Determining the Microbial Community Dynamics of Anaerobic Digestion Using Metagenomics

Henry Charles George Nicholls

PhD

University of York

Biology

October 2015

Abstract

Anaerobic Digestion (AD) is a biologically mediated technology that is used as a method for managing and obtaining energy from organic waste materials. Through the biological action of Bacteria, Fungi and Archaea, in the absence of oxygen, the organic waste is converted to biogas, mainly methane, which can be used as a fuel source. This gas can be burned to generate electricity, heat, injected into the grid or used to fuel vehicles.

I have developed a single stage, lab scale anaerobic digester that is a reflection of full-scale process systems. This model reactor facilitates the collection of samples for metagenomic sequencing, along with process data, providing an insight to the AD process. Three experiments were carried out (using the lab model) to determine (i) the dynamic changes that occur in microbial AD communities, (ii) the rate at which these communities change and (iii) if the observed changes are comparable between numerous systems run under the same conditions.

The use of amplicon sequencing appears to be a common method used to study the composition of microbial communities, especially in AD, but this method is prone to inaccuracies and so alternative methods were developed, as described in this thesis. By applying the use of shotgun metagenomic sequencing, combined with various contig assemblers and a custom clustering method, more detail on the microbes present and their functions in AD is obtained compared to targeted sequencing. Pipelines to interpret large datasets generated through Next Generation Sequencing (NGS) have been developed and utilised in this project. We have identified microbes that are present within the AD system, and the time-scale of the dynamic changes. This method has also revealed novel methanogens that are important in the AD process.

Table of contents

List of figures

List of tables

Acknowledgements

I would like to thank both the sponsors for this very interesting project – Biotechnology and Biological Sciences Research Council (BBSRC) and the industrial partners, Clearfleau Ltd.

I would also like to thank both my supervisors, James Chong and Kelly Redeker, for their guidance and insight during the project, along with the members of the Chong group, particularly Alex and Emma. Thanks also to Mark Bentley for his input in the design and construction of the lab scale AD system. Furthermore I would like to thank the members of my Thesis Advisory Panel, Thorunn Helgason and Phil Ineson, and Peter Ashton in the Technology Facility.

Finally, I would like to thank my parents for always encouraging me to achieve the best I possibly can and to Hannah for her continued support.

Declaration

I, Henry Nicholls, declare that this thesis is a presentation of original work and I 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 contig assembly and clustering, as described in Chapters 5.3.4.1 and 5.3.4.2, was performed by Dr Peter Ashton
- The contig assembly and clustering, as described in Chapter 5.3.5.2 was performed by Dr James Chong
- The script for automated BLAST was written by Dr Peter Ashton
- The phylogenetic trees and files for genome mapping (Chapters 5.3.5.5 and 5.3.5.6) were produced by Dr James Chong

1 Introduction

Alternatives to fossil fuels are now being sought due to limited resource availability and the heavy dependence on fossil fuels in modern society. Adding to this, global political pressures have been adopted to ensure that carbon emissions are reduced and renewable energy technology presence and supply is increased.

The supply of renewable energy is now a priority in the UK to ensure that there is a constant, reliable, sustainable and cleaner supply of energy. The supply of electricity generated from renewable sources now contributes a significant proportion (18 %), with this number increasing from previous years, but the dependence on fossil fuels, such as coal, gas and oil is still high (Figure 1.1). An increased presence of renewable technologies therefore would ensure that there is a lower requirement for fossil fuels. The UK has a target to ensure that the total energy supplied amounts to 15 % from renewable technologies by 2020.

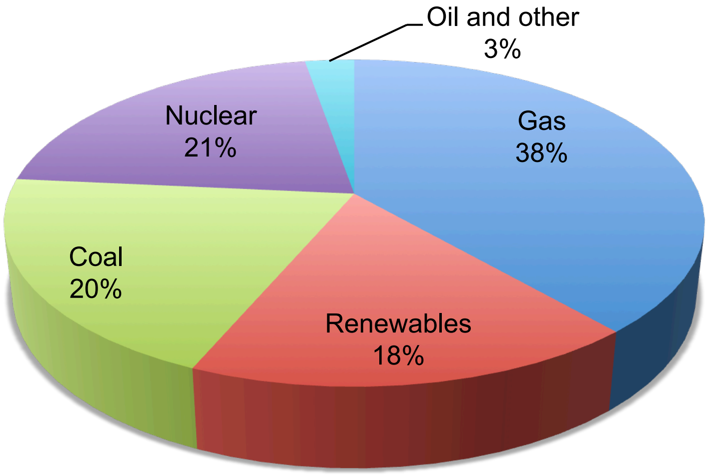


Figure 1.1. UK electricity generated from fuel source in 2014. Data taken from DEFRA (www.gov.uk).

Along with reducing the reliance for fossil fuels, renewable technologies are cleaner, as fewer carbon dioxide emissions are produced. Carbon dioxide is classed as a greenhouse gas and is linked to climate change.

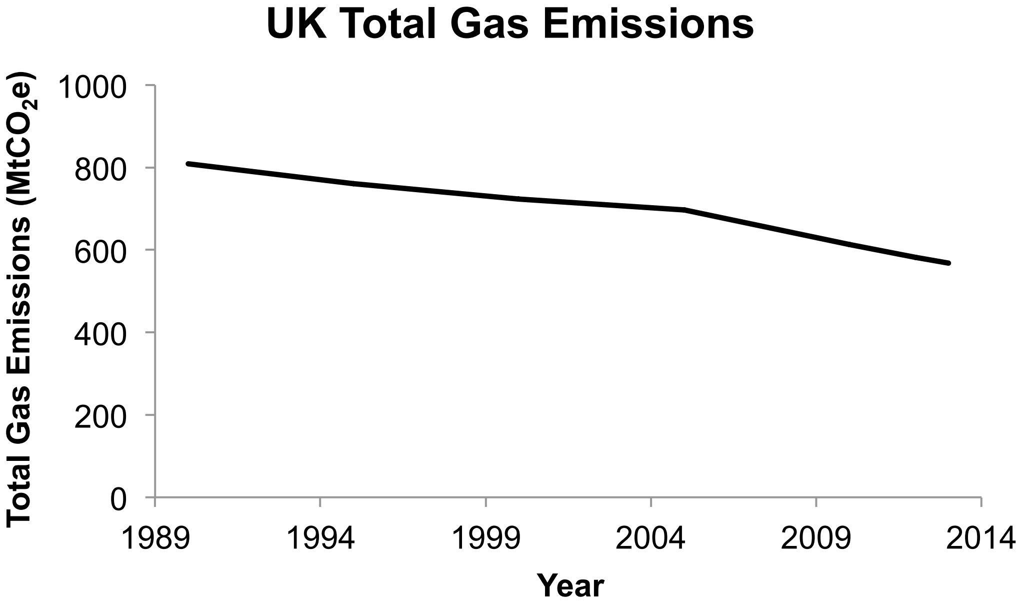


Figure 1.2. Total gas emissions for the UK. Adapted from Department of Energy & Climate Change 2014.

Total gas emissions in the UK have been decreasing (Figure 1.2) and this is partly due to a drive towards using renewable energy technologies. In 2013, the generation of electricity accounted for 33 % of total gas emissions, and therefore renewable energy technologies have the potential to significantly reduce this figure, especially as some have little to no carbon dioxide emissions. The transport sector accounted for around 21 % of total gas emissions in 2013, mainly through the use of petrol and diesel, so alternative sources would reduce the level of emissions from this sector, along with less fossil fuel use.

There are numerous renewable, low carbon energy technologies available such as wind, tidal, thermal and solar, each with their own advantages.

Anaerobic Digestion (AD) is a biologically mediated technology that is gaining more prominence as a process that can be used for the generation of renewable energy sources. The Anaerobic Digestion process involves a complex community of microbes that breakdown organic waste materials in the absence of oxygen, which results in the production of biogas, mainly methane, which is a fuel source (Figure 1.3). AD can contribute towards achieving increasingly stringent targets for a greater proportion of energy to be derived from renewable sources as well as a solution to the increased need for waste materials to be processed in a responsible manner. Currently there are around 320 AD sites in the UK.

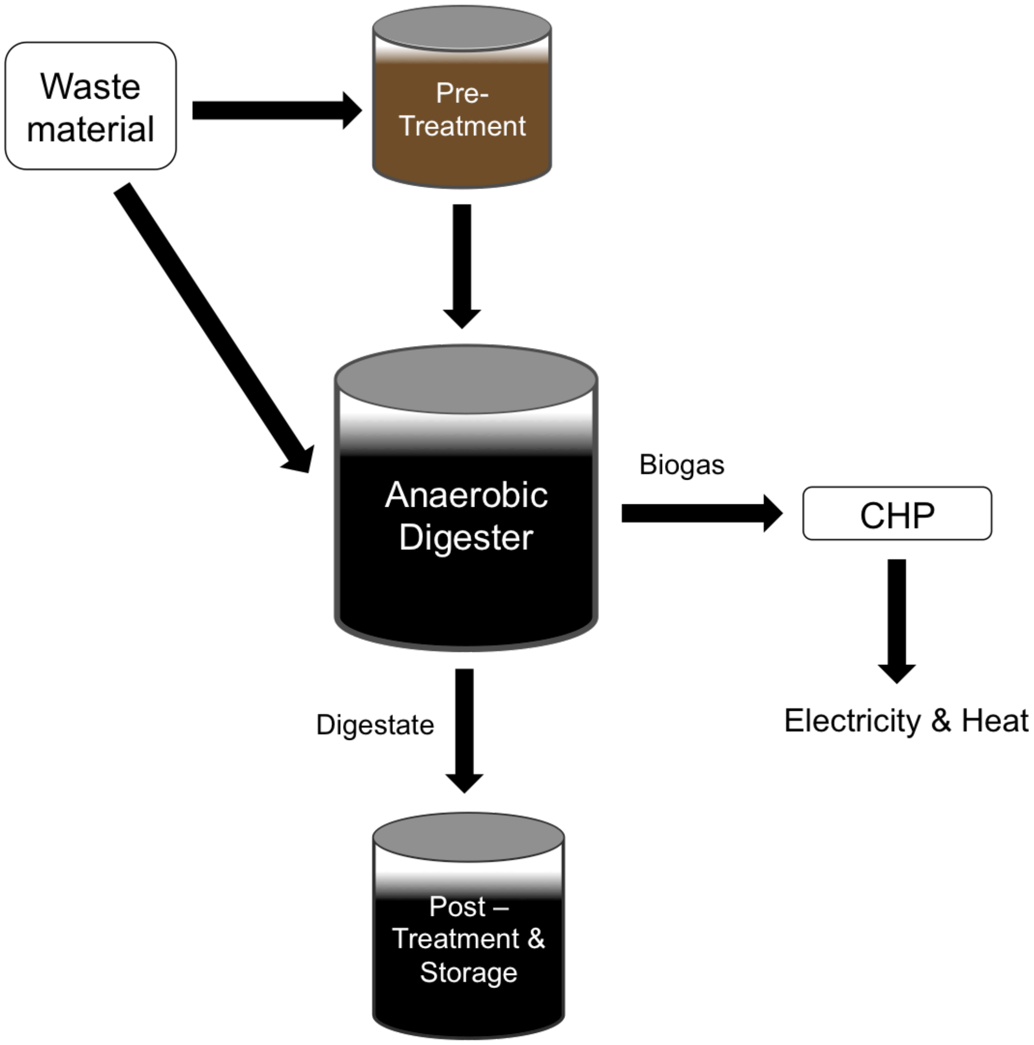


Figure 1.3. Overview of the AD process where organic waste materials are pre-treated before being placed into the digester. The microbes in the AD system utilise these compounds to produce biogas. The remaining digestate can be further processed and used e.g. as a fertiliser.

Along with the requirement for alternatives to fossil fuels, there is a need for managing waste materials in a responsible manner to ensure that less is placed into landfill. It is estimated that the UK alone generates up to 15 million tonnes of food waste annually, and AD has the potential to ensure much of this waste is not placed in landfill, whilst generating a fuel source from this. In 2013, 4 % of UK total gas emissions were from waste management (mainly landfill), with 91 % of this being methane. Therefore AD has the potential to reduce this number.

1.1 Benefits and Challenges

AD has many benefits that allow it to be a predominant contender for use in the generation of renewable fuels. The main advantage of this technology is that as the complex community of microbes break down the organic waste, the production of biogas results, which is mainly comprised of methane and carbon dioxide. The biogas produced can be burned for the generation of electricity via a Combined Heat and Power unit (CHP), with low grade heat as a by-product. If the AD unit is, for example, on a manufacturing site, this generated electricity and heat can sometimes be used on site ensuring that the manufacturer uses less fossil fuel generated electricity. As fossil fuel resources are limited (Krakat et al., 2011), the move towards greener technologies is important. Also, if fewer resources are taken from the national grid, this means that company overheads would be reduced, and the generated gas or electricity could even be sold back into the grid. The use of AD therefore has clear environmental and financial benefits. It has been previously estimated that AD has the potential to provide up to 50 % of the UK gas residential requirements (National Grid, 2009). The gas can also be upgraded and placed into the grid or can be used as a fuel source for vehicles (Goulding and Power, 2013).

Another advantage of AD is that it is a method for managing waste materials, and the preparation of the waste to be placed into the digester is often minimal. There are large costs associated with the disposal of solid waste and the treatment of liquid wastes. Any technologies that can prevent solid materials from being placed in landfill are clearly important. Any solid material that is removed from the AD system, digestate, can be further processed, according to

EU legislation, such as being pasteurised and stored (Ariunbaatar et al., 2014), and this can be used as a fertiliser due to the high nitrogen levels, meaning that potentially fewer synthetic fertilisers are required. The versatility of AD systems to accept a variety of waste streams is advantageous meaning the technology can be used in a wide number of applications. These can include solid and liquid based wastes from the foods and drinks industry, such as dairy, brewery and confectionary, along with bio-diesel waste, sewage and farm wastes. Other advantages include that there is a low sludge yield (biomass) (Chen et al., 2008), especially compared to aerobic treatments, the systems require low nutrient input and there are low operational and maintenance associated costs (De Vrieze et al., 2012, Wijekoon et al., 2011).

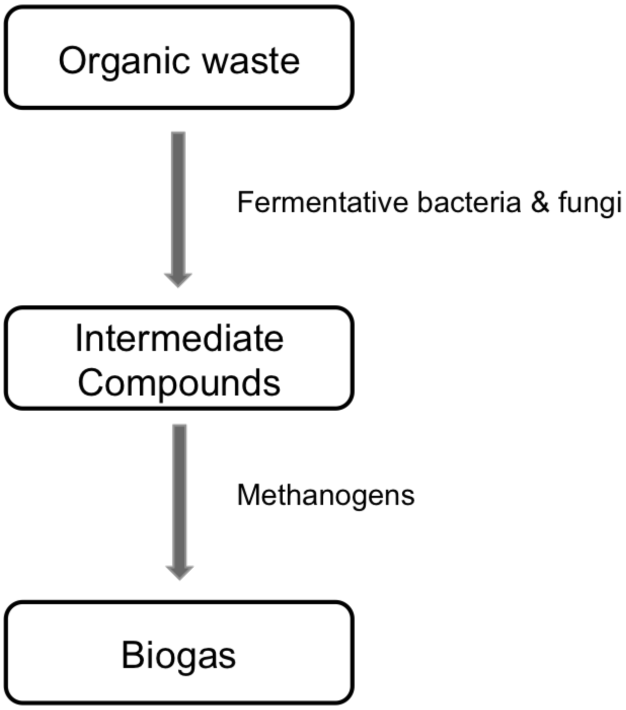


Figure 1.4. Overview of the conversion of organic waste materials to biogas mediated by different microbes.

There are challenges associated with the technology, with the main one being that as AD systems are comprised of a complex community of microbes (Figure 1.4), which if disrupted, can cause the AD system to stall, such as reactor acidification. This is especially true for the methanogen community, which can be sensitive to slight changes in operational and environmental conditions, such as temperature, pH, or organic loading rates (De Vrieze et al., 2012). If the methanogen activity is reduced, or the organic loading rates are too high, an accumulation of organic acids can result. If methanogens do not utilise these, this in turn can cause the pH to become more acidic and high levels of organic acids can have a toxic effect on the methanogens (Franke-Whittle et al., 2014), further stalling the system. For this reason, it is common that theoretical maximum organic loading rates are not reached. Long start up times are also another drawback of this technology, as the microbes in the system require time to acclimatise to the specific components of the waste. Although the systems have versatility to different waste materials, the feedstock composition cannot be abruptly changed, as again, the microbial community in the system would need time to adapt to this. Another challenge associated with AD is that systems have been reported to have foaming issues. Foam is a liquid-gas dispersion (Kougias et al., 2014) that forms in the reactor and can result in operational disruption. Ultimately, any disruption to the AD process could have both environmental and economical costs.

1.2 Biochemistry and Microbiology

Anaerobic Digestion is a process that involves a complex community of Bacteria and Archaea, all with different metabolic capacities, working in a syntrophic fashion (Pind et al., 2003) to break down organic wastes to produce biogas, mainly comprising of methane (50-70 %) and carbon dioxide (30-50 %). There is an arbitrary division of the biochemical reactions the substrates are subjected to in the AD process; hydrolysis, where polymers are hydrolysed to monomers, acidogenesis, where these monomers are converted to intermediate compounds, acetogenesis, where intermediates are further broken down to acetate and finally methanogenesis (Heeg et al., 2014), the formation of biogas (Figure 1.5).

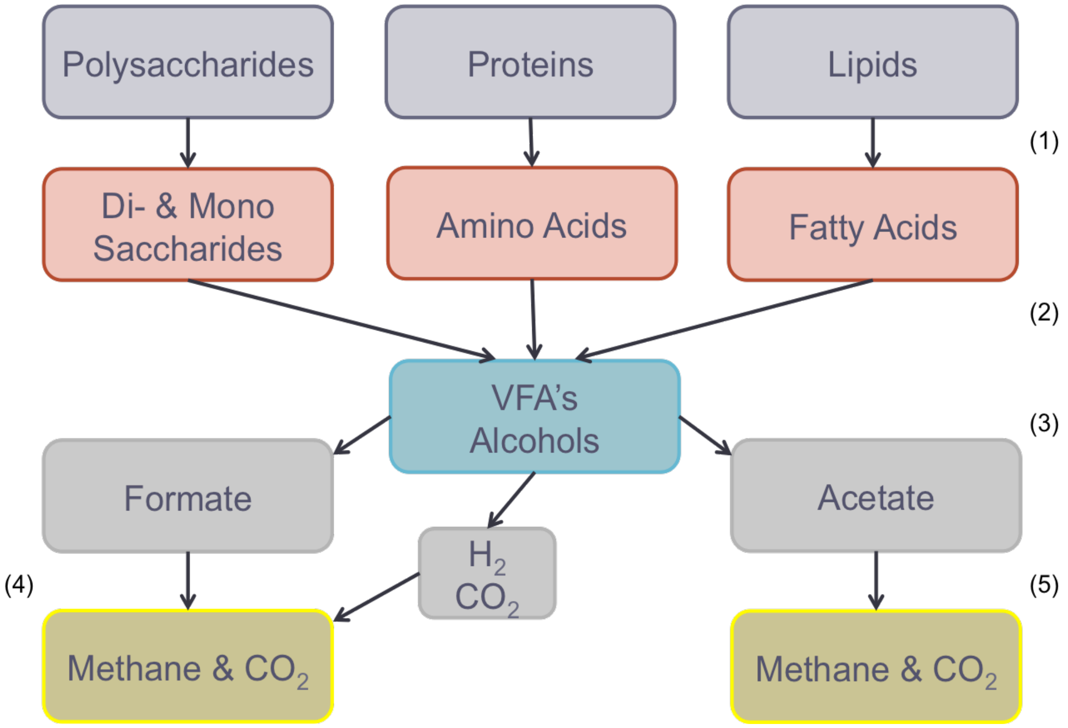


Figure 1.5. Key processes involved in the AD process. 1. Hydrolysis 2. Acidogenesis 3. Acetogenesis 4. Hydrogenotrophic methanogenesis 5. Aceticlastic methanogenesis (adapted from Madsen et al. 2011).

1.2.1 Hydrolysis

Hydrolysis is the breakdown of polymers to soluble monomers, such as polysaccharides, proteins and lipids to monosaccharides, amino acids and fatty acids, often performed by fermentative bacteria and fungi. This step is carried out by specific extracellular enzymes such as lipases, proteases and amylases, produced by these fermentative microbes (De Francisci et al., 2015). Hydrolysis is regarded as a rate limiting step in the AD process (Ge et al., 2011). This is particularly important for wastes containing high levels of insoluble particular matter as these have to be solubilised (Gavala et al., 2003) and then hydrolysed from polymers to monomers (Xue et al., 2015).

There are numerous examples of microbes that can perform the hydrolysis step. There are known organisms belonging to the Phylum Firmicutes, including the genus *Clostridium*, which can degrade complex polymers, e.g. cellulose (Nelson et al., 2011) and Bacteroidetes that are proteolytic bacteria (Rivière et al., 2009).

1.2.2 Acidogenesis

Acidogenic microbes use the soluble monomers produced during hydrolysis to form reduced intermediate products, such as a variety of organic acids, (including acetate and formate), alcohols (such as methanol and ethanol), ketones, H_2 and CO_2 (Franke-Whittle et al., 2014), in the process known as acidogensis. Other byproducts such as ammonia and hydrogen sulphide (H_2S) may also be formed in this step (Appels et al., 2008).

Eq. 1.1. $C_6H_{12}O_6 + 2H_2O \longrightarrow 2CH_3COOH + 2CO_2 + 4H_2$

Eq 1.2. $C_6H_{12}O_6 + 2H_2 \longrightarrow 2CH_3CH_2COOH + 2H_2O$

Examples of acidogenic reactions include the formation of acetic acid (Eq. 1.1) and propionic acid (Eq. 1.2) from glucose.

Organisms that perform the acidogenic reactions include *Spirochaetes*, specifically the genus *Cloacamonas*, which is commonly found in AD systems and it has been proposed that this organism has the capacity to ferment amino acids (Pelletier et al., 2008) and *Cloacamonas* species are syntrophic fermentation bacteria (Razaviarani and Buchanan, 2014).

1.2.3 Acetogenesis

In the third stage, acetogenesis, a variety of syntrophic microbes perform numerous reactions and are closely linked with the methanogens.

Reactions that occur include the oxidation of organic acids, such as butyrate (Eq. 1.3) and propionate (Eq. 1.4) to acetate.

Eq 1.3. $CH_3CH_2CH_2COO^- + 2H_2O \longrightarrow 2CH_3COO^- + H^+ + 2H_2$

Eq 1.4. $CH_3CH_2COO^- + 2H_2O \longrightarrow CH_3COO^- + CO_2 + 3H_2$

Eq 1.5. $CH_3COO^- + 4H_2O \longrightarrow 2HCO_3^- + H^+$

The oxidation of butyrate and propionate is thermodynamically unfavourable and therefore a close association with methanogens and other syntrophic microbes that utilise acetate and hydrogen is important (Wang et al., 2013). This interspecies hydrogen transfer between hydrogen producers and consumers ensures that hydrogen remains in low concentrations (McInerney et al., 2009). Examples of known organic acid oxidisers include *Syntrophomonas wolfei* and *Syntrophobacter fumaroxidans* that degrade butyrate and propionate respectively (Stams and Plugge, 2009). It is estimated that propionate, when converted to acetate, hydrogen and carbon dioxide can account for up to 35 % of methane produced (Wagner et al., 2014).

Acetate oxidation (Eq. 1.5) can also occur, producing hydrogen (Lee et al., 2015). This reaction can only take place when the hydrogen concentration is low and this requires hydrogen consuming hydrogentropic methanogens (Moestedt et al., 2014), such as *Methanoculleus*. Of the characterised syntrophic acetate oxidisers, most belong to the class *Clostridia* (Kampmann et al., 2012).

1.2.4 Methanogenesis

The final stage in the AD process is conducted by methanogens, a group of organisms belonging to the domain Archaea. As well as being common to the AD process, methanogens have been found in a variety of anaerobic environments (Wilkins et al., 2015). The methanogen community is less diverse than the bacterial community and often accounts for a smaller proportion than the bacterial population, with previous reports suggesting around 10 % relative abundance in AD systems (Wirth et al., 2012). Numerous factors often influence the Archaea community such as the methanogen diversity and activity. One example from a study by Franke-Whittle et al. (2014), showed that mesophilic digesters have a greater diversity of methanogens compared to thermophillic digesters. As with the bacterial community, having a diverse methanogen community in AD systems is beneficial for system stability as this group of organisms can be disrupted by several environmental factors, and it not unusual for conditions to vary in AD systems. This could include variations in feedstock composition and digester temperature.

Methanogens can utilise a range of substrates such as acetate, formate, and other one carbon compounds, as well as carbon dioxide and hydrogen (Ziganshin et al., 2011). Methanogens are grouped based on the substrates that they utilise; aceticlastic or hydrogenotrophic (Razaviarani and Buchanan, 2014).

Aceticlastic methanogens utilise acetate to form methane (Eq. 1.6). Both *Methanosarcina* and *Methansaeta*, belonging to the class *Methanomicrobia* are acetate utilising methanogens. Interestingly, *Methanosarcina* can switch between the aceticlastic and hydrogenotrophic pathways (Qu et al., 2009, De Vrieze et al., 2012), so can use a range of compounds such as acetate, methanol (Jäger et al., 2009) (Eq 1.8), along with hydrogen and carbon dioxide (Yu et al., 2014). The aceticlastic methanogens have shown to be affected by various conditions in the AD systems, such as dominating when hydrogen levels are low, but different aceticlastic methanogens thrive under varying acetate concentrations. For example, *Methanosarcina* has been shown to utilise acetate over *Methanosaeta* when the acetate concentration is over 1 mM, but below that concentration *Methanosaeta* is the dominant methanogen (Razaviarani and Buchanan, 2014).

Eq 1.6. $CH_3COO^- + H^+ \longrightarrow CH_4 + CO_2$

Eq 1.7. $4H_2 + CO_2 \longrightarrow CH_4 + 2H_2O$

Eq 1.8. $4CH_3OH \longrightarrow 3CH_4 + CO_2 + 2H_2O$

The hydrogenotrophic methanogens utilise hydrogen and carbon dioxide (Eq. 1.7), along with other one carbon compounds, such as formate to produce methane. Examples of these methanogens include the genera *Methanomicrobium*, *Methanospirillium* and *Methanoculleus*, amongst others belonging to the class *Methanomicrobia*. These methanogens tend to be found in lower numbers in AD systems, but when the hydrogen levels increase, the proportion of these methanogens can increase (Razaviarani and Buchanan, 2014). An example of a hydrogenotrophic methanogen is *Methanospirillium*, which uses hydrogen and carbon dioxide preferentially over formate (Nelson et al., 2011). The pathways (and enzymes) involved in methanogenesis are displayed in Figure 1.6.

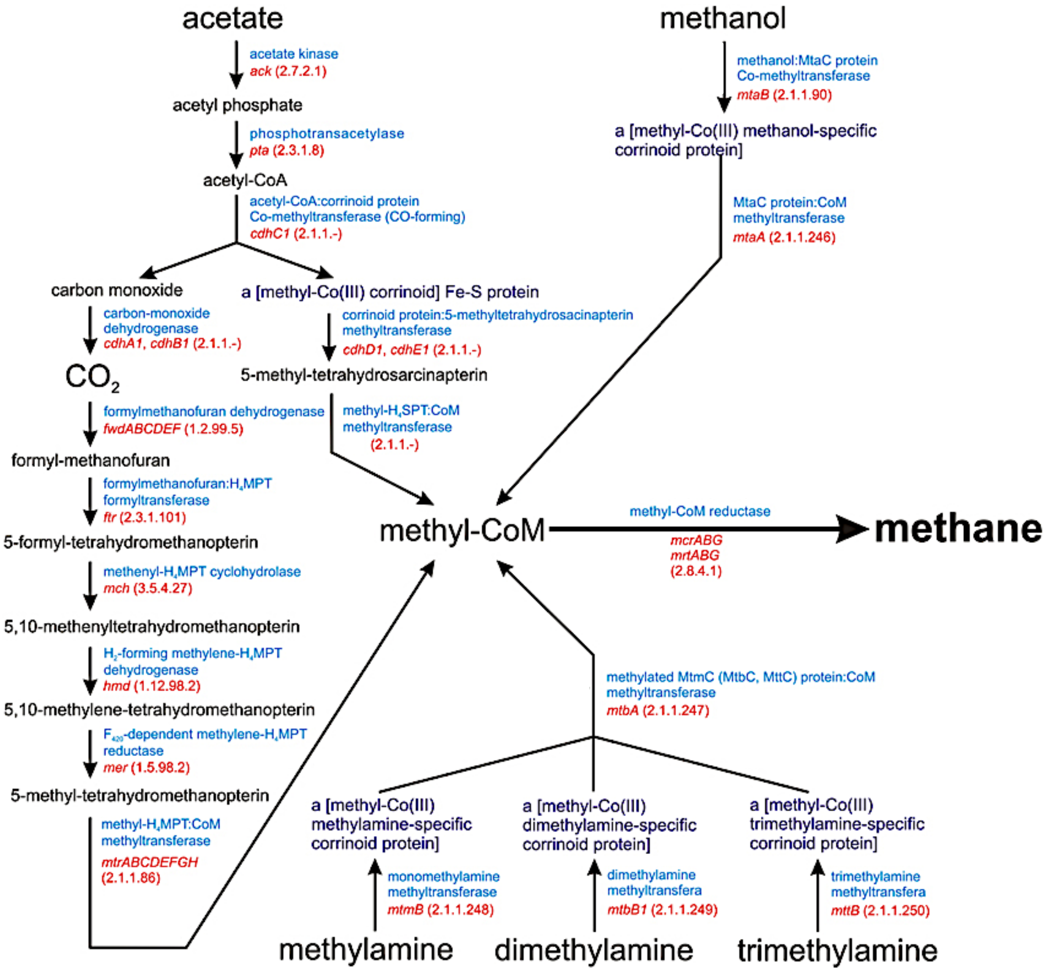

acetate
acetate kinase
ack (2.7.2.1)
acetyl phosphate
phosphotransacetylase
pta (2.3.1.8)
acetyl-CoA
acetyl-CoA:corrinoid protein Co-methyltransferase (CO-forming)
cdhC1 (2.1.1.-)
carbon monoxide
carbon-monoxide dehydrogenase
cdhA1, cdhB1 (2.1.1.-)
CO2
formylmethanofuran dehydrogenase
fwdABCDEF (1.2.99.5)
formyl-methanofuran
formylmethanofuran:H4MPT formyltransferase
ftr (2.3.1.101)
5-formyl-tetrahydromethanopterin
methenyl-H4MPT cyclohydrolase
mch (3.5.4.27)
5,10-methenyltetrahydromethanopterin
H2-forming methylene-H4MPT dehydrogenase
hmd (1.12.98.2)
5,10-methylene-tetrahydromethanopterin
F420-dependent methylene-H4MPT reductase
mer (1.5.98.2)
5-methyl-tetrahydromethanopterin
methyl-H4MPT:CoM methyltransferase
mtrABCDEFGH (2.1.1.86)
a [methyl-Co(III) corrinoid] Fe-S protein
corrinoid protein:5-methyltetrahydrosacinapterin methyltransferase
cdhD1, cdhE1 (2.1.1.-)
5-methyl-tetrahydrosarcinapterin
methyl-H4SPT:CoM methyltransferase
(2.1.1.-)
methanol
methanol:MtaC protein Co-methyltransferase
mtaB (2.1.1.90)
a [methyl-Co(III) methanol-specific corrinoid protein]
MtaC protein:CoM methyltransferase
mtaA (2.1.1.246)
methyl-CoM
methyl-CoM reductase
mcrABG
mrtABG
(2.8.4.1)
methane
methylated MtmC (MtbC, MttC) protein:CoM methyltransferase
mtbA (2.1.1.247)
a [methyl-Co(III) methylamine-specific corrinoid protein]
a [methyl-Co(III) dimethylamine-specific corrinoid protein]
a [methyl-Co(III) trimethylamine-specific corrinoid protein]
monomethylamine methyltransferase
mtmB (2.1.1.248)
dimethylamine methyltransfera
mtbB1 (2.1.1.249)
trimethylamine methyltransfera
mttB (2.1.1.250)
methylamine
dimethylamine
trimethylamine

Figure 1.6. Schematic diagram of the methanogenesis pathways (Dziewit et al., 2015)

1.3 Core Group of Microbes

Many studies have recently focused on trying to get a better understanding on the microbial communities involved in the AD process e.g. Jang et al. (2014), Lim et al. (2013) and Tian et al. (2015), amongst others, as shown in Table 1.1 and 1.2. AD is a technology that has been developed from an engineering perspective, but the microbiology still remains relatively unknown. Although numerous measurements are taken when running an AD system (Chapter 1.6) which act as indicators of system performance, there is the potential that microbial markers could also reveal and predict system performance.

In recent years our understanding of complex microbial communities, dynamics and function has increased significantly due to the development of Next Generation Sequencing (NGS) technologies. NGS technology provides the ability to sequence most complex communities, at more depth (Whiteley et al., 2012), giving more comprehensive information, as the volume of data generated is so much greater compared to previous technologies. Prior to NGS, the sequencing of complex communities was only possible using expensive, low resolution technologies, but as this technology has developed, the cost per base has dramatically decreased. With increased sequencing output, available at lower costs, our ability to understand complex communities has increased, ultimately giving greater understanding of these subject areas. Recent examples include research in AD (Schlüter et al., 2008), human microbiome (Belda-Ferre et al., 2012) and soil (Souza et al., 2013).

The microbiology and microbial dynamics of AD is still relatively unknown. The roles played by methanogens in AD has been more widely studied, but the bacterial species that are responsible for hydrolysis and acidogenesis are not well understood (Keating et al., 2016). The nature in which the systems are run and samples taken also makes understanding the microbiology more challenging. Most AD systems will be run under slightly different conditions, which are optimised to the input material, most likely shaping the community structure. Additionally, different inoculum will have been obtained, giving a different starting microbial population. These factors often make identifying important microbes involved in the process difficult and therefore drawing conclusions from recent research can be challenging, as demonstrated in Table 1.1. Nonetheless it is possible to demonstrate that there are a common group of microbes across the studied AD systems, albeit at the phylum level.

Type of Digester	Temperature	Inoculum	Feedstock	Detection method	Description	Ref
6 litre, single stage lab scale digester	Mesophilic (35 °C)	Mesophilic sewage sludge AD system	Food waste	16S rRNA using 454 & qPCR	Increasing organic loading rates and monitoring microbial changes	Jang et al., 2014
5 litre, single stage lab scale digester	Mesophilic (35 °C)	Anaerobic sludge from mesophilic wastewater treatment plant	Food and sewage waste	16S rRNA	Investigating microbes involved in the co-digestion of food and sewage waste	Lim et al., 2013
6 litre, single stage lab scale digester	Mesophilic to thermophilic	Wastewater treatment plant	Sewage sludge	16S rRNA using 454	Increasing temperature from mesophilic to thermophilic to investigate changes in microbial communities	Tian et al., 2015
10 litre, single stage lab scale digester	Mesophilic (37 °C)	Waste activated sludge from wastewater treatment plant	Municipal wastewater and biodiesel waste glycerin	16S rRNA using 454	Looking at microbial dynamics at various loading rates using biodiesel waste glycerine	Razaviarani & Buchanan 2015
500 litre, full scale digester	Mesophilic (35 °C)	Not stated	Untreated corn straw	16S rRNA	Investigating the microbes involved in a digester that has corn straw as sole feedstock	Qiao et al., 2013

Table 1.1. Previous research paper themes conducted in AD.

Phylum								
Firmicutes	Proteobacteria	Bacteroidetes	Chloroflexi	Actinobacteria	Synergistes	Other	Unclassified	Ref
2.4 %	3.6 %	3.2 %	63.9 %	-	2.6 %	*Thermotogae* (5.6 %)	16.8 %	Jang et al., 2014[1]
15.9 %	-	35.4 %	7.1 %	4.9 %	27.9 %	-	8.5 %	Jang et al., 2014[2]
15.5 %	23.3 %	19.1 %	12.7 %	12.6 %	-	-	-	Tian et al., 2015[1]
36.8 %	12.4 %	-	-	14.3 %	5.7 %	*Thermotogae* (21.4 %)	-	Tian et al., 2015[2]
48.3 %	7.2 %	7.7 %	20.1 %	9.1 %	1.0 %	*Planctomycetes* (1.9 %)	-	Qiao et al.,2013
89.5 %	-	2.3 %	-	6.9 %	-	-	-	Garcia-Peña et al., 2011
9 %	18 %	11 %	32 %	-	-	-	12 %	Rivière et al., 2009

Table 1.2. Relative abundance data of bacterial phyla from AD systems.

Archaea - Order			
Methanosarcinales (A)	*Methanobacteriales* (H)	*Methanomicrobiales* (H)	Ref
81 %	5 %	11 %	Jang et al., 2014[1]
54 %	16 %	25 %	Jang et al., 2014[2]
26 %	15 %	48 %	Qiao et al., 2013
51 %	0.2 %	10 %	Rivière et al., 2009

Table 1.3. Relative abundance of methanogens in AD systems. (A) Aceticlastic (H) Hydrogenotrophic.

The five most common phyla in AD studies are Firmicutes, Chloroflexi, Bacteroidetes, Proteobacter and Actinobacteria, although the proportions of each varies, as shown in Table 1.2. There are also some other phyla that have been detected such as Synergistes. Other papers that have conducted sequencing of numerous AD systems also confirm these findings. For example, De Vrieze et al. (2015) analysed 29 full scale AD systems and the most dominant phyla were Firmicutes, Bacteroidetes and Proteobacteria.

The relative abundance and variations of methanogens found in AD systems also differ, as shown in Table 1.3, and these microbes are influenced by various factors, such as digester type, size and operational conditions. For example, Regueiro et al. (2014) showed that temperature drop caused a change from *Methanosaeta* to *Methanosarcina* dominated communities. High volatile fatty acid (VFA) levels have also been shown to favour particular taxa (*Methanosarcina,* Franke-Whittle et al., 2014).

Previous research has also reported a large proportion of unassigned sequence e.g. Jang et al. (2014). This is not unexpected as the recent phenomenon in sequencing technology has allowed for more information to be obtained, but this is not without its own challenges. The extensive amount of data generated from sequencing often provides information on species where there is none in the literature, and so correctly assigning the sequencing information can often be a challenge, especially as many of the microbes involved in the process have not been cultivated (Vanwonterghem et al., 2014) or identified. Adding to this, independently isolating and culturing some of these organisms can be difficult as some microbes can only grow when co-cultured (Qiu et al., 2004).

There are two opposed theories regarding microbial community dynamics; neutral and niche. Both theories propose reasons for the observed formation of microbial communities. Neutral theory suggests that stochastic process determine the microbial dynamics, whereas niche suggests that deterministic factors influence this (Ofiteru et al., 2010). Few studies using AD systems that are run under the same conditions have been carried out to investigate which theory applies to this model. Vanwonterghem, et al. (2014) conducted such an experiment where three replicate lab scale AD systems were run and it was concluded that deterministic factors were the most important in shaping the microbial community, such as species interactions and operational conditions.

1.4 Methodologies for understanding microbial communities

1.4.1 Amplicon sequencing

The majority of research into the microbial communities and dynamics of AD using DNA sequencing reported to date use the 16S ribosomal RNA (16S rRNA) gene for bacteria, or mcrA for methanogens e.g. Razaviarani & Buchanan (2014), Jang et al. (2014), Regueiro et al. (2014), Sundberg et al. (2013) and Cardinali-Rezende et al. (2012). This targeted sequencing method is common for studies on microbial communities. The 16S rRNA is present in all prokaryotes and contains highly variable sequence regions, along with conserved ones. The conserved regions give this method the advantage of amplification using universal primers and the variable regions allow for phylogenetic analysis (Chan et al., 2011). Limitations associated with PCR amplification include that there can often be errors and bias associated with this method. These can include preferential annealing between both the primers and the template, varying copy numbers of the target, and the production of artefacts (Hongoh et al., 2003). These errors in the process can therefore sometimes give inaccurate estimations of the abundance of species in a microbial community.

1.4.2 Quantitative PCR

Quantitative PCR (qPCR) is a molecular method used to both amplify and detect changes in specific targets in DNA. Primers can be designed to target individual or groups of microbes by using specific target genes, and so this can be used to estimate the populations of the selected microbes which contain the targeted gene (Kim et al., 2013) i.e. monitor microbial dynamics. This method has been used previously e.g. Traversi et al. (2012), but remains susceptible to a variety of limitations, as discussed in Chapter 1.4.1.

1.4.3 Metagenomics

There are numerous NGS platforms available to researchers, using different methods of detection, each with their own advantages. The use of shotgun sequencing eliminates the limitations associated with targeted sequencing techniques. Selecting which platform to use is often based on a variety of required factors such as read length, speed, volume of data generated (largest throughput) or accuracy (Di Bella et al., 2013). Discussed are two platforms used in this project.

