanthipeptide-class-iii, T3 PKS, SapB, and Ectoine BGCs in S-9 isolates had a high degree of similarity (>90). *Streptomyces* strains have Terpene, lanthipeptide-class-iii, T3 PKS, and Ectoine gene clusters (Reinert et al. 2004; Juttner & Watson, 2007). Ectoine is a flexible nutrient that protects against osmotic stress (Schulz et al. 2017), and Lanthipeptid has antifungal, antibacterial, and antiviral bioactivities (Lagedroste et al., 2020).

Table 6.4: Biosynthetic gene cluster for secondary metabolites

Secondary Metabolites	Class	Cluster	%age (%)
Terpene	Terpene	Hopene	92
lanthipeptide-class-iii	RiPP: Lanthipeptide	SapB	100
T3 PKS	Polyketide	alkylresorcinol	100
Terpene	Terpene	albaflavenone	100
Ectoine	Others	Ectoine	100

6.3.10 RNA Extraction and Quality Check

The obtained result is presented in Table 6.6. Finally, RNA was checked on the Tapestation using HS RNA screentape (Agilent) to obtain RIN values. Observed results are given in Table 6.5.The final obtained cDNA is utilized for the qRT-PCR (Table 6.5)

Table 6.5: Quantification of isolated RNA samples on NanoDrop and Agilent Tape Station

SR. no	NanoDrop Readings	(ng/µl)	RIN Value NanoDrop	OD A260/280
1	Control	436.9	7.2	2.15
2	Plant with *Fusarium udum*	829.5	6.5	2.27

6.3.11 RNA Quality check

Total RNA of Control and Fusarium infected plant shown in Fig.6.7

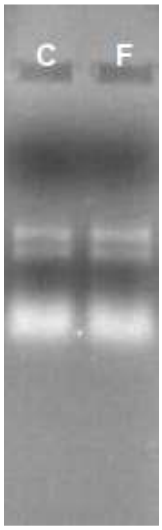


Figure 6.7: QC of isolated RNA samples on 1% denatured agarose gel

6.3.12 Result of miRNA

Finally, to obtain RIN values RNA was checked on the Tapestation using HS RNA screentape (Agilent). Observed results are given below in Table 6.6. Also, the final obtained cDNA utilized for qRT-PCR are presented in Table 6.7 and Fig. 6.8.

Table 6.6: RIN value of RNA

SR. no	Sample Name	RINe	28S/18S (Area)	Conc. [pg/µl]
1	*Fusarium udum*	6.7	0.8	8240
2	Control Pigeon pea	7.6	1.8	3840

Table 6.7: cDNA value for qRT-PCR

Sr.No	Sample name	ng/ul
1	*Fusarium udum* infected plant	184.2
2	Control Pigeon pea	182.3

F L P 2PF L

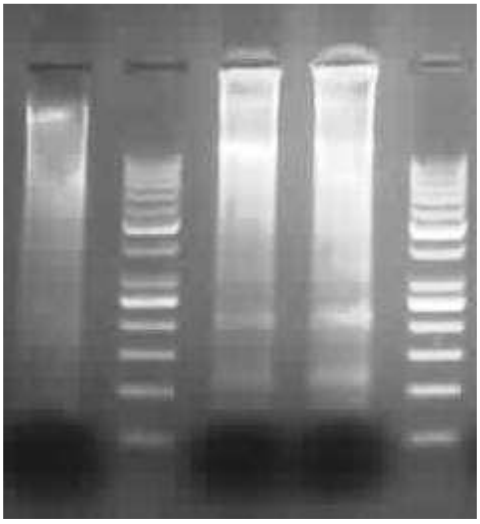


Figure 6.8: Total RNA quality Check

F = Fusarium slant (fresh submission_13032021); L = Ladder; 2P = Control Plant; 2PF= Treated Plant (Plant+fusarium)

6.3.13 qRT-PCR

The final obtained Ct values are provided in the below Table 6.8. Analysis of the result cannot be done as the internal control genes are not amplified as shown in Fig. 6.9.

