


Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 2515042Y	Student Name: Zhen Yuan
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): Civil Engineering MSc	
Course Code: ENG5059P	Course Name: MSc Project
Name of First Supervisor: Dr. Umer Zeeshan Ijaz	Name of Second Supervisor: Dr. Ciara Keating
Title of Project: Whole genome functional analysis of Pseudomonas putida	
Declaration of Originality and Submission Information	
<p><i>I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements</i></p> <p>Signed (Student) : ZHEN YUAN</p>	 E N G 5 0 5 9 P
Date of Submission : 20 August	

<i>Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator</i>	
Grade Awarded: Feedback (as appropriate to the coursework which was assessed): 	
Lecturer/Demonstrator:	Date returned to the Teaching Office:

Whole genome functional analysis of *Pseudomonas putida*

Civil Engineering MSc
Zhen Yuan (Student ID: 2515042y)
Supervisor: Dr Umer Zeeshan Ijaz
Dr Ciara Keating

August 20,2021

**A report submitted in partial fulfillment of the requirements
for Civil Engineering Degree
at the University of Glasgow**

Acknowledgement

Firstly, I would like to thank my two tutors, Dr. Umer Zeeshan Ijaz and Dr. Ciara Keating, for their time and patience in guiding our studies and monitoring our progress in spite of their busy schedules. Secondly, I would like to thank my university, the University of Glasgow, for giving us the opportunity to improve ourselves during such a difficult year and for making sure that we have a good day-to-day study and life. I would also like to thank my fellow students for their help and advice during the project, and for your patience and guidance. Last but not least, I would like to thank my family for their support and encouragement, which has helped me to study in a foreign country and you will always be my strongest support.

Abstract

Since the development of oil in the 19th century, petroleum products have become the main fuel consumed by human society. Petroleum consists of a mixture of different hydrocarbons, the main component of which is alkanes, and it also contains elements such as sulphur, oxygen, nitrogen, phosphorus and vanadium. Petroleum is mainly used as fuel oil and gasoline, which is currently one of the most important and widely used energy sources in the world. It is also a raw material for many chemical industry products such as solutions, fertilisers, pesticides, mineral oil base oils for lubricants and plastics. Oil is present in almost every part of our lives, from the transport and production of food, clothing, materials and pharmaceuticals to the plastics used in the manufacture of a large number of products.

Some of the harmful compounds produced during the extraction process of petroleum are extractable hydrocarbons, known as polycyclic aromatic also known as polycyclic aromatic hydrocarbons or polyaromatic hydrocarbons (PAHs) compounds. PAHs are pervasive environmental pollutants that are potentially toxic, mutagenic and carcinogenic, and have a significant impact and harmful effect on the environment. These hazardous substances are considered to be petroleum pollutants and have become widespread in the environment. The contamination of soils and waters has become a major problem worldwide as the pollutants produced by oil development can be very harmful to human health and the environment in which they live. When soil is contaminated, it affects the growth of plants and transmits harmful substances. When it reaches the sea, it poses a threat to humans and fish.

The treatment of oil-contaminated soils and waters by micro-organisms has been systematically studied on a large scale. They play a major role in the degradation of polluting compounds. And among them *Pseudomonas putida*

is the most promising microbial strain for degrading contaminants from petroleum.

This project will analyse the whole genome sequencing of *Pseudomonas putida* through a series of genetic analyses, focusing on the degradation of petroleum by *Pseudomonas putida* and their gene sequences. By integrating, comparing and analysing these microorganisms, the impact of pollutants on the microbial community will be assessed and the feasibility of biodegradation will be judged.

Key Words: crude oil, degradation, *pseudomonas putida*, Whole genome sequencing

Content

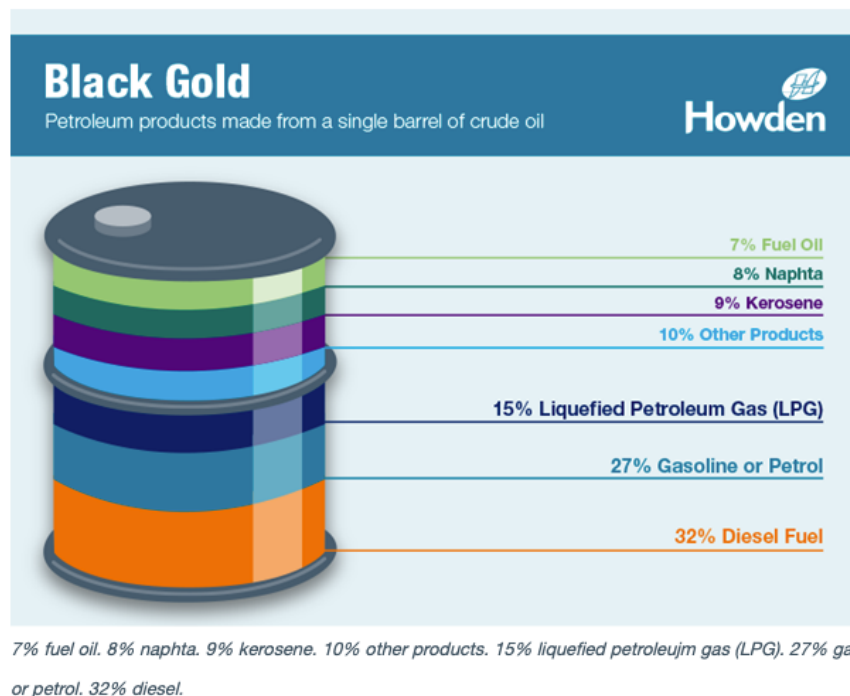
1.Introduction	1
1.1 Backgroud of Crude Oil.....	1
1.2 Hazard of Oil Applications.....	3
1.3 Role of Microbial Crude Oil Aspects	5
1.4 Pesudomonas Putida	8
1.5 Aims and Objectives.....	8
2.Method.....	10
2.1 Initial Preparation Works	10
2.1.1 Literature Search	10
2.1.2 The Acquisition and Storage of Data	11
2.1.3 Summary.....	11
2.2 Prokka	12
2.2.1 Function of Prokka	12
2.2.2 Prokka Workflow.....	12
2.3 Roary	13
2.3.1 Function of Roary.....	13
2.3.2 Roary Workflow	13
2.4 Coinfinder	14
2.4.1 Function of Coinfinder	14
2.4.2 Coinfinder Workflow.....	14
2.5 Metabolic	15
2.5.1 Function of Metabolic	15
2.5.2 Metabolic Workflow.....	15
2.6 R Studio.....	16
3.Result	17
3.1 Prokka Output	17
3.2 Pan-genome	18
3.3 Network and Heatmap of Pangenomes.....	22
3.4 The Degradation of Psudomonas Putida.....	24
3.5 Psudomonas Putida Crude Oil Degradation	25

3.5.1 Overview of Existing Research	26
3.5.2 Statistic of Pangenome activity	26
<i>4. Discussion</i>	<i>29</i>
<i>5. Conclusion</i>	<i>30</i>
<i>Reference</i>	<i>31</i>
<i>Appendix</i>	<i>36</i>

1. Introduction

1.1 Background of Crude Oil

Crude oil is a naturally occurring fossil fuel that is made up of hydrocarbons, leased layers of mainly a mixture of hydrogen and carbon atoms. It is found in liquid form in subsurface reservoirs in tiny spaces within sedimentary rocks and can also be found as a component on the surface of oil sands and in nearby areas. When crude oil is found it is usually found together with natural gas and brine. Crude oil can also be referred to as petroleum, as it includes both unrefined crude oil and refined petroleum products. It is a non-renewable energy source with limited reserves. Gragh 1 shows the petroleum products that can be processed from a single barrel of crude oil.



Gragh 1 Petroleum products made from a single barrel of crude oil

(<https://www.howden.com/en-gb/articles/pcog/where-does-crude-oil-come-from>)

Crude oil is a liquid formed from the remains of dead organisms such as algae and zooplankton that existed in the marine environment millions of years ago. During their lives, these organisms absorbed energy from the sun and stored it in their bodies in the form of carbon molecules. Once they die, their remains are buried underground or at the bottom of the sea and stored in underground layers of sand, mud and rock. Over millions of years of accumulation and deposition, these remains are buried deeper and deeper in sediment and organic matter. The accompanying great pressure, high temperatures and lack of oxygen can transform the organic material into a waxy substance known as kerogen. As time progresses, more heat is built up and greater pressures are brought into play, with the kerogen being catalytically transformed into hydrocarbons. Hydrocarbons also exist in many different types together, and the hydrocarbon component of these forms crude oil. As crude oil is a liquid product, it moves from areas of high pressure to areas of low pressure through tiny voids in the surrounding rock, i.e. it moves from deep subsurface structures and submarine structures towards the surface until it flows into impermeable layers of rock or clay where it forms underground reservoirs.

Crude oil is found underground in layers of impermeable rock and clay, and as people exploit and use it, drilling for oil becomes necessary to go deeper and deeper, meaning that oil reserves become less and less available until they are depleted. Since the exploitation and use of oil in the 19th century, energy resulting from the development of petroleum products has become the main fuel consumed by human society. It is mainly stored in the upper regions of the earth's crust. Petroleum consists of a mixture of different hydrocarbons, the main component of which is alkanes, and it also contains elements such as sulphur, oxygen, nitrogen, phosphorus, and vanadium. Oil is mainly used as fuel oil and petrol, which is one of the most important and widely used energy

sources in the world today. It is also a raw material for many chemical industry products such as solutions, fertilisers, pesticides, mineral oil base oils for lubricants and plastics. Oil is present in almost every part of our lives, from the transport and production of food, clothing, materials and medicines to the plastics used in the manufacture of a large number of products. Today 88% of the oil extracted is used as fuel and the other 12% as a raw material for the chemical industry. All in all, oil is one of the essential energy sources for the human world.

