Coursework Declaration and Feedback Form

Student Number: 2562963C	Student Name: Steffy Anna Cherian
Program of Study: MSc in Civil and Management	
Course Code: ENG5059P	Course Name: MSc Project
Name of First Supervisor: Dr.Umer Ijaz	Name of Second Supervisor: Dr. Ankush Aggarwal
Title of Project: Whole Genome Analysis of Acetotrophic and Hydrogenotrophic Methanogenic Archaea	
Declaration of Originality and Submission Information	
I affirm that this submission is all my own workin accordance with the University of Glasgow Regulations and the School of Engineering requirements Signed (Student):	ENG5059
Date of Submission: 25/08/2021	

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

University of Glasgow | School of Engineering

WHOLE GENOME ANALYSIS OF ACETOTROPHIC AND HYDROGENOTROPHIC METHANOGENIC ARCHAEA

AUGUST 2021

A thesis submitted in partial fulfilment of the requirements for the degree of
MASTERS IN CIVIL ENGINEERING AND MANAGEMENT

ABSTRACT

Increase in the production of wastewater on the daily basis calls for wastewater treatment which is inevitable in all parts of the world. There are many common ways for treating wastewater like chemical process including ion exchange, neutralisation as well as biological methods like activated sludge anaerobic digestion. This thesis focuses on methanogens, a group of organisms found in the euryarchaeotic phylum involved in the methanogenesis process of anaerobic digestion reactors which is involved in the decomposition of organic matter in the sludge under the absence of oxygen. Methanogenesis is the process of producing methane by breaking down organic matter by methanogens.

Specific species of methanogens like *Methanosarcina Barkeri, Methanothrix soehngenii and Methanobacterium formicicum, Methanococcus maripaludis* are selected from two distinct pathways: the acetolactic and hydrogenotrophic pathway. The former two methanogenic archaea belong to acetolactic pathway while the latter two belong to hydrogenotrophic pathway.

A method called whole-genome analysis is adopted where the analysis of the entire genome set of the above-mentioned species is done by downloading the genome from a public repository known as NCBI and then running it in Prokka (to annotate the genomes) as well as Roary (takes the annotated assemblies and calculates whole genome). The annotated data is then integrated into metabolic pathways, these are later on analysed with various visualisation tools. This way the genetic characteristics of all the four species are closely studied to understand their capabilities and potential to produce methane. The methanogens belonging to the same pathway are compared for clarity. But these methanogens may not function well in reactor environments with extreme conditions like varying temperatures, PH levels, acetate stress. The ability of the above said species to resist these extreme conditions are tested in Orion cluster using operating systems like Linux and R scripts for its genetic capacity.

ACKNOWLEDGMENT

I would like to take this opportunity to express my sincere gratitude to my head supervisor Dr. Ulmer Ijaz who guided me throughout my project. In addition to it I would like to thank my assistant supervisor Dr. Ciara Keating who gave me insights about bioinformatics and helped me in all the small and big steps. They took time out their busy schedules to provide guidance to us which is the reason why we could complete our project on time and smoothly. The guidance provided to us during the meetings by both of them were of paramount importance as it paved new direction for my project.

I am glad I was accepted for this project as I learned many new skills like software proficiency, importance of bioinformatics and whereabouts of methanogens. Being foreign to this field these three months were fascinating and interesting to me as I gained new knowledge every day. I am also thankful for the support of my project mates who helped me out whenever I faced any difficulty.

KEYWORDS AND ABBREVIATIONS

- AD- anaerobic digestion
- Methanogenesis
- *M. barkeri – Methanosarcina Barkeri*
- *M. soehngenii – Methanothrix soehngenii*
- *M.marpaludis – Methanococcus maripaludis*
- *M.formicicum- Methanobacterium formicicum*
- NCBI – National Centre for Biotechnological Information
- Archaea
- DNA- Deoxyribonucleic acid
- PCR- polymer chain reaction
- ddNTPs - dideoxy nucleotides
- NGS- New generation sequencing
- Whole genome sequencing

TABLE OF CONTENTS

1. INTRODUCTION

Around 1×10^6 m^3 and 7×10^6 m^3 of domestic and industrial wastewater is produced every day. This combined with the surface run-off as well as infiltration during rain creates a necessity of water treatment which is met with immense number of treatment plants in and around the UK (Gray & N.F, 2004). It is necessary to treat sewage water for a number of reasons including pollution control, environmental welfare, safeguarding public health, and so on. The aim of wastewater treatment is to clarify the water from suspended solids and organic matters which can be done by simple physical processes like sedimentation in huge reservoirs or tanks and to treat it into a reusable form so that they can be released back into a natural medium like rivers or used for some other purposes. Wastewater treatment can be carried out in many ways: both for organic wastes (done by biological processes) and inorganic wastes (done by chemical processes). The chain of processes to treat wastewater includes removing the debris by means of some physical process like passing it through a grit followed by sedimentation, it is then treated by methods ranging from activated sludge to anaerobic digestion before passing through the disinfector for further purification (Gray & N.F, 2004).

Anaerobic digestion is gaining light as it is one of the most productive and important stages in wastewater treatment. The skeleton of the process is simple, the wastewater passes through the anaerobic digestor probably an up-flow anaerobic sludge blanket reactor (UASB) where it interacts with an anaerobic sludge layer enriched with anaerobic archaea so that the break-down of organic waste happens in the absence of oxygen (Lettinga, 1995). The enclosed anaerobic chamber (UASB) is built under special specifications like standard PH, temperature, substrate composition, etc to enhance the life and growth of the anaerobic bacteria and methanogens in the reactor. The discharge from the AD reactor is then sent for tertiary treatment. Meanwhile, the anaerobic digestion gives a nitrogen-carbon rich sludge that can be used as manure as well as biomethane which is used as a renewable cleaner source of energy as its output (Lettinga, 1995).

AD is one of the main key players in the UK government's department of energy and climate change strategy. It is put forth as a vision to generate a clean source of energy from wastes generated paving the path towards a zero-waste plan. With already above 650 AD reactors in the country, this process is gaining a wide range of popularity and is a subject of study in the research field (GOV.UK, 2011).AD also serves as the best replacement to other renewable energy sources like wind energy, solar energy which unlike these are a constant and cost-effective source. Another major advantage of AD is that it can intake a huge amount of input waste or wastewater and treat it incessantly thus reducing the amount of waste build-up, unlike other waste treatment processes where the input of waste should be at a controlled rate before the next batch can be refilled for treatment (GOV.UK, 2011).

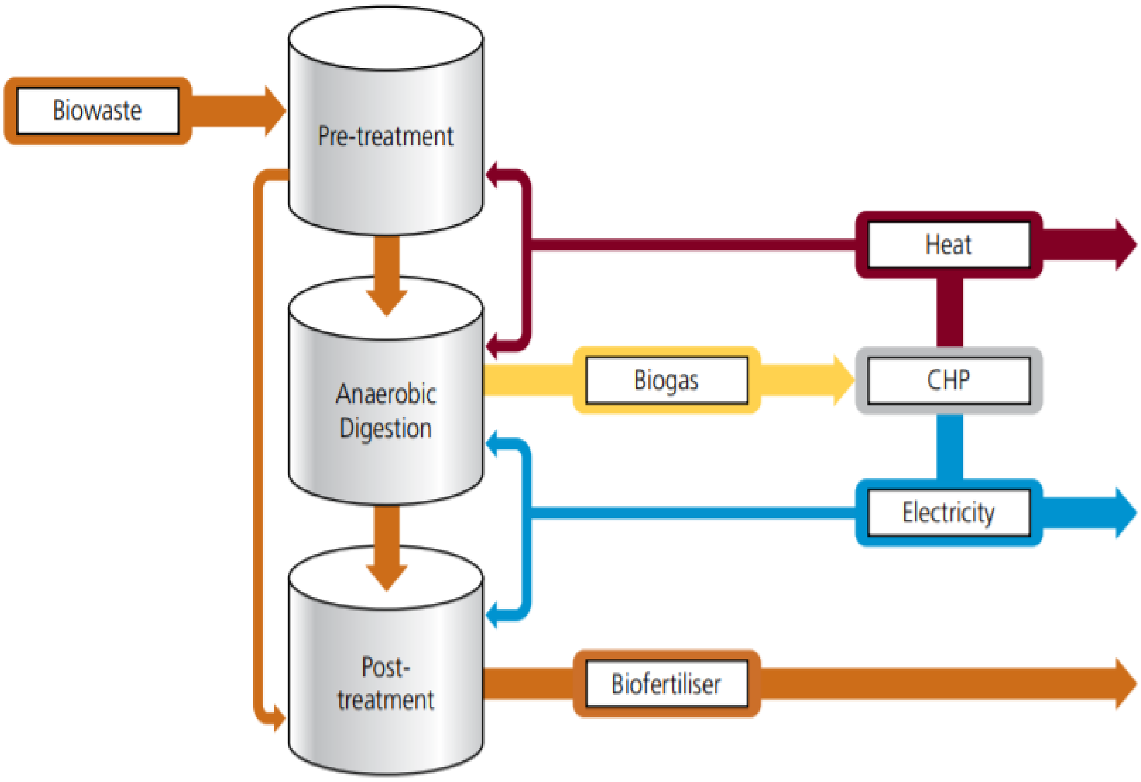


Figure 1. a schematic flow diagram showing the input/output discharges of anaerobic digestion

Anaerobic digestion can be subdivided into four biochemical reactions hydrolysis, acidogenesis, acetogenesis and methanogenesis.

- **Hydrolysis:**

This is the first stage in anaerobic digestion where complex organic compounds are broken down into simpler carbon compounds. Hydrolytic bacteria produce certain hydrolase enzymes that promote the digestion of carbohydrates, lipids, proteins into accessible forms of sugars and acids like amino acid and fatty acid for the next stage of AD. The efficacy of this process depends on the medium conditions like substrate size, PH, temperature, and so on.

Hydrolytic bacteria like Streptococcus and Enterobacterium function the best under 30-50^{0}C and PH 5-7 (Shah, et al., 2014).

- **Acidogenesis:**

Here the acidifying bacteria converts the end products of hydrolysis and water-soluble compounds into organic acids (acetic, formic acid), alcohol compounds like ethanol, methanol, aldehydes, CO_2, and hydrogen. Most of the products of acidogenesis can be used directly by methanogens for methane production like hydrogen while amino acid produced can be utilised as an energy source by anaerobic organisms. This is comparatively a faster stage of AD. Ammonia synthesised from amino acids if produced in excess can inhibit the AD process as the methanogens are extremely sensitive in nature (Shah, et al., 2014).

- **Acetogenesis:**

This is the stage that determines the capability of the AD reactor as the decomposition of the substrates and methane production depends on the amount of clean acetate produced. A certain amount of acetate generation happens in the previous step acidogenesis where acetate is created from lipids and other substrates. The remaining major portion of acetate development happens in this stage where all the carbon compounds get converted to acetate which acts as the substrate for acetate consuming methanogens. The hydrogen formed during acetogenesis is used by another group of methanogens for producing methane. Around 25% of acetate is produced in this stage (Shah, et al., 2014).

- **Methanogenesis:**

The stage where methanogenic archaea produce methane, carbon dioxide, and other by-products. Methane generation can either be done by using acetate, hydrogen, or by using some other compound depending on the type of methanogens. This is the most crucial at the same time-sensitive step in AD. A theoretical equation was developed for the production of the end products methane, CO_2, ammonia, and hydrogen sulphide in methanogenesis as given below:

$$\boxed{\begin{aligned} &C_nH_aO_bN_xS_y + \left(n - \frac{a}{4} - \frac{b}{2} + \frac{3x}{4} + \frac{y}{2}\right)H_2O \\ &\qquad \rightarrow \left(n - \frac{a}{4} + \frac{b}{2} + \frac{3x}{4} + \frac{y}{2}\right)CO_2 + \left(n + \frac{a}{4} - \frac{b}{2} - \frac{3x}{4} - \frac{y}{2}\right)CH_4 + xNH_3 + yH_2S \end{aligned}}$$

This equation is used in the design of AD reactors but does not give the right value in practical examples. Thus, the breakdown of waste in anaerobic digestion is done with the help of multiple micro-organisms of which methanogens are the lead participants (Meegoda, et al., 2018). This report is a focused study, analysis and comparative genomic survey of these methanogenic organisms which will be discussed in the later sections.

1.1 METHANOGENS AND METHANOGENESIS

Methanogens have enzyme activators producing methane that is distinguished from bacteria due to the absence of peptidoglycan in their cell wall. These methanogenic archaea are crucial in the class of micro-organisms due to their unique ability to convert carbon compounds in organic waste to methane, which is used a replacement for fossil fuels as it is cleaner and a renewable source of energy. The organic compounds used for energy generation (methane production) include acetate, lipids, hydrogen, CO_2, formate, methanol so on. Methanogens are distinctive in terms of their metabolic pathways and capability to decompose biodegradable waste while producing energy as well as ammonia-rich manure out of it (Shah, et al., 2014). They are strictly anaerobic microbes that can function in the absence of oxygen and can be found in extreme conditions making them distinctive from other microbes.

These methanogenic archaea belong to the eukaryotic phylum with five classes *Methanobacteriales, Methanococcales, Methanomicrobiales, Methanopyrales, and Methanosarcinales* ranging from mesophiles to thermophiles. They vary in shape from filamentous, rod-like to coccoid or spherical depending on their genus. The size of these micro-organisms ranges from 0.3 to 7.4 micrometres. They occur naturally in many habitats like freshwater, marine water, marshlands, other cold and hot environments as free-living organisms or in symbiotic relations with other microbes that help in sustaining their longevity.

The metabolic process of methane production is called methanogenesis (Nishio & Nakashimada, 2013). Depending on the pathway adopted by the methanogens and the substrate utilised for methane production they classified can be into three major groups: the acetoclastic methanogens, hydrogenotrophic methanogens, and methylotrophic methanogens which is explained in detail below:

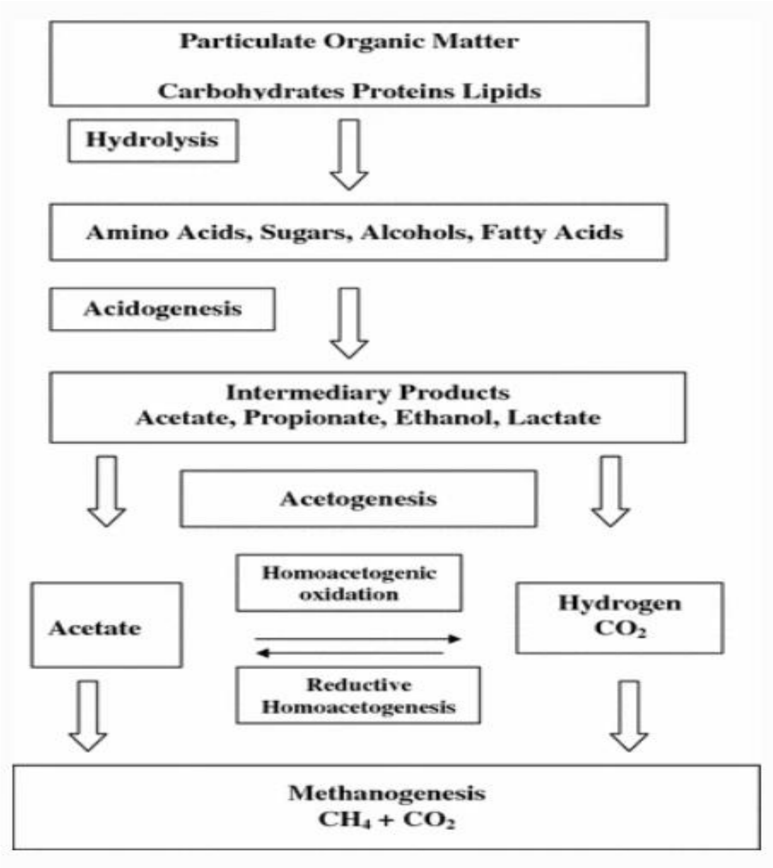

Particulate Organic Matter
Carbohydrates Proteins Lipids
Hydrolysis
Amino Acids, Sugars, Alcohols, Fatty Acids
Acidogenesis
Intermediary Products
Acetate, Propionate, Ethanol, Lactate
Acetogenesis
Homoacetogenic oxidation
Acetate
Hydrogen CO₂
Reductive Homoacetogenesis
Methanogenesis
CH₄ + CO₂

Figure1.1. a. Flow chart all the biochemical pathways in AD including methanogenesis

1.1.1 Acetoclastic Pathway:

Acetate is the most important predecessor in methane formation in acetoclastic pathways. This type of methanogenesis is responsible for about two thirds of the total energy production (methane generation) (Shah, et al., 2014). The methanogens belonging to acetoclastic pathway genus acts upon simple carbonic acid that is acetic acid to synthesize acetate which is later catalysed to form methane and carbon dioxide (Nishio & Nakashimada, 2013).This main biological pathway can be chemically represented as:

$$\boxed{CH_3COOH \longrightarrow CH_4 + H_2O}$$

There are only two main genera of acetoclastic methanogens *Methanosarcina* and *Methanothrix* belonging to the genus *Methanosarcinales. M. barkeri, M. mazei, M. acetivorans, M. baltica, M. frisia, M. horonobensis, M. lacustris, M. semesiae, M. siciliae, M.*

soligelidi, M. vacuolata, and M. thermophila are all Methanosarcina species which are of the order mesophilic to thermophilic. Mesophiles M. soehngenii, M. concilii, and M. harundinacea, as well as thermophiles *M. thermoacetophila* and *M. thermophila*, make up the *Methanothrix* genus. Both of them are at their peak performance under a PH around 5 (slightly acidic) and a temperature around 35-40^{o}C .

The genus Methanosarcina are of utmost importance as they are called generalist amongst all the methanogens as they have the ability to grow on not only acetate (carbon source) but also methylated amines, hydrogens, methanols while the genus *Methanothrix* is a specialist one as they can only grow acetate. The latter one is quite environment sensitive that is, it is sensitive to the salt concentration, temperature variation, excessive acetate, and ammonia so on. Meanwhile *Methanosarnia* group adapts to almost all conditions and grows unaffected (Stams, et al., 2019).

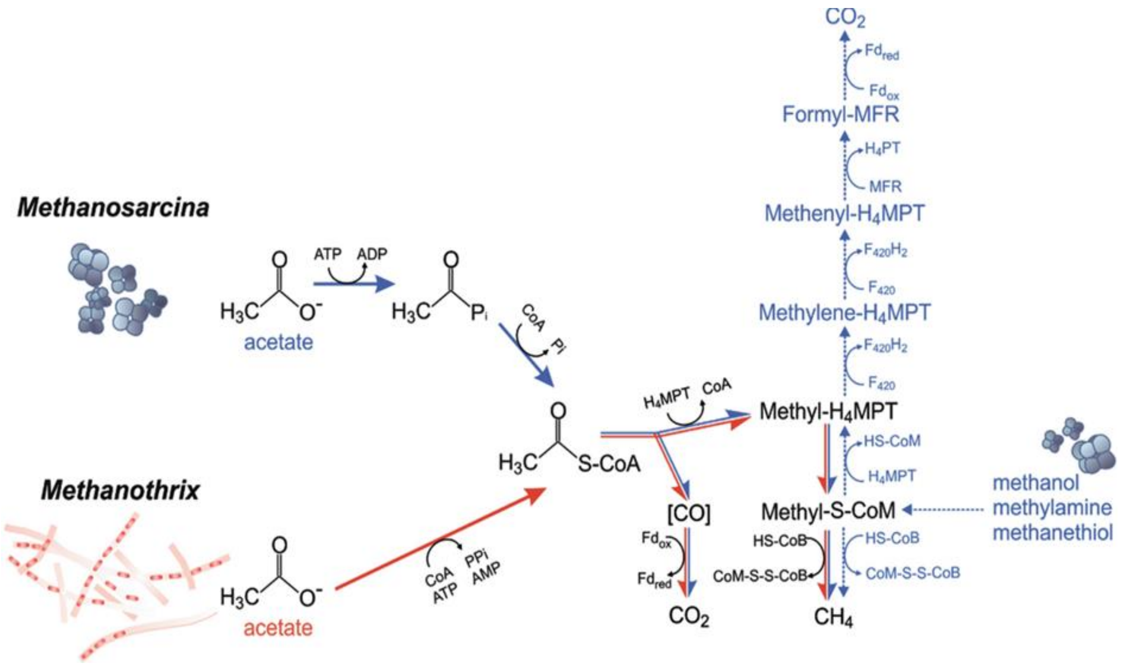


Figure 1.1.1 Acetoclastic pathway of Methanosarcina and Methanothrix

The figure 1.1.1 shows acetate translation pathways of the two genera with blue arrows for *Methanosarcina* while red arrows for *Methanothrix* where following abbreviations MFR stands for methanofuran, H4MPT for tetrahydromethanopterin, HS-CoM for coenzyme M, HS-CoB for coenzyme B, F420H2 for reduced form of electron-carrying coenzyme F420, Fd for ferredoxin, and CoA for coenzyme A. Except for the first step of acetyl CoA conversion all the rest of the metabolic pathway is the same for both the methanogenic groups.