Table 6.8: Ct value of qRT-PCR

Sr.no	**Sample Name (Primer pair)**	**well**	**1**	**2**	**3**
1	C (7-8)	A	45	45	45
2	T (7-8)	B	23.2	23.4	22.9
3	C (9-10)	C	38	37	38.3
4	T (9-10)	D	38.6	38.4	37.4
		well	4	5	6
5	C (11-12)	A	45	45	45
6	T (11-12)	B	45	45	45

A1 A2 A3 L B1 B2 B3

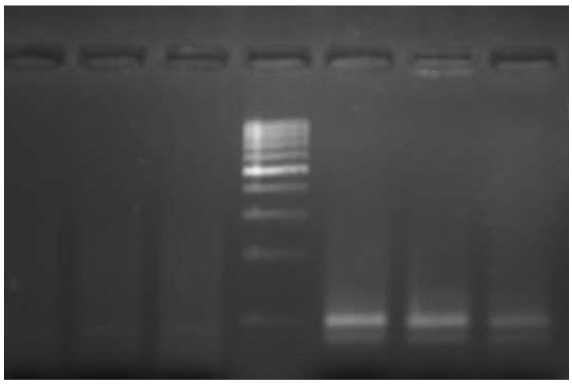


Figure 6.9: qRT-PCR gel

Starting from Left to right, well-1 =A1, well-2 = A2, well-3 = A3; Well-4 = Ladder,

Well-5 = B1, Well-6 = B2& Well-7=B3.

A=Control Pigeon pea; B= Pigeon pea infected *Fusarium udum*

6.3.14 Transcriptome analysis

The nucleotide sequence of the predicted novel genes was extracted from NCBI from the whole shotgun sequence of *Cajanus cajan*. Further, nucleotide BLAST was performed, which shows that one predicted novel gene translates to some known protein while the other three relate to some hypothetical proteins, which could be further analyzed by protein modelling.

6.3.15 Correlation Analysis

The correlation analysis shows no significant correlation between control and Fusarium infected plants. This may be attributed to the fact that genes from fusarium infected plants might deviate from their normal expression mechanism. For a significant result, the value between the two samples must be closer to 1. The closer the value to 1, more significant the expression correlation. It can be concluded here that fusarium attack leads to drastic change in expression pattern of some genes.

6.3.16 Differential Expression Analysis

The differential expression analysis of genes revealed a significant change in the expression pattern of control and Fusarium infected plant samples. The fold change between the samples indicates a higher or lower expression with respect to control. The number of upregulated and downregulated genes based on the cut-off (padjust<0.005 &log2foldchage>1) were 583 and 754, respectively. The maximum number of downregulated genes in the infected plant sample indicates its susceptibility to Fusarium infection and which can lead to head blight.

6.3.17 Novel gene

The “novel.1189” differentially expressed gene shows 10-fold change (approx.) increase in expression with respect to control. The sequence is similar to known ABD1 small subunit

ribosomal RNA gene of fusarium (MT640287.1).ABD1 gene translates to a protein named mRNA cap guanine-N7 methyltransferase. It catalyzes the transfer of a methyl group from S-adenosylmethionine to the GpppN terminus of capped mRNA. Being a nuclear protein it relocalizes to the cytosol in response to hypoxia. The following network shows the interaction between ABD1 gene and significant association with other partner genes. The higher the incomisong edges, the stronger and more vital position of that particular gene shown in Fig. 6.10.

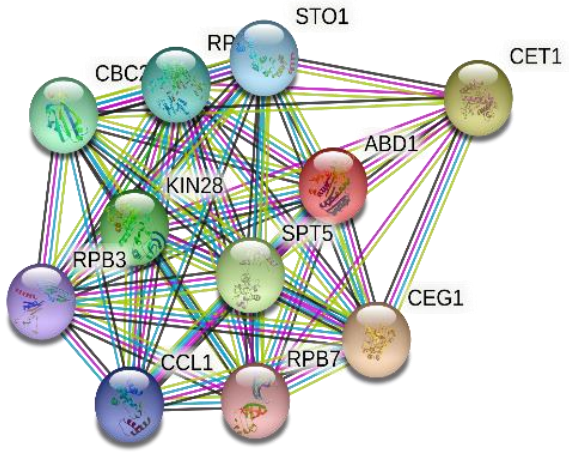


Figure 6.10: Network shows the interaction between ABD1 gene and other genes

6.3.18 Network Statistics

Number of nodes:11

Number of edges:53

Average node degree:9.64

Average local clustering coefficient:0.968

Expected number of edges:15

PPI enrichment p-value: 4.77e-14

6.3.19 Enrichment Analysis

GO Scatter plot suggests the maximum number of genes corresponding to enzyme inhibitor activity and ubiquitin transferase activity. The KEGG pathway shows that a large set of genes correspond to Aminoacyl-tRNA biosynthesis.

