Analysing Microbial Communities

Kimberley Barnes

MSc by Research

University of York

Biology

March 2018

Abstract

Anaerobic digestion, the decomposition of organic matter to biogas and digestate in the absence of oxygen, is carried out by diverse communities of microorganisms. Until recently, 16S rRNA gene amplification has been the main focus towards better understanding of these communities, ultimately for their exploitation in industry and waste management. Metagenomics and shotgun whole genome sequencing now offers a different approach, allowing for the functional analysis of individual members of the community without the need for cell culturing. But metagenomics is not without its own pitfalls. Currently there are limited tools and methods available for use with large and complex datasets from sequencing of anaerobic digestion communities. Here we present the development of a rapid fully automated software pipeline for the large-scale identification and functional analysis of quality genomes extracted from anaerobic digestion metagenomic datasets. The pipeline consists of two new tools for the analysis of metagenomic data; the MCCR tool for reducing contamination in proposed genomes formed from metagenomic data, and the MPP tool for simultaneously predicting metabolic pathways across the large numbers of organisms found in metagenomes. The tools and pipeline were tested on both synthetic and real datasets during their development, and while further development will be needed in the future, this pipeline shows high potential to be both viable and extremely useful in understanding complex metagenomic datasets.

List of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank James Chong for the opportunity to undertake a Masters by research and for his support and guidance during this project. I would also like to thank all the members of the Chong group as well as members of the bioinformatics team for making me feel welcome and supporting me in this project.

Declaration

I declare that this thesis is a presentation of original work and I, Kimberley Barnes, am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

All the work in this thesis is my own with the following exceptions:

- The DNA extraction was performed by Dr Anna Alessi.
- The metagenome assembly and contig binning of the combined dataset used in Chapter 3 was performed by Dr James P. J. Chong.

Abbreviations

%GC – G+C content

3GS – 3rd generation sequencing

AD – anaerobic digestions

GHG – greenhouse gas

NGS – next generation sequencing

OTU – operational taxonomic unit

PCR – polymerase chain reaction

rRNA – ribosomal RNA

TNF – tetranucleotide frequency

VFA – volatile fatty acid

WWT – waste water treatment

1. Introduction

Increasing concerns over a changing climate require us to take a new approach to resource management in regard to both implementing alternate energy production strategies and better protection of the natural environment. Anaerobic digestion (AD) of organic waste material provides a solution to both of these challenges, capable of exploiting multiple waste streams as resources in energy generation, and by preventing the various environmental harms that are associated with our current waste management[1]. The second largest contributor to global warming after carbon dioxide (CO_2) is methane, having ~85 times the potency of CO_2 as a greenhouse gas (GHG) over the short term of 20 years[2]. Atmospheric methane concentrations have increased by 150% since the 1750's and although only contributing ~17% towards the total effects of GHGs, 50-65% can be attributed to human activity[2,3]. Reduction of methane emissions will be necessary in order to control global temperature increases and keep them below the 2°C rise laid out in the Paris Agreement.

Methane is generated largely through uncontrolled AD in intensive farming, waste water treatment and municipal waste landfills where it is released into the atmosphere[1,4]. However, methane is also the main component in natural gas used for industry, heating and electricity generation, and biogas generated from AD plants has the potential to cost effectively replace fossil fuels in grid balancing[5]. Volatile renewable energy sources such as wind and solar are predicted to provide a large proportion of the electrical energy demand in Western Europe over the coming years, but renewable energy systems from wind and solar exhibit large temporal fluctuations in output[5,6]. By combining variable and intermittent renewable resources with those renewable resources offering high levels of predictability, for example electricity generation from burning biogas, a larger proportion of renewable energy can be integrated into energy systems[6]. Biomethane, upgraded from biogas, is also important for direct energy generation rather than conversion to electricity. In 2011, 52% of gas in the UK was used for heat generation compared to 34% to generate electricity[7]. The potential value of AD has not been overlooked, and as of 2012 over 13,800 biogas plants had already been built in Europe, with the United Kingdom producing 1764 kilo tonnes of petroleum equivalent of biogas per year[8]. For biogas and biomethane to become leading energy sources, better understanding of the processes and the complex microbial community behind AD is needed to optimize biogas yields. Metagenomics helps to provide insight into this community where before much of it was unknown, but better tools to reconstruct these complex communities computationally are needed before genetic analysis can be used to positively alter the microbial community for robustly increased methane yields.

1.1 Anaerobic digestion

1.1.1 The role of uncontrolled environmental anaerobic digestion

AD is a natural part of carbon cycling in which organic matter is degraded by microorganisms in environments where oxygen is limiting. 35-50% of the global methane is from biotic sources, including wetlands, ruminant animals and even some termite mounds[3,9]. In these environments the methane produced cannot be harnessed and is simply released into air, and so understanding and manipulation of the microbial community to reduce methane production is desirable.

1.1.1.1 Anaerobic digestion in wetlands

Natural wetlands, such as marshes, peatlands and swamps, are estimated to produce ~30% of global methane emissions[3,10,11]. As global temperatures increase, it predicted that methane emissions from wetlands will increase[3]. This is partially due to larger areas of northern tundra annually and perennially being released from permafrost, which have the potential to release an additional 63% of stored carbon in the region for decomposition to CO_2 or methane [3,12,13]. The release of soil carbon creates a positive feedback loop, where by carbon is released in response to rising temperature, which rises in response to increased release of soil carbon as CO_2 and methane. Peatlands show the highest ratio of carbon release of the three tested ecosystem types (boreal forest, tundra and peatland) in response to a 10°C rise in temperature from the northern tundra[12], and are therefore of particular interest in helping to reach global GHG emission targets[14].

1.1.1.2 Anaerobic digestion in agriculture

An additional ~20% of global methane production can be attributed to agriculture, the two largest contributors being ruminant livestock (~13%) and rice paddies (~5%)[3]. Global food demand over the last 50 years has tripled, and several modelling scenarios of future global food demand indicates that both plant- and animal-based demand will continue to strongly increase[15]. Within the stomachs of many ruminant livestock, including sheep, cattle and goats, plant material is fermented generating the H_2 and CO_2 necessary for methane production. Not only does this generate GHGs but loss of energy via methanogenesis, which utilizes 2-12% of that ingested by the animal, is also undesirable[16]. Manipulation of the enteric microbial community to supress methane production has been met with mixed results *in vitro*[17–19], and so many have turned to a bioinformatic approach to better understand methanogenesis in the ruminant microbiome[16,20–23].

Methane production in rice paddy agriculture is a result of similar conditions to those found in wetlands due to substantial amounts of submerged anoxic organic carbon in soil. Similarly to wetlands, methane emissions from paddies are expected to increase in response to increasing atmospheric CO_2 and global temperatures[24]. Methane emissions in rice paddies can be far more easily controlled than emissions by ruminants or wetlands and there is clear experimental evidence to support this[9]. Manipulation of the microbial community through intermittent drainage or irrigation of waterlogged paddies, a system often used throughout Asia, has been shown to significantly decrease methane production and alterations in nitrogen fertilizer usage can also have a large impact[9]. But the conditions in rice paddies are highly variable and methane production unpredictable. To mitigate the likely rise in methane output in response to increases in global population and food requirements, a number of computational models have been built to more accurately simulate methanogenesis in rice paddies[25–27].

1.1.2 The role of controlled anaerobic digestion in industry

CO_2 and methane represent the 2 largest contributors to GHGs. The exponential use of fossil fuels since the start of the Industrial Era (1750), in addition to their well-known increasing of atmospheric CO_2 levels, also accounted for 30% of anthropogenic methane emissions between 2000-2009[3]. Methane emissions from landfills and waste accounted for an additional 23% of anthropogenic sources between 2000-2009, and in the 6 years between 2005 and 2011 alone, atmospheric methane concentrations have increased by 1.5%[2,3]. In contrast to understanding microbial communities in biotic methane generation to reduce methane emissions, the appeal of understanding abiotic methane producing communities is to increase end product yields and decrease reliance on fossil fuels.

1.1.2.1 Anaerobic digestion in biotechnology

While a large focus of industrial AD lies in biogas generation from bio-waste, AD is also being investigated as a potential cost-effective method to produce a wide variety of high value products. In addition to methane, AD can be manipulated to produce a number of fermentative products including alcohols, aldehydes and organic acids. The production of biofuel has become a booming industry across the Americas[28]. Crops such as maize (corn) or sugarcane are grown specifically for fermentation into biofuel precursors such as ethanol (or biogas). Volatile fatty acids (VFA) produced by fermentation in AD systems, such as acetic acid, propionic acid or butyric acid, are another area of interest for biotechnology. Propionic acid and it's salts are primarily used in the food industry as food preservatives, although it is also used on a smaller scale for the production of cellulose derived biodegradable plastics[29]. Acetic acid is also used as a preservative in the food industry as the main compound in vinegar, and the production of various plastics[30]. Despite 100 years of research into utilizing a microbial community for large-scale production of VFAs, they are still largely produced using fossil fuels[29–31]. Although the potential in using microbial communities for the production of value added products is high, currently in many instances it is not cost effective to do so due to low yields and/or high downstream processing costs[28,29,32]. Optimization of operating conditions combined with genetic manipulations to utilize alternate waste resources are paving a way for lower costs and higher use[8,28,30,33].

1.1.2.2 Anaerobic digestion in waste management

Industrialised AD in waste management utilizes biodegradable waste as a resource by converting it to biogas. It is classed as a low carbon impact process responsible for dealing with a wide range of different organic waste products, ranging from human and animal waste to organic waste from cheese production[1,34,35]. As a result of the low cost implications, AD has been gaining in popularity and functionality over the last 40 years as both a solution to biological waste management from agricultural, industrial and municipal waste, and as a clean energy source[1,8]. The study of AD of solid bio-waste has existed for several decades, long before the need to reduce global carbon emissions became apparent. One of the earliest papers on AD by Cooney and Wise in 1975 saw the potential of AD as both a solution for the disposal of organic waste and as a means of converting waste into fuel[36]. During the 1970's the main focus of AD was for the treatment of organic waste material rather than energy generation, part of its attractiveness being that the gaseous end products could be easily disposed of by venting or burning, but also the stabilization of various organic substrates and decreases in volume before disposal[36]. Nowadays biogas production is a burgeoning industry.

To further reduce the volume of solids produced from waste water treatment (WWT), approximately 75% of solids from WWT now undergoes AD in the UK generating biogas as a byproduct[37]. Biogas is a combination of methane (50-70%), CO_2 (30-50%) and trace gases such as hydrogen sulphide that can be used in combined heat and power plants, or requires costly upgrading to biomethane before it can be directly injected into the national grid[5,6]. To sustainably balance the environmental and economic costs of WWT and

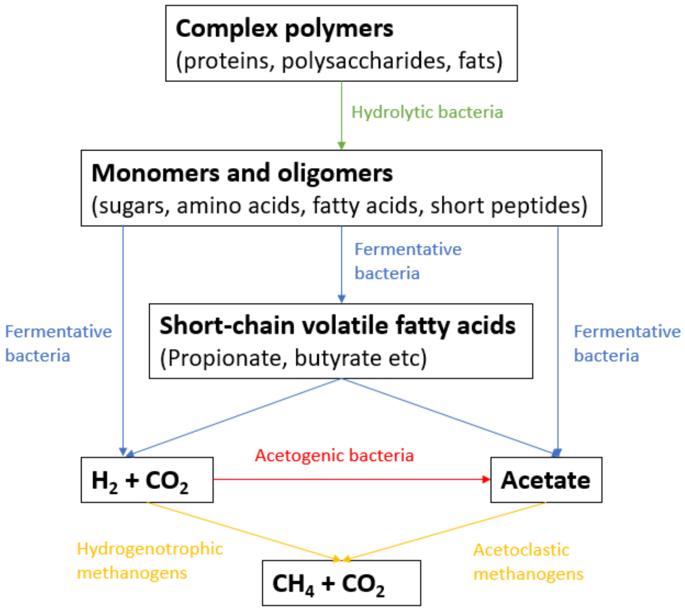


Figure 1.1 The 4 key steps of AD

A simplified overview of key steps involved in AD: hydrolysis, fermentation/acidogenesis, acetogenesis and methanogenesis. Bacteria are responsible for 3 of the 4 key steps, while methanogenic archaea are responsible for the final step.

converting it to the required >90% biomethane, far higher yields of biogas and concentrations of methane are needed. From a biochemical point of view this is achievable: it is estimated that the amount of energy that can be generated from AD of WWT is 10 times higher than the energy currently used to treat it[38]. Better understanding of the microbial community is needed in order to optimise energy recovery from waste water treatment.

1.1.3 Biochemical Steps

AD and biomethanation are carried out by a complex community of microorganisms, with each species contributing to one or more stages of the syntrophic process. Simply put, AD can be split into 4 steps: hydrolysis, fermentation, acetogenesis and methanogenesis, and illustrated as in Figure 1.1 [1,39].

1.1.3.1 Hydrolysis

The hydrolysis stage of AD relies upon a multitude of extracellular enzymes and reactions to hydrolyse large polymeric compounds into readily available substrates for the entire microbial community (Equation 1.1). As such, hydrolysis can be considered the rate limiting step in many digesters[40].

Depending on the origin, municipal or agricultural, biomass added to AD systems can consist of a high percentage of plant or lignocellulosic material made up of complex insoluble polymers including cellulose and hemicelluloses. Because the polymers that make up lignocellulose are so large and chemically inert, they require specialized extracellular glucosidases to be hydrolysed into their component sugars for uptake. Hemicelluloses are a wide class of many branched polysaccharides including xylans and glucomannans which consist of sugar monomers such as glucose, xylose, mannose, galactose and arabinose. Cellulose is a linear polysaccharide of hundreds to thousands of glucose molecules. Extracellular glucosidases such as cellulases and xylanases are in part responsible for the hydrolytic decomposition of lignocellulosic material. Hydrolysis is also important for hydrolysing proteins into amino acids and lipids into glycerol and long chain fatty acids.

Complex polymers → tri, di and monomers

Equation 1.1

1.1.3.2 Fermentation/Acidogenesis

The released sugars, amino acids and glycerol from hydrolysis are fed into fermentative pathways producing CO_2, hydrogen (H_2), ammonia and a variety of reduced mono or poly-carbon compounds including alcohols, aldehydes and VFAs. Pathways in fermentation can be homofermentative, only producing a single end product like acetate (Equation 1.2), or heterofermentative, for example with acetic acid as a coproduct to propionic or butyric acid (Equation 1.3)[30].

$$C_6H_{12}O_6 + 2H_2O \rightarrow 2CH_3COOH + 2CO_2 + 4H_2$$

Equation 1.2

$$2C_6H_{12}O_6 \rightarrow 2CH_3COOH + 2CH_3CH_2COOH + 2CO_2 + 2H_2$$

Equation 1.3

1.1.3.3 Acetogenesis

Acetogenesis is the formation of acetate from single or poly-carbon compounds. The simplest form of acetogenesis is the stepwise combination of H_2 and the acetyl groups from two single carbon compounds to form acetate[41]. Often this is CO_2 (Equation 1.4), but can also be formate, methanol or methyl groups from methoxylated aromatic compounds[41].

$2CO_2 + 4H_2 \rightarrow CH_3COOH + 2H_2O$

Equation 1.4

Acetogenesis is a combination of the reductive acetyl-CoA or Wood-Ljungdahl pathway which fixes carbon into acetyl-CoA, and the acetate kinase pathway which converts acetyl-CoA into acetate coupled to ATP synthesis[41]. The Wood-Ljungdahl pathway is split into the Eastern (or Methyl) and Western(or Carbonyl) branches. All microbes have the Eastern branch of the Wood-Ljungdahl pathway as it's important in one-carbon metabolism, while only those using the full Wood-Ljungdahl pathway for carbon fixation have the Western branch[41].

1.1.3.4 Methanogenesis

Methanogenesis consists of a few select pathways that convert simple carbon molecules, most often CO_2 generated from the previous steps of AD into methane, and represents the final step in carbon reduction. It is not a particularly thermodynamically favourable, and will only take place in the absence of alternate electron acceptors[42]. Methanogenesis can be split into 3 classes based on their terminal electron acceptor, hydrogenotrophic, acetoclastic and methylotrophic methanogenesis. Hydrogenotrophic and acetoclastic represent the predominant pathways of methanogenesis although there is increasing evidence that the importance of methylotrophic methanogenesis to global methane emissions has been underestimated[43].

In hydrogenotrophic methanogenesis H_2 and CO_2 generated during fermentation/acidogenesis are combined to generate methane and energy (Equation 1.5). H_2 acts as an electron donor, which can sometimes be replaced by formate, with CO_2 as the electron acceptor.

$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$

Equation 1.5

Acetoclastic methanogens use acetic acid produced from either the fermentation/acidogenesis or acetogenesis steps as a substrate, generating methane and CO_2 (Equation 1.6)[41]. As a result the gaseous products of AD are not purely methane, but a combination of methane and CO_2 as a result of the combined efforts of acetoclastic and hydrogenotrophic methanogens [5].

$CH_3COOH \rightarrow CH_4 + CO_2$

Equation 1.6

Methylotrophic methanogenesis encompasses all the methanogenesis pathways utilizing methyl-compounds as electron acceptors. The substrates used include methanol, methylated-amines or methylated-sulphides instead of CO_2 or acetate. Despite being considered one class of methanogenesis each substrate uses a slightly different pathway and require substrate specific methyltransferases. Only 8 sequenced methanogens are known to obligately use a combination of H_2 and methyl-compounds[43].

Although there are 3 classes of methanogenesis, the terminal step in methane generation is always the same utilizing the methyl-coenzyme M reductase complex McrABCDG in the conversion of methyl-coenzyme M into methane and a heterodisulphide of coenzyme M and coenzyme B[44].

1.1.4 The role of microbes

Both 16S rRNA gene and metagenomic sequencing of anaerobic digesters have emphasized the complexity of the microbial community within them[40,45–48]. The competition and syntrophy between microorganisms carrying out different pathways and steps is essential for a balanced AD community.

1.1.4.1 The Bacteria

Bacteria can make up 95% of the diversity/biomass in AD communities and are responsible for carrying out all the steps of AD except methanogenesis[45]. Unlike the specificity of methanogenesis, the metabolic pathways for hydrolysis, fermentation and acetogenesis are spread throughout the bacterial phyla. Bacterial diversity in AD largely depends on the physical conditions and substrates within the system however a few phyla tend to be more abundant regardless. Members of Firmicutes, Bacteroidetes and Proteobacteria often dominate in AD and can be considered as the leading decomposers, having prominent roles in all 3 bacterial stages of AD[40,45–48].

Aerobic cellulose hydrolyser genera such as Bacillus (Firmicutes) or Cytophaga (Bacteroidetes) secrete multiple extracellular enzymes, whereas anaerobic hydrolysers such as Clostridium (Firmicutes) and Bacteroides (Bacteroidetes) produce stable enzyme complexes tightly attached to the cell containing cellulases, xylanases and chitinases[40].

Metagenomic studies of AD provide evidence for members of Bacteroidetes and Firmicutes to also be key fermentative bacteria along with members of Proteobacteria[30,48]. Proteobacteria such as the *Acetobacter* and *Gluconacetobacter* are some of the most important acetate fermenters in biotechnology. The heterofermentation to propionic acid by the Proteobacteria *Acidipropionibacterium* (formerly *Propionibacterium*) has been the subject of research for a century, and members of Clostridium have been researched for their production of butyric acid[29,30]. Heterofermentation of amino acids is a metabolism only found in two phyla: Firmicutes and Synergistetes[49].

Members of Firmicutes, Proteobacteria, Bacteroidetes and Thermotogae are known acetogens [40,47]. Acetogens, or homoacetogens which only generate acetate, are obligate anaerobic bacteria and typically have a flexible metabolism, able to utilize a number of carbon sources[41]. For example, in high partial pressures of H_2 and low acetate concentrations they metabolise H2 and CO_2 into acetate, but at low partial pressures of H_2 and high acetate concentrations, acetogenesis is reversed producing H2 and CO_2 from acetate[50]. They are also able to use a variety of electron acceptors other than CO_2 often found in AD including fumarate and nitrate[41].

1.1.4.2 The Archaea + methanogenesis

The final step of AD in waste treatment, methanogenesis, is unique to a small number of highly specialised archaea: the methanogens, which form syntrophic relationships with fermentative bacteria. The archaea are phylogenetically distinct to both Eukaryotes and Bacteria, representing the 3rd, and most recently defined, domain of life[51,52]. The first prokaryotes assigned to the new kingdom of archaea were often isolates from extreme environments, such as acidic mud ponds, that had been previously thought incompatible with life[53,54]. Advances in culturing and sequencing technologies have proved this to be inaccurate. In fact, archaea have been found in a wide variety of niches, from the human gut[55] to the ocean[56], but are often much less abundant than bacteria and/or have highly specific requirements for culturing[39,45,56]. As a result, the archaeal domain is less understood than its counterparts.