1.1.2 Hydrogenotrophic Pathway:

It is most the shared and common pathway for energy production in methanogens. They grow on the hydrogen (electron acceptors) and CO_2 to synthesize methane, they also use formate as the electron donor to break down formic acid instead of hydrogen (Demirel & Scherer, 2008). Though they are responsible for methane production than acetoclastic group there are less literature study on them as they are not easy to culture in artificial conditions (Nishio & Nakashimada, 2013). Typically, it is believed that acetate is decomposed by acetoclastic methanogens, and the CO_2 produced is synthesized with the help of hydrogen to produce methane. The hydrogenotrophic pathway can be symbolized as follows (Nishio & Nakashimada, 2013):

$$\boxed{4H_2 + HCO_3^- + H^+ \longrightarrow CH_4 + 3H_2O}$$

But in conditions where the acetoclastic methanogens fails to translate acetate to the desired end product (due to the presence of excessive concentration of acetate or other stress conditions) hydrogenotrophic archaea comes in action. Syntrophic acetate oxidising bacteria convert some of the acetate to hydrogen and carbon dioxide which is then coupled by hydrogenotrophic ones for further process (Shah, et al., 2014).This syntrophic relationship between the hydrogen producing acetogenic bacteria and that of hydrogen consuming hydrogenotrophic methanogens where the hydrogen produced by one group is used as the hydrogen donor by the other group is called inter-hydrogen transfer (Nishio & Nakashimada, 2013). The class of archaea identified and classified in this group includes *Methanobacterium, Methanobrevibacter, Methanosprillum, Methanococcus, Methanogenium, and Methanoculleus.* The genera belonging to hydrogenotrophic methanogens are more towards thermophilic side, they also function well in alkaline PH easily up to 9 and even more in certain cases (Nishio & Nakashimada, 2013).

The figure shows the six-step breakdown of the substrate carbon dioxide obtained from formic acid. Here, the reduced ferredoxin serves as an electron donor while CO2 is reduced and activated to create formyl-methanofuran. In the second step, the formyl group is transferred to tetrahydromethanopterin (H4MTP), resulting in formyl-H4MTP. Using reduced F420 (F420H2) as an electron donor, the formyl group is dehydrated and reduced to methylene-H4MTP, followed by methyl-H4MTP. The methyl group is transferred to coenzyme

M after that (HS-CoM). Finally, methyl-CoM is reduced to methane utilising coenzyme B (HS-CoB) as an electron donor (Enzmann, et al., 2018).

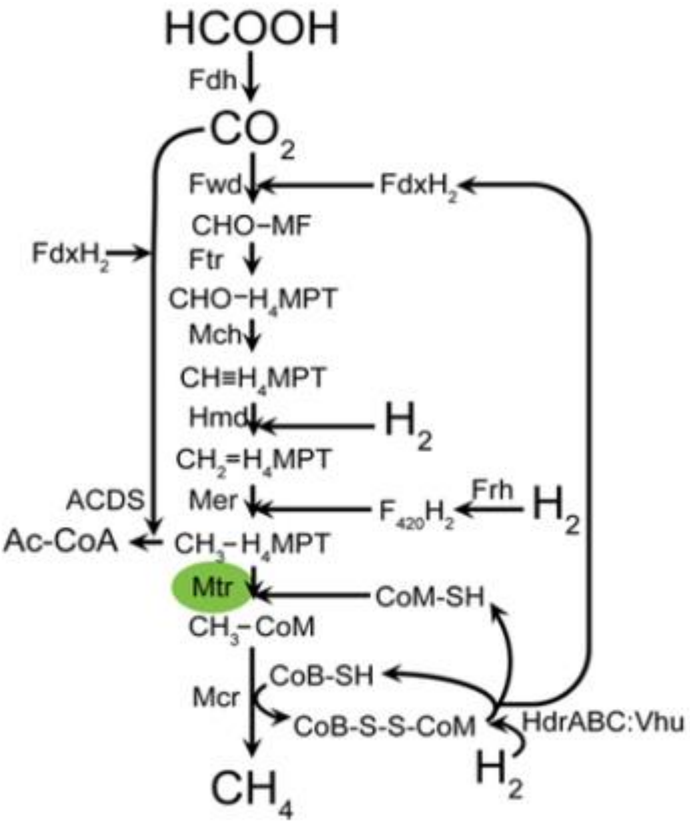


Figure 1.1.2 Steps showing the biochemical pathway for methanogenesis of Methanococcus (a hydrogen utilising methanogen) (Lieber, et al., 2014)

1.1.3 Methylotrophic Pathway:

Methylotrophic pathway has very little study done on them but is gaining popularity in the years. In this pathway the microbes grow on small, methylated compounds like methanol, methylated amines to produce methane (Kurth, et al., 2020). This type of methanogenesis is also responsible for a major share of energy production and grow syntrophically with the rest of the methanogens and disproportionate most of the organic waste present coupling with other methanogens as they take up the methylated compounds and grow on them, leaving the hydrogen, carbon dioxide and acetate to other methanogenic pathways. The pathway can be chemically represented as:

C_1COMPOUND (R-CH_3) + FORMALDEHYDE (O=CH_3) $\longrightarrow$ CO_2/ BIOMASS [13]

Methanosarcinales, Methanobacteriales, and *Methanomassiliicoccales* are the main three orders of methylotrophic methanogens. Depending on the presence of cytochromes (found in the electron transfer chain in mitochondria), methanotrophic methanogens are classified into two groups. Methylotrophs without cytochromes must rely on H2 while the ones with cytochromes (*Methanosarcinales* members) may oxidise methyl groups to CO2 using a membrane-bound electron transport chain (Mosin & Ignatov, 2014). These methanogens show high production of biomass by methanol conversion, having the ability to yield high protein product and essential amino acids. Owing to almost 37-65% of methanol transformation (Mosin & Ignatov, 2014).

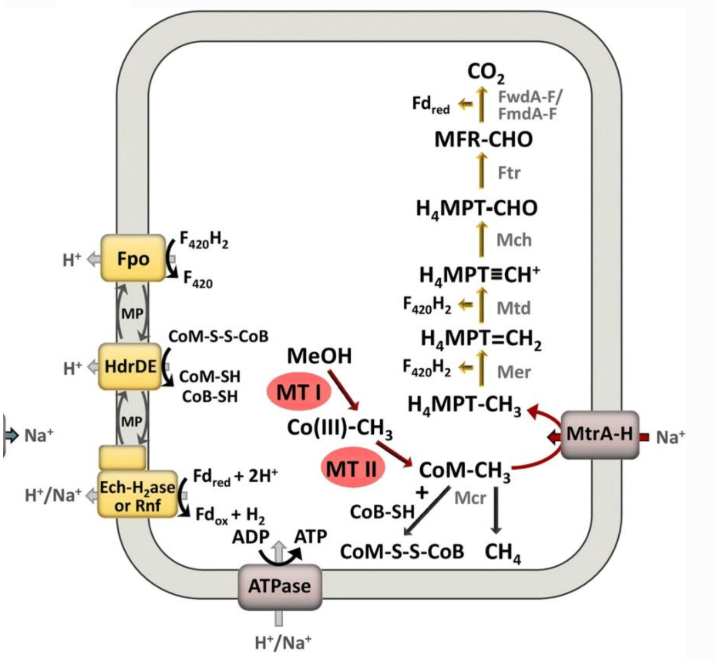


Figure 1.1.3 Pictorial representation of the chemical pathways in methylotrophic methanogens

These methanogens possess substrate targeted methyltransferase enzymes for the consumption of methanol and methylated amines as shown in the figure 1.1.3. This enzyme is constituted by three parts: a methyltransferase substrate specific group that converts methylated compounds into simpler proteins and a subsequent methyl transferase that channels methyl group into methanogenic pathways. Of the total methyl groups converted approximately three quarters are transferred to methane and the remaining one quarter is oxidised to CO_2 (Kurth, et al., 2020).

1.2 STRESS CONDITIONS AND THE RESPONSES OF METHANOGENS

The type of methanogens found in a reactor system, or any environment depends on the external environmental conditions as well as the capability of the methanogens to withstand these conditions. Methanogens some exhibit sensitivity towards surroundings while some are extremophiles found in exceptionally hostile conditions. The many stress factor that could be faced by the methanogens could even create engineering constraints on reactor, thus making it important to study stress genes, it's potential and functionality in methanogens. Some of the factors that causes stresses in methanogens include:

- **OXYGEN STRESS:**

For anaerobic digestion, oxygen is widely recognised to be the most harmful component. Although the presence of oxygen signals the end of methanogens in the medium, certain methanogenic organisms may survive a tiny amount of oxygen stress by using a specific enzyme (Borisov, et al., Jun 2021).

- **ACETATE AND AMMONIA STRESS:**

The acetate concentration in the substrate determines the pathway dominant in a methanogenic medium to a great extent. A rise in the accumulated acetate concentration up to 3000 mg COD L1 can be considered as the threshold for a shift in the methanogenesis pathway. Due to the presence of an excessive quantity of acetate content, acetoclastic methanogens fail to breakdown acetate after reaching a saturation limit, creating acetate stress. Acetate oxidising comes into play and oxidises acetate to useable form for hydrogenotrophic methanogens. However, by adapting, certain methanogens in the acetoclastic route show a greater tolerance to increasing acetate stress. Similarly, is the situation for ammonia stress certain hydrogenotrophic methanogen develop ability to break down ammonia even though the concentration is more (Shah, et al., 2014).

- **TEMPERATURE AND PH**

As far as temperature is concerned acetoclastic methanogens are usually mesophiles (ones that grow in moderate temperature), 25-40^{0}C being the optimum temperature while hydrogenotrophic ones are towards the thermophilic (exhibit high functionality at high temperature) side. So, the temperature of the reactor medium is an important factor that determine the pathway dominant in the reactor. The growth of various methanogens is also affected by PH stress, which is dictated by whether the medium is alkaline or acidic. Despite the fact that most anaerobic microbes prefer a pH of 6.8-7.2, there are species that thrive at the extremes (Shah, et al., 2014).

- **FORMATE AND HYDROGEN STRESS**

Mostly found in hydrogenotrophic methanogens where the excessive amount of formate or hydrogen electron inhibits the ability of hydrogenotrophic methanogens to produce energy. But there are stress genes present in certain methanogens that withstand the excessive amount of these substrate (Costa 1, et al., 2013).

- **SALT STRESS**

Salt concentration is also a physiological parameter, most of the methanogens under salty or halophilic conditions have to find a way survive the osmotic pressure variation and from drying. The cell withstands without drying out by two methods mainly: one by producing glutamate in the presence of lower salt concentration and by producing N-acetyl-βlysine or alanine along with glutamate at high concentration. The second way to withstand the salt concentration is by the influx on potassium and chloride ions into the cytoplasm of the cell so that the osmolarity of both the medium are maintained (Enzmann, et al., 2018).

1.3 BACKGROUND STUDY ON WHOLE GENOME ANALYSIS:

Over the past few decades, the techniques to manipulate the micro-organisms in these microbial communities have evolved into the latest techniques. The microbial communities comprise both prokaryotic and eukaryotic micro-organism in a single celled or multi-celled format including archaea, bacteria, fungi, and the rest. The earliest technique involved manually culturing the required species, identifying their DNA sequencing, and classifying them into phylogenetic trees based on different metabolic and other pathways followed by them. But this traditional method was time consuming due to the slow growth rate of the microbes. Moreover, these microbial colonies may not sustain in artificial or hostile environments, even if they manage to there may be the presence of foreign bodies cultured along with them. Hence more precise techniques were developed in the 20th century.

1.3.1 Sangar Sequencing:

This led to the development of first-generation sequencing called the sangar sequencing in 1977, the first ever method to sequence an entire genome (Merch, n.d.). Sangar sequencing can be done either manually or automated by reproducing the DNA strands using PCR method but by making use of dideoxy nucleotides(ddNTPs) along with the regular deoxynucleotides where the ddNTPs are coloured bases which act as labels as well as terminators to the sangar sequencing. All the four base pairs in ddNTPs are dyed with different colours for easy identification . The ddNTPs lack O in the 3’-OH end thus unable to form the phosphodiester bond between the 5’-phosphate of one nucleotide and 3’-OH of the next one. This results in multiple copies of the desired DNA strand in different lengths of base pairs. The new copies are separated from each other in the terms of their sizes by passing through a gel electrophoresis. An electric current is passed from one end of the gel matrix, the negatively charged copies of DNA will move towards the positively charged electrode, thus the smaller strands moving further towards the positive electrode in the gel followed by bigger ones. These strands are then read using laser reader like chromatogram to determine the sequence of the input DNA (Merch, n.d.).

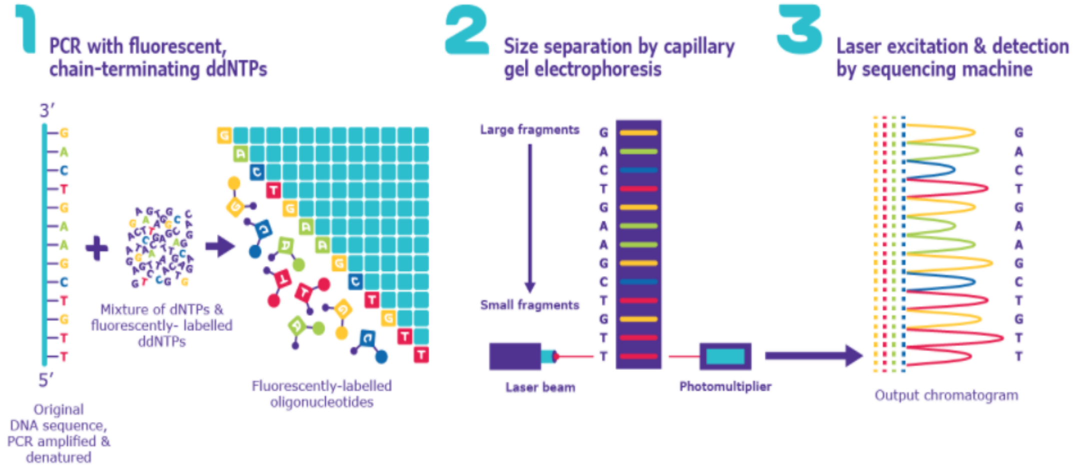

1 PCR with fluorescent, chain-terminating ddNTPs
2 Size separation by capillary gel electrophoresis
3 Laser excitation & detection by sequencing machine
3′
5′
Original DNA sequence, PCR amplified & denatured
Mixture of dNTPs & fluorescently- labelled ddNTPs
Fluorescently-labelled oligonucleotides
Large fragments
Small fragments
Laser beam
Photomultiplier
Output chromatogram
G A C T G A A G C T G T T

Figure 1.3.1 All the steps of sangar sequencing are shown in sequential order

Despite this breakthrough in genome sequencing, it had a major limitation that at times the detected waveforms are low which shows uncertainty of the right result. Moreover, sangar sequencing is not able to find the DNA combinations of species a greater number of genomes or many microbial species.

1.3.2 Shot-Gun Sequencing:

This defect paved the way for discovery of shot-gun sequencing method by ion torrent and 454 also collectively called as NGS (New Generation Sequencing). Shot gun metagenomic sequencing provides not only helps in the annotation of each gene sequence in organisms but also gives insights regarding the metabolic pathways, gene functionalities, microbial interaction between communities, genetic prediction, and so on. Another advantage of this is that it can provide all the genomic information of an organism without having to isolate the community of organisms and culture it (Novogene, n.d.). This is important because it is difficult to culture all organisms in an artificial environment.

In this method of shot-gun sequencing, the entire strand of DNA is broken down into small fragments by using chemical or physical processes like ultra-sonification, adapters are then attached to either end of these fragments with a unique identification barcode for differentiating the micro-organism from each sample. These small fragments then get attached to the surface of ISP enriched beads as the beads are provided with complementary adapters, ideally, it is considered that one DNA fragment attaches to one bead. This is accomplished by placing both the beads and fragments of DNA in oil emulsion so that they are localised as one bead-DNA pair in each globule. When the DNA strands connect to the DNA on the beads, the double-strand bond breaks resulting in a single-stranded DNA. The DNA fragments are then replicated several times on each bead using the polymerase chain reaction technique (PCR). Thus, creating millions of copies. The beads are then trapped in the well-like structures of the sequencer (mostly 454) where nucleotides are added to these wells in the form of one base type at a time like all A's, G's, C's, and finally the T's. As the base gets attached to each complementary base in the DNA strand light is given out which is recorded by a camera which helps to decipher how many bases of each type are there in a strand thus giving the DNA sequence (yourgenome, n.d.).

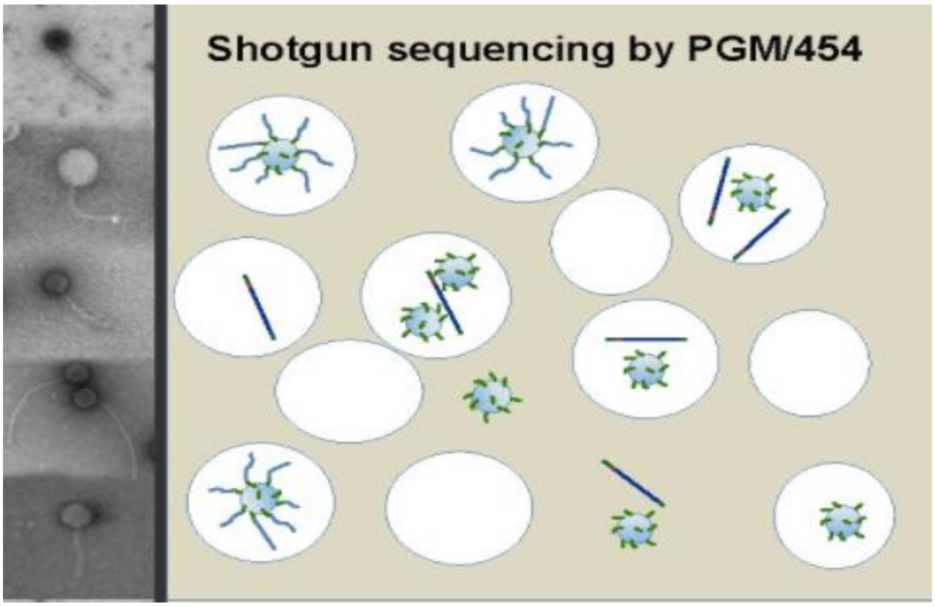


Fig 1.3.2. Single stranded DNA strands trapped in emulsion globules

The major limitation for this method adding to its high cost is that as shown in the figure 1.3.2 all emulsion bubbles cannot trap one bead-DNA pair perfectly some traps more than one bead or DNA fragments while some globules remain empty.

