DEPARTMENT OF CHEMISTRY AND BIOMOLECULAR SCIENCES

MACQUARIE UNIVERSITY

10th October 2014

METAGENOMIC AND MICROSCOPIC INVESTIGATION OF A COAL DEGRADING MICROBIAL CONSORTIUM

SILAS VICK

SUPERVISOR: PROFESSOR IAN PAULSEN

Contents

Disclaimer

Referencing and citations are written in the form laid out by The International Society of Microbial Ecology Journal.

I wish to declare the following assistance in the research described in this report:

- MP09 coal formation water chemistry was provided by Dr David Midgley (CSIRO).
- 16S rarefactions were generated by Dr David Midgley (CSIRO).
- Amplification, library preparation, and 16S rDNA amplicon sequencing using the Illumina MiSeq platform was performed by Mr DNA (Molecular Research LP).
- Assembly of metagenomic reads using IDBA-UD was performed by Dr Liam Elbourne.
- Kraken and MetaPhlAn programs were run by Dr Liam Elbourne.
- Library preparation and sequencing of metagenomic DNA was performed by The Ramaciotti Centre for Genomics (University of New South Wales).
- Sanger sequencing of clone sequences was performed by the Australian Genome Research Facility (AGRF).

All other research presented in this report is my own work.

Acknowledgements

I would like to thank Professor Ian Paulsen for the opportunity to undertake this research as well as his time, support and supervision during this project, Dr. Sasha Tetu, Dr David Midgley and Dr. Liam Elbourne for their time, support and assistance and my colleagues within the Paulsen research group and at CSIRO North Ryde for welcoming me into their research groups and their help. I would also like to thank Debra Birch and Nicole Vella from the Macquarie University Microscopy department for their training and expertise. Finally I would like to thank my MRes cohort, friends and family for helping support me through this year.

Abstract

Coal seam gas (CSG) is of increasing interest internationally due to its potential as a relatively clean bridging fuel for energy generation. CSG can originate from either thermogenic or biogenic methanogenesis. A growing awareness that biogenic processes may contribute significantly to CSG reserves worldwide has fuelled interest in the microbial communities performing this transformation. In this study a microbial consortium sourced from Australian coal seam formation water, capable of methanogenesis using coal as the sole carbon source has been examined. A combination of 16S rRNA amplicon surveys, metagenomic sequencing and scanning electron microscopy was performed on the community over a time-course of inoculation to maturation on fresh coal surfaces. This research has identified key microbial components of the community as it develops over time and used the latest informatics tools to match genetic functions, such as the three methanogenesis pathways, to these community members. Microscopic observation coupled with DNA sequencing experiments of inoculated coal surfaces over a time-course has provided the first insights into the distinct stages of the coal colonisation process, including the first observation of a biofilm attached community and the first reports of specific taxa including *Geobacter spp.* involved in the early colonisation of coal surfaces.

List of Figures and Tables

Figures

Tables

Abbreviations

GWP	global warming potential
CSG	coal seam gas
CSM	coal seam methane
CBM	coal bed methane
NSW	New South Wales
MDA	multiple displacement amplification
CDS	coding sequence
PCR	polymerase chain reaction
IDBA-UD	iterative De Bruijn Graph De Novo Assembler for Short Reads Sequencing data with Highly Uneven Sequencing Depth
Contigs	contiguous sequence
N50	The length for which the sum of all contigs of that length or greater makes up 50% of the total sum of the lengths of all the contigs.
PROKKA	prokaryotic annotation
KEGG	Kyoto Encyclopedia of Genes and Genomes
TIGRFAM	The Institute for Genome Research families
MG-RAST	Metagenome Rapid Annotation using Subsystem Technology
IMG/M	Integrated Microbial Genomes with Microbiome Samples
CAMERA	Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis
OTU	operational taxonomic unit
RDP	ribosomal database project
dsDNA	double stranded DNA
rRNA	ribosomal RNA
MetaPHlAn	Metagenomic Phylogenetic Analysis
KAAS	KEGG automatic annotation server
HMM	hidden markov model
BWA	Burrows-Wheeler Aligner
BLAST	Basic Local Alignment Search Tool
NCBI	National Center for Biotechnology Information
SEM	scanning electron microscopy
TEM	transmission electron microscopy
HMDS	hexamethyldisilazane

LB	Luria-Bertani medium
SOC	super optimal broth with catabolite repression medium
AGRF	Australian Genome Research Facility
Mcrα	Methyl-coenzyme M reductase alpha subunit
Mcrβ	Methyl-coenzyme M reductase beta subunit
mttB	Trimethylamine methyltransferase
MtmB	monomethylamine:corrinoid methyltransferase
MtbC	dimethylamine:corrinoid methyltransferase
MtbB	Dimethylamine methyltransferase
CdhD	acetyl-CoA decarbonylase/synthase complex subunit delta
MTmer	F_{420}-dependent N5, N10-methylenetetrahydromethanopterin reductase
L	litres
ml	millilitres
µl	microlitres
mm	millimetres
µm	micrometres
µg	micrograms
ng	nanograms
M	molar
µM	micromolar
s	seconds
min	minutes
U	enzyme units
bp	base pairs
Rpm	rotations per minute
xg	times gravity