Methanogens are typically strict anaerobes, able to grow at both mesophilic or thermophilic temperatures, and tend to have highly specialised and restricted energy metabolisms revolving around methanogenesis. They are phylogenetically diverse and split across 7 orders within the phyla Euryarchaeota[44]. 5 of the 7 orders utilize hydrogenotrophic methanogenesis while members of Methanosarcinales are able to use a broader spectrum of substrates, capable of utilizing hydrogenotrophic, acetoclastic and methylotrophic methanogenesis[44,57].

1.1.5 Anaerobic digestion: A summary

Energy yields from AD are much lower than the predicted energy present[58]. Large portions of this potential energy is either not readily available due to limiting steps of hydrolysis or fermentation, or converted to unusable levels of CO_2 disproportionate to that of H_2 for methanogenesis[10,39]. Bioaugmentation of the AD community with specific cellulose-degrading bacteria is one approach that has been investigated to increase yields and in vivo experiments show increases in pH can lead to increases of up to 697% in methane production, likely as a result of increased fermentation[10,39,59]. Acetogenesis and methanogenesis are both competitive and syntrophic steps in that they compete for H_2, but the acetate formed by acetogenesis is ultimately used in acetoclastic methanogenesis generating CO_2. This competition over H_2, and the use of acetoclastic rather than hydrogenotrophic methanogenesis, is one of the reasons for the mixed composition of biogas from AD and typically methane yields do not exceed 50-70% of total gas, with CO_2 making up a large percentage of the rest. The initial competition depends largely on H_2 uptake kinetics, where acetogens out compete methanogens for the uptake of H_2[50]. By augmenting AD systems with engineered hydrogenotrophic methanogens with higher H_2 uptake kinetics than acetogens, higher methane: CO_2 ratios in biogas could in theory be achieved.

Currently only a small number of organic waste streams are directed into AD and a large amount of organic matter is still sent to landfill. In the future it is likely AD will grow into a global scale process for reclaiming material, energy and nutrients. Interestingly the issues Cooney and Wise highlighted in 1975 limiting greater application of AD, system instability and long digester residence time, are still issues today 4 decades later[36]. Ultimately better understanding of the microbial communities and metabolic diversity in AD is needed for what in future will most likely be a vital process.

1.2 Metagenomics

It is estimated that over 99% of microorganisms are unculturable using classical methods, making whole genome sequencing and functional annotation impossible[60]. Until recently, the diversity of environmental samples was often measured using phylogenetic marker genes, and many bacterial and archaeal species are only known by their 16S rRNA gene sequence. In fact, studies of amplicon PCR of the 16S rRNA gene have shown the most abundant microorganisms in a sample can be unculturable and we can only guess at their metabolic contribution to the microbiome, hindering our understanding of the complexity of microbial communities [39,61–63] . The advent of high throughput sequencing in 1970s paved the way for PCR amplification of marker genes, while advances in next generation sequencing (NGS) and third generation sequencing (3GS) has allowed for the growth of a new discipline: metagenomics, where the total DNA of a sample is sequenced and the microbial community reconstructed computationally. Metagenomics provides an alternative to classical sequencing from pure culture and has the potential to yield near complete genomes allowing for functional analysis and increased understanding of individual contributions to the community metabolism[60].

1.2.1 -omics

Meta-omics, a discipline that includes metagenomics, metatranscriptomics, metaproteomics and metabolomics, attempts to view the microbial community as a whole, rather than the sum of its culturable parts. Metagenomics is the study of total DNA while metatranscriptomics, metaproteomics and metabolomics are the study of total mRNA, total protein and total metabolites respectively. While metagenomics, metatranscriptomics and metaproteomics is the study of a community of micro-organisms, metabolomics can also refer to the metabolites in a single species culture. Collectively they have the potential to provide a complete picture of the metabolic activities within a microbial community. Metagenomics may also provide better measures of microbial diversity, by negating primer bias that can occur with PCR amplification, and of relative abundance, by negating the varying copy numbers of target genes.

1.2.2 Assembly and binning

The reads from sequencing can be analysed directly using programs such as SSuMMo[64], assembled into contigs using programs such as Megahit[65], or binned into operational taxonomic units (OTU's) by programs such as COCACOLA[66] or MetaBAT[67]. Unassembled reads allow for the relative quantification of taxa or functional information, while assembly into contigs allows for the reconstruction of genes or genomes for more reliable phylogenetic or functional assignment. Binning, or clustering, aims to reconstruct either full or partial fragmented genomes from contigs and provides the best way of viewing a microbial community as a whole.

Individual sequencing reads are assembled into longer contigs using overlapping sequences, and the contigs can then be binned into clusters of contigs using a variety of parameters, but often relative abundance. Each cluster represents a single genome as an operational taxonomic unit, or OTU. However, the recovery of genomes from metagenomic data is a complex task and binning is an error prone process. Often contigs are falsely assigned to OTU's during the binning process leaving the user with a choice: either re-bin the contigs, ignore the data from those OTU's, or accept that their data is inaccurate.

1.2.3 Metabolic analysis

One of the largest advantages of shotgun metagenomic sequencing over 16S rRNA gene amplification and sequencing is the functional information gained that can be used to infer and assign metabolic pathways to individual species.

Pathway mapping is becoming an increasingly useful tool for metabolic analysis and a variety of databases are available to link individual genes into pathways. EcoCyc is an extensive database, for probably the most comprehensively studied organism in history, ideal for detailed pathway mapping in *Escherichia coli*[68]. It's sister site MetaCyc contains a repertoire of 2609 pathways from 2914 organisms designed for understanding metagenomic data, as well as a tool for the building of metabolic networks from annotated genomes[69]. Kyoto Encyclopaedia of Genes and Genomes (KEGG) also contains a pathway mapping database, KEGG Mapper, comprising of manually drawn diagrams built from published literature[70]. However, to analyse the metabolism of a novel genome using KEGG annotation using their web- based genome annotation services BlastKOALA or GhostKOALA is required[70,71]. Unfortunately, while comprehensive, these databases and pathways are often built only for specific species or specific purposes and use web-based servers, unpractical for use with large metagenomes from AD which can have hundreds or thousands of OTU's[1,40]. There is not a tool currently available that allows quick and automated mapping for large metagenomic data sets.

1.3 Aims

Anaerobic digesters have rich and complex communities consisting of hundreds of different potential species and many suspected intra-/inter-species interactions[39,45]. Low abundance species that are indiscernible during "normal" conditions can provide robustness to AD systems during environmental changes, and communities can vary greatly in taxonomic complexity between systems complicating attempts to reconstruct the AD microbiota in its entirety[1,46]. Often only a small number of relatively complete genomes are recovered from a metagenome sample, even in samples that are less taxonomically diverse than those in AD[62]. As a result, the microbiology behind AD communities remains relatively unknown.

In the face of ever increasing datasets, better tools and pipelines for the automated analysis of metagenomic data from AD are required. Towards this aim, a pipeline has been built for use with UNIX multicore workstations consisting of 5 custom Python scripts, CheckM[72] and Prokka[73]. The pipeline consists of a number of steps helping to identify near complete genomes, remove contamination and assign metabolic pathways to individual genomes.

2. Development of Tools and Materials

2.1 Multi-contig contamination removal tool

Metagenomic binning represents a “best guess” approach and the miss-binning of contigs is of particular concern when reconstructing genomes[63]. Before attempting to analyse the metabolic capabilities of an OTU, multi-species bins must be split into their component genomes and contamination removed. Genome completeness and contamination are typically measured using a number of universal single copy genes. CheckM provides information on genome completeness and contamination that could be used to identify contaminating contigs containing these genes and remove them into a separate bin[72]. However the single copy genes used for estimating completeness and contamination typically constitute less than 10% of genes and are unevenly distributed across the genome meaning that simply removing contigs identified by CheckM is not enough to ensure an uncontaminated OTU metabolically[63].

For the identification and removal of contaminating contigs the multi-contig contamination removal (MCCR) tool, (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/Multi-Contig_Contamination_Removal.py), a custom Python script utilizing both sequence composition and taxonomy, was designed.

GC content (%GC) is a compositional tool that has been used for many years in the analysis of DNA and genomes. Individual species have each evolved their own specific %GC, although we know little about why[74]. %GC does vary throughout a genome, sometimes due to the acquisition of genetic elements through horizontal gene transfer, but over long stretches of DNA %GC is relatively consistent. The same can be said for the frequencies of individual tetranucleotides as shown in Figure 2.1[75]. This makes both %GC and the frequencies of tetranucleotides (TNF) potentially good methods of identifying miss-binned contigs through sequence composition and these compositional statistics already feature in some binning software[66,75]. However, the contigs within a metagenome are not always of a length where the %GC or the TNF are representative of the genome as a whole, and therefore it is unwise to make binning decisions on sequence composition alone. For this reason, the MCCR tool uses sequence composition to identify potentially miss-binned contigs but uses taxonomy and alignments to known organisms to ultimately make the decision of where a contig belongs.

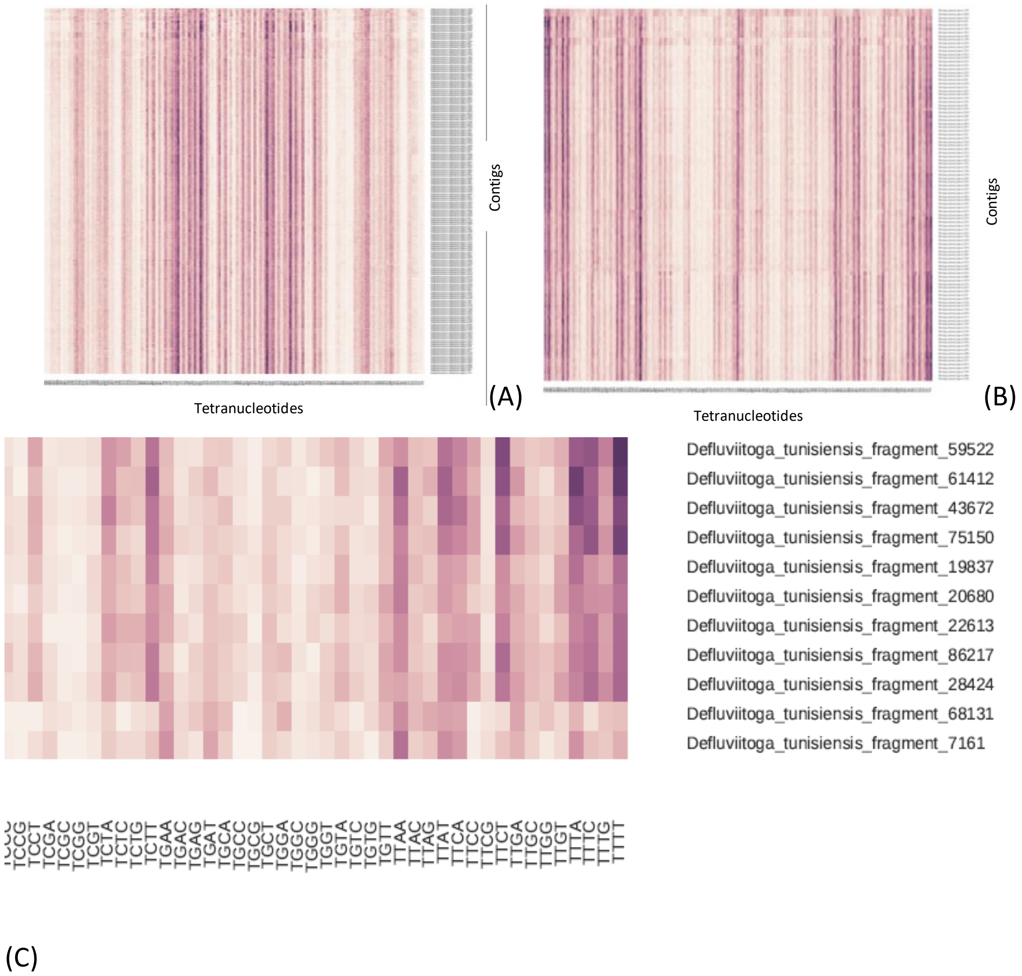


Figure 2.1 Example tetranucleotide frequency profiles

Tetranucleotide frequency profiles generated from clustered heatmaps of the tetranucleotide frequencies from 20 kb contigs of *Acidipropionibacterium acidipropionici* (A) and *Defluviitoga tunisiensis* (B)(C). Darker colours represent a higher frequency of a specific tetranucleotide within a contig, better shown in an expanded section of (B) in (C). Each contig has a specific tetranucleotide frequency pattern, which is similar to the other contigs of that genome but different to that of a different genome. *A. acidipropionici* has a %GC of 68.9 whereas *D. tunisiensis* has a %GC of 31.4. As such in (A) containing a GC-rich genome, the centre columns containing tetranucleotides starting with C and G tend to be darker than those in (B), while in (B), an AT-rich genome, the outer columns containing tetranucleotides starting with A or T tend to be darker than those in (A).

The MCCR tool takes an OTU in FASTA format generated during binning of contigs from a metagenome and builds any number of putative genomes from it depending on the level of contamination. An overview of the process can be found in Figure 2.2. The core script consists of 3 key steps:

Assignment of a putative phylum: The putative genome is assigned a putative phylum based on a blastn[76] search of the longest contig. The phylum is assigned from the blastn result using searches of the genus name against 53 phylum TXT files adapted

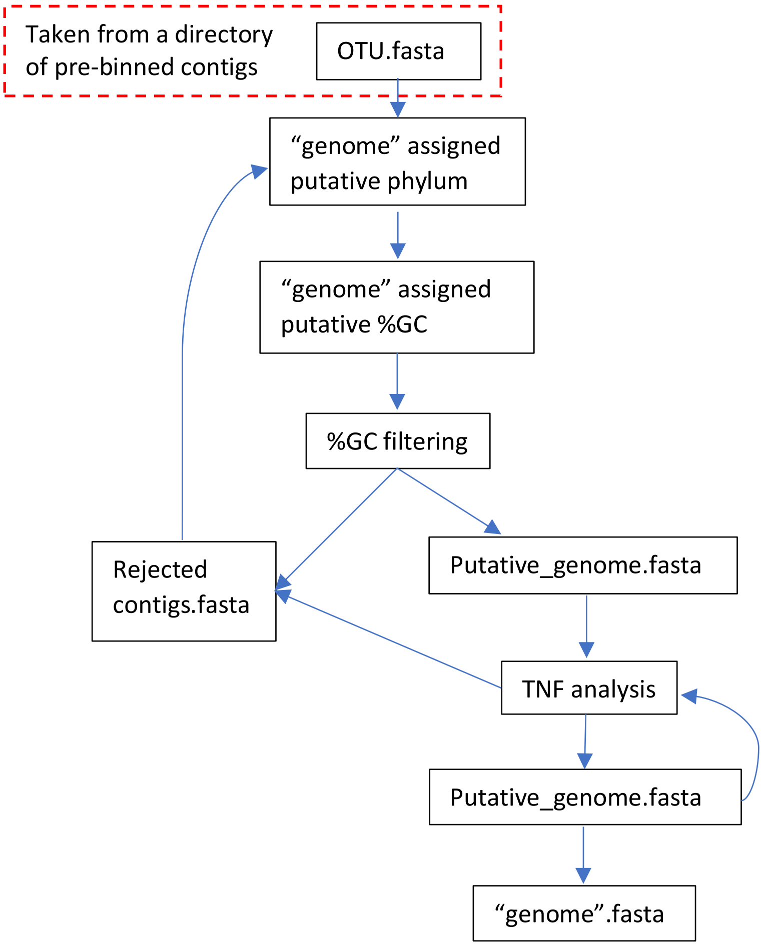


Figure 2.2 Overview of the MCCR tool to remove contamination

Each OTU of pre-binned contigs is assigned a putative phylum and putative %GC. Contigs with a dissimilar %GC to that assigned to the OTU are analysed and either accepted as part of the new putative genome, or rejected and removed. The TNF of each contig is calculated, and those with a TNF profile least similar to the rest are analysed in a looping fashion until all the contigs can be assigned to the putative phylum via blastn searches, creating an uncontaminated genome. All those contigs that have been rejected now act as a new OTU to be analysed, looping until all the contigs can be assigned to an uncontaminated genome.

from NCBI taxonomy browser[77], each containing lists of all the known genera within that phylum.

Filtering by %GC: The putative genome is assigned a putative %GC based on the %GC of the longest contig. Any contig with a %GC too dissimilar to that of the longest contig are identified for taxonomic analysis using the top 10 hits (arbitrarily chosen) from a blastn search of the contig. If the assigned putative phylum of the genome appears within the phyla of the top 10 hits, the contig is accepted as part of the genome, otherwise it is rejected.

Tetranucleotide frequency analysis: The TNF percentages of the remaining contigs are calculated and clustered on a heatmap as in Figure 2.1. The 10 (arbitrarily chosen) least similar contigs based on this TNF percentage clustering are identified for taxonomic analysis as described above. If a rejected contig is found in the outliers, it is removed, the contigs re-clustered and the new outlying 10 analysed iteratively until all 10 are accepted.

After analysis, two FASTA files are generated from the original OTU, one containing all the contigs that have been rejected from the OTU and one containing the contigs that have not. The FASTA file of rejected contigs then feeds back into the script as an OTU to be analysed, and this continues until all contigs have been built into a putative genome.

Numerous versions of the MCCR tool were built during development in order to maximise the efficiency and accuracy with which the contigs were binned through this script.

2.1.1 Testing

To test the accuracy of this tool in identifying and removing contamination and in splitting OTU's of contigs from different phyla/species, mock community 1 was created. 11 randomly chosen genomes of varying %GC and phyla were downloaded from Genbank[78] (Table 2.1(A)). These were split into contigs of varying length using a custom python script, Contig_Creator.py (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/ContigCreator.py), to better model data from metagenome datasets. Contig_Creator.py split whole genomes into contigs modelling metagenomic genomes with a bias towards shorter contigs, as these are typically harder to bin and analyse correctly. Contig lengths started at arbitrarily chosen 2 kb, and every 10th contig the contig length would increase by 80 bp, creating a wide range of contig lengths from a single genome. Genomes split into contigs were then combined into 15 highly contaminated synthetic OTU's such that each OTU contained 2 full genomes and the genomes were from different phyla (Table 2.1(B)). Since the MCCR tool relies on the sequence composition differences between contigs and genomes, a variety of differences in %GC between the two genomes were created. 3 synthetic OTU's for each difference in %GC, 0-1.5%,1.5-5%,5-10%,10-15%, and 15-20% were created (Table 2.1(B)). Since they are known organisms, to prevent alignments with the original genome in the NCBI database, a negative GI list was used containing the GI numbers of all samples from metagenome datasets and from the organisms in the synthetic dataset.

Kingdom	Phylum	Species Name	GenBank Reference	%GC
Bacteria	Actinobacteria	*Acidipropionibacterium acidipropionici*	GCA_000310065.1_ASM31006v1	68.9
Bacteria	Synergistetes	*Cloacibacillus evryensis*	GCA_000585335.1_ASM58533v1	56.0
Bacteria	Thermotogae	*Defluviitoga tunisiensis*	GCA_000953715.1_DTL3	31.4
Bacteria	Chloroflexi	*Dehalococcoides mccartyi*	GCA_000009025.1_ASM902v1	47.0
Bacteria	Bacteroidetes	*Leadbetterella byssophila*	GCA_000166395.1_ASM16639v1	40.4
Archaea	Euryarcheota	*Methanobrevibacter ruminantium*	GCA_000024185.1_ASM2418v1	32.6
Archaea	Euryarcheota	*Methanoregula formicica*	GCA_000327485.1_ASM32748v1	55.2
Archaea	Euryarcheota	*Methanothermobacter wolfeii*	GCA_900095815.1_SIV6	48.9
Bacteria	Proteobacteria	*Orrella dioscoreae*	GCA_900089455.2_OrrDiv2	67.4
Bacteria	Firmicutes	*Paenibacillus borealis*	GCA_000758665.1_ASM75866v1	51.4
Bacteria	Aquificae	*Thermocrinis albus*	GCA_000025605.1_ASM2560v1	46.9

(A)

Synthetic OTU	Genome 1	Genome 2	Difference in %GC
A	*Defluviitoga tunisiensis*	*Methanobrevibacter ruminantium*	1.2 (0-1.5%)
B	*Cloacibacillus evryensis*	*Methanoregula formicica*	0.8 (0-1.5%)
C	*Thermocrinis albus*	*Dehalococcoides mccartyi*	0.1 (0-1.5%)
1A	*Paenibacillus borealis*	*Methanoregula formicica*	3.8 (1.5-5%)
1B	*Paenibacillus borealis*	*Cloacibacillus evryensis*	4.6 (1.5-5%)
1C	*Orrella dioscoreae*	*Acidipropionibacterium acidipropionici*	1.5 (1.5-5%)
2A	*Defluviitoga tunisiensis*	*Leadbetterella byssophila*	9.0 (5-10%)
2B	*Cloacibacillus evryensis*	*Methanothermobacter wolfeii*	7.1 (5-10%)
2C	*Dehalococcoides mccartyi*	*Cloacibacillus evryensis*	9.0 (5-10%)
3A	*Cloacibacillus evryensis*	*Acidipropionibacterium acidipropionici*	12.9 (10-15%)
3B	*Paenibacillus borealis*	*Leadbetterella byssophila*	11 (10-15%)
3C	*Leadbetterella byssophila*	*Methanoregula formicica*	14.8 (10-15%)
4A	*Defluviitoga tunisiensis*	*Methanothermobacter wolfeii*	17.5 (15-20%)
4B	*Leadbetterella byssophila*	*Cloacibacillus evryensis*	15.6 (15-20%)
4C	*Paenibacillus borealis*	*Acidipropionibacterium acidipropionici*	17.5 (15-20%)

(B)

Table 2.1 Building of a synthetic mock community

11 complete genomes were downloaded from Genbank (A) and combined into 15 highly contaminated synthetic OTU’s (B), each OTU containing 2 full genomes, based on the differences in %GC between their genomes to create a variety of testing conditions. All OTU’s fit in one of 5 difference in %GC brackets: 0-1.5%, 1.5-5%, 5-10%, 10-15% and 15-20%, and all 5 difference in %GC brackets had 3 representatives.