1.2 Hazards of Oil Applications

Crude oil is a complex hydrocarbon formed from the decomposition of Carboniferous plant remains under high temperature and pressure, and is a naturally occurring product of extractable petroleum hydrocarbons (EPHs), i.e. polycyclic aromatic also known as polycyclic aromatic hydrocarbons or polyaromatic hydrocarbons (PAHs) compounds. Amongst these, PAHs are a widespread environmental contaminant that is potentially toxic, mutagenic and carcinogenic, with significant environmental impacts and hazards. The contamination of soils and waters has become a major problem in the world that needs to be addressed immediately, as the pollutants from oil development can be very harmful to human health and the living environment. In terms of soil research, shuguang et al. (2017) point out that petroleum has a low density, high viscosity, and low emulsification capacity, which is easily absorbed by the soil surface and can affect the permeability and porosity of the soil (Wang, 2009; He et al., 1999). Petroleum is rich in carbon and small amounts of nitrogen compounds, so it can alter the composition and structure of soil organic matter to some extent, while affecting soil components such as C/N, C/P, salinity, pH, EH and electrical conductivity (Li et al. 2009). Thirdly, it can hinder plant growth, e.g. the concentration of oil and the germination rate and individual height of

plants can show a negative correlation, which has a significant impact on causing the plants themselves to be less resistant to pests and diseases. (Zhu, 2010; Shan et al., 2014). In addition, oil in the soil can lead to a significant reduction in the content of elements such as phosphorus, potassium and nitrate that are available to plants in the soil, as well as an increase in the concentration of other trace elements, which can lead to an increase in the elements absorbed by the plant that are harmful to it and cannot be processed, leading to its growth and development being affected and even widespread death and extinction (Navjot.K et al. 2017). Ingestion of petroleum contaminants by humans can have serious consequences. Due to the volatility of the aromatic compounds in petroleum, it is more likely to be hazardous to human health in the environment. It can enter the body of humans and animals through breathing, skin contact and diet. It reduces the normal function of the liver and kidneys, creates respiratory problems and immune dysfunction. And the resulting social panic and economic problems and environmental pollution over the next few decades are unsustainable. The solution to petroleum contaminants is therefore a matter of great urgency for human health and survival (Pena. P et al. 2020). Finally, petroleum contaminants in soil affect not only the soil circle but also water bodies. The sources of oil pollution in the marine environment include 2 types, natural and anthropogenic, and anthropogenic pollution is the main cause of marine pollution. Offshore oil extraction, oil spills from offshore oil transportation accidents pose a great danger to the sea. Oil spills can cause great harm to deep sea and coastal fisheries. When oil enters the ocean, it spreads rapidly into a shimmering film of oil. Typically, the film formed by one tone of oil can cover an area of 12 square kilometers of sea surface. The large oil film reduces the amount of energy that can be put into the sea by solar radiation and blocks the interaction between sea and air, directly affecting the photosynthesis of

marine plants and the cycle of the entire marine food chain, thus seriously disrupting the normal ecological balance in the marine environment. Oil pollution greatly reduces the ability of marine organisms to feed, reproduce and grow, disrupting the normal physiological behavior of cells and causing abnormal development of embryos and larvae of many marine organisms. Oil can easily block the respiratory organs of sea animals and fish, causing them to suffocate and die. A 0.1% diesel emulsion in seawater can completely block the photosynthesis of seaweed seedlings - and more seriously, vulnerable links in the marine ecosystem that, once damaged, are difficult to recover from for decades. Oil waste poisons marine and coastal organic substrates, interrupting the food chain on which fish and marine life depend and on which their reproductive success is based. Commercial fishing enterprises may be permanently affected. (Carls, M. G. 2001. and Pezeshki, S. R. 2000)

1.3 Role of Microbial Crude Oil Aspects

The treatment of oil-contaminated soils and waters has been systematically studied on a global scale. The main methods included are physical methods, chemical methods, and bioremediation methods. Bioremediation techniques have evolved considerably in recent years, with the aim of effectively restoring crude oil-contaminated environments and organisms in an eco-friendly manner and at very low cost (Azubuike, C. C et al. 2016). To achieve this goal, researchers have developed and modelled a number of different biotechnological tools, however, the varying properties and composition of contaminants lead to the fact that there is no single biotechnology that can effectively address all contamination, where the applicable environmental conditions of microorganisms in the contaminant's native region with their growth and metabolism are key to addressing the biodegradation

and bioremediation of contaminants (Verma and Jaiswal 2016). Bioremediation has become one of the most valuable and viable treatment technologies in the field of eco-environmental protection compared to the other two remediation methods and is considered as the measure of choice for the remediation of crude oil pollution. The most extensively studied is the bioremediation of petroleum hydrocarbons, which uses the metabolic diversity of microorganisms to degrade crude oil contaminants. Bacteria are often chosen as the primary choice for bioremediation (Prakash and Irfan 2011) because of their rapid metabolic rates, the fact that they fit into a variety of different degradation pathways, and the fact that their genetic sequences and properties can be manipulated to enhance their bioremediation capabilities. Of these, alkane degradation is the most common phenomenon, and it is possible to isolate and characterise prokaryotic and eukaryotic microorganisms capable of using these products as carbon sources, and degradation of long-chain n-alkanes has emerged as an important approach to oil pollution of the marine environment (Kumar. S et al. 2011). It has also been shown that microorganisms can be used to induce extracellular and intracellular enzymes, with extracellular enzymes playing a major role in the degradation of petroleum hydrocarbons (Barnabas et al. 2013). Extracellular enzymes break down the oil into simpler compounds that are readily absorbed and utilized by microorganisms. The biodegradability of aromatic hydrocarbons and the presence of molecular oxygen in the form of molecular oxygen can also be used to initiate enzymatic attack on the PAH ring to form cis-dihydrodiols for the purpose of degrading pollutants (Kumar et al. 2011). Also, co-culture techniques can be used to enhance biodegradation efficiency (Kadali et al. 2012). The complementary effects of microorganisms on each other may lead to a significant increase in growth and viability between different microorganisms

(Sampath et al. 2012), and this approach can directly or indirectly increase biodegradation rates. It is also possible to exploit the hydrophobicity of the cells (Babita.k et al. 2012), the biosurfactant produced by petroleum degrading bacteria facilitates the uptake and production of hydrocarbons from crude oil by the bacterial cells and converts pollutants into less toxic products with high biodegradability, therefore hydrocarbon degrading bacteria with the ability to produce biosurfactant are widely recommended for the rapid degradation of crude oil. Rapid degradation, which can lead to increased emulsification of the oil, can also modify the adhesion of hydrocarbons to other bacterial cell surfaces (Manoj, k. et al. 2006). Transgenic microorganisms can also be used in bioremediation processes to effectively remove pollutants that cannot be degraded by indigenous microorganisms. It plays an important role in remediating industrial wastes, reducing the toxicity of harmful compounds, removing pollution from hydrocarbons and gasoline emissions. (Kumar et al. 2018)

Environmental factors can also be used to optimise microbial biomass. The growth rate of microorganisms is influenced by various environmental factors and can be aided by controlling the pH, temperature, moisture content, aeration, and nutrient utilisation of a particular environment (Kumari, B et al. 2020). Biostimulants can be used to enhance biodegradability using different types of biofortification and biostimulation products (Kumari, B et al. 2020). Fillers provide optimum free air space and regulate the water content of the soil. They are usually divided into structural and organic fillers, which when added to the soil, increase soil aeration and microbial activity. And they also have a microbial biomass that can enhance soil fertility. That is, they treat crude oil by increasing the biodegradation rate in composting applications (Nakles and Ray 2002, Koolivand et al. 2013. Kumari et al. 2016).

1.4 Pseudomonas Putida

In the biodegradation of crude oil, microorganisms are considered to be the only biological source of hydrocarbon degradation (Atlas, 1981). Bacterial species have been shown to degrade hydrocarbons. Pure cultures of *Pseudomonas aeruginosa* are obtained by enrichment techniques from effluents that have been contaminated for some time (Raghavan and Vivekanandan, 1999) and due to their powerful capabilities in the degradation and biotransformation of biotic and xenobiotic pollutants, *Pseudomonas* has great potential for different biotechnological applications, especially in the field of bioremediation and biocatalysis. (Loh.k.c, et al. 2008). A proteomics-based theoretical approach allows the conversion of aromatic compounds into dihydroxylated intermediates and their catabolism by *Pseudomonas aeruginosa* during degradation via metabolite specific upstream pathways. (Loh.k.c, et al. 2008). *pseudomonas putida* has been shown to improve the degradation of crude oil and *pseudomonas putida* has been found to show the best performance in degrading alkanes (zheng.M.Y et al. 2018). Similarly, *Pseudomonas putida* can degrade hydrocarbons such as benzene (Munoz et al. 2007), toluene, p-xylene (Yu et al. 2001), biphenyl (Ohta et al. 2001) and phenol (Juang & Tsai 2006). *Pseudomonas putida* has great potential for the degradation of alkanes and aromatic hydrocarbons in crude oil.

1.5 Aims and Objectives

The main objective of this study was to obtain the capacity of the existing strain of *Pseudomonas putida* in terms of oil degradation through data analysis. Analysis of the potential of this strain for oil degradation. All genomes of *pseudomonas putida* were located based on the database of existing studies, the collected data were genome-wide annotated using Prokka and then a pangenome was constructed using

Roary to identify core and accessory genes. Finally, the genomic data were analysed using METABOLIC (METabolic And BiogeOchemistry anaLyses In miCrobes) and analytical reports and graphs were output to study the metabolism and function of the genome.