Ion-Torrent PGM is a sequencing technology where pH changes are detected in picowells on integrated circuits. Nucleotides are washed over the well and when a nucleotide is incorporated during stand synthesis, hydrogen ions are released, and this change in pH is detected (Whiteley, et al. 2012). This technology has the advantage that as light detection systems are not required, which are often expensive, it makes this technology more affordable. This in turn means that this technology is available for individual research groups, resulting in a greater reach within the academic research community towards complex microbial communities. Drawbacks associated with this sequencing platform include short read lengths, at around 200-400 bases, low output compared to other platforms (around 2 Gb) and a greater error frequency in homopolymer tracts.

The Illumina HiSeq is another sequencing platform that employs a sequencing by synthesis approach. The sample DNA is fragmented, adapters ligated, denatured, attached to a flow cell, followed by bridge PCR, resulting in clusters. Fluorescently tagged nucleotides are washed over the flow cell and if incorporation occurs, the cluster fluoresces at a particular wavelength, associated with a nucleotide. This is then cleaved to allow for the next nucleotide to be incorporated (Di Bella et al., 2013). Pair end reads are produced from the forward and reverse strands. This sequencing platform has the advantages of low reagent costs and the highest output of sequencing technology (up to 1000 Gb), but the main drawbacks are the long run times and short reads, around 2 x 150 bp.

There are other sequencing platforms available such as Roche 454, Sequencing by Oligo Ligation Detection (SOLiD), Illumina MiSeq, along with third generation sequencing technologies, such as MinIon.

The main limitation associated with the generation of large datasets using NGS is the method for analysing these. Generally, sequencing platforms produce short reads, and these have to be assembled into contigs. Contigs are overlapping sections of DNA, generated using contig assemblers, that attempt to join short fragments of DNA together, producing a longer section of DNA. The ultimate aim is to create complete draft genomes that would be highly informative about the microbes sequenced. Long read sequencing platforms such as PacBio are a useful technology that can improve genome assemblies as the short reads obtained from other sequencers can be scaffold onto these.

1.5 Types of Anaerobic Digesters

There are a variety of AD reactor designs that can be used to break down organic wastes to biogas. These systems can vary in size, number of digestion vessels, process temperature, and the vessel design is very much dependent on the characteristics of the waste materials. Vessel design and operation have a significant influence on microbial communities.

1.5.1 Reactor design

1.5.1.1 Single stage digesters

A single reactor vessel is used and the environment is maintained to ensure that the microbes in the system are in relatively favourable conditions. This type of digester has the advantages such as ease of operation and lower initial capital required to construct the system. The drawbacks of single stage digesters are that the diverse community of microbes have different optimum pH levels, such as the acidogenic bacteria which prefer a pH of around 4, whereas the methanogens optimal pH is 7 (Appels et al., 2008). Consequently as the methanogens are slow growing organisms, the pH is often tailored to their requirements. The sponsors of the project, described within this thesis, Clearfleau Ltd, use single stage AD systems (Figure 1.7).

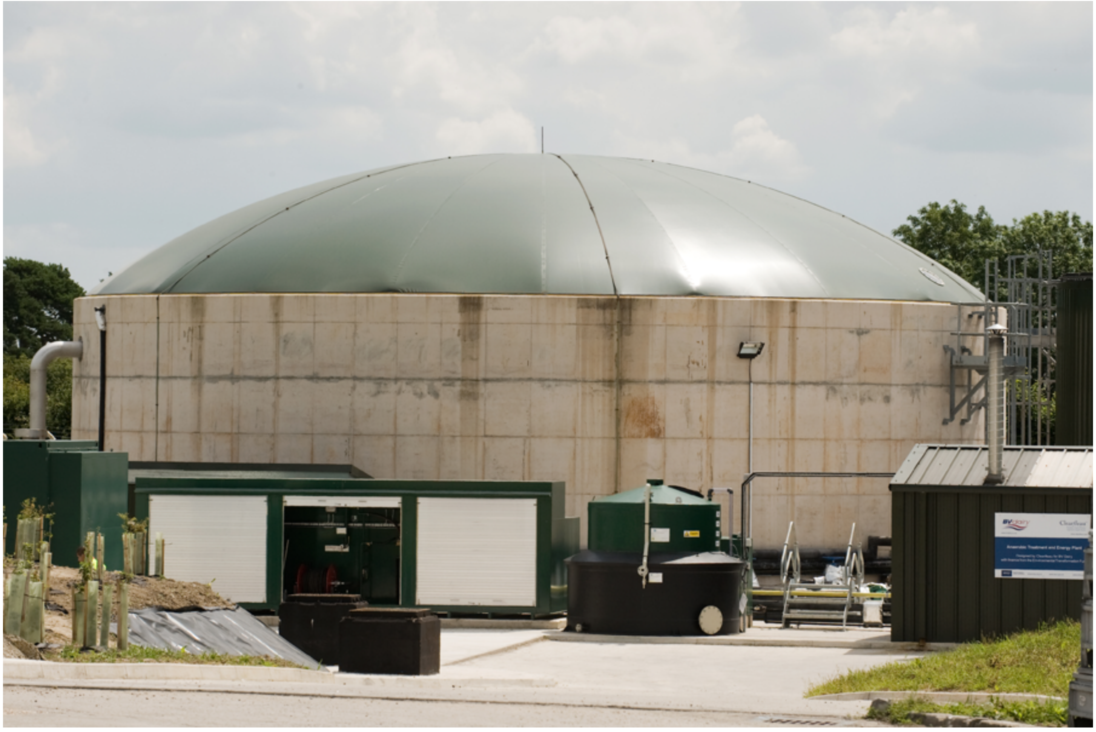


Figure 1.7. A process scale AD facility, located at a dairy facility. The digester has a 1000 m^3 capacity, taking in 200 m^3 of effluent daily.

1.5.1.2 Two (or more) stage digesters

Two or more vessels can be used to anaerobically break down organic waste. In this type of system, each vessel can be operated under different conditions as the microbes involved in the process specialise in different vessels (Lindner et al., 2016). For example, those microbes involved in hydrolysis and acidogenesis (discussed in Chapter 1.2) often accumulate in one vessel, at a pH of around 4, whereas those involved in the later stages of the process, e.g. methanogens can accumulate in the last digestion vessel (with a pH around 7), i.e. community partitioning. This community self organisation means microbes can be in favourable environments, which should ensure that all the biodegradable components of the waste get utilised, increasing the biogas output per unit of feedstock. Challenges associated with this type of system can be that the capital input and running costs are higher than single stage systems.

1.5.1.3 Fixed film AD

This is a process where inert materials (such as plastic) are used allowing for bacteria to attach and colonise, forming biofilms. The waste streams flow over the fixed bacterial colonies in the reactor, which break down the organic materials in the liquid. This process has the advantage that as the microbes are fixed onto a surface, they are not washed out, ensuring that the biomass in the system remains high, along with having a high surface area. Given this, the retention time of the system decreases, as there are high numbers of microbes to break down the organic waste. A drawback of this system is the low tolerance of suspended solids or particulate matter.

1.5.1.4 Plug Flow

Plug Flow AD systems are one of the most basic types of digesters. Waste materials are placed into one end of the plug, and this forces the material to flow through and is removed from the other end at the rate in which material is placed in. The advantage of this type of system is that it is a simple design and therefore lower capital and running costs than other digesters. These types of systems are more appropriate for feedstocks of high particulate or suspended solids content.

1.5.1.5 Upflow anaerobic sludge blanket (UASB)

In this process, a blanket of granular sludge is suspended in a tank and waste materials flow upwards through the blanket. As the material flows through the suspended blanket, it is broken down by the microbes. As the AD system becomes established, the microbes form granules where microbes attach and organise to form a cross-feeding network. Those microbes on the outside of the granule perform such processes as fermentation, and those in the middle perform methanogenesis (Li et al., 2015). This process is beneficial as it ensure the microbes remain in favourable conditions, e.g. ensuring low hydrogen levels.

1.5.2 Reactor considerations

1.5.2.1 Temperature

AD systems are generally run either as mesophilic (30-40 °C) or thermophilic (50-60 °C), depending on the feedstock characteristics. Mesophilic conditions have the advantage that less energy input is required to ensure the temperature of the system is maintained and there is often a greater diversity of microbes when run at this temperature (Yu et al., 2014), with microbial diversity being an important factor in AD. Thermophilic conditions have the advantage that the breakdown of the waste to biogas is quicker than mesophilic, allowing for higher organic loading rates (Moestedt et al., 2014) and there is greater pathogen kill compared to mesophilic. Thermophilic AD has been shown to reduce certain pathogens such as *Escherichia coli* and *Salmonella* species below detectable limits (Lloret et al., 2013). The drawbacks associated with thermophilic conditions are that greater amounts of energy are required to maintain the temperature, meaning that higher running costs are linked with this method. It is common for low solid waste materials to be placed into a mesophilic AD system, whereas high solid wastes tend to go into thermophilic systems.

1.5.2.2 Batch or continuous flow

Batch AD is when organic material is placed into an AD system and left until this has been broken down by the microbial community. The process is then stopped and restarted with new waste material, i.e. discontinuous. This is generally used for high solid based waste materials. Continuous AD is when waste materials are continually placed into the digester. This method has the advantage over the former that as the system is continually run, and does not require to be emptied and set up regularly, there is a consistent gas output.

1.5.2.3 Retention Time

The retention time is very much dependent on the type of digester system used, the size of the vessel and the characteristics of the feedstock. For example, feedstocks that are of low strength i.e. low levels of organic content (or low COD level), will be placed into an AD system at a high rate, therefore the material will remain the system for a shorter period of time, compared to feedstocks of high organic content. Additionally, waste materials that are of high solids content will have a higher retention time as these materials can take longer to breakdown.

1.5.2.4 Mixing

The level of mixing in AD systems is again influenced by the reactor design and the feed characteristics. Mixing is an important process in the running of AD systems to ensure that the material in the digester is homogenous, ensuring that there is an equal distribution of temperature and that the microbes can access the waste. There are various methods of mixing that can be used including hydraulic, mechanical and pneumatic. The mixing process can account for a large proportion of the AD running costs, but has been shown to be beneficial in gas output, releasing gas from the liquid phase to the gas phase (Lindmark et al., 2014).

1.6 Measured and controlled parameters in AD

There are several different industrial standard measurements that can be taken on a daily basis in AD which act as indicators that the systems are running efficiently. These can include gas flow, methane composition, pH, organic acids and chemical oxygen demand.

1.6.1 Biogas

Gas volume and biogas composition are key indicators of system performance. This is because the biogas is the end product of the digestion process. Therefore a high gas volume and high methane composition, relative to the vessel size and feedstock are often indicators that the biology of the system is working well. Methane content is dictated too by the composition of the waste material. High fat and protein levels produce a methane composition of around 60-70 %, whereas high sugar wastes usually yield around 50 % methane. This is because both fats and protein have higher specific methane yields than carbohydrates (Alves et al., 2009).

1.6.2 pH

pH is an important parameter that needs to be measured and controlled, especially in single stage systems. Methanogens in the system have an optimal pH of 6.8 – 7.2 (Appels et al., 2008), and so ensuring that the pH remains at least 7 is imperative to ensure that they are at optimal levels. If the pH drops below 7, an alkali solution such as sodium hydroxide is usually added. pH can sometimes act as an indicator of system imbalance. As organic waste is broken down to acetic acid, along with other volatile fatty acids (VFAs), there is only a small group of microbes that can utilise these, such as acetogens and methanogens. If there are insufficient microbes to use these products, then VFAs can start to accumulate, again causing the pH of the system to become more acidic. pH alone is not always a reliable measurement of system stability as the buffering capacity of the waste material input into the system, along with the AD system buffering capacity, could ensure that even though there potentially could be high VFA levels, the pH remains around 7 or above.

1.6.3 Temperature

AD systems can be run at different temperatures according to the waste that is placed into the system. As discussed in Chapter 1.5.2.1, AD systems are generally run either as mesophilic or thermophilic. It is important to ensure that the AD system is run at a constant temperature as the microbes in the system have acclimatised to that particular temperature, and changes in this could affect the trophic network of the microbial community. This in turn could, for example, cause an accumulation of organic acids, due to these not being utilised. It is however, not uncommon for fluctuations in the AD system temperature to occur, especially due to seasonal temperature variations. Regueiro et al. (2014) showed how a change in temperature altered the microbial community of a stable digester. In this experiment the digesters were run at a stable temperature (35 °C), before the temperature was dropped (17 °C) and then subsequently increased to 35 °C. During stable operating conditions the bacterial community was represented mainly by the phyla Firmicutes and Bacteroidetes, with *Syntrophomonas* and *Clostridium* the main genera. When the temperature was decreased, the Bacteroidetes phyla increased, but Firmicutes decreased, mainly *Clostridium* and *Syntrophomonas*. The methanogen community also showed changes, where *Methanosaeta* dominance converted to *Methanosarcina* during the temperature change.

1.6.4 Organic Acids

Organic acid, or VFA measurements in AD provides information on the process efficiency for the AD systems. Organic acids are the intermediate products in the process, and some are directly utilised by methanogens, so having an insight into the concentrations in the system is key. Increased levels of organic acids can be an indication of system imbalance, usually with the methanogens, causing a decrease in pH (Franke-Whittle et al., 2014). High levels of these organic acids can be toxic to methanogens (De Vrieze et al., 2012), so generally keeping the levels low is usually preferred. AD operators favour low levels of organic acids as it shows that the organic waste is being utilised efficiently by the microbial consortia. High organic acid levels are often perceived as unfavourable, but there can be exceptions to this. It has been noted that high organic acid levels actually enhance system performance, but this again can be dependent on the composition of the waste. Commercial kits, usually colourimetric, are available to measure the levels of organic acids, where fatty acids in an acidic environment react with diols to produce fatty acid esters, which are then reduced by iron salts to form a red colour. These kits provide a rapid measurement, but they do not offer detailed information on individual organic acids, usually just the acetic acid content. Measuring individual organic acids is an insightful parameter as it allows for detailed information on each VFA, and these can act as indicators of system performance. For example, Wang et al. (2012) reported that when the ratio of propionate to acetate was greater than 1 the methane output stopped, but at a value of <0.08 the methane output continued. Therefore, more detailed information on each VFA can be an indication on AD performance and efficiency.

1.6.5 Chemical Oxygen Demand

Chemical oxygen demand (COD) is the measure of a liquid that uses oxygen in the decomposition of organic matter. This test, usually colourimetric, is used to measure both the COD of the feed and the digestion vessel, where a sulphuric acid and potassium dichromate solution reacts with oxidisable material, resulting in a green colouration. The COD of the feed is required as this value can be used to determine the rate at which the organic waste is loaded into the digestion vessel (calculated as a feed to mass ratio, F:M), if a continuous feed system. A high COD feedstock would be added in at a lower rate, therefore increasing the retention time compared to a low COD of organic material. The COD of the reactor is a measurement that is used to determine if the organic waste placed in the system has been utilised. Typically a 95 % reduction in COD of the feedstock compared to the liquid removed from the system would be expected.

1.6.6 Feed Rate & Composition

The organic loading rate (OLR) is important to control to ensure that the AD systems are not overloaded with organic waste. Overloading the system often results in reactor acidification, and therefore a decrease in performance (Akuzawa et al., 2011). This in turn causes a build up of VFAs meaning that the balance between the VFA producers and consumers is disrupted. Where possible, controlling the composition of the feedstock being placed into the system is another important factor. The microbial community has the ability to adapt to a wide range of feedstocks, if given adequate time to do so. It is common practice for AD systems to be fed with mixed composition feedstocks (co-digestion) as this can provide additional nutrients (Park and Li, 2012), often resulting in increased biogas outputs. However, ensuring that feedstocks remain relatively constant and that there are no abrupt changes is essential. De Francisci et al. (2015) showed that when AD systems are overloaded with different substrates, e.g. carbohydrates, proteins or lipids, the microbial communities change and this can have a negative effect on the system performance. For example, when the feedstock was supplemented with carbohydrate, *Lactobacilli* numbers greatly increased, but such large changes in microbial populations were not observed with the additions of either proteins or lipids.

In some applications, waste streams can be processed before being placed into the system, referred to as pre-treatment methods. These can include thermal, chemical and mechanical pre-treatment methods, that are designed to enhance the digestion process, such as hydrolysis (Ariunbaatar et al., 2014). Some feedstocks can also be pasteurised prior to being placed into the digester. This practice would mean that the AD would be a closed system as there is no input of microbes into the digester, whereas open systems have a continual influx of microbes, which could change the microbial community composition.

Co-digestion is often used to ensure that there is a sufficient mixture of both macro- and micronutrients. The C/N ratio is measured and controlled as high levels of nitrogen can lead to an accumulation of ammonia, which has an inhibitory effect on methanogens (Chen et al., 2008), resulting in a decrease in methane output and increase in VFAs (Rajagopal et al., 2013). There are numerous proposals on how high ammonia levels have an effect on methanogens, such as that hydrophobic ammonia may diffuse into the cell causing a proton imbalance or that the ammonia may inhibit specific enzymes involved in methanogenesis process (Rajagopal et al., 2013). High ammonia levels have also been suggested to affect other microbes as well, such as aceticlastic microbes (Calli et al., 2005) In contrast, if the nitrogen level is too low, then there is not enough for microbial growth requirements, and so nitrogen additions need to be made to the system. Ammonia can also act as a buffering agent allowing for stable operating conditions, even if there are large fluctuations in VFAs, shown in Eq 1.9 (Zhang et al., 2013).

Eq. 1.9. $C_xH_yCOOH + NH_3 \cdot H_2O \longrightarrow C_xH_yCOO^- + NH_4^+ + H_2O$

1.6.7 Micronutrients

Micronutrients are elements needing to be controlled in the AD process as these have key functions. Micronutrients (or trace elements) are essential as these elements are found in enzymes that catalyse the fermentation and methane production reactions (Zandvoort et al., 2003), leading to methane formation. Some micronutrients that are added include Nickel, Iron, Cobalt, Selenium and Tungsten (Banks et al., 2012, Jiang et al., 2012). For these elements to be effective they must be in a bioavailable form and generally in low concentrations, e.g. 0.16 and 0.22 mg kg^{-1} of feedstock for Selenium and Cobalt respectively (Banks et al., 2012).

1.7 Aims

The aim of this project is to determine the microbiology of AD systems using metagenomics and to potentially correlate this with process data. To date there is a limited understanding on what the microbes and their dynamics are in anaerobic digesters. When operating these systems, a number of parameters are measured and controlled, and these are used to identify if the AD system is running efficiently. It is possible however that these data might not be providing a true reflection of system performance. Understanding of the microbes involved in the AD process could develop the potential to allow for better control, such as improving stability, efficiency, mainly gas output and breakdown of waste based on feed composition as well as microbial markers that are indicative of system performance. Additionally, an understanding of the microbial dynamics and interactions could be beneficial in improving AD systems, therefore the following questions need to be addressed:

- Is there a core group of microbes common to the AD process?
- How do the microbial communities adjust to particular feedstocks?
- Are there microbial markers that indicate system performance?

To investigate the microbial communities involved in AD, along with other process parameters, one objective was to design and develop a lab scale AD system that modelled a full scale digester, as those designed by the sponsors of the project, Clearfleau Ltd. This company uses single stage anaerobic digesters for the management of liquid waste materials. It was important that the digester closely mirrored the larger systems to ensure that it was a fair reflection of the full scale industrial process. An advantage of using a lab scale system, instead of collecting samples from an established AD system was that it ensured samples could be collected as required and processed immediately, along with allowing unique experiments to be performed.

Secondly, the monitoring of the microbial changes along with other parameters that are required to be measured in the AD process can include pH, individual VFAs, Organic Acids, Chemical oxygen demand (COD), gas flow and gas composition. All these parameters offer indications of system performance. Methods to measure process parameters needed to have local protocols put in place, such as those required for gas and VFA analysis.

Finally, samples needed to be sequenced using metagenomics to determine the microbial communities involved in the AD process. Methods to analyse the large datasets generated by DNA sequencing technologies were also required to be developed.

2 Materials & Methods

2.1 Digester set up and operation

2.1.1 AD operation

The lab scale, single stage anaerobic digester (Figure 2.1), with a 30 L working volume, was constantly mixed via a peristaltic pump (Rapide R8, Verderflex), with a flow rate of 300 L/h (10 x turnover/hour). The temperature was maintained at 35 °C (± 0.1 °C) via a heating plate (UC150, Stuart) and controlled via a custom designed temperature feedback system. The pH of the digester was logged and controlled (Hach Lange) using a 50:50 mix of 32 % NaOH solution and saturated Na_2CO_3 solution via a peristaltic pump to ensure the system remained at a constant pH 7. The feedstock was continuously added using a peristaltic pump (i150, ipumps). Water and small molecules, below 20 nm were removed from the system by using tangential flow filtration (3006805, Berghof).

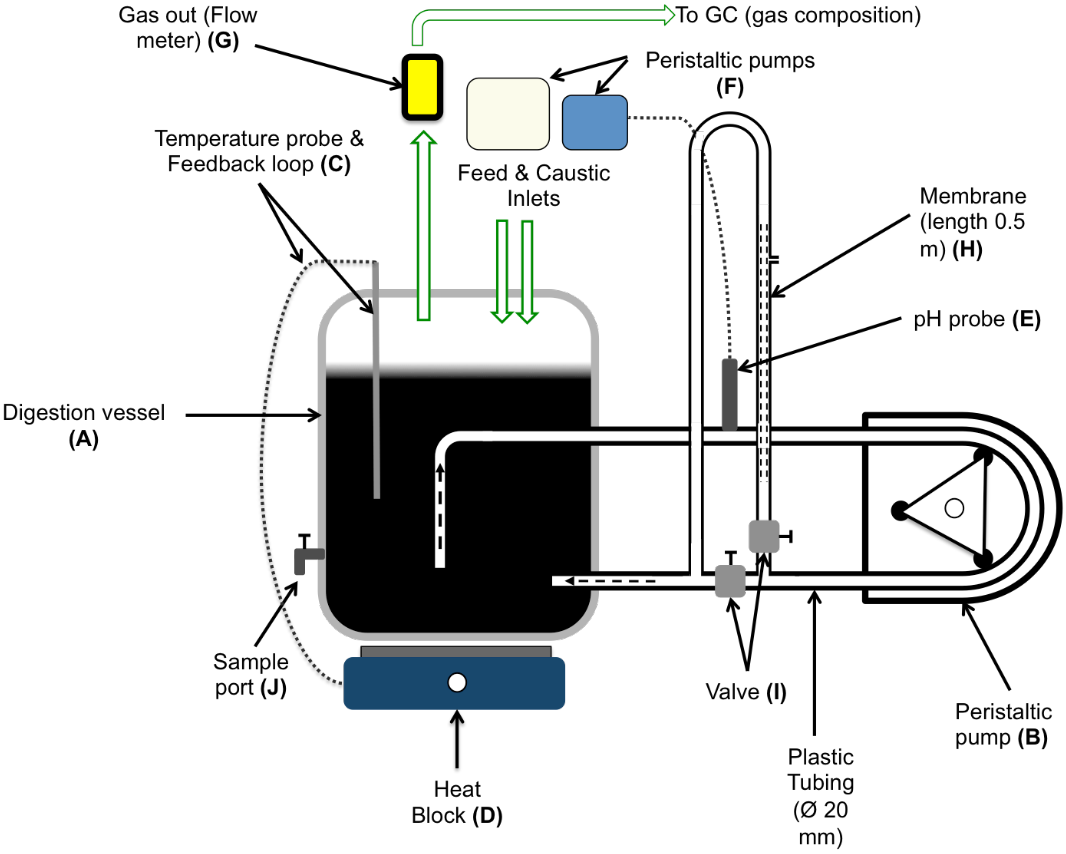


Figure 2.1. Cross section of the lab scale AD system final design. The 30 Litre stainless steel digestion vessel (A) is connected to a peristaltic pump (B) which mixes the material in the digester. A temperature probe (C) is connected to a heat block (D) to maintain a constant temperature and a pH probe (E) is connected to a pump that inputs caustic into the system (F). Feed is also introduced into the system via a peristaltic pump (F). The biogas is released via a gas out line (G) connected to a gas flow meter and GC. Water is removed from the system using a membrane (H) and the direction of flow in controlled via a valve (I). Digester samples can be collected from the sample port (J).

2.1.2 Inoculum

For each experiment, anaerobic digestate was collected from a local domestic wastewater treatment facility; Yorkshire Water Naburn, York. The sample was collected from an outlet pipe coming from a post-digestion holding tank, which contained the digestate (prior to a dewatering step). 30 litres of the AD material was collected in a sealed plastic barrel. The material was immediately transported to the lab where it was poured into the digestion vessel. The AD material was poured over a metal sieve (mesh size of 5 mm) to ensure that no large solid material was introduced into the digester.

2.1.3 Feedstock storage and input

2.1.3.1 Storage

Feed obtained from manufacturing sites was collected in sealed food grade barrels and stored at 4 °C.

2.1.3.2 Pasteurisation

The biodiesel waste was pasteurised prior to use. The 30 L barrel of waste was heated to 60 °C using an immersed heating element, held for 30 minutes, then chilled back to 4 °C.

2.1.3.3 Artificial feedstock

The artificial feedstock was made using a mixture of skimmed milk powder (400 ml/L), malt extract powder (300 ml/L), Coffee-mate ® powder (a mixture of glucose and vegetable fats, 200 ml/L) and yeast extract powder (100 ml/L). All components were made up at 100 g/L using tap water.

2.2 Process Data

2.2.1 Gas Flow and composition

For experiment one (Chapter 4.4.1) and two (Chapter 4.4.2), the flow of biogas produced by the AD system was measured using a gas flow meter (EW-32707-02, Cole-Parmer) and recorded via a data logger at 10-minute intervals (ACD-20, Picolog). Gas composition was measured once daily by GC-TCD (8610C, SRI) using a 3′ x 1/8″ 5A molecular sieve column (8600-PK2A, SRI). The GC was calibrated over a range of methane concentrations, varying from 30 % to 80 %. Methane standards were created using Nitrogen and Methane in Teflon bags. The initial oven temperature was set at 40 °C, held for 2 minutes after injection, increased to 200 °C at a rate of 20 °C/min. 100 µl of sample was injected using a syringe that has been flushed three times with sample prior to loading. Helium (Grade A, BOC) was the carrier gas at a flow rate of 10 ml/min. Peak analysis was carried out using PeakSimple software (SRI).

For experiment three (Chapter 4.4.3), the biogas flow was measured using a gas flow meter calibrated to a methane and carbon dioxide mix (WZ-32648-06, Cole-Parmer) and was logged at 10-minute intervals (ACD-20, Picolog). Gas composition was measured as described above, with the exception that the initial oven temperature was set at 60 °C, held for two minutes, and then increased to 220 °C. 100 µl of sample was loaded onto the column by an auto sampler (SRI). The GC was calibrated using a mix of Carbon Dioxide and Methane, at three different concentrations, which were run in triplicate (displayed in Figure 2.2 and Table 2.1).

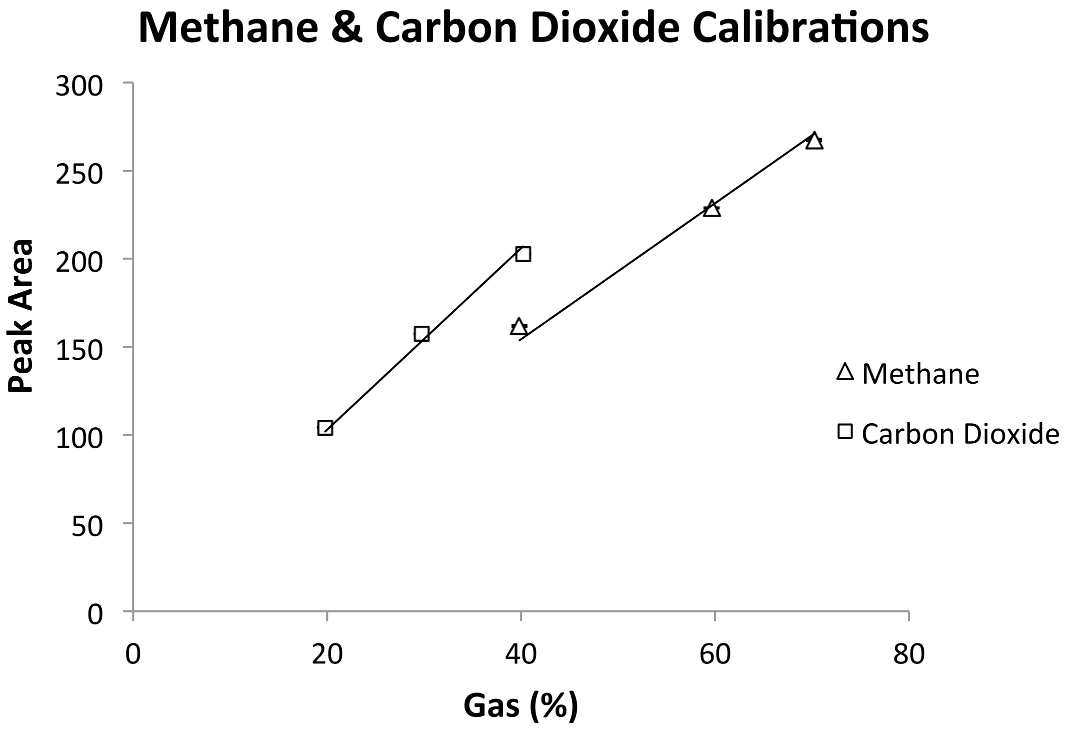


Figure 2.2. Calibration data for methane and carbon dioxide standards, run in triplicate.

Gas	Range (%)	R² Value
Methane	39.8 – 70.2	0.986
Carbon Dioxide	19-8 – 40.2	0.991

Table 2.1. Calibration data for methane and carbon dioxide showing the percentage range of each gas standard and R^2 value, when using the selected method.

2.2.2 Volatile Fatty Acids

Samples were obtained from the lab scale reactor using the sample tap. 30 ml of digestate was discarded before another 30 ml collected. Digester sample aliquots (2 x 2 ml) were centrifuged at 12,000 x g for 8 minutes, after which the supernatant was filtered through a 0.45 μm filter (Millex). The pellet fraction was used for DNA extraction (Chapter 2.3) while 1 ml of the supernatant was acidified using 15 µl phosphoric acid, then 2 µl was injected onto a GC-FID (Chrompack 9000).

A Nukol Capillary column, 30 m x 0.25 mm (size x I.D.), df 0.25 um (24107, Sigma) was used to separate VFAs. The initial oven temperature was 100 °C, increased to 200 °C at a rate of 8 °C/min, then held for 10 minutes. Helium (Grade A, BOC) was used as the carrier gas with a flow rate of 5 ml/min. The injector and detector were set at 230 °C. Peak analysis was carried out using PeakSimple software (SRI).

This method is similar to that used by Elbeshbishy & Nakhla (2012) and Cysneiros et al. (2012), and was capable of detecting Acetic acid, Propionic acid, Isobutyric acid, Butyric acid, Isovaleric acid, Valeric acid, Isocaproic acid, Caproic acid and Heptanoic acid. The system was calibrated using dilutions of Acetic acid (Figure 2.3) as well as dilutions of a Volatile Acid Mix (46975-U, Sigma) composed of Acetic acid, Propionic acid, Isobutyric acid, Butyric acid, Isovaleric acid, Valeric acid, Isocaproic acid, Caproic acid and Heptanoic acid (Figure 2.4 and Table 2.2).

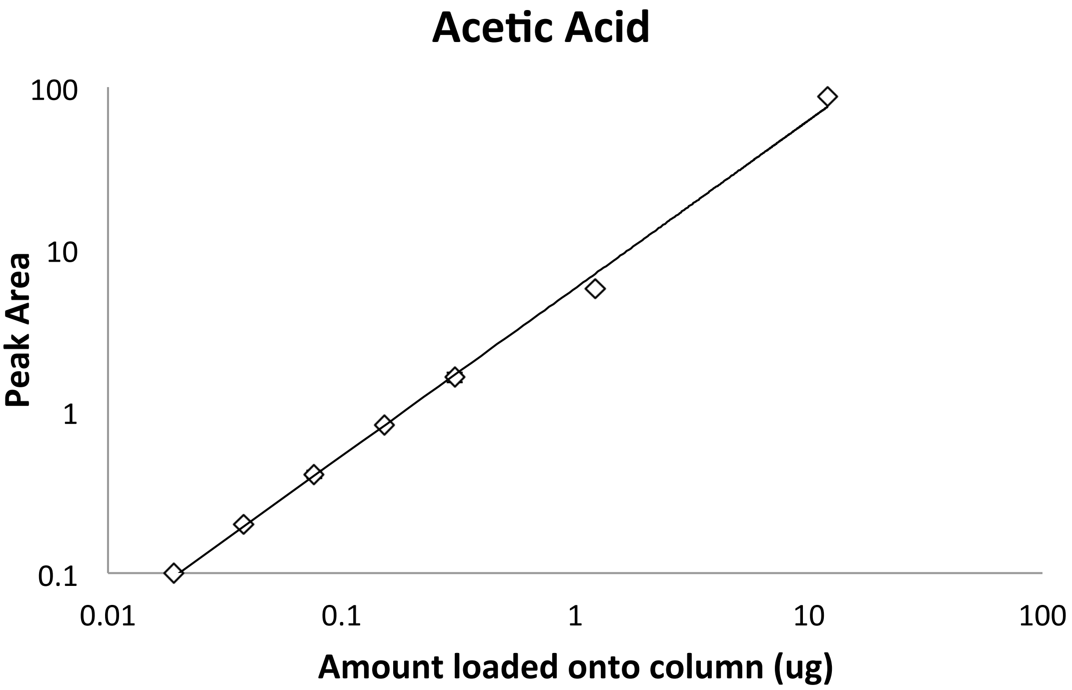


Figure 2.3. Calibration graph for acetic acid showing a linear response, on a log scale.

VFA	Range (ug)	R² Value
Acetic acid	0.02 - 12.01	0.998
Propionic acid	0.02 - 1.55	0.998
Isobutyric acid	0.03 - 1.78	0.998
Butyric acid	0.03 - 1.78	0.998
Isovaleric acid	0.03 - 2.11	0.999
Valeric acid	0.03 - 2.11	0.999
Isocaproic acid	0.04 - 2.33	0.998
Caproic acid	0.04 - 2.33	0.998
Heptanoic acid	0.04 - 2.66	0.997

Table 2.2. Individual VFA GC analysis showing the range of concentrations used to calibrate the instrument, and R^2 values, when run in triplicate.

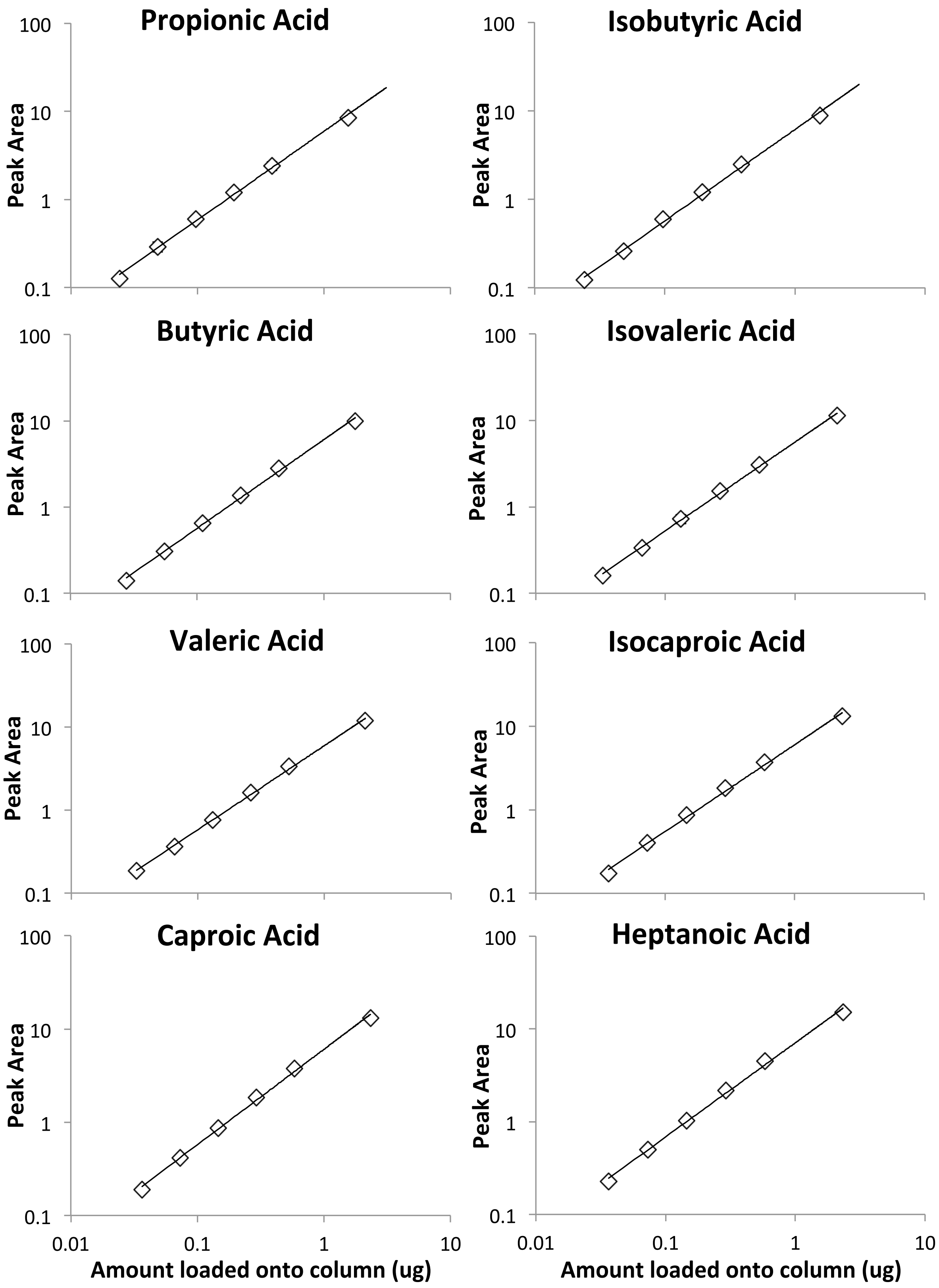


Figure 2.4. Calibration graphs for each Volatile Fatty Acid, on a log scale using dilutions of the standard VFA mix.

2.2.3 Organic Acids and Chemical Oxygen Demand

The test to determine concentration of organic acids was performed using the LCK 365, and the COD was measured using LCK 514 and LCK 914 kits (Hach Lange) according to the manufacturers instructions using filtered digestate samples.

2.3 Molecular methods for community analysis

2.3.1 DNA Extraction

2.3.1.1 QIAamp & SoilMaster DNA extractions

Genomic DNA was extracted using the QIAamp DNA Stool Minikit (51504, Qiagen) and SoilMaster DNA extraction kit (SM02050, Epicentre), according to the manufacturers instructions.

2.3.1.2 Mo-Bio Powersoil DNA extraction

Digester samples were taken from the sample port and were processed as described in Chapter 2.2.2. DNA was extracted using the PowerSoil DNA extraction kit (12888, MO-BIO), according to the manufacturers instructions with the following exceptions. During the bead beating stage samples were vortexed for 15 minutes instead of 10 minutes. Additionally, at step number 16 of the protocol, the centrifugation time was extended to 1 minute instead of 30 seconds. At step 18, the centrifugation was extended to 2 minutes instead of 1 minute. At step 21, the centrifugation time was extended to 1 minute. DNA was quantified by absorbance at 260 nm using a Nanodrop Spectrophotometer (Thermo Scientific). Samples were stored at -80 °C.

DNA quality was checked on a 50 ml 1 % agarose gel, using 0.5 g agarose, 50 ml 1x TAE and 2.5 ul EtBr (10 mg/ml). 10 µl of purified DNA was loaded into each well, along with 2 µl DNA loading dye (6x). The gel was run in 1x TAE at 100 Volts for 40 minutes.

2.3.2 Ion Torrent PGM Sequencing

Samples were sent to the Technology Facility, University of York to be sequenced. All reagents and equipment were obtained from Life Technologies unless stated otherwise. Libraries were prepared for metagenomic sequencing using the Ion Xpress Plus gDNA fragment library kit, with Ion Shear Plus reagent fragmentation for 300 base read libraries, according to the manufacturers instructions. Briefly, 1 µg of each DNA sample was fragmented for 6 minutes using the Ion Shear Plus protocol, followed by an enzymatic fragmentation and adapter ligation. Fragment sizes were determined using the Agilent Bioanalyzer with Agilent High Sensitivity DNA Kit. The average fragment size was 400 bp. Barcoded adapters were ligated and nick repair performed, and run on an E-gel SizeSelect 2% Agarose Gel (Invitrogen), with DNA fragments of 400 bp extracted. Five rounds of PCR amplification was performed and was accessed using the Agilent High Sensitivity DNA Kit with the Agilent 2100 Bioanalyser. Libraries were then pooled at eqimolar concentrations and diluted to 26 pM, ready for sequencing. Sequencing template preparation was performed using the Ion OneTouch system in conjunction with the Ion PGM Template OT2 400 Kit, and sequencing was performed on an Ion Personal Genome Machine System, using an Ion 318 Chip v2 with the Ion PGM Sequencing 400 Kit.