Chapter 1. Introduction and background

1.1.1 Methane, energy and the environment

Understanding the various aspects of global carbon cycling is an important aim of contemporary scientific research, particularly considering that the movement of carbon between the biosphere, geosphere and atmosphere has important climatic and environmental effects. The biggest source of carbon transfer into and out of the atmosphere is biogenic with carbon being assimilated into the biosphere primarily through photosynthesis and cycled back into the atmosphere through respiration. While this cycling is ongoing and largely balancing, biomass can also be removed from the cycle for much longer periods of time when it becomes stored underground and enters the geosphere. This trapped biomass is eventually degraded through geochemical processes to become fossil fuels. This stored carbon can then enter back into the biosphere and atmosphere in several ways: the first is through anthropogenic burning of fossil fuels for energy, releasing the carbon back into the atmosphere as CO_2, the second is through thermogenic methane generation in which biomass is slowly broken down under pressure and high temperature to methane underground over thousands of years and the third is through biogenic methanogenesis, a process whereby buried organic matter is anaerobically converted to methane by microbes.

This process of microbial methane generation known as methanogenesis is an important aspect of global carbon cycling and represents the main way in which biomass trapped in anaerobic environments can be transferred back into the atmosphere. Globally methanogenesis returns around 1.65 % of the carbon fixed by photosynthetic organisms to the atmosphere each year and accounts for 80% of the methane present in the atmosphere (Hedderich and Whitman, 2006). Roughly 10 % of this methane is derived from fossil fuel sources which have been trapped in the geosphere for long periods of time making this an important additive contribution to greenhouse gasses in the atmosphere (Hedderich and Whitman, 2013). In addition methanogenesis is an important process due to the high global warming potential (GWP) of methane which has twenty one times the GWP of CO_2 over 100 years, meaning even relatively small contributions in terms of total carbon release can have significant climatic effects (Forster et al, 2007).

In addition to the climatic and environmental importance of methanogenesis it is also of considerable industrial interest as a bridge fuel for energy generation. The use of methane derived from fossil fuels has benefits over traditional fossil fuel sources due to both its higher calorific value per unit of carbon compared to more complex fossil fuels and the relative

cleanness of its combustion emissions. A major source of fossil fuel derived methane originates from coal seams and is synonymously termed coal seam gas (CSG), coal seam methane (CSM) and coal bed methane (CBM) although here it will be referred to as coal seam gas (CSG). As coal seams make up 85% of the carbon stored as fossil fuels and extraction and remediation of coal mines is often very costly, CSG is being investigated as an alternative cleaner, lower impact energy source which can be utilised after simple retrofitting of existing power plants (Falkowski et al, 2000).

1.1.2 The coal seam as a methanogenic environment

Coal originates from terrestrial plant material which has accumulated in anoxic sediments. This differs from other fossil fuels such as oil and natural gas which are the product of aquatic microbial deposition. The starting biological material is of importance in the fossil fuels they create as plant based material contains high quantities of structural polymers such as cellulose and lignin which are more recalcitrant to degradation than microbial cellular debris. Within coal there is also a certain amount of heterogeneity due to the types of plant, algal and fungal material from which it is derived, as well as the different maturation processes which it has undergone. In Australia the major coal types are the mature bituminous coals found through the Bowen, Surat and Sydney basins deposited in the Permian period when the cycads, conifers and lower vascular plants like the lycophytes were predominant and the less mature brown coals found in southern New South Wales (NSW) and Victoria deposited in the Neogene period after the appearance of the angiosperm plants (Ahmed and Smith, 2001, Papendick et al, 2011, Strąpoć et al, 2011).

Initially it was thought that CSG was purely of thermogenic origin but the discovery of biogenic methane production in coal seams came after the detection of significant methane deposits in coal seams thought too immature for thermogenic methane production. This observation was confirmed by analysis of the methane which showed C^{13}:C^{12} isotopic ratios characteristic of biogenic methanogenesis and a higher ratio of methane (C_1 alkane) to longer chain alkanes (C_2 & C_3) which supported a biogenic methane origin (Strąpoć et al, 2011).

1.2 Microbial population structuring and biochemistry in methanogenic environments

Methanogenic communities are complex assemblages of bacterial and archaeal organisms living symbiotically and relying on several mutualistic interspecific trophic interactions. The general community structure thought to be common to all methanogenic environments includes three broad metabolic groups: the primary fermenters, the secondary fermenters and the methanogens (Figure 1.1). These methanogenic communities can only occur in strictly anoxic environments where inorganic electron accepters are depleted. In the presence of

alternative electron acceptors such nitrate, nitrite or sulphate these molecules will be preferentially used by other bacterial groups including the sulphate reducing bacteria which will out-compete the methanogens (Elferink Oude et al, 1994). Because of the absence of O_2 and inorganic electron accepters, methanogenic communities are only able to release a small portion of the potential chemical energy in the organic matter they degrade and are thought to survive at the limits of what is energetically possible in a biological system. As an example, degradation of hexose to its final product, methane, in a methanogenic environment releases only 15% of the energy which could be released under oxic conditions. Because of this, methanogenic communities rely on various mutualistic metabolic interactions to utilise the small amount of energy which can be obtained in these environments (Stams, 1994).