2.1.2 Development of the script

2.1.2.1 Assignment of a putative phylum

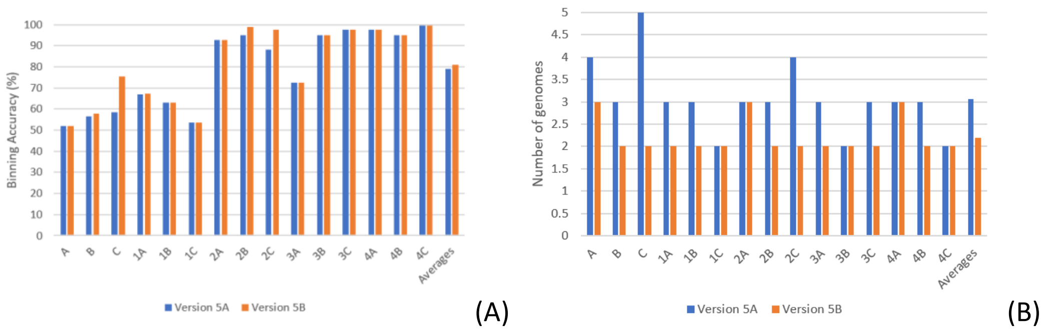


Figure 2.3 Comparing Version 5A and Version 5B of the MCCR tool
Comparison of binning accuracy (A) and number of genomes (B) created from each OTU in mock community 1 between Version 5A and Version 5B of the MCCR tool. In Version 5A, the putative phylum of a genome was calculated using a blastn search of the longest contig within an OTU, regardless of the result. In Version 5B, the blastn search had to return a result with an evalue of 0 for the contig to be used for assigning a putative phylum, else the next longest contig was used until an appropriate result was found.

The longest contig in an OTU is assumed to create the most reliable alignment through its blastn search of the NCBI database and therefore can be most reliably assigned the correct taxonomy. Because of this it is used as an anchor point in which to build the new "genome" around from a contaminated OTU. Initial measurements of accuracy were promising. Over the 15 samples, Version 5A was able to correctly bin 79% of contigs accurately (Figure 2.3(A)). However unnecessary numbers of genomes were created, an average of 3, due to genomes being created around less reliable BLAST results (Figure 2.3(B)).

For example, OTU 2C generated 4 "genomes" rather than the 2 it should. The first genome was built around a contig from *C. evryensis*, correctly assigned to Synergistetes from a blastn result with an evalue of 0. The second genome was built around a second contig from *C. evryensis*, incorrectly assigned to Proteobacteria from a blastn result with an evalue of 0.092. The third genome was built around a contig from *D. mccartyi*, correctly assigned to Chloroflexi from a blastn result with an evalue of 0. The forth genome was built around a contig that was unable to be assigned a phylum, as well as having an evalue of 0.049 from its blastn result. Although the contigs were binned with 88% accuracy for this OTU, the generation of so many genomes is misleading as to how many organisms were present in the OTU.

To prevent the generation of misleading numbers of genomes, in Version 5B, for a contig to act as an anchor point, the contig must fulfil the condition that the BLAST result must have an evalue of 0. If the condition is not met, the next longest contig is used iteratively until a result that meets the condition is found. By adding this condition, it prevents new genomes being generated around contigs that can't be firmly assigned to an existing species. This increases accuracy, measured using Equation 2.1, from 79% to 81% and reduces the number of genomes generated from a sample from 3 to 2.2 (Figure 2.3).

$$\text{Accuracy} = \frac{\text{Total number of correctly binned contigs}}{\text{Total number of contigs within an OTU}} *100$$

Equation 2.1

If an anchor point can't be found, or there is only a single contig left that hasn't been assigned to the genomes previously built, the remaining contigs will be placed in a FASTA file of unassigned contigs.

From these initial results in Figure 2.3(A) it is obvious that both Version 5A and Version 5B finds OTU's where the difference in %GC is largest easier to separate more accurately. On average there was a large jump in accuracy once the difference in %GC was greater than 5%. For Version 5B those with a difference of less than 5% in %GC had an average accuracy of 62% whereas those with a difference of greater than 5% in %GC had an average accuracy of 94%.

2.1.2.2 Size vs %GC

To measure the effect of contig size on binning accuracy and to assess whether contig length needs to be a considered variable when binning, 5 additional mock communities were generated using contigs of a specific size. Each genome from Table 2.1(A) was split into contigs of length 2kb, 4kb, 6kb, 10kb and 20kb using a slight variation of Contig_Creator.py and combined as in Table 2.1(B), generating 75 OTU's of varying differences in %GC and contig length.

Averages for the 3 synthetic OTU's in each difference in %GC bracket were taken and compared across the different contig lengths in Table 2.2. The binning accuracy increases slightly on average in response to longer contigs, however binning accuracy was only increased by 4% in response to a 10-fold increase in length between 2kb and 20kb contigs. As can be seen in Figure 2.4(A), when contig length is plotted against accuracy, length doesn't appear to have a clear effect on accuracy in comparison to Figure 2.4(B) where the

		Length of Contigs					
		2 kb	4 kb	6 kb	10 kb	20 kb	Average
Difference in %GC	0-1.5%	64%	53%	59%	77%	75%	65%
	1.5-5%	81%	85%	74%	82%	88%	82%
	5-10%	82%	82%	91%	90%	95%	89%
	10-15%	93%	96%	97%	89%	99%	95%
	15-20%	95%	97%	96%	98%	98%	97%
	Average	83%	84%	83%	84%	87%	

Table 2.2 Binning accuracies of Version 5B: %GC vs contig length

Analysing the effect of contig length on binning accuracy across differing difference in %GC brackets using Version 5B. Mock communities 2-6 were created as in Table 2.1(B) using contigs of specific sizes: 2 kb, 4 kb, 6 kb, 10 kb and 20 kb respectively to assess if contig length effects binning accuracy.
Since it has previously been established that differences in %GC affect binning accuracy, averages of accuracy scores for each difference in %GC bracket containing 3 representatives are shown and compared across 5 contig lengths. Scores of >90% accuracy are highlighted in yellow.

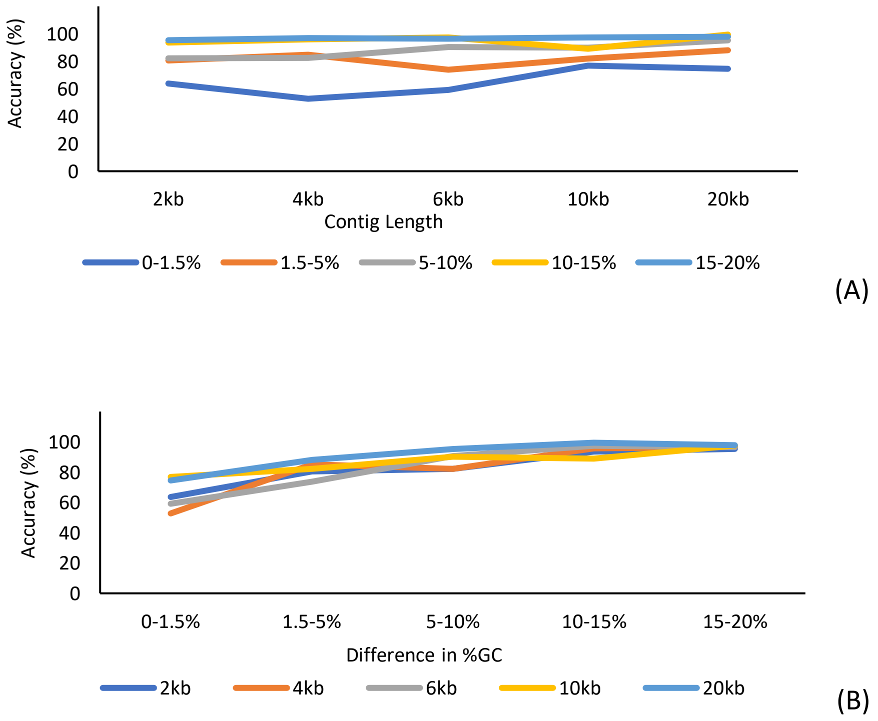


Figure 2.4 Comparative graphs of the effect of contig length (A) and difference in %GC (B) on binning accuracy using Version 5B.

Graphical representation of data taken from Table 2.2. 5 additional mock communities were created as in Table 2.1(B) using contigs of specific sizes: 2 kb, 4 kb, 6 kb, 10 kb and 20 kb respectively to assess if contig length effects binning accuracy across the 5 difference in %GC brackets. (A) Binning accuracy of the 5 differences in %GC against contig length using Version 5B (B) Binning accuracy of 5 different contig lengths against differences in %GC using Version 5B.

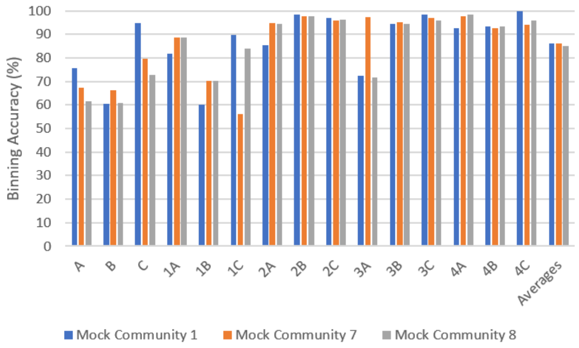


Figure 2.5 Binning accuracy of Version CE2 in respect to contig length

Mock community 7 and 8 were created as in Table 2.1(B) containing variable but increasingly smaller average contig lengths compared to mock community 1.

difference in %GC between the two genomes has a clear effect on binning accuracy regardless of contig length.

An additional 2 mock communities, mock community 7 and 8, were created using contigs of varying but increasingly smaller lengths rather than contigs of all the same length to further analyse the effect on contig length against accuracy using Version 5CE2. The lengths of contigs of mock community 7 and 8 were ~65% and ~50% of those in mock community 1 respectively. As shown in Figure 2.5, on average accuracy only marginally decreased with decreasing average contig length. Even using a later and more accurate version of the MCCR tool and using mixed length contig OTU's, average contig length still has little effect on accuracy.

2.1.2.3 Development of GC filter

By calculating the likely %GC content of one species in a OTU using the longest contigs, particularly different contigs from potentially different species, can be easily identified for further analysis. Initially in Version 5A and Version 5B the %GC filter was the longest contig %GC +/- 5% such that any contig with a %GC higher than the %GC of the longest contig + 5% or less than the %GC of the longest contig – 5%, was analysed. However different combinations of genomes within mock community 1 differ in the range of %GCs the contigs produce. The %GC of contigs from OTU's generated from genomes with a 0-1.5% difference were only spread across 26%, whereas those from a 15-20% were spread across 35%.

OTU's generated from genomes that are more similar in their %GC will have far more of their contigs within that 10% window that aren't analysed as highlighted in Figure 2.6(A). This one of the reasons why contigs from those with only 0-1.5% difference in %GC are

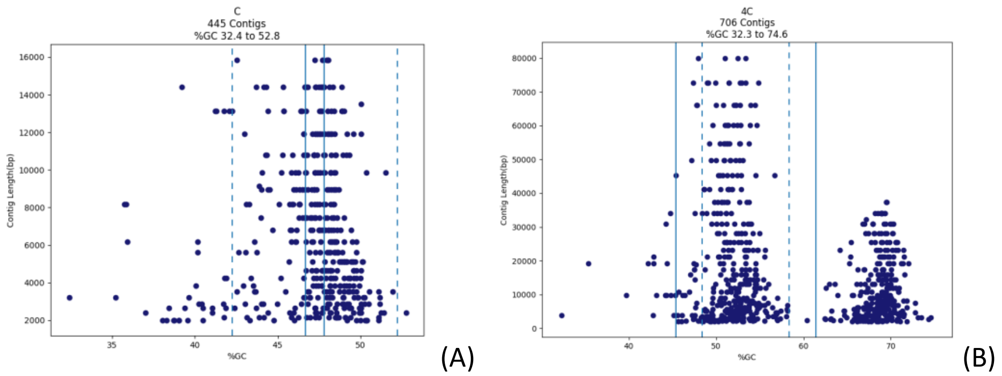


Figure 2.6 Graphical comparison of %GC filter between Version 5A/B and Version 5CE2.

A graphical representation of the distribution of contigs across length and %GC for OTU's C(A) and 4C(B) showing the upper and lower bounds of the %GC filter of the MCCR tool. The upper and lower bounds of Version 5A/B are shown using dashed lines, while those of Version 5CE2 are shown with solid lines. For contigs outside of the lines, blastn is used to assign a taxonomy and assess whether the contig belongs in the genome. (A) OTU C has a difference in %GC of 0.1 and a range of %GC of 20.3 The use of Version 5CE2, rather than Version 5A/B, increases the number of contigs analysed. (B) OTU 4C has a difference in %GC of 17.5% and range of %GC of 42.3. The use of Version 5CE2, rather than Version 5A/B, slightly decreases the number of contigs analysed.

binned ~30% less accurately than those with a 15-20% difference as shown in Table 2.2.

Changing the %GC filter to reflect the differences in the total range of %GCs in an OTU seemed more appropriate. To identify a better method of identifying contigs based on %GC, the %GC filter was changed to be +/- either 5%, 10%, 20% or 40% of the %GC range, calculated as in Equation 2.2.

%GC range = highest %GC – lowest %GC of the OTU

Equation 2.2

Overall the most accurate %GC filter proved to be using +/-5% of the %GC range (Figure 2.7(A)), but this created an additional problem. Speed has not been measured as part of the development of this tool, however the most computationally exhaustive part of this script is the use of blastn even when using a local database rather than web interface. Therefore, for these analyses speed is roughly inversely equal to the number of alignments that take place. When using +/- 5% of the %GC range, the number of alignments that took place was much higher, making the script quite slow. Steps towards ensuring the script was more efficient had been taken in creating a TXT file of all the blastn results that could be searched through for the results first, before any additional alignments took place,

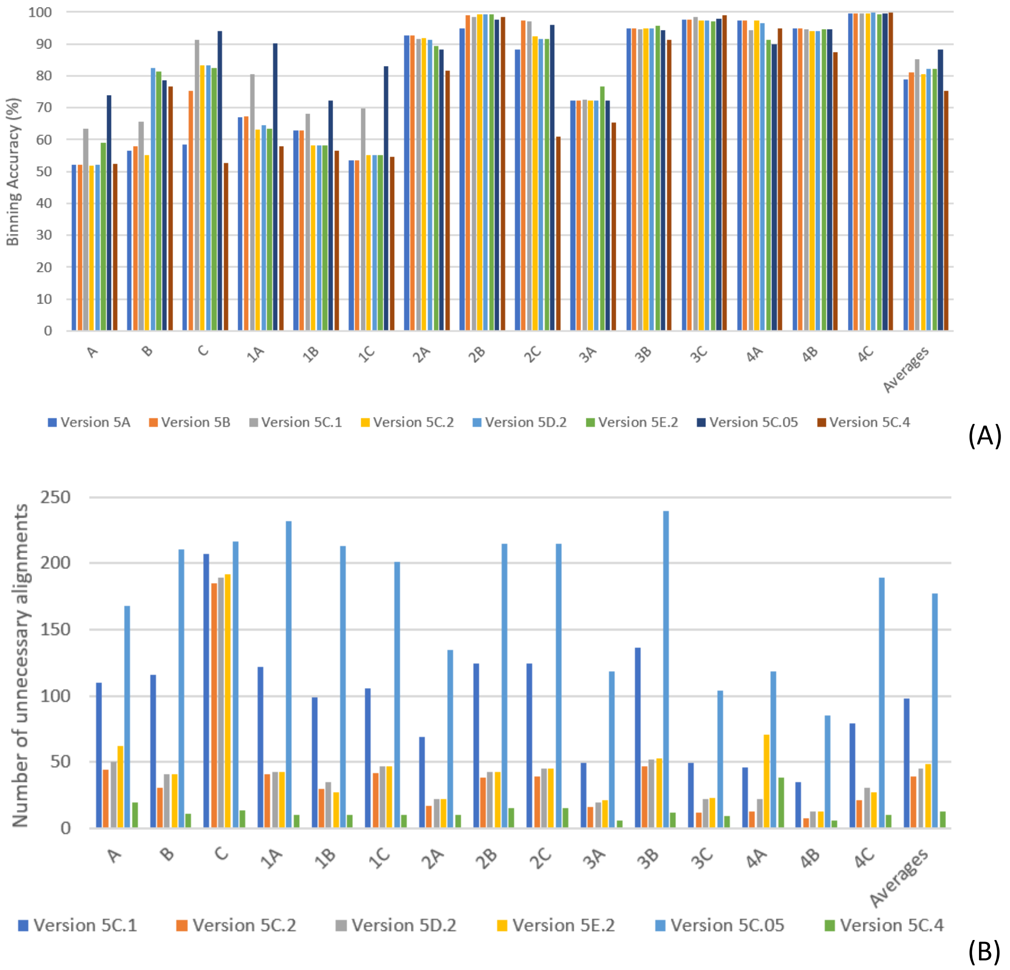


Figure 2.7 Comparison of binning accuracies and efficiencies
A number of different percentages of the %GC range were used for development of the %GC filter. 5%, 10%, 20% and 40% for 5C.05, 5C.1, 5C.2 and 5C.4 respectively. Binning Accuracy (%) denotes the accuracy with which contigs are binned as a result of that %GC filter, while number of unnecessary alignments is the number of BLASTS that had no effect on the outcome of the accuracy (%).

preventing anything needing to be analysed twice. But still a balance needed to be struck between accuracy and efficiency.

The percentage of the %GC range that gave the highest accuracy for each synthetic OTU was taken and compared to the %GC range. A scatter graph of preferred percentage of %GC range vs %GC range indicated that the smaller the %GC range, the smaller the percentage of the %GC that should be used as shown in Figure 2.8.

Two exponential functions were devised using a trendline so that where the %GC range is very large or small, the %GC filter reflects this, creating a balance between efficiency and accuracy. These were the longest contig %GC +/- 0.07*e(0.13*%GC range) in Version 5CE and the longest contig %GC +/- 0.04*e(0.14*%GC range) in Version 5CE2.

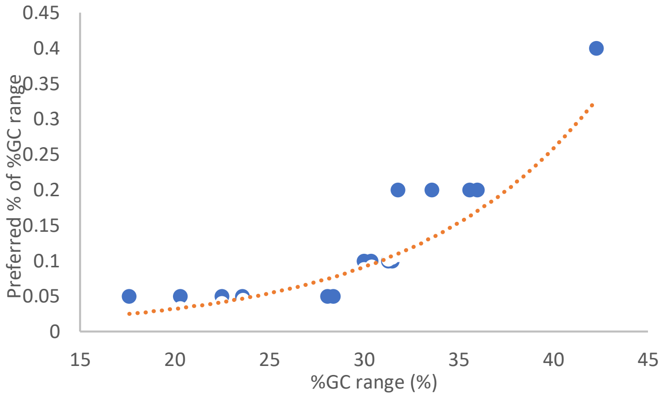


Figure 2.8 Calulating the exponential function used in the %GC filter of Versions 5CE and 5CE2

The %GC filter was changed so that it was calculated using a percentage of the %GC range of an OTU highlighted in Figure 2.4. Plotting of the %GC range against the % of the %GC range that gives the highest result from Table 2.5 finds an exponential relationship used in the %GC filter of Version 5CE and Version 5CE2 of the MCCR tool.