6.3.20 Coexpression Venn Diagram

The coexpression Venn diagram presents the number of genes uniquely expressed within each group/sample, with the overlapping regions showing the number of co-expressed genes in two or more groups/samples (Figure 6.11).

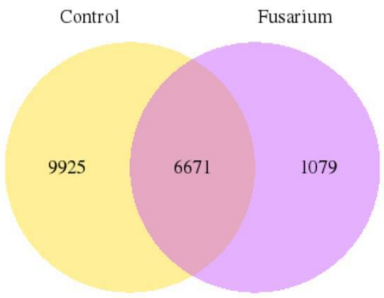


Figure 6.11: CoexpressionVenn diagram

6.3.21 Differential Expression Analysis

Readcount obtained from Gene Expression Analysis is used for differential expression analysis. For samples with biological replicates, differential expression analysis of two conditions/groups was performed using the DESeq2 R package (Anders et al., 2010). It provides statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. Therefore, if the readcount of the i-th gene in j-th sample is Kij, there is: Kij∼NB(μij,σij2) And the resulting P values were adjusted using the Benjamini and

Hochberg's approach for controlling the false discovery rate. All clean reads were deposited in the NCBI Short Road Archive (SRA) database and can be accessed with accession numbers-SRR15059311 and SRR15059312. The raw reads for all the Illumina sequenced transcriptome used for the analysis have been deposited to NCBI with the BioProject ID PRJNA743724. All raw data from the DGE library sequencing has been deposited in SRA (NCBI BioSample Accessions Number: SAMN20060467 and SAMN20060468)

Kij～NB(μij,σij2)

For the samples without biological replicates, the readcount was adjusted by TMM, then differential expression analysis was performed by using the EdgeR R package. The result of differential expression analysis is shown in Table 6.9

Table 6.9: Result of differential expression analysis

gene_length	gene_biotype	gene_description	tf_family	Fusarium_readcount	Control_readcount	Fusarium_fpkm	Control_fpkm	gene_length	gene_biotype	gene_description	tf_family	Fusarium_readcount	Control_readcount	Fusarium_fpkm	Control_fpkm	gene_length	gene_biotype	gene_description	tf_family
1098117796	2432.715	0.694607	11.49985	3.89E-46	6.58E-42	LOC109811796	NW_017984086.1	551968	552937	+	774	protein_coding	exansin-likeB1	-	921	2	16639.01	2.975882	109811796
novel.11189	1611.291	0.694607	10.905488	1.10E-41	9.30E-38	-	NW_018018115.1	1	213	-	213	-	-	-	610	2	40046	10.81377	novel.11189

6.3.22 DEGs in Response to *Fusarium* infection in *Cajanus cajan*

The genes with a false discovery rate (FDR) <0.005 and an estimated absolute log2foldchange (log2FC) >1 in sequence counts between libraries were considered significantly differentially expressed (Figure 6.12).Volcano plots are used to infer the overall distribution of differentially

expressed genes. The horizontal axis for the fold change of genes in different samples. The vertical axis for a statistically significant degree of changes in gene expression levels, the smaller the corrected p-value, the bigger -log10(corrected p-value), and the more significant the difference. In the pair of infected and control plants, 754 genes were down-regulated and 583 genes were up-regulated after the *Fusarium* infection shown in Fig. 6.12.