1.2.1 The primary fermenters

Metabolically the first group of organisms in a methanogenic environment are the primary fermenters, they are responsible for carrying out the first stages of anaerobic metabolism in the community. Extracellular enzymes are excreted by these organisms to break down large organic polymers to simpler monomers. These monomers are then transported into the cells and fermented to simpler reduced products such as fatty acids, alcohols, lactate and succinate as well as undergoing partial degradation directly to H_2, CO_2 and acetate. This group of organisms is probably the most metabolically diverse group among different methanogenic environments and the species present and enzymes they produce are likely to be highly dependent on the organic polymers present. Anaerobic primary fermenters present in the gastrointestinal tracts of animals are specialised for the diet of the host organism as well as the metabolic capability of the host whereas those present in wetlands, peat bogs and other sediments would be required to degrade a different set of organic polymers probably including higher concentrations of cellulose and lignin, the highly aromatic polymer found in woody plant material (Morgavi et al, 2010). In contrast, the primary fermentative group in fossil fuel degrading environments such as in coal seams likely includes a group of organisms with the ability to degrade the complex aromatic and aliphatic structures which dominate mature coals. Although the enzymes used for this process remain unknown it is thought that some aspects of coal degradation may show similarity to that of anaerobic lignin degradation due to the chemical similarities between the materials (Strąpoć et al, 2011).

Lignin is usually very recalcitrant to degradation due to its aromatic ring structure and usually requires oxygenases to cause ring cleavage (Fuchs et al, 2011). As atmospheric oxygen is not present in these anaerobic environments lignin degradation is much more difficult and only a handful of specialised enzymatic strategies are known for anaerobic ring cleavage. This

suggests that primary fermenters in lignin rich environments and potentially coal are likely to include members with these specialised metabolic pathways as well as genes involved in aromatic and aliphatic hydrocarbon degradation (Berdugo-Clavijo et al, 2012, Chang et al, 2006, Fuchs et al, 2011, Hirschler-Réa et al, 2012, Strąpoć et al, 2011, Zengler et al, 1999, Zhang et al, 2012).

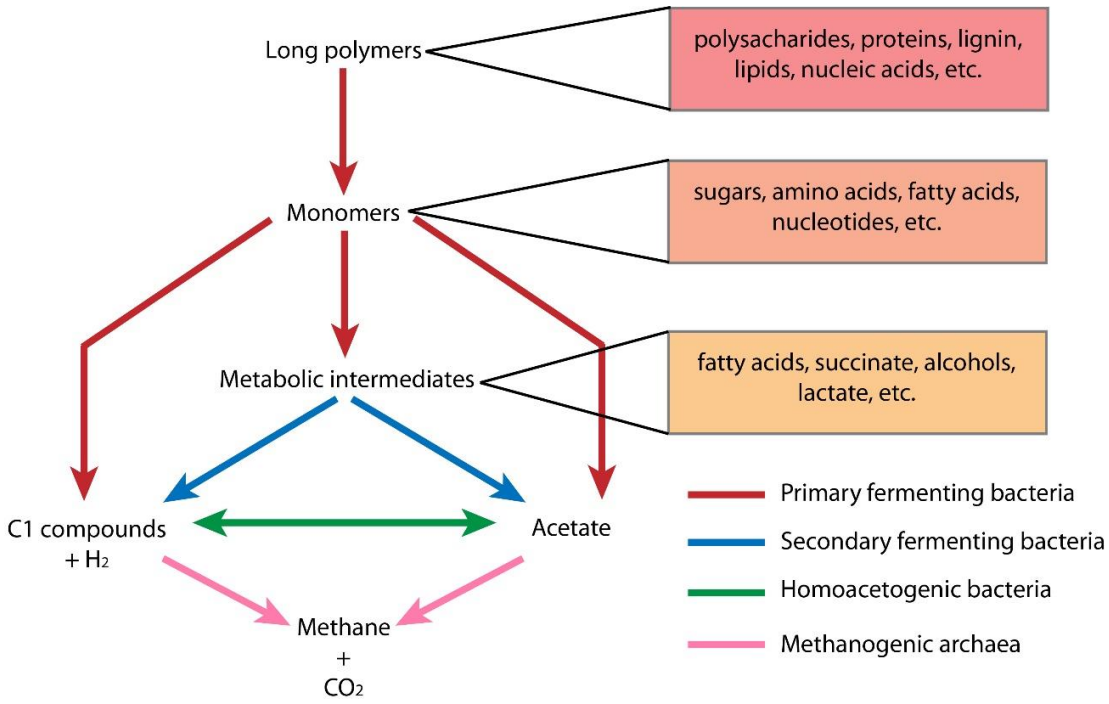


Figure 1.1 Illustration of the basic trophic community structure observed in methanogenic communities.