2. Method

All the data and models involved in this project were completed with the assistance of Dr Umer Zeeshan Ijaz (<http://userweb.eng.gla.ac.uk/umer.ijaz/>) and Dr Ciara Keating at the University of Glasgow, who provided relevant training to help me with the processing and analysis of the relevant data.

2.1 Initial Preparation Works

2.1.1 Literature Search

As this project investigates the analysis of gene sequences corresponding to specific bacteria, specific keywords were used in the selection of the relevant studies. These keywords include: "pseudomonas putida", used to select specific bacterial strains; "oil degradation/ degrading crude oil", used to qualify the topic of interest; and "pathway/review", used to search for gene functions and methods. "pathway/review" for finding the function and method of a gene. By adding the above keywords to relevant academic search engines (Google Scholar: <https://scholar.google.com/> and University of Glasgow library: <https://www.gla.ac.uk/myglasgow/library/>), 54 relevant studies were initially screened. Then, by reviewing the content of these publications, the methods and conclusions given in the articles were used to compare the corresponding sample data and information about their gene sequences with the subject of this project, and the relevant requirements of this project for these. As there are not many studies involved in the strains of this project that degrade crude oil, the final selection of literature was minimal. However, the direction of this project has great potential in the field of bioremediation

2.1.2 The Acquisition and Storage of Data

The strains involved in this project were provided by NCBI (National Center for Biotechnology Information), which has a range of databases related to biotechnology and biomedicine and is a great resource for bioinformatics tools and services, the main one used in this case being the GenBank database of DNA sequences. All downloaded data is stored in a remote cluster, also provided by Dr Umer Zeeshan Ijaz. The above procedures were operated via the relevant command line under Linux operating system conditions set up on the MobaXterm software (<https://mobaxterm.mobatek.net/>) platform, and the gene sequences downloaded from the database were stored in the files corresponding to their NCBI login numbers. The specific workflow is:

- 1) Create the exclusive *Pseudomonas putida* folder on the MobaXterm software platform.
- 2) connect to the bacterial data on the ncbi on the platform (ftp ftp.ncbi.nlm.nih.gov)
- 3) then log in to anonymous and find the bacterial data on genbank in genomes
- 4) Find the list of *Pseudomonas_putida* data and download its data to a local storage tool.
- 5) Download the bacterial genome to the created file using the relevant commands. (Refer to Appendix 1 for the exact procedure steps).

2.1.3 Summary

Preliminary data collation has been completed by reading the literature on the project and downloading the relevant genomes. The next step is to analyze the genomes studied: prokka -> roary -> coinfinder -> metabolic.

2.2 Prokka

2.2.1 Function of Prokka

Prokka is an excellent solution, a command line software tool that can fully annotate a bacterial genome component in a very short time and can be used for further analysis and to view the genome information and results on relevant software. information and analysis results (Seemann, T 2014).

2.2.2 Prokka Workflow

The workflow of Prokka is also done under the Linux operating system set up on the MobaXterm software platform and is as follows.

1) Set up the installation path and environment variables in the linux operating environment, enable the minconda2 environment which is available for all tools and set up pang genome.

2) Type in the prokka command to enter the annotation of the bacteria.

3) Run prokka and annotate all sequences.

It takes 5-10 minutes to run prokka for each genome, and as *Pseudomonas_putida* has 194 genomes, this stage of running prokka takes 3 working days to carry out. (See Appendix 2 for the exact procedure steps).

2.3 Roary

2.3.1 Function of Roary

Roary is a tool for the rapid construction of large-scale pangenomes, identifying core and helper genes. Roary can build pangenomes of thousands of prokaryotic samples on a standard desktop without compromising the accuracy of the results. As the software requires the use of annotated spliced sequences per sample, all strains analysed must be from the same species. The coding regions in the sequences are converted to protein sequences by the CD-HIT tool and then the BLASTP tool is used to search for protein sequences in all strains, which are then divided into groups using the Markov cluster algorithm (MCL) and combined with the previously converted sequences from CD-HIT to produce the final result (Page, A et al. 2015). By running Roary, the core and common genes of a strain can be found and the present and absent of the genome corresponding to each gene can be seen.

2.3.2 Roary Workflow

- 1) Create the roary folder, find all the .gff files in the genome file and copy them to the roary file.
- 2) Set up the installation path and environment variables in the linux operating environment, enable the minconda2 environment for all tools and set up pangenome.
- 3) Set the correct path to PRERL5LIB.
- 4) Enter the ROOAY command to enter the identification of the genome and the creation of the pangenome.
- 5) As *Pseudomonas_putida* has 194 genomes, it took 3 working days to run this stage of ROARY. (The exact steps can be found in Appendix 3)

2.4 Coinfinder

2.4.1 Function of Coinfinder

Coinfinder is an algorithm and software tool that detects genes that are associated and segregated from other genes more frequently than expected in the pan-genome. Written primarily in C++, Coinfinder is a command line tool that generates text, gexf and pdf output for the user. The given gene pairs are overlapping (Whelan, F. J. 2020). Where the gexf file outputs a networkmap and the pdf file outputs a heatmap.

2.4.2 Confinder Workflow

- 1) Set the installation path and environment variables in the linux operating environment, enable the minconda2 environment for all tools, and enable coinfinder on Orion cluster.

- 2) Create the coinfinder-test folder and use the "git clone" command to download the data associated with the coinfinder manuscript (<https://github.com/fwhelan/coinfinder-manuscript.git>)

- 3) This project is interested in the gene_presence_absence.csv generated by Roary and core_gps_fasttree.newick. Suggest that the phylogeny should be in Newick format with no zero-length branches. They suggest using core gene information to construct such phylogenies (e.g. as suggested in the Roary pipeline <https://sanger-pathogens.github.io/Roary/>). Therefore copy the core_gps_fasttree.newick and gene_presence_absence.csv files to provide preparation for running the coinfinder.

- 4) Run according to the coinfiner command and view the resulting files (of interest are the PDF files and GEXF files that can be viewed at <https://gephi.org/users/download/>. (Appendix 4).

2.5 Metabolic

2.5.1 Function of Metabolic

Metabolic (<https://github.com/AnantharamanLab/METABOLIC>) is able to predict the metabolic and biogeochemical functional profile of any given genomic dataset. These genomic datasets can be macrogenome assembled genomes (MAG), single cell amplified genomes (SAG) or pure culture sequenced genomes. METABOLIC has two main implementations, METABOLIC-G and METABOLIC-C. METABOLIC-G.pl allows the generation of metabolic profiles and biogeochemical cycling maps of input genomes and does not require input sequencing reads. METABOLIC-C.pl generates the same output as METABOLIC-G.pl, but as it allows input of macro-genomic reads it will generate information related to community metabolism. It can also calculate genome coverage. Parses the information and generates graphs of elemental/biogeochemical cycling pathways (Zhou, Z et al. 2020).

2.5.2 Metabolic Workflow

- 1) Set up the metabolic environment, enable metabolic software on Orion and directory.
- 2) Set up the path to the METABOLIC repository so that the project can start using the software
- 3) Dump the test genome into a local directory
- 4) Ensure that the genome has a .fasta extension
- 5) run the software
- 6) Check the input genome folder. You will notice that additional annotation files have been generated
- 7) Check the output folder. You are interested in

Nutrient_Cycling_Diagrams and METABOLIC_result.xlsx, which summarises everything (refer to Appendix 5 for detailed steps)

2.6 R Studio

In this project, R studio will be used to analyse and count genomic similarity and gene relatedness. The jaccard similarity in R is used to enforce the similarity of the genome. It ranges from 0 to 1. The higher the number, the more similar the data is. and use Coincidence Analysis (CNA) to model and provide a comprehensive analysis of causality. can is a method for causal data analysis. It allows you to see if there is an association between genes.

3. Result

3.1 Prokka Output

When a Prokka run of a genome is completed, 10 output files will be generated, such as table2.

Table 2. Description of Prokka output files

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

Table 2 (Seemann, T 2014)

The focus of this project is on the genbank database, and the gff file is almost identical to the gbk file. As the gff file will need to be received for subsequent analysis, the gff file was chosen for this project. As *Pseudomonas_putida* has 194 genomes, one of the genomes was chosen as an example and after completing the prokka run, its output file is Figure 1:

```
PROKKA_07072021.err PROKKA_07072021.fsa PROKKA_07072021.sqn
PROKKA_07072021.faa PROKKA_07072021.gbk PROKKA_07072021.tbl
PROKKA_07072021.ffn PROKKA_07072021.gff PROKKA_07072021.tsv
PROKKA_07072021.fna PROKKA_07072021.log PROKKA_07072021.txt
```