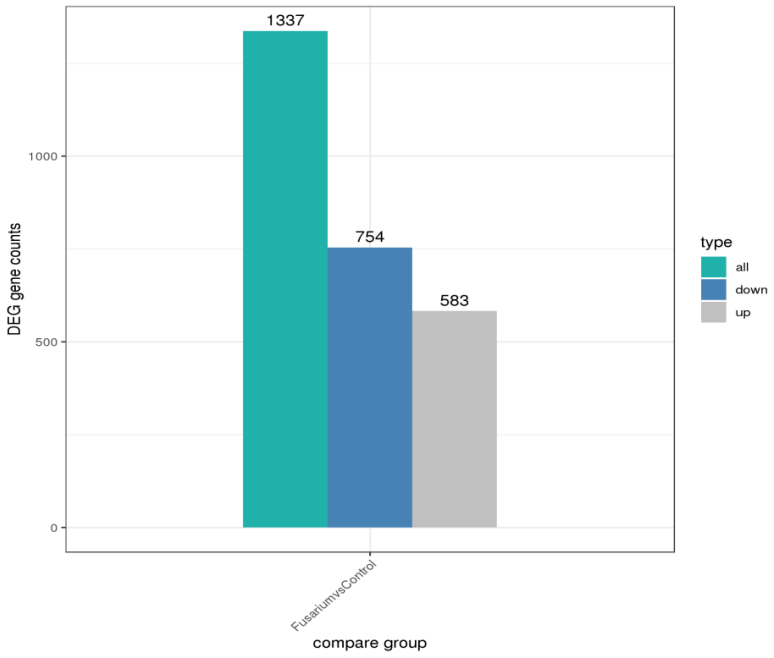


Fig. 6.12: DEGs Count in Control vs Fusarium infected plant

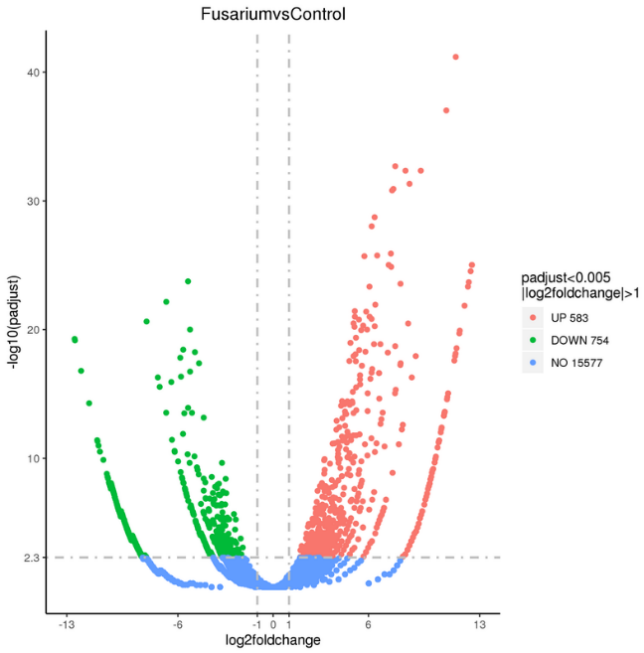


Figure 6.13: Volcano plot for DEGs between *Fusarium*-infected and healthy *C. cajan* Plants

Up-regulated genes are shoun in red, down-regulated genes are shown in gree, the genes without significant expression change are shown in blue in Figure 6.13.

6.3.23 Cluster Analysis

Cluster analysis on differential expression indicates genes with similar expression patterns under various experimental conditions. By clustering genes with similar expression patterns, it is possible to predict unknown functions of previously characterized genes or unknown genes. Hierarchical clustering analysis is carried out of log2 (FPKM+1) of union differential expression genes, within all comparison groups. Genes within the same cluster show the same trends in expression levels under different conditions.Expressipon pattern clusterin heat map analysis shon in Fig. 6.14.