1.2.2 The secondary fermenters

Once the initial stages of organic polymer fermentation have occurred the secondary fermenters can utilise the reduced carbon compounds produced by the primary fermenters for secondary fermentation to simple compounds such as CO_2, H_2, formate and acetate. These fermentation reactions however are energetically unfavourable to perform, having a positive Gibbs free energy change under standard conditions. For these reactions to occur the products must be kept at a low concentration by constant removal from the environment. To achieve this secondary fermenters form an obligatory syntrophic relationship with the acetogenic and hydrogenotrophic methanogens. The H_2 and acetate formed by the secondary fermenters is kept at a concentration (or partial pressure in the case of H_2) low enough for the secondary fermentation reactions to proceed but high enough for the methanogenic reactions to also proceed. There is also some evidence that the secondary fermenters can form direct electron

transport with methanogens through specialised electrically conductive pili termed “microbial nanowires” allowing for faster more direct electron transfer which does not rely on transport across the cell membranes and diffusion across the space between cells (Gorby et al, 2006, Morita et al, 2011, Reguera et al, 2005, Stams, 1994, Stams and Plugge, 2009).

While the homoacetogenic bacteria (Figure 1.1) do not play as clear a role in syntrophic metabolism as the methanogens and primary and secondary fermenters, they appear to have important roles under different physicochemical conditions. Under normal conditions H_2 is consumed by the hydrogenotrophic methanogens with higher energy yields than can be achieved by the homoacetogens leading to out-competition of the homoacetogens. However, it has been observed that under acidic and low temperature conditions the homoacetogenic utilisation of H_2 appears to be more favourable than hydrogenotrophic methanogenesis and the homoacetogens outcompete the hydrogenotrophic methanogens (Conrad and Wetter, 1990, Kotsyurbenko et al, 2001, Phelps and Zeikus, 1984, Schink, 1997). Homoacetogenic bacteria have also been shown to carry out the acetogenic reaction in reverse depending on substrate concentrations (and temperature). This suggests that they can play a role in influencing which methanogenic pathway predominates, acetogenic methanogenesis or hydrogenotrophic methanogenesis, by shuttling between the $C_1 + H_2$ and acetate carbon pools (Lee and Zinder, 1988, Schink, 1997).

1.2.3 The methanogens

The biological production of methane is limited to a subset of the archaeal domain and is only possible in strictly anoxic environments. Within the archaeal domain there are five well distinguished methanogenic orders, Methanobacteriales, Methanococcales, Methanomicrobiales, Methanosarcinales, and Methanopyrales, although recent culture independent microbial studies in rice paddies (Lueders et al, 2001) and rumen (Nicholson et al, 2007) have identified methanogens with sufficient phylogenetic differences to suggest the existence of further orders. The monophyletic grouping of these five orders within the Euryarchaeota, along with the biochemical complexity of methanogenesis and the common pathway components present across these orders suggests a single common origin of methanogenesis (Hedderich and Whitman, 2006). The unique nature of the methanogenic process has allowed these archaea to fill an ecological niche in a wide range of carbon rich and oxidant poor, anaerobic environments ranging from cold arctic through to high temperature and extreme high salinity environments (Franzmann et al, 1992, Franzmann et al, 1997, Hedderich and Whitman, 2006, Jones et al, 1983, Kurr et al, 1991).

Despite the phylogenetic diversity of the methanogens there are only three known methanogenic pathways; CO_2 reduction, methyl-group reduction and acetate reduction (Liu and Whitman, 2008). CO_2 reduction, performed by the hydrogenotrophic methanogens, is the most common pathway and is seen in all five of the methanogen orders. The reducing agent used is most commonly H_2 although formate, secondary alcohols and CO can also be used by some individuals (Daniels et al, 1977, O'brien et al, 1984). Methyl-group reduction, performed by the methylotrophic methanogens, involves reduction of methanol, methylated amines or methylated sulphides with electrons donated by simultaneous oxidation of other methyl-groups to CO_2. This method is unique to members of the order Methanosarcinales with the exception of methanol reduction to methane with H_2 as the electron donor, which can be performed by species of the genus Methanosphaera from the order Methanobacteriales. Thirdly is the acetate splitting or acetoclastic pathway which accounts for ~67% of the biogenic methane production on the planet. This pathway is known to be performed by the two genera Methanosarcina and Methanosaeta within the order Methanosarcinales although it is thought that there are some differences in the pathways utilised by the two groups as Methanosaeta is capable of performing this reaction at far lower acetate concentrations (Liu and Whitman, 2008). The phylogenetic diversity of the methanogens and the substrates they utilise are summarised in Figure 1.2.