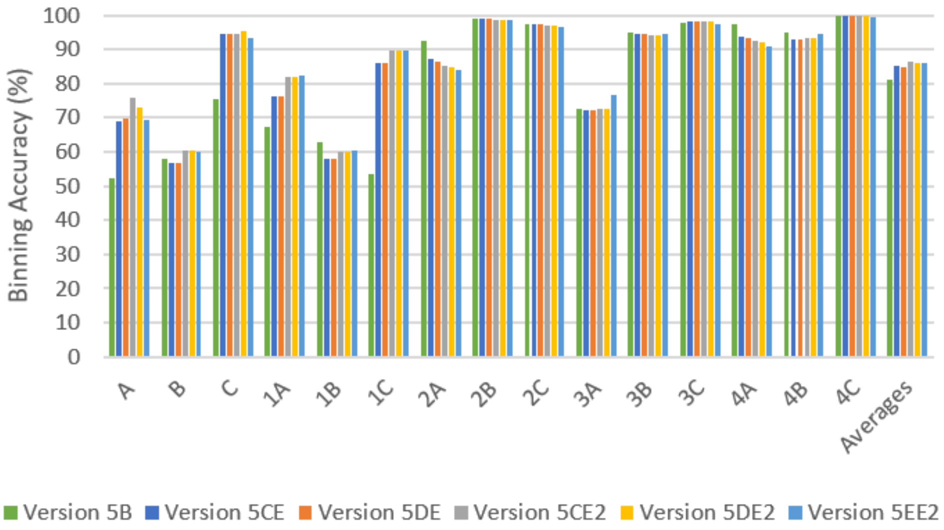


Figure 2.9 Development of the MCCR tool

An additional 6 versions of the MCCR tool were created and their binning accuracies measured. 5CE and 5CE2 use two different exponential functions for the %GC filter, while 5DE2 explores increasing the depth of the TNF analysis to 20 contigs, and in 5DE2-D removing the TNF step entirely. 5EE2 explores setting a minimum evalue of 1e-05 for all blastn analysis.

By using an exponential function, the average accuracy was increased by 4% to 85% using Version 5CE and by 5% to 86% using Version 5CE2 (Figure 2.9). The exponential function had a marked increase on OTUs with a difference of less than 5% in %GC, increasing average accuracy of 62% using Version 5B to 77% using Version 5CE2.

2.1.2.4 Development of TNF profiling

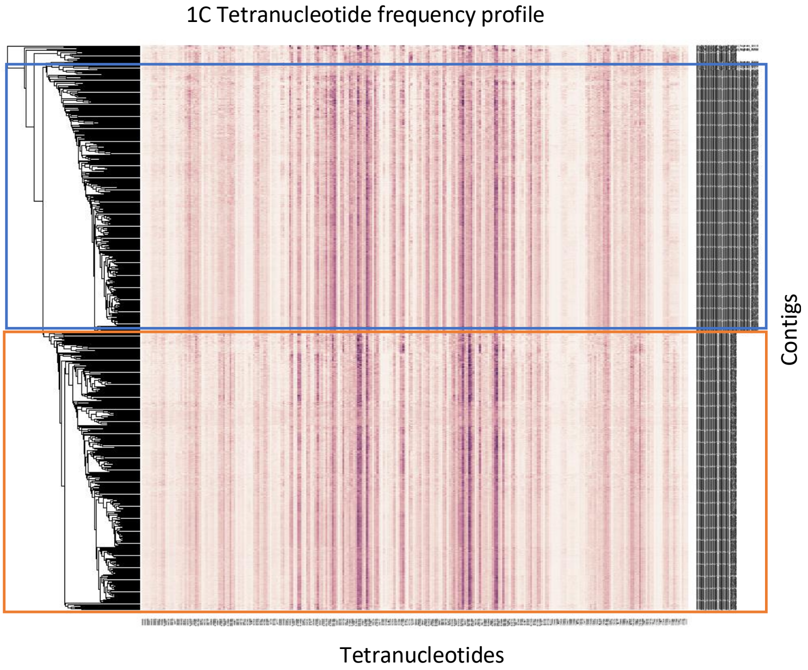


Figure 2.10 TNF profiles highlight compositional differences in sequences between highly similar genomes.

The TNF profile of OTU 1C with a 1.5 difference in %GC between the %GC of *A. acidipropionici* and *O. dioscoreae* genomes within it. The contigs from *A. acidipropionici* (highlighted in blue) and *O. dioscoreae* (highlighted in orange) although very similar in terms of %GC clearly separate according to the TNF profiles of their contigs.

In addition to the %GC filter, a second measure of sequence composition is used: the percentage frequency of tetranucleotides within a given contig. Similar to the %GC filter, it is based on the assumption that the majority of a genome will have a similar percentage frequency across the 256 tetranucleotide combinations[66,75]. This is far more sensitive than using %GC content alone for analysis, and can be used to separate species that have similar

%GC, but have different TNF profiles, shown in Figure 2.10. TNF is calculated as a percentage, the frequency of a given tetranucleotide divided by the total length of the contig.

Initially the 10 least clustered contigs, or more specifically the top 5 and bottom 5 contigs of the clustered heatmap, are analysed. To see if increasing or decreasing the number of contigs analysed in this way increased accuracy, the top 10 and bottom 10 contigs were analysed in Version 5DE2 and this step was removed entirely in Version 5DE2-D (Figure 2.9). Surprisingly by increasing the number of contigs taken, the average accuracy marginally decreased. Comparing Version 5CE2 and Version 5DE2, the accuracy of only 4 synthetic OTU's increased when increasing the number of contigs analysed by this step, 3 decreased while 8 had no change in accuracy. Unsurprisingly, removing this step altogether also resulted in a decrease in accuracy to 85% and it is better to use both %GC and TNF to identify potentially contaminating contigs.

Finally, in Version 5EE2 a minimum evalue was placed on the blastn results from both the %GC filtering and TNF analysis to ensure higher quality blastn results. A minimum evalue of 1e-05 resulted in a marginal decrease in binning accuracy (Figure 2.9).

2.1.3 Discussion

The MCCR tool version with the highest average binning accuracy developed was Version 5CE2 at 86.23%. In this version a contig must have a blastn evalue of 0 to act as an anchor point for building a genome from an OTU, the %GC filter is calculated using the longest contig %GC +/- 0.04*e(0.14*%GC range), 10 contigs are used for each loop of the TNF analysis, and there is no minimum evalue required for the blastn results from either the %GC filter or TNF analysis.

As shown in Table 2.5(A), on average the MCCR tool is far more effective at splitting synthetic OTU's with a difference in %GC of > 5% than < 5%, likely due to the difficulties with identifying contigs in a sample that is made of sequences that are compositionally similar. A rough estimate of potential accuracy of the script and synthetic OTU's was calculated by using the MCCR tool on each individual genome split into contigs before combining into synthetic OTU's and taking an average of the two genomes that constitute each synthetic OTU. This estimate is not OTU specific and doesn't take into account the possibility of the blastn results of contigs from one genome containing the phyla of the opposing genome, and only calculates the percentage of contigs that would be rejected from their own genome. The calculated potential accuracies vs current accuracies shown in Table 2.5(B) highlight 4 OTU's where the accuracy of the script is far less that what it could be. 3 of these OTU's were in the <5% difference in %GC bracket. The 4th, 3A, is low because the blastn results from a large amount of *C. evryensis* contigs have various Actinobacteria species in them, and so the contigs are kept with the rest of the *A. acidipropionici* genome.

The issue of genetic material showing high levels of similarity to that of another phylum is an issue that could be quite prevalent in genomes that have aquired genetic elements through horizontal gene transfer. These additional pieces of DNA take time to acclimatise to their new host genomes's %GC/TNF, and so would likely be singled out using these compositional tools and potentially not binned with the genome with which it truly belongs

Difference in %GC (%)	Potential Accuracy (%)	Current Accuracy (%)	Current - Potential Accuracy (%)
0-1.5	93	77	-16
1.5-5	99	77	-22
5-10	94	94	0
10-15	98	88	-10
15-20	95	95	0

(A)

OTU	Potential Accuracy (%)	Current Accuracy (%)	Current - Potential Accuracy (%)
A	85	76	-9
B	97	61	-36
C	97	95	-2
1A	99	82	-17
1B	98	60	-38
1C	100	90	-10
2A	88	85	-3
2B	98	99	1
2C	96	97	1
3A	98	72	-26
3B	99	94	-5
3C	98	98	0
4A	89	93	4
4B	97	93	-4
4C	100	100	0
Average	96	86	-10

(B)

Table 2.3 Potential vs current binning accuracies of the MCCR tool

(A) Averages of potential and current accuracies are taken for each difference in %GC bracket highlighting areas for improvement.

(B) Potential and current accuracies of each OTU in mock community 1. Potential accuracies are calculated by determining the percentage of all contigs in an OTU that return results for their own phylum. In theory this is the maximum binning accuracy the MCCR tool could achieve. Current accuracy is the current binning accuracy of the tool. OTU's with current accuracies of < 75% are highlighted in yellow.

resulting in further genome fragmentation. Although this is an area that will need some investigation in the future, it is important to remember that the MCCR tool was designed to be used on OTUs that would otherwise be unuseable rather than as a preferred binning method.

Kingdom	Species Name	Phylum	# genera within the phylum	# species within the genus
Bacteria	*A. acidipropionici*	Actinobacteria	424	6
Bacteria	*C. evryensis*	Synergistetes	17	4
Bacteria	*D. tunisiensis*	Thermotogae	14	1
Bacteria	*D. mccartyi*	Chloroflexi	38	39
Bacteria	*L. byssophila*	Bacteroidetes	421	2
Archaea	*M. ruminantium*	Euryarcheota	120	112
Archaea	*M. formicica*	Euryarcheota	120	19
Archaea	*M. wolfeii*	Euryarcheota	120	14
Bacteria	*O. dioscoreae*	Proteobacteria	889	1
Bacteria	*P. borealis*	Firmicutes	587	~4170
Bacteria	*T. albus*	Aquificae	14	9

Table 2.4 Taxonomy statistics for each genome of mock community 1 according to the NCBI taxonomy browser[77]

Both A and 2A have relatively low accuracies, both potential and current, because of *D. tunisiensis*. The MCCR tool bins the *D. tunisiensis* genome alone with only 78% accuracy, compared to the >90% for the others, meaning only 78% of its contigs have alignments to species within Thermotogae. Both this low score and that from 3A with *C. evryensis* is likely a result of how limited the diversity of the phyla and genera are and their current under-representation in the tree of life of cultured organisms. *C. evryensis* belongs to Synergistetes, with only 17 genera within the phylum and only 4 species in *Cloacibacillus*. *D. tunisiensis* belongs to Thermotogae with only 14 genera and is the only species within *Defluviitoga* (Table 2.6). Attempts were made to circumvent this problem by introducing greater flexibility around the blastn e-value condition by allowing alignments with up to a maximum of 1e-05. In the case of A and 2A this instead decreased the binning accuracy by up to ~6%, although it is valuable to note that a decrease in accuracy was not shown for all synthetic OTU's and as such should perhaps be a flexible and user defined parameter. These results highlight a very specific problem with using taxonomy to bin contigs, in that it works well if the genome belongs to a phylum that is very well represented such as Proteobacteria, but it will always be less accurate if the phylum is smaller like Synergistetes or Thermotogae.

2.2 Metabolic pathway prediction tool

In order to quickly find genomes with pathways of interest in a metagenome the metabolic pathway prediction (MPP) tool, written in Python, was created (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/Metabolic_Pathway_Prediction.py). This tool uses gene names from a number of manually built

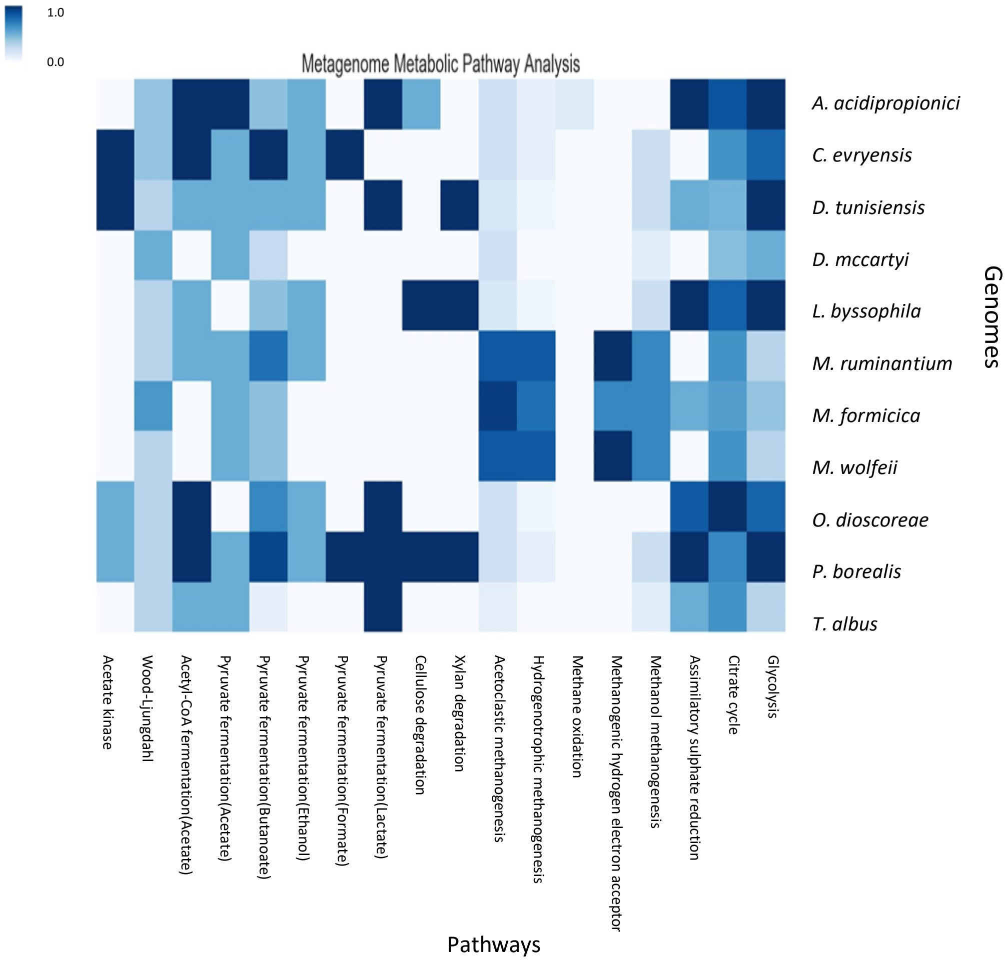
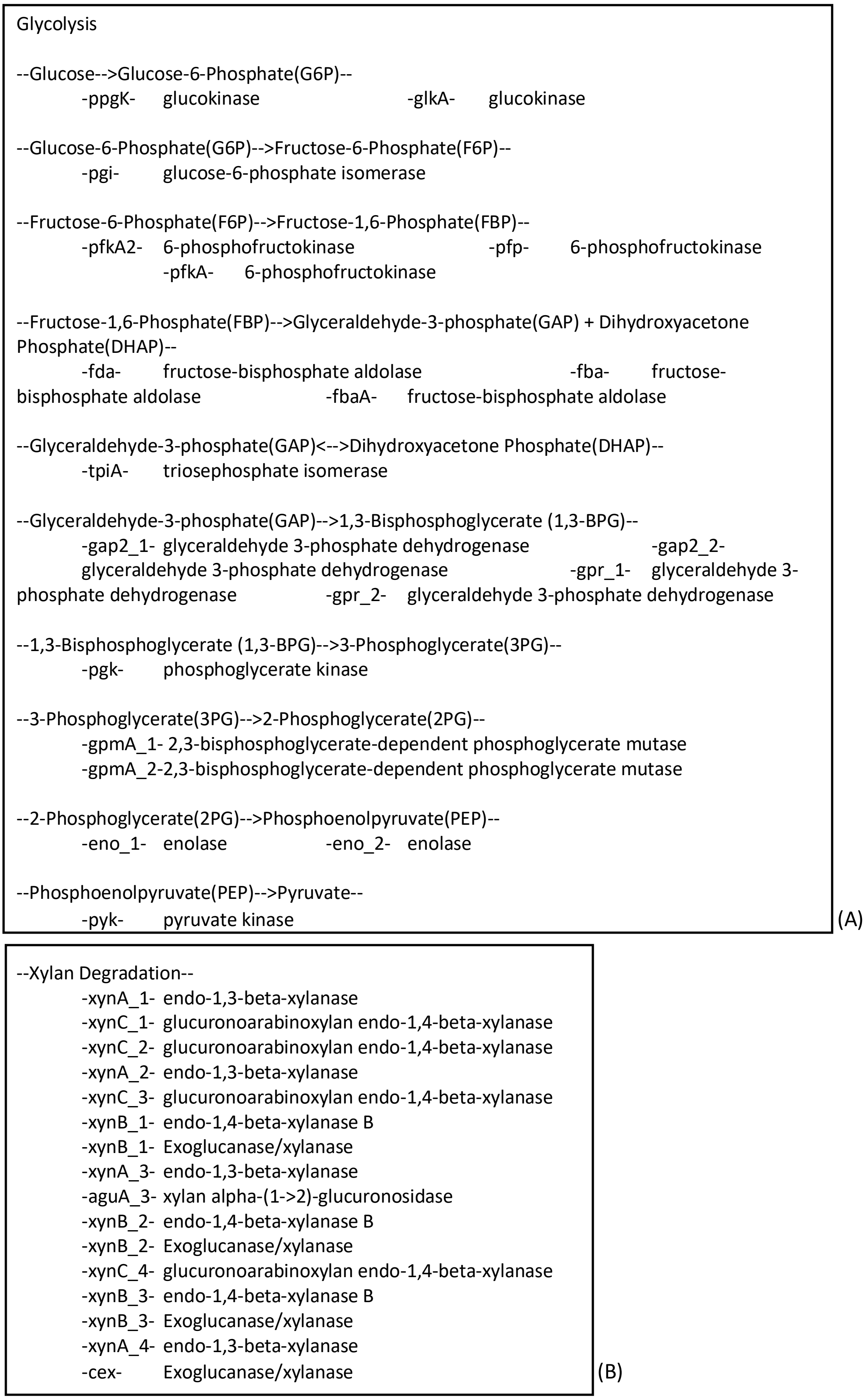


Figure 2.11 MPP heatmap of known organisms

A heatmap of the given metabolic pathways against each genome, where darker colours indicate more complete pathways. This provides an easy method of searching for pathways of interest within a metagenome

pathway files each containing a metabolic pathway and its enzymes. These were adapted from KEGG[70], MetaCyc[69] and literature searches[57] and are used in combination with GBF files generated from genome annotation using Prokka[73]. By searching for the enzyme name, or names, involved in each step of a given pathway within an annotated genome, an estimated measure of pathway completeness can be given and shown on a heatmap where the darkest colours indicate more complete pathways (Figure 2.11). In this way, by searching up the column of a pathway in the heatmap, genomes containing that pathway can be easily identified. The heatmap itself is built from a tab delimited TSV file containing all the numerical pathway percentages, which can be used instead of or in combination with the heatmap.