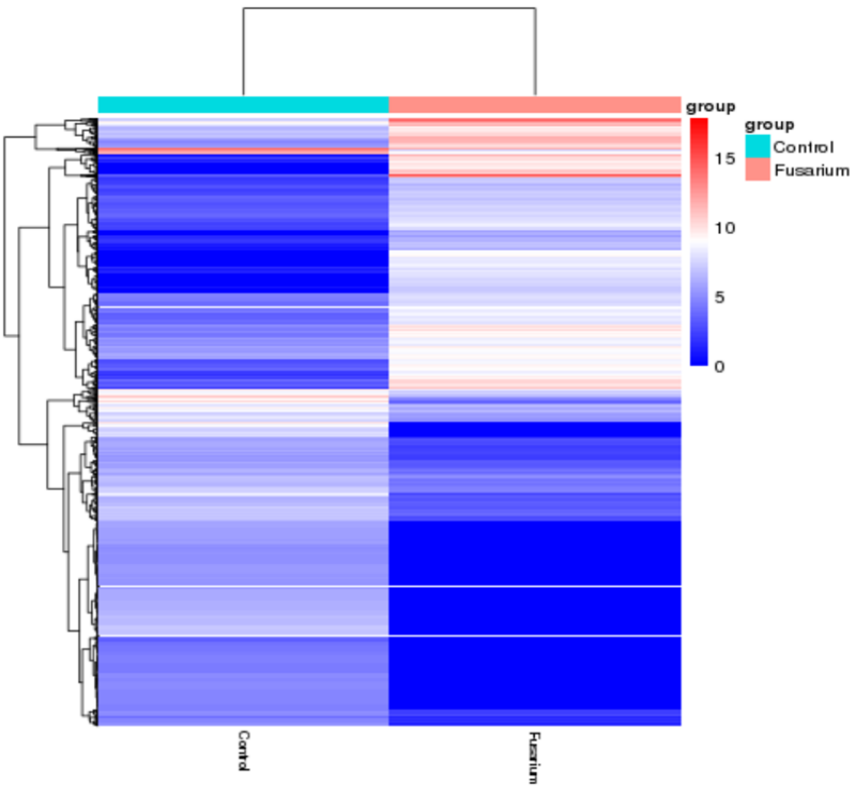


Figure 6.14: Hierarchical Clustering Heatmap between Control and Infected plant

The overall results of FPKM cluster analysis, clustered using the log2 (FPKM+1) value. Red color indicates genes with high expression levels, and blue color indicates genes with low expression levels. The color ranging from red to blue indicates that log2 (FPKM+1) values where from large to small.

6.3.24 GO Enrichment Analysis

Top 20 significantly enriched terms in the GO enrichment analysis are displayed below. If the enriched pathway is less than 20, all the terms will be displayed in Fig.6.15.In the scatter plot shown in Fig. 6.16, the horizontal axis is customized as GeneRatio and the Vertical axis is customized as the Term's Description. The size of every dot represents the number of the differential expression genes and the color of every dot represents the range of Qvalue.

"Cell wall organization (GO: 0071555)," "polysaccharide metabolic process (GO: 0005976)," and "glucosyl transferase activity cell periphery (GO: 0046527)" were dominant within each sub-

ontology. The terms "Endopeptidase complex (GO: 1905369)" and "Enzyme inhibitor activity O: 0004857)" were most prevalent in BP, MF, and CC, respectively. Similar to up regulated genes, down regulated genes have dominating keywords related to the proteasome core complex (GO: 0019773).