1.2.4 Coal associated methanogenic communities in the literature

Over the past decade investigations have begun into the microbial communities associated with methanogenesis in coal seams leading to the production of biogenic coal seam gas (CSG) (Dawson et al, 2012, Fry et al, 2009, Green et al, 2008, Guo et al, 2012, Li et al, 2008, Midgley et al, 2010, Penner et al, 2010, Shimizu et al, 2007, Singh et al, 2011, Strapoc et al, 2008, Tang et al, 2012, Wei et al, 2013). These studies have employed both culture independent phylogenetic analysis such as 16S ribosomal RNA (rRNA) amplicon sequencing as a well as examinations of methanogenic enrichment cultures taken from coal seams or associated formation waters. These studies have reported a wide range of bacterial and archaeal taxa with differing taxonomic diversity and abundances associated with different sites around the world (Figure 1.3). It should be noted that these reported abundances are not necessarily directly comparable due to a variety of differences in experimental design between the various studies including the use of 16S rRNA clone libraries or amplicon

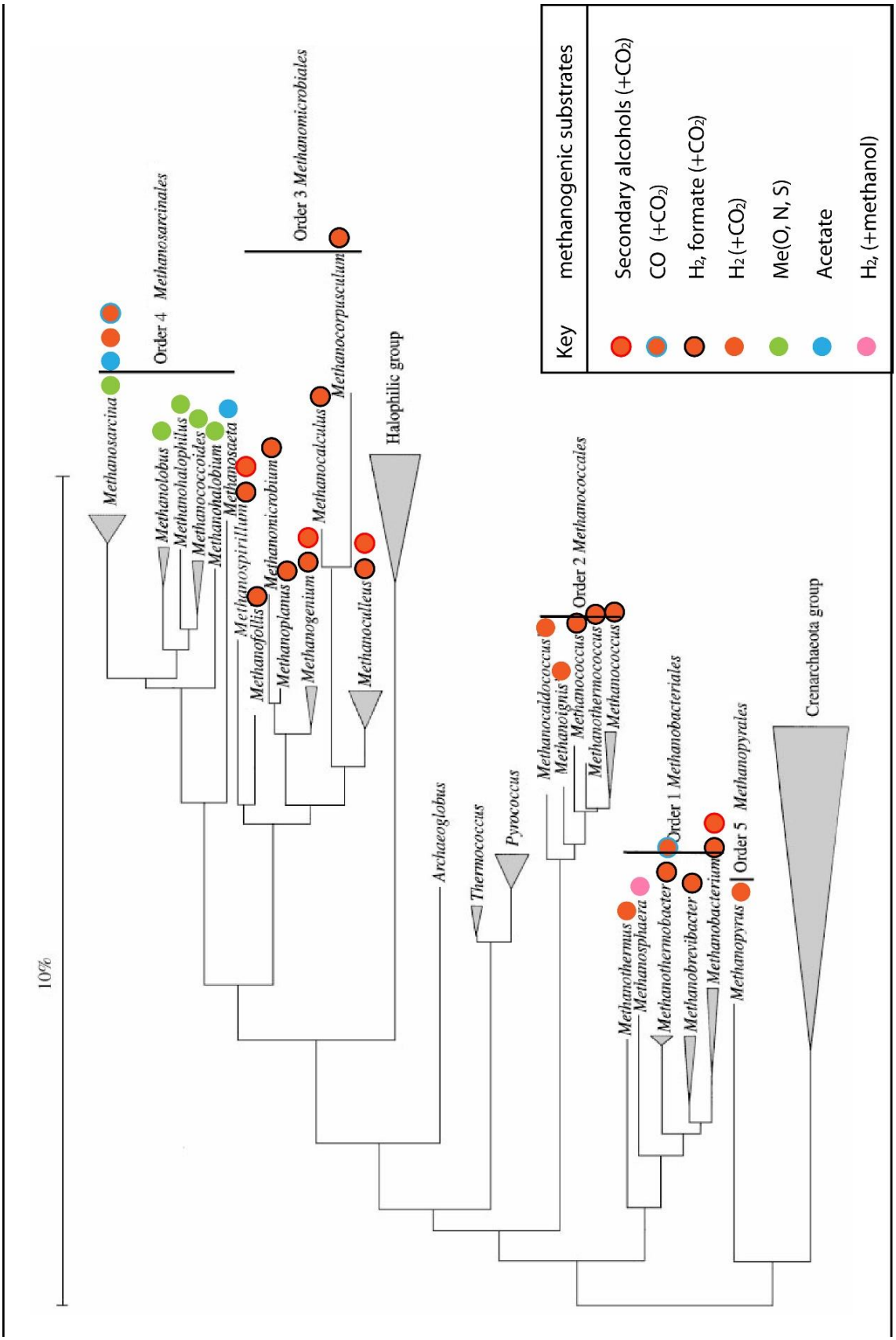


Figure 1.2 Phylogenetic tree based on 16S rRNA gene sequences depicting the methanogenic archaea and their close relatives, adapted from (Garcia et al, 2000). Substrates utilised by the various archaea for methanogenesis are indicated by a coloured circle next to the family or genus name.