2.3.3 Quantitative real-time PCR

2.3.3.1 Primer Design

Primers were designed to detect eight different microbial species (Table 2.3), using Primer-Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/).

End point PCR was carried out with each primer pair to ensure a single band was formed for each selected target. The total volume for each PCR reaction was 50 µl, which consisted of 100 ng DNA, 0.1 mM forward primer, 0.1 mM reverse primer, 25 µl 2x Mastermix (Thermo Scientific) and made up to 50 µl using water. PCR amplification was carried out using a Tpersonal thermocycler (Biometra), where initial denaturation was at 95 °C for 2 minutes, followed by

94 °C for 30 seconds, 57 °C for 30 seconds and 72 °C for 30 seconds for 30 cycles and a final extension at 68 °C for 5 minutes. The PCR products were run on a 50 ml 2 % agarose gel, using 1 g agarose, 50 ml 1x TAE and 2.5 µg of EtBr (10 mg/ml). 10 µl of sample was loaded into each well with 2 µl of 6x DNA loading dye, and run in 1x TAE for 40 minutes at 100 Volts.

2.3.3.2 qPCR

For qPCR, 10 µl SYBR Green mastermix (Applied Biosystems), 0.5 mM forward primer, 0.5 mM reverse primer, 25 ng DNA and made up to 20 µl using water, assembled in the qPCR plate. Each DNA sample was run in triplicate on a StepOnePlus Real Time PCR system (Applied Biosystems). The holding stage was 95 °C for 20 seconds, followed by the cycle stage, 95 °C for 3 seconds, then 60 °C for 30 seconds, for 40 cycles.

Species	Forward Primer 5′ ➔ 3′	Reverse Primer 5′ ➔ 3′
Methanosarcina mazei	GCCCCTTCCCCTGACTTTAC	CCTGCCCTCAAAGTAACCGT
Methanoculleus marisnigri	AGGTGGAGGTGAGCATCATG	TGTGGTCCCGTATCTCCTCT
Cloacamonas acidaminovorans	CCGTGGTCTGATTGCCAATG	GCTCGTTCATAGAATCCCGTG
Syntrophus acidotrophicus	AAGATCCCGTCATTGCCGTT	GCCGACAGTGTGTTGCATTT
Bacteroides 3_1_19	TGTTCAGGCTTCTCTCGGTG	CGGATCGGCAGGGTTATGAA
Dyadobacter fermentans	TGTTCAGCCACATCCCATCC	AGAATGCCGAGAGTGTTGCA
Bacteroides vulgatus	AACATGAGGCCCGAAGTCAG	TCTTTCCCATCCATCACCGC
Pedobacter heparinus	CATGCCGGGAGATGTCAAGT	GCGAGAAAACCACTACCCCA

Table 2.3. Primer sequences used for each of the eight selected microbes.

2.3.4 Illumina HiSeq Metagenomic Sequencing

DNA samples were sent to the Leeds Institute of Molecular Medicine to be sequenced using two lanes on the Illumina HiSeq 2500 sequencing platform. The libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina (E7370, NEB), according to the manufacturers instructions. DNA samples were sheared to an average size of 200 bp using a Covaris S2 system before Library Prep Kit was used. After the PCR amplification step, fragment sizes were determined using the Agilent Bioanalyzer with Agilent High Sensitivity DNA Kit. The prepared libraries were run using a HiSeq2500 PE flow cell on the Illumina HiSeq 2500 platform.

2.3.5 PacBio Sequencing

DNA samples were sent to the Centre of Genomic Research, University of Liverpool to be sequenced using the PacBio sequencing platform. The genomic DNA samples were firstly purified using AMPure beads (Agencourt) and the quality and quantity was measured using both the Nanodrop and Qubit assay. The Fragment Analyser (AAT), using the high sensitivity genomic DNA kit was used to determine the average size of the DNA. DNA was treated using the SMRTbell library kit (PacBio). Briefly, the DNA was treated with Exonuclease V11 at 37 °C for 15 minutes and the ends of the DNA were repaired according to the Pacific Biosciences protocol. Sample was treated using DNA damage repair mix for 20 minutes at 37 °C, followed by a 5 minutes incubation using end repair mix at 25 °C. DNA was ligated to the adapter overnight at 25 °C. Ligation was terminated by incubation at 65 °C for 10 minutes followed by exonuclease treatment for 60 minutes at 37 °C. The library was then purified using 0.5 x AMPure beads and size selected using a 0.75 % blue pippin cassette (SAGE), in the range of 15000-20000 bp. The recovered fragments were damage repaired again. A Qubit assay determined the recovery of DNA and the Fragment analyser determined the average fragment size. SMRTbell library was annealed to the sequencing primer (determined by the Binding calculator) and a complex made using DNA Polymerase (P6/C4 chemistry). The complex was bound to Magbeads and sequencing was done using 360 minute movie time on the Pacific Biosciences RS11 instrument.

2.4 Bioinformatics

2.4.1 MG-RAST

Sequencing data obtained from the Ion Torrent PGM sequencing platform was uploaded into MG-RAST (Meyer et al., 2008). The data was analysed using default settings. Briefly, the annotation database used was M5NR, the maximum e-Value cut off was 1e-5 and the minimum identity cut off was 60 %.

2.4.2 Contig assembly

Contigs were assembled from the sequencing data using Megahit (Liu et al., 2015), using the default parameters. For the alternative assembly strategy, Megahit, IDBA-UD (Peng et al., 2012), Newbler (Roche) and Minimus2 (Sommer et al., 2007), were all used. Briefly, the contigs were formed by dividing the Illumina sequencing data into three groups (1-4, 5-8, 9-11) and then run through IDBA-UD, in parallel with all the 11 samples through Megahit. The contigs that were over 2 Kb were merged with the PacBio reads using Minimus 2. Any contigs smaller than 2 Kb were reassembled using Newbler before being reassembled with the other contigs.

2.4.3 Clustering

2.4.3.1 K-means clustering

Contigs were clustered using the Sci-Kit Learn module for Python, using an input cluster number of 256.

2.4.3.2 Custom clustering method

A custom script was written for clustering the contigs. Briefly, the standard deviation of the difference of the log value of abundance for all time points between pairs of contigs was calculated, based on the normalised data. If the standard deviation was greater than 0.035 then the query and test are judged to be colinear. See Appendix A for custom script.

2.4.4 Prokka

Open Reading Frames (ORFs) from the clusters were assigned using to Prokka (Seemann, 2014), using the default settings.

2.4.5 Metabolic activity

The ORFs from each cluster were uploaded to the KASS KEGG (Moriya et al., 2007) server (http://www.genome.jp/tools/kaas/) to determine the genes within these clusters, denoted to be involved in methane metabolism.

2.4.6 BLAST

2.4.6.1 Manual BLAST

Manual BLAST searching of ORFs, as assigned by Prokka, was conducted using the NCBI standard protein BLAST (Altschul et al., 1997).

2.4.6.2 Automated BLAST

The automated BLAST search was carried out using a custom script whereby the top match for each ORF was returned, displaying the function of the potential gene, the organism it belonged to and the e value. Any search that was returned as a hypothetical protein was excluded. See Appendix B for custom script.

2.4.7 Phylogeny

A core set of genes associated with 30 methanogens (see Appendix C) was selected using MicroScope (Vallenet et al., 2013) (http://www.genoscope.cns.fr/agc/microscope/home/) using an MICFAM parameter of 50/80. A BLAST database of these sequences was generated and a list of common core genes for each cluster to be placed in a phylogeny was retrieved and edited to form a FASTA file of concatenated genes for both reference and query species. This file was loaded into Clustal X (Larkin et al., 2007), to generate an alignment and tree file that was visualised using FigTree (v1.4.2).

2.4.8 Genome mapping

The specified reference genomes were downloaded from the NCBI database and loaded into Double Act (http://www.hpa-bioinfotools.org.uk/pise/double_act.html) along with the cluster sequence data. The output file from Double Act, along with the reference genome and cluster sequence data were loaded into Artemis Comparison Tool (ACT) (Carver et al., 2005), to visualise the comparison.

3 Lab scale digester construction

One aim of the project was to design and construct a lab scale anaerobic digester that would provide a means to monitor the dynamic changes of the microbial communities involved in AD, along with measuring other factors, such as process data.

To do this, a lab scale anaerobic digester was designed and constructed, with the aim of ensuring that it modelled the full scale system, built by the sponsors of the project, Clearfleau Ltd. This company uses single stage anaerobic digesters for the management of low solids liquid waste materials e.g. confectionary, biodiesel and dairy wastewaters. These systems are designed to have a solids content ranging between 2 – 3 %, and can vary in size, depending on the output volume of wastewaters and composition of the feedstock. The size of Clearfleau process scale digesters ranges between 1500 – 5800 m^3, with feed inputs ranging from 42 – 275 m^3 per day (data taken from recent Clearfleau built digesters). Clearfleau Ltd also have a pilot scale AD system that is used on site to trial the suitability of feedstocks, which is a scaled down version of the process scale digesters, with a volume of 8 m^3. Importantly, as the majority of feedstocks are from food processing sites, the microbial load in the feedstock is minimal. This means that essentially no microbes are added into the system, which is beneficial when studying the microbial compositions of AD, as this ensures there are no fluctuations or changes occurring because of external microbes added. This in comparison to other AD systems that have an influx of microbes from the input material (e.g. domestic wastewater), and therefore the community structure would be highly influenced by the continual addition of microbes. It was important that the lab digester closely mimicked the larger systems to ensure that it was a fair reflection of the full scale industrial process.

A critical part of this process is the solids and microbe retention within the system. This is achieved on the pilot scale unit using a Cavitation Air Flotation Tank (CAFT) system (as described in Chapter 3.2), where reactor sludge is vigorously mixed with air and polymer, causing the solids and microbes to aggregate and float. This material can then be returned into the digestion vessel to ensure that the microbes remain in the AD system and the solids are returned as they require a longer retention time to be biologically degraded.

An advantage of using a lab scale model, instead of collecting samples from an established AD system, is that samples collected can be processed immediately, which is important if transcriptomics or metabolomics studies are to be conducted, especially as RNA and metabolites are unstable. Additionally, the lab scale system allows for unique experiments to be performed, where process parameters can be varied, such as increasing feed rates beyond normal practice. These sorts of experiments are not feasible at pilot or process scales due to costs, whereas the lab scale systems can be easily restarted if required.

3.1 Lab Scale AD system

The design and construction of the lab scale system, in collaboration with the Biology Mechanical and Electronic Workshops, University of York was an extensive process that involved much experimentation to ensure the lab scale system mimicked those of process scales in terms of layout and how they were run. The suggested digester layout information, based on the process scale system, was provided by Clearfleau Ltd.

The digester unit, which is effectively a membrane bioreactor (Xiao et al., 2015), was a modified stainless steel stockpot, with a 36 litre capacity, giving a liquid working volume of 30 litres. To this, a pH probe, with feedback connected to a pump that added an alkali solution, to ensure the pH remained around 7. A temperature probe, connected to a heat block was used to ensure the temperature remained at the set temperature. A gas out pipe and inlet ports for feed and caustic solution were all installed. A peristaltic pump was selected as the method for mixing the system. Figure 2.1 shows a diagrammatic view of the digester layout including the dewatering system used in the final design, tangential flow filtration, discussed in Chapter 3.2.2, and Figure 3.1 is a photo of the system.

The main challenge associated with the designing of the lab system was the issue of scaling down to a 30 litre unit. Such challenges include that at process scale a chopper pump is used to circulate and mix the reactor content, whereas at lab scale a peristaltic pump had to be used, as this gave the required flow rates. The two pumps work in a different manner, potentially having a different effect on the sludge material. Additionally a heat exchanger is used on larger systems, but for the lab unit, a heat block had to be used, as this was an appropriate size for a scaled down system. Level sensors are also used in the larger systems as an estimation of the volume within the reactor, but this did not work at lab scale, and finally a CAFT system was used as a method of dewatering, but again, this did not work at the lab scale (discussed in Chapter 3.2.1). Tangential Flow Filtration (TFF) was used as an alternative for dewatering. The continuous addition of feed is another key feature of the AD systems, which is a challenge when scaling down, especially as a low flow pump and small tubing is required. These factors demonstrate some limitations of scaling down and therefore on the lab scale AD unit.

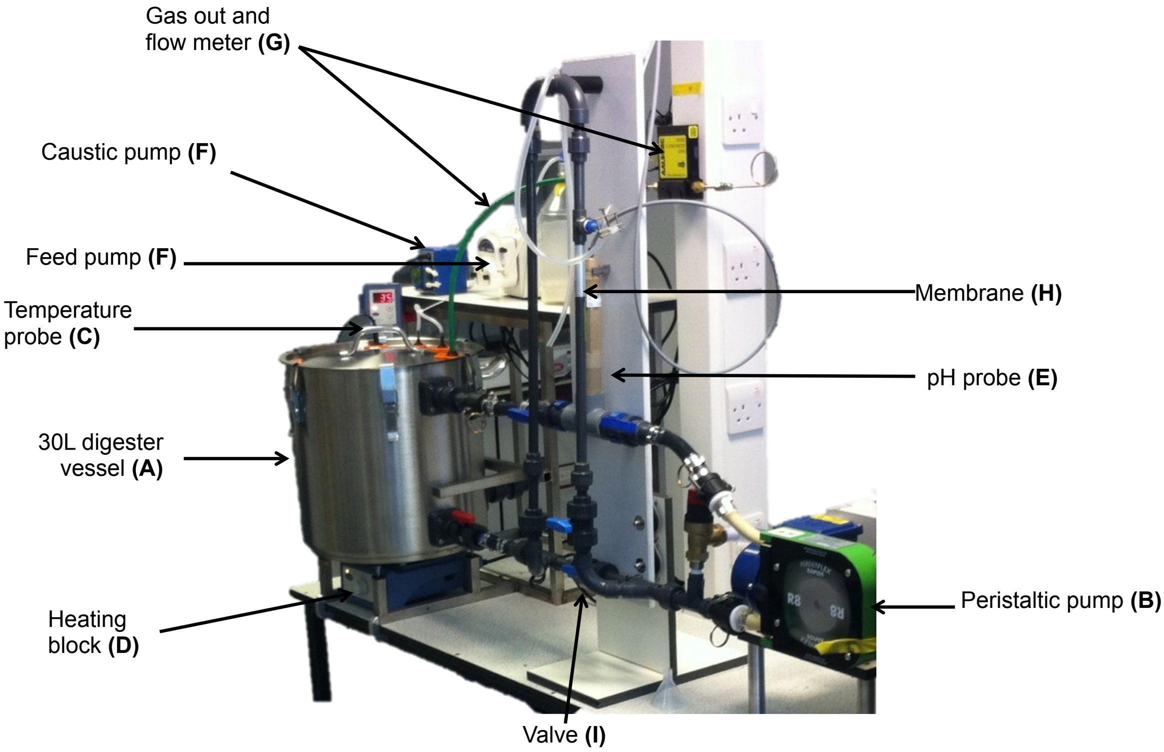


Figure 3.1. The lab scale anaerobic digester. The 30 Litre stainless steel digestion vessel (A) is connected to a peristaltic pump (B) which mixes the material in the digester. A temperature probe (C) connected to a heat block (D) is used to maintain a constant temperature and a pH probe (E) is connected to a pump which inputs caustic into the system (F). Feed is also introduced into the system via a peristaltic pump (F). The biogas is released via a gas out line (G) connected to a gas flow meter. Water is removed from the system using a membrane (H).

3.2 Dewatering methods

The designing and construction of the lab scale dewatering system required several different methods to be trialled. This is because Clearfleau's digesters use a CAFT system, but at the lab scale, this did not operate as expected (discussed in Chapter 3.2.1), and so an alternative method, tangential flow filtration was trialled (Chapter 3.2.2). The removal of water from the system is to ensure the volume of the digester remains constant, and is especially important when dealing with low solids/high volume liquid wastes. Retaining the microbes in the system is important, especially for AD systems that have a liquid material input, which have a low solids content and a short HRT. This is because methanogens are slow growing organisms, and if digester material was just removed directly from the system, the methanogen abundance would decrease due to microbial washout and this could reduce the biogas output. Discussed in this section are the two different dewatering methods that were investigated.

3.2.1 CAFT

The initial plan was to replicate a method that Clearfleau Ltd use on both pilot and process scale systems. This is known as a CAFT system. Figure 3.2 displays a representation of this system.

The CAFT system is a method for dewatering which utilises air and polymer to ensure that microbes and solids can be returned back into the AD system. Methanogen activity is reduced when exposed to oxygen (Fetzer et al., 1993) and so it could be assumed that using this method could potentially be detrimental to the methanogens and therefore methane production. It is possible that when flocculation occurs, the majority of organisms are protected from air exposure, except those on the surface, explaining why this technology can be used. It is also possible that granules could form (McHugh et al., 2003), a process which polymer could encourage, which again offers an explanation to why this technology can be used. Along with oxygen exposure, another drawback of using this method for dewatering is the use of a polymer. Polymers added are usually either cationic or anionic. These cause the solids and microbes to flocculate. Whilst the polymer is effective at this for dewatering, the returned aggregated solids and microbes can potentially fall out of suspension in the system, and when more polymer is used, can exacerbate the issue. This in turn means that the microbes in the AD unit could either be at the bottom of the reactor or floating on the surface, reducing both the capacity of the reactor and its performance. Another drawback of this technology can be the associated polymer usage costs, increasing the overall running costs of the AD unit.

The use of a CAFT system has the advantage that when the solids and microbes flocculate, they can be scraped off the surface and returned back into the system. This ensures that there is no microbe washout, keeping the microbial numbers high within the system.

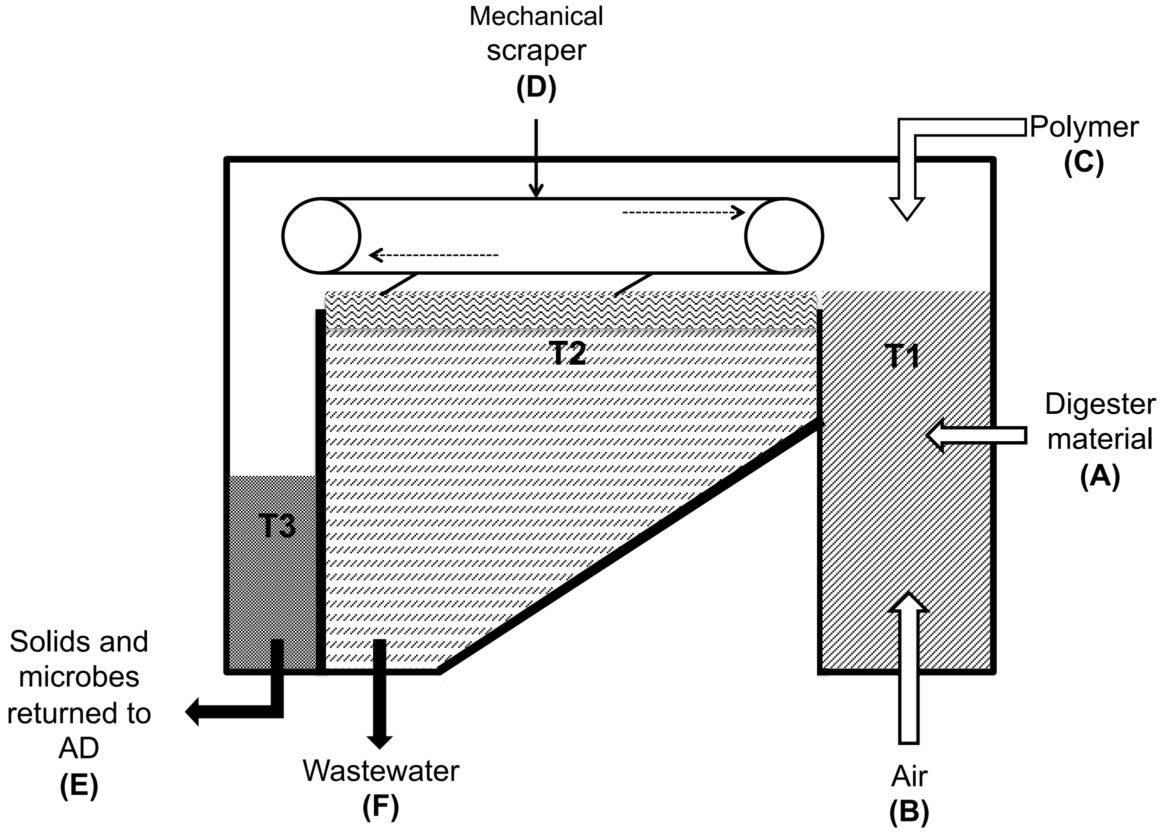


Figure 3.2. Cross-sectional representation of the CAFT system used to retain solids and microbes whilst removing water. Digester material from the reactor (A) is pumped into the first tank (T1), where it is mixed with both air (B) and polymer (C). This material then flows over into the second tank (T2) where the aggregated and flocculated material (solids and microbes) can be mechanically removed from the surface of the liquid by a scraper (D) into the third tank (T3). The solids and microbes can then be pumped back into the AD reactor (E), whilst the remaining wastewater (F) can be discharged down the drain.

A lab scale CAFT system was constructed and trialled, as the aim of the project was to mimic that of a full scale system. The lab scale CAFT had a working volume of 7 litres. Although several modifications were made to the CAFT to ensure it worked as intended, the process was not as efficient at flocculation as the full scale systems, as many of the solids remained in suspension. It is likely that the aeration on the lab system was not vigorous enough to cause flocculation. Examples of modifications made included altering the heights of the weir, along with using a more powerful motor to draw the air into the system for flocculation, but these changes did not prove to be sufficient. Because of this, an alternative method had to be sought.

3.2.2 Membrane Filtration

Tangential flow filtration was investigated as an alternative to the CAFT system. The principle of tangential flow filtration is that the material of interest is passed through a membrane tube under pressure. Molecules smaller than the selected membrane pore size can cross the membrane, such as water, but those greater cannot, for example, solids and microbes (Figure 3.3). When used in AD, this ensures that the solids and microbes remain in the system. The rate at which liquid passes across the membrane, as filtrate, depends upon the rate of flow through the membrane, which determines the pressure. Increased flow is beneficial in TFF as this ensures there is scouring of the membrane, preventing it from becoming blocked. An advantage of using this method is that the pore size of the membrane can be selected, giving more control of what material is emitted in the filtrate. This ensures particulates such as biomass (Xiao et al., 2015) and suspended solids are retained in the system, and this is beneficial for slow growing anaerobes (Smith et al., 2012). Another clear advantage of this method is that it is a closed system, unlike using the CAFT method, and so no oxygen or polymer is introduced.

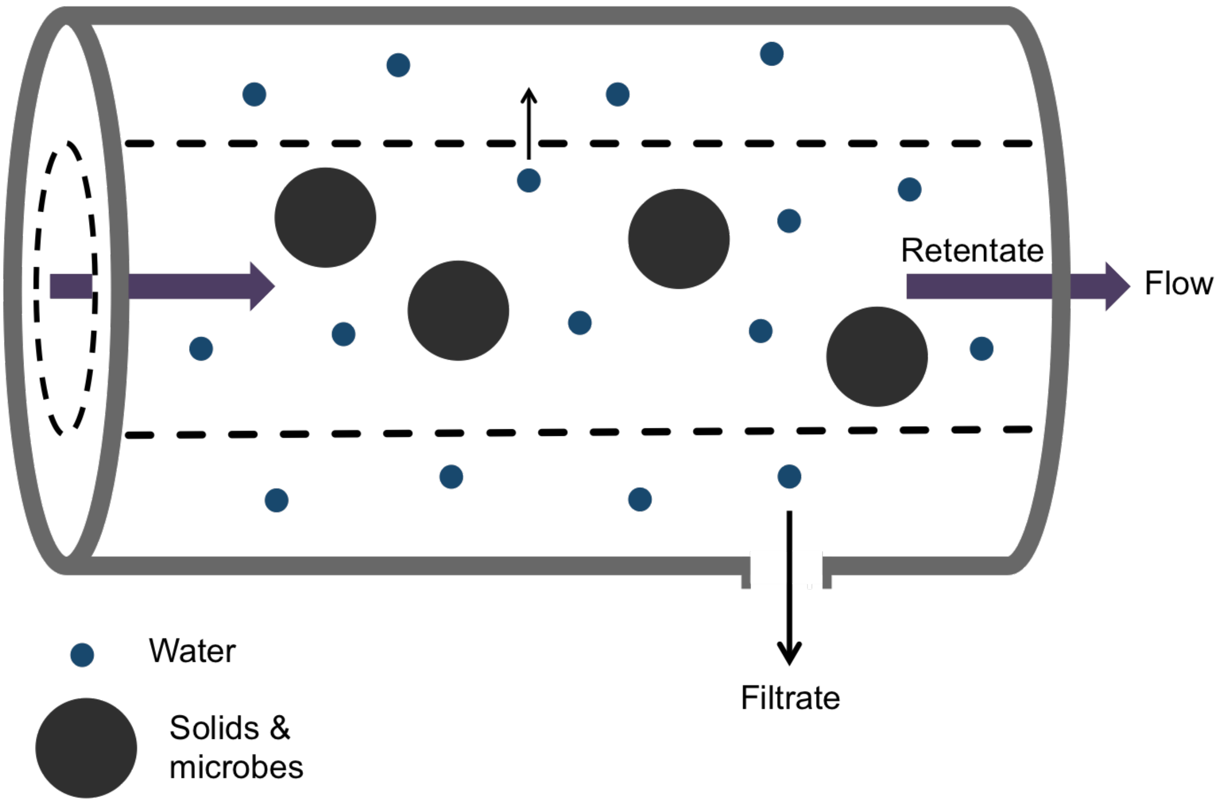


Figure 3.3. Cross sectional view demonstrating the principle of tangential flow filtration. Digester material is passed along the membrane tube where material smaller than the pore size is removed as filtrate, but anything greater is retained in the system.

Three different membranes were trialled to determine if any were suitable for water removal in AD (Table 3.1). Subsequently two were found to have a membrane inner diameter that was too small, and the rate of water removal was not sufficient for the lab scale system.

Membrane	Material	Length (cm)	I.D (mm)	Pore size
WaterSep Explorer 12	M-PES	31.2	1	750 kD
MidiKross TC Filter Module	M-PES	23	1	500 kD
Berghof	PVDF	50	8	30 nm

Table 3.1. A comparison of the membrane properties including the material, module length, inner diameter of the membrane and pore size and those trialled to determine which was most suitable for water removal in anaerobic digestion.

A third membrane was trialled, which was suitable for the lab system. The membrane (Berghof), had an I.D of 8 mm, 0.5 m long and a pore size of 30 nm, and performance tests demonstrated sufficient water removal was achieved using this membrane.

3.3 Conclusions

The lab scale AD system design enabled it to be used for numerous experiments. To ensure that the system was working prior to conducting the various experiments, a short commissioning experiment was carried out to ensure that all the components were working as intended (data not shown). In this run, the AD system was inoculated with sludge taken from a local wastewater AD system and fed using biodiesel waste. The various process parameters were monitored to ensure that the system was working as anticipated. This included checking that the data were being logged correctly, the temperature and pH remained within the set limits, the peristaltic pump mixed the digester material, the gas composition was monitored to ensure the lab scale system remained anaerobic and finally that the dewatering method was suitable.

The greatest challenge was developing a scaled down version of an AD system. There are some clear limitations to be acknowledged. These include that on the full scale systems, mixing is achieved by a chopper pump, heating maintained by heat exchange, dewatering using a CAFT system, all the tubing is oxygen impermeable and micronutrients are added. In comparison, for our lab system, mixing is maintained by a peristaltic pump. The use of this pump can be a drawback because as the rollers go over the tubing it gets compressed, which will also compress the digestate. The set temperature was maintained by a heat block, and so potentially the heat distribution could be slightly uneven within the system. Dewatering was achieved using membrane filtration, which ensures the system remains closed and no polymer is used. This is a distinct difference between the two systems, but both methods for dewatering retain the microbes, ensuring the lab system mimics the process scale. Additionally, silicon based tubing was used for the smaller peristaltic feed pump which is oxygen permeable. As the experiments described in this work were not for long periods (maximum 57 days), micronutrients were not added.

Regardless of the limitations, the lab unit has been proven to be a robust system that is comparable to larger systems, and therefore useful for conducting predictive experiments.

4 Process Data from Lab Scale AD Trials

4.1 Introduction

The use of laboratory scale AD systems are important to address a wide range of issues in this field. These systems therefore are an essential tool in understanding the processes behind AD, and ultimately ensuring this information is translated to the process scale systems, with the aim of improving the AD process.

An advantage of having lab scale systems is the opportunity for novel experiments to be conducted, whilst collecting process data and samples for metagenomics. Such experiments include, as described in this chapter, (a) feeding a lab scale system beyond the normal industrial practice rate, (b) starting the feed rates higher than expected and increasing this at a much faster rate and (c) running three systems in parallel. The availability of lab scale systems ensured that AD samples could be collected as required, often on a daily basis and processed immediately for DNA extraction, ensuring that the DNA is a reflection of the AD system at that immediate time point. The variety of process data available to be collected is also an advantage as this offers an insight into system performance. An important parameter to measure is the individual and total organic acid concentrations. These compounds are intermediate metabolites in the system, and acetic acid is especially important as this is utilised by methanogens, and so levels of these can be indicators of system performance. The concentration of other VFAs are also important to measure as these have been shown to be inhibitory to the process, such as propionic acid, at concentrations over 900 mg/L (Wang et al., 2009). The measured COD value acts as an indicator of how well the components of the feedstock are being utilised, along with the gas flow rate and gas composition, which can all be measured on our lab systems. The pH is another important parameter, especially as the maximum biogas yield has a pH around 7 (Liu et al., 2008). Furthermore, the pH has been shown to affect which methanogenesis pathway is dominant, with the aceticlastic pathway being most dominant when the pH is around 6.5, but in a more acidic environment, the hydrogenotrophic pathway is more dominant (Hao et al., 2012). Therefore the measurement of these parameters is beneficial in AD to determine the system performance. Additionally, by collecting all these data, it has the potential to correlate this with the metagenomic analyses, and therefore understand how the microbial population responds to the system environment or vice versa. Process data indicates what has already happened in the system, as a response, therefore is likely to have a poor predictive ability. It is hypothesised that microbial indicators will predict system stability better than process data.

4.2 Aims

- To run the lab scale AD systems conducting different experiments - under varying conditions, such as using different feedstock compositions and varying the feed rates
- Measure process parameters including gas flow, composition, pH and volatile fatty acids, as key measureable components of system performance
- Take DNA samples for next generation sequencing to investigate the microbial communities

4.3 Experimental Design

For each experiment the lab scale AD systems were seeded using 30 litres of inoculum taken from an established domestic wastewater treatment site, with an anaerobic digester, located in Naburn (Yorkshire Water), York. A different feedstock was added into the systems continuously for each experiment as detailed below and process measurements including gas flow, composition, pH, organic acids were all collected, along with DNA samples for sequencing.

4.4 Results

4.4.1 Biodiesel Waste

The first experiment was carried out to show that the lab AD unit was a robust system to use for the monitoring of various process parameters, and to provide samples for DNA sequencing, allowing for the development of molecular tools to investigate the microbial communities.

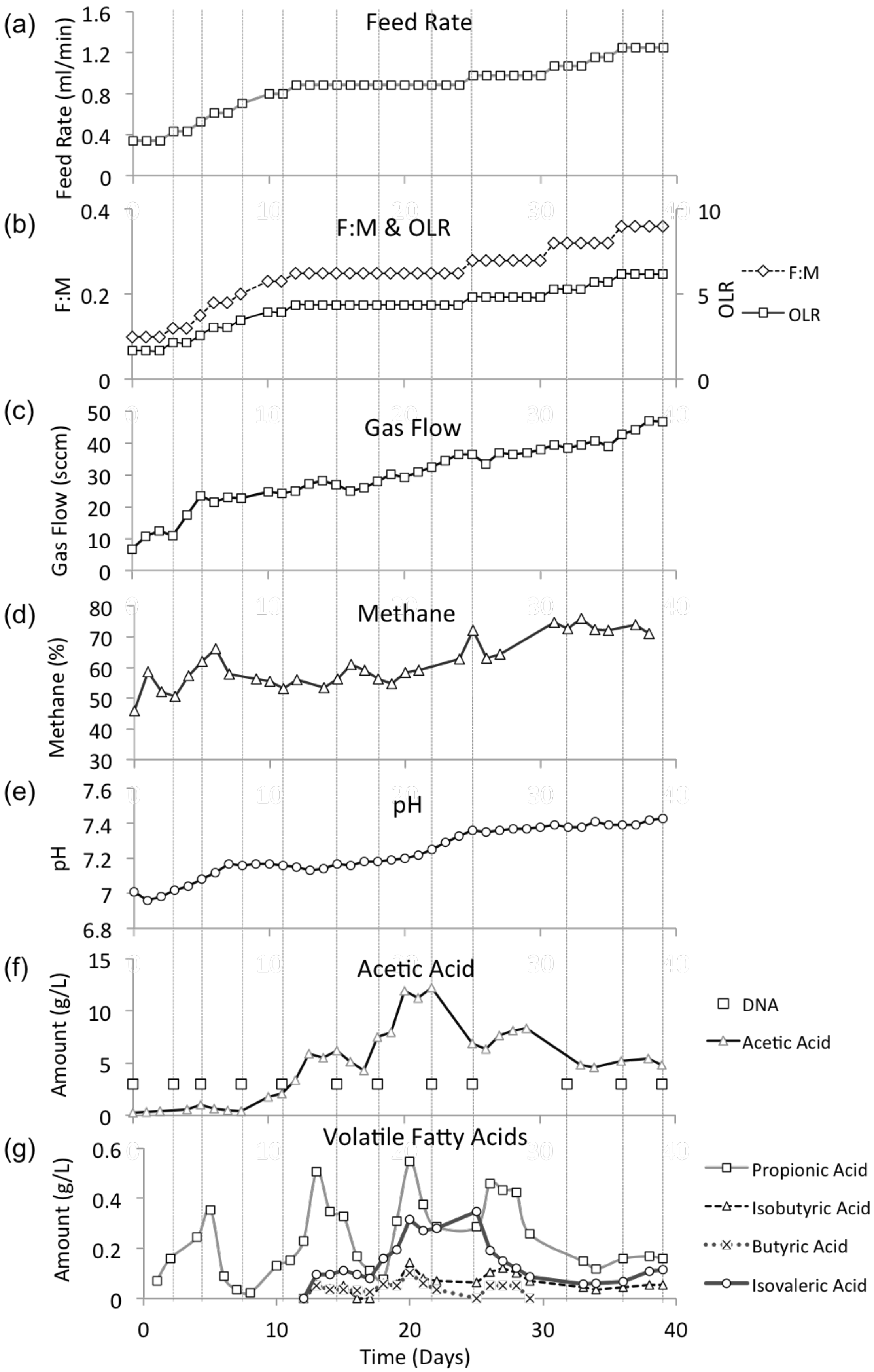
Process waste, collected from a biodiesel production facility was used as the feedstock. It was expected that this waste process water contained (amongst others) some methanol and glycerol (Siles López et al., 2009). The process waste was pasteurised before use (Chapter 2.1.3.2), to ensure the microbial load was low, and therefore not introducing microbes via the feed into the AD system.

The experiment was run for 39 days (Chapter 2.1.1). The solids content of the inoculum collected from the established AD system was around 2.2 % (according to the Yorkshire Water on-site monitoring equipment) and the COD of the feed when collected from the site (before being pasteurised and refrigerated) was 92.5 g/L. It is therefore possible to estimate the F:M ratio and OLR. The starting feed rate for the experiment was 0.34 ml/min (F:M 0.1). This was steadily increased throughout the trial, reaching a maximum of 1.25 ml/min (Figure 4.1 a), giving an estimated F:M of 0.36 (Figure 4.1 b). The estimated starting OLR was 1.68 and reached a maximum of 6.17 Kg COD.m^3.d.

The system total volatile fatty acid concentration when sampled (see Chapter 2.2.2) responded to the introduction of feed and subsequently increased, reaching a peak of 12.92 g/L after 22 days, although this decreased thereafter. The main VFA that accumulated in the system was acetic acid (Figure 4.1 f), and this accounted for 95% of total VFAs (12.25 g/L) when the VFAs reached a peak of 12.92 g/L. Other VFAs were detectable in the system, such as propionic acid, isobutryic acid, butyric acid and isovaleric acid (Figure 4.1 g), but at much lower concentrations than acetic acid, e.g. a peak of 0.54 g/L was measured for propionic acid.

Initial pH of the digester was 7.01 and reached 7.43 at the end of the 39 day trial (Figure 4.1 e), suggesting that the digester or the feedstock had sufficient buffering capacity to ensure high VFA levels did not cause system acidification (Murto et al., 2004). High levels of VFAs are reported to be a common cause of system failure (Franke-Whittle et al., 2014), but in this trial the high VFA concentrations appeared not to exert any negative effect.

The gas flow rate responded with an upward trend to the increased feed rates throughout the trial, starting at 6 standard cubic centimetres per minute (sccm), and reached a peak of 47.1 sccm at day 38 (Figure 4.1 c). The gas flow output continually increased regardless of increasing total VFA levels, further suggesting that high VFA levels are not a true indication of system performance and stability. Additionally, the methane composition of the biogas increased throughout the trial (Figure 4.1 d), with an average of 61 %. The starting methane composition was 45.9 %, but reached 71 % at day 39.


(a)
Feed Rate
Feed Rate (ml/min)
1.6
1.2
0.8
0.4
0
(b)
F:M & OLR
F:M
0.4
0.2
0
OLR
10
5
0
F:M
OLR
(c)
Gas Flow
Gas Flow (sccm)
50
40
30
20
10
0
(d)
Methane
Methane (%)
80
70
60
50
40
30
(e)
pH
pH
7.6
7.4
7.2
7
6.8
(f)
Acetic Acid
Amount (g/L)
15
10
5
0
DNA
Acetic Acid
(g)
Volatile Fatty Acids
Amount (g/L)
0.6
0.4
0.2
0
Propionic Acid
Isobutyric Acid
Butyric Acid
Isovaleric Acid
0
10
20
30
40
Time (Days)

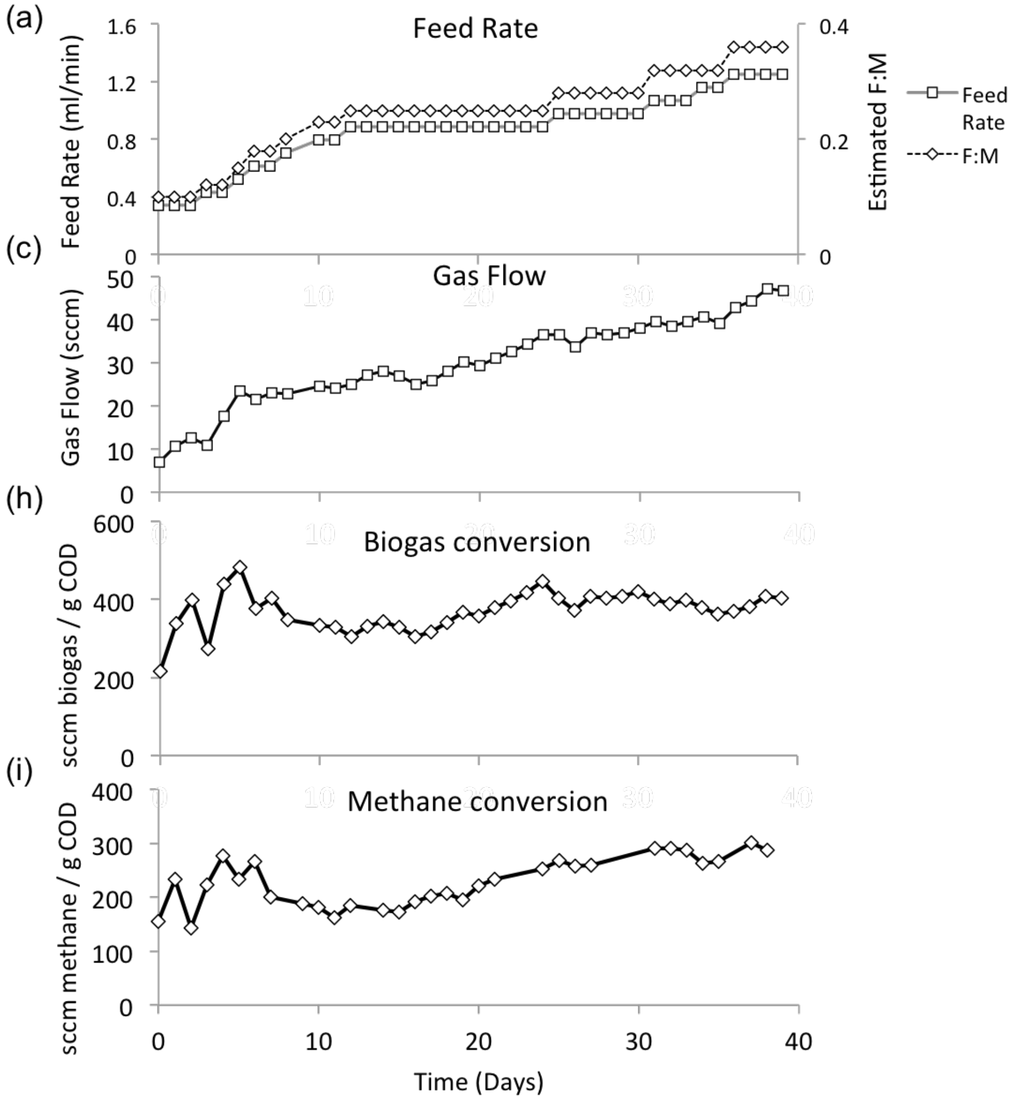


Figure 4.1. AD process data for experiment one showing feed rates (a), estimated F:M ratios and OLR over 39 days (b), gas flow (c), methane as percentage of gas production (d), pH (e), points at which DNA samples were taken and acetic acid concentration (f), other detectable volatile fatty acids (g), biogas conversion rates (h) and methane conversion rates (i).