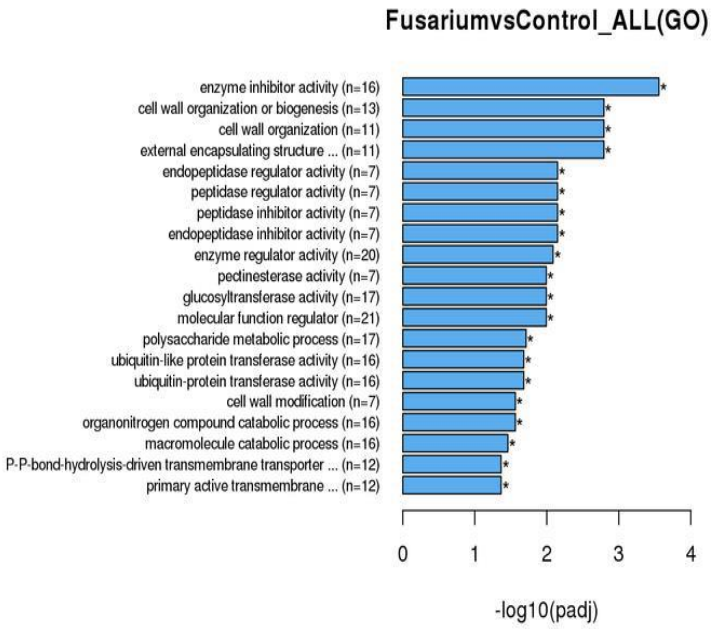


Figure 6.15: GO Enrichment Histogram

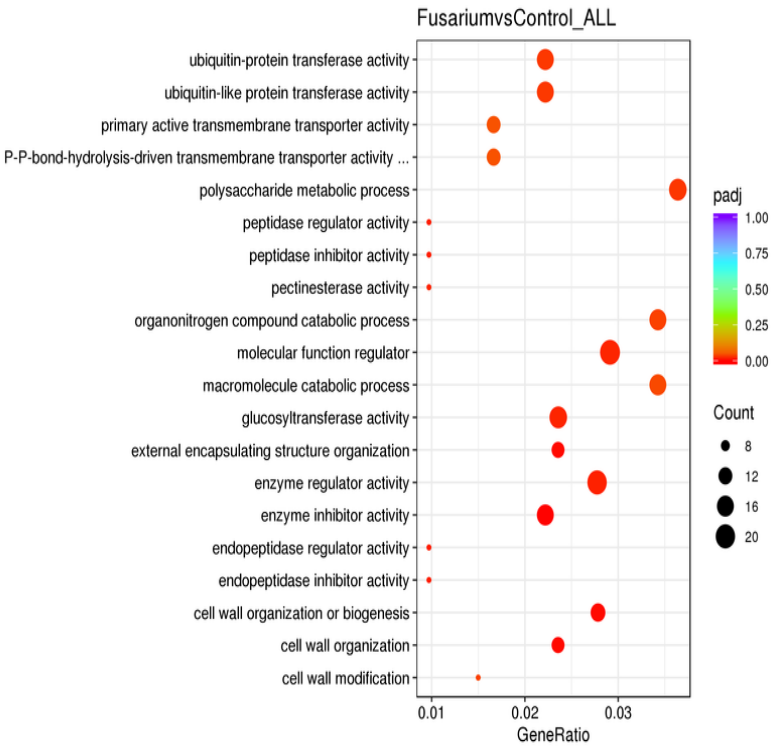


Figure 6.16: GO Enrichment Scatter Plot

6.3.25 KEGG Enrichment Analysis

Top 20 significantly enriched terms in the KEGG enrichment analysis were displayed in

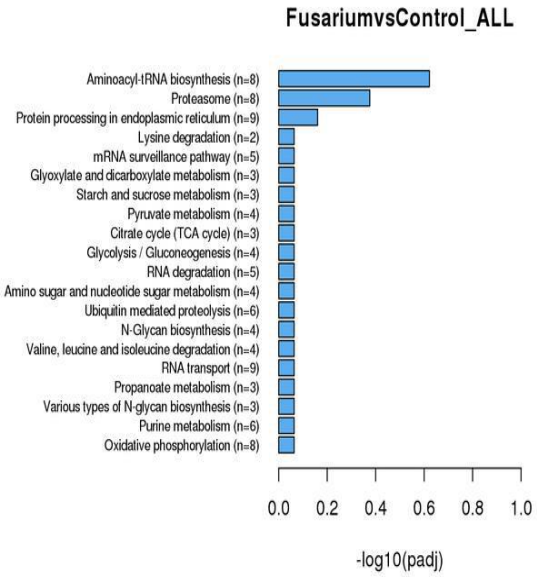


Figure 6.17: KEGG Enrichment Histogram

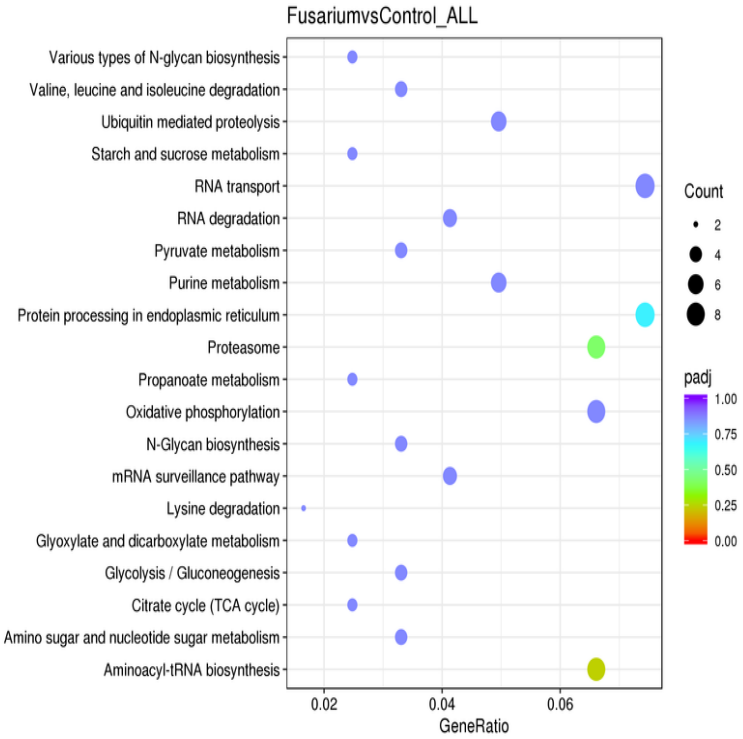


Figure 6.18: KEGG Enrichment Scatter Plots

The differentially expressed genes have been labeled in the pathway map. Some of the pathway maps across different sample groups are displayed below. In the Figure 6.18, the green box indicates the species-specific gene or Enzyme.