1.3.3 Ion Torrent Sequencing:

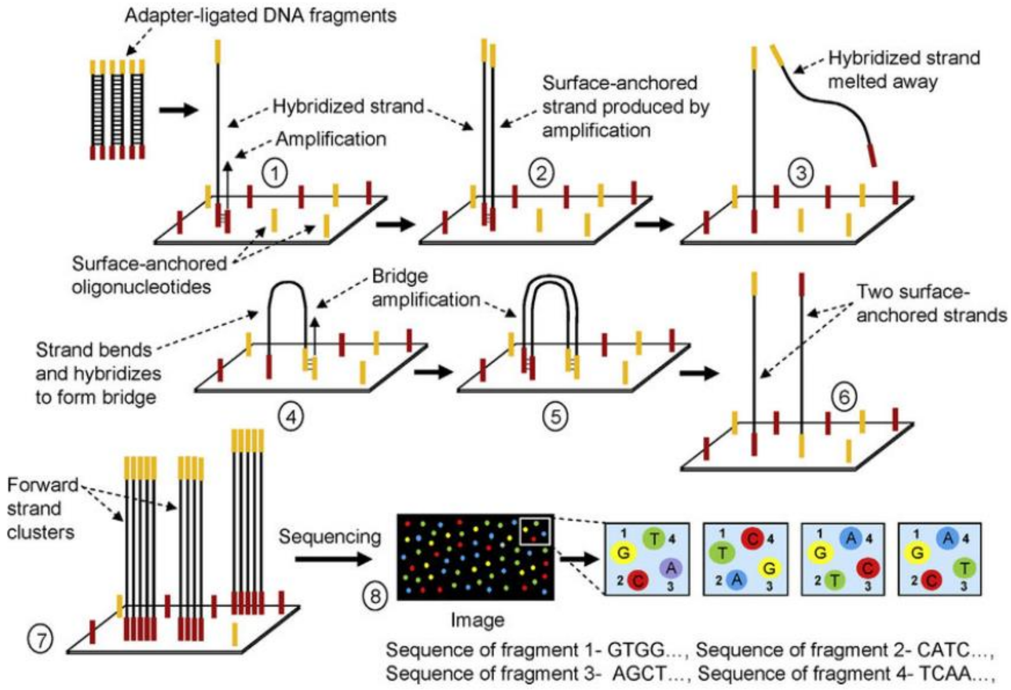
Ion torrent sequencing is similar as 454 but unlike the light signals generated in 454 sequencing Ion torrent makes use of the change in the PH in the solution due to the release of hydrogen ion (a by-product) produced when a nucleotide is introduced into a DNA strand by a polymerase which can be measured by using a proprietary ion sensor. This chemical information is then translated into digital information as 0 and 1. Major advantage of this method is the cost effectiveness in completing the sequencing (EMBL-EBI, n.d.).

1.3.4 Illumina Sequencing

Illumina sequencing entered the picture in late 2000 as an amendment to Ion torrent as well as 454, as both could not properly detect homopolymer region. Illumina sequencing makes use of a method called bridge amplification. In bridge PCR, the adapter connected to the Flow cell's surface serves as a template for bridge amplification. The flow cell has two types of oligos fixed to it, the P5 and P7 ones, which get attached to the complementary adapters (either adapter 1 or adapter 2) on the fragment. Once this happens, by polymerisation a reverse strand is created and then the original strand in the double-stranded DNA is washed away. The long single strand the bends over to hybridise with adjacent complementary oligos leading to next polymerisation resulting in a double stranded bridge. The double strands are then denatured into two single long strands which then hybridises with two different complementary oligos forming two bridge strands and the cycle recurs as seen in the fig 1.3.4.

After numerous rounds of amplification and denaturation, each DNA fragment will be condensed into a bundle at its own location, with each bundle containing multiple copies of a single DNA template. The resulting library should next be placed in a sequencer to be sequenced. DNA fragments can be put together using a specific area on the sequencing machine. The sequencer can identify all additional DNA sequence information at the same time since each fragment has its own connecting space. Each sequence is read twice giving a clear picture of the genome (Bioquest, n.d.).

The information sequenced using these sequencing methods are then stored in a public directory called NCBI (National centre for biotechnology Information) where the database of almost all micro-organisms is stored in the library in the form of GenBank files. This information can also be assessed in the form of FASTA or FASTQ files depending on the nature of work. NCBI library contains information ranging from microbiology, molecular biology, biochemistry, genetics so on. The detail of genomes, proteins, different areas of study can be found in this library so that there is no need to culture an organism in an artificial environment under additional requirements.


Adapter-ligated DNA fragments
Hybridized strand
Amplification
1
Surface-anchored strand produced by amplification
2
Hybridized strand melted away
3
Surface-anchored oligonucleotides
Bridge amplification
Strand bends and hybridizes to form bridge
4
5
Two surface-anchored strands
6
Forward strand clusters
7
Sequencing
8
Image
Sequence of fragment 1- GTGG..., Sequence of fragment 2- CATC...,
Sequence of fragment 3- AGCT..., Sequence of fragment 4- TCAA...,

3 Principles of Illumina Solexa sequencing. The DNA-sequencing library is prepared by ligating adapters to the end-polished

Fig 1.3.4. detailed figure showing the steps in illumina sequencing (Bridge amplification)

1.4 WHOLE GENOME ANALYSIS

The whole-genome sequencing is a critical tool in sequencing the entire genome of eukaryotic and prokaryotic organisms to study the species in general or detail as per the user's requirement as well as helps in mapping the different genomes in a species to each other so as to find the differences and similarities between them. This method cannot be overlooked as it provides a more diversified outlook which is extremely important in the field of research (Novogene, n.d.).

Genome refers to the total genetic material in an organism that provides all the data regarding the prominent genes and pathways in an organism. The different set of genomes in an organism is got as a result of experimenting and extracting samples of the same species from the different natural medium. Another whole set of genomes can also be generated by culturing the species of interest under an artificial environment to study specific characteristics or mutations in them.

The whole genome analysis has the ability to analyse all the given strains of genomes in a given species throughout its entire length. This adds multiple advantages to this method like superior quality data for scientific and professional research. In addition to this, since it annotates the entire genetic material of the multiple genomes belonging to the same species it helps in identifying the mutations, adaptations as well as stress responses of the different genomes of that particular genus. This method cut down the expense of genome sequencing to a large extend. It was also able to produce a large amount of data in a shorter time span in a more efficient manner.

The complete gene sequence of a specific specimen is mainly pre-recorded and preserved in a public database. The approach of whole genome analysis is used in this study in a variety of ways, and interrogations are conducted to gain a deeper understanding of methanogens. This technique is gaining popularity in recent years as a result of its various advantages.

1.5 OBJECTIVE

This project focuses on the whole genome analysis performed on the prominent species of acetoclastic and hydrogenotrophic methanogens belonging to the group of anaerobic archaea which are crucial in wastewater treatment. They are studied in detail with properly organised tools in a user-friendly operating system for the presence and absence of relevant metabolic pathways along with the response of their respective accessory genes to external stresses. Methanogens and their pathways are gaining popularity nowadays as they are hot topics of study not only in wastewater treatment but also in renewable cleaner energy production as they have the potential to give out biomethane. The objectives of the project are broken down as follows:

- To successfully conduct the whole-genome analysis of important species in acetoclastic and hydrogenotrophic pathways using Orion cluster.
- To examine the capability within the chosen methanogen genomes (like *Methanosarcina barkeri*) for the isolates to withstand stress like ammonia and sulphate stress resistance as well as their resistance to varying temperatures and PH.
- To find the important methanogenic as well as active stress response genes so that the growth rate of the methanogen can be manipulated to improve wastewater treatment and harvest renewable energy.

2.METHODOLODY: THE WHOLE-GENOME ANALYSIS WORK-FLOW

This workflow is done in the Orion cluster (http://userweb.eng.gla.ac.uk/umer.ijaz/#orion) which is a bioinformatics cluster that has been developed and managed by Dr.Umer Ijaz where different software and workflows can be done genomes even with high number of data sets. This cluster has a number of servers of which howe.gla.ac.uk and becker.gla.ac.uk have been utilised in this project. These servers can be accessed remotely by MobaXterm, it is basically a recommended application which acts as a virtual interface that can access a remote server like howe in this case and work in it by logging into it using SSH client for example "ssh studentprojects@becker.eng.gla.ac.uk". The operating system used to work through MobaXterm in this project is the Linux.

Using the user-friendly environment mentioned above, the workflow for whole-genome analysis is carried out in the following basic steps as shown below:

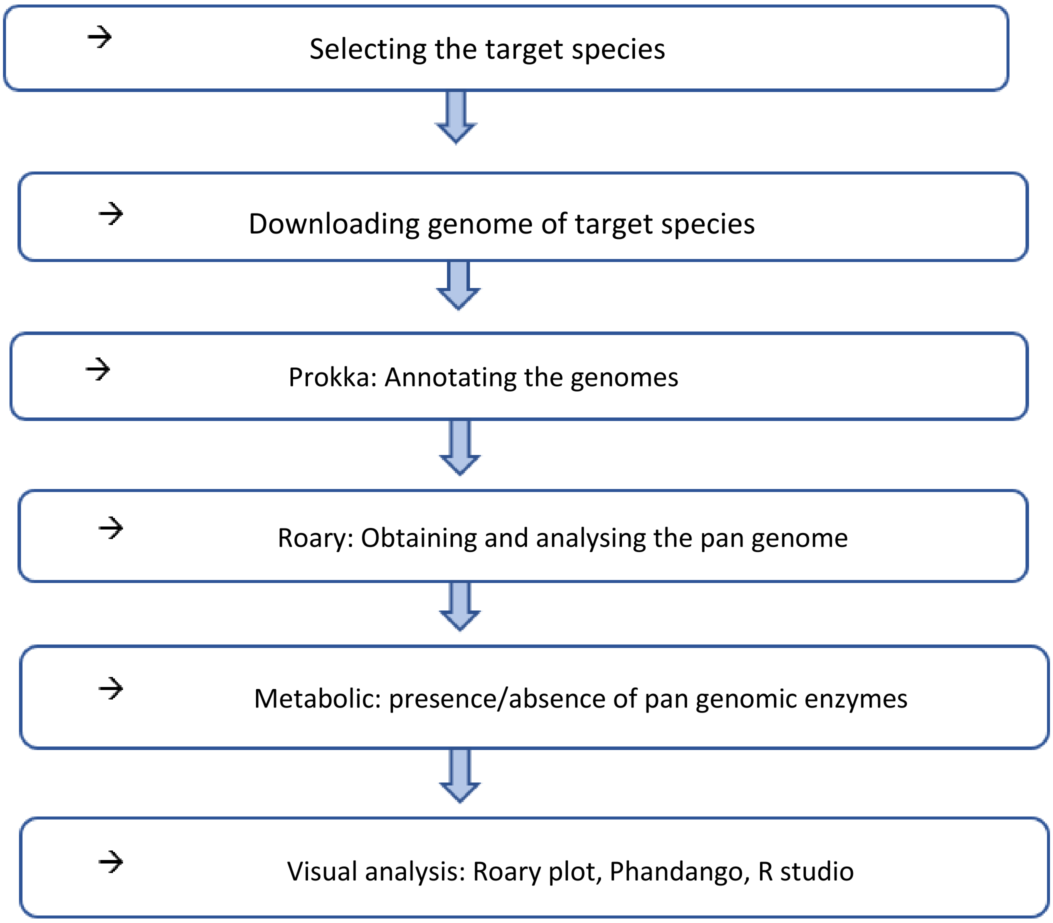

Selecting the target species
Downloading genome of target species
Prokka: Annotating the genomes
Roary: Obtaining and analysing the pan genome
Metabolic: presence/absence of pan genomic enzymes
Visual analysis: Roary plot, Phandango, R studio

Figure 2. Workflow for whole-genome analysis

2.1 Selecting the Target Species

The first step and foremost step in whole genome analysis was to decide the species on which the study should be conducted on. My focus of study being on the methanogens, was narrowed down to the important methanogenic species take parting in both acetoclastic as well as hydrogenotrophic pathways, which are the dominant pathways in anaerobic digestion. Based on some of the peculiarities and features of certain acetoclastic and hydrogenotrophic species, the species were finalised. The species selected are as follows:

2.1.1 Acetoclastic Methanogenic Species

Amongst the acetate decarboxylating methanogens two species were selected for study and comparison in this project to find out the which one amongst these will be better suitable for wastewater treatment in a reactor or for other organic waste decomposing purposes. For this study the following species were targeted for the following reasons:

- ***Methanosarcina barkeri:***

  These methanogens belonging to the genus *Methanosarcina* are one of the second largest in genome length in archaea. They are found in different environment like freshwater medium, marine medium and also in rumens of cow. These methanogens are mesophiles seen in an optimum temperature of 35-40^{0}C and a pH of 6-7 (Demirel & Scherer, 2008). *Methanosarcina barkeri* is an extremely relevant species as these methanogens are called the generalist amongst the methanogens. Though it an acetate oxidising methanogen it is known for its ability to act on other substrates like formate, hydrogen, ammonia, and sulphide (Shah, et al., 2014). They are also seen in places which require the certain genes to exhibit properties of halophilic or thermophilic depending on the environment these methanogens thrive in. The enzymes/genes responsible for these special characteristics are studied by performing whole-genome analysis on them.

- ***Methanothrix soehngenii:***

  Another prominent genus in the group is *Methanothrix* formerly known as *Methanosaeta* isolated usually from the sewers in mesophilic conditions. The single cells belonging to the species *Methanothrix soehngenii* are rod-shaped, grouped together in long filaments wrapped in a sheath-like structure. These acetoclastic methanogens unlike *Methanosarcina barkeri* which is a generalist only can grow on acetate medium. They cannot act on any other substrates like hydrogen, formate, methanol, so on. They have higher affinity towards acetate compounds and gets inhibited when the acetate saturation in the substrate increases from a particular level. The ideal temperature for growth and methane production is 37°C, with a pH range of 7.4–7.8. they stop function in the presence of oxygen though they remain viable (Huser, et al., 1982).

2.1.2 Hydrogenotrophic Methanogenic Species

Certain formate decomposing methanogens were also selected for study and comparison. The features of these methanogens based on the occurrence of certain metabolic pathways are closely observed to derive their importance as well as stress tolerances in a reactor. The species of this category selected were:

- ***Methanococcus maripaludis:***

  Methanococcus maripaludis has been recently gaining attention for its ability to fix nitrogen compounds. This micro-organism is cocci shaped found in a harsh environment like sea marshes which is a clear indicative that these are halophiles. Their group varies from mesophilic to thermophilic where *Methanococcus maripaludis* is a mesophilic methanogen with an optimum temperature of 35-40^{o}C and a pH 7-9, well there are exceptions where the strains of maripaludis have shown extreme characters of thermophilic and alkalinity resistant archaea (Goyal, et al., 2016). This methanogen uses hydrogen ion as electron acceptor to break down the CO_2 used as substrate. These are synthesised from formate which is obtained from acetic acid produced by acetyl-Co enzyme. But these species are said to produce methane by reducing CO_2 even in the absence of acetate (Ladapo & Whitman, 1990). This species of methanogens is checked here for formate stress as well.

- ***Methanobacterium formicicum***:

  Methanobacterium formicicum is a filamentous methanogen which uses formate or hydrogen and CO_2 as the substrate to produce methane. The strains were found to have good tolerance to abiotic stresses of the surroundings like osmolarity, high temperature stress tolerance (Maus, et al., 2014). They are found to grow in standard conditions like 35-56°C and 6.5-8 maximum pH. Besides this the formate dehydrogenase active in both the selected species *Methanococcus maripaludis* and *Methanobacterium formicicum* are different and checked in this project. Moreover, the enzymes responsible for carbon fixation could inhibit under formate limitation as well as sometimes in case of hydrogen starvation, which means stress responses could be poor in this case which is also later analysed (Costa 1, et al., 2013).

2.2 Downloading The Genome of Target Species

Once the species were finalised, their genomes have to be downloaded from NCBI (National centre for biotechnology Information) which is a public directory which contains all the genetic information of a genome in the form of GenBank files. GenBank file (Gbk format) contains all the publicly available DNA sequence information from gene to base pairs belonging to a particular organism (NCBI, 2021). The information stored can be retrieved any format depending on their usage using a ftp client (ftp://ftp.ncbi.nlm.nih.gov/genbank.) like:

• fna	- used to specify the nucleic acid
• faa	- Fasta amino acid, contains sequences for amino acid
• fasta	- generic fasta file with stored info of sequence fragments after mapping. Other extensions include fna, faa, fa, frn.
• gff	- a tab limited file containing all the genetic sequences.
• fastq	– file that store sequence fragments before mapping

Table 1. Examples of different types of files in Genbank

Here the genomes were downloaded in the form of fna files by using wget ftp command line which can be found in appendix. Thus, genomes of all the four species are downloaded in separate screens in Linux.

➢ *Methanosarcina barkeri*	- has 7 genomes
➢ *Methanothrix soehngenii*	- has 2 genomes
➢ *Methanococcus maripaludis*	- has 24 genomes
➢ *Methanobacterium formicicum*	- has 6 genomes

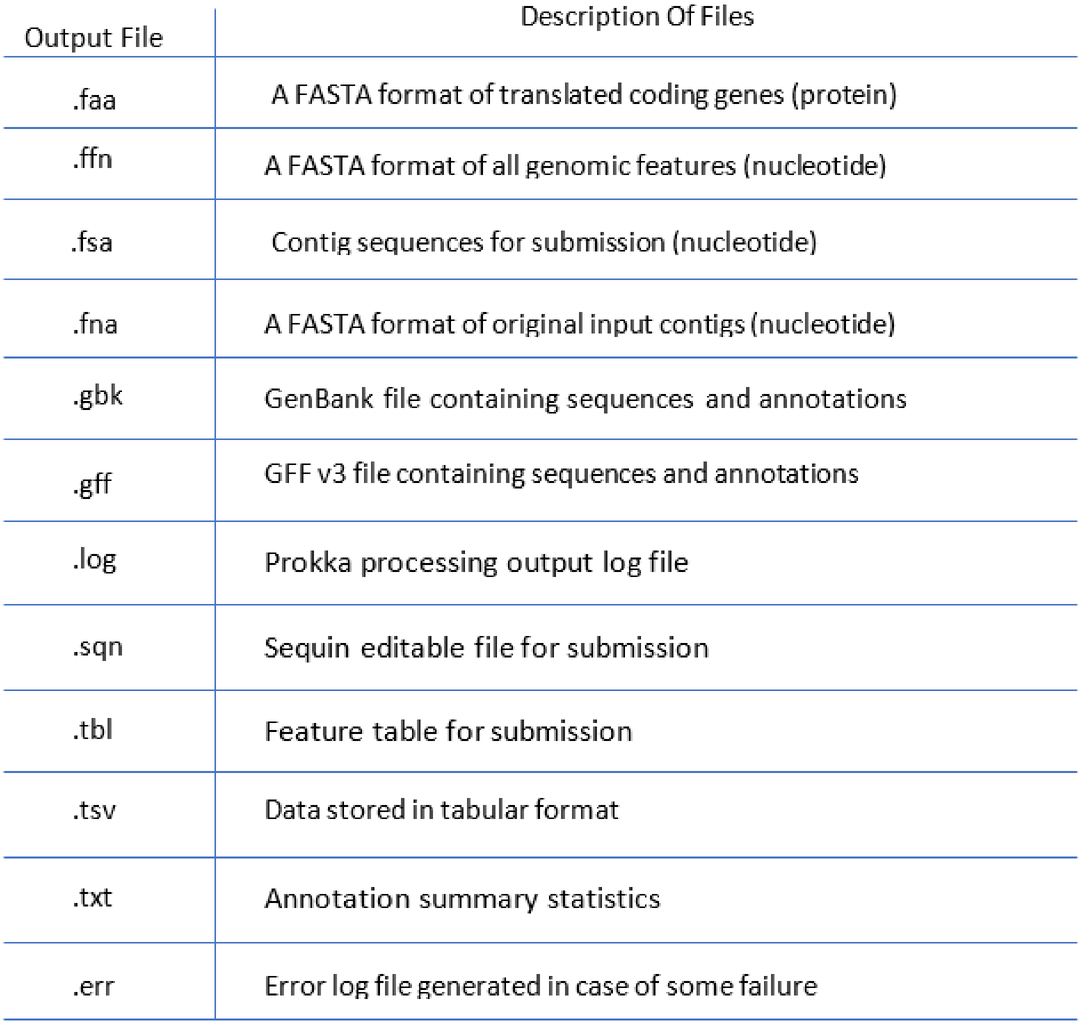
These genomes in fna format are used for further analysis in this project.