The efficiency of conversion can be derived as another measure of system performance. This was determined as the amount of gas produced (sccm) per gram of COD put into the system from the feedstock. This can be measured as both the conversion to biogas or methane. The conversion efficiency to methane is important as this is the key output from the AD process. The rate of conversion to biogas was high in the first 7 days of the experiment (Figure 4.1 h), peaking at 483 sccm/g COD on day 5, but this decreased until day 12, at which the value starts to gradually increase. The same trend is observed for methane conversion rates (Figure 4.1 i), supporting the notion that the microbial community is acclimatising to the components of the feed and so the efficiency of the conversion of feedstock components to biogas, most importantly methane, is improving. The starting conversion rate to methane at day 1 was 155.2 sccm/g COD, and peaked at 301.2 sccm/g COD at day 38, showing nearly a doubling in the efficiency of conversion. This suggests that methanogen abundance has increased throughout the experiment as the volume of methane output increased, and this is expected to be reflected from the DNA sequencing of samples (Chapter 5). Additionally, as TFF is used as the dewatering method, the microbes are retained within the system, and so the slower growing methanogens would not be washed out. This would allow for the number of these microbes to gradually increase, which would be beneficial for the AD system and methane production.

4.4.2 One Week Trial

The initial experiment and sequencing of samples using the lab AD system for digesting biodiesel waste (Chapter 4.4.1) highlighted that there were dynamic changes occurring within the microbial community (Chapter 5.3.2.2). The changes that were observed from this trial had been monitored using the starting population and a sample after 25 days. Additionally, the qPCR data from the same experiment displayed the dynamic changes of selected microbes that were occurring over days.

The aim of this experiment was to gain a more comprehensive insight into the speed with which the microbial populations changed and whether these changes can be observed and detected. This experiment was conducted for 7 days using the lab scale AD system, where samples were taken daily and hourly. The feedstock used was malt wastewater. This process waste was collected from a manufacturing site and refrigerated until used. Figure 4.2 displays the process data.

The solids content of the inoculum collected from the established AD system was around 2 % (according to the on-site monitoring equipment) and the COD of the feed when collected from the site (before being refrigerated) was 42.5 g/L. It is therefore possible to estimate the F:M ratio. The feed was introduced into the system at a rate of 0.9 ml/min (estimated F:M 0.14) before being increased to 1.3 ml/min (estimated F:M 0.21) after 92 hours (Figure 4.2 e, f), since the VFA levels were decreasing, suggesting the feed rate could be increased. The normal practice of the industrial partners was to start with a feed rate at a F:M of 0.1 or below to allow the microbial community to adapt.

The measured concentrations of acetic acid and propionic acid indicate a change when the liquid waste was introduced into the system. The organic acid concentration increased to 0.68 g/L and 0.35 g/L for acetic and propionic acid respectively after 24 hours. The concentration of these two acids decreased until the feed rate was increased again at 92 hours where a slight increase was observed, then the level of acetic acid remained constant for the remainder of the experiment (Figure 4.2 a). The gas flow rate changed with the introduction of the waste material and continued to increase throughout the experiment, reaching a peak of 14.4 sccm after 147 hours (Figure 4.2 b). The methane content of the biogas fluctuated throughout the experiment, giving an average of 59 %. DNA extractions were carried out every 24 hours with the exception of samples being taken more frequently at 92, 94, 97 and 100 hours, when the feed rate was increased at 92 hours, to determine if the microbial community responded to the rapid increase in the rate of feed.

The conversion efficiency is varied in this experiment. After 24 hours this value was 173.7 sccm/g COD, but decreased to 88.8 sccm/g COD after 92 hours, when the feed rate was increased, before reaching 126.6 sccm/g COD at 147 hours (Figure 4.2 c, d). It would be expected that if the feed rate increased then the gas output would also increase, in a well running system. This does not necessarily mean the conversion efficiency would increase quickly when the feed rate is turned up as the microbial communities would need time to adapt to the components in the feedstock. The change in the feed rate does not appear to affect the VFA levels, suggesting the methanogens are utilising the acetic acid to produce methane, and the methane percentage is relatively consistent. The conversion efficiency to total biogas remained consistent also, but the methane conversion did decrease until the feed rate was increased. This correlates with the acetic acid levels that were decreasing until that point. Therefore, one suggestion could be that the increase in feed rates could be having a negative impact on other microbes in the process e.g. those involved in hydrolysis or acidogenesis, and this affects the efficiency of converting the waste to biogas. Another reason could be that parts of the microbial community do not specialise as quickly as others, or those required were low in abundance in the inoculum, therefore impacting on the efficiency of conversion. The measurement of VFA concentrations alone does not provide enough information, and other components, such as nitrogen, or other nutrients could be limited and therefore affecting the microbial community. Determining the microbial community would reveal if any of the above mentioned suggestions would explain the results.

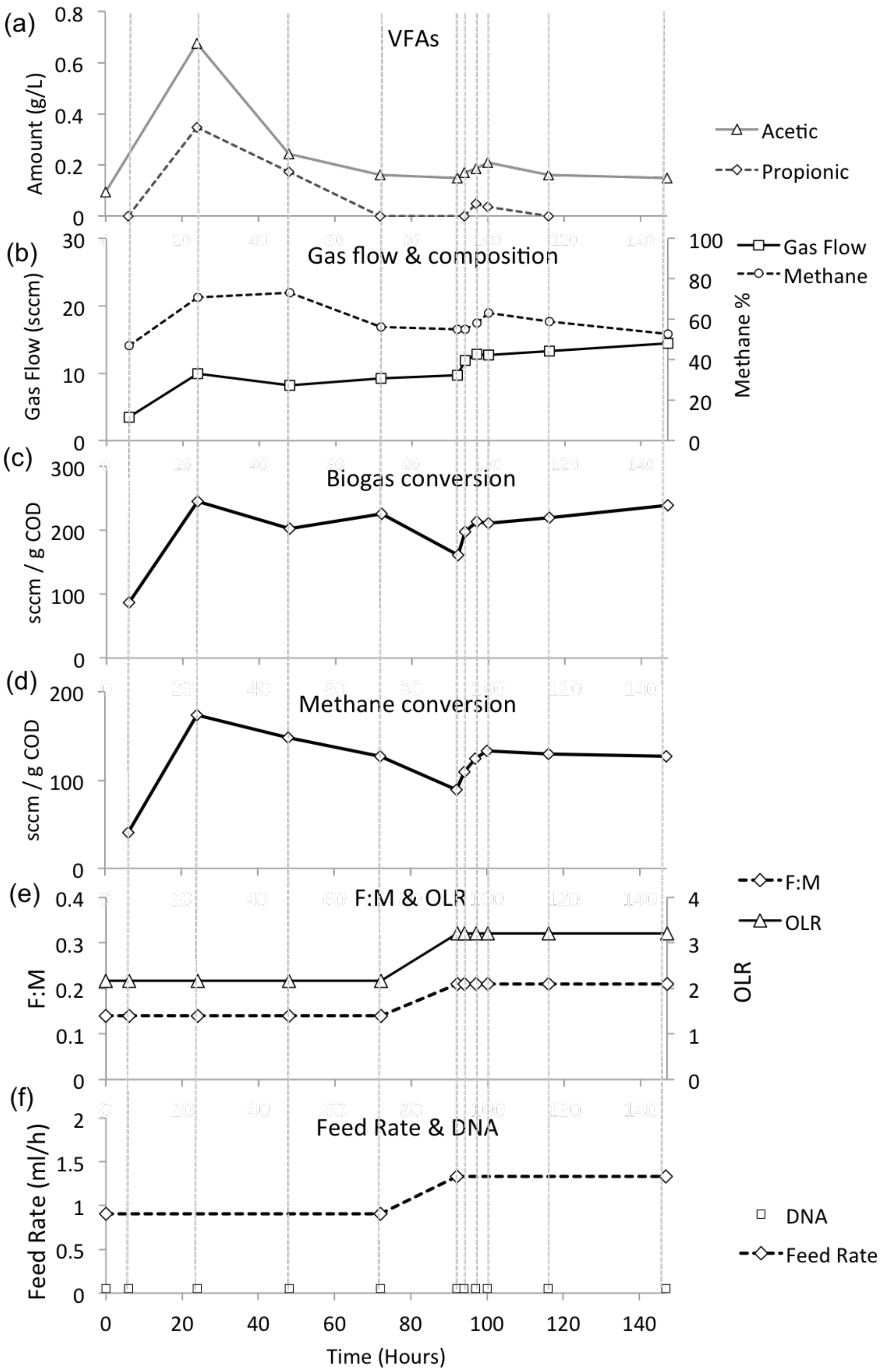


Figure 4.2. AD experiment two process data showing the concentration of organic acids (a), gas flow & methane levels (b), biogas conversion rates (c), methane conversion rates (d), estimated F:M & OLR (e) and feed rates & DNA extractions (f).

4.4.3 Triplicate Artificial Waste Trial

In the third experiment, three AD systems were run for 57 days until the maximum feeding rate was achieved (F:M 0.3 and OLR of 4.7 Kg COD.m^3.d.), under the same conditions. This was to investigate the microbial community change throughout the experiment, and if these were replicated in each of the three vessels, along with the process data. If the results gave the same outcome for the three vessels then this would suggest the microbial communities in AD are predictable when using a membrane bioreactor.

Two opposing theories exist to explain community formation: stochastic and deterministic. Deterministic (or niche) theory argues that such factors as competition and environmental parameters determine the community structure (Ofiteru et al., 2010). The stochastic (or neutral) model assumes all species to be ecologically equivalent along with the same demographic rates (Dumbrell et al., 2010), probabilistic dispersal and random birth-death events (Stegen et al., 2012). It has been suggested that both these theories could play a role in microbial community structures (Ofiteru et al., 2010). A recent study that investigated which theory best explains the shaping of a microbial community in AD was carried out by Vanwonterghem et al. (2014), where three 2 litre replicate systems were run over 362 days. This report states that niche factors are responsible for shaping the microbial communities and such factors as operational conditions and substrate availability are very important, causing a synchronisation in the microbial population. As stated in this report, targeted sequencing was used, and this therefore limited the reporting of the microbial community structure down to the genus level. Additionally, the systems that were used in the experiment did not retain the microbes. Although this can be reflective of some process scale AD systems, there is likely to be some microbial washout, especially of the slower growing microbes, and so therefore this can influence the microbial communities. To fully understand the changes that are occurring within the AD system, monitoring of the species and strain level are required, as these different microbes might be changing, and so greater resolution on the microbial community is required. This third experiment serves to run three systems in parallel, where microbial retention is a key part of the process, along with taking samples for sequencing, and then analysing these at the strain level.

The measured process parameter data for the three systems are displayed in Figure 4.3.

The three AD systems were run for 57 days under the same conditions, using the same feedstock. This was a mixture of four components: skimmed milk powder, malt extract, Coffee-mate ® (a mixture of glucose and vegetable fats) and yeast extract (Chapter 2.1.3.3). This mixture ensured there was a variety of nutrients available and provided a broad range of biological polymers: carbohydrates, proteins and fats. Importantly, the mixture was made frequently to ensure that few microbes were introduced into the system, preventing the community structure from being influenced by new microbes that are introduced. All of the systems were seeded using material taken from an established anaerobic digester, Naburn, York.

Initially the feedstock was added at a low rate (F:M of 0.05) to ensure that the systems were not overwhelmed with the addition, along with allowing the microbial community to acclimatise. After 12 days, the feed rate was increased, and this was increased further throughout the experiment, although the rate was decreased at the weekends, due to limitations regarding feed bottle size. The maximum feed rate target of 0.89 ml/min was reached (Figure 4.3 f) towards the end of the trial, giving a F:M of 0.3 and an OLR of 4.7 Kg COD.m^3.d. (Figure 4.3 g). No artificial mix was added from day 26 to 27 as a foaming event occurred. To overcome this, 50 g of rock salt was added to each system. Beyond this point, the foaming event did not occur again. High loading rates of proteins and lipids, along with high concentrations of acetic and butyric acid can lead to an increased occurrence of foaming within systems. At the time of the foaming event, the organic acid levels were high.

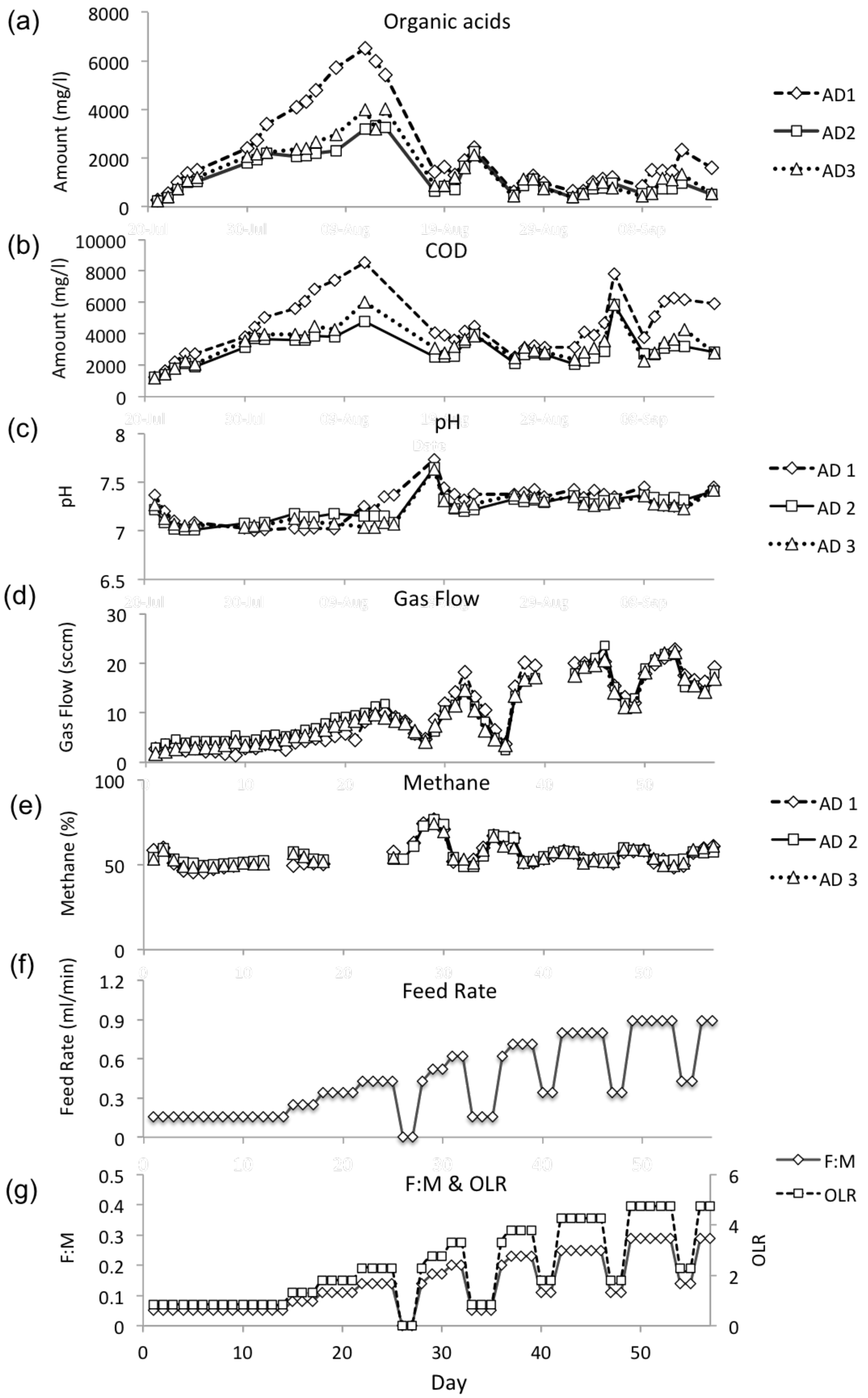


Figure 4.3. AD process data over 57 days for experiment three showing organic acid concentration (a), Chemical oxygen demand (b), pH (c), Gas flow (d), Methane percentage of gas (e), feed rates (f) and F:M and OLR (g).

The total measured organic acid concentration (Figure 4.3 a) changed with the introduction of the artificial mix, and steadily increased at a comparable level for each of the three systems, until day 10 where the organic acid measurements for AD 1 rapidly increased, reaching a peak of 6.54 g/L at day 22, whereas AD 2 and 3 had lower values of 3.19 g/L and 4.01 g/L respectively. The Standard Deviation (SD) values were highest during this time, with the highest value being 1.81, largely due to the high organic acid measurement of AD 1. When the feed was stopped, due to the foaming, these values rapidly decreased to 1.43, 0.66 and 0.89 g/L for AD 1, 2 and 3. The organic acid values thereon reflected the rate at which the feed was added. When the feed was increased, the organic acid levels increased, and when this was decreased, the organic acids did the same. The values of organic acids for each of the systems remained comparable from day 29 onwards, with a maximum SD reaching 0.70, and the lowest 0.05, with the largest values occurring towards the end of the experiment. This compared to AD 2 and 3, which appear to have much closer values, giving an SD maximum of 0.36 and lowest of 0.001. The digester COD values changed in a similar fashion to the organic acid measurements (Figure 4.3 b).

The pH (Figure 4.3 c) of all systems started above 7, but as the artificial mix was added, this decreased (to exactly pH 7 for all the systems). Caustic solution was added to ensure that the pH remained at 7. It is important that the pH remains at around 7 as this value has been shown to give a maximum biogas yield (Liu et al., 2008), and variations from this could decrease the efficiency of the system, as the organic acids would accumulate and not get converted to biogas. Once the feed was resumed beyond day 28, no caustic was added as the pH remained above 7, suggesting that the systems have sufficient buffering capacity and that the microbes are metabolising the artificial mix, and this is further supported by the data from the organic acid measurements.

The gas flow rate changed to the feed rate (Figure 4.3 d), with a low degree of variation between the three systems (highest σ 2.1). Although AD 1 appeared not to be performing as well as AD 2 and 3, in regards to higher organic acid and COD values, the average gas flow throughout the trial was 10 sccm, compared to 10.5 and 9.7 sccm for AD 2 and 3 respectively. The quality of the gas appeared also to be comparable between the three systems, with average methane compositions of 55.7, 56.2 and 55 % for AD 1, 2 and 3 (Figure 4.3 e). The SD of gas quality between the three systems reached a maximum of 3.3, indicating the variability between the three systems is low. The methane content of the gas decreased slightly when the systems had artificial mix added, but when this was decreased, the methane content increased. This could be possibly explained that as the mix had a high sugar content, this would influence the methane output, as high sugar waste has a theoretical yield of 50 % (Alves et al., 2009). When the system is having the mix added at a higher rate, the simple carbohydrates, such as sugar gets more readily utilised before other components, whereas when the feed rate is decreased, the other components in the mix are utilised, as the feed becomes more limiting.

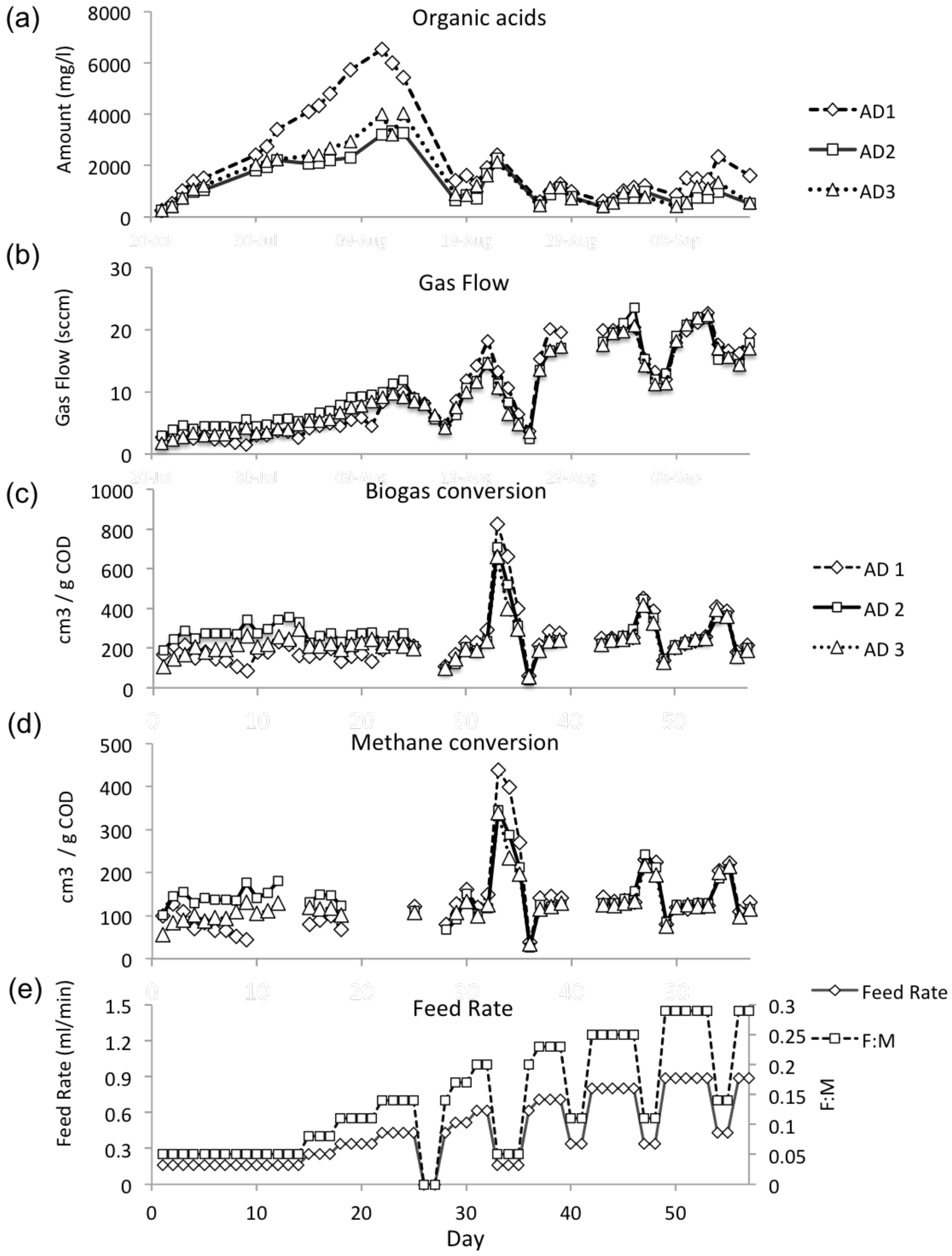


Figure 4.4. AD process data from the triplicate systems showing the organic acid concentrations (a), gas flow (b), biogas conversion rates (c), methane conversion (d), feed rates and F:M (e).

The conversion rates for this experiment appear to be comparable between the three systems, although there is more variation at the start of the experiment. The conversion rates for both the biogas (Figure 4.4 c) and methane (Figure 4.4 d) appears to peak when the feed is reduced over the weekend, which coincides with a decrease in organic acids and COD. This suggests that the systems were being over fed during the five days, and the decrease at the weekends proved to be beneficial during the experiment. There is a noticeable peak around day 34, and this also coincides with when the feed was reduced. This suggests that the feed rates (OLR) during the week could either be too high, or the feed rate could be having some inhibitory effect on the microbes, possibly the methanogens. This is because the total organic acids increase during the week, but decrease at the weekend, and the methane levels are higher during the lower feeding rates. High or rapidly increased OLR have been reported to have a negative effect on AD systems, such as resulting in increasing VFAs and lowered gas output (Hori et al., 2015). It is also possible that ammonia levels could increase when the feed rates were high, as protein was put into the system. Increased ammonia concentrations have been shown to have an inhibitory effect in AD systems (Moestedt et al., 2016), but as the feed rate was decreased at the weekends, the microbial community could use the ammonia and so the inhibition reduced. It could also be possible that long chain fatty acids could be present within the systems, again having an inhibitory effect. LCFA have also been shown to be inhibitory to AD systems, especially acetate utilising methanogens (Ma et al., 2015). It has also been reported that food waste contains low levels of micronutrients, such as Selenium and Cobalt, with the former required for coenzymes in the reduction of formate, preventing propionate accumulation (Yirong et al., 2014). Therefore the high levels of VFAs could be reduced with the addition of micronutrients. Measurements of all these parameters would have to be taken to prove these.

4.5 Conclusions

The three experiments conducted using the lab scale digesters demonstrated that they are useful systems for testing a variety of aspects in AD, such as measuring process parameters and taking DNA samples whilst varying process conditions.

The first experiment, using biodiesel waste shows that the system can be run at high feed rates (estimated F:M and OLR of 0.36 and 6.17 Kg COD.m^3.d respectively), where the levels of VFAs are high and reactor acidification does not result as a consequence. Therefore pH alone is not necessarily a useful proxy for system stability, especially when digesters and/or feedstocks have high buffering capacity (Franke-Whittle et al., 2014). Buffering capacity was not measured directly during this experiment, and so it is hypothesised that the feed or digester has a high buffering capacity, based on the pH. The conversion rates to biogas and methane at the end of the experiment are high compared to other points throughout the experiment, and these generally increased throughout the trial, suggesting that the microbial communities are becoming more specialised. Although the acetate levels increase during the trial, they decrease during the latter stages of the experiment, again suggesting that the microbial communities are changing. More specifically, decreasing acetate levels suggest that the methanogen numbers are increasing as this gets used, and this correlation between an increase in *Methanosarcina* abundance and decreasing acetate has been reported (Hori et al., 2006). There is an initial increase in the methane conversion efficiency during this experiment for the first seven days before a decrease. It would be interesting if the metagenomic analysis of samples during those time points reflects a change in the microbial community, mostly the methanogens. It is also possible that each digester has an optimal operation conditions (Franke-Whittle et al., 2014), such as the ability to process high VFA levels, suggesting that process data alone is not truly reliable, and that understanding the microbiology could be more informative. If this experiment were continued it would have been interesting to see if the conversion rate would continue to increase or level out and if the feed rate could be further increased.

The second experiment potentially demonstrated that the AD systems acclimatise to the feedstock much quicker than has been suggested as the feed rate was higher (estimated starting F:M and OLR of 0.14 and 2.17 Kg COD.m^3.d respectively) than would be normally carried out in industry (F:M of 0.05 – 0.1), as there was no significant accumulation of VFAs over the time measured. However, as mentioned from experiment one, VFA concentrations alone do not provide a true reflection of system performance, but an indication. Again, if the experiment was continued, it would be interesting to determine if the feed rate could be further increased at a faster rate, without causing system instability or inefficiencies. The conversion efficiency to methane appears to have a downwards trend, suggesting that actually the increased feed is not beneficial to the system and that possibly the microbial communities in the system need more time to adapt. Intriguingly, as previously reported, low methane output levels would usually coincide with high VFAs (Xiao et al., 2013), but for this experiment, both parameters were low. This information is conflicting with the VFA measurements, which remained in low concentrations in the system, and therefore analysis of the microbial communities may explain the reasoning for low VFAs but also low efficiencies.

In the experiment where three systems were run in triplicate, there was a similar output in process data. Although one system appears to differ for some process measurements, especially during early stages of the experiment (organic acids and COD), the three systems generally track each other, suggesting that the microbial community composition could be similar. It is also interesting to note that the decreased feed rates during the weekends appeared to have a positive effect on the systems, as the process data for the three systems aligned beyond these points. An explanation for variation could be that the starting material is not homogenous, giving variability in the starting microbial communities between the three systems. This variability in community could still result in the same process data, or the community all tends to shape in the same way, even if there is variability at the start. Furthermore, digester differences could be a factor of experimental variability. Although the three systems were run in the same way, using the digesters that were built and designed in the same way, some variability could occur. Examples include that the mixing and/or heating could be more efficient in one system compared to the others. A longer experiment would be beneficial to truly demonstrate if the three systems converge. The data from this experiment initially suggests that deterministic factors shape the microbial communities, as suggested by Vanwonterghem et al. (2014). Even though there is variation in the process data at the start of the experiment, this variability between systems decreases, and the AD systems appear to track each other closely, which would suggest that the environment is having an impact on the community structure. The sequencing data from this experiment looking at the microbial community structure and dynamics would reveal if this hypothesis is true.

A comparison between the three experiments is somewhat difficult as the three independent experiments were run under differing conditions i.e. feed composition. It is possible to compare the systems based on an important consideration – the methane conversion efficiency, when at a comparable F:M ratio, in this instance, 0.2. For experiment one this was 201 sccm/g COD, experiment two, 119 sccm/g COD and experiment three, AD1 149 sccm/ g COD, AD 2 117 sccm/g COD and AD 3 126 sccm/g COD. These results are suggestive that the feedstock ultimately determines the methane composition of the biogas and therefore influences the conversion efficiency value. The limitation with using the efficiency value is that there is an assumption that the feed added (as grams COD) is equal, but feedstock composition affects the methane output. This is because the conversion efficiency value is a measure of the amount of feed added to the system (grams COD), the gas flow (sccm) and the methane composition (%). Therefore, a feed that produces a high methane composition is likely to have a higher methane conversion efficiency than one that produces a lower methane output. Experiment two and three have comparable conversion efficiencies, except for AD 3 of experiment three, and it should be noted that the composition of these feedstocks could be somewhat similar. This is because the feed taken for experiment two was collected from a malting facility, so high in sugars, and the feed made for experiment three was composed of milk and malt extract (amongst others), which again is high in sugar. Therefore the conversion efficiencies would be expected to be similar, although experiment three has other components and so would explain why these conversion efficiencies are marginally higher. The conversion efficiency of experiment one is almost double that of experiment two and three. This feed was collected form a bio-diesel processing facility, and so was expected to contain notable amount of fats and glycerol, which would explain the high methane output. This is due to the theoretical yields, as fats and proteins give a higher methane yield (69.5 % and 68.8 % respectively), in comparison to carbohydrates, at 50 %. The volume of biogas generated from fats is the highest, due to the high energy value, and carbohydrates are the lowest (Alves et al., 2009). This would explain why experiment one has the highest methane output and the highest methane conversion efficiency of the three experiments. It can be concluded that using the efficiency of conversion is a valuable method for determining the performance of a system, but the feed composition must be accounted for also, and so measuring the composition of the feedstock would be important.

5 Molecular Tools for Determining the Microbial Community Structure and Dynamics

5.1 Introduction

Previous research has characterised the microbes involved in the AD process using a variety of techniques. These have mainly focused on using 16S rRNA amplification for the process e.g. Li et al. (2013), and this method has the potential to provide biased results. The use of shotgun sequencing in AD appears to be uncommon, with few articles published e.g. Yang et al., 2014. This can be because sequencing technology has exponentially improved, with increased output and accuracy (Solomon et al., 2014). This in turn has the potential for a greater understanding of complex microbial communities but the pitfalls associated with this technology is that the processing and interpretation of large volumes of data generated can be challenging. There also appears to be a lack of suitable pipelines available that can process such large datasets. There are numerous contig assemblers that can be used to handle these datasets, e.g. Megahit (Liu et al., 2015), Newbler (Roche), SPAdes (Bankevich et al., 2012), Metaray (Boisvert et al., 2012) and IBDA-UD (Peng et al., 2012), amongst others. The choice of assembler is specific to each dataset generated, and often trial and error is used to determine the assembler that produces the best results. A challenge with some contig assemblers is that these expect equal coverage for genomes, which would generally be found in the sequencing of single organisms, but metagenomes do not have this. Therefore if assemblers assume that everything is equal, contigs that have a lot of coverage would get discarded (Reddy et al., 2014). The choice of assembler is therefore important, so trying numerous ones often appears to be the best option, or using numerous assemblers to generate contigs from a dataset. The aim of assembling contigs is to ultimately reconstruct complete microbial genomes and gain a greater understanding of microbial functions.

This chapter describes the analytical pipelines that can be used to understand the microbial communities involved in the AD process. The microbial communities and the dynamic changes are monitored using a variety of molecular techniques such as Ion-Torrent, qPCR, Illumina HiSeq and PacBio sequencing platforms, and analysed using different bioinformatic analysis techniques. The resulting sequencing data has been processed using several methods, such as contig assembly, clustering and gene annotation. Examples of tools used include Megahit and IDBA-UD for contig assembly. Additionally, custom scripts have been developed to process the data, such as clustering the contigs and searching the databases. It is hypothesised that shotgun metagenomic sequencing provides more informative data regarding microbial community dynamics and functions compared to targeted sequencing.

5.2 Aims

- Determine which DNA extraction kit was most suitable for use on anaerobic digester samples
- Ensure the extracted DNA quality is sufficient for Next Generation Sequencing by Ion Torrent and to establish initial data analysis of those data
- Investigate qPCR as a means to measure dynamic changes in populations for selected organisms
- Investigate short and long read technologies
- Develop pipelines to process and assess the utility of such data

5.3 Results and Discussion

5.3.1 DNA Extraction

Three different DNA extraction kits were trialled to determine which of these was most suitable for the extraction of genomic DNA from anaerobic digester samples (Chapter 2.3.1). The extracted DNA was checked for quality on a 1 % agarose gel (Figure 5.1).

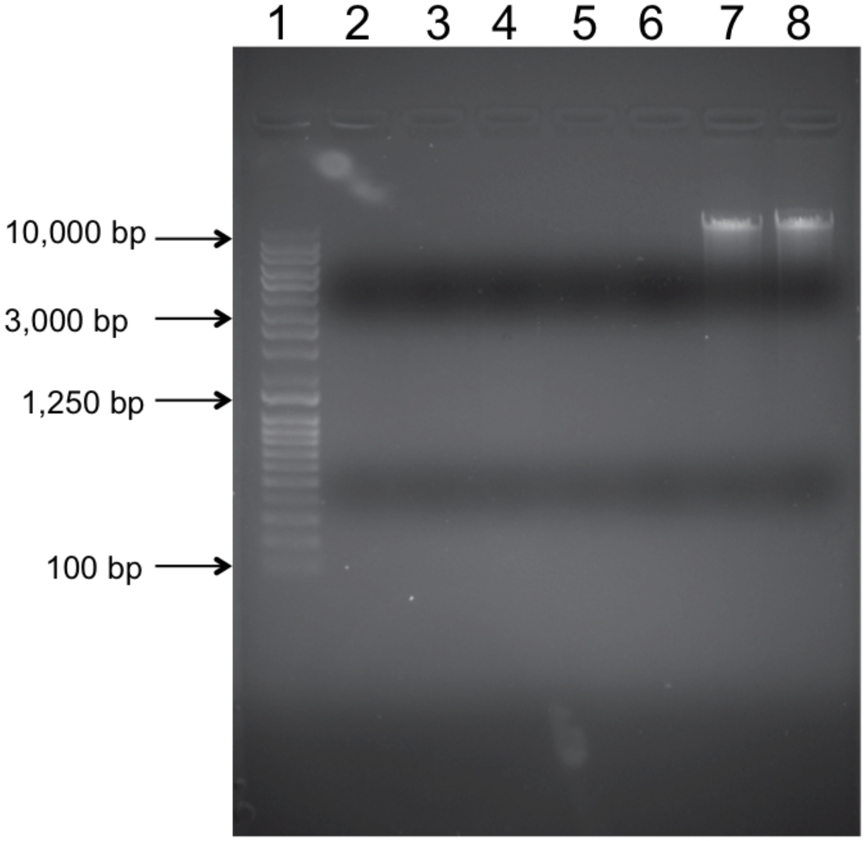


Figure 5.1. DNA quality from the three different DNA extraction kits using the same AD sample. Lane 1, Q-step 4 ladder (Yor Bio). Lane 2, 3 and 4, DNA extracted using Qiagen. Lanes 5 & 6, Epicentre. Lanes 7 & 8, MO-BIO.

Using the same anaerobic digester sample to determine which kit yielded optimal amounts of DNA with a high quality demonstrated that the MO-BIO Powersoil Kit was most suitable. The other kits showed that low levels of DNA were recovered from the samples, whereas the PowerSoil kit produced a distinct band. A possible drawback of the PowerSoil kit is the slight level of DNA shearing that has occurred during the extraction process, as shown in Figure 5.1.

5.3.2 Ion Torrent Metagenomic Sequencing

5.3.2.1 Ion Torrent PGM

Two samples (Day 0 and Day 25), from experiment one (Chapter 4.4.1), were sequenced to determine whether changes occurring in the microbial populations from the starting sample to the microbial community that had acclimatised to the particular feedstock (taken from a biodiesel refinery site), could be measured. This initial sequencing run using the Ion Torrent was also used as a trial to demonstrate that the extracted DNA using the selected method was suitable for metagenomic sequencing. The Ion Torrent PGM platform, using a 318 chip and the 400 bp kit was used to sequence the two DNA samples (Chapter 2.3.2). The read length distribution obtained from both samples is displayed in Figure 5.2.

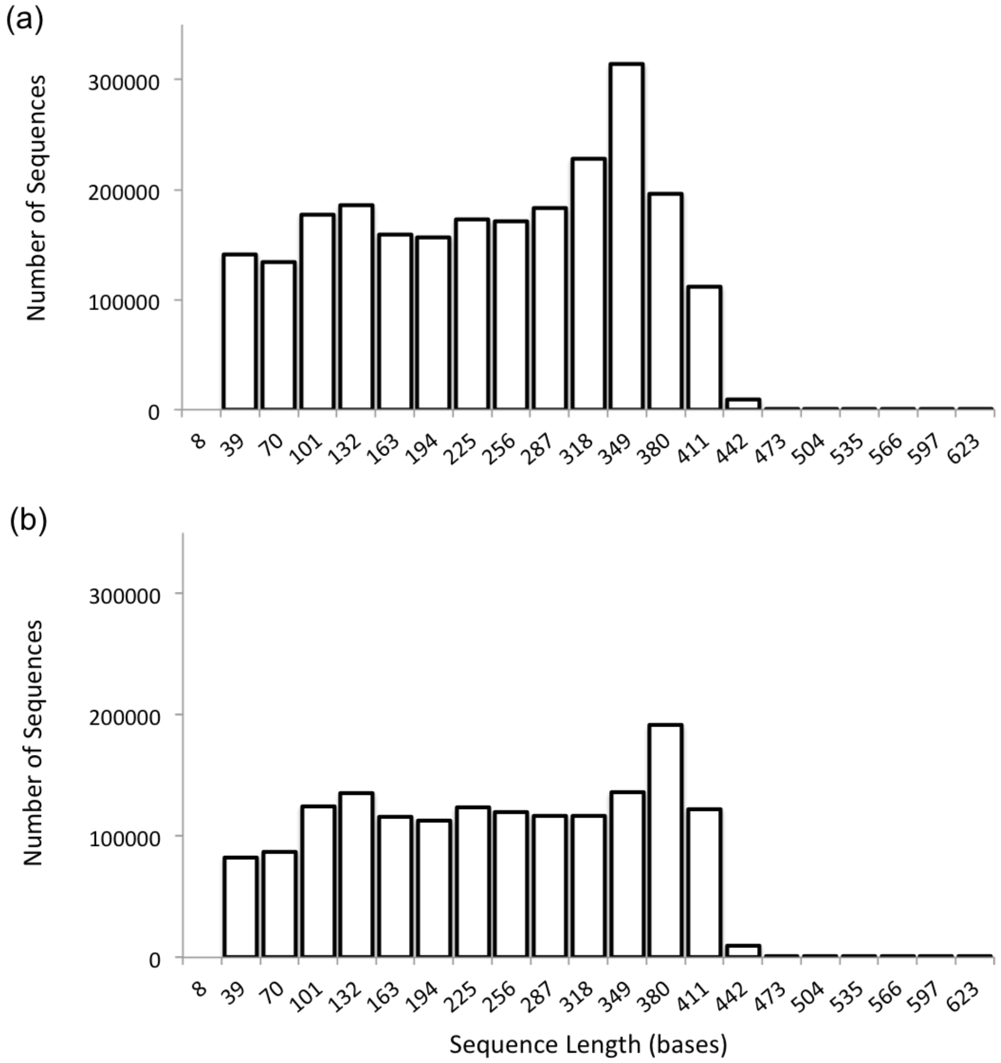


Figure 5.2. Sequence length (bases) distribution obtained from the Ion Torrent PGM sequencing platform for Day 0 (a) and Day 25 (b) samples.