The maximum number of DEGs within the metabolism category were associated to "Protein processing in endoplasmic reticulum," followed by "Aminoacyl t- RNA biosynthesis Metabolic pathways," and then "Amino sugar and nucleotide sugar metabolism," according to KEGG pathway enrichment analysis of the DEGs shown in Fig. 6.19.

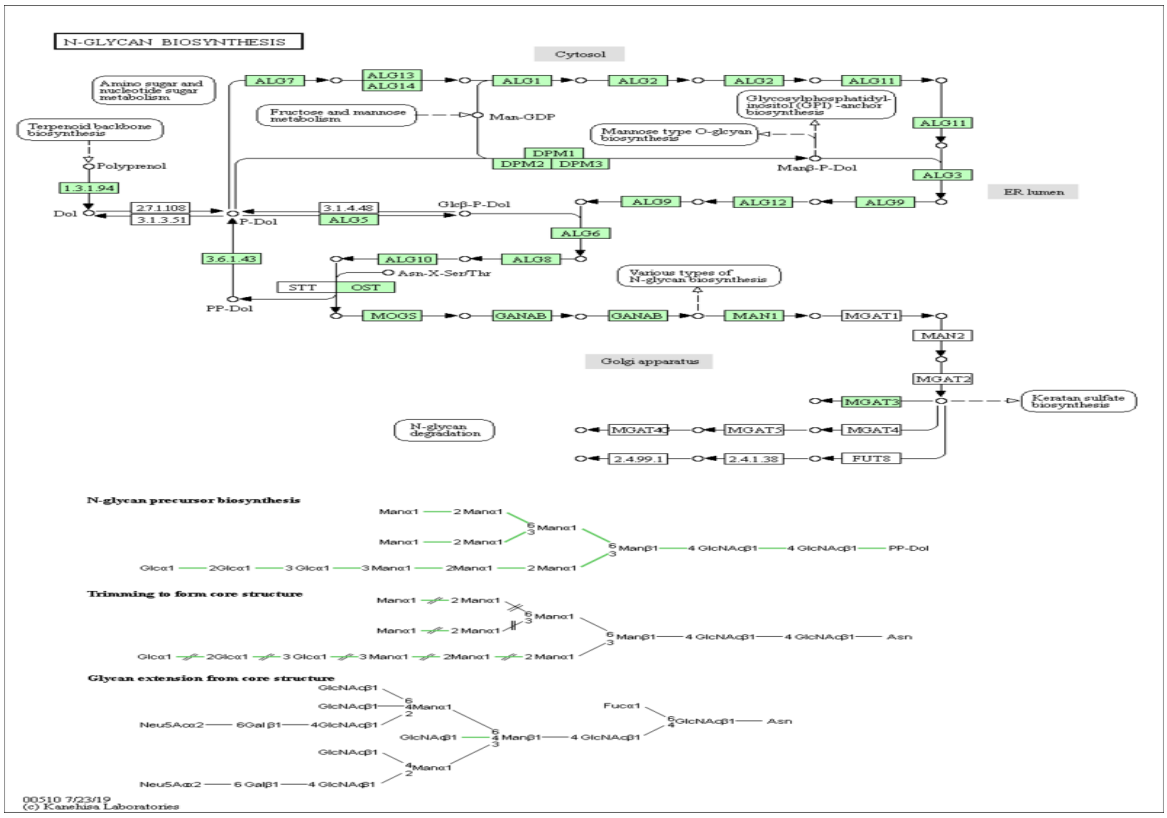


Fig. 6.19: Differential expression gene using KEGG

6.4 Discussion

The genus *Streptomyces*, which is the most common type of actinobacteria, is responsible for fifty % of the entire population of soil actinobacteria (Parte et al. 2020). Because of their capacity to produce biologically active chemicals as well as secondary metabolites, *Streptomyces* strains are of critical value to the commercial sector. A relatively small number of actinobacterial genera that are important to agriculture have been investigated at the entire genome level in order to look for novel genes related with the generation of secondary metabolites.Here we sequence *Streptomyces* species S-9, isolated form Pigeon pea (*Cajanus cajan*) plant. The examination of the genome provides in-depth knowledge of genes that play significant roles in a variety of key physiological, biochemical, and molecular processes that facilitates secondary metabolite production in plants.In a previous investigation, it was demonstrated that these bacteria had the capacity to create antifungal chemicals as well as a large number of hydrolytic enzymes (Zitouni et al. 2007). Because of their activity profiles, which include antibacterial, antifungal, and