The MPP tool is designed to give a general overview of a metagenome through its heatmap, however it also generates a 3rd type of output: metabolism files for each



```
Glycolysis

--Glucose-->Glucose-6-Phosphate(G6P)--
        -ppgK-   glucokinase                -glkA-    glucokinase

--Glucose-6-Phosphate(G6P)-->Fructose-6-Phosphate(F6P)--
        -pgi-    glucose-6-phosphate isomerase

--Fructose-6-Phosphate(F6P)-->Fructose-1,6-Phosphate(FBP)--
        -pfkA2-  6-phosphofructokinase             -pfp-    6-phosphofructokinase
                 -pfkA-   6-phosphofructokinase

--Fructose-1,6-Phosphate(FBP)-->Glyceraldehyde-3-phosphate(GAP) + Dihydroxyacetone
Phosphate(DHAP)--
        -fda-    fructose-bisphosphate aldolase              -fba-     fructose-
bisphosphate aldolase          -fbaA-    fructose-bisphosphate aldolase

--Glyceraldehyde-3-phosphate(GAP)<-->Dihydroxyacetone Phosphate(DHAP)--
        -tpiA-   triosephosphate isomerase

--Glyceraldehyde-3-phosphate(GAP)-->1,3-Bisphosphoglycerate (1,3-BPG)--
        -gap2_1- glyceraldehyde 3-phosphate dehydrogenase            -gap2_2-
        glyceraldehyde 3-phosphate dehydrogenase           -gpr_1-   glyceraldehyde 3-
phosphate dehydrogenase        -gpr_2-   glyceraldehyde 3-phosphate dehydrogenase

--1,3-Bisphosphoglycerate (1,3-BPG)-->3-Phosphoglycerate(3PG)--
        -pgk-    phosphoglycerate kinase

--3-Phosphoglycerate(3PG)-->2-Phosphoglycerate(2PG)--
        -gpmA_1- 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase
        -gpmA_2-2,3-bisphosphoglycerate-dependent phosphoglycerate mutase

--2-Phosphoglycerate(2PG)-->Phosphoenolpyruvate(PEP)--
        -eno_1-  enolase             -eno_2-   enolase

--Phosphoenolpyruvate(PEP)-->Pyruvate--
        -pyk-    pyruvate kinase
```

```
--Xylan Degradation--
        -xynA_1- endo-1,3-beta-xylanase
        -xynC_1- glucuronoarabinoxylan endo-1,4-beta-xylanase
        -xynC_2- glucuronoarabinoxylan endo-1,4-beta-xylanase
        -xynA_2- endo-1,3-beta-xylanase
        -xynC_3- glucuronoarabinoxylan endo-1,4-beta-xylanase
        -xynB_1- endo-1,4-beta-xylanase B
        -xynB_1- Exoglucanase/xylanase
        -xynA_3- endo-1,3-beta-xylanase
        -aguA_3- xylan alpha-(1->2)-glucuronosidase
        -xynB_2- endo-1,4-beta-xylanase B
        -xynB_2- Exoglucanase/xylanase
        -xynC_4- glucuronoarabinoxylan endo-1,4-beta-xylanase
        -xynB_3- endo-1,4-beta-xylanase B
        -xynB_3- Exoglucanase/xylanase
        -xynA_4- endo-1,3-beta-xylanase
        -cex-    Exoglucanase/xylanase
```

Figure 2.12 Individual metabolism file output of the MPP tool

- *(A)* The glycolysis pathway of *A. acidipropionici*. For each step in the pathway the substrates and products are shown, as well as any enzymes found in the genome that are known to catalyse that step.
- (B) The *P. borealis* genome encodes for many different xylanases, indicating that this could be an important pathway in *P. borealis* metabolism.

individual genome. Each metabolism file includes a breakdown of each compound and enzymatic step within each analysed pathway so the user can make independent informed decisions on the completeness of that pathway (Figure 2.12). These metabolism files can be particularly useful in identifying the genes present/missing when pathways are partially complete or when pathways are highly similar with similar completeness percentages, but differ in a couple of genes, e.g. acetoclastic and hydrogenotrophic methanogenesis. They can also show where pathways and genes are present in high copy numbers in genomes (Figure 2.12(B)), as well as in linking pathways together, for example linking glycolysis to pyruvate fermentation or intermediates from the citrate cycle into glycolysis, as each metabolism file shows all compounds used and produced during each step in a pathway.

2.2.1 Testing

For testing of the MPP tool, the same 11 genomes used for building mock community 1 in testing of the MCCR tool were used (Figure 2.11). Genome information can be found in Table 2.1(A). To measure how effectively the MPP tool was able to predict the presence of certain pathways the results were compared to those of KEGG, a highly curated database, in Tables 2.7-11. If a pathway was >80% complete, it was counted as a complete pathway. In total 18 pathway files were built and used for analysis with several from each of the key steps in AD: 2 involved in hydrolysis (Table 2.7), 2 involved in acetogenesis (Table 2.8), 6 involved in fermentation (Table 2.9), and 5 pathways involved with methane production and use (Table 2.10) as well as other miscellaneous pathways of interest such as hydrogen sulphide production and 2 pathways in central carbon metabolism (Table 2.11).

In comparison to KEGG, the MPP tool is 93% accurate in predicting the presence of the given pathways. Of the 198 results gained from the MPP tool, 18 were excluded from the comparison because *C. evryensis* isn’t in the KEGG database, and an additional 20 were excluded as KEGG doesn’t have a pathway for xylan degradation or the use of hydrogen as an electron donor during methanogenesis (Tables 2.7 and 2.10). Of the remaining 160, only 11 results were inconsistent with the annotated data from KEGG.

From the literature *C. evryensis* has been shown to be an anaerobic, amino acid utilizing bacterium, unable to grow on carbohydrates[49]. It appears to have a proteolytic heterofermentative metabolism showing growth on several amino acids and producing acetate, butyrate, H_2 and CO_2 as well as propionate and valerate in some cases[49]. It is unable to use sulphate, thiosulphate or sulphite as electron acceptors[49]. The results from the MPP are supported by the literature. Through the MPP tool, *C. evryensis* is shown to be unable to hydrolyse cellulose or xylan (Table 2.7), both polysaccharides, nor does it contain the citrate cycle or the assimilatory sulphate reaction (Table 2.11). Contrary to expected for a proteolytic metabolism it does contain 80% of the glycolysis pathway (Table 2.11), but many of these enzymes could also be used in the pentose phosphate pathway for building cell carbon. The MPP tool also found the expected fermentation pathways, showing that *C. evryensis* has the pathways for producing acetate via the acetate kinase pathway, and butanoate from pyruvate (Table 2.9).

	Hydrolysis pathways			
	Cellulose degradation		Xylan degradation	
Species Name	KEGG	MPP	KEGG	MPP
A. acidipropionici	-	-	NA	-
C. evryensis	NA	-	NA	-
D. tunisiensis	+	-	NA	+
D. mccartyi	-	-	NA	-
L. byssophila	+	+	NA	+
M. ruminantium	-	-	NA	-
M. formicica	-	-	NA	-
M. wolfeii	-	-	NA	-
O. dioscoreae	-	-	NA	-
P. borealis	+	+	NA	+
T. albus	-	-	NA	-

Table 2.5 Comparison table of hydrolysis pathways between KEGG and MPP tool

	Acetogenesis			
	Acetate kinase pathway		Wood-Ljungdahl pathway	
Species Name	KEGG	MPP	KEGG	MPP
A. acidipropionici	-	-	-	-
C. evryensis	NA	+	NA	-
D. tunisiensis	+	+	-	-
D. mccartyi	-	-	-	-
L. byssophila	-	-	-	-
M. ruminantium	-	-	-	-
M. formicica	-	-	-	-
M. wolfeii	-	-	-	-
O. dioscoreae	+	-	-	-
P. borealis	+	-	-	-
T. albus	-	-	-	-

Table 2.6 Comparison table of acetogenesis pathways between KEGG and MPP tool

	Fermentation pathways											
	Acetyl-CoA fermentation (ethanol)		Pyruvate fermentation (acetate)		Pyruvate fermentation (Butanoate)		Pyruvate fermentation (Ethanol)		Pyruvate fermentation (Formate)		Pyruvate fermentation (lactate)	
Species Name	KEGG	MPP	KEGG	MPP	KEGG	MPP	KEGG	MPP	KEGG	MPP	KEGG	MPP
A. acidipropionici	+	+	+	+	-	-	-	-	-	-	+	+
C. evryensis	NA	+	NA	-	NA	+	NA	-	NA	+	NA	-
D. tunisiensis	-	-	-	-	-	-	-	-	-	-	+	+
D. mccartyi	-	-	-	-	-	-	-	-	-	-	-	-
L. byssophila	-	-	-	-	-	-	-	-	-	-	-	-
M. ruminantium	-	-	-	-	-	-	-	-	+	-	-	-
M. formicica	-	-	+	-	-	-	-	-	-	-	-	-
M. wolfeii	-	-	-	-	-	-	-	-	-	-	-	-
O. dioscoreae	+	+	-	-	+	-	-	-	-	-	+	+
P. borealis	+	+	-	-	+	+	-	-	+	+	+	+
T. albus	-	-	-	-	-	-	-	-	-	-	-	-

Table 2.7 Comparison table of fermentation pathways between KEGG and MPP tool

	Methane associated pathways									
	Acetoclastic methanogenesis		Hydrogenotrophic methanogenesis		Methane oxidation		Hydrogen electron donor		Methanol methanogenesis	
Species Name	KEGG	MPP	KEGG	MPP	KEGG	MPP	KEGG	MPP	KEGG	MPP
A. acidipropionici	-	-	-	-	-	-	NA	-	-	-
C. evryensis	NA	-	NA	-	NA	-	NA	-	NA	-
D. tunisiensis	-	-	-	-	-	-	NA	-	-	-
D. mccartyi	-	-	-	-	-	-	NA	-	-	-
L. byssophila	-	-	-	-	-	-	NA	-	-	-
M. ruminantium	-	+	+	+	-	-	NA	+	-	-
M. formicica	+	+	+	-	-	-	NA	-	-	-
M. wolfeii	+	+	+	+	-	-	NA	+	-	-
O. dioscoreae	-	-	-	-	-	-	NA	-	-	-
P. borealis	-	-	-	-	-	-	NA	-	-	-
T. albus	-	-	-	-	-	-	NA	-	-	-

Table 2.8 Comparison table of methane related pathways between KEGG and MPP tool

	Other energy metabolism					
	Assimilatory sulphate reduction		Citrate cycle		Glycolysis	
Species Name	KEGG	MPP	KEGG	MPP	KEGG	MPP
A. acidipropionici	+	+	+	+	+	+
C. evryensis	NA	-	NA	-	NA	+
D. tunisiensis	-	-	-	-	+	+
D. mccartyi	-	-	-	-	-	-
L. byssophila	-	+	+	+	+	+
M. ruminantium	-	-	-	-	-	-
M. formicica	-	-	-	-	-	-
M. wolfeii	-	-	-	-	-	-
O. dioscoreae	-	-	+	+	-	+
P. borealis	+	+	+	-	+	+
T. albus	-	-	-	-	-	-

Table 2.9 Comparison table of other energy metabolism pathways between KEGG and MPP tool

2.2.2 Discussion

One of the problems with metabolic pathway analysis using annotation servers such as KEGG is the very rigid structure of the pathway maps. It can be difficult trying to build a metabolic network of organisms with unusual metabolisms from each of these rigid pathways. However, with the genome metabolism files of the MPP tool, since each step in a metabolic pathway is shown including the intermediates, the genome pathway file also allows for easy linking between pathways. For example, the intermediates from the TCA cycle enter into many different pathways.

Metabolism can be difficult to predict simply from a genome, especially when enzymes are not specific to one pathway like in methanogenesis. The MPP tool shows *M. ruminantium* to contain similar percentages of completeness for acetoclastic and hydrogenotrophic methanogenesis, while KEGG and the literature shows it to be purely hydrogenotrophic[16]. Hydrogenotrophic methanogenesis has far more steps, consisting of multi subunit enzymes and so from the heatmap acetoclastic and hydrogenotrophic methanogenesis can look equally complete. The breakdown of the pathway in the genomes metabolism file shows acetoclastic to be missing an essential gene, while hydrogenotrophic is only missing a couple of subunits from multi-subunit complexes. *M. formicica* has been shown to be a hydrogenotrophic methanogen rather than acetoclastic experimentally[79]. However, from analysing its genome both through the MPP tool and KEGG, it does contain all the genes required for acetoclastic methanogenesis.

It is also important to remember with fragmented genomes that assumptions shouldn't be made based on the presence of a small number of genes, even with pathways as specific as methanogenesis. Several of the genomes in the order Archaeoglobales, which are not methanogens, contain a few genes involved in hydrogenotrophic methanogenesis, which are instead used in lactate utilization[44]. The MPP tool attempts to circumvent this issue by creating a pathway completeness percentage that can be used in the context of how complete the genome is and how complete other pathways in the genome are.

2.3 Development of a pipeline for the analysis of metagenomic data

Metagenomic datasets, particularly from AD, are often large and complex. The analysis of such datasets can be time consuming with many different steps. In response to this, a modular pipeline for the analysis of pre-binned contigs was designed (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/Metagenome_Analysis_Pipeline.sh). The automated pipeline consists of 5 custom Python scripts and their associated TXT files, CheckM[72] and Prokka[73] and an overview of this can be seen in Figure 2.13. The aim of the pipeline is such: to be easily usable to those with limited bioinformatics experience, to analyse the quality of genomes and where possible or necessary increase that quality, annotate genomes of a suitable quality and build metabolic pathways from the annotated genomes.

To fulfil the first aim, variables within the pipeline such as which dataset to analyse and what levels of completeness and contamination are acceptable are all controlled using a single TXT file, "Parameters.txt" (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/Parameters.txt), meaning the user does not need to enter variables directly into each separate script.

Metagenome binning is difficult and, regardless of the algorithm used, can result in a large number of OTU's being formed containing only a few short contigs. Towards the second aim and to avoid unnecessary analysis on OTU's too small to contain a full genome, the first step is to create a directory for all the analysis to take place in, containing only those genomes large enough to contain a genome. This is achieved using FileSize_Filter.py (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/FileSize_Filter.py). The user is able to specify the minimum size of an OTU in kilobytes, since 1 kilobase is approximately equal to 1 kilobyte, through the Parameters.txt file. Depending on the quality of the dataset, measuring the size of an OTU in kilobytes rather than kilobases is 8 times faster. If the OTU only contains a few long contigs, measuring in kilobases would be faster, however the complexity of AD datasets can result in highly fragmented genomes

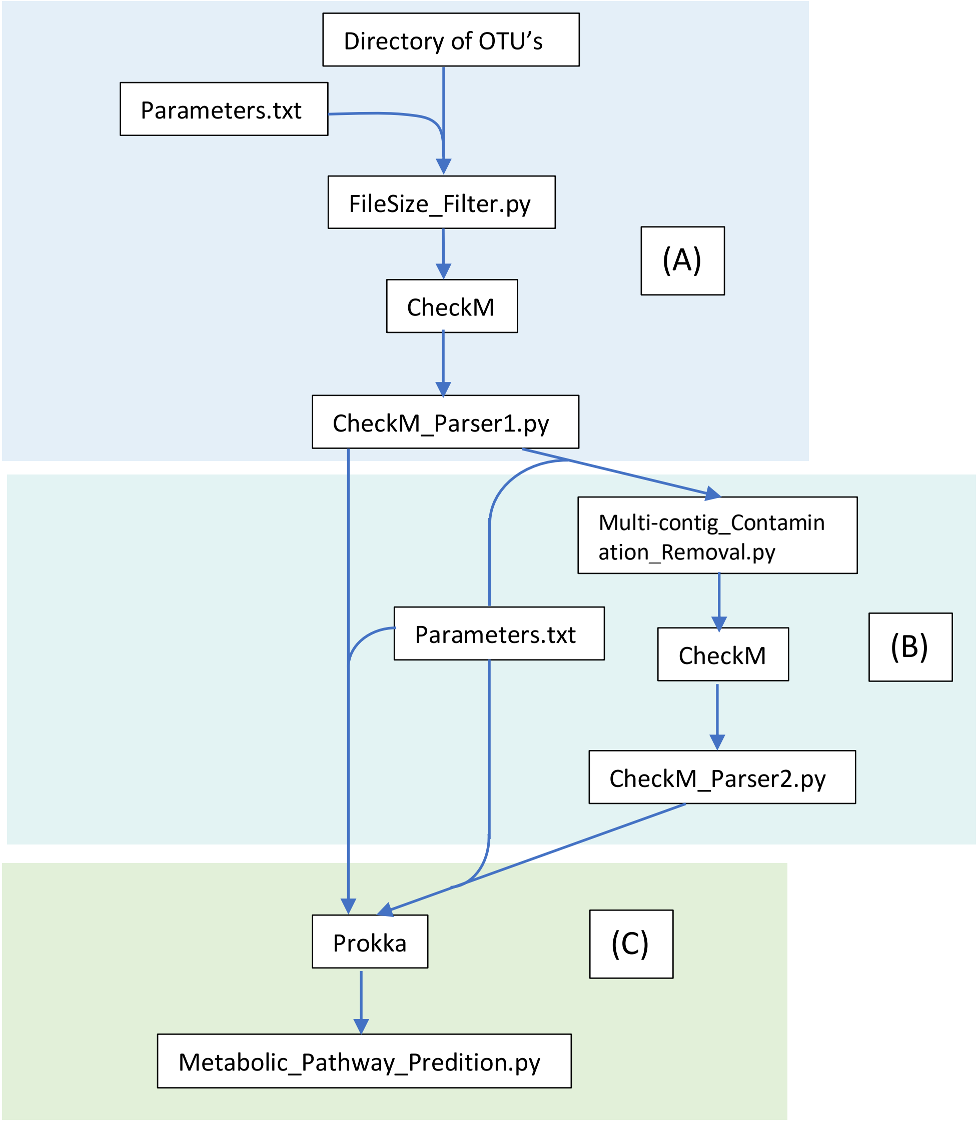


Figure 2.13 Overview of the metagenone analysis pipeline.

A basic overview of pipeline workflow consisting of 8 steps, 7 programs, and one .txt file to provide user submitted variables.

(A) A directory containing a metagenome, pre-binned into OTU's, is filtered by size using FileSize_Filter.py and the minimum size dictated by a user in the Parameters.txt file. The completeness, contamination and taxonomy of these OTU's is then measured and interpreted using CheckM and CheckM_Parser1.py respectively.

(B) User defined completeness and contamination thresholds in Parameters.txt direct OTU's to different parts of the pipeline depending on the OTU's contamination. If an OTU is above the completeness and below the contamination threshold in Parameters.txt, it is deemed acceptable and annotated. If the contamination is too high, the MCCR tool attempts to decrease it using the MCCR tool. The contamination and completeness is remeasured using CheckM and CheckM_Parser2.py. If the contamination is decreased and the completeness still above the threshold from Parameters.py, the OTU is annotated.

(C) OTU's are annotated using Prokka as either bacteria or archaea depending on the taxonomy assigned by CheckM. The GBF output files generated by Prokka feed into Metabolic_Pathway_Prediction.py which screens OTU's for pathways of interest.

into thousands of contigs and measuring the size of the file rather than the cumulative size of each contig is much faster for roughly the same results.

Next CheckM is used to assess genome contamination and completeness using single copy marker genes. CheckM_Parser1.py (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/CheckM_Parser1.py) rewrites some of the results from

CheckM into a more readable format containing OTU name, length, assigned taxonomy, completeness and contamination. The minimum acceptable completeness and maximum acceptable contamination is specified by the user in Parameters.txt, and CheckM_Parser1.py uses this information to direct OTU's, or what could now be considered genomes, with acceptable completeness and contamination towards genome annotation.

OTUs with acceptable completeness but too high contamination are directed into the MCCR tool, where attempts to decrease contamination are made using sequence composition and taxonomy as described in Chapter 2.1. After this CheckM and CheckM_Parser2.py (https://github.com/KimBarnes/Metagenome_Analysis_Pipeline/blob/master/CheckM_Parser2.py) are used to search for OTU's, or genomes, within the results of the MCCR tool with acceptable completeness and contamination and direct them to genome annotation.

Towards the third aim, genomes are annotated using Prokka[73] as either bacteria or archaea depending on the taxonomy assigned by CheckM. The GBF output file of Prokka is then fed into the MPP tool to create graphical and text representations of the metabolic pathways present within each genome.

2.3.1 Testing

2.3.1.1 Results

Mock community 1, shown in Table 2.1(B), consisting of 30 genomes in 15 "highly contaminated" synthetic OTU's was used for the testing of the pipeline. Since all the OTU's contained full genomes the file size threshold was set to an arbitrary 1 kb, completeness threshold was set to 90%, and the contamination threshold set to 10%. Of the 15 synthetic OTU's, all were over 1 kb, had a completeness over 90% and contamination over 10% and were pushed into the MCCR tool to reduce the contamination. 16 genomes with a completeness of >90% and contamination of <10% were created as a result, although an

Genome	Species
C-A	*T. albus*
C-B	*D. mccartyi*
2A-A	*L. byssophila*
2B-A	*C. evryensis*
2B-B	*M. wolfeii*
2C-A	*C. evryensis*
2C-B	*D. mccartyi*
3B-A	*P. borealis*
3B-B	*L. byssophila*
3C-A	*L. byssophila*
3C-B	*M. formicica*
4A-C	*M. wolfeii*
4B-A	*L. byssophila*
4B-B	*C. evryensis*
4C-A	*P. borealis*
4C-B	*A. acidipropionici*

Table 2.10 16 genomes created using the MCCR tool from mock community 1

16 genomes were created using the MCCR tool with completeness and contamination of >90% and <10% respectively, 13 bacterial and 3 archaeal

additional 2 had a completeness of >75%. The 16 genomes consisted of 3 archaeal genomes and 13 bacterial genomes noted in Table 2.13, that were then annotated using Prokka and analysed using the MPP tool. A comparison of the MPP output from the pipeline with that from Figure 2.11 of the whole genomes showed that the results were fairly consistent and that, for these metabolic pathways at least, use of the MCCR tool did not only decrease contamination of the single copy marker genes used by CheckM to estimate contamination.