Figure1 Prokka Output File

In this genbank file generated from the genome, as in Figure 2, you can see its output (command=less PROKKA_07072021.gbk), including

the size of the genome, the different kinds of genes it contains, its origin and its gene sequence. This provides valid data for subsequent computational analysis. It can be said to provide a detailed annotation of the whole genome.

```

LOCUS      AE015451.2          6181873 bp    DNA      linear      07-JUL-2021
DEFINITION Genus species strain strain.
ACCESSION
VERSION
KEYWORDS   .
SOURCE     Genus species
ORGANISM   Genus species
           Unclassified.
COMMENT    Annotated using prokka 1.14.6 from
           https://github.com/tseemann/prokka.
FEATURES   Location/Qualifiers
            source          1..6181873
                               /organism="Genus species"
                               /mol_type="genomic DNA"
                               /strain="strain"
            CDS             complement(147..1019)
                               /gene="parB"
                               /locus_tag="GCA_000007565.2_ASM756v2_00001"
                               /inference="ab initio prediction:Prodigal:002006"
                               /inference="similar to AA sequence:UniProtKB:Q83AH2"
                               /codon_start=1
                               /transl_table=11
                               /product="putative chromosome-partitioning protein ParB"
                               /db_xref="COG:COG1475"
                               /translation="MAVKKRGLGRGLDALLSGPSVSALEEQAVKIDQKELQHLPVELV
                               QRGKYQPRRDMPEALEELAHSIRNHGVMQPIVVRPIGDNRYEIIAGERRWRATQQAG
                               LDTIPAMVREVPDEAAIAMALIENIQREDLNPLEEAMALQRLQQEFELTQQQVADAVG
                               KSRVTVANLLRLITLPDAIKTMLAHGDLEMGHARALLGLDENRQEEGARHVVARGLTV
                               RQTEALVRQWLSDKPDVPESKPDPIARLEQRLAERLGSQVQIRHGNKKGQQLVIRY
                               NSLDELQGVLAHIR"

```

Figure2 Genome Kernel

3.2 Pan-genome

When run roary is complete, an analysis graph of the pangenome can be output, analysing the different organism colonies by comparing the presence and absence of genes in the genome with each other to give a clearer picture of the presence of genes in the genome.

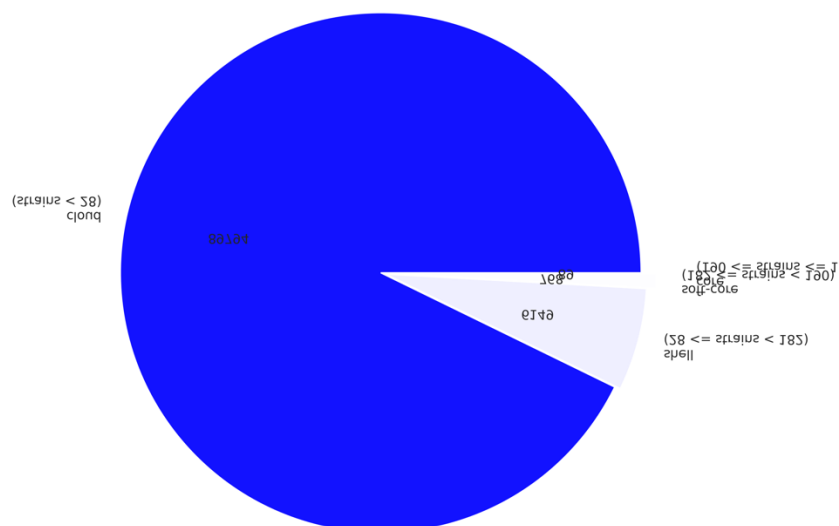


Figure 3 pangenome pie

As shown in Figure 3, The total number of genomes possessed by *Pseudomonas putida* is 194, of which the classification of genes is as follow :

Core genes	(190% <= strains <= 194%)	89
Soft core genes	(182% <= strains <190%)	768
Shell genes	(28% <= strains < 182%)	6149
Cloud genes	(strains < 28%)	89794
Total genes		96800

From the genome analysis report provided by Roary, it appears that the only core genes associated with degradation in the genes of *Pseudomonas putida* is uao, which is Uric acid degradation bifunctional

protein.

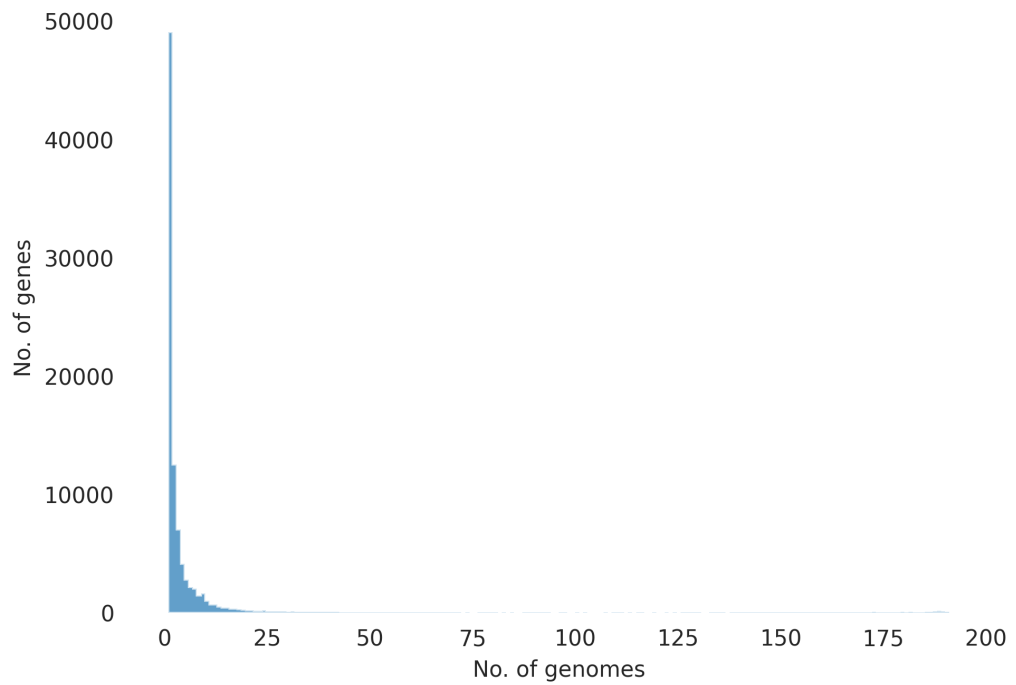


Figure 5 pangenome frequency

An analysis of the frequency of the pangenome is shown in Figure 5 above, and it can be seen that the frequency of activity of genomes in no.1-25 is frequent.

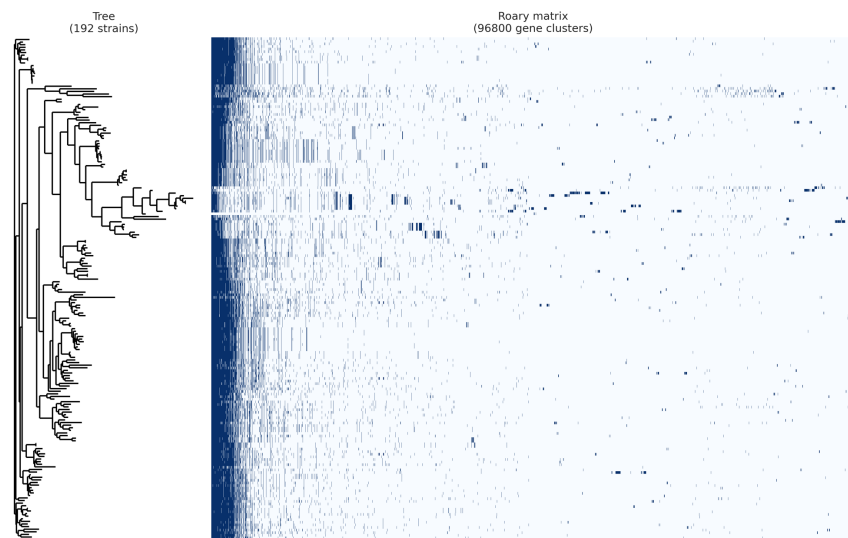


Figure 6 pangenome matrix

A pangenome matrix plot is shown in Figure 6, with the classification of the gene represented on the left and the frequency of its presence on the right. The tree file and present_absence file from the roary output file can be added via the phandango (Hadfield, J et al. 2018) website (<https://jameshadfield.github.io/phandango/#/>) to view the detailed distribution of genes. Figure 6 shows the higher frequency of genes over time, as shown in Figure 7.

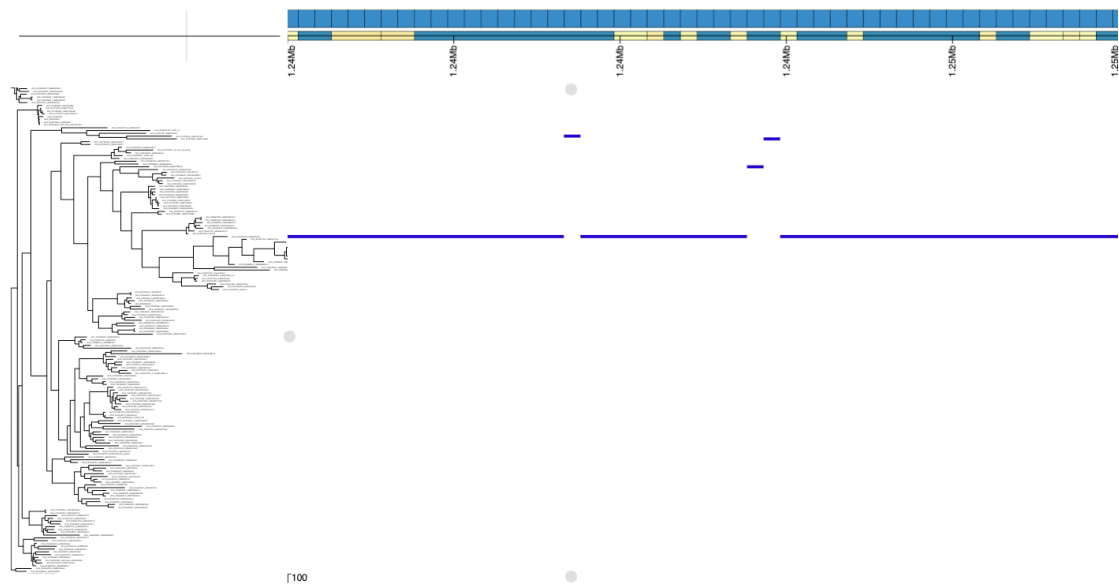


Figure 7 Phandango Output

3.3 Network and Heatmap of Pangenomes

Figures 8 and 9 show a schematic representation of the network and heatmap in the coinfinder output file. After clusters of homologous genes have been clustered together coinfinder can identify overlapping sets of genes in a pan-genomic dataset.