enzymatic activities (Dave & Ingle, 2021), the strains S-9 were chosen for further study. S-9 was found to have a genetic similarity of 99 % to *Streptomyces* which had been isolated from the Pigeon pea plant (a tropical climate with temperatures that average 33-37°C). This allowed us to designate S-9 as a species of *Streptomyces*.The phylogenetic analysis examines the evolutionary development of a species, group of organisms, or organism trait. A considerably more accurate perspective of the species and strain phylogeny in *Streptomyces* has been demonstrated as a result of the phylogenetic analysis that was carried out as part of the current study. This view takes into account many components of the genome.

Most genome databases employ Illumina short-read. *Streptomyces* genomes are large, repetitive, and heavy in G+C, making them difficult to fully assemble from short reads. As a result, 90% of the known genomes are only in draft form; hundreds of contigs with an average N50 of thousands of nucleotides. We assembled high-quality genome sequences with PacBio and Illumina data, where the >8 Mb chromosome assembled as a single contig in one strain.An examination of the gene annotations shared by all of the *Streptomyces* strains revealed a large number of apparent lineage-specific gene families. It is possible that these gene families originated in the *Streptomyces* clade's last common ancestor. The species were found to match its most closely strains of *Streptomyces pseudogriseolus* sharing 99 % of their genetic material.

We evaluated the ANI values of the S-9 species with their related species in order to precisely recognise and comprehend them. Following the analysis of the genomic data, the genes responsible for primary and secondary metabolism were assigned annotations.To perform a better resolution in elucidating taxonomic structure beyond the species rank Average Amino-acid Identity (AAI) was performed for the S-9 species and found the similarity of S-9 species was with the *Streptomyces* species with highest similarity index of other species. Moreover, the genetic foundation for the production of antibacterial and plant growth-promoting compounds by these species as PGPR was also predicted. Comparisons of chromosomal maps conclusively identify the species as *Streptomyces*.

The subsystem analysis of S-9 species genes is majorly associated with nitrogen (N), phosphorus (P), and potassium (K), as well as iron acquisition and metabolism which clearly showed their association with secondary metabolite or plant growth-promoting activities. The production of Lanthipeptidesmeansits capacity to improve the absorption of iron by plants and also repress the phytopathogens.

The antibiotic gene annotation of S-9 species revealed that it is associated with macrolide antibiotic resistance following the previous results (Fyfe al. 2016; Dinos, 2017).

Streptomyces possesses a wide variety of stress-specific sigma factors, and the majority of the possible combinations of these sigma factors are what are responsible for the stress responses, adaptation to energy limitation, and development of the organism (Bentley, et al. 2002). The potential of *Streptomyces* to withstand high levels of stress is demonstrated by the presence of 31 stress-specific sigma factors, a complete ectoines biosynthesis pathway, and two related heterocyclic amino acids (Sadeghi et al. 2014). These stress-specific sigma factors are responsible for osmotic, heat, cold, draught, and pH stresses (Van-Thuoc et al. 2013).

Another significant discovery was made when researchers sequenced the entire genome of the *Streptomyces* S-9 species and then mined the genome for information on its biological and biotechnological possibilities. It was demonstrated that gene clusters for lanthipeptide, ribosomallyand non-ribosomally produced peptides, ectoine, and xenobiotic degradation pathways as well as heavy metal resistance were present in the organism.

In general, having knowledge on a genome scale about the potential for secondary metabolism in *Streptomyces* sp. will make the further characterization of bioactive molecules with a wide variety of actions much simpler. In addition, the genome that has been sequenced helps us gain a better understanding of this organism in terms of the production of antibiotics and other characteristics that are important to biotechnology. These characteristics include the bioremediation potential of this strain, as well as the production of new secondary metabolites.

Wilt is a biotic stress that severely affects various plant growth stages to different extents by targeting several physiological and biological processes related to the respective growth stage.Genes involved in “Oxidative phosphorylation” showed suppressed expression. Earlier, several DEGs enriched in “Oxidative phosphorylation” were obtained based on transcriptome sequencing. The “novel.1189” differentially expressed gene shows 10-fold change (approx.) increase in expression with respect to control. The sequence is similar to known ABD1 small subunit ribosomal RNA gene of Fusarium (MT640287.1).ABD1 gene translates to a protein named mRNA cap guanine-N7 methyltransferase.