The mean sequence length from the Ion Torrent platform was 222 ± 112 bases and 225 ± 115 bases for Day 0 and Day 25 samples respectively. The total number of bases from Day 0 sample was 520,061,463 and 360,310,163 bases for Day 25 sample. Although this technology has the advantage that longer reads can be obtained, compared to other short read sequencing technologies, the amount of sequence data generated using this platform is not sufficient for complex community metagenomic studies. That said, this method provided a useful tool to start investigating the microbial communities.

5.3.2.2 Annotation software

The sequence data from the two samples (Day 0 and Day 25) were uploaded to MG-RAST (Meyer et al., 2008), using default settings for analysis (Chapter 2.4.1). Post QC sequencing information is displayed in Table 5.1.

Sample	**Day 0**	**Day 25**
Bases	296,885,752	198,597,836
Number of sequences	1,819,527	1,239,582
Mean Length (bases)	163 ± 82	160 ± 82
Alpha Diversity	592	538

Table 5.1. Post QC sequencing information for Day 0 and Day 25 samples according to MG-RAST.

For the Day 0 sample, Bacteria accounted for 92.6 %, and Archaea 2.4 %, whereas the Day 25 sample, Bacteria accounted for 89.6 % and Archaea 9.2 %, exhibiting an increase in the methanogens. The four most dominant phyla in the Day 0 sample were Proteobacteria (33.2 %), Bacteroidetes (28.2 %), Firmicutes (12.6 %) and Actinobacteria (4 %). A small proportion of the data was categorised as unclassified (4.3 %). In contrast to this, the most abundant phylum in the Day 25 sample was Bacteroidetes (29.6 %), which is consistent with the Day 0 sample. The second most abundant was Proteobacteria (23.1 %), showing that microbes belonging to this phylum have decreased. Firmicutes (15.5 %) was the third most abundant phylum, showing an increase, and Actinobacteria (3.1 %) showing a decrease. Again the unclassified accounted for a sizeable proportion at 5 % (Figure 5.3).

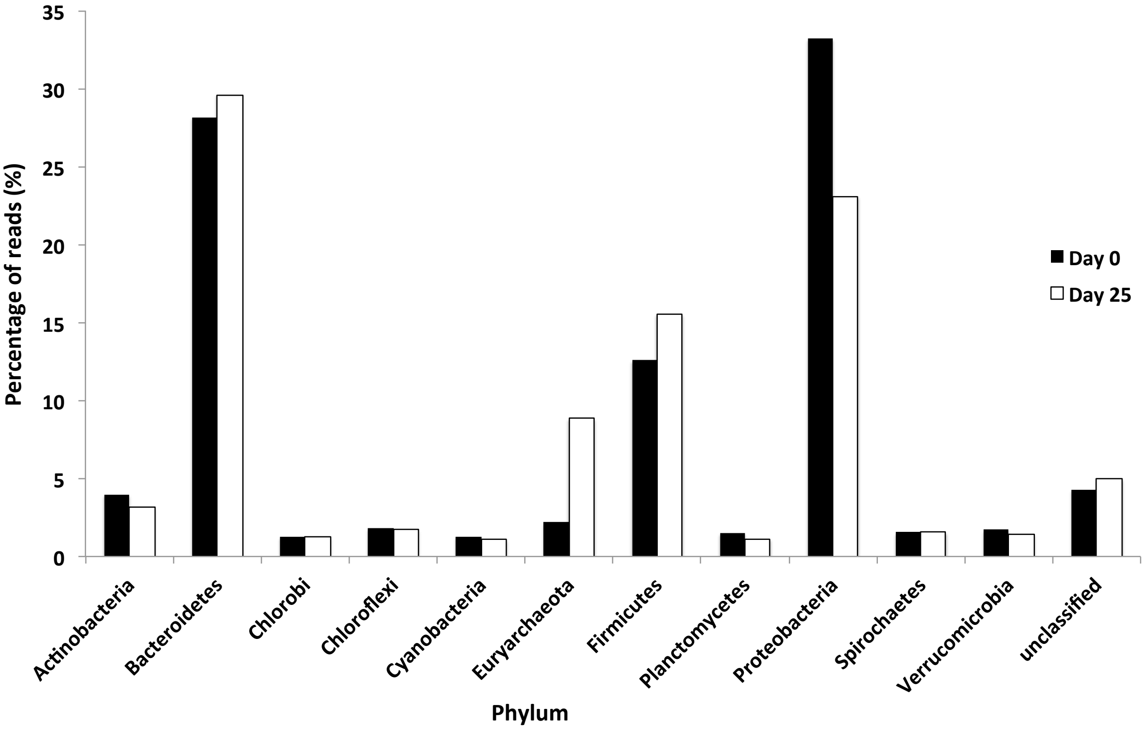


Figure 5.3. The percentage of reads that are annotated to known organisms at phyla level for the Day 0 and Day 25 samples from experiment one, according to MG-RAST.

The MG-RAST annotation software also provides the alpha diversity from each sample. For the starting sample (Day 0), the alpha diversity was 592 species, whereas after 25 days, the alpha diversity was 538 species, suggesting there is a simplification of the microbial communities.

An approach to look at the most abundant organisms present is to impose a cut-off threshold. The cut-off for organisms that were classed as most abundant was 0.5 % or over of total reads. The most significant change in the microbial community is the increase in abundance of methanogens. At Day 0 they account for 0 % of the most abundant organisms, but after 25 days, methanogens (*Methanosarcina* and *Methanoculleus*), account for 10 % (Figure 5.4), suggesting these microbes are involved in the digestion process, as these have significantly increased in abundance. *Methanosarcina* are known acetate using methanogens (Jäger et al., 2009) and during this experiment the acetate concentrations were high, and so the increase in abundance of this organism correlates with the process data. *Methanoculleus* uses hydrogen and carbon dioxide to produce methane (Anderson et al., 2009) and the increase in this organism could also possibly be due to the increased acetic acid levels, where hydrogen is produced. Hydrogen utilising methanogens are required for acetate production and so for acetate levels to increase, an increase in such organisms as *Methanoculleus* could be expected. Generally the abundant organisms do not appear to change drastically in number, when comparing the two samples, but more detail from samples in-between these time points is required. *Syntrophus* was an organism that showed a decrease in abundance, from 10 % to 4 % from Day 0 to Day 25. A large proportion of the data is grouped as unassigned; 45 % and 37 % for Day 0 and Day 25 respectively.

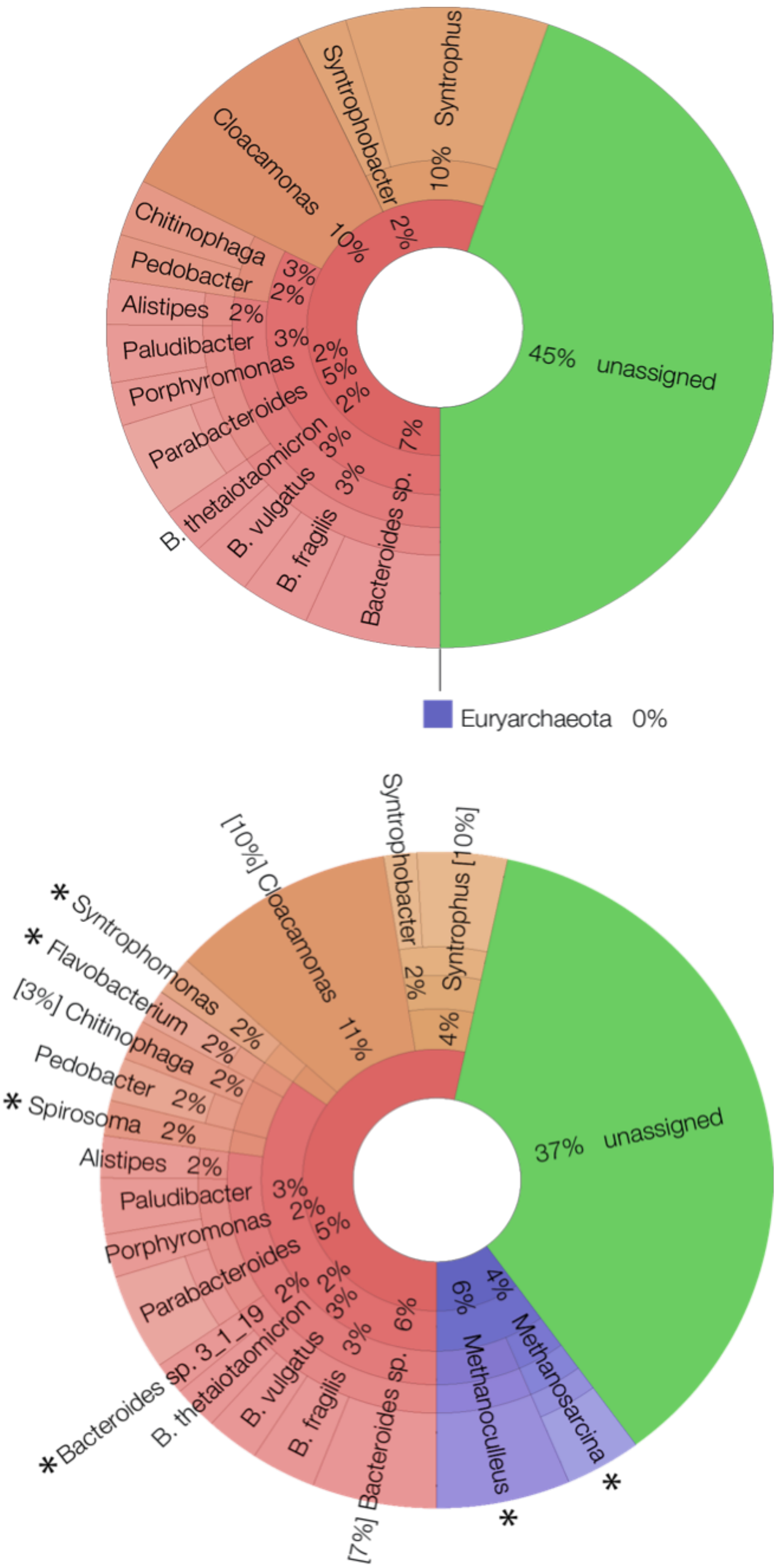


Figure 5.4. Krona graph representation of the most abundant organisms at the genus level where 0.5 % and above of total reads were used as the cut-off for Day 0 (a) and Day 25 (b). [ ] displays percent in Day 0 sample. * displays those organisms that are present in Day 25, but not Day 0 sample.

Changes in the microbial community structure would be expected as the naïve microbial community had been exposed to a change in feedstock, therefore the microbial numbers and composition would be predicted to change.

The amount of sequence data obtained from the Ion Torrent platform did not provide enough of a comprehensive insight and complete coverage of all the microbes present in the sample, as the read depth of data from the Ion Torrent was low. Although MG-RAST is a good tool to use for metagenomic studies, there are limitations. It has been reported (R. Randle-Boggis, Personal communication) that MG-RAST is highly accurate at correctly assigning sequence to known organisms at phylum level, but at species level, the assignments become more inaccurate. This is important because incorrect assignments can give an inaccurate sense of the microbial communities. Unassigned data is another limitation associated with MG-RAST. For example, as displayed in Figure 5.4, there is a large proportion of data that is unassigned (45 % and 37 % for Day 0 and Day 25 samples respectively). Therefore the sequencing platform, coupled with MG-RAST, does not provide the best overall interpretation of the microbial community.

5.3.3 Quantitative PCR

Quantitative Polymerase Chain Reaction (qPCR) is a method that can be used to monitor the relative changes that are occurring for selected microbes. This method was used on the time course samples collected from experiment one. Primers for qPCR were designed based on the initial sequencing data obtained from the Ion Torrent sequencing platform (Day 0 and Day 25) that was uploaded and assigned using MG-RAST. Eight targets were selected, based on species that appeared to show different population dynamics. Table 2.3 displays the primers used.

End point PCR was carried out (Chapter 2.3.3.1) with the primer pairs to ensure a single band was formed for each target, when run on a 2 % agarose gel (Figure 5.5), along with a melt curve to ensure there was only a single product.

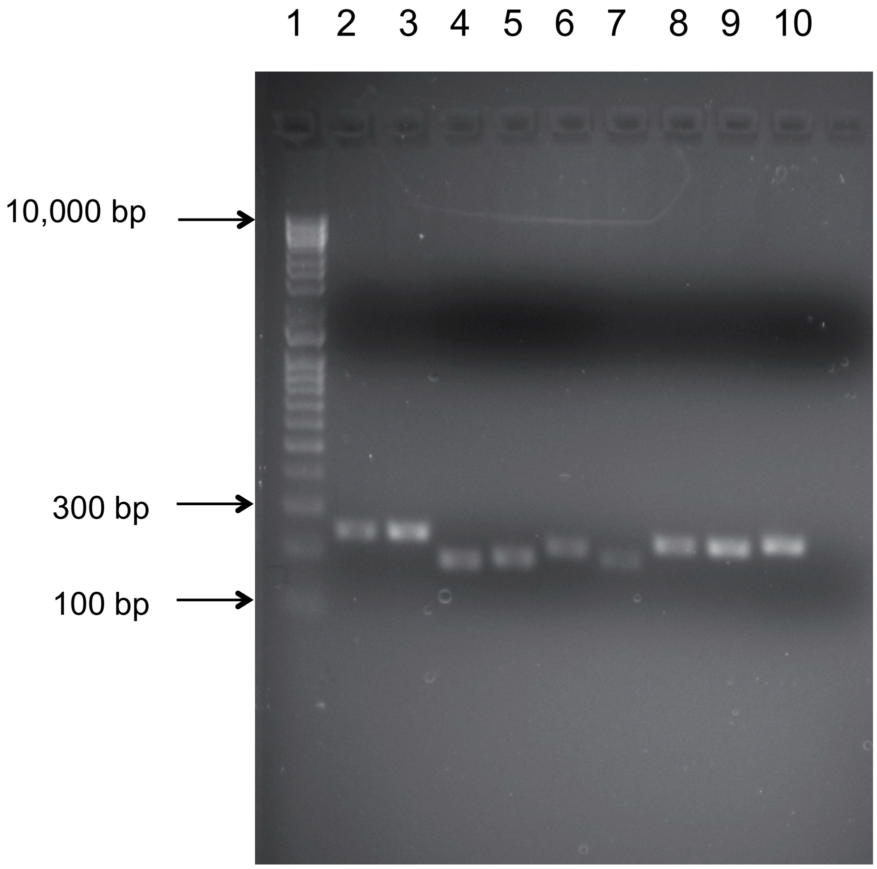


Figure 5.5. PCR products showing single bands from the selected targets on a 2 % agarose gel from experiment one. Lane 1. Q-step 4 ladder (Yor Bio), 2. *Methanoculleus marisnigri* 3. *Dyadobacter fermentans* 4. *Syntrophomonas wolfei* 5. *Bacteriodes* 3_1_19, 6. *Bacteriodes vulgatus* 7. *Methanosarcina mazei* 8. *Candidatus Cloacamonas* 9. *Pedobacter heparinus* 10. *Syntrophus acidotrophicus*

qPCR was carried out (Chapter 2.3.3.2) to determine the dynamic changes of specific organisms that are occurring over the 39 day experiment in more detail. The two samples that were sent for sequencing provided a snapshot of the changes at those time points, but qPCR allows for the dynamic changes that are occurring in-between and beyond those time points to be monitored. Each measurement was run in triplicate and the data from each target was normalised against the first sample (Day 0), to display the relative changes compared to the starting population. The results from qPCR showed that the fold changes for the majority of the organisms selected are low (Figure 5.6 a).

A significant change occurs for *Methanoculleus* species, which after 5 days had over a 15-fold increase, but after 11 days, this number had drastically reduced (Figure 5.6 b). During this point the VFA levels increased, although were initially low. Increasing VFA levels can increase the hydrogen concentrations within a system, and there appears to be a correlation between VFA levels and *Methanoculleus* abundance. Hori et al. (2006) showed that the numbers of *Methanoculleus* declined during the accumulation of VFAs. In fact, a different hydrogenotrophic methanogen dominated during this time, but the detection of other hydrogen consuming methanogens was not carried out in this experiment. Interestingly, *Methanoculleus* species and *Syntrophus aciditrophicus* increase at comparable rates during the trial. *Syntrophus* species ultilise benzoate and certain fatty acids in association with hydrogenotrophic methanogens to ensure the hydrogen levels remain low (Kim et al., 2013). This possibly explains the comparable increase in abundance of these two organisms during the trial, especially after day 11, as these microbes require may a close syntrophic association. *Methanosarcina* species, which are known aceticlastic methanogens (Sousa et al., 2013), had the largest increase of the eight microbes, with a fold increase after 39 days reaching over 1000 (Figure 5.6 c), although this number varied drastically throughout the experiment. *M.mazei* has the ability to grow on a variety of substrates including acetate and methanol (Jäger et al., 2009), and acetate was in abundance, possibly along with methanol, potentially explaining why the numbers of this organism increased throughout the trial. This correlation between high VFAs and *Methanosarcina* dominance has also been reported (Franke-Whittle et al., 2014). Additionally, *Methanosarcina* has been stated to have a higher growth rate during high acetate levels, especially compared to other acetate utilising methanogens, such as *Methanosaeta* (Walter et al., 2012). This would explain the significant increase of this organism during high VFA levels and the subsequent decrease of the acetate concentration. Overall, the qPCR data shows that abundance changes for each of the microbes do not have a gradual change, but instead demonstrate that fluctuations in the microbial abundance is occurring over days.

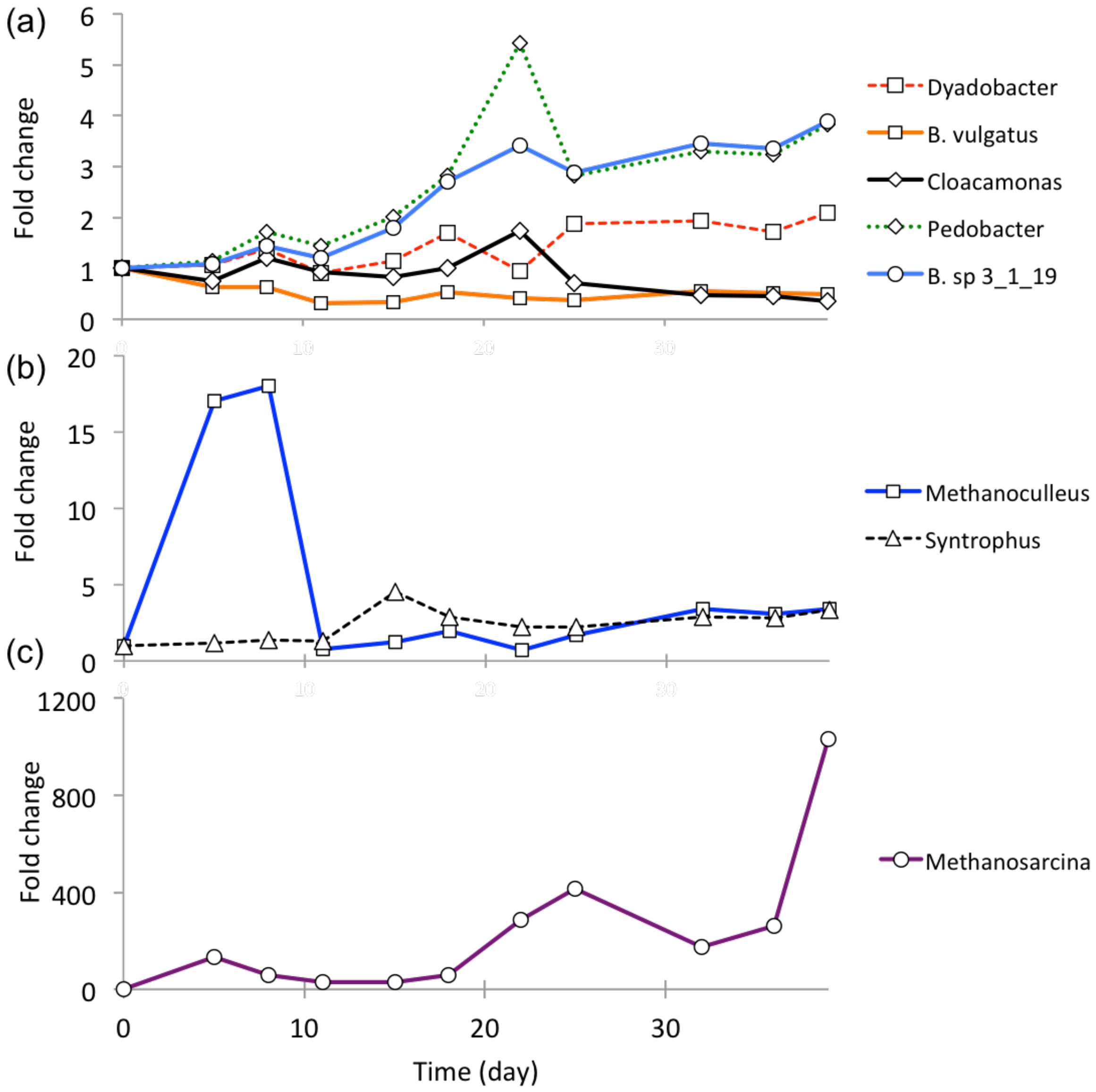


Figure 5.6. qPCR data showing the relative change in abundance of markers believed to be associated with eight different organisms, (a) *Dyadobacter*, *B.vulgatus*, *Cloacamonas*, *Pedobacter* and *Bacteroides* sp 3_1_19, (b) *Methanoculleus* and *Syntrophus*, (c) *Methanosarcina*.

Although most organisms exhibited some change in abundance throughout the experiment it is probable that the MG-RAST program assigned sequences to organisms, even though they were not exact matches. A random gene from an assigned organism by the program was selected to form the probe. This could explain why such organisms as *Methanoculleus* and *Methanosarcina* have such a large increase in numbers – the possibility that the probe could have selected for two or more different strains of these organisms.

Although the qPCR data was highly informative on the dynamic changes occurring within the system over the experiment, the main challenge associated with this technology is the limitation on the number of targets and the number of reactions. Therefore this limits the resolution. Additionally, only information on dynamic changes are obtained, not function. DNA sequencing offers a more cost-effective alternative to this, where not only the dynamic changes occurring throughout the experiment are obtained, but detailed information on the microbial functions.

5.3.4 Illumina HiSeq Metagenomic Sequencing

In addition to the initial sequencing using the Ion Torrent platform and qPCR data (from experiment one), DNA samples from the three experiments were sent to be sequenced using the Illumina Hi-Seq platform (Chapter 2.3.4). This sequencing platform allows for more in-depth sequence coverage to be obtained on the microbes and the dynamic changes. A total of 39 samples were sequenced (Table 5.2), from three independent experiments, as discussed in Chapter 4.

Experiment	Sample (day)	Concentration (ng/µl)
One	0	146
	5	59
	8	123
	11	151
	15	116
	18	82
	22	152
	25	104
	32	123
	36	143
	39	127
Two	1	126
	2	129
	3	132
	4	130
	5	134
	5 (+2h)	140
	5 (+5h)	148
	5 (+8h)	142
	6	137
	7	138
Three	1	1 – 98, 2 – 123, 3 – 97
	10	1 – 79, 2 – 97, 3 – 109
	21	1 – 95, 2 – 88, 3 - 76
	31	1 – 80, 2 – 98, 3 - 83
	43	1 – 140, 2 – 154, 3 – 149
	52	1 – 156, 2 – 195, 3 - 139

Table 5.2. Samples taken throughout the three experiments and DNA concentration of these samples sent for sequencing using Illumina HiSeq platform.

5.3.4.1 Contig assembly using Megahit

The sequencing data obtained from the Illumina platform provides short read lengths, with 2 x 100 bp reads. These short reads need to be assembled into contigs, which are overlapping DNA segments. The main aim of assembling the short reads is to eventually reassemble the entire genome of organisms.

Contigs were generated by Megahit (Chapter 2.4.2) using the sequencing data obtained from the three experiments that were sequenced using the Illumina HiSeq platform. This assembler was selected because it was the only one that could process the large dataset. This assembler produced 11,618 contigs over 10 kb, the largest contig being 415.6 kb, for the three experiments.

To begin to visualise and interpret the data, the 250 longest contigs from experiment one were plotted and it was noted there were groupings of these contigs that exhibit the same pattern of change throughout the trial, suggesting that they could belong to the same organism (Figure 5.7). The data were normalised to the starting sample so to display the relative changes occurring throughout the experiment.

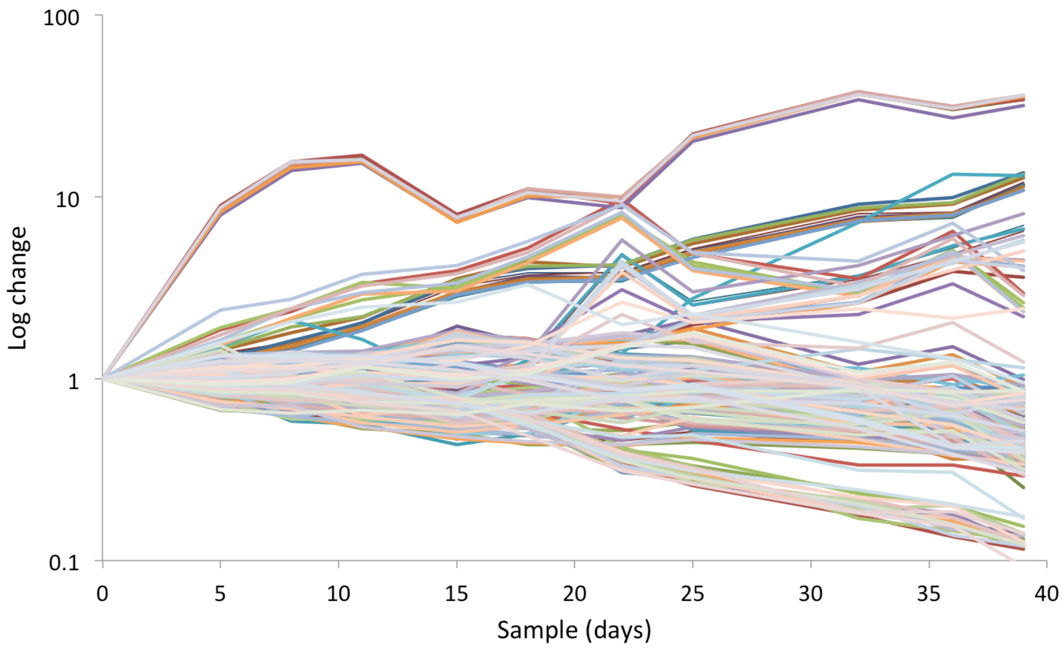


Figure 5.7. Log change of the 250 longest contigs assembled using Megahit when normalised to the starting sample.

5.3.4.2 K-means clustering

Plotting the contigs into graph format enables a good visualisation of the data, and an appreciation for how the groups of contigs begin to cluster, showing distinct patterns. The limitation of this method however is that there are more than 11,000 contigs that were over 10 kb, and plotting these is time consuming.

Therefore, an alternative method that grouped contigs based on similarities in change was required. The contigs that were generated using Megahit from the three experiments were assembled into clusters. Clustering was achieved by k-means using SciKit-Learn module for Python (Chapter 2.4.3.1). The data is displayed in graphical form in Figure 5.8, where 64 of the 256 clusters are displayed as an example of the data generated. The graphs presented display the normalised change in abundance for each contig (y-axis), against the sample number (x-axis). The number of contigs in each graph is displayed. Samples 1-10 are for experiment 2, 11-21 for experiment 1 and 22-39 for experiment 3.

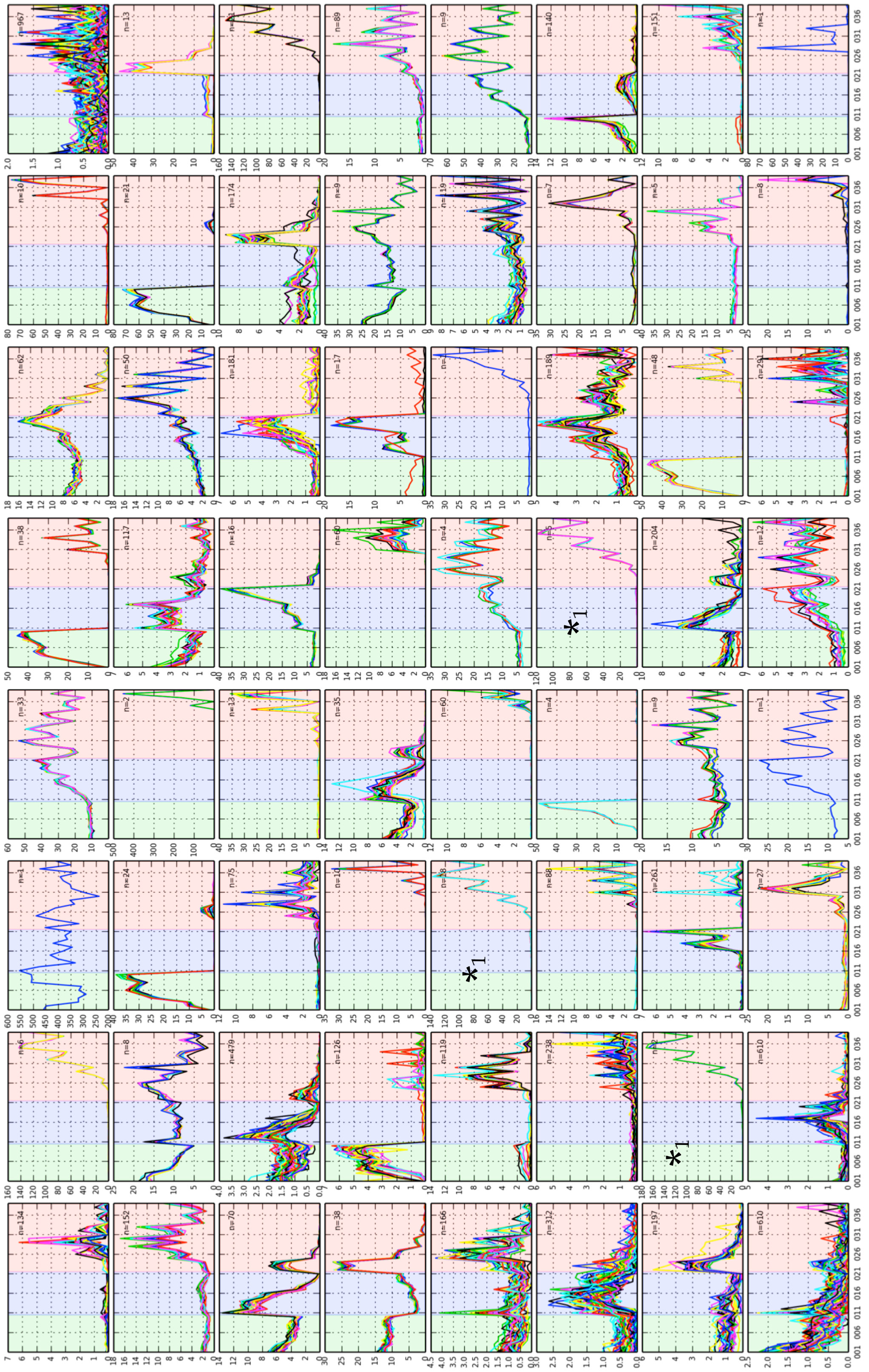


Figure 5.8. The first 64 clusters formed using K-means clustering displaying the relative change of each cluster for each sample. Experiment one (blue), two (green) and three (red) are displayed. n=number of contigs assigned to that cluster.

The K-means clustering from the three experiments is a useful method to display the changes in the contigs throughout each experiment, along with giving a comparison of those potential microbes that are present in one experiment but not others. This method however has the limitation that an arbitrary number of clusters have to be selected, in this instance 256. By selecting the number of groupings for the contigs this essentially forces contigs together that do not necessarily follow the same pattern. Furthermore, some contigs clearly exhibit the same pattern of change throughout the experiments (e.g. marked as *1 on Figure 5.8), but these are plotted in different graphs as the change is not on the same scale. Therefore, a different method for clustering is required, one that does not require an imposed limit on the number of clusters.

5.3.5. Additional Sequencing and Alternative Assembly

5.3.5.1 PacBio sequencer

In addition to the Ilumina HiSeq sequencing, the same eleven samples from experiment one were sent for sequencing using the PacBio platform (Chapter 2.3.5). This sequencing platform provides long read lengths compared to other technologies, and so has the advantage that the short read lengths obtained from such platforms as Illumina can be scaffold onto the longer reads, producing longer contigs. Longer reads also have the potential to close gaps on draft genomes. This is advantageous as it means more detailed information on the organisms function is reported.

The total number of bases obtained from this technology was 135,590,359, and the total number of reads was 48,012. The longest read generated was 27 kb. The read lengths obtained from the technology are smaller than expected, as the majority of reads (68 %) were between 200-1999 bases, although the majority of the base sequences (76 %) were in the 2000-19999 group (Figure 5.9). This could have resulted from using the DNA extraction kit previously mentioned, as there is shearing of the DNA, and so to ensure longer DNA is obtained, a different extraction method might be required.

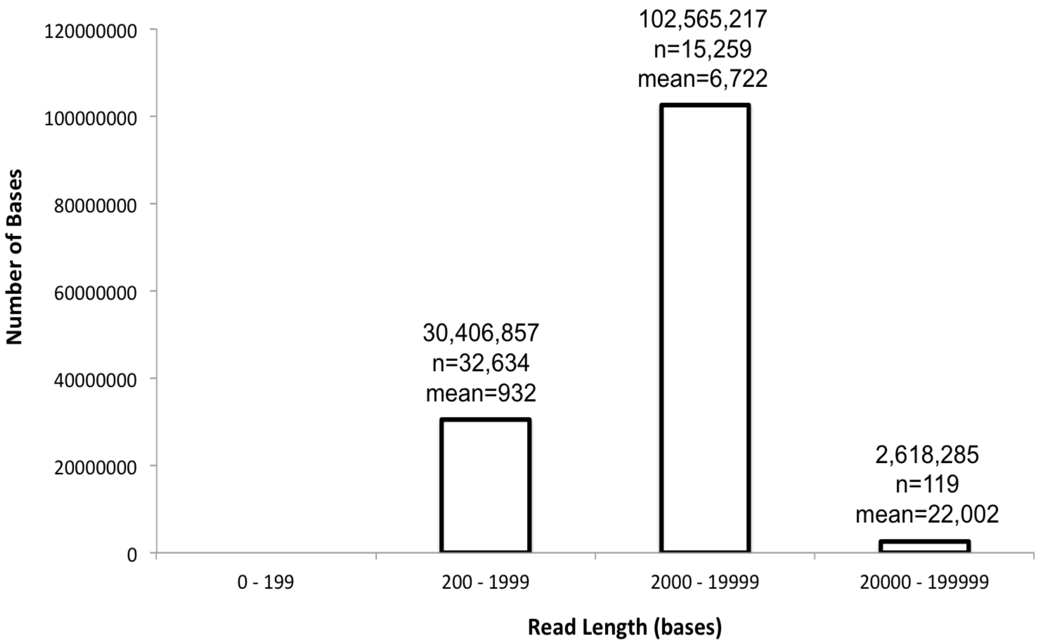


Figure 5.9. The read length distribution of sequences obtained from the PacBio sequencing platform using samples from experiment one. n=number of reads in each size group and mean=average length of each read.

5.3.5.2 Alternative Assembly Strategy

The sequencing data from experiment one that was generated using the Illumina HiSeq and the PacBio platforms were assembled using an alternative method to form contigs and these were then clustered (Chapter 2.4.2). The methodology is displayed in Figure 5.10.

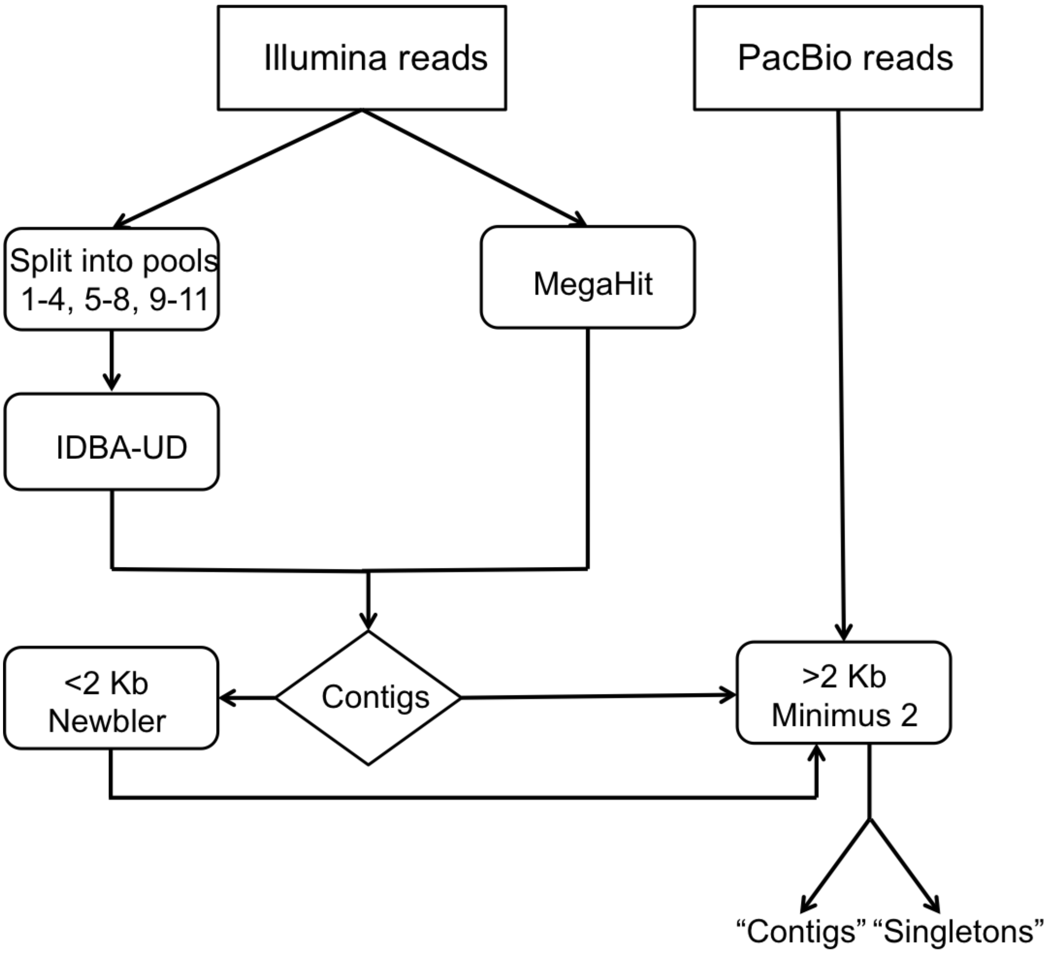
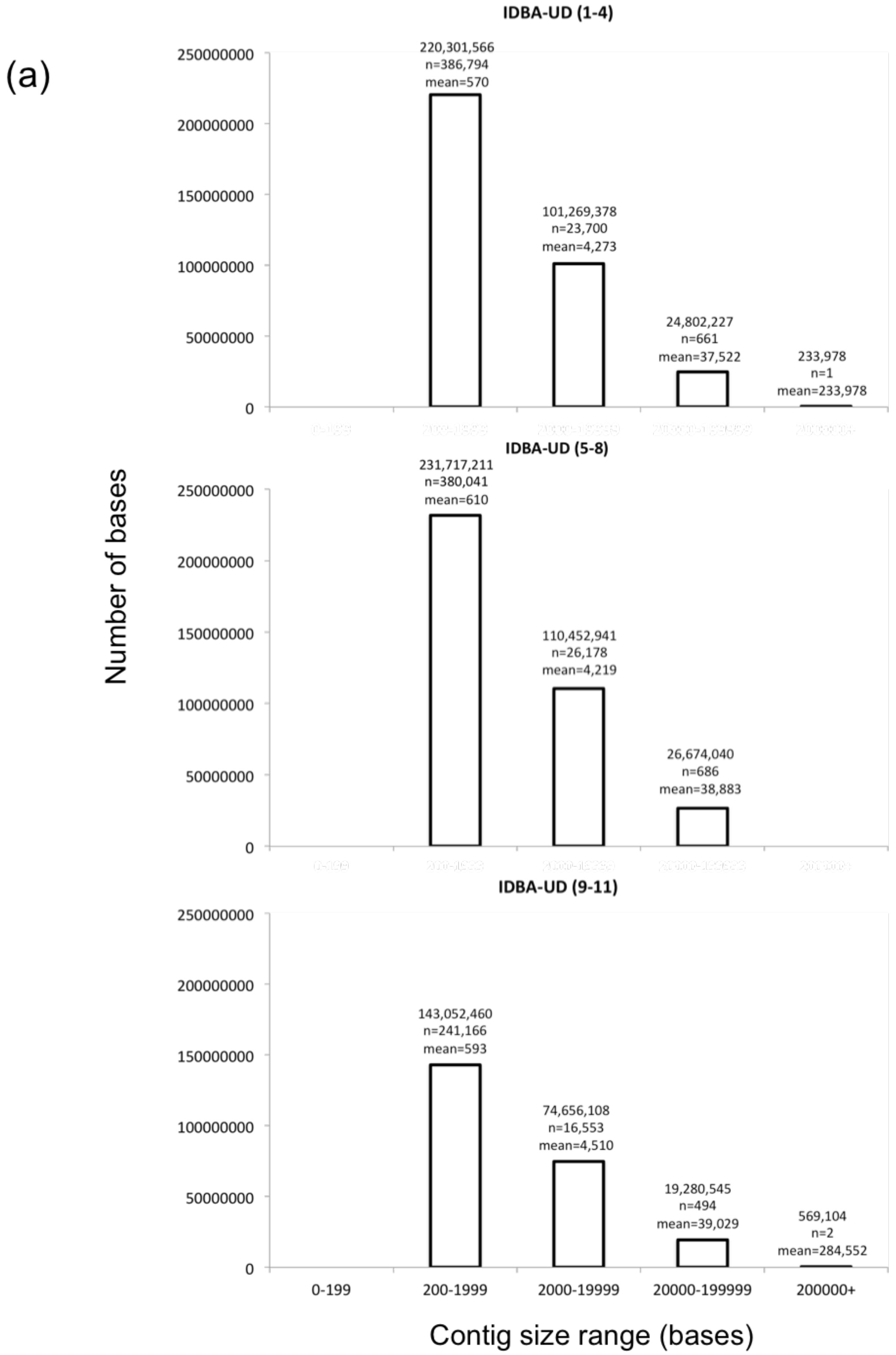


Figure 5.10. The method used for contig building (based on Scholz et al. 2014) for sequencing data from experiment one. The Illumina reads were divided into three pools and assembled using IDBA-UD, whilst in parallel the same Illumina reads were assembled using MegaHit. The resulting contigs were filtered, comblined and assembled using Newbler if smaller than 2 Kb. Those greater than 2 Kb were loaded into Minimus 2, with the PacBio data, to form merged contigs and unincorporated singleton contigs.