2.3 Prokka: Annotating the Genomes

Next step in the workflow was annotating the genome sequence using a one-line command tool called Pokka in a very small amount of time. In simple words, it identifies and labels all the relevant features in a genome sequence (Seemann , 15 July 2014). To obtain a comprehensive and reliable annotation of genomic archaeal/bacterial sequences, Prokka integrates a number of external prediction tools to obtain high accuracy output like Prodigal, RNammer, Aragorn and so on. It will take advantage of as many processor cores as possible, and a typical bacterial genome may be annotated in approximately 10 minutes on. It is suitable for integrating pipelines into genomic applications and iterative sequence analysis (Seemann , 15 July 2014).

Here, Prokka had been already pre-installed in the nodes of Orion cluster which was made to run using mobaXterm using a single command line. The downloaded genomes of all the methanogenic species (*Methanosarcina barkeri, Methanothrix soehngenii*, *Methanococcus maripaludis, Methanobacterium formicicum*) were run in Prokka separately. The one-line command used to run prokka in Linux terminal is a for loop where all the downloaded genomes of one species (for example the 7 genomes of *Methanosarcina barkeri*) from the previous step are inputted in fna file format. For annotating the genomes, a two-step process is adopted by prokka where, the coordinates of candidate genes are determined using a tool Prodigal. Prokka then compares it to a huge database of known sequences, generally at the protein sequence level, and transfers annotation of the best significant match.

The following result files were obtained after annotation is done on all the fna input genome files.



Output File	Description Of Files
.faa	A FASTA format of translated coding genes (protein)
.ffn	A FASTA format of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.fna	A FASTA format of original input contigs (nucleotide)
.gbk	GenBank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Prokka processing output log file
.sqn	Sequin editable file for submission
.tbl	Feature table for submission
.tsv	Data stored in tabular format
.txt	Annotation summary statistics
.err	Error log file generated in case of some failure

Table. 2 Table showing the description of output files in prokka (Seemann , 15 July 2014)

2.4 Roary: Obtaining and Analysing the Pan Genome

Roary is a pipeline which takes calculates the pan genome of the given annotated data from Prokka. The input file accepted by this pipeline is in the GFF format. It cannot be used for comparing genes like in metagenome instead it gives data about all the set of genes available in the genome set of a species. It is the only criteria of roary tool, these GFF files must have nucleotide sequence provided at the end of the file. The most efficient way to generate GFF files with nucleotide sequence is by generating it through Prokka (Page, 2018).

The workflow used here follows the same path. One of the output files obtained from Prokka in the previous step is GFF files with nucleotide sequences. Once the Prokka is ran, it generates one output .gff files for every input genome. fna files. That is, for all the 7 genomes in *Methanosarcina barkeri*, 7 GFF files will be generated after Prokka. These GFF files generated were then moved to a folder created called Roary after activating the Conda environment, the Roary command line (given in appendix 1) was ran in the folder to generate another set of output files which are of prime importance in the further workflow.

The output files generated in this pipeline is shown in appendix 2.a, of all the files gene_presence_absence.csv file and fastree.newick file are the two important output files. The gene_presence_absence.csv file gives an excel sheet with the presence and absence spreadsheet where the names, function and occurrences are shown. While the newick tree file is a phylogeny file showing the relationship between the genomes of the species based on shared ancestors and point of divergence

The two output files discussed above are used for further visual analysis in software tools like roary plots and Phandango. Roary does not produce a newick tree file output file directly; instead, the user uses a fastree command to create it from the core_gene_alignment.aln file (another roary output file) (Page, 2018).

2.5 Metabolic: Presence/Absence of Pan Genomic Enzymes

Metabolic is an abbreviation of METabolic And BiogeOchemistry anaLyses In miCrobes. This tool helps in gives the entire metabolic and biochemical functions of the given genome isolates that has been already downloaded from the NCBI directory. Apart from the metabolic profiles metabolic is also designed to output the biochemical cycling diagrams (Madison, 2021). It annotates the genes and integrates all the metabolic information provided by the inputted genome data in an easy readable format.

There are two ways in which the metabolic results are expressed. One in the METABOLIC-G and METABOLIC-C format. Both of them generates almost the same output the only difference is that METABOLIC-C accepts metagenomic read data as input files too. The type of implementation used in this workflow is METABOLIC-G (Madison, 2021).

In order to do metabolic the .fna genomic files should be converted to .fasta files and moved to a new directory called Metabolic (created by the user). The software is worked out in metabolic environment, so to proceed further the user disables the pangenome path and sets metabolic as its current path. Then metabolic tool is run using simple command lines.

The software processes several output files like METABOLIC_result_xlsx, METOBOLIC_FIGURES so on as shown in appendix 2 b. These result files have all the annotations with presence and absence of metabolic pathways along with nutrient cycling diagrams. The Metabolic spreadsheet was used further to identify the presence of important genes having the potential to partake in methanogenesis and stress responses.

2.6 Visual analysis: Roary plot, Phandango, R studio

Visualisation analysis is done by using a number of tools to get a better understanding of the pan genome data. The following visualisation tools are utilised to get results:

2.6.1 Roary Plots:

Three illustrative plots are obtained by running roary plot:

- **A phylogenic tree matrix**: It gives a tree classifying the set of genomes from the same species based on ancestral lineage. It shoes the point and generation of divergence based on which the genome isolates most similar and distinct ones can be found out easily. The matrix shows area of clustered genes that is areas where genes are present and gene devoid areas in the genome isolates. The matrix form also helps compare each genome within the species (Page, 2018).
- **A pie chart:** The pie chart depicts the total number of genes in the genome sets, as well as their general gene composition, which comprises core genes, soft genes, could genes, and shell genes, and the total number of genome strains in which they may be found.
- **A frequency chart:** This is a bar chart showing the frequency of gene present in each genome.

The procedure to get roary plot is simple; roary plot.py python file is downloaded from https://raw.githubusercontent.com/sangerpathogens/Roary/master/contrib/roary_plots/roary_plots.py using a wget command (Page, 2018). Then a newick tree file is created from a core_gene_alignment.aln result file in roary, this newick file along with the gene_presence_absemce.csv file is run in a single lime command to get the three visualisation plots from roary_plot.py.

2.6.2 Phandango:

It is an interactive software available in the web browser where the client can drag and drop his data to play with the results and get the desirable output. Here also the output is produced in the form of visual representation. Phandango allows the user to zoom in into genomic regions and clusters closely for analysing the visual data. Moreover, by uploading more than one set of genome isolates it helps to create a comparison study within them in the same environment. Phandango expects simple text formatted input files like CSV, GFF files (Hadfield, et al., 2018).

This project made use of this interactive software for visualisation of the roary results. The gene_presence_absence.csv file and the newick tree file from roary output of each species separately are dragged and dropped into the user interface website of Phandango. This software in turn gives an accurate phylogeny along with the pan genomic profiles of the genome strains uploaded in a species, creating an easy comparison. A line graph is also produced beneath the phylogeny- matrix plot indicating the total number of isolates in the given species carrying a certain gene.

2.6.3 Visual plots by R Studio:

R studio is an integrated platform developed to analyse date using R programming language. It gives a visual representation in the form a co-ordinate graph showing the Jaccard distance (distance that shows the similarities and dissimilarities of the consequent genomes i.e., how close or different they are to each other) and clustering of isolates. This is done by making a gene-genome presence absence CSV format file from the METABOLIC_RESULT.xlsx file and replacing the present and absent with 1 and o respectively. It is done in using a R studio by installing packages like vegan, ggplots2, and ggrepel, later choosing the directory where the CSV file of species are present and running it in the R script. Since this report is study of four species, two belonging to the acetoclastic pathways and the remaining belonging to hydrogenotrophic pathway, the Jaccard distance of the species *Methanosarcina barkeri and Methanothrix soehngenii* are compared using one CSV file and the other two *Methanococcus maripaludis and Methanobacterium formicicum* are compared and analysed using one CSV file in R studio.

All these visualisation aims to provide a clear understanding of the strains of all species as well to compare the differences of the species that have common ancestral background by getting a clear understanding of genes that are unique and core or shared within the species that share a common pathway of energy production

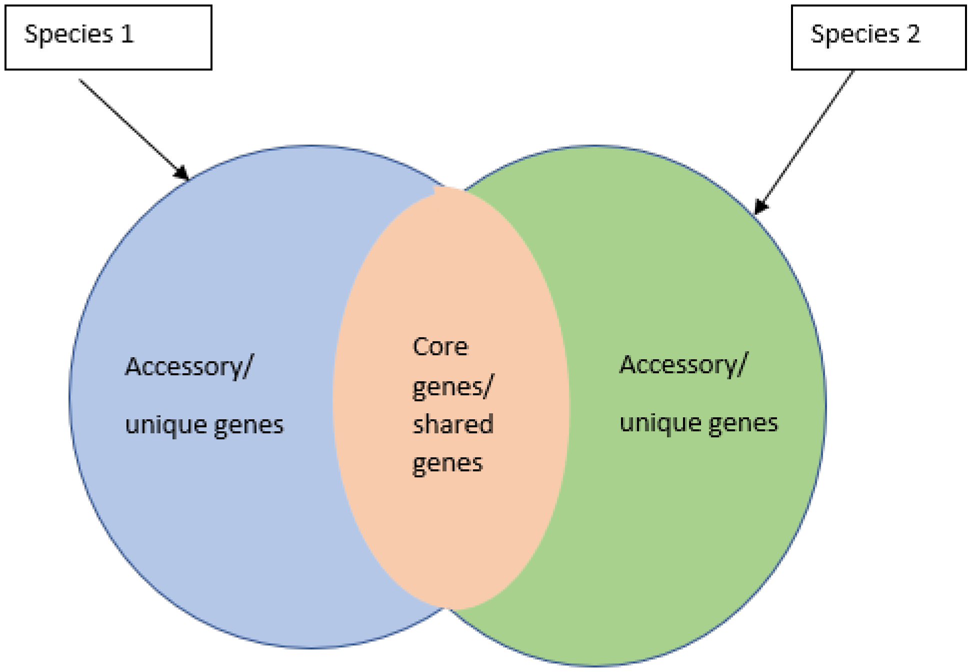


Figure 2.6.3 showing the summary of expected findings through this analysis

Similarly, it is expected for species 3 and 4 as well, here species 1- *Methanosarcina barkeri,* species 2- *Methanothrix soehngenii,* species 3- *Methanococcus maripaludis,* species 4- *Methanobacterium formicicum.*

The script for this workflow is shown in appendix 1.

3. RESULT AND DISCUSSION

The visualisation results obtained from all the software tools are analysed closely to derive close observation of all the species. The metabolic files are then filtered manually to find out the potential genes capable of doing the core functions as well as the peculiar genes to find the unique characteristics of all the methanogenic species like their stress responses, hence finding which species have an upper hand over the other isolates and thus manipulating the microbe growth in AD reactors accordingly.

3.1 INVESTIGATION ON *METHANOSARCINA BARKERI* STRAINS

3.1.1 Result from Roary plots:

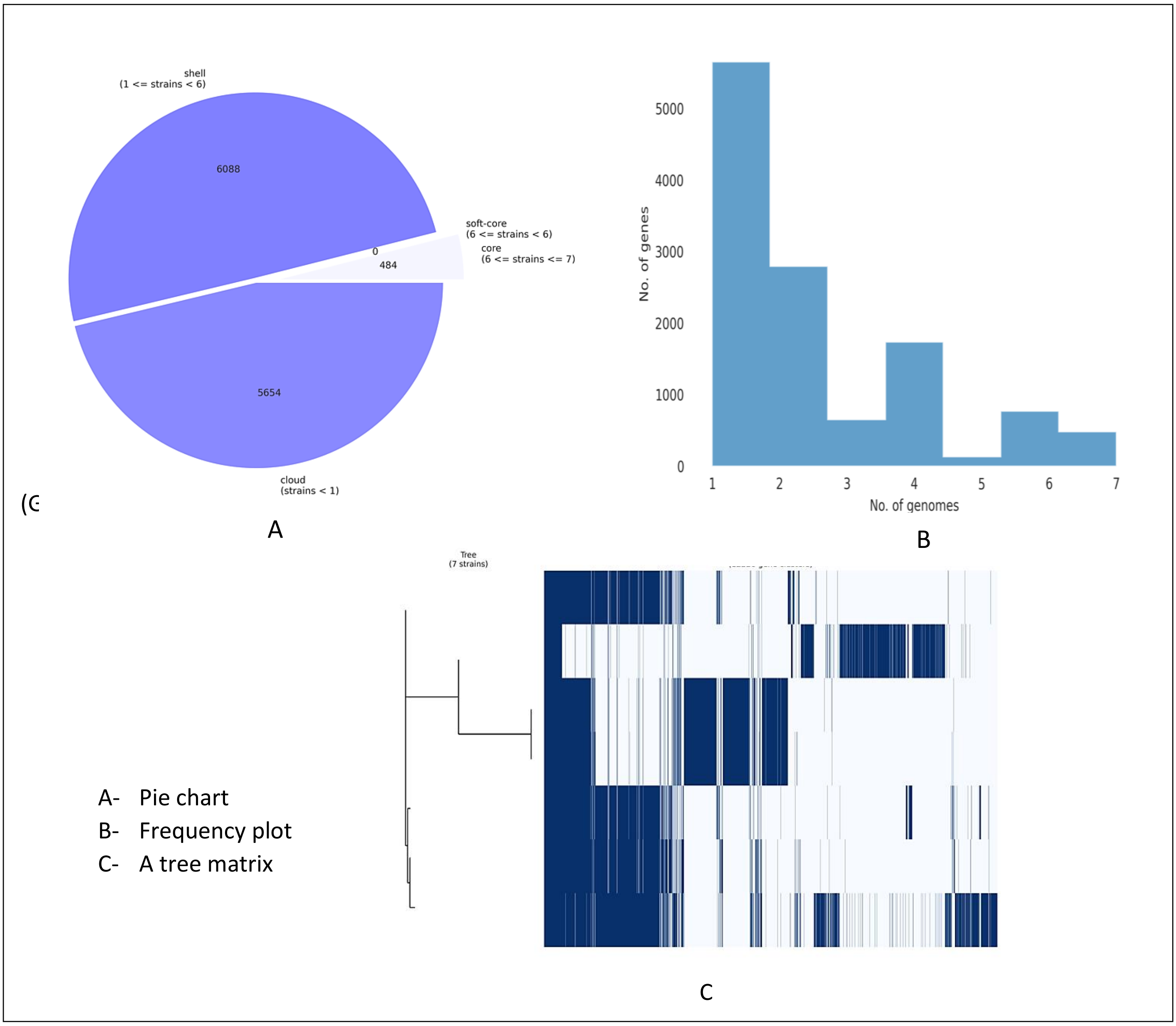


Figure 3.1.1 Visual plots from Roary plots of *M.barkeri*

- The pangenome pie chart as shown (figure 3.1.1 A) gives the total genomic composition of *M.barkeri.* This species has a total of 7 genome isolates. According to the pie chart pangenome of total 12226 genes present there are about 484 core genes and zero soft genes, the core genes is present in around all the 6-7 strains. Similarly, there are 6088 shell genes present in 15% to 95% of the species while cloud genes constitute 5654 genes of the total genes and are just found in 0-15% of the strains, given in appendix 4, table 1.
- The graph (figure 3.1.1.B) shows the Frequency of genes in each genomic strain of *Methanosarcina barkeri*, that the software plot as it runs through the length of each genome strain. It reads the number of genes present in the respective genome isolates thus giving the number of gene frequency in each genome. In *Methanosarcina barkeri* the frequency graph shows a zigzag trend with maximum gene reads of over 5000 in 1st genome to the lowest value around 200 in genome 5 and increasing again for the next two genomes.
- The next result plot is a phylogenic tree plot showing the similarities and dissimilarities of the 7 genome isolates. Genome strains of the same species become slightly different from each other due the presence of certain genes in some genomes while the absence of those other strains, the reason for this is divergence and evolution from ancestral strain which will be further explained in the Phandango plot as they illustrate the same.

3.1.2 Phandango Plot:

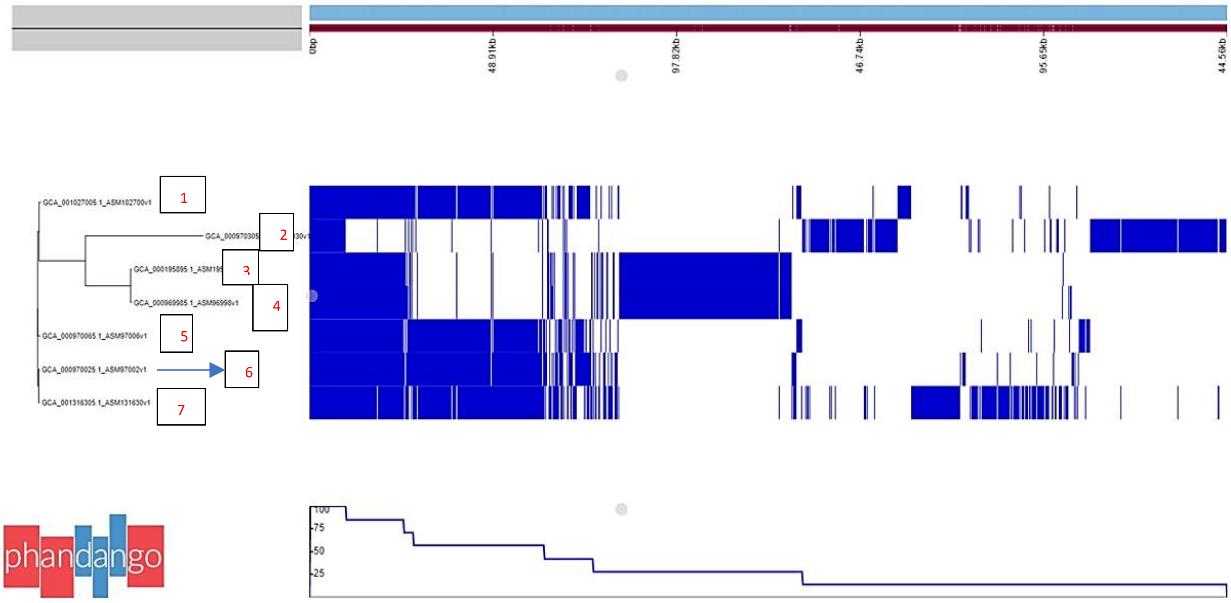

0bp
48.91kb
97.82kb
46.74kb
95.65kb
44.56kb
GCA_001027005.1_ASM102700v1
1
GCA_000970305
2
30v1
GCA_000195895.1_ASM195
3
GCA_000969985.1_ASM96998v1
4
GCA_000970065.1_ASM97006v1
5
GCA_000970025.1_ASM97002v1
6
GCA_001316305.1_ASM131630v1
7
phandango
100
75
50
25

Figure 3.1.2 plot from Phandango (Hadfield, et al., 2018), genomes are numbered for clarity

Phandango the plot (figure 3.1.2) shows the hierarchical evolution and how similar or dissimilar each strain is from each other. The phylogeny tree shows that genome isolate 3 and 4 of *M.barkeri* resembles each other as they have a shared point of divergence in the tree and have the same of presence and absence of gene base pairs in the gene matrix throughout the length of both the strains. Similarly, genomes 1,6 and 7 are similar but different from genomes 3,4 as well genome 2. Genome 2 shows slightly different characters. There is a line graph provided at the bottom showing the collective frequency of gene clusters available throughout the genome (all 7) lengths. The reason for variation is that some functional genes may be present in some set of genome isolates while the same genes could be non-functional in few other genomes creating the variations. Certain accessory gene availability in the species also depends on the environment these genome strains are extracted from.