sequencing, the use of different primers for 16S rRNA amplification which can impose taxa specific amplification biases, taxonomic assignments based on different regions of the 16S rRNA sequence, differences in sampling and sequencing coverage between the different studies and differences in bioinformatic analyses undertaken for taxonomic assignment. Despite these issues some clear trends in the reported microbial populations can be observed. The coal seam communities appear to contain very diverse prokaryotic groups spanning multiple phyla within the bacterial and archaeal domains. Members of the bacterial phylum Proteobacteria, particularly Alpha, Beta and Gamma, make up significant abundances (around 50 %) in all the reported studies except two in North America where the Firmicutes appear as the dominant group. In addition to the highly abundant members there are several taxonomic groups which are consistently seen in all or most of the studies which are likely to play key roles in the communities. These include the Bacteroidetes and Actinobacteria which are both found in low abundances, around 5%, in the majority of studies. It is difficult to comment on the functional roles in degrading coal to methane played by the various groups of bacteria observed due to the unique carbon source and the fact that previous studies have focussed solely on taxonomic identification. It has been observed though that some groups of organisms have defined roles in other methanogenic communities such as those in sediments and animal guts. The Bacteroidetes are known to anaerobically degrade biopolymers such as cellulose, protein and polysaccharides in sediment and gut environments and the Spirochaetes have been shown to degrade plant biopolymers in rumen environments and perform acetogenesis in the termite gut (Haack and Breznak, 1993, Leadbetter et al, 1999, Paster and Canale-Parola, 1982, Strapoc et al, 2008). In addition the Firmicutes have diverse roles but have been known to degrade amino acids as a sole carbon source suggesting their role as secondary fermenters (Rogosa, 1969, Strapoc et al, 2008).

In the archaeal domain there is typically less diversity seen in the individual coal seams, instead one or two methanogenic archaeal groups dominate. There does not appear to be any specific methanogenic group responsible for CSG generation in all coal seams and the methanogenic group present is likely due to chance inoculation combined with some physicochemical characteristic of the environments which lends itself to a particular methanogenic pathway. All three methanogenic pathways have been observed in coal seams through both direct observation of methane generation, implication from the methanogenic species present and C^{13}:C^{14} ratio observations (Papendick et al, 2011, Strąpoć et al, 2011).