The aim of using a variety of assemblers such as IDBA-UD and Megahit for contig formation was because these have slightly different assembly algorithms, based on the de Bruijn graph approach. This has the potential that the longest contigs possible were formed, in addition to the other reassemblies that were carried out to further enhance this. Minimus2 was used as this can merge contigs generated from numerous assemblers. The distribution of the contig sizes using IDBA-UD, Megahit and the final assembly is displayed in Figure 5.11, and the mean contig lengths in Table 5.3.

The selected assembly method for the experiment one sequencing data produced 21,162 contigs, accounting for 237,497,501 bases. The longest contig formed was 398,305 bases, and the minimum was 2,000 bases.


(a)
IDBA-UD (1-4)
250000000
200000000
150000000
100000000
50000000
0
220,301,566
n=386,794
mean=570
101,269,378
n=23,700
mean=4,273
24,802,227
n=661
mean=37,522
233,978
n=1
mean=233,978
IDBA-UD (5-8)
250000000
200000000
150000000
100000000
50000000
0
231,717,211
n=380,041
mean=610
110,452,941
n=26,178
mean=4,219
26,674,040
n=686
mean=38,883
IDBA-UD (9-11)
250000000
200000000
150000000
100000000
50000000
0
143,052,460
n=241,166
mean=593
74,656,108
n=16,553
mean=4,510
19,280,545
n=494
mean=39,029
569,104
n=2
mean=284,552
0-199
200-1999
2000-19999
20000-199999
200000+
Number of bases
Contig size range (bases)

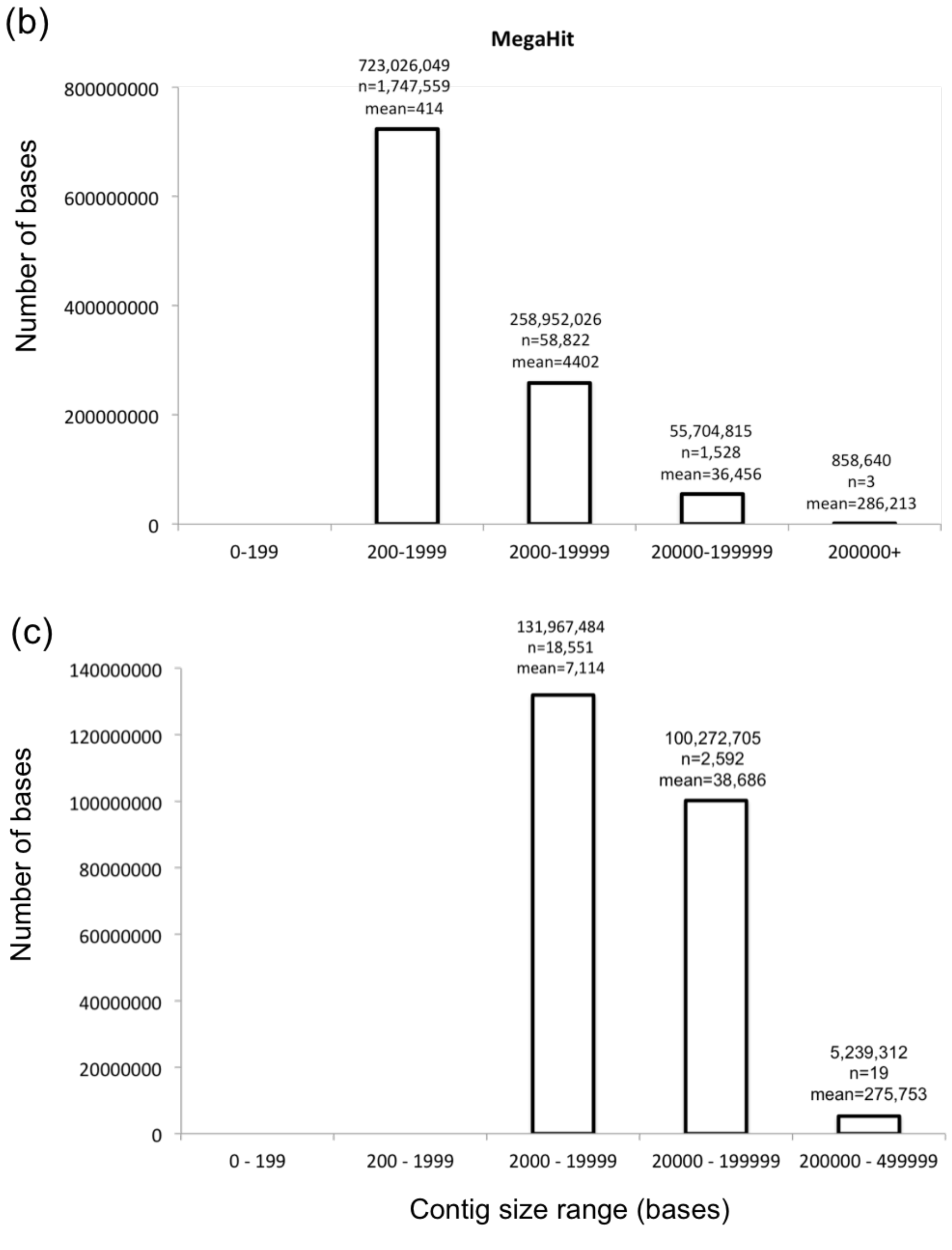


Figure 5.11. Contig size distribution when using different assemblers, (a) IDBA-UD, (b) MegaHit, and the (c) final assembly. n=number of clusters and mean=the average contig length (bases).

	Contig Size (bases)			
Assembly	200-1,999	2,000-19,999	20,000-199,999	200,000+
IDBA-UD (1-4)	570	4,273	37,522	233,978
IDBA-UD (5-8)	610	4,219	38,883	0
IDBA-UD (9-11)	593	4,510	39,029	284,552
Megahit	414	4,402	36,456	286,213
Final Assembly	0	7,114	38,686	275,753

Table 5.3. The mean contig lengths generated using different assemblers, showing that the mean size of the final assembly produces on average longer contigs.

The assembled contigs were then clustered, as described in Chapter 2.4.3.2. This method produced a total of 1,929 clusters, with an average of 11 contigs per cluster.

Figure 5.12 displays the log change in the 50 clusters that had the most sequence coverage, when normalised to the starting sample and Table 5.4 shows the parameters associated with these clusters. There are distinct clusters that follow others closely in a similar fashion, suggesting that these could potentially be one organism, or possibly two or more syntrophic organisms. It is also evident there are some clusters of contigs which are increasing in abundance, others which remain at a relatively similar level, and others which are decreasing. Plotting the data in a graph format alone only shows the dynamic changes and so further analysis of the data set was required.

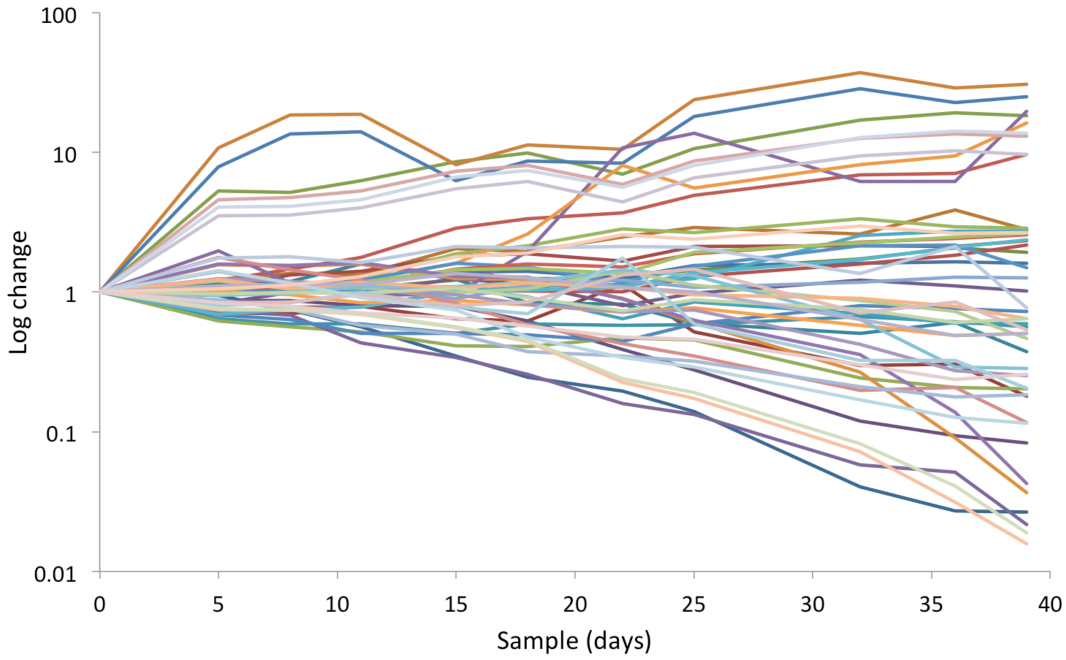


Figure 5.12. The top 50 clusters, when normalised to the starting sample that contain the most sequence data.

Cluster name	Number of contigs	Total DNA content (bases)	Average contig length (bases)	G+C content (%)
272	422	6534294	15484	48.1 ± 3.9
2147	295	4814055	16319	52.7 ± 3.7
2030	73	3340573	45761	62.2 ± 3.5
11549	77	2419280	31419	45.3 ± 1.9
279	149	2225819	14938	47.8 ± 2.9
13	46	2154827	46844	44.6 ± 2.2
19	21	2009106	95672	71.9 ± 2.5
278	45	1668273	37073	47.6 ± 4.6
3513	48	1445705	30119	31.7 ± 3.4
11549	42	1433558	34132	39.9 ± 1.2
5989	132	1416866	10734	54.2 ± 4.2
94	64	1209450	18898	38.8 ± 3.6
3001	6	1174578	195763	59.5 ± 0.7
1610	27	1083290	40122	63.2 ± 2.8

Cluster name	Number of contigs	Total DNA content (bases)	Average contig length (bases)	G+C content (%)
3519	35	1075844	30738	44.7 ± 2.3
92	33	978464	29650	36.6 ± 1.7
438	49	884516	18051	45.6 ± 2.9
7044	28	881686	31489	69.0 ± 1.8
2308	48	881333	18361	63.6 ± 2.2
95	35	871097	24888	63.2 ± 3.1
8526	26	806975	31038	46.9 ± 3.2
11280	33	764984	23181	60.8 ± 7.1
3435	21	659978	31428	51.9 ± 1.4
91	17	639820	37636	45.2 ± 1.0
3861	76	637004	8382	62.7 ± 2.8
274	77	631337	8199	60.7 ± 4.4
10885	40	605383	15135	50.1 ± 3.4
1939	20	568579	28429	64.7 ± 0.8
312	7	516981	73854	54.2 ± 1.7
11547	14	508412	36315	41.1 ± 1.9
2688	24	505721	21072	44.7 ± 1.4
1614	57	490367	8603	62.2 ± 3.8
271	38	483608	12727	42.2 ± 7.29
2460	29	481497	16603	37.9 ± 1.1
97	15	479149	31943	30.2 ± 1.2
5304	35	467644	13361	49.2 ± 4.3
1620	47	453656	9652	63.0 ± 4.1
1636	11	448588	40781	47.5 ± 3.5
3436	10	440982	44098	52.9 ± 2.8
99	19	428005	22527	35.6 ± 1.2
1345	3	391485	130495	55.1 ± 2.7
3023	7	381606	54515	41.6 ± 1.6
2682	10	376067	37607	45.2 ± 1.7
2309	5	362442	72488	41.5 ± 1.2
3003	19	338410	17811	60.4 ± 1.5
7042	14	337372	24098	29.3 ± 0.9

Cluster name	Number of contigs	Total DNA content (bases)	Average contig length (bases)	G+C content (%)
3860	20	329752	16488	61.1 ± 3.7
6873	11	273981	24907	47.6 ± 2.2
2686	25	272809	10912	52.7 ± 4.8
3514	13	270420	20802	42.9 ± 1.5

Table 5.4. The data for the top 50 clusters which have the most coverage, displaying the cluster name, the number of contigs in each cluster, the total length of all the contigs in the cluster, average length of each contig and the average G+C content.

The total number of contigs in the top 50 set is 2,362, out of the total 21,162 from the assembly, that account for 11.2 %. The average number of contigs per cluster for the top set was 49, and these top clusters accounted for 50,913,658 bases, out of the total set of 237,479,501 bases (21.4 %). The average number of bases per contig across the top 50 contigs was 1,060,701 bases. The G+C content of each cluster varies which suggests that each cluster could belong to a different organism. There is generally also a low standard deviation value for each cluster, giving confidence that the clustering method for the contigs works. Interestingly, there are clusters that have low G+C values such as cluster 97 and 7042, with values of 30.2 % and 29.3 % respectively. There are also clusters with high G+C content values such as cluster 19, at 71.9 %.

5.3.5.3 Metabolic activity and markers

The nucleotide sequences of the top 50 clusters were analysed by Prokka (Chapter 2.4.4), where they were firstly translated to amino acids sequences, and then the potential open reading frames (ORFs) were determined.

The results obtained using Prokka were uploaded to KEGG Automatic Annotation Server (Chapter 2.4.5) to determine which genes were present in each cluster, with the aim of obtaining an indication of the function of each putative organism. Genes that were associated with methane metabolism, according to KASS (M00567, M00357, M00356, M00563) in Figure 5.13, were selected as a method for determining if any of the clusters were methanogens (Table 5.5). Examples of genes associated with methane production include formythanofuran dehydrogenase, from the hydrogenotrophic pathway and acetyl-CoA decarboxylase/synthase complex, involved in the acetotrophic pathway (Li et al., 2013). The first 20 clusters are displayed as an example of those enzymes potentially involved in the methane metabolism pathways (so likely to be a gene found in a methanogen), is present in those clusters.

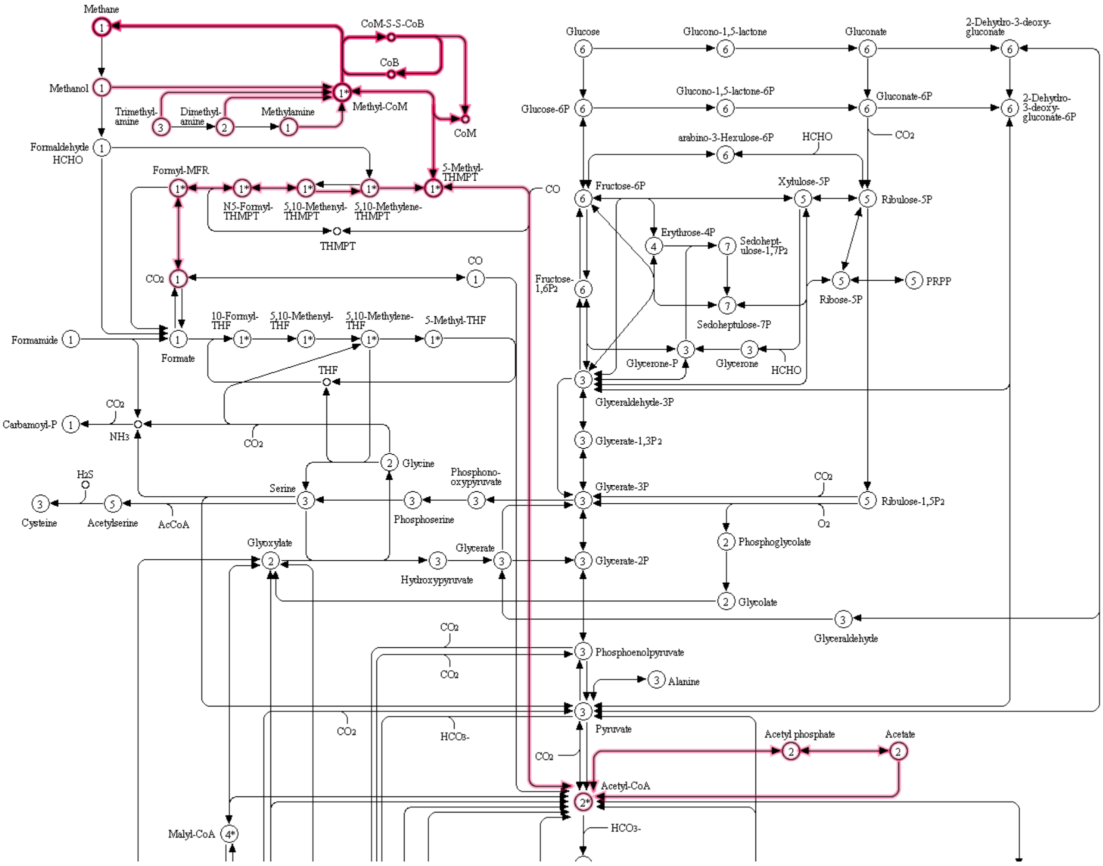


Figure 5.13. The carbon metabolism pathways and those metabolic pathways associated with methanogen metabolism (highlighted in red) according to KASS (taken from www.kegg.jp).

				Enzyme										
Cluster	Contigs	ORF's	Methanogenesis genes	1	2	3	4	5	6	7	8	9	10	11
272	422	8366	16	✓✓✓(a) ✓(b) ✓✓✓(c)										
2147	295	5532	9											
279	149	3041	5											
5989	132	1421	10		✓ (e)									
3861	75	787	16	✓(b) ✓(c)	✓(a) ✓(b) ✓(c)	✓✓		✓	✓	✓	✓			
274	77	789	14		✓✓(a) ✓(c) ✓(e)		✓✓(α) ✓(β) ✓(γ)							
2030	73	2538	9									✓		
94	65	1297	4											
1614	57	577	16		✓✓(e)		✓(β)							
3513	52	1195	4											
438	49	764	5											
13	46	1779	8	✓✓✓(a) ✓(c)										
1620	47	490	10	✓✓✓(a)	✓(b) ✓(d)									

				Enzyme										
Cluster	Contigs	ORF's	Methanogenesis genes	1	2	3	4	5	6	7	8	9	10	11
2308	48	842	4											
278	45	1422	9											
11549	43	1114	6											
3519	35	886	10											
10885	39	784	2											
271	38	653	1											
95	35	734	10	✓✓(a) ✓(b) ✓(c)										

Table 5.5. Information according to the KASS assignment – the cluster number that was uploaded, the number of contigs in that cluster, the number of ORF's assigned by Prokka in that cluster, the number of methanogen genes identified by KASS, along with the specific genes (1) Heterodisulphide reductase (subunits a, b, c), (2) Formylmethanofuran dehydrogenase (subunits a, b, c, d, e) (3)Ferredoxin, (4) Methyl – coenzyme M reductase (subunits alpha, beta, gamma), (5) Formylmethanofuran – tetrahydromethanopterin N – formyltransferase (6) Methanyltetrahydromethanopterin cyclohydrolase, (7) Methanyltetrahydromethanopterin dehydrogenase (8) Methanyltetrahydromethanopterin reductase (9) Trimethylamine – corrinoid protein co-methyltransferase, (10) Tetrahydromethanopterin S-methyltransferase (subunits a, b, c, d, e, f, g, h), (11) Acetyl-CoA decarbonylase/synthetase complex (subunits alpha, beta, gamma).

The KASS assignments, as shown in Table 5.5, offer an indication if a methanogen is present for each cluster, but this is not enough detail alone. This method also demonstrated that genes assigned to the methane metabolism pathway are not all strictly found in methanogens, and so this data has to be treated with caution. For example, Heterodisulphide reductase (HDR) is an enzyme that is found in methanogens, being a key enzyme in both the aceticlastic and hydrogenotrophic pathways. HDR-like proteins are also found in other archaea as well as some bacteria (Refai et al., 2014). Therefore detection of this enzyme does not directly suggest if the organism is a methanogen. It was noted that those clusters found to contain the highest number of methane metabolism associated genes did actually represent methanogens, as discussed below. It was concluded that using KASS in the mentioned way (for metagenomic studies) is perhaps not the best method for data interpretation. Therefore using this program would be more appropriate for investigating complete draft genomes or metagenomic datasets that have more sequence, to determine exactly what enzymes are present in the organism, concluding what functions they are potentially carrying out.

To further investigate this, the protein sequences of a selected group of clusters (274, 1620, 1614 and 3861), that had genes assigned to known methanogen metabolism pathways by KASS were BLAST (Altschul et al., 1997) searched to identify the closest database match.

The results from the BLAST search are suggestive that the four clusters represent methanogens as the majority give the top result of *Methanoculleus marisnigri* JR1. Further investigation into the BLAST results for these clusters indicate that actually the top search result is not the organism present, but a close relation to. For example, cluster 3861, where 15 genes were BLAST searched, 93 % returned a result of *Methanoculleus marisnigri* JR1, with 7 % being an Uncultured archaeon. The mean identity value for the *Methanoculleus* hits was 85 %, suggesting that the searched sequences are not that particular strain, but are likely to be from that genus. Similarly, cluster 1620 indicates comparable results, where 85 % of top search results were *Methanoculleus marisnigri* JR1, and of these the mean identity value was 91 %. Furthermore, as none of the searches resulted in identical matches, it indicates that the clusters could actually be a previously undescribed organism.

Interpretation of the results in this fashion starts to reveal interesting information regarding the clusters. This method however is time consuming and the number of genes searched is limited, especially as the number of ORFs in each cluster is high, such as 8,366 in cluster 272. Therefore an automated method for performing BLAST searches on all the clusters was required.

5.3.5.4 Automated BLAST searches

The writing of a custom script that allowed for a large number of BLAST searches to be conducted enables the interpretation of large datasets, as the top closest match is displayed (Chapter 2.4.6.2).

The clusters that were BLAST searched were those that appeared to display interesting changes throughout the trial. In this instance, those showing an increase throughout the experiment were investigated (Figure 5.14). Distinct groups of clusters appeared to follow similar patterns of change, and so these were divided up into five groups: NC1, NC2, NC3, NC4 and NC5. Table 5.6 displays the statistics for each cluster, the BLAST result and the group it belongs to.

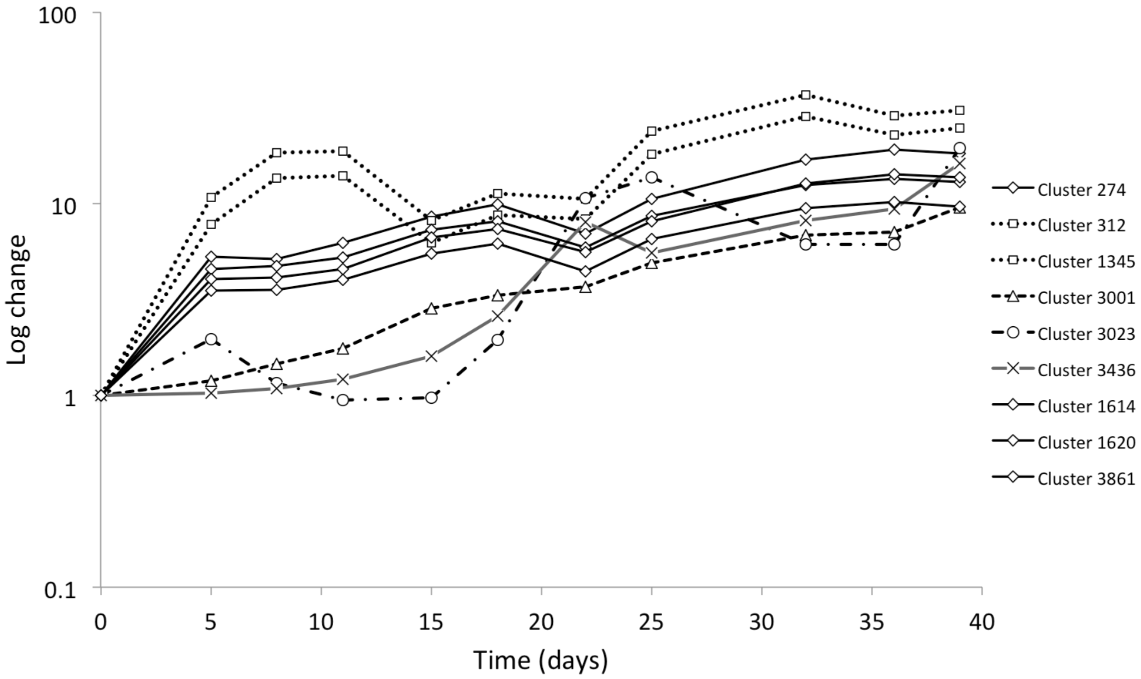


Figure 5.14. The log change of nine clusters throughout experiment one. Clusters that have a similar behaviour have been grouped together: NC1 (Cluster 312, 1345), NC2 (Cluster 274, 1614, 1620, 3861), NC3 (Cluster 3001), NC4 (Cluster 3023) and NC5 (Cluster 3436).

Group	Cluster	Number of contigs	Mean contig size (inc min and max)	Mean G+C content	Number of matches	Top matches
NC1	312	7	73854 (4257 & 188006)	54.2 ± 1.7	551	184 (33.4 %) *Candidatus Methanomethylophilus alvus* Mx1201 106 (19.2 %) *Thermoplasmatales archaeon* BRNA1
NC1	1345	3	48228 (17087 & 100112)	55.1 ± 2.7	368	109 (29.6 %) *Candidatus Methanomethylophilus alvus* Mx1201 64 (17.4 %) *Thermoplasmatales archaeon* BRNA1
NC2	274	77	8204 (2060 & 26463)	60.7 ± 4.4	713	193 (27.1 %) *Methanoculleus marisnigri* JR1 150 (21%) *Methanoculleus bourgensis* MS2
NC2	1620	47	9652 (2304 & 22135)	63 ± 4.1	454	126 (27.8 %) *Methanoculleus marisnigri* JR1 98 (21.6 %) *Methanoculleus bourgensis* MS2
NC2	1614	57	8635 (2367 & 23419)	62.2 ± 3.8	528	148 (28 %) *Methanoculleus marisnigri* JR1 148 (28 %) *Methanoculleus bourgensis* MS2
NC2	3861	75	8493 (2194 & 36858)	62.7 ± 2.8	736	217 (29.5 %) *Methanoculleus marisnigri* JR1 209 (28.4 %) *Methanoculleus bourgensis* MS2
NC3	3001	6	195763 (35709 & 398305)	59.5 ± 0.7	1144	112 (9.8 %) *Candidatus Methanomethylophilus alvus* Mx1201
NC4	3023	7	49439 (28470 & 106615)	41.6 ± 1.8	312	101 (32.4 %) *Methanosarcina mazei* Go1 82 (26.3 %) *Methanosarcina mazei* Tuc01
NC5	3436	10	54386 (20808 & 101546)	53.1 ± 1.5	404	158 (39.1 %) *Methanosaeta concilii* GP6

Table 5.6. The five groups and associated clusters that appear to show an increase throughout experiment one. The number of clusters in each contig is displayed, the average contig size, G+C content, the total number (and %) of BLAST matches using the custom script, and the closest known organism that had the highest number of matches

The results from the BLAST search are clearly indicative that those microbes increasing in number are methanogens. Group NC1, which included the clusters 312 and 1345 both gave comparable information. For example, both clusters had long contigs, with the mean contig size being 73,854 bp for cluster 312 and 48,228 bp for cluster 1345. Additionally the mean G+C content of the clusters were similar, 54.2 ± 1.7 and 55.1 ± 2.7, indicating they both could belong to the same organism. The G+C content is the measure of the guanine – cytosine content in the DNA sample. Each organism has a different G+C content and therefore could be reflective of how closely related organisms are, but is not a direct measure, just suggestive. *Candidatus Methanomethylophilus alvus* Mx1201 accounted for 33.4 and 29.6 % of the top BLAST matches for clusters 312 and 1345 respectively and *Thermoplasmatales archaeon* BRNA1 accounted for 19.2 and 17.4 %. This strongly suggests that this group is a methanogen. The genome of *Candidatus Methanomethylophilus alvus* Mx1201 has recently been described by Borrel et al. (2012), isolated from the human gut. This organism was enriched using methanol as a substrate and the genome contained genes that encoded for methylotrophic methanogenesis. It is therefore suggestive that as *Candidatus Methanomethylophilus alvus* Mx1201 is the top result from the search, the organism in this cluster could be a methanol utiliser. It was suggested that as a result of the processing methods at the biodiesel facility, there would be methanol present within the feedstock. It is therefore expected that methanogens that can utilise methanol would be identified in the samples.

Group NC2 included the clusters 274, 1620, 1614 and 3861. In comparison to NC1, the mean contig length was smaller: 26,463, 22,135, 23,419 and 36,858 bases for clusters 274, 1620, 1614 and 3861 respectively. The mean G+C values were comparable at 60.7, 63, 62.2 and 62.7, again consistent with the notion they belong to one organism. The top BLAST result for the ORFs in each of these clusters all gave the same organism as *Methanoculleus marisnigri* JR1 at 27.1, 27.8, 28 and 29.5 % for clusters 274, 1620, 1614 and 3861. *Methanoculleus bourgensis* MS2 was the second top hit for the four clusters, showing that the genus *Methanoculleus* account for a significant proportion of the top results. Although the top BLAST result was *Methanoculleus*, it is likely that the organism belongs to this genus, but is not in the database. *Methanoculleus marisnigri* JR1 is a previously described organism that was noted to metabolise H_2/CO_2 and formate to methane. It was also reported that the G+C content was 62.1 % (Anderson et al., 2009), which is consistent with the average G+C content of the four clusters, at 62.1 %. This is highly suggestive that the organism in question is highly related to *Methanoculleus marisnigri*.

Group NC3 had large contigs with a mean size of 195,763 bases. The mean G+C content was 59.5 ± 0.7 and the top BLAST result was *Candidatus Methanomethylophilus alvus* Mx1201, but only at 9.8 % of returned top matches. This indicates that the organism in question is not closely related to the top match.

NC4 contigs had an average of 49,439 bases and mean G+C of 41.6 ± 1.8. *Methanosarcina mazei* Go1 was the top match, at 32.4 %, closely followed by *Methanosarcina mazei* Tuc01 at 26.3 %. This is suggestive that the organism in question is highly related, or even a different strain of *Methanosarcina mazei*. *Methanosarcina mazei* species are capable of metabolising a wide range of compounds including acetate, methanol and methylamines to methane (Jäger et al., 2009).

Group NC5 had a mean contig length of 54,386 bp and an average G+C content of 53.1 %. The top BLAST match was *Methanosaeta concilii* GP6 at 39.1 %. This organism is a known slow growing aceticlastic methanogen and the circular chromosome has a G+C content of 51.03 % (Barber et al., 2011), which is a similar value to that of the group. The presence of both *Methanosarcina* and *Methanosaeata* is expected as both these organisms can utilise acetic acid and during this experiment the acetic acid levels were high.

5.3.5.5 Phylogeny

Although the BLAST searches of the ORFs give an indication if a cluster closely matches a known microbe, it only presents the top match, which in most cases is not an exact match. Furthermore, the G+C content of the clusters also provides an indication of the number of organisms present in each cluster, and can also be suggestive if the organism is closely related to the top match. Therefore it is possible to estimate the number of organisms present in the cluster, and further demonstrates that the chosen clustering method works, but this data only shows the closest known organism, not how closely related they are.

Determining exactly where the organisms fit in a phylogenetic tree is essential to understanding what the organisms are, and also suggesting what the potential functions could be. Phylogenetic trees were drawn as detailed in Chapter 2.4.7. Figure 5.15 (a) displays a methanogen phylogenetic tree for groups NC1 and NC2 and Figure 5.15 (b) for groups NC3, NC4 and NC5.

When the ORFs of group NC1 were BLAST searched, the top result was *Candidatus Methanomethylophilus alvus*, but based on the phylogenetic tree, *Candidatus Methanoplasma termitum* is a close known organism. This organism belongs to the order *Methanomassiliicoccales*, and reduces methanol using hydrogen, to produce methane (Lang et al., 2015). *Methanoculleus marisnigri* JR1 was the top result for group NC2, and the phylogenetic tree suggests that the closest relation to group NC2 is this organism, and so almost certainly belongs to the *Methanoculleus* genus.

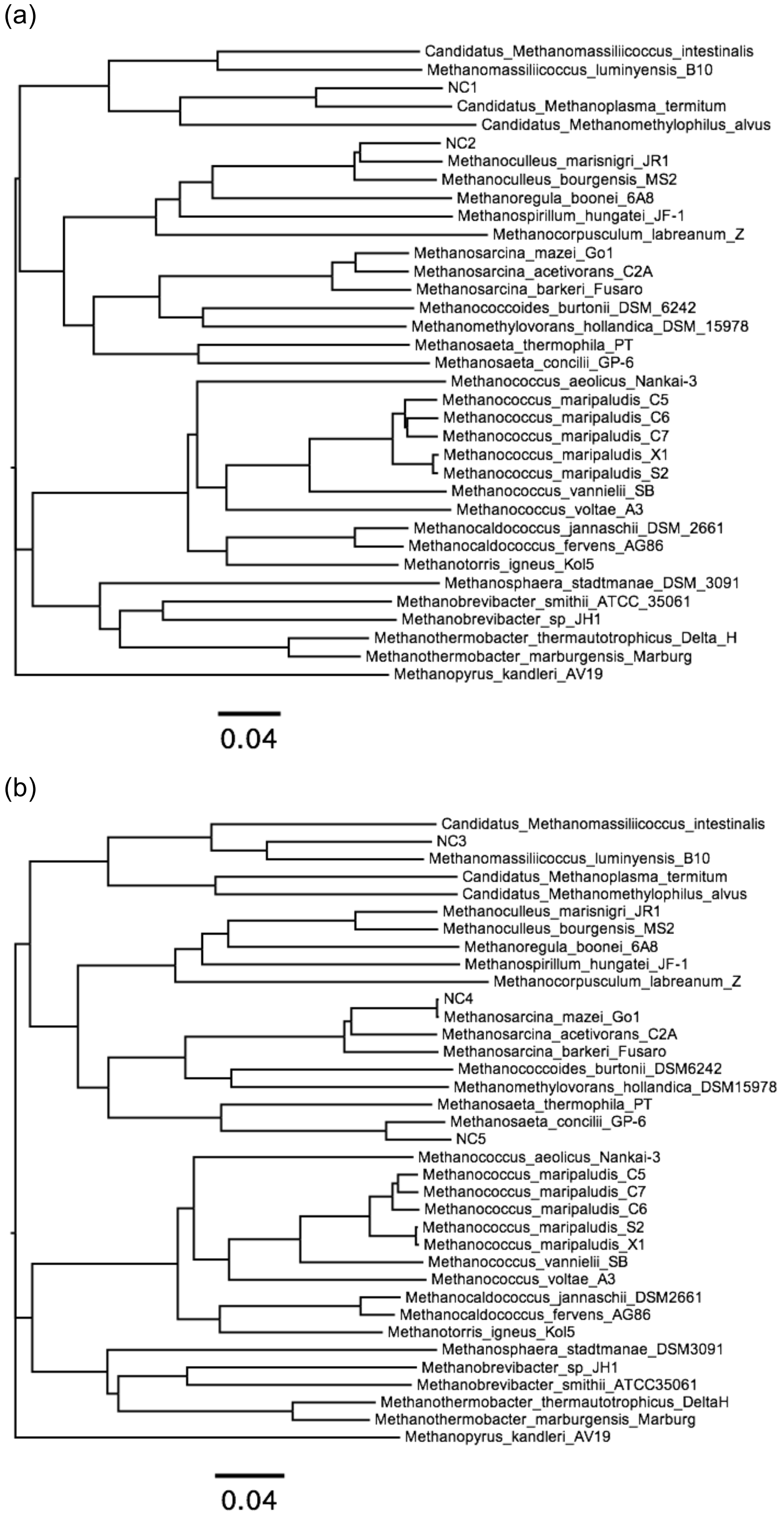
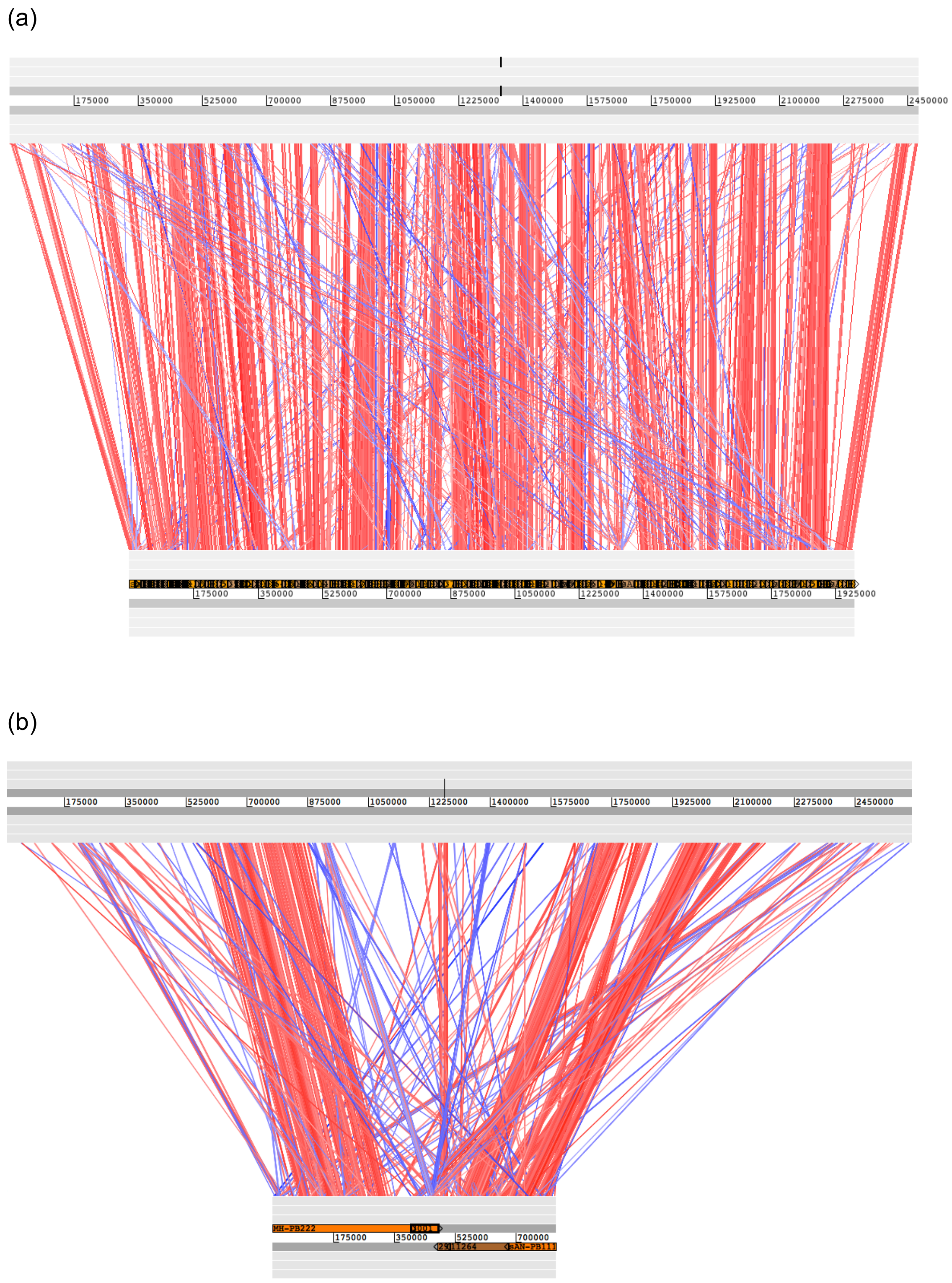

(a)
Candidatus_Methanomassiliicoccus_intestinalis
Methanomassiliicoccus_luminyensis_B10
NC1
Candidatus_Methanoplasma_termitum
Candidatus_Methanomethylophilus_alvus
NC2
Methanoculleus_marisnigri_JR1
Methanoculleus_bourgensis_MS2
Methanoregula_boonei_6A8
Methanospirillum_hungatei_JF-1
Methanocorpusculum_labreanum_Z
Methanosarcina_mazei_Go1
Methanosarcina_acetivorans_C2A
Methanosarcina_barkeri_Fusaro
Methanococcoides_burtonii_DSM_6242
Methanomethylovorans_hollandica_DSM_15978
Methanosaeta_thermophila_PT
Methanosaeta_concilii_GP-6
Methanococcus_aeolicus_Nankai-3
Methanococcus_maripaludis_C5
Methanococcus_maripaludis_C6
Methanococcus_maripaludis_C7
Methanococcus_maripaludis_X1
Methanococcus_maripaludis_S2
Methanococcus_vannielii_SB
Methanococcus_voltae_A3
Methanocaldococcus_jannaschii_DSM_2661
Methanocaldococcus_fervens_AG86
Methanotorris_igneus_Kol5
Methanosphaera_stadtmanae_DSM_3091
Methanobrevibacter_smithii_ATCC_35061
Methanobrevibacter_sp_JH1
Methanothermobacter_thermautotrophicus_Delta_H
Methanothermobacter_marburgensis_Marburg
Methanopyrus_kandleri_AV19
0.04
(b)
Candidatus_Methanomassiliicoccus_intestinalis
NC3
Methanomassiliicoccus_luminyensis_B10
Candidatus_Methanoplasma_termitum
Candidatus_Methanomethylophilus_alvus
Methanoculleus_marisnigri_JR1
Methanoculleus_bourgensis_MS2
Methanoregula_boonei_6A8
Methanospirillum_hungatei_JF-1
Methanocorpusculum_labreanum_Z
NC4
Methanosarcina_mazei_Go1
Methanosarcina_acetivorans_C2A
Methanosarcina_barkeri_Fusaro
Methanococcoides_burtonii_DSM6242
Methanomethylovorans_hollandica_DSM15978
Methanosaeta_thermophila_PT
Methanosaeta_concilii_GP-6
NC5
Methanococcus_aeolicus_Nankai-3
Methanococcus_maripaludis_C5
Methanococcus_maripaludis_C7
Methanococcus_maripaludis_C6
Methanococcus_maripaludis_S2
Methanococcus_maripaludis_X1
Methanococcus_vannielii_SB
Methanococcus_voltae_A3
Methanocaldococcus_jannaschii_DSM2661
Methanocaldococcus_fervens_AG86
Methanotorris_igneus_Kol5
Methanosphaera_stadtmanae_DSM3091
Methanobrevibacter_sp_JH1
Methanobrevibacter_smithii_ATCC35061
Methanothermobacter_thermautotrophicus_DeltaH
Methanothermobacter_marburgensis_Marburg
Methanopyrus_kandleri_AV19
0.04

Figure 5.15. Phylogenetic tree of methanogens for groups NC1 and NC2 (a) and groups NC3, NC4 and NC5 (b), drawn using FigTree.