3.2 INVESTIGATION ON *METHANOTHRIX SOEHNGENII* STRAINS

3.2.1 Roary Plot Results:

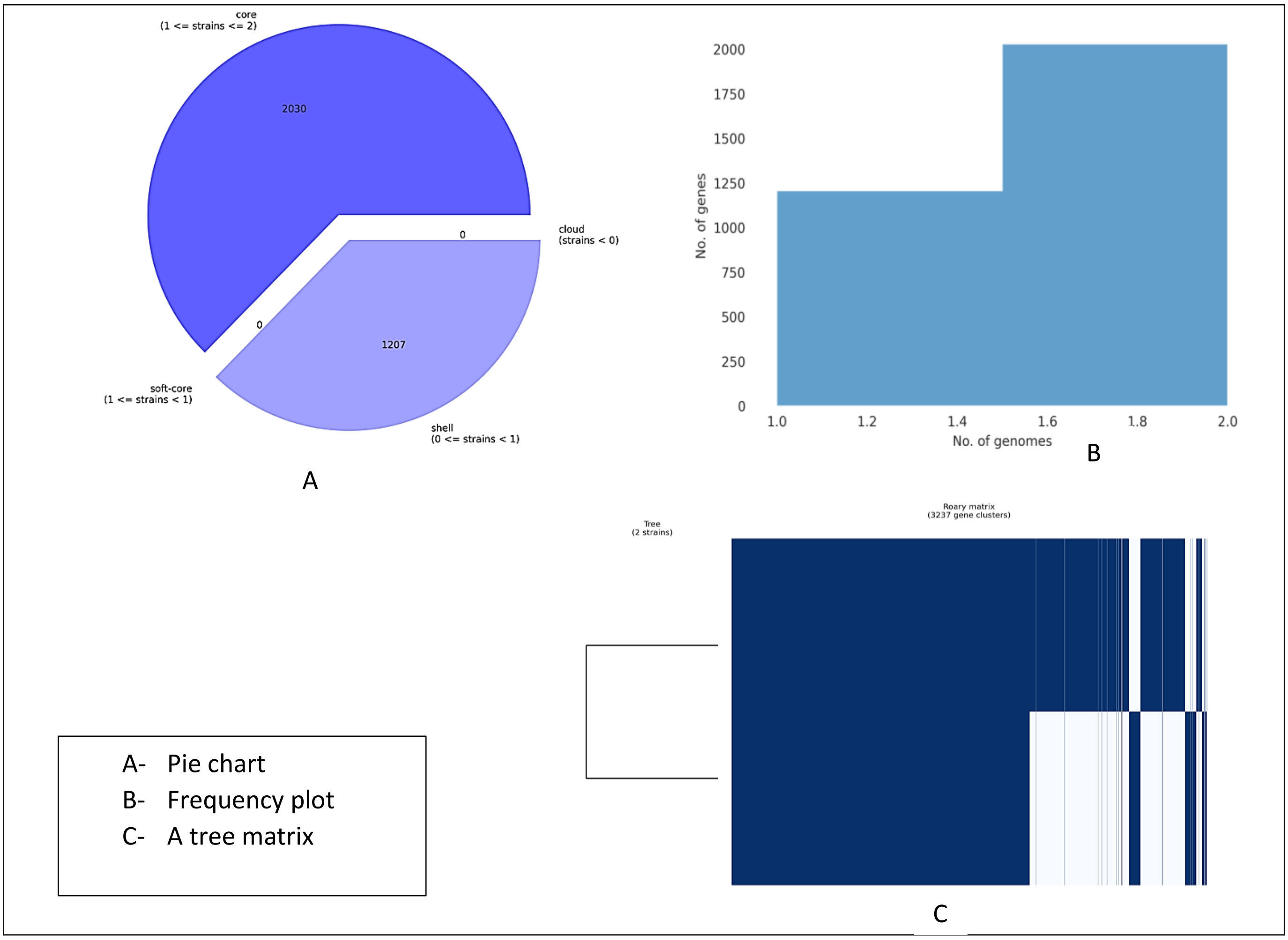


Figure 3.2.1 Visual plots from Roary plots of *M.soehngenii*

- The pie chart representing the gene composition of Methanothrix soehngenii is shown in figure 3.2.1, the same composition is represented in tabular format in appendix 4 table 2. According to the pie chart there is a total of 3237 functional genes with 2030 core gene statistics shared by the two genomes. The remaining 1207 genes have a probability of being present in one or both the strain of the species.
- The pangenome frequency plot in this case shows that first genome though has larger number of genes, the gene frequency through the strain length is lesser compared to second genome in *M.soehngenii,* thus giving a frequency graph in the given shape as shown in figure 3.2.1 B.
- As seen in figure 3.2.1. C, Both the genome strains are quite different from each other due the presence and absence of functional biochemical enzymes which may be available in one genome but not in the other like some accessory genes involved in salt stress. Some enzymes are available in both like the ones responsible for acetate oxidation while there are few absent in both like certain genes used for formate reduction.

3.2.2 Phandango Plot:

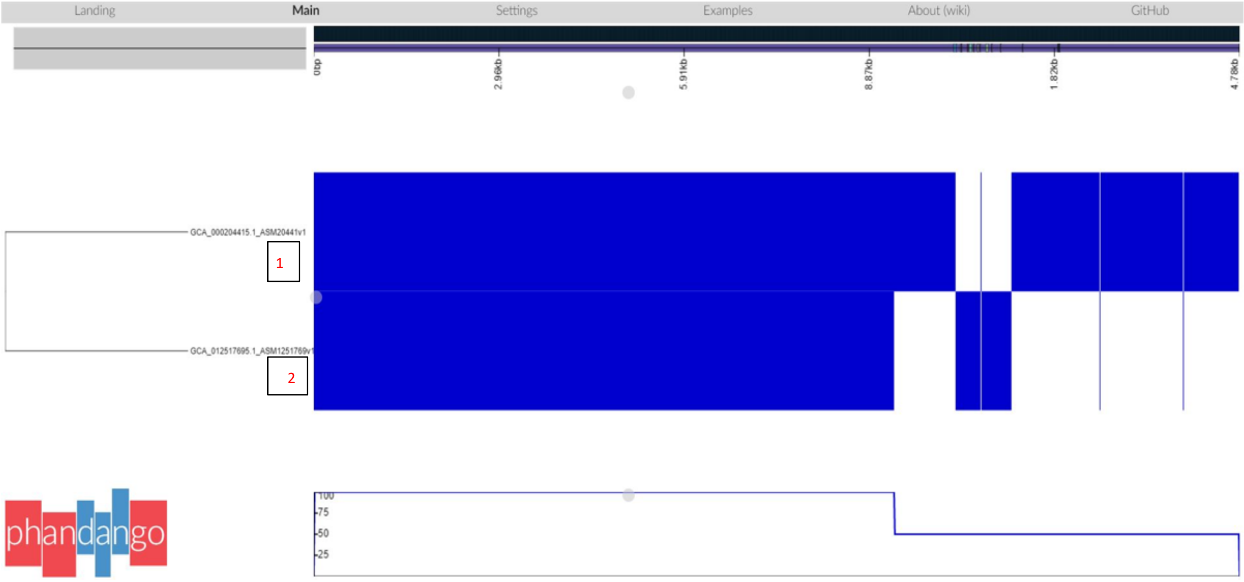


Figure 3.2.2 Plot from Phandango, numbering given to genomes for (Hadfield, et al., 2018)

The comparison of all the attributes of both the genomes and even the clustering of genes is evident in the Phandango plot, figure 3.2.2. Both genome 1 and genome 2 shows difference in the set of genes present in both. Initially, Both the genome isolates showed similar genetic composition and later shows variation. This is also shown by the graph in the Phandango plot where initially the line graph shows a maximum value implying core genes are present in that region shared by both species while further through the length the line graph decreases to half of its initial value implying that genomic material is now not common or shared by both.

3.3 COMPARATIVE STUDY OF THE ACETOCLASTIC METHANOGENS

- *M.barkeri* and *M.soehngenii*

Basic study is done on the group of methanogens in the previous sections. In order to do a comparative study of both a closer look is taken into these species. For this purpose, certain functional genes are selected from the Metabolic result.xlsx (the file gives all the metabolic pathways and gene present as well as absent in all the genome strains) file based on their potential to perform the core as well as special functions. The genes selected and their availability in each genomic strain of the species is listed below:

Gene Name	*Methanosarcina barkeri*	*Methanothrix soegenii*
Total genome available	**7**	**2**
methyl-coenzyme M reductase alpha subunit	7	1
acetyl-CoA decarboxylase	7	2
Glucoamylase	7	0
Isoamylase	2	0
Chitinase/amylase	4	2
methane monooxygenase regulatory protein B	4	0
Ni-Fe Hydrogenase, H2-uptake/unidirectional	7	0
Ni-Fe Hydrogenase, H2-evolving	7	0
dissimilatory sulphite reductase alpha subunit	4	2
nitrogenase iron-iron protein, alpha subunit	4	2
nitrous oxidase accessory protein	7	0

formate dehydrogenase major subunit	7	2
formate C-acetyltransferase	0	0
acetate => acetaldehyde	0	0
acetaldehyde => ethanol	5	0
branched-chain amino acid aminotransferase (acetate)	6	1
acetyl coenzyme A synthetase (ADP forming), alpha domain/ acetyl-CoA synthase (ammonia)	7	1
cytochrome bd ubiquinol oxidase subunit I	7	0
histidine-phosphate/aromatic aminotransferase	6	1
serine-pyruvate aminotransferase	2	0

Table 3. presence and absence of genes in the species

A comprehensive literature review of all the selected genes and their functions are studied and compared between both *Methanosarcina barkeri* and *Methanothrix soehngenii*.

M barkeri as well *M soehngenii* shows the presence of methyl-coenzyme reductase which is a basic core gene present in methanogens that promotes methane reduction (Zhang, et al., 2019). The enzymes Acetyl-CoA are the precursor enzyme for acetate activation by carbon fixation [5], being acetoclastic methanogens this enzyme is their basic characteristic, and the analysis shows the functional presence of this gene in both the species. Acetate oxidising methanogens uses different kinds of enzyme for breaking down complex carbon compounds like glucoamylase, isoamylase and chitinase to later get acetate substrates. *Methanosarcina barkeri* makes use of all the three while glucoamylase is dominant (seen in all these genome strains), on the other hand chitinase is the only enzyme functional in *Methanothrix soehngenii.* Apart from breaking down acetate *M barkeri* also shows the potential of oxidising methane to methanol another substrate that is utilised by *Methanosarcina barkeri* as well the first step of Methylotrophic methanogens (Smith, et al., 2011). *M soehngenii* shows the complete absence of this functional gene. The enzyme formate dehydrogenase shown in the table is a formate oxidising substrate helping in the breakdown of complex carbon compounds. Formate is the major substrate utilised by hydrogenotrophic methanogens and requires hydrogen ion for to complete the reaction. Hydrogen oxidisation is promoted by Ni-Fe hydrogenase enzymes. *Methanosarcina barkeri* is seen to have the potential of oxidising formate with the help of hydrogen ions as all the enzymes needed for this proceed is active in this species. The unidirectional form of Ni-Fe enzyme has oxygen tolerance which otherwise becomes inactive, surprisingly this function is seen in the results of *M barkeri* (Shafaat, et al., 2013). Methanogens require both sulphate and nitrogen for growth, the enzyme used for oxidation of both the substrates are present in all the genomes of both the species. These are the basic or core characteristics of the species.

There are certain genes that display certain special attributes that help in stress responses. One amongst that is the branched-chain amino acid transferase enzyme, this enzyme apart being amino acid breakdown enzyme also has shown the potential to help Methanogens like *Methanosarcina barkeri* to assimilate with increasing acetate stress (He, et al., 2019). Another inhibitor of acetogenesis is ammonia stress, increase in the ammonia concentration can stop the growth of acetate utilising methanogens. This has paved way for the adaptation of a new alternative pathway in case of high concentrations of ammonia stress, this pathway involves the participation of acs (acetyl-CoA synthase) instead of pta gene to continue carbon degradation (Yu, et al., 2020). A remarkable feature of *Methanosarcina barkeri* being an anaerobe is the presence of cytochrome bd ubiquinol oxidase whose primary function is the reduction of oxygen in molecular form found in the medium to H_2O (Borisov, et al., Jun 2021). This feature sets aside *Methanosarcina barkeri* aside from all the other anaerobes due to its unique ability to not lose its capability even under the presence of oxygen. Another stress gene found in methanogens is aromatic aminotransferase which shows the potential to be heat resistant making the species mildly thermophilic (Weigent DA, 1976 Nov 25). This property is shown by majority of the genomes in *Methanosarcina barkeri* and in some of *Methanothrix soehngenii* as shown in table 3. Serine-pyruvate aminotransferase has two stress properties to it, they have shown the ability to make a species hyper thermophilic. This enzyme was found in all the microbes obtained from very hot medium (Sakuraba H, 2004 Aug). This enzyme if found in a very few genomes of *M.barkeri* giving them this attribute depending on the environment Methanosarcina has to grow. Methanogens survive hypersaline conditions in very unique way. Until a smaller salt concentration, methanogens produce glutamate as ions to balance that salt concentration in the outside medium with the internal environment of the isolates. As the salt in the medium increases to hyper-salinity a compound called alanine is produced along with glutamate (Enzmann, et al., 2018). The product alanine is produced by the enzyme serine-pyruvate aminotransferase; hence this enzyme favours the growth of methanogens in hyper halophilic conditions while glutamate is produced by histidine-phosphate aminotransferase which helps the organism to survive salt concentrations of a particular level. Here 2 M.barkeri isolated have shown great potential against hyper-salinity and thermophilic conditions in the table 3.

3.3.1 Visual Comparison of *M.barkeri* and *M.soehngenii*

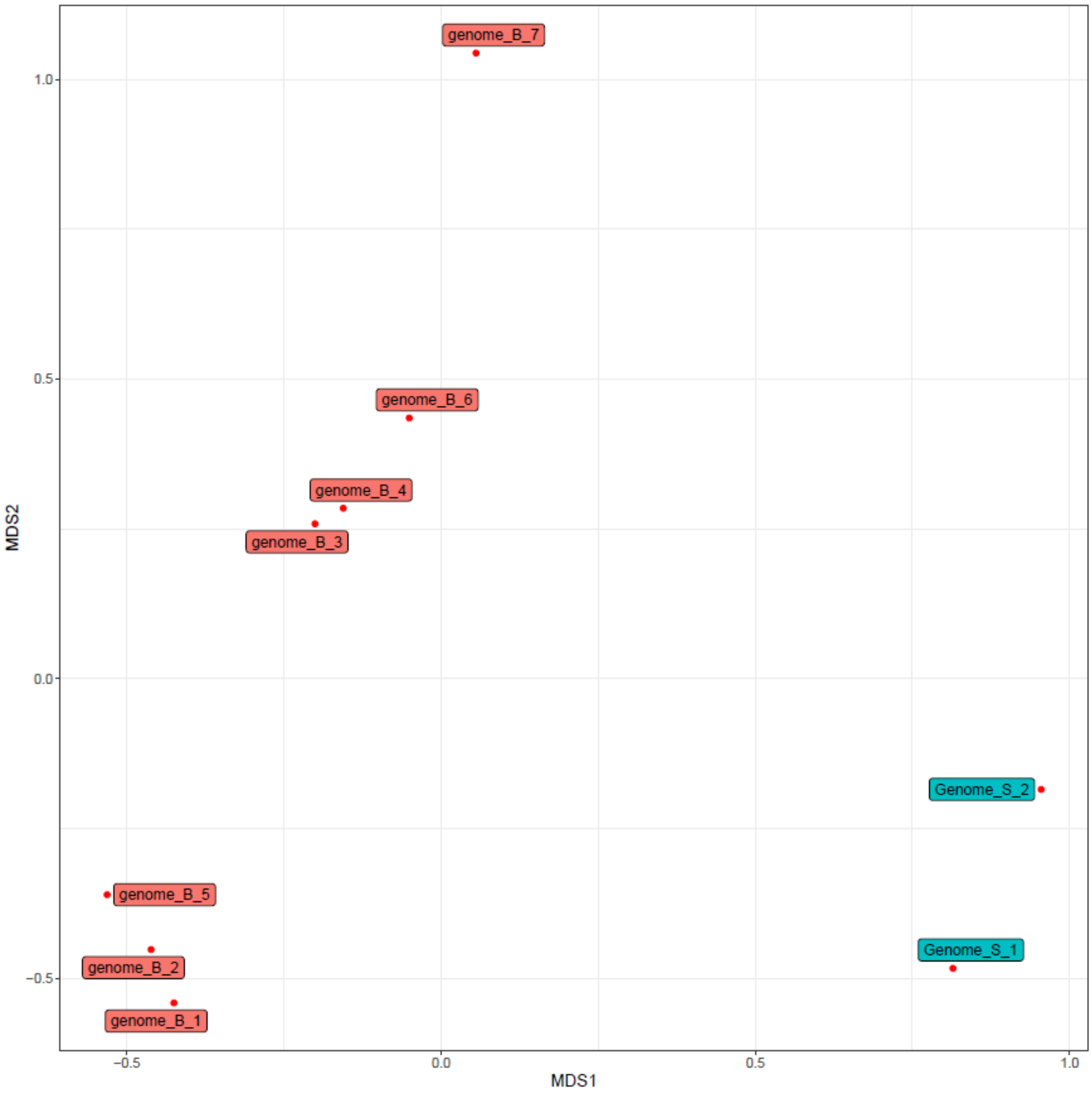


Figure 3.3.1 Plot obtained from R studio for *M barkeri* -red and *M soehngenii*-blue

The plot obtained from R studio shows the clustering in both the species. This is measure by using Jaccard index which is the distance from the centre of one genome to the centre of the next genome. The distance is calculated based on the comparison of the presence and absence of selected genes in each genome isolates from both the species and mapped in one plot. Here all the strains of *Methanosarcina barkeri* shown in red are clustered towards one side while all that of *Methanothrix soehngenii* are towards the other side in figure 3.3.1. This difference in Jaccard distance of the genomes is because of the presence of some genes in one set of strains but its absence in the remaining sets. Also depending on the medium from which each individual strains are extracted the genes in the genetic material of the species can show variation and may also develop new stress tolerance enzymes, thus the reason why strains of the same species exhibit some value of Jaccard distance between them.