2.3.1.2 Genome Quality

7 OTU's were split only into 2 FASTA files, generating two genomes with completeness and contamination within the accepted bounds. These were OTU's C, 2B, 2C, 3B, 3C, 4B and 4C, while OTU's 2A and 4A both only created 1 genome. However, both 2A and 4A contained *D. tunisiensis* which can't form a genome with >90% completeness due to its propensity to split itself into two genomes as discussed in section 2.1.3. It is interesting to note that there were small variations in amount of contamination found in each OTU by CheckM. All the 15 OTU's pre-MCCR tool should have a contamination of 100% exactly since they're made of 2 full genomes, however this was the case for only 1 of the 15 OTU's, 1B. 6 OTU's had contamination of less than 100%, and 8 OTU's had contamination greater than 100%.

Deceases and increases in contamination could occur from the single copy genes CheckM uses to measure contamination being split into two, and either both halves are too small to be recognised making the contamination <100% or both halves are big enough to be recognised as two copies of the same gene making the contamination >100%. This would also explain why CheckM found small amounts of contamination, up to 3.5%, for C-B, 1A-B, 1C-B, 2A-C, 2B-B, 2C-B, 3B-B, 3C-B, 4A-C, 4B-C and 4C-B, despite them only containing contigs from a single species.

OTU 3C has a completeness of 100% and generates 2 genomes, 3C-A and 3C-B, of completeness >98% but only has a contamination of 36.57%, far lower than the other 14 OTU's of the mock community. For some reason CheckM only identifies 90 and 47 marker genes for *L. byssophila* and *M. formicica* respectively within 3C, however when split into 3C-A (*L. byssophila*) and 3C-B (*M. formicica*), CheckM finds 393 and 203 marker genes respectively. There appears to be no explanation for why this happens.

Unsurprisingly those OTU's that showed the lowest binning accuracies also showed the highest levels of contamination post MCCR. For 1B-A, which has the lowest binning accuracy at 60%, the contamination measured by CheckM wasn't decreased at all, despite the removal of 170,720 bp in 28 contigs. Opposite to this CheckM found only 11% contamination in 1A-A, despite it containing the whole of the *P. borealis* genome, and 119 (1,269,738 bp) of the 287 contigs (2,820,858 bp) of *M. formicica*.

2.3.1.3 Metabolic Pathway Mapping

Of the 16 genomes extracted, 13 were annotated as bacteria and 3 annotated as archaea based on the taxonomy assigned by CheckM. The GBF output from Prokka of each genome was then searched for the 18 metabolic pathways involved in AD.

Comparing Figure 2.14 of the genomes extracted from mock community 1 and the results from Figure 2.11 of the individual genomes, all 16 genomes formed by the MCCR tool produced a heatmap that looked visually the same as their original genomes. Although the 16 genomes, 13 bacterial and 3 archaeal, appeared to have no changes in pathway completeness there were a few differences that could be seen in either the TSV file of numerical pathway completeness values, or the individual metabolism files. For example, the *L. byssophila* genome from 4B-A contained 10% more of the citrate cycle and 16% more of the pyruvate fermentation to butanoate pathway than the original genome. This does not appear to be to the detriment of its sister genome 4B-B from *C. evryensis* which still contains the same pathway completeness for those two pathways as the original genome.

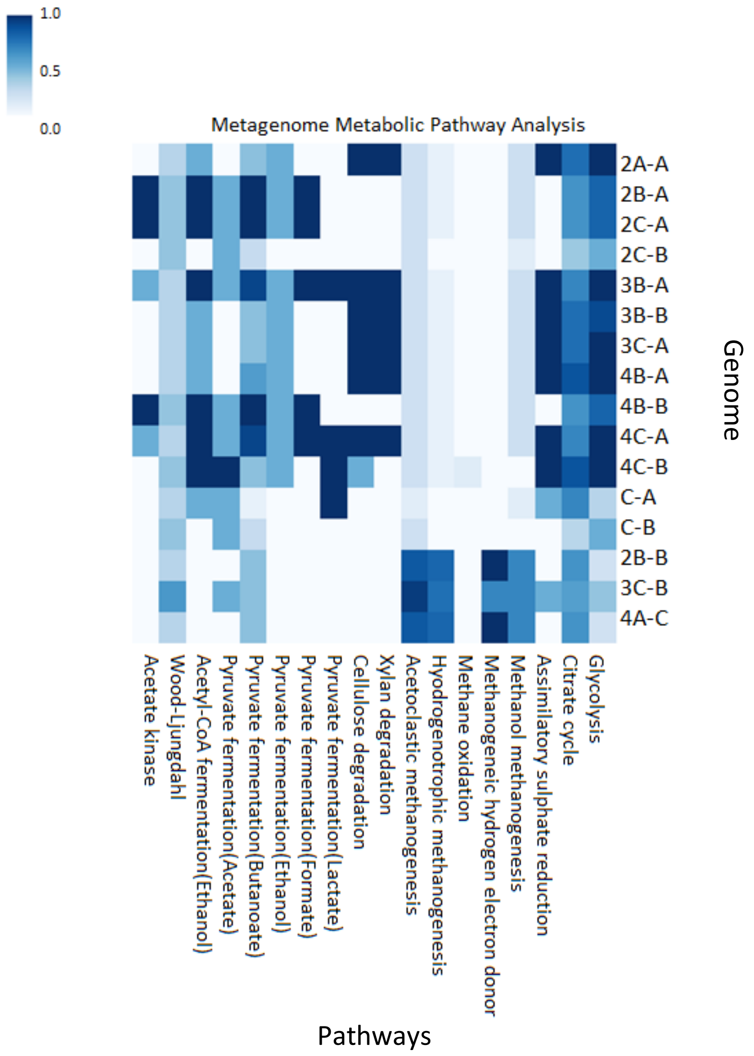


Figure 2.14 Metabolic pathway analysis of mock community 1.

16 genomes were formed from the analysis of 15 OTU's by the metagenome analysis pipeline, which were annotated using Prokka and 18 metabolic pathways searched for. Representative pathways from each step in AD are shown: (left to right) acetogenesis, fermentation, hydrolysis and methanogenesis as well as 3 other pathways involved in energy metabolism. Darker colours indicate more complete pathways.

The *D. mccartyi* genome from C-B lost 10% of its citrate cycle, which caused the *T. albus* genome in C-A to gain 5% of a citrate cycle.

2.3.2 Discussion

It is perhaps unsurprising that only 16 genomes were created from the 15 OTU's when a completeness threshold of >90% is set, as the MCCR tool had a binning accuracy >90% for only 8 OTU's. 7 of the OTU's with a binning accuracy of >90% created two genomes while 1, 4A, only formed 1 genome, 4A-C, that was >90% complete. The second genome present in 4A, *D. tunisiensis*, split into two "genomes" and although 90% of contigs and 92% of sequence from *D. tunisiensis*, were binned into the same genome, 4A-A, this only amounted to 77% of the genome in terms of completeness. This obviously highlights the issue with using single copy marker genes in measuring genome completeness, as genome completeness measured in this way is not always indicative of functional completeness or how much of an organisms' metabolism is still intact. However it is still the most commonly used method for assessing genome completeness[72,80,81]. The creation of the metabolic pathway analysis heatmap made it easy to identify potential genomes of interest. For example, it is easy to pick out the 3 methanogenic archaea 2B-B, 3C-B and 4A-C, or the hydrolysers 2A-A, 3B-A, 3B-B, 3C-A, 4B-A and 4C-A simply by looking for darker squares in the columns of each pathway of interest.

Although only 16 genomes were extracted from the 15 OTU's at a completeness of >90%, a success rate of only 53%, these are still 16 genomes that would otherwise be considered too contaminated to be functionally analysed in a metagenome. There is still much work to be done on the MCCR tool in regards to increasing its accuracy, however the remainder of the tool was both easy and efficient to run and understand the results.

3. Analysis of large metagenomic datasets from anaerobic digestion

A previously unpublished AD metagenome dataset resulting from investigations into the effect of DNA extraction methodology was used to better understand both the capabilities, and pitfalls, of the pipeline on a real metagenome and to better understand the community within the dataset.

DNA was extracted from 4 samples for each of two commercial mesophilic (35°C) wastewater AD systems (Naburn, York, UK and Blackburn Meadows, Sheffield, UK) and one lab-scale (5 litre) thermophilic (55°C) AD reactor inoculated with sludge from a waste water treatment plant and acclimatised to thermophilic conditions for two weeks (Millbrook, Southampton, UK, 50°54'33.4"N 1°26'44.6"W). Paired end sequencing on an Illumina HiSeq 3000 resulted in an average of 32 million pair-end reads per sample and can be found in the European Nucleotide Archive under accession number PRJEB20855. Reads from all 12 samples were pooled and assembled using Megahit with a minimum contig length of 1 kb[65]. Binning was done using a custom Python script using differential coverage of reads between different samples resulting in 15,025 OTU's.

3.1 Aims

To understand the effectiveness of the pipeline on real metagenomic datasets from AD, and start to understand the unique role each genome plays in the complex community of AD.

3.2 Methods

The pipeline was run using Python 2.7 on a UNIX multicore workstation. The size filter was set at 500 kb based on the smallest known bacterial genome at 530 bp[82], completeness arbitrarily set to a minimum of 75% and contamination arbitrarily set to a maximum of 10%. A negative GI list containing environmental and metagenome samples was used with the MCCR tool to ensure BLAST hits to known organisms.

3.3 Results

The AD dataset consisted of 15,025 OTU's varying between 2 kb and 131,495 kb in length. Cluster_k99_1504826 was excluded from the analysis due to its size. It contained 131,495,035 bp in 26,600 contigs with 2855% contamination and likely contained contigs from a variety of bacterial and archaeal species with similar differential coverage of reads.

Not only is the pipeline not designed to deal with that amount of contamination, but many of the analysis files would be difficult to open due to their size.

Excluding Cluster_k99_1504826, of the 15,024 OTU's in the metagenome, only 85 were over 500 kb, and of those 85 only 23 had a completeness greater than 75%. 9 of the 23 OTU's, 1 archaeal and 8 bacterial, had contamination of less than 10%, while the remaining 14 had contamination greater than 10% and were directed through the MCCR tool.

3.3.1 Reducing Contamination

Of the 14 OTU's directed into the MCCR tool in an attempt to reduce contamination, none were altered to the required parameters of completeness greater than 75% and contamination to less than 10%. The tool was successful in reducing contamination in many cases, but either did not decrease it to below 10% or the completeness decreased to below 75%. This was largely the result of three problems: the OTU containing multiple species from the same phyla, the %GC filter was not narrow enough, or the resulting genomes were not considered complete enough. One example of each issue is discussed below.

3.3.1.1 Single phyla OTU's

Cluster_k99_382050 is a 6.6Mb, 81% complete OTU with relatively small 24% contamination. Both the amount of contamination measured by CheckM and the distribution of contigs as shown in Figure 3.1(A) indicates the presence of two organisms, however the MCCR tool was unable to decrease the contamination. The longest contig showed highest similarity to Planctomyces sp. SH-PL14, and 56% of contigs BLAST results returned this species. Of the 808 contigs analysed, 83% were related to a Planctomycetes species, including those clustering around ~50% %GC that could be assumed to be a different species (Figure 3.1(A)). This data combined with the distribution of contigs over %GC indicate that there were two species of Planctomycetes, and both a %GC of ~50% and ~59% are well within the range of currently known Planctomycetes species[83,84]. The MCCR tool would not be able to distinguish between since it is currently only designed to differentiate species on their phylum.

3.3.1.2 Incorrect %GC filter

A second problem that had not been suitably anticipated lies in the exponential function that the %GC filter used. Bacterial genomes typically have a %GC within the range of 20-75%, and some OTU's contained almost the whole spectrum. Cluster_k99_6047478

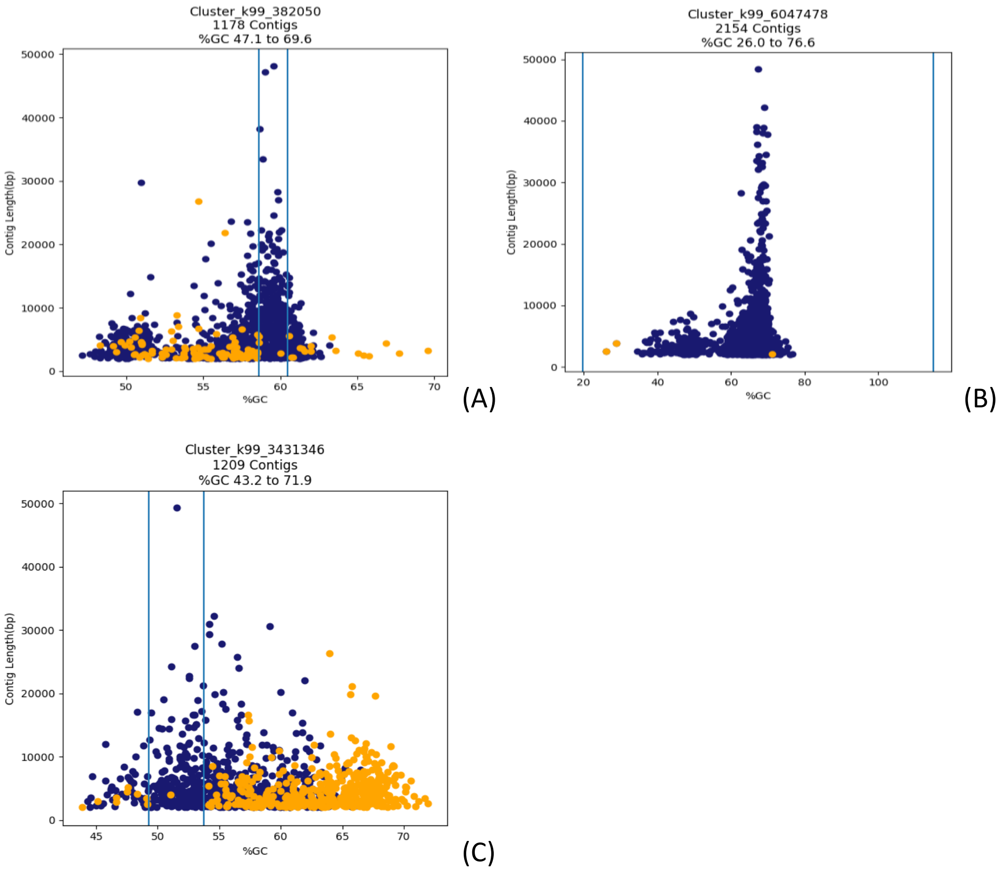


Figure 3.1 Distribution of contigs in OTU's analysed by the MCCR tool from a real metagenome dataset

(A) Contigs from Cluster_k99_382050-A, related to Planctomycetes species are highlighted in blue. Contigs from Cluster_k99_382050-B, unrelated to Planctomycetes, are highlighted in orange. Lines show the upper and lower bounds of the %GC filter. The overall distribution of contigs, with one cluster around 50% %GC and another around 60% %GC indicates the presence of two genomes within the OTU.
(B Contigs from Cluster_k99_6047478. Lines show the upper and lower bounds of the %GC filter. The overall distribution of contigs, with one cluster around 50% %GC and another around 70% %GC indicates the presence of two genomes within the OTU. The 3 contigs removed from the OTU are in orange.
(C) Contigs from Cluster_k99_3431346-A, related to Synergistetes, are highlighted in blue. Contigs from Cluster_k99_3431346-B, related to Proteobacteria, are highlighted in orange. Lines show the upper and lower bounds of the %GC filter. The overall distribution of contigs, compared to Figures 3.2 and 3.3, do not indicate the presence of two genomes, however the orange ones are largely distributed to the right, and blue to the left.

contained 10.7 Mb in 2154 contigs with a contamination of 127%. From the distribution of contigs shown in Figure 3.1(B), it is obvious that those clustering around 48% %GC are likely from a different genome. However, the contigs were spread from %GC of 26-77%, which meant that the lower bound of the %GC filter was 20% while the upper bound was 115% and no contigs were removed by this step in the analysis. In total only 3 contigs were removed via the TNF analysis.

3.3.1.3 Incomplete genomes

Cluster_k99_3431346 is a 6.4Mb, 96% complete genome with 147% contamination indicating 2.5 genomes. This is supported by the presence of 3 copies of RNA polymerase alpha subunit, and 2 gene clusters of RNA polymerase beta and beta' subunits. Since there were no 16S rRNA gene sequences, blastn of the RNA polymerase subunits indicated 2 closely related members of Synergistetes and a member of Alphaproteobacteria, or more specifically *Rhodobacter*. The MCCR tool was able to pull out a large portion of the Alphaproteobacteria genome, reducing the contamination from 147% to 112%, however the resulting Alphaproteobacteria genome was only 37% complete. Analysis of the binning of the RNA polymerase genes and ribosomal protein genes used by CheckM to assign completeness and contamination[72], show that where possible the genes were largely redistributed in a 2:1 ratio, with those with BLAST hits to Alphaproteobacteria successfully pulled out. Plotted on the same graph in Figure 3.1(C) it is clear that those with higher %GC tended to be part of the Alphaproteobacteria genome, while those with lower %GC were part of the Synergistetes genomes in keeping with the literature[85–87]. In this respect the MCCR tool worked as well as could be expected.

3.3.2 Metabolic analysis

Only 9 OTU's, or genomes, from the metagenome went on to have their metabolism analysed. From Figure 3.2 generated as part of the metabolic analysis we can see that Cluster_k99_3668352 is likely the only methanogen as none of the other genomes have high pathway completeness for any of the methane related pathways. There are 3 genomes with complete cellulose and xylan degradation pathways: Cluster_k99_1276485, Cluster_k99_4934154 and Cluster_k99_466860. Far more prevalent were the fermenters. All 8 of the bacterial genomes had at least one fermentative pathway, with the acetate kinase pathway, forming acetate and ATP from acetyl-CoA, the most prevalent. No complete pathways for hydrogen sulphide production by the assimilatory sulphate reduction pathway were found.

By using the graph in combination with the individual metabolic genome files describing each step in the pathway for each genome, a clearer picture of the metabolisms in each OTU is given.

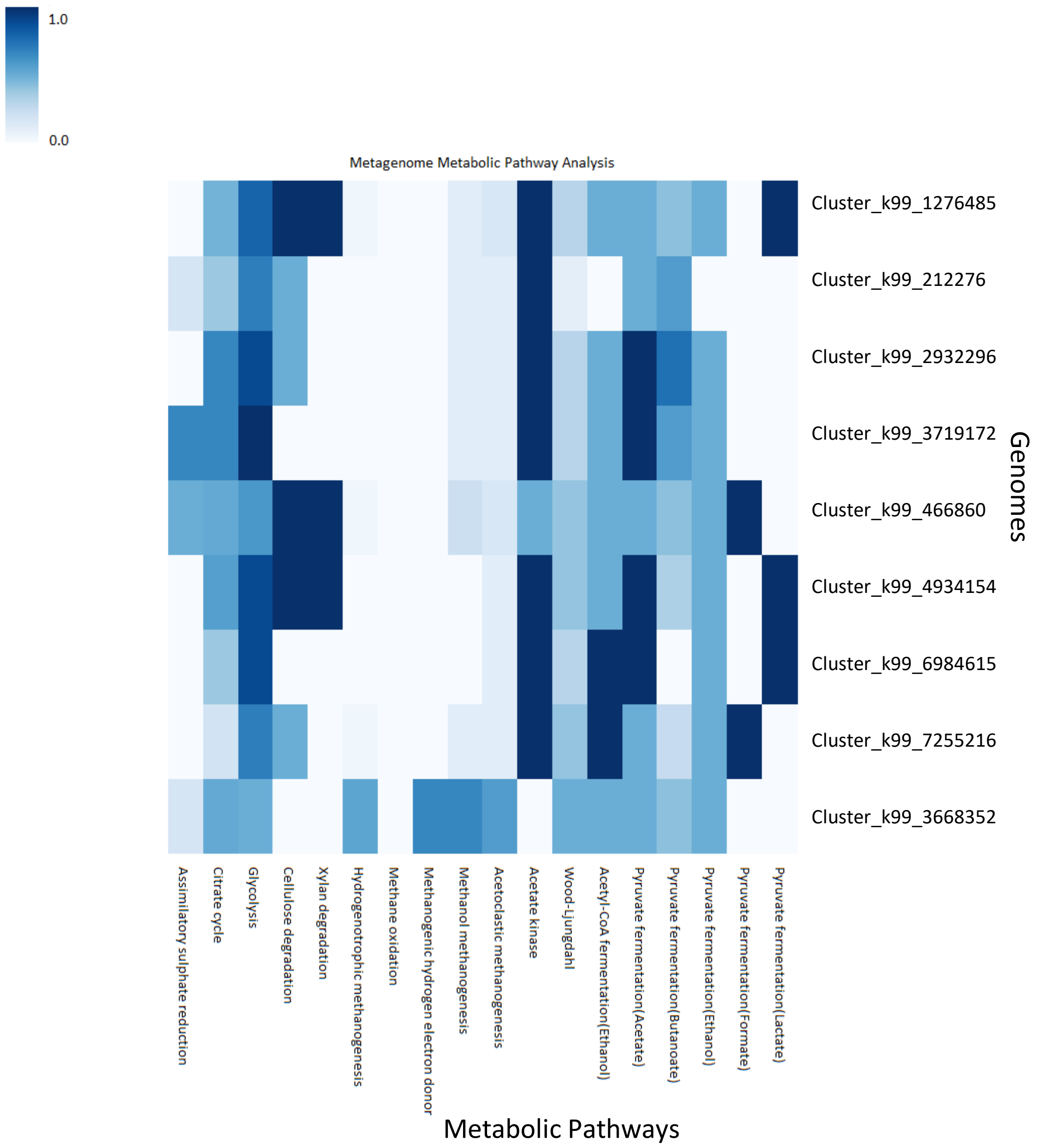


Figure 3.2 Metabolic analysis of 9 genomes from an AD metagenome.