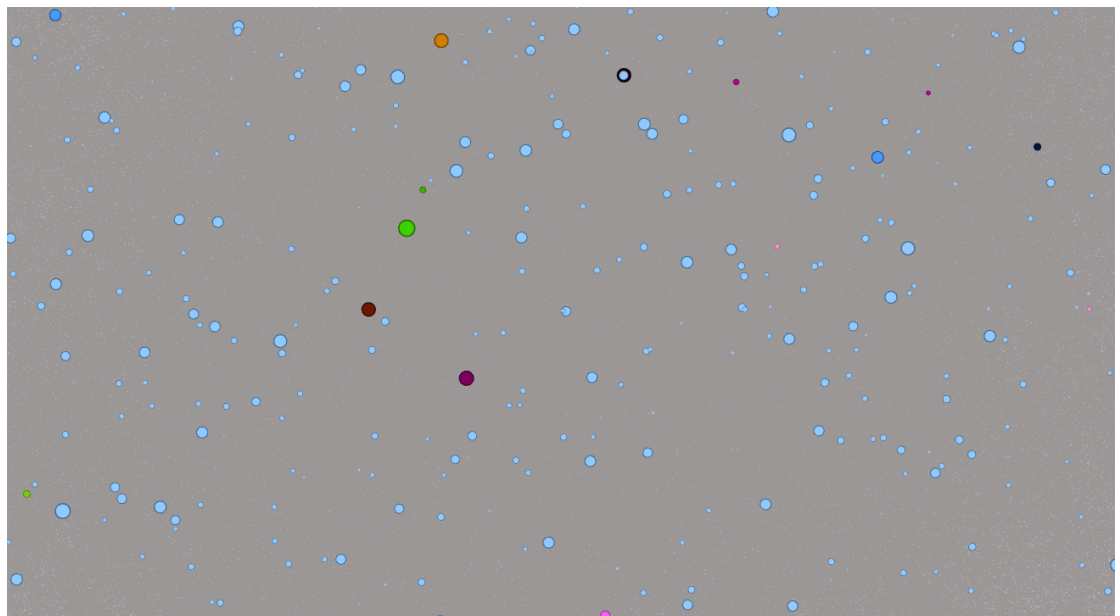


Figure 8 Network of Pangenomes

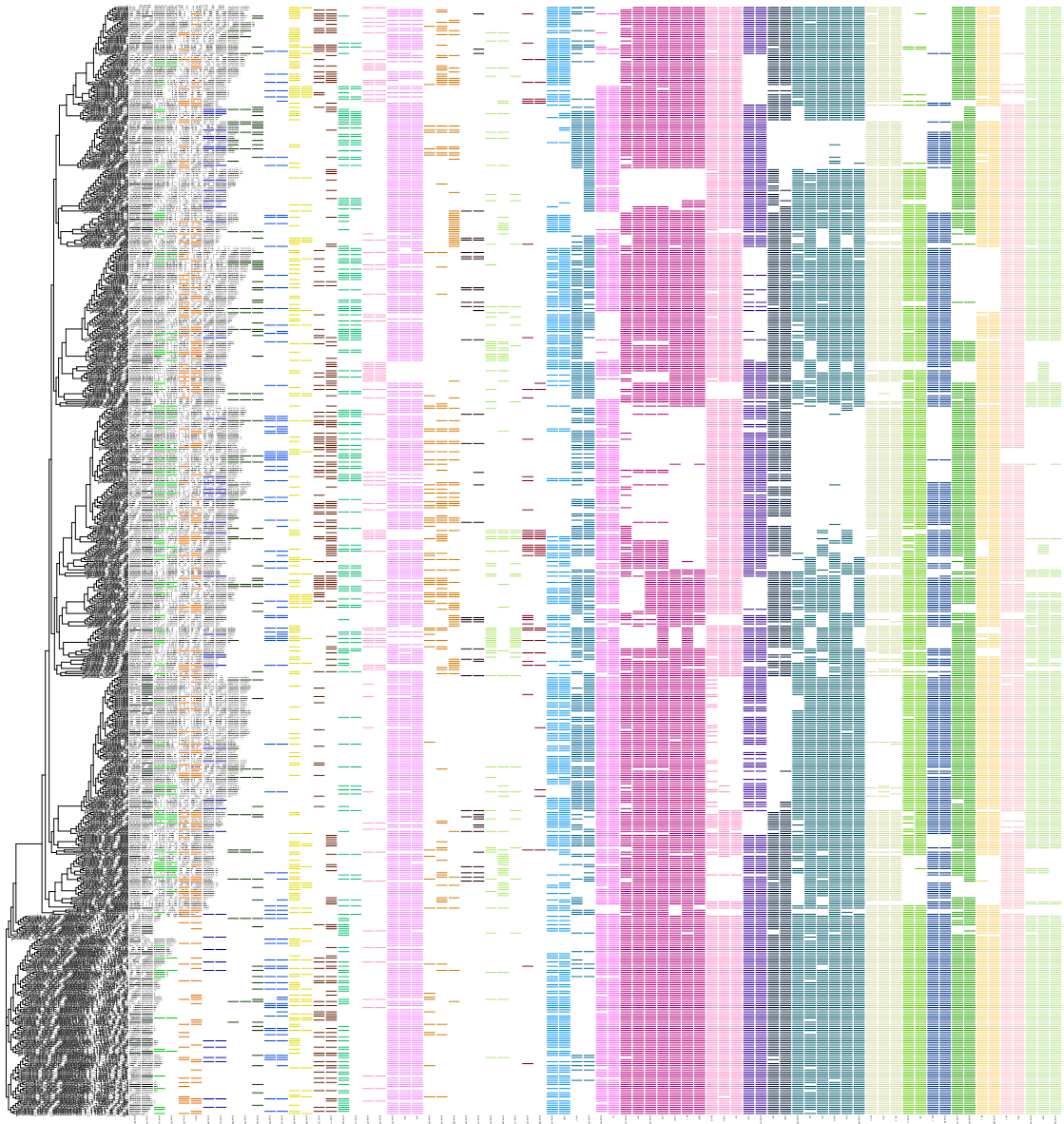


Figure 9 Heatmap of Pangenomes

A network and heatmap of the 194 genomes is shown in Figure 8 and Figure 9, and if each gene node is statistically related to each other in the pan-genome, they connect to another gene. The nodes are colored by the connected component (i.e. the set of reassociated genes) in colors corresponding to those used in the heatmap output. Similar to networks, each set of overlapping genes is co-colored. Genes are shown in relation to the input core gene phylogeny.

3.4 The Degradation of *Pseudomonas Putida*

In the analysis tables generated by metabolic, all the genomes associated with *pseudomonas putida* can be found along with the gene's category, function and the associated set of analysis results. Of these, the genomes associated with degradability and not completely absent are shown in Table 3 below.

Gene.name	Category	Function
acyl-CoA dehydrogenase	Fatty acid degradation	Fatty acid degradation
cellulase	Complex carbon degradation	Cellulose degrading
beta-glucosidase	Complex carbon degradation	Cellulose degrading
beta-galactosidase	Complex carbon degradation	Other oligosaccharide degrading
alpha-amylase	Complex carbon degradation	Amylolytic enzymes
isoamylase	Complex carbon degradation	Amylolytic enzymes
hexosaminidase	Complex carbon degradation	Chitin degrading
catechol 1,2-dioxygenase	Aromatics degradation	Protocatechuate/Catechol degradation
flavin prenyltransferase	Aromatics degradation	Phenol => Benzoyl-CoA

Table 3 Degradation of *Pseudomonas Putida*

From Table 3 it can be seen that in terms of degrading microorganisms, *pseudomonas putida* species are mainly found in three areas, Fatty acid degradation, complex carbon degradation and Aromatics degradation, after deleting genes that are completely absent from the pan-genome, a total of There were nine genes. Of these, acyl-CoA dehydrogenase oxidises short-chain aliphatic substrates of fatty acids (Brian McMahon et al. 2005); in the case of cellulose degradation, *pseudomonas putida* is surface bound as a cellulase to achieve co-hydrolysis of cellulose substrates (Tozakidis, I.E.P et al. 2016), hemicellulose debranching has a similar role (Wang, Y et al. 2019); the ability of bacteria to use low molecular weight lignin suggests that bacteria have a number of unique and specific enzymes that catalyse the production of a variety of useful compounds, lignin decomposes *Pseudomonas malodorosa* mt-2 (Tahir et al. 2019); other oligosaccharides are not yet mentioned in the degradation of *pseudomonas putida*, and only alginate lyase is described as degrading (qian.I et al. 2020); microbial cells can carry or secrete amylolytic enzymes, and Shuba et al. (2010) reported that soil samples

isolated from *Pseudomonas putida* containing chitin were characterised by chitin degradation; Urszula G (2011) noted that activated sludge samples from wastewater treatment plants were isolated from a strain of *Pseudomonas putida* designated as N6. Catechol 1,2-dioxygenase in strains designated as N6 up to a certain concentration can completely degrade phenol; flavin prenyltransferase and vanillate can whole-cell regulate the mode supporting the hypothetical activity of the enzyme in anoxic aromatic metabolism and degradation of the compound through intermediates from Phenol to Benzoyl-CoA (Werner D. et al 1991); benzoyl coenzyme A reductase (BCR) catalyzes the dearylation of benzoyl coenzyme A (benzoyl-CoA), the central step in the anaerobic degradation pathway of various aromatic compounds. Benzoyl coenzyme A reductase (BCR) catalyzes dearomatization of benzoyl coenzyme A (benzoyl-CoA), which is the central step in the anaerobic degradation pathway. The BCR catalyzes dearomatization of benzoyl coenzyme A (benzoyl-CoA), which is the central step in the anaerobic degradative pathways for a variety of aromatic compounds.