3.4 INVESTIGATION ON *METHANOCOCCUS MARIPALUDIS* STRAINS

3.4.1 Roary Plot Results:

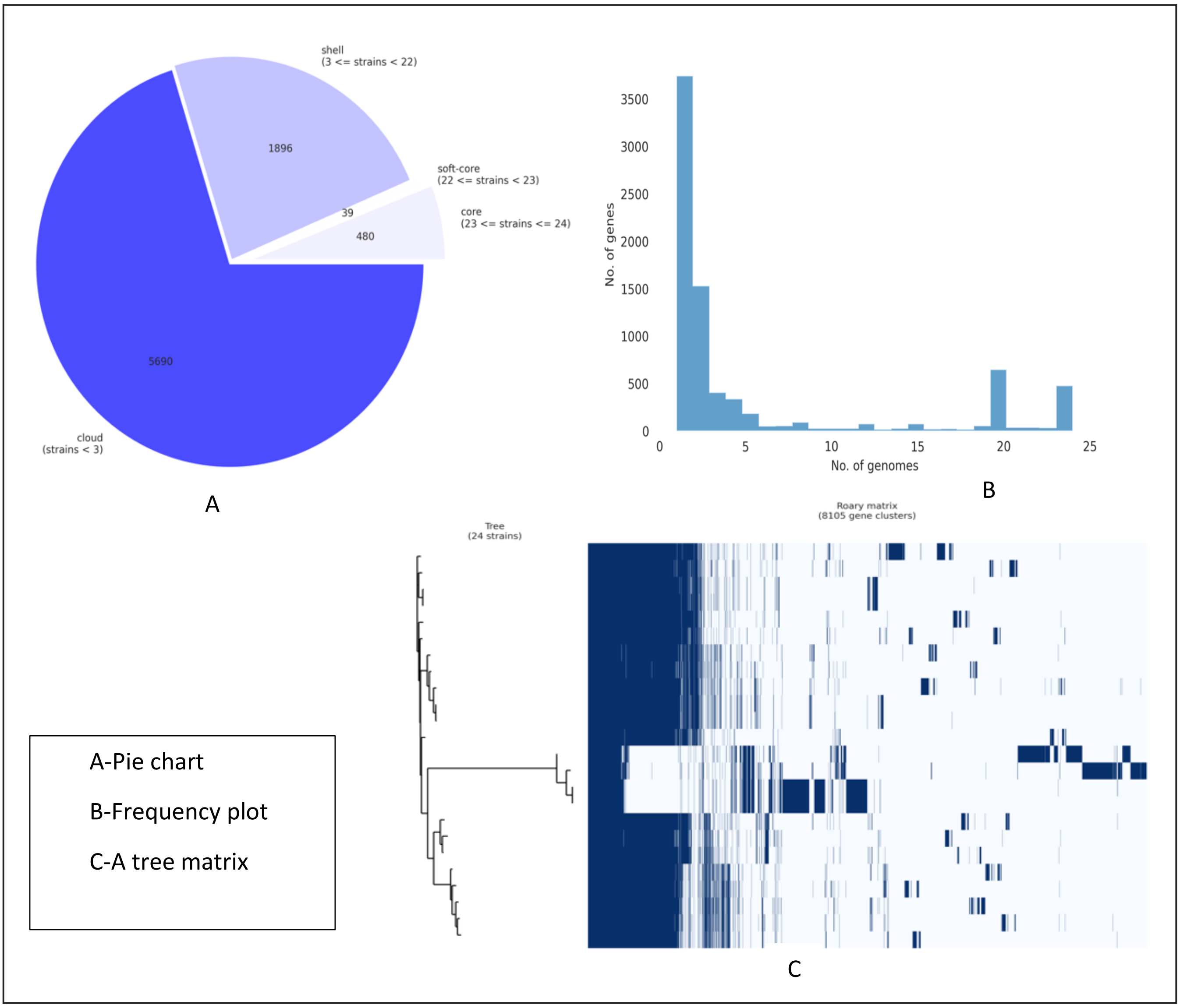


Figure 3.2.1 Visual plots from Roary plots of *M.maripaludis*

- Figure 3.2.1 A represents the gene compostion of the 24 genome strains (8105 total genes) in *M. maripaludis.* Only a proportion of 480 genes is shared by all the 24 strains of maripaludis while the scarcely shared or unique genes are shown to be 5690 in number.Almost 1896 shell genes are shared by atleast 3 to maxiumum 22 strains while soft core genes make a very samll contribution of 39 genes to the total of 8105 functional genes found.
- The frequency graph starts with a surge of more than 3500 genes in the first genome but the frequency of the genes per genome decreases rapidly to around 1500 in the second strain to the lowest of less than 500. The plot as in figure 3.2.1 B shows more frequency of genes towards the end around the 20th and 24th genome.
- Then final plot obtained fromroary plots is a phylogeny tree matrix, figure 3.2.1 C, which compares the clustering and area of genes present in all the 24 genomes based on phylogeny of the strains. The plot is very similar to the result produced by phandango and will be explained further in the section below.

3.4.2 Phandango Plot:

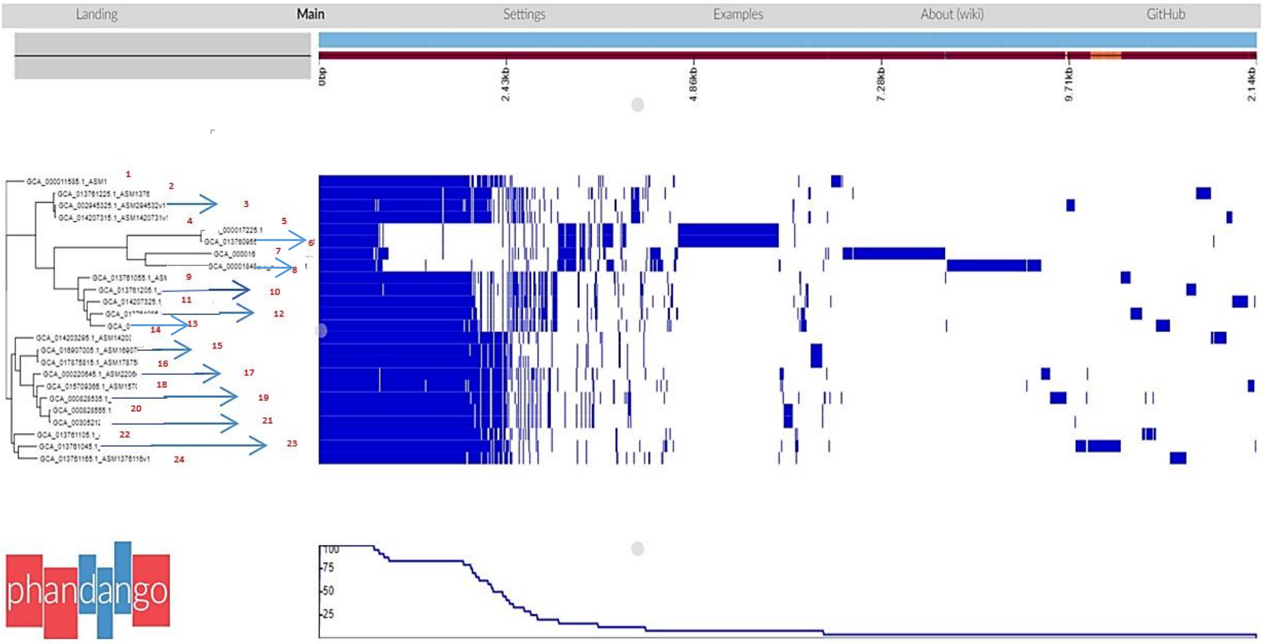


Figure 3.4.2 plot from phangando. Genomes given numberings for clarity (Hadfield, et al., 2018)

A lot of hierarchical evolution is shown by figure 3.4.2. The most diverse strains in the species are the genome isolate 5,6,7,8 as these variants took form after a number of divergences from the ancestral species. Amongst these the genome strain 5,6 show close similarities in the matrix while the isolates 7 and 8 slightly differ from each especially in the 7.2kb to 9.7kb length. Another set of newly formed variants are the genomes 9 to 14, they have a common ancestral point hence shows some similarities but again is different from each other as genome isolates evolves over time or due to development some stress gene that helps that genome isolate survive so on. Thus, the reason for minute variations and clustering of genes in the plot 3.4.2. The rest of the isolates from genome 14 to genome 24 in the map share a common point of divergence at some point showing that they share some core as well as shell genes. On the whole genome 1 seems to be the most different as it belongs to an older basic type of *M.maripaludis* and is located at the beginning of the phylogeny tree.

3.5 INVESTIGATION ON *METHANOBACTERIUM FORMICICUM* STRAINS

3.5.1 Roary Plot Results:

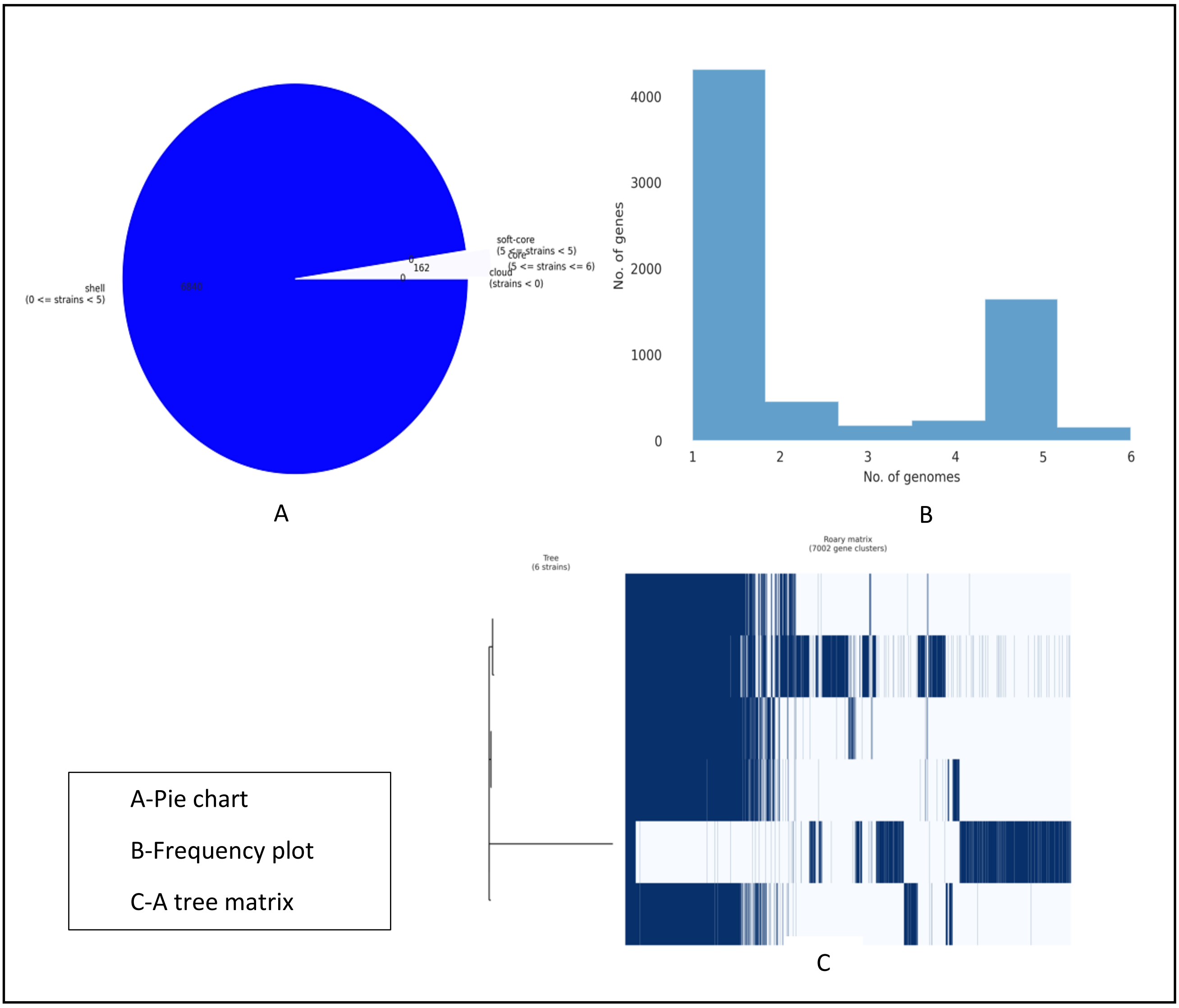


Figure 3.5.1 Shows the plots of *M.formicicum*

- Of the total 7002 genes present in the 6 genomes of *Methanobacterium formicicum* as seen in figure 3.5.1 A, it has 162 genes that are shared by all the genome isolates called core genes and the remaining 6840 genes are shared genes. Cloud genes as well as soft core genes are completely absent making the strains have lesser variations compared to other species so far discussed.
- The highest frequency of genes seems to be present in genome 1 around 4000 and genome 5 where the number of genes is less than 2000 in the frequency graph 3.5.1 B, the rest of the genomes seems to have lesser frequency of genes.
- Figure 3.5.1 C shows the phylogeny tree matrix where the phylogenetic classification as well as the matrix showing clusters and areas of similarities of each genome is shown.

3.5.2 Phandango Plot

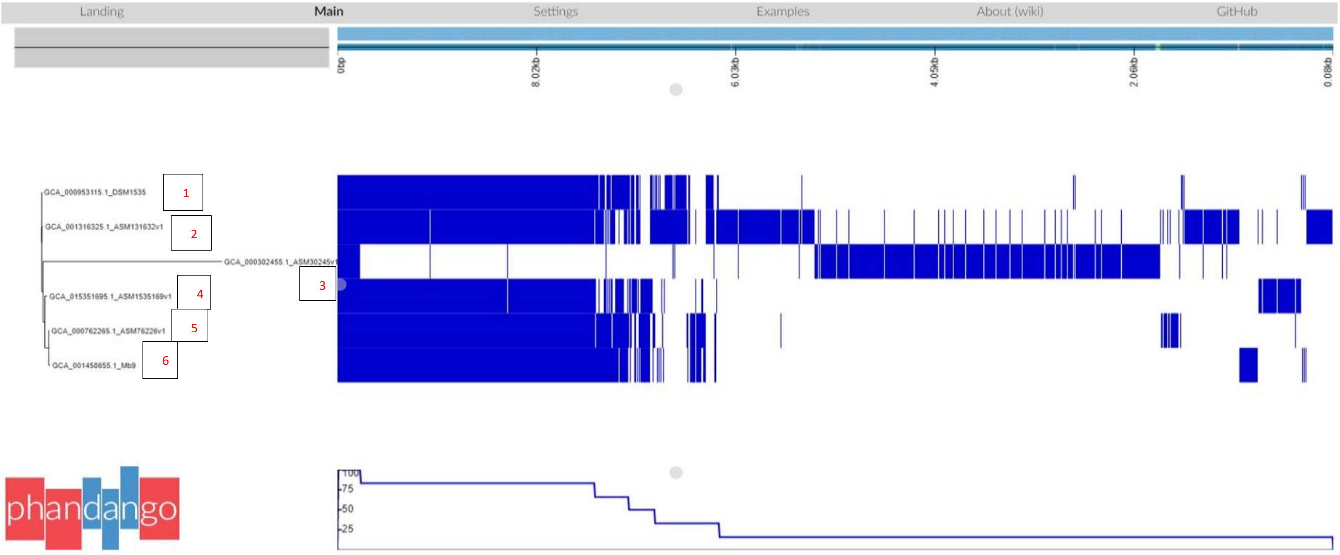


Figure 3.5.2 plot from Phandango, the genomes given separate numberings (Hadfield, et al., 2018)

According to the phylogenetic tree obtained the genome strain 3 shows the is most variation as seen in the matrix figure 3.5.2, It shows the presence of a lot of cloud genes responsible for this variation from the rest of the genomes. Whereas the genome strains 1,2,4,5,6 shows core gene similarities in the beginning of the genome length as they share a common point of ancestral divergence but shows slight variations as they are divergent strains of the predecessor genes. Though the genome isolates 1,2,4,5,6 show similarities genome 5,6 are similar in terms of gene presence absence to each other which can be seen in the gene matrix. A gene frequency versus genome length graph is also shown in the figure.

3.6 COMPARATIVE STUDY OF HYDROGENOTROPHIC METHANOGENS-

- *M.maripaludis* and *M.formicicum*

Genes of functional importance are filtered and chosen from the metabolic result file for comparative study of both the species. Following are the genes taken under consideration:

Gene Name	*Methanococcus maripaludis*	*Methanobacterium formicicum*
Total genome available	**24**	**6**
methyl-coenzyme M reductase alpha subunit	24	6
Glucoamylase	24	0
Chitinase/amylase	24	0
formate dehydrogenase major subunit	2	5
formate dehydrogenase (coenzyme F420) alpha subunit	2	5
formate dehydrogenase (coenzyme F420) beta subunit	24	6
Ni-Fe hydrogenase, Bidirectional	24	1
Ni-Fe hydrogenase, H_2 evolving	0	6
Ni-Fe hydrogenase, H_2 uptake (unidirectional)	24	6
acetyl-CoA decarboxylase	24	6
methane monooxygenase regulatory protein B	0	0
dissimilatory sulphite reductase alpha subunit	0	0
nitrogenase iron-iron protein, alpha subunit	24	5
nitrous oxidase accessory protein	17	2
acetate => acetaldehyde	0	0
acetaldehyde => ethanol	0	5
branched-chain amino acid aminotransferase (acetate)	24	6

acetyl coenzyme A synthetase (ADP forming), alpha domain/ acetyl-CoA synthase (ammonia)	2	5
cytochrome bd ubiquinol oxidase subunit I	0	0
histidine-phosphate/aromatic aminotransferase	24	6
serine-pyruvate aminotransferase	24	6

Table 4: Gene availability in *M.maripaludis* and *M.formicicum*

Additional formate decomposing genes are selected apart from other mainstream genes in case of hydrogenotrophic methanogens to analyse the validity of their main pathway. The metabolic characteristics of both the methanogens is discussed based on the potential and availabilities of each gene in table 4. Th function of all the genes taken for analysis is given separately table 5 of appendix 5.

Methanococcus maripaludis as well as *Methanobacterium formicicum* shows the presence of methyl coenzyme reductase which is the key player in methane formation and should be available in all methanogenic species to qualify as one. Glucoamylase, isoamylase and chitinase are different type of genes used during carbon fixation, amongst them glucoamylase and chitinase are found in all the strains of *M.maripaludis* while *M.formicicum* uses a different pathway for carbon fixation. Hydrogenotrophic methanogens grow mainly on formate, they have the ability to synthesise methane from CO_2 or CO and hydrogen. Formate dehydrogenase enzymes has the potential to break down formate to perform methanogenesis. Referring to table 4*, M.formicicum* shows higher potential than *M.maripaludis* as the formate oxidising enzyme formate dehydrogenase is seen in almost all the strains of the former species. While both the species shows high traces of coenzyme 420 that acts as the catalysing agent in breaking down formate by assisting the supply of hydrogen ions. These hydrogen ions are converted to usable forms by Ni-Fe hydrogenase enzyme seen in decent number of genomes of both the species. The enzyme acetyl-CoA decarboxylase shows full functionality in these species, the enzyme participates in acetate activation to convert acetate to simpler carbon compounds or for the generation of formate. Hydrogenotrophic methanogens does not oxidise acetate for methane generation. Unlike acetoclastic methanogens as discussed before these species of methanogens does not grow on other substrates like sulphite or ethanol. Here these species are seen to lack any sulphite oxidising enzyme while around 5 genome strains of *M.formicicum* shows traces of ethanol synthesising enzyme. *M.formicicum* as well as *M.maripaludis* both gets energy by growing on nitrogen compounds and almost all the genomes in both these species share genes that are capable of nitrogen fixation and nitrogen oxidation.