9 genomes were extracted from the AD metagenome with a completeness >75% and contamination <10%. Each were screened for 18 different pathways and given a score of 0-1 of how complete the pathway was. A heatmap was created for a graphical representation of pathway completeness across the metagenome, where darker colours indicate more complete pathways.

Cluster_k99_1276485 – Bacteria, 96% complete, 2% contamination

A hydrolyser able to hydrolyse both cellulose and xylan, though likely having a preference towards xylan as it has 9 different xylanases compared to 3 cellulases. The glycolysis pathway is 80% complete, only missing a glucokinase to phosphorylate glucose to glucose-6-phosphate which is not necessary if the bacterium contains a phosphotransferase transport system for importing glucose, and glyceraldehyde 3-phosphate dehydrogenase. It is able to ferment pyruvate from glycolysis into lactate and into acetate via acetyl-CoA. It also has 3 different alcohol dehydrogenases, so while it doesn't appear to be able to ferment pyruvate into ethanol through the specific pathways tested, the organism could potentially generate an alcohol end product. The TCA cycle is only 50% complete, and in the context of the genome being 96% complete, it is difficult to predict whether the pathway is likely to be complete or whether it is used for creating metabolic intermediates.

Cluster_k99_212276 – Bacteroidetes, 80% complete, 3% contamination

The only pathway that is complete is the acetate kinase pathway. The glycolysis pathway is 70% complete, and in the context of a genome completeness of 80% it is likely the pathway could be complete. The genome contains one cellulase, but no xylanases. The pathway for pyruvate fermentation to butanoate is 60% complete, with the genes for last 2 steps for butanoate synthesis present, indicating it may be able to produce butanoate, but not necessarily from pyruvate

Cluster_k99_2932296 – Bacteroidetes, 96% complete, 2% contamination

Another potential hydrolyser with the genome containing 2 cellulases. Glycolysis is 90% complete, only missing the last gene of the pathway: pyruvate kinase. The citrate cycle is 67% complete, missing the genes for the 3 steps in converting acetyl-CoA to 2-oxoglutarate. This organism has several fermentation pathways from pyruvate including pyruvate directly to acetate as well as to acetate via acetyl-CoA. The pyruvate fermentation to butanoate pathway is 75% complete. Of the 7 steps in fermenting pyruvate to butanoate the genome has representatives for 5, although not all subunits are present, indicating the bacterium is able to ferment pyruvate to butanoate. This is unsurprising, many bacteria only heteroferment pyruvate to butanoate with acetate and other VFA's as sub-products[30].

Cluster_k99_3719172 – Bacteria, 84% complete, 1% contamination

This bacterium appears to have a variety of different energy producing metabolisms. The assimilatory sulphate reduction pathway is 67% complete, with all the genes present that are required for sulphate reduction to sulphite. The microbe also likely has a heterofermentative metabolism, having all the genes for glycolysis and both acetate fermentation pathways, as well as a 60% complete pyruvate fermentation pathway to butanoate, including the last step in the pathway generating butanoate from butanoyl phosphate.

Cluster_k99_466860 – Clostridiales, 92% complete, 0% contamination

This bacterium is another hydrolyser, with genes for both cellulose and xylan degradation, instead specialising in cellulose with 9 cellulases to 3 xylanases. Despite its high genome completeness, only the two hydrolysis pathways and pyruvate fermentation to formate are complete. The assimilatory sulphate reduction pathway contains the genes to reduce sulphate to PAPS, but no further. 4 of the 6 fermentative pathways are 50% complete and glycolysis is only 60% complete.

Cluster_k99_4934154 – Bacteria, 84% complete, 6% contamination

The third hydrolyser/fermenter containing 7 cellulases and 6 xylanases. The glycolysis pathway is 90% complete, and the fermentative pathways of pyruvate to acetate and lactate as well as acetyl-CoA in the acetate kinase pathway. The acetate kinase pathway has several copies of the genes involved, in comparison to the other pathways where there are single copies.

Cluster_k99_6984615 – Bacteria, 88% complete, 4% contamination

This bacterium could be considered a dedicated fermenter. The glycolysis pathway is 90% complete, and the genome contains 4 different complete fermentative pathways: acetate kinase, acetyl-CoA to ethanol and pyruvate fermentation to acetate and lactate.

Cluster_k99_7255216 – Clostridiales, 92% complete, 2% contamination

This bacterium only contains 70% of the glycolysis pathway despite being 91% complete, however the presence of two cellulases indicate that glucose is a likely growth substrate. The organism also has 2 of the least common fermentative pathways in this metagenome, pyruvate fermentation to formate, and acetyl-CoA fermentation to ethanol. The organism is also able to produce acetate via the acetate kinase pathway.

Cluster_k99_3668352 – Euryarchaeota, 76% complete, 1% contamination

At only 75.7% complete it is perhaps unsurprising that the genome has no complete pathways. From the heatmap of pathway completeness it is difficult to tell which methanogenesis pathways the genome contains as they all look roughly the same completeness. From analysis of the individual steps in each pathway, there is at least one subunit for each step in hydrogenotrophic and acetoclastic methanogenesis, but only the methane producing step that all three classes of methanogenesis have in common is present in the methanol methanogenesis pathway.

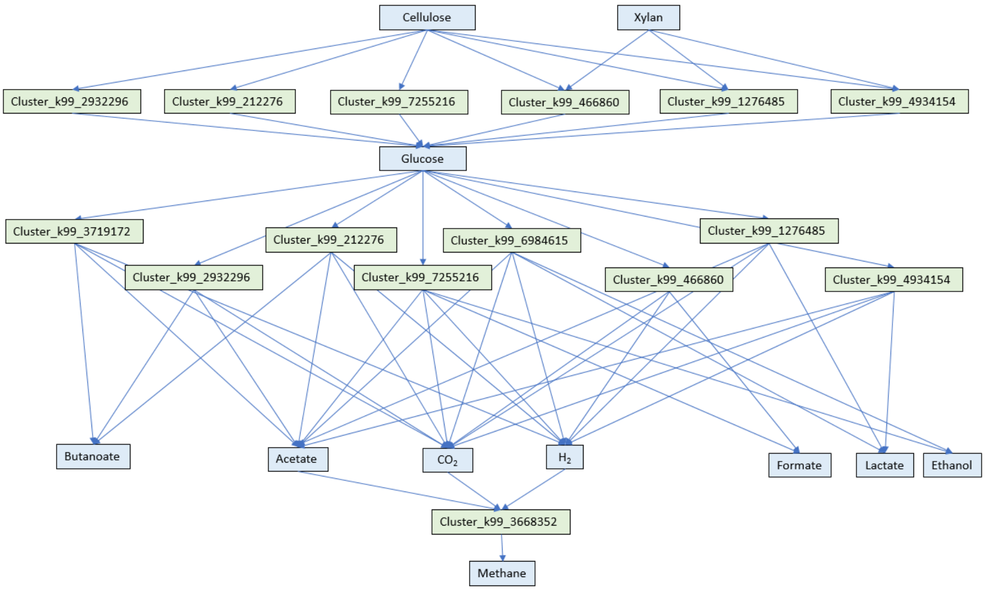


Figure 3.3 Metabolic networking of a metagenome

Detailed functional information gained using the MPP tool can be used to build metabolic networks that shown the interactions between microorganisms within a microbial community. Extracellular enzymes from 6 microorganisms hydrolyse cellulose and xylan into glucose and other soluble sugars which are taken up by all 8 bacterial species for fermentation into various products. The by-products of fermentation, acetate, CO_2 and H_2, are used by the archaeal methanogen to produce methane.

The genome contains 4 copies of the *acsA* gene, indicating that acetoclastic methanogenesis might be preferentially used.

All 9 OTU's had varying percentages of the citrate cycle present. Although not shown on the heatmap, none had cytochrome oxidases for aerobic respiration and so the citrate cycle likely performs a different function than in aerobic respiration. All had genes involved in the conversion of 2-oxoglutarate to succinyl-CoA which can replenish the supply of NAD+ from NADH, and many had genes involved in the conversion of succinyl-CoA to succinate for the release of CoA and direct generation of ATP. Pyruvate fermentation to propanoate was not a pathway covered by the MPP tool, but succinyl co-A acts as a precursor in this pathway. 2-oxoglutarate, another intermediate in the citrate cycle, is a precursor in the biosynthesis of several amino acids. The intermediates produced by the citrate cycle are involved in many metabolic pathways, and high pathway completeness in this pathway is not just indicative of a complete pathway.

The detailed information from the MPP tool can be used to generate basic metabolic networks between organisms within a metagenome like in Figure 3.3. While this map is not an accurate representation of the metagenome, which contains many more organisms than the 9 analysed, it helps to provide clear linkages between each organism.

3.3.3 Adjusting the completeness and contamination thresholds

The completeness threshold for genomes for metabolic analysis was set to 75%. However as already pointed out, the single copy genes CheckM uses only constitute ~10% of the genome giving a potentially lower measure of completeness than is actually there. This is supported by Figure 3.4, where all 86 OTU's over 500 kb were analysed for the 13 bacterial metabolic pathways regardless of completeness or contamination.

Although on average, those with a lower % completeness had fewer complete pathways, that's not always the case. Cluster_k99_10585819 has a measured completeness of 47.3%, but has a higher total percentage of pathway completeness than 8 of the 25 OTU's with a completeness >75% (data not shown). Pathways such as cellulose or xylan degradation or many of the fermentation pathways only consist of one or two steps, so they're far more likely to be "complete" compared to longer pathways like glycolysis or the Wood-Ljungdahl pathway. An additional 15 hydrolysers, either for cellulose, xylan or both, could be found from genomes >500 kb and <75% completeness. Even for longer pathways such as glycolysis with 10 steps, examples can be found. Cluster_k99_2622967 has a measured completeness of 32%, but has an 80% complete glycolysis pathway, 70% complete assimilatory sulphate reduction pathway (the 3rd highest) and almost half of a pyruvate to butanoate pathway, all of which are above average compared to the other 86 OTU's, while the completeness was below average.

3.4 Discussion

AD communities consist of a wide variety of microbes predominantly from only a few phyla working in tandem. It is not unexpected for many of these organisms to be closely related and have similar relative abundances, resulting in OTU's containing more than one species although this was perhaps exacerbated by combining reads from 3 dissimilar sources running on vastly different parameters (thermophilic vs mesophilic, lab vs commercial scale). This incorrect binning by using differential coverage creates several problems. Many species in AD are still unknown, making binning contigs from closely related species based on taxonomy and BLAST searches much harder. Secondly, when assembling short read data, highly conserved genes such as ribosomal RNA can co-assemble both creating hybrids of multiple closely related organisms and vastly decreasing the number of these genes, which are often used in estimating genome completeness[88,89]. Only 6 full and 6 partial 16S rRNA genes were identified by Prokka in the entire metagenome of 15,025 bins, all of which were most closely related to uncultured organisms and none of which were in the 9 good quality genomes (completeness >75%, contamination <10%) identified using CheckM. This made it impossible to compare the assignment of a phylum by the MCCR tool to the assignment of a phylum based on the 16S rRNA gene sequence and better understand how practical the MCCR tool is in a more realistic setting. CheckM also attempts to assign a taxonomy based on homology of the core genome rather than just the 16S rRNA gene sequence but was unable to assign anything more specific than a kingdom for 4 of the 9 OTU's.

Arguments can be made both for and against setting completeness and contamination thresholds. On the one hand it is important to use high quality genomes in analysis, but it is also important to include as much of the functional annotation as possible to really understand the complexities of microbial communities in metagenomes. Although only 9 OTU's underwent functional analysis 3 of the key steps of AD, hydrolysis, fermentation and methanogenesis were represented by these 9.

In this case based on the scarcity of 16S rRNA gene sequences, the number of bins, and the levels of completeness/contamination of some of those bins, it is likely that a large number of the contigs contain chimeric sequences and perhaps the analysis should be restarted from the assembly stage, pooling data from the 4 different extraction methods but assembling each of the 3 sources independently.

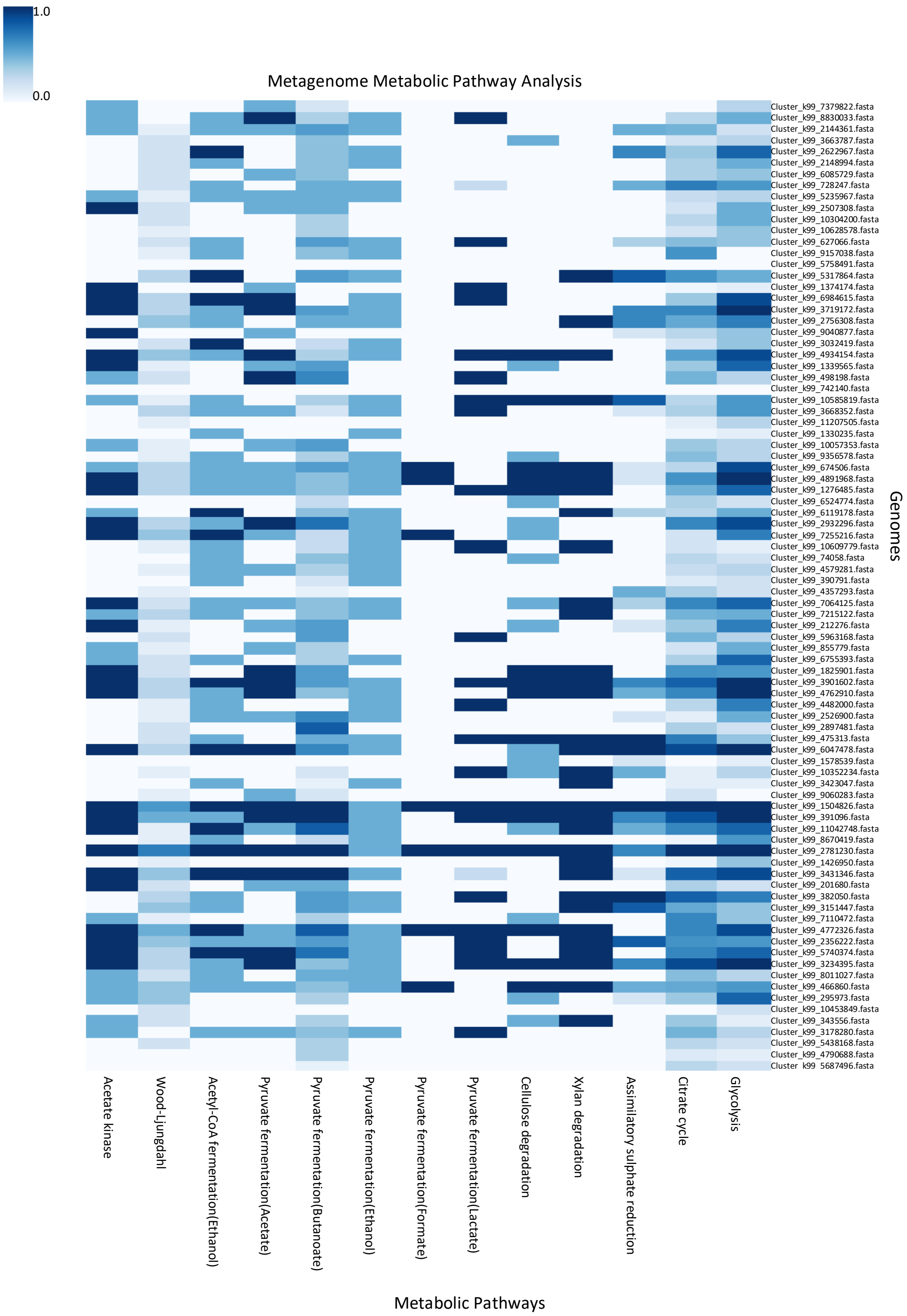


Figure 3.4 Metabolic analysis of a metagenome from AD.

Metabolic pathway analysis for all 86 OTU's over 500 kb. Only bacterial pathways are shown. Each were screened for 18 different pathways and given a score of 0-1 of how complete the pathway was. A heatmap was created for a graphical representation of pathway completeness across the metagenome, where darker colours indicate more complete pathways.

4. General Discussion

4.1 Assembly quality is essential to accurate genomes

Better assemblies, creating high quality contigs for binning into OTU's are essential for better understanding of the AD microbial communities. The AD metagenome used in this study could be considered low quality, only 86 OTU's were more than ¾ complete according to CheckM out of 15,000 and there were very few 16S rRNA gene sequences indicating likely co-assembly of these genes into chimeras based on their highly conserved nature. This co-assembly of conserved genes into chimeras and highly fragmented genomes are two of the issues that prevent the extraction of complete and contamination free genomes from AD metagenomic datasets. Long reads, such as those from MinION, go a long way to solve this, and can assemble whole genomes into single contigs[90,91]. Fewer contigs would also result in more accurate binning. Although the MCCR tool has been shown to remove contamination up to 100% in testing, when used on a real metagenome it is difficult to tell if its inability to extract genomes from contaminated OTU's is a result of the algorithm used by the MCCR tool or the quality of the genome. There were several OTU's from the AD metagenome in which the MCCR tool was able to significantly reduce the amount of contamination, but this was always at the cost of genome completeness. Additional testing on a wide variety of different datasets will be needed to properly understand the capabilities of this tool and pipeline, however two things are clear: when using single copy marker genes for measuring genome completeness a high quality assembly is essential, and while the MCCR tool can potentially help to polish OTU's there is a limit to which the tool can act as binning software.

4.2 Taxonomy-dependent binning can be misleading for unknown organisms

Taxonomy dependent binning, used by the MCCR tool, relies upon alignments of nucleotide sequence to known organisms and this method can create inaccurate results if the nucleotide sequence in question is relatively novel. This became clear with the test genomes *D. tunisiensis* and *C. evryensis,* both of which were binned less accurately than their counterparts likely as there were few close relatives within their respective phyla. It is difficult to say if this is an issue that would be prevalent in AD metagenomes. On the one hand a large proportion of microbes from AD are completely novel since their syntrophic inter/intra species interactions largely prevent isolation and sequencing. However, on the other hand the typically most prominent phyla in AD, Proteobacteria, Firmicutes and

Bacteroidetes, are quite diverse and all contain several hundred genera within them which would lend itself to more accurate binning (Table 2.6) [40,45–48].

Horizontal gene transfer would also be a potential area in which this method of binning would fall down and recently acquired genetic material would likely be removed from a genome unnecessarily. Despite the potential issues the MCCR tool might create or be unable to solve, it is important to remember that in this pipeline it is only used on OTU's that might be considered too contaminated, and although the resulting genomes may not be complete, that doesn't mean that the functional information in their genome is insignificant to the understanding of a metagenome.