3.5 *Pseudomonas putida* crude oil degradation

A comprehensive analysis of the genome and an overview of the literature has been presented above, and this section will present and analyse the functions of genes in *pseudomonas putida* that are capable of contributing to the crude oil degradation field and characterise them by analogy.

3.5.1 Overview of existing research

From the analysis reported above and from literature studies on crude oil degradation, four of the genes in *pseudomonas putida* have been shown to play a role in the field of crude oil degradation. By table 4.

Gene.name	Function
alkanesulfonate monooxygenase	Sulfite production from organic sulfur Alkylsulfonate/Isethionate -> Sulfite
alpha-amylase	Amylolytic enzymes
isoamylase	Amylolytic enzymes
catechol 1,2-dioxygenase	Protocatechuate/Catechol degradation

Table 4 Existing Psudomonas Putida of Oil Degradation

In a study conducted on crude oil-contaminated soil (Abbasian, F et al. 2016), alkane monooxygenase was involved in hydrocarbon degradation, and alkane sulfonate monooxygenase was involved in the degradation of alkanes and sulfonated alkanes (Pal, S et al. 2017). And (Pal, S et al. 2017) noted that microorganisms isolated from crude oil sludge with growth substrates containing alpha-D-glucose had very good growth in crude oil. Genes related to degradation functions alpha-amylase and isoamylase were present in the results of the analysis. catechol 1,2-dioxygenase is an aromatic hydrocarbon and aromatic hydrocarbons are the main compounds in crude oil and aromatic hydrocarbons can be biodegraded to less toxic compounds in several steps for the purpose of degrading crude oil (Tavakoli, A et al. 2017).

3.5.2 statistic of pangenome activity

Figure 10 shows the statistics for the active regions of the genome, from which it can be seen that there are seven points in the regions where the genes occur frequently, which means that the genomes in these regions share the same functions and properties. Taking genome 53 and 70 (GCA_001467195.1 and GCA_001904555.1) as an example, it can be seen through the ncbi database that their size is 5.13991 6.76743, and GC percentage: 57.7% ,62.3%, and in genome20 and genome126, their sizes and GC percentages are 5.83022 and 5.52011;

61.4% and 63.3%, respectively, and they share a common core gene, uao, in the roary core genes.

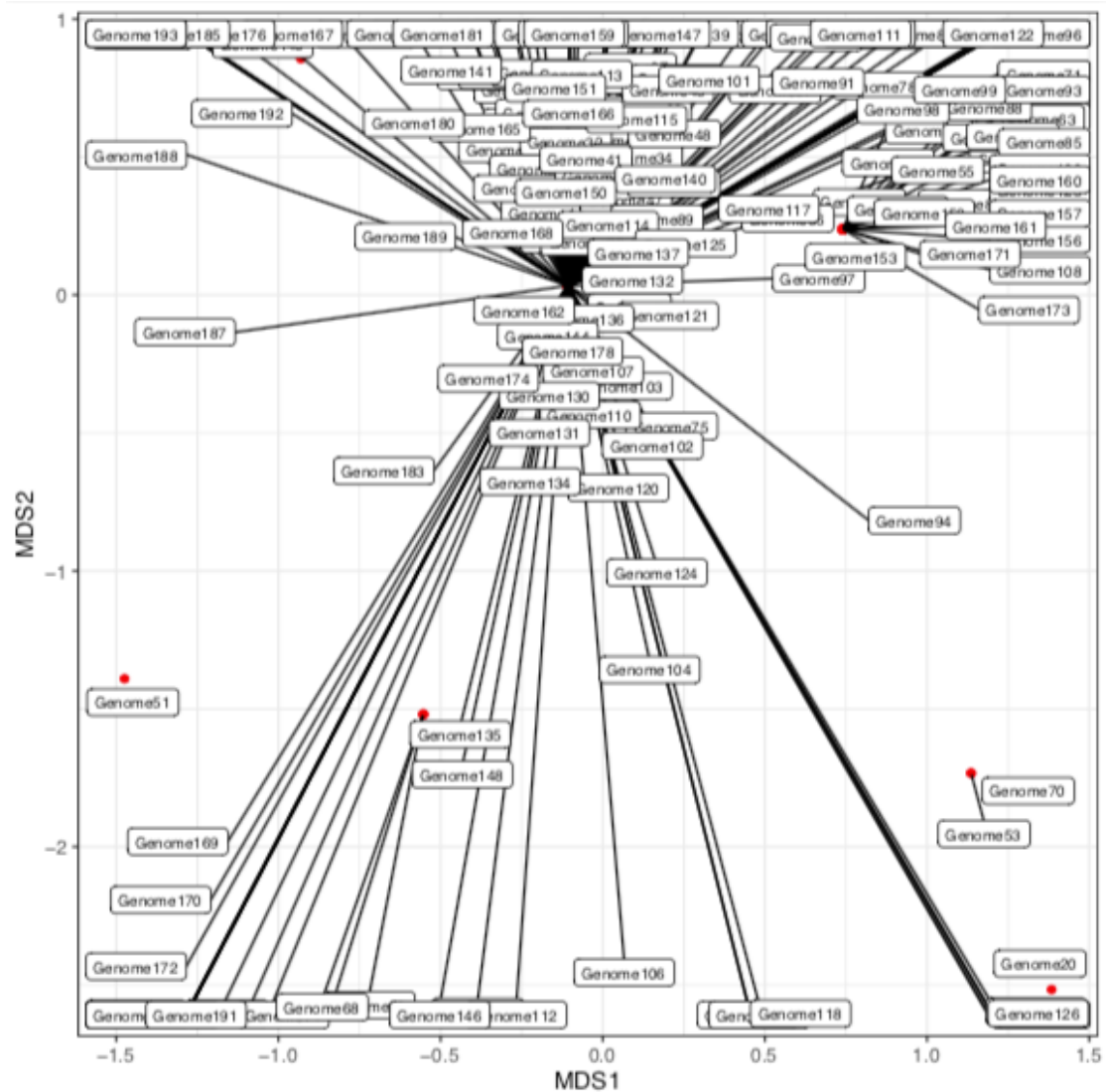


Figure 10 Jaccard Output

Table 5 CAN Output

MCF						
	outcome	condition	consistency	coverage	complexity	minimal
1	GENE1	gene4->GENE1	1	0.141361 26	1	TRUE
2	GENE1	gene4->GENE1	1	0.005235 6	1	TRUE
3	GENE2	GENE3->GENE 2	1	0.96875	1	TRUE
4	GENE2	GENE4->GENE 2	1	0.864583 33	1	TRUE
5	GENE2	gene1->GENE2	1	0.010416 67	1	TRUE
6	GENE4	gene1->GENE4	1	0.012048 19	1	TRUE
ASF						
	outcome	condition	consistency	coverage	complexity	minimal
1	GENE2	GENE3<->GEN E2	1	0.96875	1	TRUE
CSF						
	outcome	condition	consistency	coverage	complexity	minimal
1	GENE2	GENE3<->GEN E2	1	0.96875	1	TRUE

Table 5 shows the output of the analysis in the R package which is a configuration comparison method for causal data analysis. There is a strong correlation between genes 2 and 3, and that the

corresponding genes are alpha-amylase and isoamylase, which are genes belonging to the same role and species.

4. Discussion

The aim of this project was to carry out a genome-wide analysis of *Pseudomonas putida* through a series of software processes and data analysis. The results of the analysis provide a clear picture of the type, form and function of the entire genome and gene sequences of *Pseudomonas putida* and the role of the different genes. It is also possible to see that genome belonging to the same genotype have different gene sequences. It is also possible to see the metabolic and reactive forms of each of their genes.

As can be seen in Figure 8, the pan-genome grid shows that the coinfinder aggregates the pan-genomes into homologous clusters of genes identified as overlapping sets of genes, aggregated into seven different coloured points, meaning that the pan-genomes are statistically related or separated at the same point. And the same definition is used in the hotspot map. At the same time, the genes with degradation roles in the *Pseudomonas putida* present have seven different functions. And, as shown in Figure 10, the genome is clustered at seven points as well, and these regions of the genome are similar or related. Whether this means that the genomes have the same properties in the seven degradation modes remains to be further investigated.

In the genome-wide analysis of *Pseudomonas putida*, complex carbon degradation accounted for one-half of the genes in the gene species used for degradation, but this has not been argued in the study. However, hydrocarbon degradation is the main form of degradation in *Pseudomonas putida*. And of the genes already present that degrade crude oil, they are all in the form of enzymes or enzyme species. This suggests that *Pseudomonas putida* still has great potential for degrading crude oil.