The research shows both *M.maripaludis* and *M.formicicum* has a few functional stress genes showing the potential of adapting to hostile environments. One of the most common type of stress that could be faced by hydrogenotrophic methanogens is the hydrogen stress or formate stress. During this period the coenzyme 420 becomes more active capture hydrogen ions but the growth of *Methanococcus maripaludis* and *Methanobacterium formicicum* is seen to decrease after a point of hydrogen starvation and formate starvation (Costa 1, et al., 2013). The branched chain amino acid apart from acting on amino acid compounds helps to withstand acetate stress and helps break down acetate for formate formation. Though this enzyme is found to be functional all the strains here, hydrogenotrophic methanogens does not have to deal with acetate stress as syntrophic acetate oxidising bacteria breaks down acetate into hydrogen and CO_2 further used by hydrogenotrophic methanogens and the utilisation of hydrogen is necessary for the growth of acetate oxidising bacteria creating a syntrophic relationship between them (Nishio & Nakashimada, 2013). Both *M.formicicum* and *M.maripaludis* has genes that shows the remarkable potential to adapt to extreme hot or saline environments. Here both the species shows the presence of aromatic aminotransferase and serine pyruvate aminotransferase responsible for giving these methanogens these unique qualities, thus the species selected could be identified as thermophilic and halophilic. Both *M.maripaludis* and *M.formicicum* shows no traces of cytochrome bd ubiquinol oxidase, hence unlike *Methanosarcina barkeri* both these methanogens is extremely hostile to oxygen and cannot function even in the presence of small quantities of oxygen.

3.6.1 Visual Comparison of *M.maripaludis* and *M.formicicum*

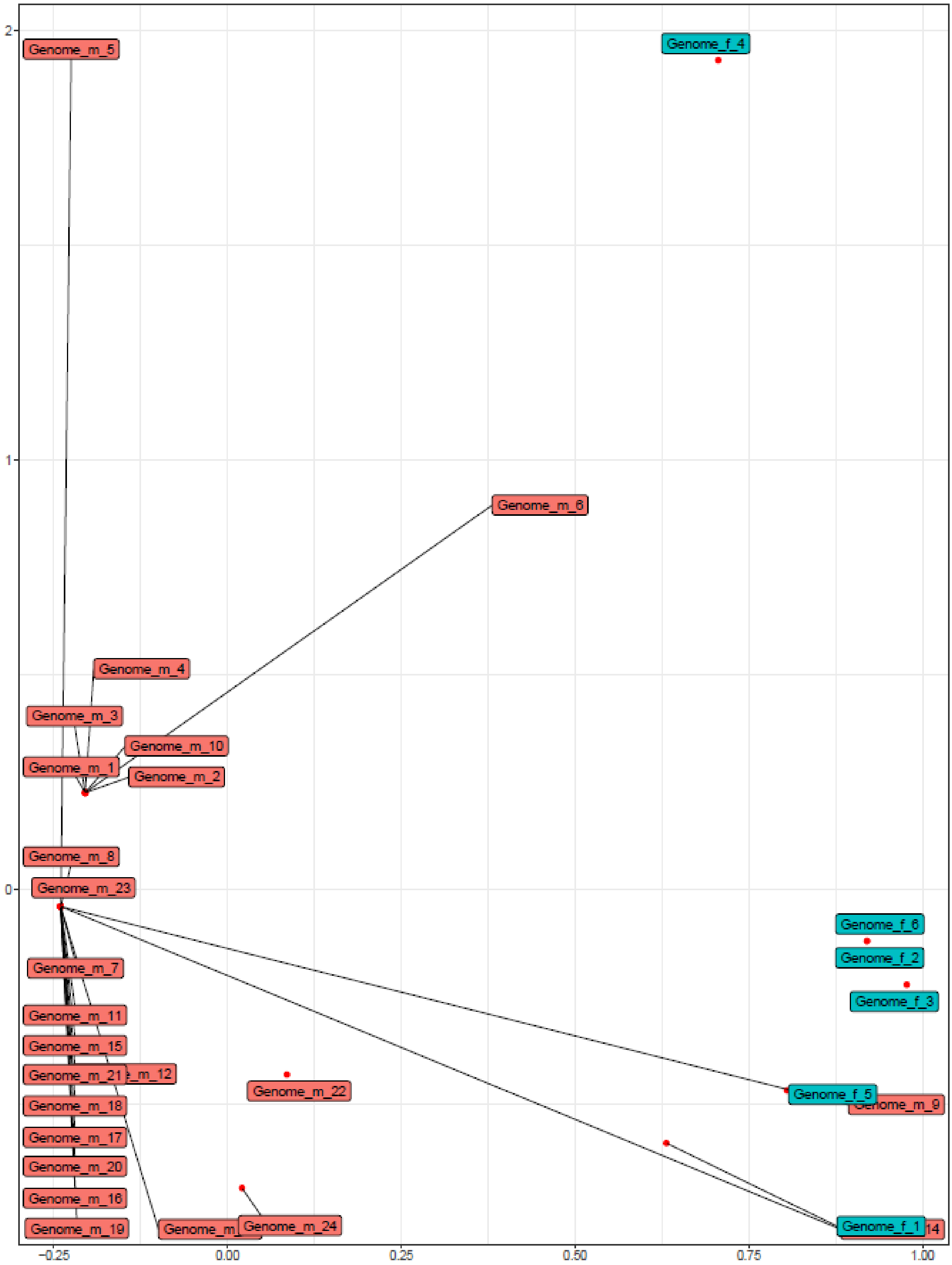


Figure 3.6.1 Plot obtained from R studio for *M maripaludis* -red and *M formicicum*-blue

The clustering of *Methanococcus maripaludis* and *Methanobacterium formicicum* are checked for visual comparison by using the Jaccard distance of each genome from other genomes. The Jaccard index is calculated taking the similarities and dissimilarities that is, the presence and absence of genes into consideration. The plot here is a pictorial representation of the gene availability in table 4.Here *Methanococcus maripaludis s*hown by red in the plot 3.6.1 shows a lot of genome cluster at one point indicating that the genomes are almost similar with less phylogeny. It also shows one more cluster with lesser genome strains and the genomes in this cluster are closely similar with no Jaccard distance between them. While the genomes of *Methanobacterium formicicum* in blue in the figure shows its genome on the other extreme side of the plot indicating that these genome strains are different from the ones of *Methanococcus maripaludis.* Nonetheless, Two genome strains of *M.formicicum* forms cluster with the *M maripaludis* genes as they have the same potential functional genes available.

4. CONCLUSION AND FUTURE CONSIDERATION

This study helped to get a closer view into four methanogens following two different pathways. By conducting the whole genome analysis the important biochemical pathways of acetoclastic as well as hydrogenotrophic methanogens were able to be analysed. This helped understand the potential and functionality of each gene in the species. These are studided in this research so that the methanogens present in an anaerobic reactor can be manipulated in such a way that they survive the extremities without losing its viability and at the same time produce methane and breakdown organic matter.

After analysing both *Methansarcina bakeri* and *Methanothrix soehngenii* belonging to acetoclastic pathway, it could be seen that though *Methanothrix soehngenii* has more affinity towards acetate the genes and metabolic pathway present in *Methanosarcina barkeri* shows the ability to withstand acetate stress and not lose its functionality with increasing acetate concentration. Apart from having the potential of tolerance to other hostile conditions like increase in ammonia content these methanogenic species can even change from mesophilic to thermophilic as well as can be seen survive in salt medium. Surprisingly, this species shows the ability to survive even in the presence of oxygen. *Methanosarcina barkeri* is also called all-rounder in methanogenic species as they have the ability to grow on almost all substrates, though acetate is the main substrate for *M.barkeri* it can grow on sulphite, ammonia, methanol and formate substances. These qualities make this species of methanogens dominant and one of the most relevant in the acetoclastic methanogens group.

Amongst the hydrogenotrophic methanogens studied, *Methanobacterium formicicum* seems to show stronger formate-hydrogen affinity than *Methanococcus maripaludis* which makes them dominant in this pathway. The strains of both these methanogens shows the potential of having high heat stability as high osmotic stability making them halophilic. They exist in a syntrophic relationship with acetate oxidising bacteria that oxidises acetate to carbon compounds and hydrogen that is used by hydrogenotrophic bacteria to produce methanogen.

In an anaerobic digestor a community of micro-organisms are seen, it is important is to promote the growth of most efficient group of organisms in this reactor so that the reactor functions in an effective manner. In future, it is advisable to promote the culture of *Methanosarcina barkeri* in AD reactors as they have the ability to act on almost all the organic matter. This microbe could be cultured in close relationship with the two hydrogenotrophic methanogens(*Methanococcus maripaludis* and *Methanobacterium formicicum*) which could live in a syntrophic relationship with the acetate oxidising bacteria. In this way when *Methanosarcina barkeri* is faced with acetate stress, the acetate oxidising bacteria breaks down acetate relieving the environment of this stress. The products from acetate oxidising bacteria can be utilised by both acetoclastic as well as hydrogenotrophic methanogens in the reactor.

The reactor can be set in high or low temperature for energy generation as all these three methanogens can thrive in the medium even if it is hot. In an anaerobic digestor a mixed community of anaerobes are present, and the reactor conditions should be in such a way that all the species present in the medium should thrive equally and complement each other.

5. References

Borisov, V. . B. et al., Jun 2021. Bacterial Oxidases of the Cytochrome bd Family: Redox Enzymes of Unique Structure, Function, and Utility As Drug Targets. *Antioxidants & Redox Signaling,* pp. 1280-1318.

Costa 1, K. . C. et al., 2013. *Effects of H2 and formate on growth yield and regulation of methanogenesis in Methanococcus maripaludis.*

Enzmann, F., Mayer , F. & Rother, M., 2018. *Methanogens: biochemical background and biotechnological applications.*

Goyal, N., Zhou , Z. & Karimi, . I. . I. A., 2016. *Metabolic processes of Methanococcus maripaludis and potential applications,* 10 June.15(107).

Hadfield, J. et al., 2018. *genomics, Phandango: an interactive viewer for bacterial population,* 15 January, Volume 34(Issue 2), p. 292–293.

He, P. et al., 2019. Responses of Methanosarcina barkeri to acetate stress. *Biotechnol Biofuels 12, 289.*

Huser, B. A., Wuhrmann , K. & Zehnder, . A. J. B., 1982. *Methanothrix soehngenii gen. nov. sp. nov., a new acetotrophic non-hydrogen-oxidizing methane bacterium,* July.1–9 (1982)(132).

Kurth, J. . M., Op den Camp , H. J. M. & Cornelia , 2020. *Several ways one goal—methanogenesis from unconventional substrates,* June.Issue 6839–6854 (2020).

Ladapo, J. & Whitman, W. B., 1990. *Method for isolation of auxotrophs in the methanogenic archaebacteria: role of the acetyl-CoA pathway of autotrophic CO2 fixation in Methanococcus maripaludis.,* Issue 87(15): 5598–5602.

Lieber, D. J., Catlett, J., Madayiputhiya, N. & Nandakumar, R., 2014. *A Multienzyme Complex Channels Substrates and Electrons through Acetyl-CoA and Methane Biosynthesis Pathways in Methanosarcina,* september.

Maus, I. et al., 2014. Complete genome sequence of the methanogenic neotype strain Methanobacterium formicicum MFT. *Journal of Biotechnology,* Volume Volume 192, Part A,.

Meegoda, J. N., Li, B., Patel, K. & Wang, L. B., 2018. *A Review of the Processes, Parameters, and Optimization of Anaerobic Digestion,* 11 October.Volume no. 10: 2224.

Mosin , O. & Ignatov, I., 2014. Journal of Medicine, Physiology and Biophysics. *Metabolism and Physiology of Methylotrophic Microorganisms,* Volume Vol 6.

Nishio, N. & Nakashimada, . Y., 2013. Manufacture of Biogas and Fertilizer from Solid Food Wastes by Means of Anaerobic Digestion.

Page, A., 2018. *GitHub.* [Online]
Available at: https://sanger-pathogens.github.io/Roary/
[Accessed 22 Jan 2018].

Seemann , T., 15 July 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* Volume 30(Issue 14), p. 2068–2069.

Shah, F. A. et al., 2014. *Microbial Ecology of Anaerobic Digesters: The Key Players of Anaerobiosis,* Volume vol. 2014, Article ID 183752, p. 21 pages.

Stams, A. J. M., Teusink, B. & Sousa, D. . Z., 2019. *Ecophysiology of Acetoclastic Methanogens,* Volume In: Stams A., Sousa D. (eds) Biogenesis of Hydrocarbons. Handbook of Hydrocarbon and Lipid Microbiology. .

Yu, D. et al., 2020. Ammonia stress decreased biomarker genes of acetoclastic methanogenesis and second peak of production rates during anaerobic digestion of swine manure. *Bioresource Technology,* Volume 317(124012).

Anon., n.d. *Merch.* [Online]
Available at: https://www.sigmaaldrich.com/GB/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing

Bioquest, A., n.d. *www.aatbio.com.* [Online]
Available at: https://www.aatbio.com/resources/faq-frequently-asked-questions/How-does-bridge-amplification-work-in-next-generation-sequencing-NGS

Demirel, B. & Scherer, . P., 2008. *The roles of acetotrophic and hydrogenotrophic methanogens during anaerobic conversion of biomass to methane: a review,* Volume Rev Environ Sci Biotechnol 7, 173–190.

EMBL-EBI, n.d. *EMBL-EBI.* [Online]
Available at: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/ion-torrent-proton-pgm-sequencing/

GOV.UK, 2011. *Anaerobic digestion strategy and action plan.* [Online]
Available at: https://www.gov.uk/government/publications/anaerobic-digestion-strategy-and-action-plan

Gray & N.F, 2004. *Biology of wastewater treatment (Vol. 4). World Scientific..* s.l.:s.n.

Lettinga, G., 1995. *Anaerobic digestion and wastewater treatment systems,* March.

Madison, A. M. L. a. W., 2021. *Github.* [Online]
Available at: https://github.com/AnantharamanLab/METABOLIC
[Accessed 24 Aug 2021].

Merch, n.d. *Merch.* [Online]
Available at: https://www.sigmaaldrich.com/GB/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing

NCBI, 2021. *NCBI.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/genbank/

Novogene, n.d. *en.novogene.com.* [Online]
Available at: https://en.novogene.com/services/research-services/metagenomics/shotgun-metagenomic-sequencing/#overview

Novogene, n.d. *en.novogene.com.* [Online]
Available at: https://en.novogene.com/services/research-services/genome-sequencing/whole-genome-sequencing/microbial-whole-genome-resequencing/

Sakuraba H, K. R. T. H. O. T., 2004 Aug. Novel Archaeal Alanine:Glyoxylate Aminotransferase from Thermococcus litoralis. *J Bacteriol,* Issue 186(16): 5513–5518.

Shafaat, H. . S., Rüdiger, O., Ogata, H. & Lubitz, W., 2013. [NiFe] hydrogenases: A common active site for hydrogen metabolism under diverse conditions. *Biochimica et Biophysica Acta (BBA) - Bioenergetics,* Volume 1827 (Issues 8–9), pp. 986-1002.

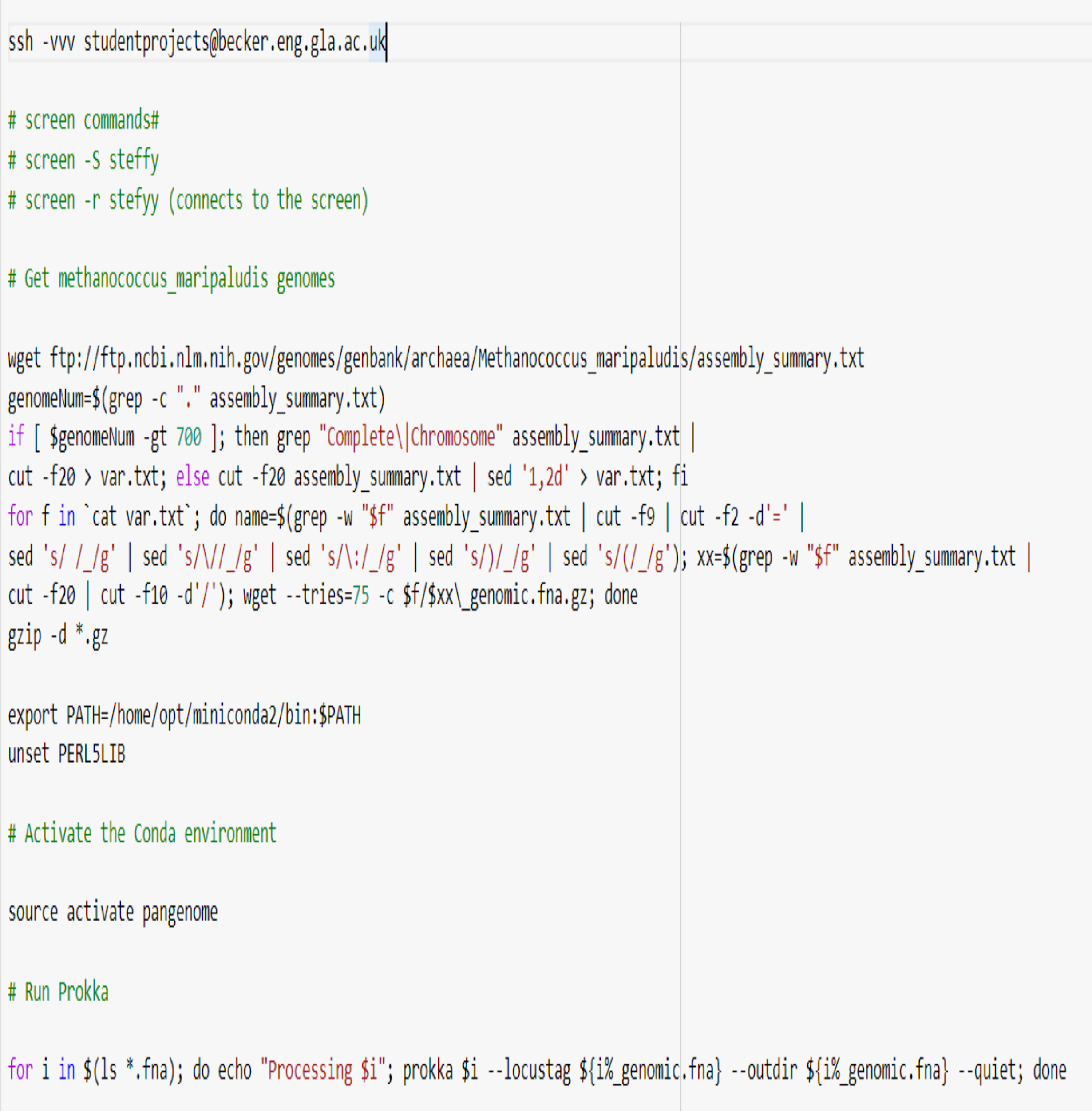
Smith, S. M., Balasubramanian, R. & Rosenzweig, A. C., 2011. Chapter thirteen - Metal Reconstitution of Particulate Methane Monooxygenase and Heterologous Expression of the pmoB Subunit. *Methods in Enzymology,* Volume 495(https://www.sciencedirect.com/science/article/pii/B9780123869050000139).

Weigent DA, N. E., 1976 Nov 25. Purification and properties of two aromatic aminotransferases in Bacillus subtilis. *J Biol Chem,* Volume 251(22):6974-80. PMID: 11213.