Group NC3 had no distinct matches when BLAST searching. In fact, the top result was *Candidatus Methanomethylophilus alvus* at under 10 %. According the placement on the phylogenetic tree, group NC3 is distant from *Candidatus Methanomethylophilus alvus*, and actually the closest known organism is *Methanomassiliicoccus luminyensis*. A recent description of *Methanomassiliicoccus luminyensis* showed that this organism uses hydrogen as an electron donor to reduce methanol and the circular chromosome has a G+C content of 60.5 (Gorlas et al., 2012), which is consistent with the G+C content of the cluster (59.5 %). NC4 is very closely related to *Methanosarcina mazei* Go1 and this data is confirmed from the BLAST data where the majority of top results were from *Methanosarcina mazei* species. *Methanosarcina* species are found in numerous environments and have an average G+C content of 42.5 %, which again is highly comparable to that of group NC4 (41.6 %). Group NC5 is related to *Methanosaeta concilii* GP6, which again agrees with the top BLAST search results.

5.3.5.6 Genome mapping

To further support the phylogenetic tree data, the sequencing data from individual groups can be mapped onto the genomes to which they are most closely related. This gives a clear indication of how much of the sequence maps onto the known genome, but also where the variations in sequence are. The gaps support that the group in question is not the organism that it is most closely related to, according to the phylogenetic tree. ACT (Carver et al., 2005) is a program that can be used to display assembled sequences against a completed genome, allowing for comparison (Chapter 2.4.8). Figure 5.16 displays the contigs from groups NC2, NC3 and NC4, when individually mapped against known genomes.



(a)
175000
350000
525000
700000
875000
1050000
1225000
1400000
1575000
1750000
1925000
2100000
2275000
2450000
175000
350000
525000
700000
875000
1050000
1225000
1400000
1575000
1750000
1925000
(b)
175000
350000
525000
700000
875000
1050000
1225000
1400000
1575000
1750000
1925000
2100000
2275000
2450000
MH-PB222
0001
175000
350000
525000
700000
2911264
sAN-PB111

(c)

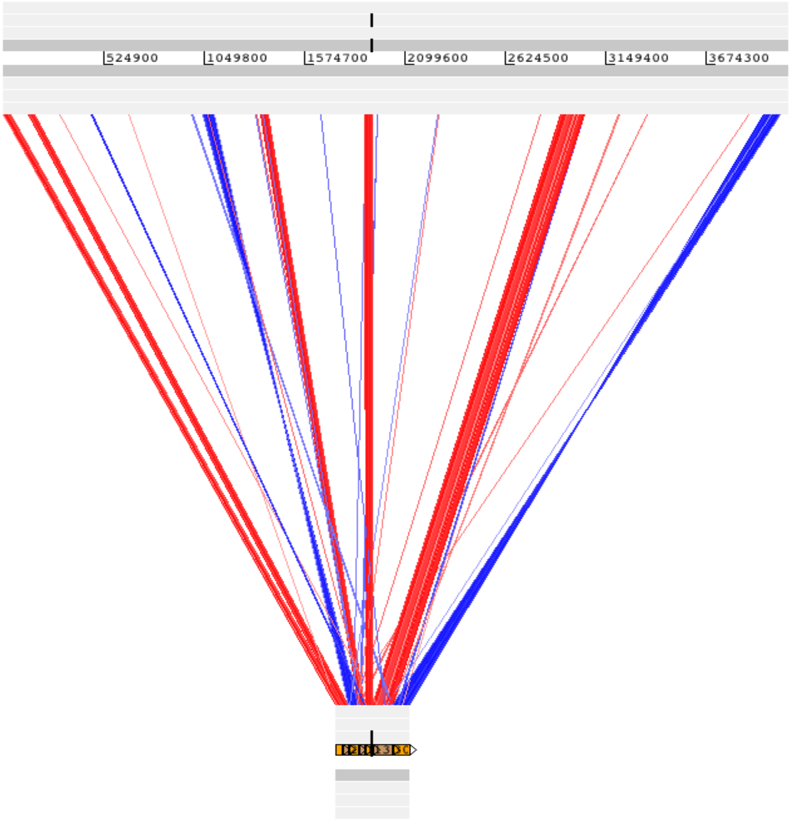


Figure 5.16. Group NC2 compared to the (a) *M.marisnigri* genome (CP000562.1), Group NC3 clusters mapped against the (b) *M.luminyensis* genome (CAJE01000001.1) and Group NC4 clusters compared to the complete genome of (c) *M.mazei* (NC_003901.1). Images generated using ACT. The red bands represent the forward matches and the blue represent the reverse matches.

Group NC2 sequence maps closely to the genome of the known organism. The gaps in NC2 sequence suggests that this cluster is not *Methanoculleus marisnigri*, but is closely related, and with so many matching regions of sequence, this is suggestive of a similar function, which backs up the data from the phylogenetic tree.

Group NC3 maps well to certain regions of the *M.luminyensis*, further suggesting, as in the phylogenetic tree, that it is related to this organism.

When group NC4 was aligned to the reference genome of *M.mazei*, it is evident that the group encoding for an organism is closely related to *M.mazei*, which further supports the phylogenetic tree. Although the group total contig size is small in comparison to the reference genome, the fact that the sequences are matching across the whole reference genome further suggests that the organism in question is highly related to *M.mazei*, but either further analysis of the metagenomic data or additional sequencing of the samples would be required.

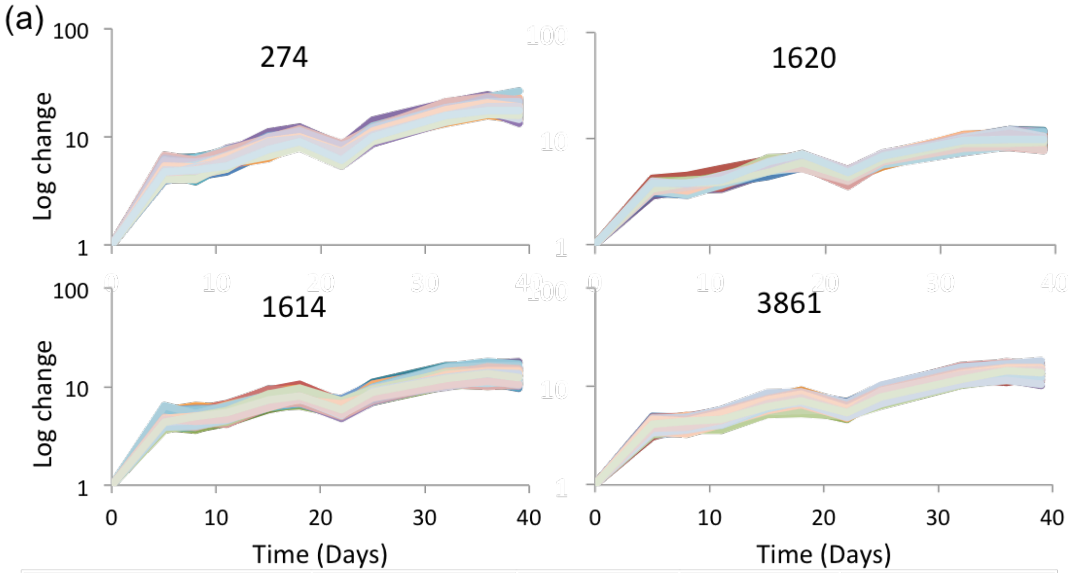
5.4 Conclusions

The advancements in sequencing technology through NGS has revolutionised our view and understanding of even the most complex habitats. The large datasets generated through this technology give a wealth of information. The challenge though is how this information can be interpreted and turned into something useful. There are a variety of tools available to researchers to begin to interpret these large datasets, but as of yet, there is not a 'complete package', in terms of just placing the sequencing data in and obtaining the results. Instead there are a variety of tools and methods that can be used or adapted, allowing for subjective input and interpretation. Therefore this makes handling large datasets very challenging. None the less, we have complemented the use of both pre-existing programs as well as custom developed scripts to obtain useful information regarding the microbes involved in anaerobic digestion and how they change throughout the digestion process. The use of these methods has also suggested that there could be previously undescribed novel organisms within the systems.

Although there was a limited investigation into the dataset for experiment one, as the focus was directed towards those clusters that appeared to increase in abundance throughout the experiment, the ultimate question is how well do we think the method used to study the microbial dynamics has worked. To address this, numerous checks have been carried out whilst using and developing tools, giving the confidence that the method used is suitable. This is because:

- Increased use of assemblers gave a higher proportion of longer contigs
- The clustering of contigs has been shown to work – that contigs with the same pattern of change are clustered together
- BLAST searches of each cluster reveal that a significant proportion of top matches are from the same organism
- The G+C content of each cluster is consistent with low levels of variation
- Large proportions of the assembled contigs (in clusters) map very well onto the closest known organisms

These points are justified in Figure 5.17 and 5.18, where the contigs for each cluster is displayed (a), along with the cluster data in comparison to a known organism (b), where the organism is placed on a phylogenetic tree (c) and finally how well it maps to known a genome (d). The ultimate aim would be to develop a pipeline that performs all the above mentioned tasks. In theory this would be that numerous contig assemblies would take place, along with cluster formation. From this, the clusters would be automatically searched and placed on a phylogenetic tree, along with been compared to a known genome. If the majority of these processes were to be automated, this would allow for large datasets to be interpreted.



(b)

Cluster	Number of contigs	Mean length (bp)	G+C content (%)	Top match
274	77	8199	60.7	*M.marisnigri*
1620	47	9652	63	*M.marisnigri*
1614	57	8603	62.2	*M.marisnigri*
3861	76	8382	62.7	*M.marisnigri*
Reference genome (*M.marisnigri*)			62.1	

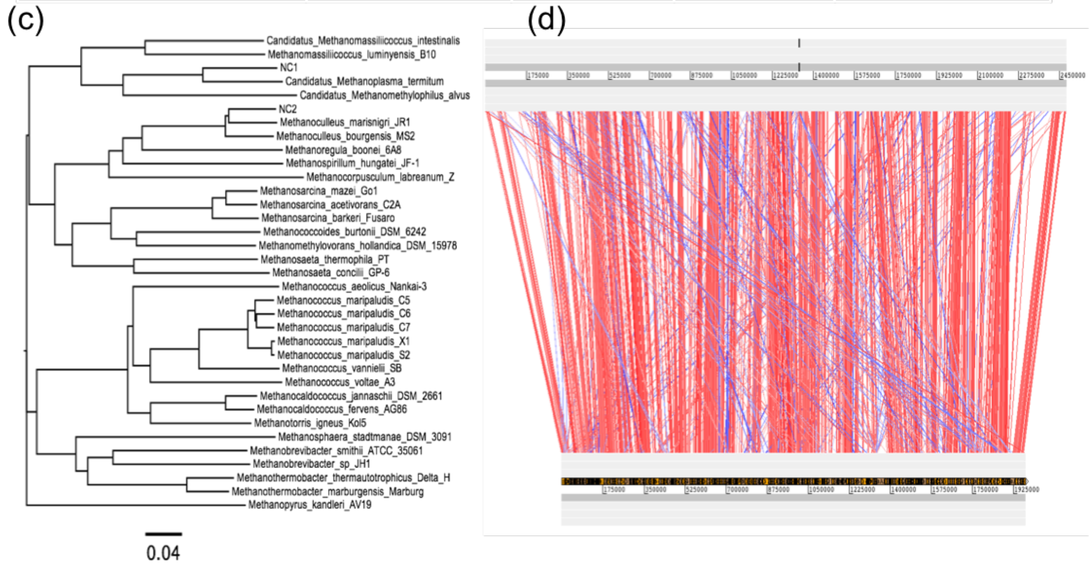
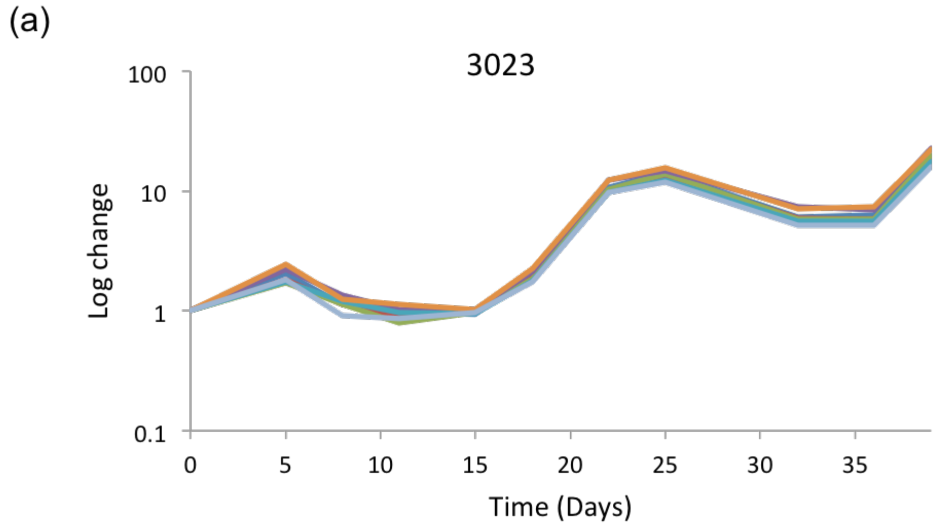

(c)
(d)
Candidatus_Methanomassiliicoccus_intestinalis
Methanomassiliicoccus_luminyensis_B10
NC1
Candidatus_Methanoplasma_termitum
Candidatus_Methanomethylophilus_alvus
NC2
Methanoculleus_marisnigri_JR1
Methanoculleus_bourgensis_MS2
Methanoregula_boonei_6A8
Methanospirillum_hungatei_JF-1
Methanocorpusculum_labreanum_Z
Methanosarcina_mazei_Go1
Methanosarcina_acetivorans_C2A
Methanosarcina_barkeri_Fusaro
Methanococcoides_burtonii_DSM_6242
Methanomethylovorans_hollandica_DSM_15978
Methanosaeta_thermophila_PT
Methanosaeta_concilii_GP-6
Methanococcus_aeolicus_Nankai-3
Methanococcus_maripaludis_C5
Methanococcus_maripaludis_C6
Methanococcus_maripaludis_C7
Methanococcus_maripaludis_X1
Methanococcus_maripaludis_S2
Methanococcus_vannielii_SB
Methanococcus_voltae_A3
Methanocaldococcus_jannaschii_DSM_2661
Methanocaldococcus_fervens_AG86
Methanotorris_igneus_Kol5
Methanosphaera_stadtmanae_DSM_3091
Methanobrevibacter_smithii_ATCC_35061
Methanobrevibacter_sp_JH1
Methanothermobacter_thermautotrophicus_Delta_H
Methanothermobacter_marburgensis_Marburg
Methanopyrus_kandleri_AV19
0.04

Figure 5.17. Summary of the NC2 group displaying the contig pattern of change throughout the experiment in each cluster (a), the cluster, G+C content and top BLAST match against the closest known organism (b), the placement of the cluster in a phylogenetic tree (c) and all the contigs from the cluster lined up against the complete genome (d).



(b)

Cluster	Number of contigs	Mean length (bp)	G+C content	Top match
3023	7	54515	41.6	*M.Mazei* Go1
Reference genome (*M.mazei* Go1)			41.5	

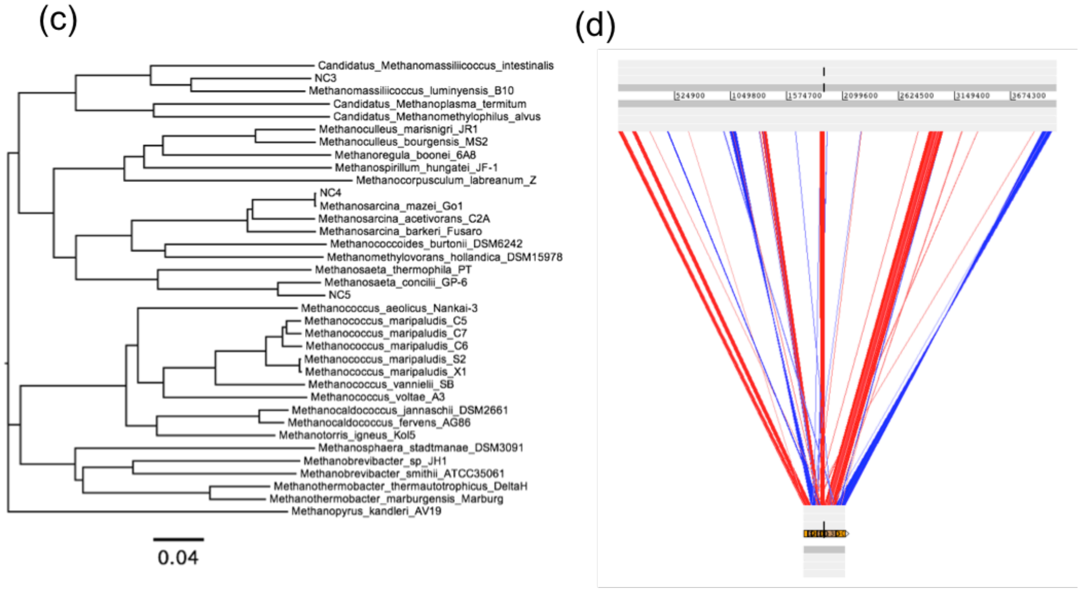

(c)
Candidatus_Methanomassiliicoccus_intestinalis
NC3
Methanomassiliicoccus_luminyensis_B10
Candidatus_Methanoplasma_termitum
Candidatus_Methanomethylophilus_alvus
Methanoculleus_marisnigri_JR1
Methanoculleus_bourgensis_MS2
Methanoregula_boonei_6A8
Methanospirillum_hungatei_JF-1
Methanocorpusculum_labreanum_Z
NC4
Methanosarcina_mazei_Go1
Methanosarcina_acetivorans_C2A
Methanosarcina_barkeri_Fusaro
Methanococcoides_burtonii_DSM6242
Methanomethylovorans_hollandica_DSM15978
Methanosaeta_thermophila_PT
Methanosaeta_concilii_GP-6
NC5
Methanococcus_aeolicus_Nankai-3
Methanococcus_maripaludis_C5
Methanococcus_maripaludis_C7
Methanococcus_maripaludis_C6
Methanococcus_maripaludis_S2
Methanococcus_maripaludis_X1
Methanococcus_vannielii_SB
Methanococcus_voltae_A3
Methanocaldococcus_jannaschii_DSM2661
Methanocaldococcus_fervens_AG86
Methanotorris_igneus_Kol5
Methanosphaera_stadtmanae_DSM3091
Methanobrevibacter_sp_JH1
Methanobrevibacter_smithii_ATCC35061
Methanothermobacter_thermautotrophicus_DeltaH
Methanothermobacter_marburgensis_Marburg
Methanopyrus_kandleri_AV19
0.04
(d)
524900
1049800
1574700
2099600
2624500
3149400
3674300

Figure 5.18. Summary of the NC4 group displaying the contig pattern of change throughout the experiment in each cluster (a), the cluster, G+C content and top BLAST match against the closest known organism (b), the placement of the cluster in a phylogenetic tree (c) and all the contigs from the cluster lined up against the complete genome (d).

The methods developed demonstrate that the metagenomic approach provides more detail than that obtained by targeted sequencing. The data from experiment one has also shown that novel organisms have been sequenced. Out of the five groups that have been investigated, it appears that five are not described in the literature. To obtain more detail on these, further information on the genome is required. Draft genomes that are produced by contig assembly are usually independent contigs, where the positions along a sequenced genome are unknown (Lu et al., 2014). Therefore complete sequenced genomes would also require further rearrangement along with increased sequencing to ensure the entire genome can be assembled, giving highly informative data.

6 Discussion and Future Work

This project has started to reveal the microbes that are involved in anaerobic digestion when run in the lab scale system, under certain conditions. It has highlighted those microbes, such as those assigned as *Methanosarcina* and *Methanoculleus* that are increasing throughout the experiment and also indicated the closest known organisms. Although only a limited number of groups have been explored, this initial analysis offers a clear indication and a confidence that the methods used are suitable.

6.1 The Metagenomic Approach

The traditional method for understanding microbes is to isolate them from their environment and culture these microbes as pure cultures. From this, the microbes can be identified using a variety of tests, e.g. Gram staining. Although these methods have proved to be successful, as well as informative, there can often be drawbacks associated with these methods. A main challenge of using these techniques is that many microorganisms cannot be grown in isolation, often requiring other microbes, possibly due to a syntrophic interaction (Qiu et al., 2004). Furthermore, the isolation and identification of microbes in a complex community environment will be challenging and time consuming due to the large variety of species. Alternative methods have been developed that are culture independent, such as DNA sequencing using NGS technologies.

The targeted sequencing (16S rRNA) method appears to be a staple for numerous studies of complex communities, especially in AD e.g. Whiteley et al. (2012), St-Pierre & Wright (2014), Ziganshin et al. (2013), Garcia et al. (2011), Heeg et al. (2014), Tian et al. (2015), Kobayashi et al. (2014) and Jang et al. (2014). These publications offer interesting results, reporting the different microbial communities found. There have also been reports to bring together sequencing data from numerous anaerobic digesters e.g. Nelson et al. (2011) and Leclerc et al. (2004). A drawback associated with this method however is that PCR is known to introduce errors, such as bias introduced by primers and the amplification process (Urich et al., 2008). This in turn has the potential to give inaccurate results that are not a true reflection of the microbes present within the samples. Furthermore, the sequencing of the 16S rRNA region does not provide information on the microbial functions.

The shotgun sequencing of DNA samples using NGS technologies has revolutionised studies into complex microbial communities. The use of this technology eliminates the associated challenges known with targeted PCR amplification and provides much more informative detail on the individual microbes e.g. not only does this method provide information on the microbial community structure, but also what the potential function of those microbes could be (Vanwonterghem et al., 2014). It also has the key advantage that NGS allows for the discovery of novel microbes by assembling draft genomes, and this could ultimately reveal those microbes that are central to the AD process. By doing this there is the potential to optimise AD systems, such as increasing the amount of methane that is derived from the input material.

High throughput sequencing is an exciting technology but is limited by computational analysis. The large data sets generated through the technology can be a challenge to analyse. There are a limited number of annotation software packages available that interpret the data, such as MG-RAST, where numerous authors e.g. Yang et al. (2014) and Kovács et al. (2015) have used this program. Although this program is a useful tool, it is considered that the assignments can be somewhat inaccurate, and other approaches many be more beneficial, especially when dealing with unknown organisms. The methodology taken in this thesis offers a different approach to analyse the sequencing data, by using a variety of contig assemblers as described in Scholz et al. (2014), and a novel clustering method that has been proven to work. These methods would need to be automated to ensure the data can be interpreted.

6.2 Automated Analysis

To build upon the current work of this thesis, the following areas require additional work. Most importantly, the automation of analysis, via the pipeline we have developed for the sequence data, would need to be established. The current pipeline requires user intervention at the majority of the stages. This is mainly because at each stage checks were carried out to ensure that the data were processed in the intended manner. As it has been demonstrated that the processes used in this work are suitable, these now need to be joined together, so all the tasks carried out in this document are completed automatically (Figure 6.1). This would ensure that the large sequencing datasets generated by NGS can be interpreted. If a pipeline were implemented, it would allow for further analysis on the remaining clusters identified in the first experiment.

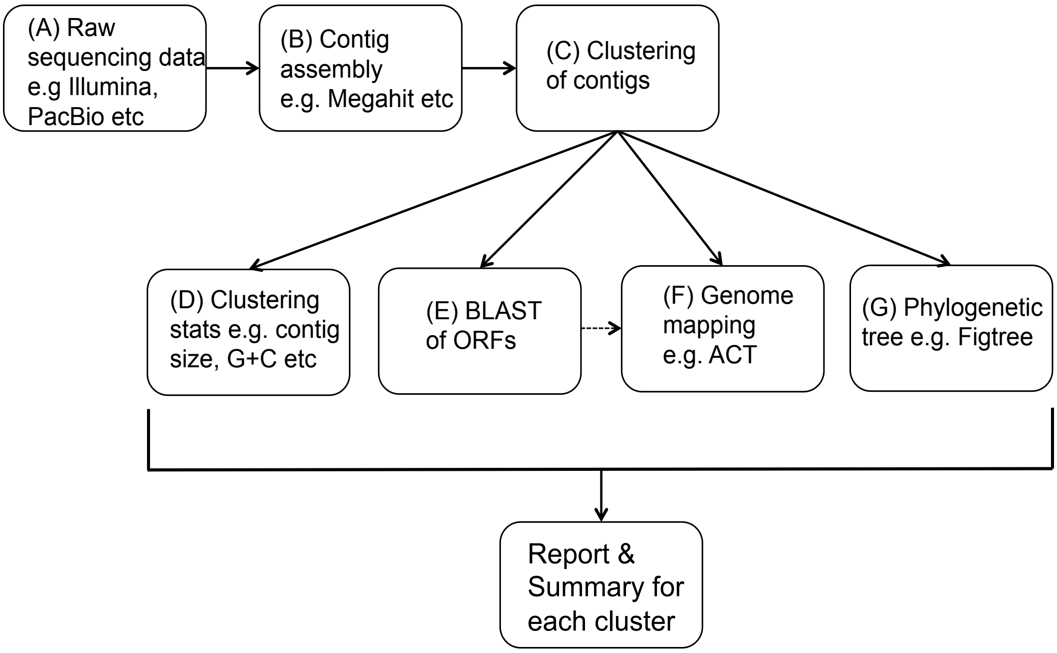

(A) Raw sequencing data e.g Illumina, PacBio etc
(B) Contig assembly e.g. Megahit etc
(C) Clustering of contigs
(D) Clustering stats e.g. contig size, G+C etc
(E) BLAST of ORFs
(F) Genome mapping e.g. ACT
(G) Phylogenetic tree e.g. Figtree
Report & Summary for each cluster

Figure 6.1. Proposed pipeline for the analysis of DNA sequencing data. The (a) initial sequencing data would be loaded into the (b) contig assemblers, where numerous assemblies would take place to generate the longest possible contigs, and then (c) clustered using the method described in this thesis. The clusters would then be used to produce an (d) output file for each of these, the ORFs (by Prokka) (e) automatically BLAST searched for the top match, (f) mapped onto the closest known genome, according to the BLAST results and (g) placed in a phylogenetic tree using a core set of genes database. The pipeline would produce a summary for each cluster, such as those in Figure 5.17 and 5.18.

The initial results produced using the methods described in this thesis have highlighted microbes that are increasing in abundance within our model system and support the notion that these organisms have not been previously described, as none of the top BLAST results had identical matches. Although a draft genome assembly was attempted, the amount of sequence data was not enough.

6.3 Additional Sequencing

Although two different sequencing platforms providing long and short reads were used for the microbial community analysis, the amount of sequence returned was not enough to reassemble a complete draft microbial genome. The assembly of a complete microbial genome gives the potential to understand what the function of that microbe is within the system. The current analysis methods allow the length of sequence in question be compared to other organisms phylogentically, but this does not necessarily mean that the microbe from which the sequence was derived from will be performing the same functions as those microbes it is most closely related to. That said, with the current dataset it might be possible to investigate the genes in an organism, indicating its potential role in the AD process. Further sequencing of samples ensures that more of the microbial genomes are sequenced, and this coupled with longer read sequencing technologies, has the potential to produce complete draft assembled genomes of novel organisms.

Furthermore, additional samples from experiment three, where three lab scale AD systems were run in parallel, would need to be sequenced to determine if the changes observed in process data are reflected in the microbial communities. The process data were suggestive that the microbial community structure outcome is shaped via deterministic factors. Further sequencing of the samples taken daily would be highly informative and would either disprove or further support this theory.

The long reads obtained using PacBio are an important tool in analysing microbial communities. In this experiment, the reads obtained were not as long as anticipated, probably due to the DNA extraction method. To ensure that longer reads are obtained when using this technology, a different DNA extraction method should be developed e.g. Bey et al. (2010) and Pang et al. (2008).

6.4 Merging process and metagenomic data

The aim of collecting both the process data from the experiments as well as investigating the microbial community dynamics using a variety of approaches e.g. qPCR and metagenomics, was to determine where correlations between system performance and the microbial community can be made (Figure 6.2).

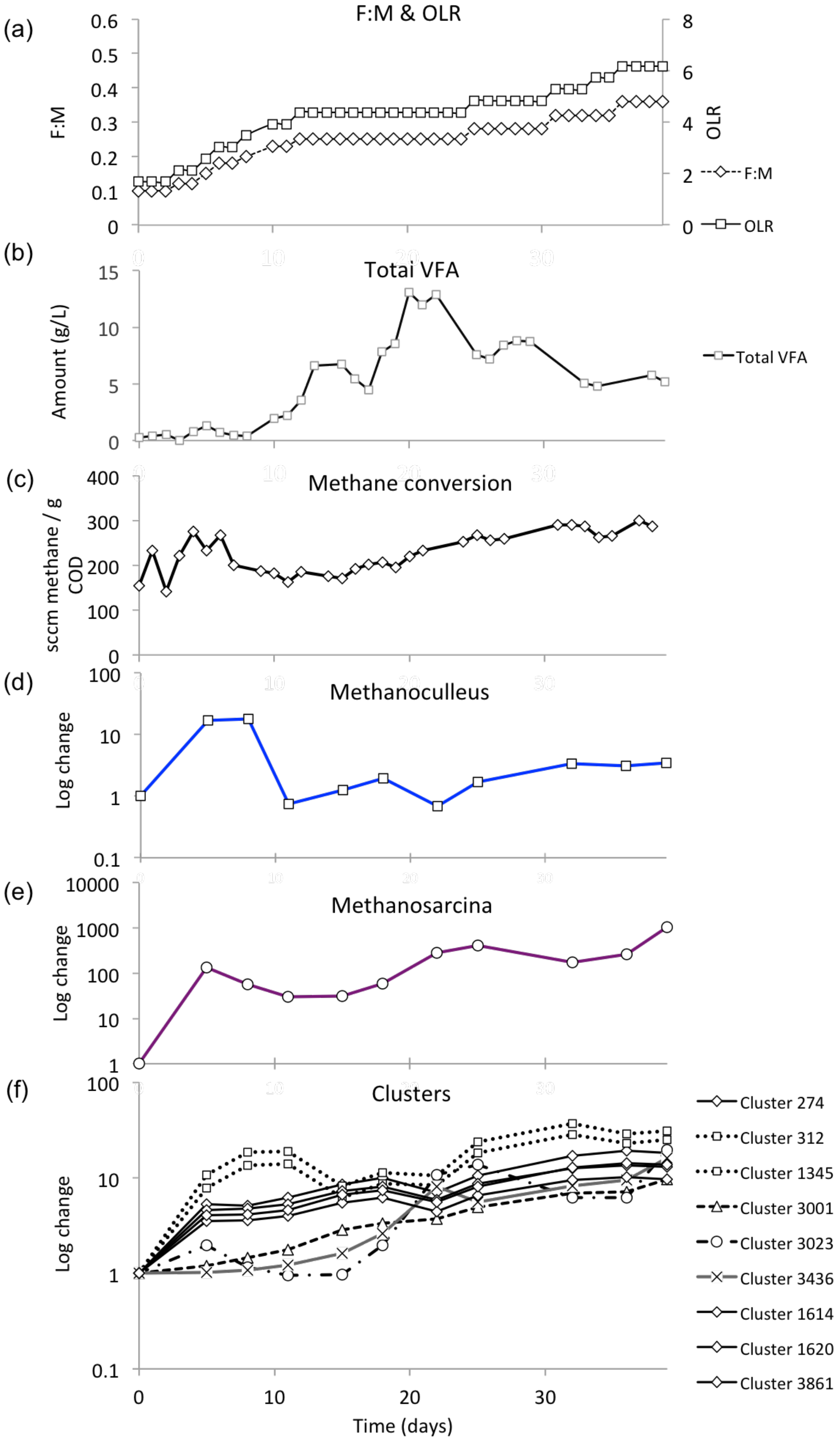


Figure 6.2. The process and metagenomic data from experiment one. The F:M and OLR (a), total VFAs (b), the methane conversion efficiency (c) compared to the qPCR data for the two measured methanogens, *Methanoculleus* (d) and *Methanosarcina* (e) and the five methanogen groups (NC1-5) (f), identified by this work.

The methane conversion efficiency is initially high in experiment one, but after day 7 decreases, and thereafter gradually increases (Figure 6.2 c). Interestingly, the methanogens appear to undergo a significant increase that coincides with this initial increase in methane conversion efficiency. The qPCR data suggests that the *Methanoculleus* species increases during this time (over 15 fold, Figure 6.2 d), but as the methane efficiency decreases, so do these organisms. The same pattern is observed for the *Methanosarcina* species (Figure 6.2 e) that have over a 100 fold increase after 5 days, but again, decrease slightly. These *Methanosarcina* generally increase in abundance throughout the experiment, with some slight fluctuations and this increase correlates with an improved methane conversion efficiency. Curiously however, the metagenomic data do not match the qPCR data entirely. Cluster 3023 has been shown to be closely related to *Methanosarcina mazei* (which the qPCR was targeted to), and a similar trend in microbial dynamics can be observed, but the fold changes of these organisms are very different. For example, at day 25, the qPCR data suggests *Methanosarcina* species have increased by over 400 fold, compared to the starting sample, whereas the metagenomic data suggest this microbe has increased only 10 fold. Furthermore our analysis suggests clusters 274, 1620, 1614 and 3861 are related to *Methanoculleus marisnigri*. The *Methanoculleus* data from qPCR suggests there is an initial increase (over 10 fold) of this organism by day 8, before a rapid decrease, but the metagenomic data imply that this organism does have an initial increase, but continues to increase in abundance throughout the experiment. Hori et al. (2006) reported that *Methanoculleus* decreased during high levels of VFAs, which correlated with the qPCR data, but Ma et al. (2015) suggested that *Methanoculleus* was not affected by changing VFA levels, as shown by the metagenomic data. A possible reason for differences between the two datasets could be the selected qPCR primers have actually targeted more than one organism showing the same behaviour, or possibly a different strain. This emphasises the need for careful primer selection. To determine if the qPCR data is accurate (especially for the *Methanoculleus*) further investigation into the primers used and therefore the product formed would be required. Although preliminary checks were conducted in the form of checking for single product formation and performing a melt curve, further investigations including sequencing of the PCR product could be carried out. It could also be that as only a small percentage of the sequencing data has been analysed, the microbes displaying this behaviour are yet to be identified e.g. *Methanoculleus* measured by qPCR could be a different strain to that group of clusters, assigned also as *Methanoculleus*. Therefore the analysis method chosen to determine microbial abundance change could actually influence the result. It can be concluded however that both analysis methods can be useful and that some comparisons between the two datasets can be drawn. It is probable that the *Methanosarcina* species should exhibit an increase in abundance during this experiment as the acetic acid concentration was high (with a maximum at over 12 g/L), and this organism utilises this substrate to produce methane. Interestingly, Hao et al. (2013) suggested that the aceticlastic pathways was inhibited when the acetate concentration was above 50mM, and that hydrogenotrophic methanogens were more tolerant to high acetate levels, but these data suggest that the high acetic acid levels are having little effect on the methanogen activity or methane production. *Methanosarcina* has been reported to have a higher growth rate than *Methansaeata* during high acetate levels (Walter et al., 2012), but the metagenomic data suggests both these organisms are able to grow, at comparable levels, under the conditions of high acetate. This could explain the increasing methane output regardless of the high VFA levels.

Complementary to sequencing DNA samples, investigating the transcriptome would begin to reveal those microbes and associated genes that are active within the AD systems.

6.5 Transcriptomics

The use of metagenomic shotgun sequencing of DNA has the advantage that it is a method to investigate microbial community structure, along with revealing the potential functions of organisms (Alneberg et al., 2014), along with discovery of new and novel genes (Urich et al., 2008). The drawback associated with this method however is that DNA sequencing and analysis can only provide information on the microbes that are present in the system and does not necessarily indicate that these microbes are active within the system for a specific function. Additionally, DNA sequencing does not provide information on the expression state of the genes and ultimately the functional roles (Urich et al., 2008) of microbes. Those microbes that are increasing within the systems, as identified in this thesis, would be assumed to be contributing towards the digestion process, whereas others that appear to have no relative increase in numbers could actually be there, but not highly active. It could be possible that the microbes showing little change could actually be metabolically active but not growing at the same rates as other organisms. An alternative to sequencing DNA is RNA. Investigation into the transcriptome could potentially reveal those microbes that are active within the system, the functions of these and give a much clearer perspective of the process occurring within AD systems.

6.6 Further lab experiments

Experiments over longer timescales using the lab scale digester would be important to further investigate process data variations and the microbial communities involved in the process. In this thesis, the longest trial was run for 57 days, although the analysis investigating the microbial communities has only been conducted on the 39 day trial (experiment one) thus far. For these experiments, the systems were run until maximum feed rates (F:M of 0.3, as done in industry) were achieved. If these systems were run for longer then it would give the opportunity to investigate the changes occurring both for process data and the microbial populations. The focus of two experiments in this thesis was to investigate the changes occurring within the microbial communities in the system, along with rate of change, until the maximum feed rate was achieved, but maximum feed rate does not necessarily mean maximum stable running conditions.

Longer experiments also have the benefit of investigating how stable and robust the microbial communities are within AD systems. It has been reported (De Francisci et al., 2015) that if the composition of the feedstock is dramatically changed in a system, then the microbial population responds to this change, and that can result in a decrease in performance. Therefore longer experiments to further investigate this issue, such as varying the change of feedstock composition levels to determine how robust and stable the systems are to slight changes, along with greater ones. Variations in feedstock composition is an important consideration as there can always be variability in feedstock compositions. Understanding how the microbial communities change under specific conditions, especially if run in triplicate (or more system) would be highly informative and an ultimate aim would be to build a predictive model of how the AD systems behave. The IWA Anaerobic Digestion Model 1 (ADM1) (Batstone et al., 2002) is a model that considers numerous processes to simulate all the reactions occurring in AD. These include both biological and physio-chemical reactions (Jeong et al., 2005). This model is widely used and has often been adapted e.g. Galí et al. (2009). Having a computational model to simulate the AD process is advantageous as it can allow for predictions to be made, and shows confidence in the process. There is a lot of embedded expertise in those who operate the AD systems to ensure the systems are run properly, but there is a requirement for an updated predictive model to make the process more predictable, especially as the use of NGS has revealed novel microbes and the biochemistry of these is better understood. Experiment three can serve to act towards the goal of building a predictive model, once the DNA sequenced samples have been analysed. Further experiments using three or more lab systems can be used to start to understand the interactions within the AD systems, as well as used to predict and model the AD process.