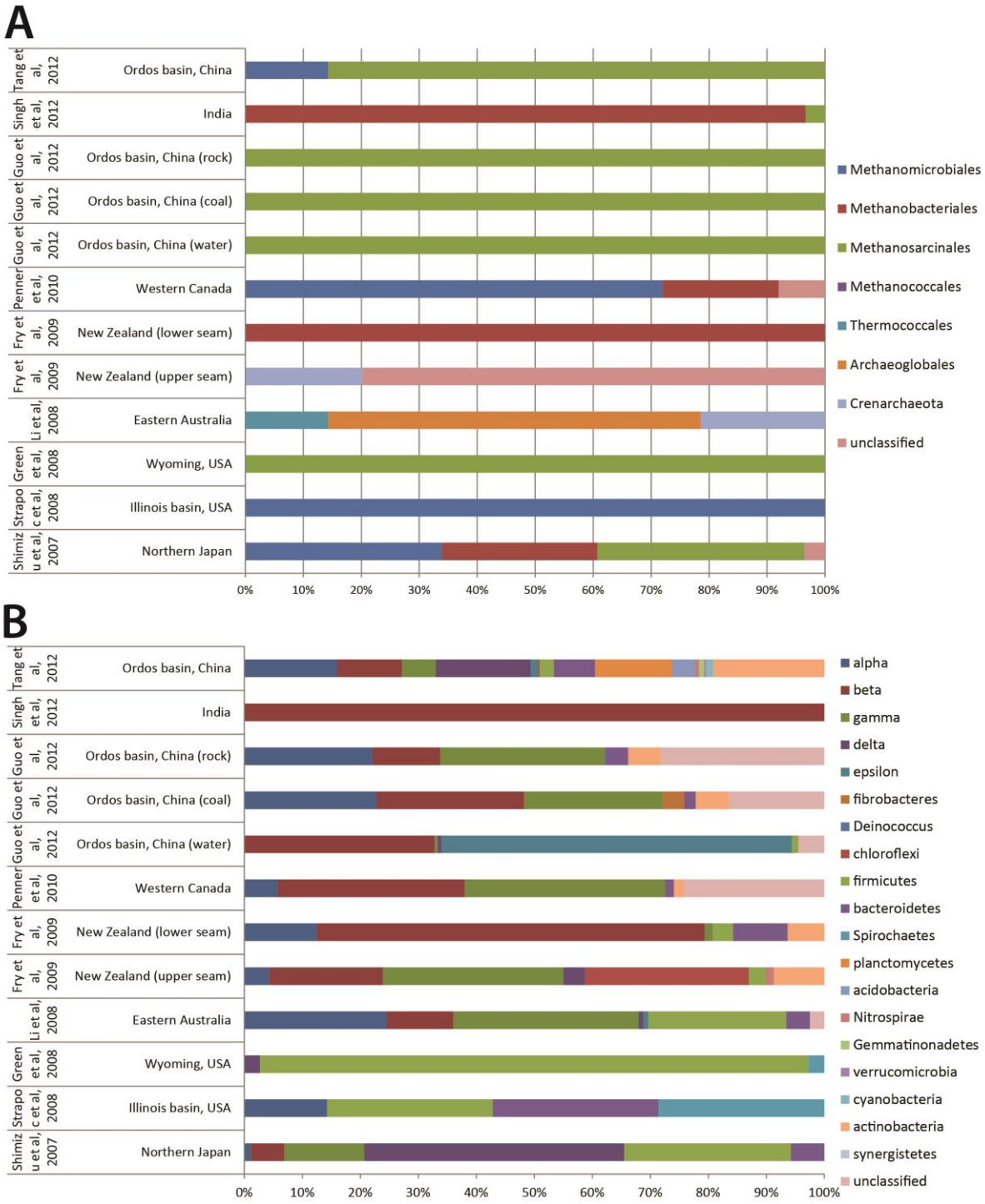


Figure 1.3 Reported taxonomic compositions of archaeal (a) and bacterial (b) taxa from previous phylogenetic studies utilising the 16S rRNA marker gene sequence. Composition is shown at the phylum level for bacteria with the exception of the Proteobacteria which are shown at the order level (b) and the order level for Archaea (a) (Fry et al, 2009, Green et al, 2008, Guo et al, 2012, Li et al, 2008, Midgley et al, 2010, Penner et al, 2010, Shimizu et al, 2007, Singh et al, 2011, Strapoc et al, 2008, Tang et al, 2012).

Another observation made in some of these studies is that the abundances of various taxa appear to vary significantly within the same coal seam depending on where sampling occurred. In (Guo et al, 2012) significant abundance differences can be seen between samples of the associated water, rock and coal components of the same coal seam and in (Fry et al, 2009) a similar observation was made when looking at the communities at different depths. This suggests that community composition depends upon micro-environments and is not homogenous throughout the coal seam. This could be due to the heterogeneous nature of coal and coal seams with their different physical and chemical compositions and suggests that community members inhabit both trophic and physical niches within the coal seam environment (Fry et al, 2009, Guo et al, 2012).

1.3 Aims, scope and direction of the current work

1.3.1 The sample community

Studying coal seam associated microbes carries with it many practical challenges due to the physical isolation of the communities up to several hundred meters subsurface, the low cell densities in these communities, the difficult chemical nature of coal as a substrate when measuring cell density, observing fluorescence and purifying biomolecules and the long anaerobic incubation periods needed to culture these communities. To study these communities, culture independent approaches have been most commonly utilised but now that several coal seam associated consortia have been isolated capable of methanogenesis from coal as the sole carbon source, approaches utilising cultured consortiums are now possible (Fry et al, 2009, Guo et al, 2012, Midgley et al, 2010, Papendick et al, 2011). While these cultures are unlikely to contain the full spectrum of members found in environmental samples or have equivalent abundances, the ability to culture coal degrading communities capable of methanogenesis in the laboratory has several benefits over environmental sampling. One advantage is that while samples can be taken from an environment capable of methanogenesis from coal, the sampled subset may not include all the members necessary for this capability due to the different physical niches in the environment. Other benefits include improvement in reproducibility of community measurements in a closed system and the ability to perform manipulative experiments on the community with replicates and controls (Beckmann et al, 2011). This study utilised a cultured coal seam associated consortium originally isolated from formation water from a CSG mine in Menangle Park, NSW, Australia grown on coal of a similar maturity from the Rangal coal measure within the Bowen Basin, QLD, Australia (Beeston, 1986, Faiz et al, 2007). This consortium has been shown to produce methane using

coal as the sole carbon source making it an ideal community for exploring this (Midgley et al., unpublished data).

1.3.2 Surveying taxonomic abundances and functional capabilities of a microbial community

Amplicon sequencing of taxonomically conserved marker genes such as the 16S rRNA gene is a technique used for studying many communities in microbial ecology, including several coal seam communities (Guo et al, 2012, Wei et al, 2013). It is a particularly useful technique in environmental samples where very little DNA can be extracted and has the benefit of being relatively low cost due to the short DNA lengths sequenced. However, while it is an informative starting point for characterising the relative abundances of taxa in a microbial community it suffers from limitations in terms of primer biases and 16S rRNA gene copy number which can bias taxonomic abundance measurements. While methods are available for normalising for 16S rRNA copy number they rely on good genomic knowledge of the taxonomic groups being studied, which is likely not the case for coal seam communities (Kembel et al, 2012).

Another powerful emerging tool for studying microbial communities is metagenomics. This involves extraction of DNA from all organisms in a sample, direct sequencing and analysis of the genetic information of the entire community as a whole. This technique can be applied to both cultured consortia or directly to *in situ* environmental samples and provides information regarding both taxonomic and functional composition of a community. Metagenomic sequencing also bypasses the issues of primer biases and 16S rRNA gene copy number as taxonomic assignments and abundances can be calculated based on a range of taxonomically specific or informative genes with varying copy numbers using software such as PhyloSift, MetaPhlAn or Kraken (Darling et al, 2014, Segata et al, 2012, Wood and Salzberg, 2014).