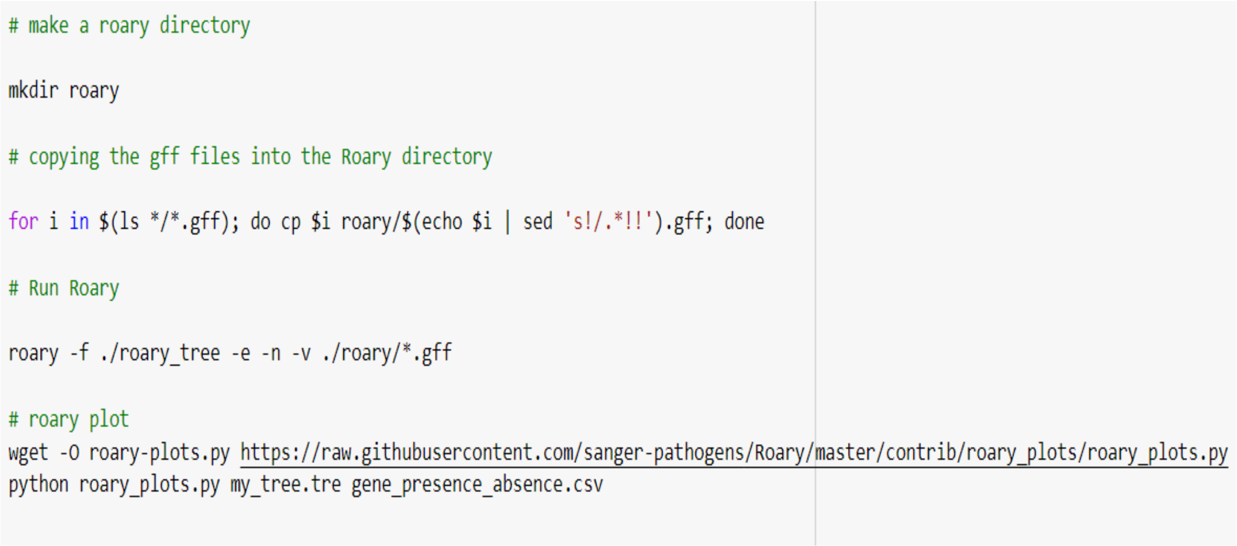
yourgenome, n.d. *yourgenome.* [Online]
Available at: https://www.yourgenome.org/facts/what-is-the-454-method-of-dna-sequencing

Zhang, Y., Ma, A., Zhuang, G. & Zhuang , X., 2019. The acetotrophic pathway dominates methane production in Zoige alpine wetland coexisting with hydrogenotrophic pathway. *Scientific Reports,* Volume volume 9.

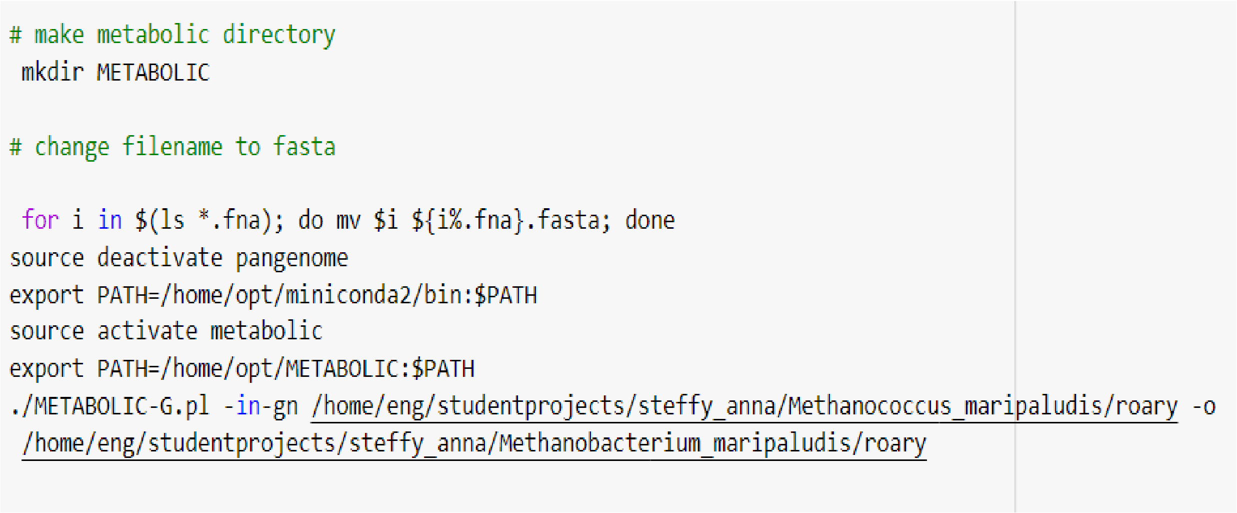
APPENDIX 1



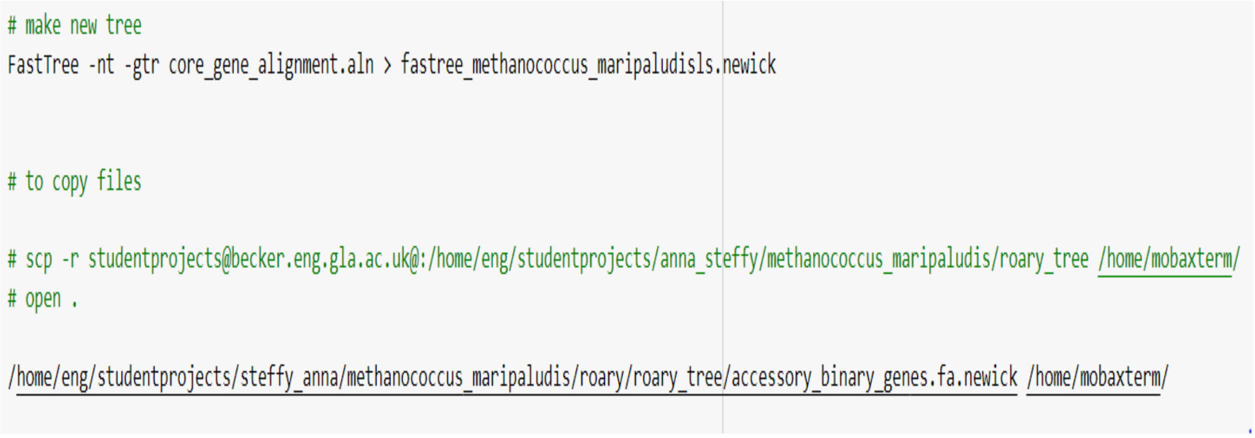
```
ssh -vvv studentprojects@becker.eng.gla.ac.uk

# screen commands#
# screen -S steffy
# screen -r stefyy (connects to the screen)

# Get methanococcus_maripaludis genomes

wget ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/archaea/Methanococcus_maripaludis/assembly_summary.txt
genomeNum=$(grep -c "." assembly_summary.txt)
if [ $genomeNum -gt 700 ]; then grep "Complete\|Chromosome" assembly_summary.txt |
cut -f20 > var.txt; else cut -f20 assembly_summary.txt | sed '1,2d' > var.txt; fi
for f in `cat var.txt`; do name=$(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
gzip -d *.gz

export PATH=/home/opt/miniconda2/bin:$PATH
unset PERL5LIB

# Activate the Conda environment

source activate pangenome

# Run Prokka

for i in $(ls *.fna); do echo "Processing $i"; prokka $i --locustag ${i%_genomic.fna} --outdir ${i%_genomic.fna} --quiet; done
```



```
# make a roary directory

mkdir roary

# copying the gff files into the Roary directory

for i in $(ls */*.gff); do cp $i roary/$(echo $i | sed 's!/.*!!').gff; done

# Run Roary

roary -f ./roary_tree -e -n -v ./roary/*.gff

# roary plot
wget -O roary-plots.py https://raw.githubusercontent.com/sanger-pathogens/Roary/master/contrib/roary_plots/roary_plots.py
python roary_plots.py my_tree.tre gene_presence_absence.csv
```



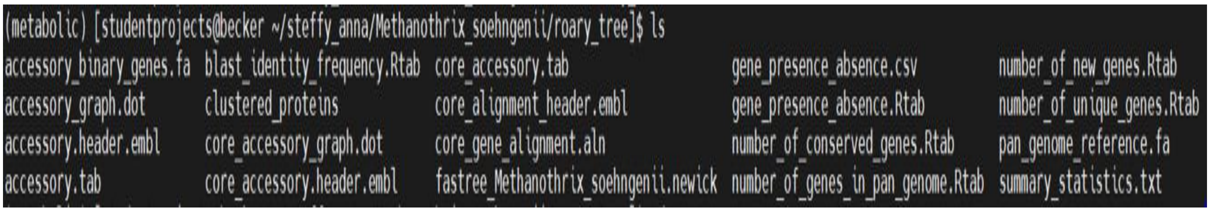
```
# make metabolic directory
 mkdir METABOLIC

# change filename to fasta

 for i in $(ls *.fna); do mv $i ${i%.fna}.fasta; done
source deactivate pangenome
export PATH=/home/opt/miniconda2/bin:$PATH
source activate metabolic
export PATH=/home/opt/METABOLIC:$PATH
./METABOLIC-G.pl -in-gn /home/eng/studentprojects/steffy_anna/Methanococcus_maripaludis/roary -o
 /home/eng/studentprojects/steffy_anna/Methanobacterium_maripaludis/roary
```



```
# make new tree
FastTree -nt -gtr core_gene_alignment.aln > fastree_methanococcus_maripaludisls.newick



# to copy files

# scp -r studentprojects@becker.eng.gla.ac.uk@:/home/eng/studentprojects/anna_steffy/methanococcus_maripaludis/roary_tree /home/mobaxterm/
# open .

/home/eng/studentprojects/steffy_anna/methanococcus_maripaludis/roary/roary_tree/accessory_binary_genes.fa.newick /home/mobaxterm/
```

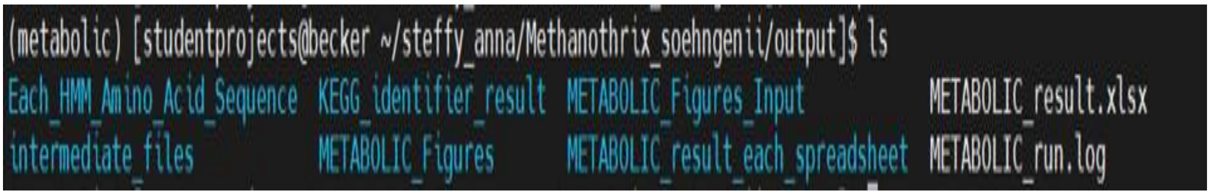
A basic script of the workflow (here shown with *Methanococcus maripaludis*)

APPENDIX 2



```
(metabolic) [studentprojects@becker ~/steffy_anna/Methanothrix_soehngenii/roary_tree]$ ls
accessory_binary_genes.fa  blast_identity_frequency.Rtab  core_accessory.tab                       gene_presence_absence.csv         number_of_new_genes.Rtab
accessory_graph.dot        clustered_proteins             core_alignment_header.embl               gene_presence_absence.Rtab        number_of_unique_genes.Rtab
accessory.header.embl      core_accessory_graph.dot       core_gene_alignment.aln                  number_of_conserved_genes.Rtab    pan_genome_reference.fa
accessory.tab              core_accessory.header.embl     fastree_Methanothrix_soehngenii.newick   number_of_genes_in_pan_genome.Rtab  summary_statistics.txt
```

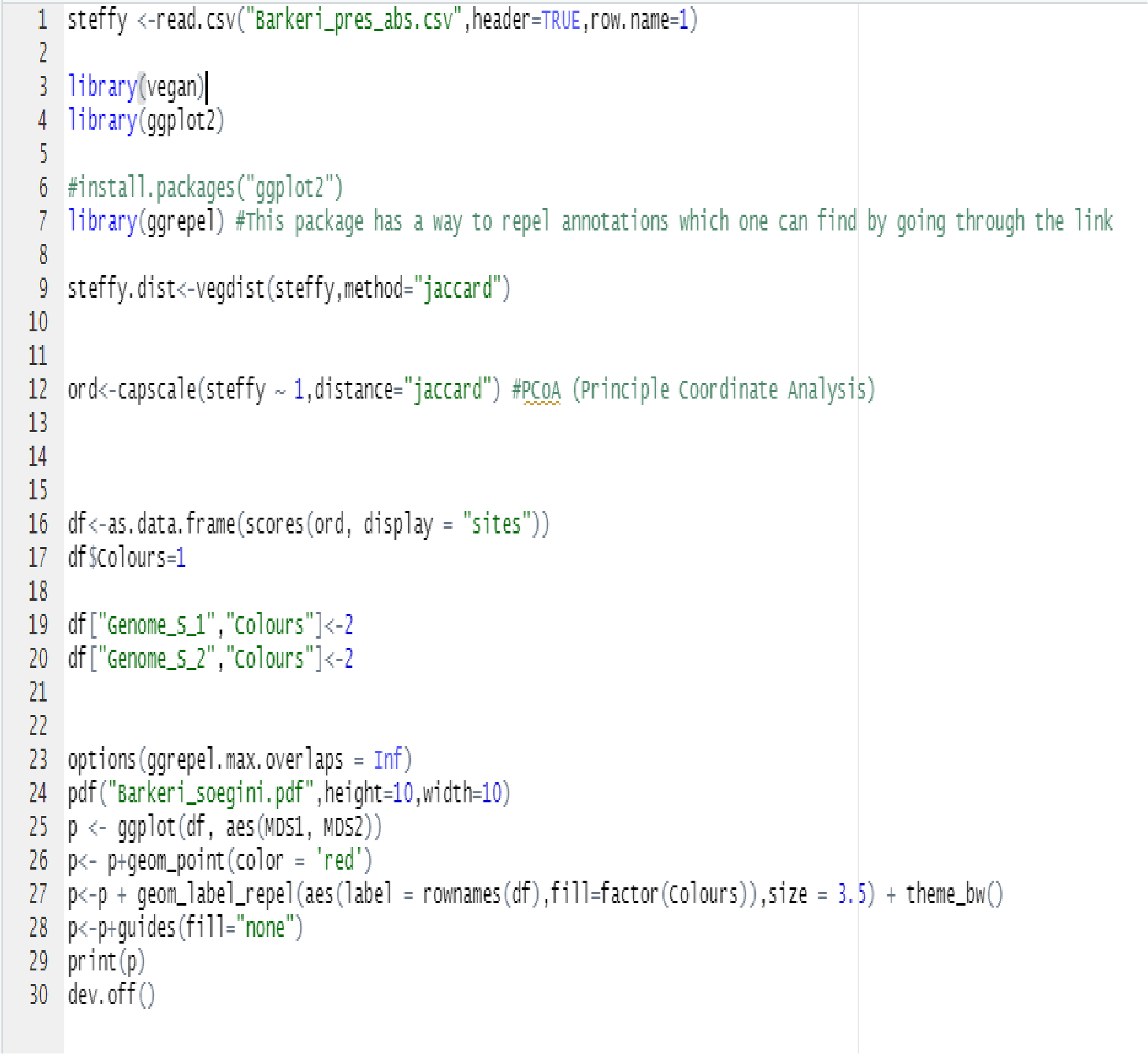
a. Figure shows all the output files of Roary command of *M.soehngenni*



```
(metabolic) [studentprojects@becker ~/steffy_anna/Methanothrix_soehngenii/output]$ ls
Each_HMM_Amino_Acid_Sequence  KEGG_identifier_result  METABOLIC_Figures_Input            METABOLIC_result.xlsx
intermediate_files            METABOLIC_Figures       METABOLIC_result_each_spreadsheet  METABOLIC_run.log
```

b. Figures showing all output files of Metabolic of *M.soehngenii*

APPENDIX 3



```
steffy <-read.csv("Barkeri_pres_abs.csv",header=TRUE,row.name=1)

library(vegan)
library(ggplot2)

#install.packages("ggplot2")
library(ggrepel) #This package has a way to repel annotations which one can find by going through the link

steffy.dist<-vegdist(steffy,method="jaccard")


ord<-capscale(steffy ~ 1,distance="jaccard") #PCoA (Principle Coordinate Analysis)



df<-as.data.frame(scores(ord, display = "sites"))
df$Colours=1

df["Genome_S_1","Colours"]<-2
df["Genome_S_2","Colours"]<-2


options(ggrepel.max.overlaps = Inf)
pdf("Barkeri_soegini.pdf",height=10,width=10)
p <- ggplot(df, aes(MDS1, MDS2))
p<- p+geom_point(color = 'red')
p<-p + geom_label_repel(aes(label = rownames(df),fill=factor(Colours)),size = 3.5) + theme_bw()
p<-p+guides(fill="none")
print(p)
dev.off()
```

Figure shows the R script done on *M.barkeri*

APPENDIX 4

Type of Gene	Probability of presence	Number of Genes
Core genes	(99% <= strains <= 100%)	484
Soft core genes	(95% <= strains < 99%)	0
Shell genes	(15% <= strains < 95%)	6088
Cloud genes	(0% <= strains < 15%)	5654
Total genes	(0% <= strains <= 100%)	12226

Table 1. Gene composition of *M.barkeri*

Type of Gene	Probability of presence	Number of Genes
Core genes	(99% <= strains <= 100%)	2030
Soft core genes	(95% <= strains < 99%)	0
Shell genes	(15% <= strains < 95%)	1207
Cloud genes	(0% <= strains < 15%)	0
Total genes	(0% <= strains <= 100%)	3237

Table 2. Gene composition of *M.soehngenii*

Type of Gene	Probability of presence	Number of Genes
Core genes	(99% <= strains <= 100%)	480
Soft core genes	(95% <= strains < 99%)	39
Shell genes	(15% <= strains < 95%)	1896
Cloud genes	(0% <= strains < 15%)	5690
Total genes	(0% <= strains <= 100%)	8105

Table 3. Gene composition of *M.maripaludis*

Type of Gene	Probability of presence	Number of Genes
Core genes	(99% <= strains <= 100%)	162
Soft core genes	(95% <= strains < 99%)	0
Shell genes	(15% <= strains < 95%)	6840
Cloud genes	(0% <= strains < 15%)	0
Total genes	(0% <= strains <= 100%)	7002

Table 4. Gene composition of *M.formicicum*

APPENDIX 5

Gene Name	Functions
methyl-coenzyme M reductase alpha subunit	- Methane production
acetyl-CoA decarboxylase	- Acetogenesis
Glucoamylase	- Carbon fixation
Isoamylase	- Carbon fixation
Chitinase/amylase	- Carbon fixation
methane monooxygenase regulatory protein B	- Methane to methanol
Ni-Fe Hydrogenase, H2-uptake/unidirectional	- Hydrogen oxidation
Ni-Fe Hydrogenase, H2-evolving	- Hydrogen oxidation
dissimilatory sulphite reductase alpha subunit	- Sulphite reduction
nitrogenase iron-iron protein, alpha subunit	- Breaks down nitrogen compound in wastewater
nitrous oxidase accessory protein	- Reduces the nitrogen compound produced in the previous step
formate dehydrogenase major subunit	- Formate oxidation
formate C-acetyltransferase	- Formate to pyruvate
acetate => acetaldehyde	- Acetate breakdown
acetaldehyde => ethanol	- Acetaldehyde breakdown
branched-chain amino acid aminotransferase (acetate)	- Breaksdown ammonia as well as helps with acetate stress
acetyl coenzyme A synthetase (ADP forming), alpha domain/ acetyl-CoA synthase (ammonia)	- Alternative pathway in case of ammonia stress
cytochrome bd ubiquinol oxidase subunit I	- Oxygen stress
histidine-phosphate/aromatic aminotransferase	- Thermophilic characteristics
serine-pyruvate aminotransferase	- Salt stress

Table 5: Shows all the function of selected genes