4.3 Towards a better understanding of functional annotation in AD metagenomes

Although the MPP tool is not always 100% accurate compared to more curated tools such as KEGG, it has been shown to be 93% accurate and provides a much faster and wide-reaching approach to analysing functional annotation in metagenomes. Many of the bacterial genomes in both the mock community and AD metagenome appeared to contain genes involved in methanogenesis, a strictly archaeal pathway. Prokka uses BLAST searches against the UniProt database to assign annotation, however protein sequences submitted to UniProt are not always consistently named. In archaea *mcr* and *mta* gene clusters encode for the final step in methane production and the first step in methanol conversion to methane respectively and are specific to the production of methane. In bacteria *mcrB* encodes for a 5-methylcytosine-specific restriction enzyme while *mtaB* encodes for threonylcarbamoyladenosine tRNA methylthiotransferase used in tRNA modification, both of which appear relatively commonly within the bacterial genomes. While each of these gene names only represent one subunit in multimeric complexes part of the multi-step pathways of methanogenesis, representing only 10% of the shortest methanogenesis pathway, they do highlight the issue of false positives when using text matching to gene names. For this reason, the tool will likely never be able to give a comprehensive and in-depth analysis of each genome in a metagenome, for that there are already many highly curated databases and annotation services. However, it is able to give a general overview of the metabolic pathways present in 100's of individual OTU's within the context of the metagenome as a whole, something that other services cannot. This allows for the relatively quick and easy identification of genomes of metabolic interest, be those methanogens, hydrolysers or even hydrogen sulphide producers, to then be analysed using a more specific database.

There are alternatives to using the current method of text matching gene names. Hidden Markov Models could instead be used to identify the genes themselves, however a model would need to be built for each protein within a complex within each step of a pathway and this was beyond the scope of this project. Alternatively, the Prokka annotation software can be supplied with a user created and curated database of genes rather than using that from UniProt, to ensure all genes were labelled consistently. This way rather than using the shorter gene name, the longer protein name could be used for searching, making the process more specific e.g. using the protein name methyl-coenzyme M reductase beta subunit rather than the gene name *mcrB* which can be confused with the bacterial gene for threonylcarbamoyladenosine tRNA methylthiotransferase, also labelled *mcrB.*

4.4 Future work

Many issues were flagged up through development of the pipeline and custom Python scripts. Analysis of Cluster_k99_1504826 highlighted the need for there to be a size limit on which files the pipeline will try and analyse. Not only do files of this size and complexity create issues with the size of some graphical outputs, the script is really not designed to split that many genomes apart. It would be far more practical in that situation to re-bin the contigs and start again.

The MCCR tool was not as effective on a real metagenomic dataset in comparison to the synthetic dataset. An alternate method of binning might be to use mathematical modelling to determine the likely number of genomes in each OTU based on the distribution of contigs and measure of contamination from CheckM. As shown in Figure 3.1, contigs from a single genome tend to follow a Gaussian distribution when length is plotted against %GC, that could be used to estimate the number of genomes and build each genome within an OTU simultaneously rather than iteratively. This method would potentially require a larger number of alignments and taking much longer. This could be negated by using a different, faster alignment algorithm than blastn. Kraken is estimated to be 909 times faster than Megablast and would significantly decrease running time[92].

CheckM provides consistent taxonomic classification at domain level for each OTU however it would be better to include a separate more specific dedicated taxonomy assignment feature in the pipeline. There are a wide range of tools able to do this, from SSuMO[64] specifically designed for 16S rRNA gene fragments to extracting the assembled 16S rRNA gene sequences and using BLAST, or perhaps both. By adding a taxonomy assignment feature at both the beginning and end of the pipeline it would reflect both the total diversity within the metagenome, as well as automatically assigning taxonomy to individual OTU's which could be graphically represented.

Finally, additional user variables could be integrated into the Parameters.txt file, as currently it only contains 4 parameters: directory, and thresholds for file size, completeness and contamination. For example, Python scripts typically only use one core at a time when run on a multicore machine, however the blastn command within the MCCR tool can be instructed to use any number of cores which should be another user specified parameter. As should the maximum evalue for the longest contig, which is currently set to 0, but could easily become a user submitted variable.

4.5 Summary

As metagenomic datasets become larger and more detailed as a result of advances in sequencing technologies, the need for automated pipelines and software to analyse this overwhelming influx of data in an efficient manner will be needed. The pipeline and tools described here attempt to address this problem through their polishing of OTU's to extract as much information from a metagenome as possible, as well as rapid and simple metagenomic metabolic pathway mapping.

Although the MCCR tool will potentially become redundant in the future as long reads, which are easier to assemble and bin, become more frequently used currently metagenome sequencing is largely done using high throughput short reads. These can be difficult to bin correctly, and so there is still much improvement to be made in this area. However, as genome quality improves, and the number of genomes able to be reconstructed from a metagenome increases, the need for rapid functional annotation and assignment of metabolic pathways will also increase. Through the MPP tool, and its approach of viewing all the functional information from a metagenome collectively, better understanding of some of the many inter/intra species interactions that occur within an AD community can be achieved.

References

1. St-Pierre, B. & Wright, A. D. G. Metagenomic analysis of methanogen populations in three full-scale mesophilic anaerobic manure digesters operated on dairy farms in Vermont, USA. *Bioresour. Technol.* **138,** 277–284 (2013).

2. Myhre, G. *et al.* Anthropogenic and Natural Radiative Forcing. *Clim. Chang. 2013 Phys. Sci. Basis. Contrib. Work. Gr. I to Fifth Assess. Rep. Intergov. Panel Clim. Chang.* 659–740 (2013). doi:10.1017/ CBO9781107415324.018

3. Ciais, P. *et al.* The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. *Chang. IPCC Clim.* 465–570 (2013). doi:10.1017/CBO9781107415324.015

4. Burger, N. *et al.* 'Summary'. in *Outcome Evaluation of U.S. Department of State Support for the Global Methane Initiative* Xi–Xviii (RAND Corporation, 2013, 2013).

5. Ahern, E. P., Deane, P., Persson, T., Ó Gallachóir, B. & Murphy, J. D. A perspective on the potential role of renewable gas in a smart energy island system. *Renew. Energy* **78,** 648–656 (2015).

6. Hahn, H., Krautkremer, B., Hartmann, K. & Wachendorf, M. Review of concepts for a demand-driven biogas supply for flexible power generation. *Renew. Sustain. Energy Rev.* **29,** 383–393 (2014).

7. Department of Energy and Climate Change. The Future of Heating : Meeting the challenge. *Decc* (2013).

8. Cecchi, F. & Cavinato, C. Anaerobic digestion of bio-waste: A mini-review focusing on territorial and environmental aspects. *Waste Manag. Res.* **33,** 429–438 (2015).

9. Smith, K. A., Smith, K. A. & Conen, F. Impacts of land management on fluxes of trace greenhouse gases. *Soil Use Manag.* **20,** 255–263 (2004).

10. Ye, R. *et al.* PH controls over anaerobic carbon mineralization, the efficiency of methane production, and methanogenic pathways in peatlands across an ombrotrophic-minerotrophic gradient. *Soil Biol. Biochem.* **54,** 36–47 (2012).

11. Mosier, A. R. *et al.* Mitigating agricultural emissions of methane. *Climatic Change* **40,** 39–80 (1998).

12. Sch'del, C. *et al.* Potential carbon emissions dominated by carbon dioxide from

thawed permafrost soils. *Nat. Clim. Chang.* **6,** 950–953 (2016).

13. Hugelius, G. *et al.* Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences* **11,** 6573–6593 (2014).

14. Minderlein, S. & Blodau, C. Humic-rich peat extracts inhibit sulfate reduction, methanogenesis, and anaerobic respiration but not acetogenesis in peat soils of a temperate bog. *Soil Biol. Biochem.* **42,** 2078–2086 (2010).

15. Bodirsky, B. L., Rolinski, S., Biewald, A. & Weindl, I. Global Food Demand Scenarios for the 21 st Century. *PLoS One* 1–27 (2015). doi:10.5281/zenodo.31008

16. Leahy, S. C. *et al.* The Genome Sequence of the Rumen Methanogen Methanobrevibacter ruminantium Reveals New Possibilities for Controlling Ruminant Methane Emissions. *PLoS One* **5,** e8926 (2010).

17. Zhou, Y. Y. *et al.* Inhibition of rumen methanogenesis by tea saponins with reference to fermentation pattern and microbial communities in Hu sheep. *Anim. Feed Sci. Technol.* **166–167,** 93–100 (2011).

18. Brown, E. G. *et al.* Effects of oral nitroethane administration on enteric methane emissions and ruminal fermentation in cattle. *Anim. Feed Sci. Technol.* **166–167,** 275–281 (2011).

19. Knight, T. *et al.* Chloroform decreases rumen methanogenesis and methanogen populations without altering rumen function in cattle. *Anim. Feed Sci. Technol.* **166–167,** 101–112 (2011).

20. Attwood, G. T. *et al.* Exploring rumen methanogen genomes to identify targets for methane mitigation strategies. *Anim. Feed Sci. Technol.* **166–167,** 65–75 (2011).

21. Kelly, W. J. *et al.* The complete genome sequence of the rumen methanogen Methanobacterium formicicum BRM9. *Stand. Genomic Sci.* **9,** 1–8 (2014).

22. Kelly, W. J. *et al.* The complete genome sequence of the rumen methanogen Methanobrevibacter millerae SM9. *Stand. Genomic Sci.* **11,** 1–9 (2016).

23. Nathani, N. M. *et al.* Comparative evaluation of rumen metagenome community using qPCR and MG-RAST. *AMB Express* **3,** 1–8 (2013).

24. Das, S. & Adhya, T. K. Dynamics of methanogenesis and methanotrophy in tropical

paddy soils as influenced by elevated CO2and temperature interaction. *Soil Biol. Biochem.* **47,** 36–45 (2012).

25. Cheng, K., Ogle, S. M., Parton, W. J. & Pan, G. Predicting methanogenesis from rice paddies using the DAYCENT ecosystem model. *Ecol. Modell.* **261–262,** 19–31 (2013).

26. Katayanagi, N. *et al.* Development of a method for estimating total CH4emission from rice paddies in Japan using the DNDC-Rice model. *Sci. Total Environ.* **547,** 429–440 (2016).

27. Huang, Y., Sass, R. L. & Fisher, Jr, F. M. A semi-empirical model of methane emission from flooded rice paddy soils. *Glob. Chang. Biol.* **4,** 247–268 (1998).

28. Rincones, J., Zeidler, A. F., Grassi, M. C. B., Carazzolle, M. F. & Pereira, G. A. G. The golden bridge for nature: The new biology applied to bioplastics. *Polym. Rev.* **49,** 85–106 (2009).

29. Parizzi, L. P. *et al.* The genome sequence of *Propionibacterium acidipropionici* provides insights into its biotechnological and industrial potential. *BMC Genomics* **13,** 562 (2012).

30. Baumann, I. & Westermann, P. Microbial Production of Short Chain Fatty Acids from Lignocellulosic Biomass: Current Processes and Market. *Biomed Res. Int.* **2016,** (2016).

31. Gonzalez-Garcia, R. *et al.* Microbial Propionic Acid Production. *Fermentation* **3,** 21 (2017).

32. Nazem-Bokaee, H., Gopalakrishnan, S., Ferry, J. G., Wood, T. K. & Maranas, C. D. Assessing methanotrophy and carbon fixation for biofuel production by Methanosarcina acetivorans. *Microb. Cell Fact.* **15,** 1–13 (2016).

33. Tokumoto, H. & Tanaka, M. Novel anaerobic digestion induced by bacterial components for value-added byproducts from high-loading glycerol. *Bioresour. Technol.* **107,** 327–332 (2012).

34. Comino, E., Riggio, V. A. & Rosso, M. Biogas production by anaerobic co-digestion of cattle slurry and cheese whey. *Bioresour. Technol.* **114,** 46–53 (2012).

35. Ferrera, I. & Sánchez, O. Insights into microbial diversity in wastewater treatment systems: How far have we come? *Biotechnol. Adv.* **34,** 790–802 (2016).

36. Cooney, C. L. & Wise, D. L. Thermophilic anaerobic digestion of solid waste for fuel gas production. *Biotechnol Bioeng* **17,** 1119–1135 (1975).

37. DEFRA. Waste water treatment in the United Kingdom. *Defra* 49 (2012).

38. Sheik, A. R., Muller, E. E. L. & Wilmes, P. A hundred years of activated sludge: Time for a rethink. *Front. Microbiol.* **5,** 1–7 (2014).

39. Sun, L., Pope, P. B., Eijsink, V. G. H. & Schnürer, A. Characterization of microbial community structure during continuous anaerobic digestion of straw and cow manure. *Microb. Biotechnol.* **8,** 815–827 (2015).

40. Tsavkelova, E. *et al.* The structure of the anaerobic thermophilic microbial community for the bioconversion of the cellulose-containing substrates into biogas. *Process Biochem.* 1–14 (2018). doi:10.1016/j.procbio.2017.12.006

41. Ragsdale, S. W. & Pierce, E. Acetogenesis and the Wood-Ljungdahl pathway of CO2 fixation. *Biochim. Biophys. Acta - Bioenerg.* **1784,** 1873–1898 (2009).

42. Schlegel, K., Welte, C., Deppenmeier, U. & Müller, V. Electron transport during aceticlastic methanogenesis by Methanosarcina acetivorans involves a sodium-translocating Rnf complex. *FEBS J.* **279,** 4444–4452 (2012).

43. Borrel, G. *et al.* Comparative genomics highlights the unique biology of Methanomassiliicoccales, a Thermoplasmatales-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* **15,** (2014).

44. Borrel, G. *et al.* Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis. *Genome Biol. Evol.* **5,** 1769–1780 (2013).

45. Sundberg, C. *et al.* 454 Pyrosequencing Analyses of Bacterial and Archaeal Richness in 21 Full-Scale Biogas Digesters. *FEMS Microbiol. Ecol.* **85,** 612–626 (2013).

46. Delbès, C., Moletta, R. & Godon, J.-J. Monitoring of activity dynamics of an anaerobic digester bacterial community using 16S rRNA polymerase chain reaction–single-strand conformation polymorphism analysis. *Environ. Microbiol.* **2,** 506–515 (2000).

47. Gagen, E. J., Padmanabha, J., Denman, S. E. & McSweeney, C. S. Hydrogenotrophic culture enrichment reveals rumen Lachnospiraceae and Ruminococcaceae

acetogens and hydrogen-responsive Bacteroidetes from pasture-fed cattle. *FEMS Microbiol. Lett.* **362,** 1–8 (2015).

48. Hanreich, A. *et al.* Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. *Syst. Appl. Microbiol.* **36,** 330–338 (2013).

49. Ganesan, A. *et al.* Cloacibacillus evryensis gen. nov., sp. nov., a novel asaccharolytic, mesophilic, amino-acid-degrading bacterium within the phylum ‘Synergistetes’, isolated from an anaerobic sludge digester. *Int. J. Syst. Evol. Microbiol.* **58,** 2003–2012 (2008).

50. Kotsyurbenko, O. R., Glagolev, M. V., Nozhevnikova, A. N. & Conrad, R. Competition between homoacetogenic bacteria and methanogenic archaea for hydrogen at low temperature. *FEMS Microbiol. Ecol.* **38,** 153–159 (2001).

51. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* **74,** 5088–5090 (1977).

52. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87,** 4576–4579 (1990).

53. Brock, T. D., Brock, K. M., Belly, R. T. & Weiss, R. L. Sulfolobus: A new genus of sulfur-oxidizing bacteria living at low pH and high temperature. *Arch. Mikrobiol.* **84,** 54–68 (1972).

54. Woese, C. R., Magrum, L. J. & Fox, G. E. Archaebacteria. *J. Mol. Evol.* **11,** 245–252 (1978).

55. Borrel, G. *et al.* Genome sequence of ‘Candidatus Methanomethylophilus alvus’ Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens. *J. Bacteriol.* **194,** 6944–6945 (2012).

56. Tamburini, C. *et al.* Distribution and activity of Bacteria and Archaea in the different water masses of the Tyrrhenian Sea. *Deep. Res. Part II Top. Stud. Oceanogr.* **56,** 700–712 (2009).

57. Kaster, A. K. *et al.* More than 200 genes required for methane formation from H 2 and CO 2 and energy conservation are present in

methanothermobacter marburgensis and methanothermobacter thermautotrophicus. *Archaea* **2011,** (2011).

58. Heidrich, E. S., Curtis, T. P. & Dolfing, J. Determination of the internal chemical energy of wastewater. *Environ. Sci. Technol.* **45,** 827–832 (2011).

59. Peng, X., Börner, R. A., Nges, I. A. & Liu, J. Impact of bioaugmentation on biochemical methane potential for wheat straw with addition of Clostridium cellulolyticum. *Bioresour. Technol.* **152,** 567–571 (2014).

60. Vollmers, J., Wiegand, S. & Kaster, A. K. *Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! PLoS ONE* **12,** (2017).

61. Godon, J., Zumstein, E., Dabert, P. & Habouzit, R. I. C. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis . Molecular Microbial Diversity of an Anaerobic Digestor as Determined by Small-Subunit rDNA Sequence Analysis. *Apllied Environ. Microbiol.* **63,** 2802–2813 (1997).

62. Podell, S. *et al.* Assembly-Driven Community Genomics of a Hypersaline Microbial Ecosystem. *PLoS One* **8,** (2013).

63. Sharon, I. & Banfield, J. F. Genomes from Metagenomics. *Science (80-.).* **342,** 1057–1058 (2013).

64. Leach, A. L. B., Chong, J. P. J. & Redeker, K. R. SSuMMo: Rapid analysis, comparison and visualization of microbial communities. *Bioinformatics* **28,** 679–686 (2012).

65. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31,** 1674–1676 (2015).

66. Lu, Y. Y., Chen, T., Fuhrman, J. A., Sun, F. & Sahinalp, C. COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* **33,** 791–798 (2017).

67. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3,** e1165 (2015).

68. Keseler, I. M. *et al.* The EcoCyc database: Reflecting new knowledge about

Escherichia coli K-12. *Nucleic Acids Res.* **45,** D543–D550 (2017).

69. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44,** D471–D480 (2016).

70. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2016).

71. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428,** 726–731 (2016).

72. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Cold Spring Harb. Lab. Press Method* **1,** 1–31 (2015).

73. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

74. Wu, H., Zhang, Z., Hu, S. & Yu, J. On the molecular mechanism of GC content variation among eubacterial genomes. *Biol. Direct* **7,** 2 (2012).

75. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5,** 163 (2004).

76. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.* **7,** 203–214 (2000).

77. Coordinators, N. R. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **45,** D12–D17 (2017).

78. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **33,** 34–38 (2005).

79. Yashiro, Y. *et al.* Methanoregula formicica sp. nov., a methane-producing archaeon isolated from methanogenic sludge. *Int. J. Syst. Evol. Microbiol.* **61,** 53–59 (2011).

80. Castelle, C. J. *et al.* Genomic expansion of domain archaea highlights roles for

organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25,** 690–701 (2015).

81. Sedlar, K., Kupkova, K. & Provaznik, I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.* **15,** 48–55 (2017).

82. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science (80-.).* **351,** (2016).

83. Bondoso, J. *et al.* Aquisphaera giovannonii gen. nov., sp. nov., a planctomycete isolated from a freshwater aquarium. *Int. J. Syst. Evol. Microbiol.* **61,** 2844–2850 (2011).

84. Scheuner, C. *et al.* Complete genome sequence of Planctomyces brasiliensis type strain (DSM 5305T), phylogenomic analysis and reclassification of Planctomycetes including the descriptions of Gimesia gen. nov., Planctopirus gen. nov. and Rubinisphaera gen. nov. and emended des. *Stand. Genomic Sci.* **9,** 1–18 (2014).

85. Bhandari, V. & Gupta, R. S. Molecular signatures for the phylum Synergistetes and some of its subclades. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **102,** 517–540 (2012).

86. Suresh, G. *et al.* Description of rhodobacter azollae sp. nov. and rhodobacter lacus sp. nov. *Int. J. Syst. Evol. Microbiol.* **67,** 3289–3295 (2017).

87. Strnad, H. *et al.* Complete genome sequence of the photosynthetic purple nonsulfur bacterium Rhodobacter capsulatus SB 1003. *J. Bacteriol.* **192,** 3545–3546 (2010).

88. Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W. & Banfield, J. F. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.* **12,** R44 (2011).

89. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9,** 1–11 (2018).

90. Davis, A. M. *et al.* Using MinION nanopore sequencing to generate a de novo eukaryotic draft genome: preliminary physiological and genomic description of the extremophilic red alga Galdieria sulphuraria strain SAG 107.79. *Doi.Org* 076208 (2016). doi:10.1101/076208

91. Batovska, J., Lynch, S. E., Rodoni, B. C., Sawbridge, T. I. & Cogan, N. O. Metagenomic arbovirus detection using MinION nanopore sequencing. *J. Virol. Methods* **249,** 79–84 (2017).

92. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** (2014).