Appendix A

```
#
# produce fasta files and some statistics for clusters
# push clusters that are less than 20kb into a single file
# JC 15/7/15
#

import csv
import re
from Bio import SeqIO
from Bio.SeqUtils import GC
from Bio.SeqRecord import SeqRecord

# put all the fasta sequences into a dictionary
# key to fasta_dict contains contig name
# entry contains length of contig, contig %GC and contig
sequence

fasta_dict = {}

print('Loading fasta files')

handle = open('final_full.fasta', 'rU')
for record in SeqIO.parse(handle, "fasta"):
      fasta_entry =
len(record.seq),GC(record.seq),record.seq,record.id
      fasta_dict[record.name] = fasta_entry

# put all the clusters into another dictionary
# this code parses the file containing contigs
# each cluster has a name in group_dict.keys
# plus an entry which contains all of the matching contigs

group_dict = {}

print('Loading cluster details')

with open('clustered_contigs.txt', 'r') as groups:
      datafile = groups.read()
      entry = ''
      for character in datafile:
            entry = entry+character
            if character == '\n':
                  details = entry.split('[')
                  group_name = details[0]
                  others = details[1]
                  other = others.rsplit(']')
                  contig_list = other[0]
                  contigs = re.sub('[\']', '', contig_list)
                  group_dict[group_name] = group_name+',
'+contigs
                  entry = ''

# how to parse the group_dict

#lumpy = group_dict.keys()
#for lister in lumpy:
```

```
#     marge = group_dict[lister]
#     simpsons = marge.split(', ')
#     for list in range(len(simpsons)):
#           print simpsons[list]


# set up a dictionary with the abundance data (Windows.csv
file)

input_file = 'output_to_sort.csv'
input_dict = {}

print('Loading abundance data')

with open(input_file, 'r') as input:
       for row in input:
              query_values = row.strip('\r\n')
              query_parts = query_values.split(',')
              query_id = query_parts[0]
              query_values = query_parts[1:12]
              input_dict[query_id] = query_values

# now make a final dictionary (big_dict)
# this contains clusters of contigs with interleaved fasta
entries
# and relative abundance data
# appended to the end of each record are statistics on number
of contigs
# in cluster and total number of bp in each cluster

big_dict = {}

print('Assigning contigs to groups')

cluster_list = group_dict.keys()
fasta_names = fasta_dict.keys()

for cluster in cluster_list:
      contig_list = group_dict[cluster]
      contig_names = contig_list.split(', ')
      number_of_contigs = 0
      length_of_contigs = 0
      for contig_name in range(len(contig_names)):
            cluster_name = contig_names[0]
            current_entry = contig_names[contig_name]
#           print ('Cluster name is %s' % cluster_name)
            fasta_record = fasta_dict[current_entry]
#           print ('Contig name is %s' % current_entry)
#           print ('Record for contig is', fasta_record)
            entry_length = fasta_record[0]
#           print ('Contig is %d bp long' % entry_length)
            entry_GC = fasta_record[1]
#           print ('with GC content %d' % entry_GC)
            if str('Group_'+cluster_name) in big_dict:

      big_dict['Group_'+cluster_name].append(fasta_record)
            else:
                  big_dict['Group_'+cluster_name] =
[fasta_record]
            abundance_data = input_dict[current_entry]
      big_dict['Group_'+cluster_name].append(abundance_data)
            number_of_contigs += 1
            length_of_contigs += entry_length
      big_dict['Group_'+cluster_name].append(number_of_contigs
)
      big_dict['Group_'+cluster_name].append(length_of_contigs
)
#print big_dict

# next step is to sort by total length of cluster and save
file 'length'
# then sort by number of contigs in cluster and save another
file 'number'
# initially save one file each for the top 20 plus one file
for the rest
# into different directories

# sort by length

print ('Sorting clusters by length')

length_list = []
list_length = big_dict.keys()
for loop1 in list_length:
      longest = 0
      for loop2 in list_length:
            record1 = big_dict[loop2][-1]
            if record1 >= longest:
                  longest = record1
                  longest_id = loop2
      length_list.append(longest_id)
      list_length.remove(longest_id)
#add subroutine to print list with length of each cluster

print ('Saving length-sorted clusters')

top_20_length = 0
for checker in length_list:
      if top_20_length < 50:
            file_name = str(checker)
            data_name = file_name+'.csv'
            file_name = file_name+'.fasta'
            print file_name
            contig_file = open(file_name, 'w')
            data_file = open(data_name, 'w')
            w = csv.writer(data_file)
            w.writerow([0,5,8,11,15,18,22,25,32,36,39])
            for outputter1 in range(0, int(big_dict[checker][-
2]), 2):
                  sequence_going_out =
big_dict[checker][outputter1]
                  numbers_going_out =
big_dict[checker][outputter1+1]
                  x = SeqRecord(sequence_going_out[2], id =
sequence_going_out[3], description = "length
"+str(sequence_going_out[0])+" GC content
"+str(sequence_going_out[1]))
                  SeqIO.write(x, contig_file, 'fasta')
#                 data_file.write(numbers_going_out)
```

```
                w = csv.writer(data_file)
                w.writerow(numbers_going_out)
            contig_file.close()
            data_file.close()
      else:
            contig_file = open('below_top_20.fasta', 'w')
            for outputter2 in range(0, int(big_dict[checker][-
2]), 2):
                  sequence_going_out =
big_dict[checker][outputter2]
                        x = SeqRecord(sequence_going_out[2],
id = sequence_going_out[3], description = "length
"+str(sequence_going_out[0]))
                        SeqIO.write(x, contig_file, 'fasta')
            contig_file.close()
      top_20_length += 1

# sort by number

#print ('Sorting clusters by number of contigs')

#number_list = []
#list_number = big_dict.keys()
#for loop3 in list_number:
#        most = 0
#        for loop4 in list_number:
#                record2 = big_dict[loop4][-2]
#                if record2 >= most:
#                        most = record2
#                        most_id = loop4
#        number_list.append(most_id)
#        list_number.remove(most_id)

#print ('Saving clusters with most contigs')

#top_20_number = 0
#for hecker in number_list:
#        if top_20_number < 20:
#                file_namer = str(hecker)
#                file_namer = 'numbers_'+file_namer+'.fasta'
#                print file_namer
#                contig_filer = open(file_namer, 'w')
#                for outputter3 in range(0,
int(big_dict[hecker][-2]), 2):
#                        sequence_going =
big_dict[hecker][outputter3]
#                        y = SeqRecord(sequence_going[2], id =
sequence_going[3], description = "length
"+str(sequence_going[0]))
#                        SeqIO.write(y, contig_filer, 'fasta')
#                contig_filer.close()
#        else:
#                contig_filer =
open('numbers_below_top_20.fasta', 'w')
#                for outputter4 in range(0,
int(big_dict[hecker][-2]), 2):
#                        sequence_going =
big_dict[hecker][outputter4]
```

```
#                        y = SeqRecord(sequence_going[2], id =
sequence_going[3], description = "length
"+str(sequence_going[0]))
#                        SeqIO.write(y, contig_filer, 'fasta')
#                contig_filer.close()
#        top_20_number += 1
```

Appendix B

```
from __future__ import print_function, division

from Bio.Blast import NCBIXML
import argparse
import re

REPORT_NO_HITS = False

def check_description(title_string):
    if re.search("[Hh]ypothetical", title_string):
        return False
    elif re.search("[Pp]utative", title_string):
        return False

    return True

parser = argparse.ArgumentParser(description="Summarise
BlastXML output files")
parser.add_argument("files", metavar="<filename>", type=str,
nargs='+', help="A blast XML output file to summarise")
args = parser.parse_args()

for filename in args.files:
    infile = open(filename, 'r')

    for blast_record in NCBIXML.parse(infile):
        query_title = blast_record.query

        for description in blast_record.descriptions:
            if check_description(description.title):
                print("\t".join([query_title,
description.accession, description.title,
str(description.e)]))
                break
        else:
            if REPORT_NO_HITS:
                print("No suitable hits for " + query_title)
```

Appendix C

The core genes were identified using MicroScope – Microial Genome Annotation & Analysis Platform

The following organisms were selected and searched using an MICFAM parameter of 50/80:

Methanobrevibacter smithii ATCC 35061
Methanobrevibacter sp. JH1
Methanocaldococcus fervens AG86
Methanocaldococcus jannaschii DSM 2661
Methanococcoides burtonii DSM 6242
Methanococcus aeolicus Nankai-3
Methanococcus maripaludis C5
Methanococcus maripaludis C6
Methanococcus maripaludis C7
Methanococcus maripaludis S2
Methanococcus maripaludis X1
Methanococcus vannielii SB
Methanococcus voltae A3
Methanocorpusculum labreanum Z
Methanoculleus bourgensis MS2 type strain:MS2
Methanoculleus marisnigri JR1
Methanomassiliicoccus luminyensis B10
Methanomethylovorans hollandica DSM 15978
Methanopyrus kandleri AV19
Methanoregula boonei 6A8
Methanosaeta concilii GP-6
Methanosaeta thermophila PT
Methanosarcina acetivorans C2A
Methanosarcina barkeri Fusaro
Methanosarcina mazei Go1
Methanosphaera stadtmanae DSM 3091
Methanospirillum hungatei JF-1
Methanothermobacter marburgensis Marburg
Methanothermobacter thermautotrophicus Delta H
Methanotorris igneus Kol 5

Abbreviations

°C	degrees celsius
AD	Anaerobic Digestion
BLAST	Basic Local Alignment Search Tool
bp	Base pair
C/N	Carbon to Nitrogen
CAFT	Cavitation Air Flotation Tank
COD	Chemical Oxygen Demand
EtBr	Ethidium Bromide
F:M	Feed to Mass
g	gram
g/L	gram per litre
Gb	Gigabase
HRT	Hydraulic Retention Time
Kb	Kilobase
kD	Kilodalton
Kg	Kilogram
L	Litre
L/h	Litre per hour
m	metre
m^3	Cubic metre
ml	millilitre
ml/min	millilitre per minute
mM	millimolar
mm	millimetre
mRNA	messenger ribonucleic acid
Na_2CO_3	Sodium Carbonate
NaOH	Sodium Hydroxide
ng	nanogram
nm	nanometre
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
qPCR	Quantitative Real Time PCR
RNA	Ribonucleic acid

rRNA	ribosomal ribonucleic acid
sccm	Standard cubic centermetres per minute
SD	Standard Deviation
TAE	Tris-Aceate-EDTA
TFF	Tangential Flow Filtration
UK	United Kingdom
ul	microlitre
VFA	Volatile Fatty Acid
VS	Volatile Solids
VSS	Volatile Suspended Solids

List of References

Akuzawa, M., Hori, T., Haruta, S., Ueno, Y., Ishii, M., Igarashi, Y., 2011. Distinctive Responses of Metabolically Active Microbiota to Acidification in a Thermophilic Anaerobic Digester. *Microbial Ecology 61*, 595–605.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C., 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods 11*, 1144–1146.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research 25*, 3389–3402.

Alves, M.M., Pereira, M.A., Sousa, D.Z., Cavaleiro, A.J., Picavet, M., Smidt, H., Stams, A.J.M., 2009. Waste lipids to energy: how to optimize methane production from long-chain fatty acids (LCFA). *Microbial Biotechnology 2*, 538–550.

Anderson, I.J., Sieprawska-Lupa, M., Lapidus, A., Nolan, M., Copeland, A., Glavina Del Rio, T., Tice, H., Dalin, E., Barry, K., Saunders, E., Han, C., Brettin, T., Detter, J.C., Bruce, D., Mikhailova, N., Pitluck, S., Hauser, L., Land, M., Lucas, S., Richardson, P., Whitman, W.B., Kyrpides, N.C., 2009. Complete genome sequence of Methanoculleus marisnigri Romesser et al. 1981 type strain JR1. *Standards in genomic sciences 1*, 189–196.

Appels, L., Baeyens, J., Degrève, J., Dewil, R., 2008. Principles and potential of the anaerobic digestion of waste-activated sludge. *Progress in Energy and Combustion Science 34*, 755–781.

Ariunbaatar, J., Panico, A., Esposito, G., Pirozzi, F., Lens, P.N.L., 2014. Pretreatment methods to enhance anaerobic digestion of organic solid waste. *Applied Energy 123*, 143–156.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a., Pevzner, P. a., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology 19*, 455–477.

Banks, C.J., Zhang, Y., Jiang, Y., Heaven, S., 2012. Trace element requirements for stable food waste digestion at elevated ammonia concentrations. *Bioresource Technology 104*, 127–135.

Barber, R.D., Zhang, L., Harnack, M., Olson, M. V., Kaul, R., Ingram-Smith, C., Smith, K.S., 2011. Complete genome sequence of Methanosaeta concilii, a specialist in aceticlastic methanogenesis. *Journal of Bacteriology 193*, 3668–3669.

Batstone, D.J., Keller, J., Angelidaki, I., Kalyuzhnyi, S. V., Pavlostathis, S.G., Rozzi, a., Sanders, W.T., Siegrist, H., Vavilin, V. a., 2002. The IWA Anaerobic Digestion Model No 1 (ADM1). *Water Science and Technology 45*, 65–73.

Belda-Ferre, P., Alcaraz, L.D., Cabrera-Rubio, R., Romero, H., Simón-Soro, A., Pignatelli, M., Mira, A., 2012. The oral metagenome in health and disease. *The ISME Journal 6*, 46–56.

Bey, B.S., Fichot, E.B., Dayama, G., Decho, A.W., Norman, R.S., 2010. Extraction of high molecular weight DNA from microbial mats. *BioTechniques 49*, 631–640.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J., 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology 13*, R122.

Borrel, G., Harris, H.M.B., Tottey, W., Mihajlovski, A., Parisot, N., Peyretaillade, E., Peyret, P., Gribaldo, S., O'Toole, P.W., Brugere, J.-F., 2012. Genome Sequence of "Candidatus Methanomethylophilus alvus" Mx1201, a Methanogenic Archaeon from the Human Gut Belonging to a Seventh Order of Methanogens. *Journal of Bacteriology 194*, 6944–6945.

Calli, B., Mertoglu, B., Inanc, B., Yenigun, O., 2005. Effects of high free ammonia concentrations on the performances of anaerobic bioreactors. *Process Biochemistry 40*, 1285–1292.

Cardinali-Rezende, J., Colturato, L.F.D.B., Colturato, T.D.B., Chartone-Souza, E., Nascimento, A.M. a, Sanz, J.L., 2012. Prokaryotic diversity and dynamics in a full-scale municipal solid waste anaerobic reactor from start-up to steady-state conditions. *Bioresource Technology 119*, 373–383.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis comparison tool. *Bioinformatics 21*, 3422–3423.

Chan, E.R., Hester, J., Kalady, M., Xiao, H., Li, X., Serre, D., 2011. A novel method for determining microflora composition using dynamic phylogenetic analysis of 16S ribosomal RNA deep sequencing data. *Genomics 98*, 253–259.

Chen, Y., Cheng, J.J., Creamer, K.S., 2008. Inhibition of anaerobic digestion process: A review. *Bioresource Technology 99*, 4044–4064.

Cysneiros, D., Banks, C.J., Heaven, S., Karatzas, K.A.G., 2012. The effect of pH control and "hydraulic flush" on hydrolysis and Volatile Fatty Acids (VFA) production and profile in anaerobic leach bed reactors digesting a high solids content substrate. *Bioresource Technology 123*, 263–271.

De Francisci, D., Kougias, P.G., Treu, L., Campanaro, S., Angelidaki, I., 2015. Microbial diversity and dynamicity of biogas reactors due to radical changes of feedstock composition. *Bioresource Technology 176*, 56–64.

De Vrieze, J., Hennebel, T., Boon, N., Verstraete, W., 2012. Methanosarcina: The rediscovered methanogen for heavy duty biomethanation. *Bioresource Technology 112*, 1–9.

De Vrieze, J., Saunders, A.M., He, Y., Fang, J., Nielsen, P.H., Verstraete, W., Boon, N., 2015. Ammonia and temperature determine potential clustering in the anaerobic digestion microbiome. *Water Research.*

Department of Energy & Climate Change, 2014. 2012 UK Greenhouse Gas Emissions.

Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., Reid, G., 2013. High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods 95*, 401–414.

Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C., Fitter, A.H., 2010. Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal 4*, 337–345.

Dziewit, L., Pyzik, A., Romaniuk, K., Sobczak, A., Szczesny, P., Lipinski, L., Bartosik, D., Drewniak, L., 2015. Novel molecular markers for the detection of methanogens and phylogenetic analyses of methanogenic communities. *Frontiers in microbiology 6*, 694.

Elbeshbishy, E., Nakhla, G., 2012. Batch anaerobic co-digestion of proteins and carbohydrates. *Bioresource Technology 116*, 170–178.

Fetzer, S., Bak, F., Conrad, R., 1993. Sensitivity of methanogenic bacteria from paddy soil to oxygen and desiccation. *FEMS Microbiology Ecology 12*, 107–115.

Franke-Whittle, I.H., Walter, A., Ebner, C., Insam, H., 2014. Investigation into the effect of high concentrations of volatile fatty acids in anaerobic digestion on methanogenic communities. *Waste Management 34*, 2080–2089.

Galí, A., Benabdallah, T., Astals, S., Mata-Alvarez, J., 2009. Modified version of ADM1 model for agro-waste application. *Bioresource Technology 100*, 2783–2790.

Garcia, S.L., Jangid, K., Whitman, W.B., Das, K.C., 2011. Transition of microbial communities during the adaption to anaerobic digestion of carrot waste. *Bioresource Technology 102*, 7249–7256.

Gavala, H.N., Yenal, U., Skiadas, I. V., Westermann, P., Ahring, B.K., 2003. Mesophilic and thermophilic anaerobic digestion of primary and secondary sludge. Effect of pre-treatment at elevated temperature. *Water Research 37*, 4561–4572.

Ge, H., Jensen, P.D., Batstone, D.J., 2011. Temperature phased anaerobic digestion increases apparent hydrolysis rate for waste activated sludge. *Water Research 45*, 1597–1606.

Gorlas, A., Robert, C., Gimenez, G., Drancourt, M., Raoult, D., 2012. Complete Genome Sequence of Methanomassiliicoccus luminyensis, the Largest Genome of a Human-Associated Archaea Species. *Journal of Bacteriology 194*, 4745–4745.

Goulding, D., Power, N., 2013. Which is the preferable biogas utilisation technology for anaerobic digestion of agricultural crops in Ireland: Biogas to CHP or biomethane as a transport fuel? *Renewable Energy 53*, 121–131.

Hao, L., Lü, F., Li, L., Wu, Q., Shao, L., He, P., 2013. Self-adaption of methane-producing communities to pH disturbance at different acetate concentrations by shifting pathways and population interaction. *Bioresource Technology 140*, 319–327.

Hao, L.P., Lü, F., Li, L., Shao, L.M., He, P.J., 2012. Shift of pathways during initiation of thermophilic methanogenesis at different initial pH. *Bioresource Technology 126*, 418–424.

Heeg, K., Pohl, M., Sontag, M., Mumme, J., Klocke, M., Nettmann, E., 2014. Microbial communities involved in biogas production from wheat straw as the sole substrate within a two-phase solid-state anaerobic digestion. *Systematic and applied microbiology 37*, 590–600.

Hongoh, Y., Yuzawa, H., Ohkuma, M., Kudo, T., 2003. Evaluation of primers and PCR conditions for the analysis of 16S rRNA genes from a natural environment. *FEMS Microbiology Letters 221*, 299–304.

Hori, T., Haruta, S., Sasaki, D., Hanajima, D., Ueno, Y., Ogata, A., Ishii, M., Igarashi, Y., 2015. Reorganization of the bacterial and archaeal populations associated with organic loading conditions in a thermophilic anaerobic digester. *Journal of Bioscience and Bioengineering 119*, 337–344.

Hori, T., Haruta, S., Ueno, Y., Ishii, M., Igarashi, Y., 2006a. Dynamic Transition of a Methanogenic Population in Response to the Concentration of Volatile Fatty Acids in a Thermophilic Anaerobic Digester Dynamic Transition of a Methanogenic Population in Response to the Concentration of Volatile Fatty Acids in a The. *Applied and Environmental Microbiology 72*, 1623–1630.

Hori, T., Haruta, S., Ueno, Y., Ishii, M., Igarashi, Y., 2006b. Dynamic Transition of a Methanogenic Population in Response to the Concentration of Volatile Fatty Acids in a Thermophilic Anaerobic Digester. *Applied and Environmental Microbiology 72*, 1623–1630.

Jäger, D., Sharma, C.M., Thomsen, J., Ehlers, C., Vogel, J., Schmitz, R. a, 2009. Deep sequencing analysis of the Methanosarcina mazei Gö1 transcriptome in response to nitrogen availability. *Proceedings of the National Academy of Sciences of the United States of America 106*, 21878–21882.

Jang, H.M., Kim, J.H., Ha, J.H., Park, J.M., 2014. Bacterial and methanogenic archaeal communities during the single-stage anaerobic digestion of high-strength food wastewater. *Bioresource Technology 165*, 174–182.

Jeong, H.-S., Suh, C.-W., Lim, J.-L., Lee, S.-H., Shin, H.-S., 2005. Analysis and application of ADM1 for anaerobic methane production. *Bioprocess and biosystems engineering 27*, 81–89.

Jiang, Y., Heaven, S., Banks, C.J., 2012. Strategies for stable anaerobic digestion of vegetable waste. *Renewable Energy 44*, 206–214.

Kampmann, K., Ratering, S., Baumann, R., Schmidt, M., Zerr, W., Schnell, S., 2012. Hydrogenotrophic methanogens dominate in biogas reactors fed with defined substrates. *Systematic and Applied Microbiology 35*, 404–413.

Keating, C., Chin, J.P., Hughes, D., Manesiotis, P., Cysneiros, D., Mahony, T., Smith, C.J., McGrath, J.W., O'Flaherty, V., 2016. Biological Phosphorus Removal During High-Rate, Low-Temperature, Anaerobic Digestion of Wastewater. *Frontiers in Microbiology 7*, 1–14.

Kim, J., Lim, J., Lee, C., 2013. Quantitative real-time PCR approaches for microbial community studies in wastewater treatment systems: Applications and considerations. *Biotechnology Advances 31*, 1358–1373.

Kim, M., Le, H., McInerney, M.J., Buckel, W., 2013. Identification and characterization of re-citrate synthase in syntrophus aciditrophicus. *Journal of Bacteriology 195*, 1689–1696.

Kobayashi, T., Tang, Y., Urakami, T., Morimura, S., Kida, K., 2014. Digestion performance and microbial community in full-scale methane fermentation of stillage from sweet potato-shochu production. *Journal of Environmental Sciences (China) 26*, 423–431.

Kougias, P.G., Boe, K., O-Thong, S., Kristensen, L.A., Angelidaki, I., 2014. Anaerobic digestion foaming in full-scale biogas plants: a survey on causes and solutions. *Water Science & Technology 69*, 889.

Kovács, E., Wirth, R., Maróti, G., Bagi, Z., Nagy, K., Minárovits, J., Rákhely, G., Kovács, K.L., 2015. Augmented biogas production from protein-rich substrates and associated metagenomic changes. *Bioresource Technology 178*, 254–261.

Krakat, N., Schmidt, S., Scherer, P., 2011. Potential impact of process parameters upon the bacterial diversity in the mesophilic anaerobic digestion of beet silage. *Bioresource Technology 102*, 5692–5701.

Lang, K., Schuldes, J., Klingl, A., Poehlein, A., Daniel, R., Brune, A., 2015. New Mode of Energy Metabolism in the Seventh Order of Methanogens as Revealed by Comparative Genome Analysis of "Candidatus Methanoplasma termitum." *Applied and Environmental Microbiology 81*, 1338–1352.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics 23*, 2947–2948.

Leclerc, M., Delgènes, J.-P., Godon, J.-J., 2004. Diversity of the archaeal community in 44 anaerobic digesters as determined by single strand conformation polymorphism analysis and 16S rDNA sequencing. *Environmental microbiology 6*, 809–19.

Lee, S.-H., Park, J.-H., Kim, S.-H., Yu, B.J., Yoon, J.-J., Park, H.-D., 2015. Evidence of syntrophic acetate oxidation by Spirochaetes during anaerobic methane production. *Bioresource Technology.*

Li, A., Chu, Y., Wang, X., Ren, L., Yu, J., Liu, X., Yan, J., Zhang, L., Wu, S., Li, S., 2013. A pyrosequencing-based metagenomic study of methane-producing microbial community in solid-state biogas reactor. *Biotechnology for Biofuels 6*, 3.

Li, Y., Zhang, Y., Xu, Z., Quan, X., Chen, S., 2015. Enhancement of sludge granulation in anaerobic acetogenesis by addition of nitrate and microbial community analysis. *Biochemical Engineering Journal 95*, 104–111.

Lim, J.W., Chen, C.L., Ho, I.J.R., Wang, J.Y., 2013. Study of microbial community and biodegradation efficiency for single- and two-phase anaerobic co-digestion of brown water and food waste. *Bioresource Technology 147*, 193–201.

Lindmark, J., Thorin, E., Bel Fdhila, R., Dahlquist, E., 2014. Effects of mixing on the result of anaerobic digestion: Review. *Renewable and Sustainable Energy Reviews 40*, 1030–1047.

Lindner, J., Zielonka, S., Oechsner, H., Lemmer, A., 2016. Is the continuous two-stage anaerobic digestion process well suited for all substrates? *Bioresource Technology 200*, 470–476.

Liu, C., Yuan, X., Zeng, G., Li, W., Li, J., 2008. Prediction of methane yield at optimum pH for anaerobic digestion of organic fraction of municipal solid waste. *Bioresource Technology 99*, 882–888.

Liu, C.-M., Luo, R., Lam, T.-W., 2015. MEGAHIT: An ultra-fast single-node solution for large and com- plex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics 31*, 11.

Lloret, E., Pastor, L., Pradas, P., Pascual, J. a., 2013. Semi full-scale thermophilic anaerobic digestion (TAnD) for advanced treatment of sewage sludge: Stabilization process and pathogen reduction. *Chemical Engineering Journal 232*, 42–50.

Lu, C.L., Chen, K.-T., Huang, S.-Y., Chiu, H.-T., 2014. CAR: contig assembly of prokaryotic draft genomes using rearrangements. *BMC bioinformatics 15*, 381.

Ma, J., Zhao, Q.B., Laurens, L.L.M., Jarvis, E.E., Nagle, N.J., Chen, S., Frear, C.S., 2015. Mechanism , kinetics and microbiology of inhibition caused by long - chain fatty acids in anaerobic digestion of algal biomass. *Biotechnology for Biofuels,* 1–12.

Madsen, M., Holm-Nielsen, J.B., Esbensen, K.H., 2011. Monitoring of anaerobic digestion processes: A review perspective. *Renewable and Sustainable Energy Reviews 15*, 3141–3155.

McHugh, S., O'Reilly, C., Mahony, T., 2003. Anaerobic granular sludge bioreactor technology. *Reviews in Environmental Science and Biotechnology 2,* 225–245.

McInerney, M.J., Sieber, J.R., Gunsalus, R.P., 2009. Syntrophy in anaerobic global carbon cycles. *Current Opinion in Biotechnology 20*, 623–632.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R., 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics 9*, 386.

Moestedt, J., Müller, B., Westerholm, M., Schnürer, A., 2016. Ammonia threshold for inhibition of anaerobic digestion of thin stillage and the importance of organic loading rate. *Microbial Biotechnology 9*, 180-194.

Moestedt, J., Nordell, E., Schnürer, A., 2014. Comparison of operating strategies for increased biogas production from thin stillage. *Journal of Biotechnology 175*, 22–30.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research 35*, W182–W185.

Murto, M., Björnsson, L., Mattiasson, B., 2004. Impact of food industrial waste on anaerobic co-digestion of sewage sludge and pig manure. *Journal of Environmental Management 70*, 101–107.

National Grid, 2009. The potential for Renewable Gas in the UK.

Nelson, M.C., Morrison, M., Yu, Z., 2011. A meta-analysis of the microbial diversity observed in anaerobic digesters. *Bioresource technology 102*, 3730–9.

Ofiteru, I.D., Lunn, M., Curtis, T.P., Wells, G.F., Criddle, C.S., Francis, C. a, Sloan, W.T., 2010. Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences of the United States of America 107*, 15345–15350.

Pang, M.F., Abdullah, N., Lee, C.W., Ng, C.C., 2008. Isolation of high molecular weight DNA from forest topsoil for metagenomic analysis. *Asia-Pacific Journal of Molecular Biology and Biotechnology 16*, 35–41.

Park, S., Li, Y., 2012. Evaluation of methane production and macronutrient degradation in the anaerobic co-digestion of algae biomass residue and lipid waste. *Bioresource Technology 111*, 42–48.

Pelletier, E., Kreimeyer, A., Bocs, S., Rouy, Z., Gyapay, G., Chouari, R., Rivière, D., Ganesan, A., Daegelen, P., Sghir, A., Cohen, G.N., Médigue, C., Weissenbach, J., Le Paslier, D., 2008. "Candidatus Cloacamonas acidaminovorans": Genome sequence reconstruction provides a first glimpse of a new bacterial division. *Journal of Bacteriology 190*, 2572–2579.

Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England) 28*, 1420–8.

Pind, P.F., Angelidaki, I., Ahring, B.K., 2003. Dynamics of the anaerobic process: effects of volatile fatty acids. *Biotechnology and bioengineering 82*, 791–801.

Qiao, J.T., Qiu, Y.L., Yuan, X.Z., Shi, X.S., Xu, X.H., Guo, R.B., 2013. Molecular characterization of bacterial and archaeal communities in a full-scale anaerobic reactor treating corn straw. *Bioresource Technology 143*, 512–518.

Qiu, Y.L., Sekiguchi, Y., Imachi, H., Kamagata, Y., Tseng, I.C., Cheng, S.S., Ohashi, A., Harada, H., 2004. Identification and Isolation of Anaerobic, Syntrophic Phthalate Isomer-Degrading Microbes from Methanogenic Sludges Treating Wastewater from Terephthalate Manufacturing. *Applied and Environmental Microbiology 70*, 1617–1626.

Rajagopal, R., Massé, D.I., Singh, G., 2013. A critical review on inhibition of anaerobic digestion process by excess ammonia. *Bioresource Technology 143*, 632–641.

Razaviarani, V., Buchanan, I.D., 2015. Anaerobic co-digestion of biodiesel waste glycerin with municipal wastewater sludge: Microbial community structure dynamics and reactor performance. *Bioresource Technology 182*, 8–17.

Razaviarani, V., Buchanan, I.D., 2014. Reactor performance and microbial community dynamics during anaerobic co-digestion of municipal wastewater sludge with restaurant grease waste at steady state and overloading stages. *Bioresource Technology 172*, 232–240.

Reddy, R.M., Mohammed, M.H., Mande, S.S., 2014. MetaCAA: A clustering-aided methodology for efficient assembly of metagenomic datasets. *Genomics 103*, 161–168.

Refai, S., Berger, S., Wassmann, K., Deppenmeier, U., 2014. Quantification of methanogenic heterodisulfide reductase activity in biogas sludge. *Journal of Biotechnology 180*, 66–69.

Regueiro, L., Carballa, M., Lema, J.M., 2014. Outlining microbial community dynamics during temperature drop and subsequent recovery period in anaerobic co-digestion systems. *Journal of biotechnology 192PA*, 179–186.

Rivière, D., Desvignes, V., Pelletier, E., Chaussonnerie, S., Guermazi, S., Weissenbach, J., Li, T., Camacho, P., Sghir, A., 2009. Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *The ISME journal 3*, 700–14.

Schlüter, A., Bekel, T., Diaz, N.N., Dondrup, M., Eichenlaub, R., Gartemann, K.H., Krahn, I., Krause, L., Krömeke, H., Kruse, O., Mussgnug, J.H., Neuweger, H., Niehaus, K., Pühler, A., Runte, K.J., Szczepanowski, R., Tauch, A., Tilker, A., Viehöver, P., Goesmann, A., 2008. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of Biotechnology 136*, 77–90.

Scholz, M., Lo, C.-C., Chain, P.S.G., 2014. Improved Assemblies Using a Source-Agnostic Pipeline for MetaGenomic Assembly by Merging (MeGAMerge) of Contigs. *Scientific Reports 4*, 6480.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics 30*, 2068–2069.

Siles López, J.Á., Martín Santos, M. de los Á., Chica Pérez, A.F., Martín Martín, A., 2009. Anaerobic digestion of glycerol derived from biodiesel manufacturing. *Bioresource Technology 100*, 5609–5615.

Smith, A.L., Stadler, L.B., Love, N.G., Skerlos, S.J., Raskin, L., 2012. Perspectives on anaerobic membrane bioreactor treatment of domestic wastewater : A critical review. *Bioresource Technology 122*, 149–159.

Solomon, K. V., Haitjema, C.H., Thompson, D. a., O'Malley, M. a., 2014. Extracting data from the muck: Deriving biological insight from complex microbial communities and non-model organisms with next generation sequencing. *Current Opinion in Biotechnology 28*, 103–110.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., Pop, M., 2007. Minimus: a fast, lightweight genome assembler. *BMC bioinformatics 8*, 64.

Sousa, D.Z., Salvador, A.F., Ramos, J., Guedes, A.P., Barbosa, S., Stams, A.J.M., Alves, M.M., Pereira, M.A., 2013. Activity and viability of methanogens in anaerobic digestion of unsaturated and saturated long-chain fatty acids. *Applied and Environmental Microbiology 79*, 4239–4245.

Souza, R.C., Cantão, M.E., Vasconcelos, A.T.R., Nogueira, M.A., Hungria, M., 2013. Soil metagenomics reveals differences under conventional and no-tillage with crop rotation or succession. *Applied Soil Ecology 72*, 49–61.

St-Pierre, B., Wright, A.-D.G., 2014. Comparative metagenomic analysis of bacterial populations in three full-scale mesophilic anaerobic manure digesters. *Applied Microbiology and Biotechnology 98*, 2709–2717.

Stams, A.J.M., Plugge, C.M., 2009. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nature reviews. Microbiology 7*, 568–577.

Stegen, J.C., Lin, X., Konopka, A.E., Fredrickson, J.K., 2012. Stochastic and deterministic assembly processes in subsurface microbial communities. *The ISME Journal 6*, 1653–1664.

Sundberg, C., Al-Soud, W. a., Larsson, M., Alm, E., Yekta, S.S., Svensson, B.H., Sørensen, S.J., Karlsson, A., 2013. 454 Pyrosequencing Analyses of Bacterial and Archaeal Richness in 21 Full-Scale Biogas Digesters. *FEMS Microbiology Ecology 85*, 612–626.

Tian, Z., Zhang, Y., Li, Y., Chi, Y., Yang, M., 2015. Rapid establishment of thermophilic anaerobic microbial community during the one-step startup of thermophilic anaerobic digestion from a mesophilic digester. *Water Research 69*, 9–19.

Traversi, D., Villa, S., Lorenzi, E., Degan, R., Gilli, G., 2012. Application of a real-time qPCR method to measure the methanogen concentration during anaerobic digestion as an indicator of biogas production capacity. *Journal of Environmental Management 111*, 173–177.

Urich, T., Lanzén, A., Qi, J., Huson, D.H., Schleper, C., Schuster, S.C., 2008. Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. *PLoS ONE 3*, e2527.

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Thil Smith, A.A., Weiman, M., Medigue, C., 2013. MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Research 41*, D636–D647.

Vanwonterghem, I., Jensen, P.D., Dennis, P.G., Hugenholtz, P., Rabaey, K., Tyson, G.W., 2014a. Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *The ISME journal* 1–14.

Vanwonterghem, I., Jensen, P.D., Dennis, P.G., Hugenholtz, P., Rabaey, K., Tyson, G.W., 2014b. Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *The ISME Journal 8*, 2015–2028.

Vanwonterghem, I., Jensen, P.D., Ho, D.P., Batstone, D.J., Tyson, G.W., 2014c. Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Current Opinion in Biotechnology 27*, 55–64.

Wagner, A.O., Reitschuler, C., Illmer, P., 2014. Effect of different acetate:propionate ratios on the methanogenic community during thermophilic anaerobic digestion in batch experiments. *Biochemical Engineering Journal 90*, 154–161.

Walter, A., Knapp, B.A., Farbmacher, T., Ebner, C., Insam, H., Franke-Whittle, I.H., 2012. Searching for links in the biotic characteristics and abiotic parameters of nine different biogas plants. *Microbial Biotechnology 5*, 717–730.

Wang, J., Liu, H., Fu, B., Xu, K., Chen, J., 2013. Trophic link between syntrophic acetogens and homoacetogens during the anaerobic acidogenic fermentation of sewage sludge. *Biochemical Engineering Journal 70*, 1–8.

Wang, L.H., Wang, Q., Cai, W., Sun, X., 2012. Influence of mixing proportion on the solid-state anaerobic co-digestion of distiller's grains and food waste. *Biosystems Engineering 112*, 130–137.

Wang, Y., Zhang, Y., Wang, J., Meng, L., 2009. Effects of volatile fatty acid concentrations on methane yield and methanogenic bacteria. *Biomass and Bioenergy 33*, 848–853.

Whiteley, A.S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., Allcock, R., Donnell, A.O., 2012a. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *Journal of Microbiological Methods 91*, 80–88.

Wijekoon, K.C., Visvanathan, C., Abeynayaka, A., 2011. Effect of organic loading rate on VFA production, organic matter removal and microbial activity of a two-stage thermophilic anaerobic membrane bioreactor. *Bioresource Technology 102*, 5353–5360.

Wilkins, D., Lu, X.-Y., Shen, Z., Chen, J., Lee, P.K.H., 2015. Pyrosequencing of mcrA and Archaeal 16S rRNA Genes Reveals Diversity and Substrate Preferences of Methanogen Communities in Anaerobic Digesters. *Applied and Environmental Microbiology 81*, 604–613.

Wirth, R., Kovács, E., Maróti, G., Bagi, Z., Rákhely, G., Kovács, K.L., 2012. Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. *Biotechnology for Biofuels 5*, 41.

Xiao, K.K., Guo, C.H., Zhou, Y., Maspolim, Y., Wang, J.Y., Ng, W.J., 2013. Acetic acid inhibition on methanogens in a two-phase anaerobic process. *Biochemical Engineering Journal 75*, 1–7.

Xiao, X., Huang, Z., Ruan, W., Yan, L., Miao, H., Ren, H., Zhao, M., 2015. Evaluation and characterization during the anaerobic digestion of high-strength kitchen waste slurry via a pilot-scale anaerobic membrane bioreactor. *Bioresource Technology 193*, 234–242.

Xue, Y., Liu, H., Chen, S., Dichtl, N., Dai, X., Li, N., 2015. Effects of thermal hydrolysis on organic matter solubilization and anaerobic digestion of high solid sludge. *Chemical Engineering Journal 264*, 174–180.

Yang, Y., Yu, K., Xia, Y., Lau, F.T.K., Tang, D.T.W., Fung, W.C., Fang, H.H.P., Zhang, T., 2014. Metagenomic analysis of sludge from full-scale anaerobic digesters operated in municipal wastewater treatment plants. *Applied Microbiology and Biotechnology 98*, 5709–5718.

Yirong, C., Heaven, S., Banks, C.J., 2014. Effect of a Trace Element Addition Strategy on Volatile Fatty Acid Accumulation in Thermophilic Anaerobic Digestion of Food Waste. *Waste and Biomass Valorization 6*, 1–12.

Yu, D., Kurola, J.M., Lähde, K., Kymäläinen, M., Sinkkonen, a., Romantschuk, M., 2014. Biogas production and methanogenic archaeal community in mesophilic and thermophilic anaerobic co-digestion processes. *Journal of Environmental Management 143*, 5460.

Zandvoort, M.H., Geerts, R., Lettinga, G., Lens, P.N.L., 2003. Methanol degradation in granular sludge reactors at sub-optimal metal concentrations: Role of iron, nickel and cobalt. *Enzyme and Microbial Technology 33*, 190–198.

Zhang, C., Xiao, G., Peng, L., Su, H., Tan, T., 2013. The anaerobic co-digestion of food waste and cattle manure. *Bioresource Technology 129*, 170–176.

Ziganshin, A.M., Liebetrau, J., Pröter, J., Kleinsteuber, S., 2013. Microbial community structure and dynamics during anaerobic digestion of various agricultural waste materials. *Applied Microbiology and Biotechnology 97*, 5161–5174.

Ziganshin, A.M., Schmidt, T., Scholwin, F., Il'inskaya, O.N., Harms, H., Kleinsteuber, S., 2011. Bacteria and archaea involved in anaerobic digestion of distillers grains with solubles. *Applied microbiology and biotechnology 89*, 2039–52.