5. Conclusion

Through the sequencing analysis of the whole genome of *Pseudomonas putida*, this project demonstrates that *Pseudomonas putida* has an important role in the field of petroleum degradation. The existence of many genomes (194) to degrade compounds, the possession of seven different functional genes that can play a role in the field of degradation, and their expression are evidence of the important role that microorganisms play in bioremediation and environmental protection. Of course, this project also has certain drawbacks, as it carries out a completely purely theoretical analysis and data statistics and does not produce experimental results in the laboratory, which lacks a certain degree of convincing and practical ability. However, it can also help the laboratory to save time by targeting genomes for experiments. Overall, with more research and practice, *Pseudomonas putida* will play a great role in the degradation of petroleum compounds.

Reference

Wang, L.C. (2009). Oil exploration impacts to environment and protection counter measure-Taking an example of Zhongyuan oilfield. *Ocean University of China*.

He, L. J., Wei, D. Z., Zhang, W. Q. (1999)Research of microbial treatment of petroleum contaminated soil. *Advances in Environmental Science*, 7(3), 110-111.

Li, C.R. (2009).Ecological effects and bioremediation of petroleum-contaminated soil. *Shaanxi: Chang'an University*.

Zhu, W. (2010). Petroleum polluted soil and sludge bio-treatment technology. *Beijing: China petrochemical press*. 26-29.

Shan, B. Q., Zhang, Y. T., Cao, Q. L. (2014). Growth responses of six leguminous plants adaptable in Northern Shaanxi to petroleum contaminated soil. *Environmental Science*.35,1125-1130.

Kaur, N.,Erickson.,Todd, E.B., Andrew, S.R., Megan,H.(2017).A review of germination and early growth as a proxy for plant fitness under petrogenic contamination. knowledge gaps and recommendations. *Science of the Total Environment*, 603-604,728-744.

Pena, P. G. L., Northcross, A. L.,Lima, M. G. d., Rêgo, R. d. Cássia, .F.(2020). Derramamento de óleo bruto na costa brasileira emergência em saúde pública em questão. *Cadernos de saude publica*,36(2), e00231019.

Carls, M. G., Babcock, M. M., Harris, P. M., Irvine, G. V., Cusick, J. A., Rice, S. D. (2001) Persistence of oiling in mussel beds after the Exxon Valdez oil spill. *Marine Environmental Research*, 51(2),167-190.

Pezeshki, S. R., Hester, M. W., Lin, Q., Nyman, J. A. (2000). The effects of oil spill and clean-up on dominant US Gulf coast marsh macrophytes: A review. *Environmental Pollution*, 108(2),129-139.

Atlas, R.M. (1981). Microbial degradation of petroleum hydrocarbon: an environmental perspective. *Microbiology Review*, 45, 185-209.

Raghavan,P.U.M. and Vivekanandan, M. (1999). Bioremediation of oil-spilled sites through seeding of naturally adapted *Pseudomonas putida*. *International Biodeterioration and Biodegradation*, 44(1), 29-32.

Siddhartha.P., Anirban.K., Tirtha.D.B., Balaram.M., Ajoy.R., Riddha.M., Pinaki.S., Sufia.K.(2017). Genome analysis of crude oil degrading *Franconibacter pulveris* strain DJ34 revealed its genetic basis for hydrocarbon degradation and survival in oil contaminated environment. *Genomics*, 109(5-6),374-382.

Loh, K. C.,Cao, B. (2008).Paradigm in biodegradation using *Pseudomonas putida*-A review of proteomics studies. *Enzyme and Microbial Technology*,43(1),1-12.

Zheng, M.Y., Wang, W.Y., Hayes, M., Nydell, A., Tarr, M. A., Van,B., Sunshine, A., Papadopoulos, K. (2018). Degradation of Macondo 252 oil by endophytic *Pseudomonas putida*. *Journal of Environmental Chemical Engineering*, 6(1), 643-648.

Munoz, R., Diaz, L.F., Bordel, S., Villaverde, S. (2007). Inhibitory effects of catechol accumulation on benzene biodegradation in *Pseudomonas putida* F1 cultures. *Chemosphere*, 68,244-52.

Yu, H., Kim, B.J., Rittmann, B. E. (2001). The roles of intermediates in biodegradation of benzene, toluene, and p-xylene by *Pseudomonas putida* F1. *Biodegradation*, 12,455-63.

Ohta, Y., Maeda, M., Kudo, T. (2001). *Pseudomonas putida* CE2010 can degrade biphenyl by a mosaic pathway encoded by the *tod* operon and *cmtE*, which are identical to those of *P. putida* F1 except for a single base difference in the operator–promoter region of the *cmt* operon. *Microbiology*,147,31-41.

Juang, R.S., Tsai, S. Y. (2006). Enhanced biodegradation of mixed phenol and sodium salicylate by *Pseudomonas putida* in membrane contactors. *Water Resources*,40, 3517-26.

Komukai, N., Syoko, S., Kei. J. Y. I., Yukie, T., Haruhisa, V., Kasthuri, Y., Satoshi, T., Hiroki, H., Shigeaki. (1996). Construction of bacterial consortia that degrade Arabian light crude oil. *Journal of Fermentation and Bioengineering*, 82(6) 570-574

Shen, Y.B., Shi, X.M., Shen, J.Z. (2019).Application of whole genome sequencing technology and bioinformatics analysis in antimicrobial resistance researches. *Chinese Journal of Biotechnology*. 35(4),541-557.

Whelan, F. J., Rusilowicz, M., McInerney, J. O. (2020). Coinfinder: detecting significant associations and dissociations in pangenomes. *Microbial genomics*, 6(3), e000338.

Brian, M., Mary, E. G., Stephen, G. M., (2005). The protein coded by the PP2216 gene of *Pseudomonas putida* KT2440 is an acyl-CoA dehydrogenase that oxidises only short-chain aliphatic substrates. *FEMS Microbiology Letters*, 250(1), 121–127.

Tozakidis, I.E.P., Brossette, T., Lenz, F. (2016). Proof of concept for the simplified breakdown of cellulose by combining *Pseudomonas putida* strains with surface displayed thermophilic endocellulase, exocellulase and β -glucosidase. *Microb Cell Fact*, 15, 103.

Wang, Y., Horlamus, F., Henkel, M. (2019). Growth of engineered *Pseudomonas putida* KT2440 on glucose, xylose, and arabinose: Hemicellulose hydrolysates and their major sugars as sustainable carbon sources. *GCB Bioenergy*, 11, 249– 259.

Tahir, A.A., Barnoh, N. F., Mohd, Y., Nurtasbiyah, S., Nuurul, N., Mohd, U., Motoo, Y., Ang, M. H., Hazni, N., Megat, J. M. M. A., Fazrena, N. M.D., Mohamad, S. E., Sugiura, N., Othman, Nor.A., Zakaria, Z., Hara, H. (2019). Microbial diversity in decaying oil palm empty fruit bunches (OPEFB) and isolation of lignin-degrading bacteria from a tropical environment. *Microbes and Environments*, 34(2), 161-168.

Li, Q., Hu, F., Wang, M.Y., Zhu, B.W., Ni, F., Yao, Z. (2020). Elucidation of degradation pattern and immobilization of a novel alginate lyase for preparation of alginate oligosaccharides. *International Journal of Biological Macromolecules*, 146, 579-587.

Schwartz, A., Bar, R. (1995). Cyclodextrin-enhanced degradation of toluene and p-toluic acid by *Pseudomonas putida*. *Applied and Environmental Microbiology*, 61(7), 2727-2731.

Das, S.N., Sarma, P.V.S.R.N., Neeraja, C. (2010). Members of Gammaproteobacteria and Bacilli represent the culturable diversity of chitinolytic bacteria in chitin-enriched soils. *World J Microbiol Biotechnol*, 26, 1875–1881.

Urszula, G., Izabela, G., Katarzyna, H. K., Danuta, W. (2011). Catechol 1,2-dioxygenase from the new aromatic compounds – Degrading *Pseudomonas putida* strain N6. *International Biodeterioration & Biodegradation*, 65(3), 504-512.

Werne, D., Ruth, B., Achim, L., Magdy, M., Jiirgen, K., Brigitte, O., Birgit, S., Andreas, T., Georg, F. (1991). Differential expression of enzyme activities initiating anoxic metabolism of various aromatic compounds via benzoyl-CoA. *Arch Microbiol*, 155, 256 – 262.

- Abbasian, F., Palanisami, T., Megharaj, M., Naidu, R., Lockington, R., Ramadass, K. (2016). Microbial diversity and hydrocarbon degrading gene capacity of a crude oil field soil as determined by metagenomics analysis. *Biotechnology progress*, 32(3), 638–648.
- Pal, S., Kundu, A., Banerjee, T. D., Mohapatra, B., Roy, A., Manna, R., Sar, P., Kazy, S. K. (2017). Genome analysis of crude oil degrading *Franconibacter pulveris* strain DJ34 revealed its genetic basis for hydrocarbon degradation and survival in oil contaminated environment. *Genomics*, 109(5-6), 374–382.
- Tavakoli, A., Hamzah, A. (2017). Characterization and evaluation of catechol oxygenases by twelve bacteria, isolated from oil contaminated soils in Malaysia. *Biological Journal of Microorganisms*, 5(20), 71-80.
- Firouz, A., Thavamani, P., Mallavarapu, M., Ravi, N., Robin, L., Kavitha, R. (2016). Microbial Diversity and Hydrocarbon Degrading Gene Capacity of a Crude Oil Field Soil as Determined by Metagenomics Analysis. *American Institute of Chemical Engineers*. 32(3),638-648.
- Azubuikwe, C. C., Chikere, C. B., Okpokwasili, G. C. (2016). Bioremediation techniques-classification based on site of application: principles, advantages, limitations and prospects. *World journal of microbiology & biotechnology*, 32(11), 180.
- Babita, K.S.N., Singh, D.P., Singh. (2012). Characterization of two biosurfactant producing strains in crude oil degradation. *Process Biochemistry*, 47(12), 2463-2471.
- Kumar, M., Leon, V., De, S. M., Angela, I., Olaf, A. (2006). Enhancement of oil degradation by co-culture of hydrocarbon degrading and biosurfactant producing bacteria. *Polish Journal of Microbiology*, 55(2), 139-146.
- Kumar, N. M., Muthukumaran, C., Sharmila, G., Gurunathan, B. (2018). Genetically Modified Organisms and Its Impact on the Enhancement of Bioremediation. *Bioremediation: Applications for Environmental Protection and Management*.
- Kumar, S., Upadhyay, S. K., Kumari, B., Tiwari, S., Singh, S.N., Singh, P. (2011). *In vitro* degradation of fluoranthene by bacteria isolated from petroleum sludge. *Bioresour Technol.*

- Kumari, K., Singh, G., Sinam, G., Singh, D.P. (2020). *Microbial Remediation of Crude Oil-Contaminated Sites*. Environmental Concerns and Sustainable Development: Volume 1: Air, Water and Energy Resources.
- Nyer, E. K., Payne, F., Suthersan, S. (2002). Environment vs. bacteria or let's play 'name that bacteria. *Ground Water Monit Remediat*, 23, 36–45.
- Koolivand, A., Naddafi, K., Nabizadeh, R., Nasser, S., Jafari, A. J., Yunesian, M., Yaghmaeian, K., Nazmara, S. (2013). Biodegradation of petroleum hydrocarbons of bottom sludge from crude oil storage tanks by in-vessel composting. *Toxicol Environ Chem*, 95(1):101–109.
- Kumari, B., Singh, S. N., Singh, D. P. (2016). Induced degradation of crude oil mediated by microbial augmentation and bulking agents. *Int J Environ Sci Technol*, 13(4), 1029–1042.
- Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22), 3691–3693.
- Zhou, Z. C., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., Anantharaman, K. (2020). METABOLIC: High-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. *BioRxiv*.
- James, H., Nicholas, J. C., Richard, J. G., Khalil, A., David, M. A., Simon, R. H. (2018). Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, 34(2), 292–293.
- Whelan, F. J., Rusilowicz, M., McInerney, J. O. (2020). Coinfinder: detecting significant associations and dissociations in pangenomes. *Microbial genomics*, 6(3), e000338.

Appendix 1

```
[studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida/test]$ ftp
ftp.ncbi.nlm.nih.gov
Name (ftp.ncbi.nlm.nih.gov:studentprojects): anonymous
your password
Password:
230 Anonymous access granted, restrictions apply
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> cd genomes
250 CWD command successful
ftp> cd genbank
250 CWD command successful
ftp> ls Pseudomonas*
ftp> cd bacteria
ftp> bye
[studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida/test1]$ for f
in `cat var.txt`; do name=$(grep -w "$f" assembly_summary.txt
| cut -f9 | cut -f2 -d=' ' | sed 's/ /_/g' | sed 's/\\/_/g' |
sed 's/\\:/_/g' | sed 's/)/_/g' | sed 's/(/_/g'); xx=$(grep -w
"$f" assembly_summary.txt | cut -f20 | cut -f10 -d'/'); wget -
--tries=75 -c $f/$xx\_genomic.fna.gz; done
--2021-08-14 18:33:45--
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/016/845/885/GCA_016
845885.1_ASM1684588v1/GCA_016845885.1_ASM1684588v1_genomic.fna
.gz
=> 'GCA_016845885.1_ASM1684588v1_genomic.fna.gz'
Resolving ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)...
130.14.250.12, 130.14.250.11, 2607:f220:41e:250::11, ...
Connecting to ftp.ncbi.nlm.nih.gov
(ftp.ncbi.nlm.nih.gov)|130.14.250.12|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done. ==> PWD ... done.
==> TYPE I ... done. ==> CWD (1)
/genomes/all/GCA/016/845/885/GCA_016845885.1_ASM1684588v1 ...
done.
==> SIZE GCA_016845885.1_ASM1684588v1_genomic.fna.gz ...
1372706
==> PASV ... done. ==> RETR
GCA_016845885.1_ASM1684588v1_genomic.fna.gz ... done.
Length: 1372706 (1.3M) (unauthoritative)
```

```
100%[=====]
=====>] 1,372,706    989KB/s   in 1.4s
```

```
2021-08-14 18:33:48 (989 KB/s) -
'GCA_016845885.1_ASM1684588v1_genomic.fna.gz' saved [1372706]
```

Appendix 2

```
[studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/test1]$ export
PATH=/home/opt/miniconda2/bin:$PATH
[studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/test1]$ unset
PERL5LIB
[studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/test1]$ source
activate pangenome
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/test1]$ prokka
-h
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/test1]$ prokka
GCA_014764425.1_ASM1476442v1_genomic.fna --locustag
GCA_014764425.1 --outdir GCA_014764425.1
```

Appendix 3

```
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
mkdir roary
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
ls -d */*.gff
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
for i in $(ls */*.gff); do cp $i roary/$(echo $i | sed
's!/.*!!').gff; done
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
unset MAFFT_BINARIES

(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
export PATH=/home/opt/miniconda2/bin:$PATH
```

```
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
source activate pangenome
```

```
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
export PERL5LIB=/usr/local/lib/perl5/site_perl/5.22.0/
```

```
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
roary -h
```

```
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]$
roary -f ./roary_tree -e -n -v -p 8 -v -r -I 60 --group limit
8000 ./roary/*.gff
```

```
(pangenome) [studentprojects@moore
/shared5/studentprojects/zhen/Pseudomonas_putida/RAW_genomes]
watch -n 1 "ls -larth | tail -5"
```

Appendix 4

```
[studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes]
$ export PATH=/home/opt/miniconda2/bin:$PATH
```

```
[studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes]
$ source activate coinfinder-env
```

```
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes]
$ mkdir coinfinder-test
```

```
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes]
$ cd coinfinder-test
```

```
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ git clone
```

```
https://github.com/fwhelan/coinfinder-manuscript.git
```

```
Cloning into 'coinfinder-manuscript'...
```

```
remote: Enumerating objects: 31, done.
```

```
remote: Counting objects: 100% (31/31), done.
```

```
remote: Compressing objects: 100% (24/24), done.
```

```
remote: Total 31 (delta 8), reused 29 (delta 6), pack-reused 0
```

```
Unpacking objects: 100% (31/31), done.
```

```
Checking out files: 100% (18/18), done.
```

```
(coinfinder-env) [studentprojects@howe
```

```

/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ ls
coinfinder-manuscript
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ ls -larth
total 203K
drwxrwxr-x 201 studentprojects studentprojects 394 Aug 15
11:25 ..
drwxrwxr-x 3 studentprojects studentprojects 3 Aug 15
11:25 .
drwxrwxr-x 3 studentprojects studentprojects 21 Aug 15 11:25
coinfinder-manuscript
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ ls
coinfinder-manuscript
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ cp coinfinder-
manuscript/gene_presence_absence.csv .
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ cp coinfinder-manuscript/core-
gps_fasttree.newick .
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ ls
coinfinder-manuscript core-gps_fasttree.newick
gene_presence_absence.csv
(coinfinder-env) [studentprojects@howe
/shared5/studentprojects/zhen/Pseudomonas_putida2/RAW_genomes/
coinfinder-test]$ coinfinder -i gene_presence_absence.csv -I -
p core-gps_fasttree.newick -o output

```

Appendix 5

```

abund_table<-read.csv("presence_absence_table.csv",header =
TRUE,row.name = 1)
#install.packages("vegan")
library(vegan)
library(ggplot2)
library(ggrepel)
#help("vegdist")

```

```

#help(package="vegan")
#vignette(package="vegan")
#vignette(package="vegan","intro-vegan ")
ggrepel.max.overlaps = Inf
abund_table.dist<-vegdist(abund_table,method="jaccard")

#ord<-metaMDS(abund_table,distance="jaccard",k=5)
ord<-capscale(abund_table ~
1,distance="jaccard")#PCoA(Principle Coordinate Analysis

#Now we use the scores() to extract the location of Genomes
df<-as.data.frame(scores(ord, display = "sites"))
pdf("myplot.pdf")

p <- ggplot(df,aes(MDS1,MDS2))
p<-p+geom_point(color = "red")
p<-p + geom_label_repel(aes(label = rownames(df)),size = 2.5)
+ theme_bw()
print(p)
dev.off()

```

Appendix 6

```

abund_table<-
read.csv("presence_absence_table.csv",header=TRUE,row.name=1)

```

```

#install.packages("cna")

```

```

library(cna)

```

```

#help(package="cna")

```

```

threshold_perc<-0.9

```

```

abund_table.cna<-cna(abund_table,cov=threshold_perc)

```

```

#Take hints from here: https://rdr.io/cran/cna/man/condTbl.html

```

```

#One of these three methods should work,if not then check help("msc"), or
help("asf"), or help("csf")

```

#All the files are stored as CSV files

```
df.msc<-as.data.frame(msc(abund_table.cna))
```

```
write.csv(df.msc,"MSC.csv")
```

```
df.asf<-as.data.frame(asf(abund_table.cna))
```

```
write.csv(df.asf,"ASF.csv")
```

```
df.csf<-as.data.frame(csf(abund_table.cna))
```

```
write.csv(df.csf,"CSF.csv")
```