While metagenomic sequencing provides more information and can overcome some of the limitations of amplicon surveys it requires a much higher quantity of input DNA not often found in environmental samples with low cell densities. One technique which can enable metagenomic analysis where very low quantities of DNA are present is multiple displacement amplification (MDA). MDA uses random hexanucleotide primed rolling circle amplification with a bacteriophage ϕ29 DNA polymerase to produce randomly amplified products of on average 12 Kb (Dean et al, 2002). MDA kits are commercially available and relatively inexpensive, making this a promising option for enabling metagenomic studies of coal seam communities. However, it should be noted that observed sequence abundance may not be

representative of actual sequence abundances, particularly if used on small metagenomes due to stochastic amplification biases of the MDA reaction (Yilmaz et al, 2010).

1.3.3 Bioinformatic analysis of metagenomic datasets

Bioinformatic tools for metagenomic analyses are constantly being developed as the metagenomic field expands but in general involve a common series of steps; assembly of sequence reads, taxonomic binning of reads or contigs and annotation of the genetic elements identified. Other steps such as phylogenetic tree construction from taxonomic assignments and single genome assemblies can also be performed depending on the aims of the study and the quality of the metagenome.

The first stage of the metagenomic data analysis is assembly, here raw reads are combined into longer consensus sequences termed contigs. Assembly can be performed by two broad methods termed reference-based assembly and *de novo* assembly. Reference based assemblers rely on assembly by alignment with known genomes and are not well suited for novel organisms or metagenomic datasets. *De novo* assemblers use alignment between reads to assemble contigs and are better suited to metagenomes, especially new *de novo* assemblers such as IDBA-UD (Peng et al, 2012) and SPAdes (Bankevich et al, 2012) which take into account assumptions of fragmented sequences at different depths of coverage to make them better suited for assembly of metagenomic reads (Thomas et al, 2012).

Many tools are available for taxonomic binning of metagenomic sequences, however most rely on variations of two techniques. The first, conserved nucleotide composition binning, assumes distinct taxonomic groups have conserved characteristics throughout their genome such as GC content and kmer profiles upon which binning can be based. The second is based on sequence alignments of reads with annotated nucleotide databases to match reads or contigs to sequences of known organisms (Teeling and Glöckner, 2012, Thomas et al, 2012). One recent novel taxonomic binning method, termed GroopM, bins contigs based on coverage between multiple sample replicates with taxonomy-specific abundance changes (Albertsen et al, 2013, Imelfort et al, 2014).

Annotation of contigs is another important component of a metagenomic workflow in which genes are identified and assigned function. This process has two steps, the first is feature prediction which involves the recognition of genes or other elements such as RNAs, the second is functional annotation of the predicted coding sequences (CDSs), usually done via sequence alignments to annotated online protein or nucleotide databases. The identified CDSs can also be annotated by categorisation based on hierarchical gene/protein families for biological function, cellular localisation, metabolic pathways or structural homology.

Pipelines such as PROKKA (Seemann, 2014) will perform all of these steps on contig sequences and provide annotation from sequence alignments as well as hierarchical categorisations such as KEGG pathways (Kanehisa et al, 2004), Pfam (Punta et al, 2012) and TIGRFAM (Selengut et al, 2007) where available. Online automated annotation pipelines and metagenomic databases such as MG-RAST (Glass et al, 2010), IMG/M (Markowitz et al, 2012) and CAMERA (Sun et al, 2011) are also available for both annotation and analysis of reads and contigs, however these fully automated pipelines may generate spurious results can without assessment of assignment quality at each stage (Teeling and Glöckner, 2012, Thomas et al, 2012).

1.3.4 Aims and scope of the present study

The identification of the microbes responsible for the various stages of degradation of coal to methane and elucidation of the biochemical pathways used by these organisms is an important, so far elusive step in understanding microbial generation of CSG. At this time there has been only one metagenomic sequence published for a coal seam community with a preliminary functional and taxonomic analysis performed. Therefore there is considerable potential for detailed metagenomic studies to lead to novel discoveries regarding coal seam communities (Ghosh et al, 2014). The aim of this study is to characterise a coal degrading methanogenic microcosm MP09 (described in 1.3.1) in terms of its taxonomic members and functional genetic and enzymatic components. To achieve this, 16S rRNA amplicon surveys and metagenomic analysis of the mature MP09 community will be performed, in addition to DNA sequence and microscopic analysis of the community over a time-course as it colonises and matures on a coal surface (experimental flow illustrated in Figure 1.4). From these experiments we hope to: 1. Develop a better understanding of the necessary functional components of a coal degrading methanogenic community, 2. provide insight into which taxonomic groups are responsible for these functional components and 3. determine how these taxonomic and functional groups develop over time from colonisation to maturation.

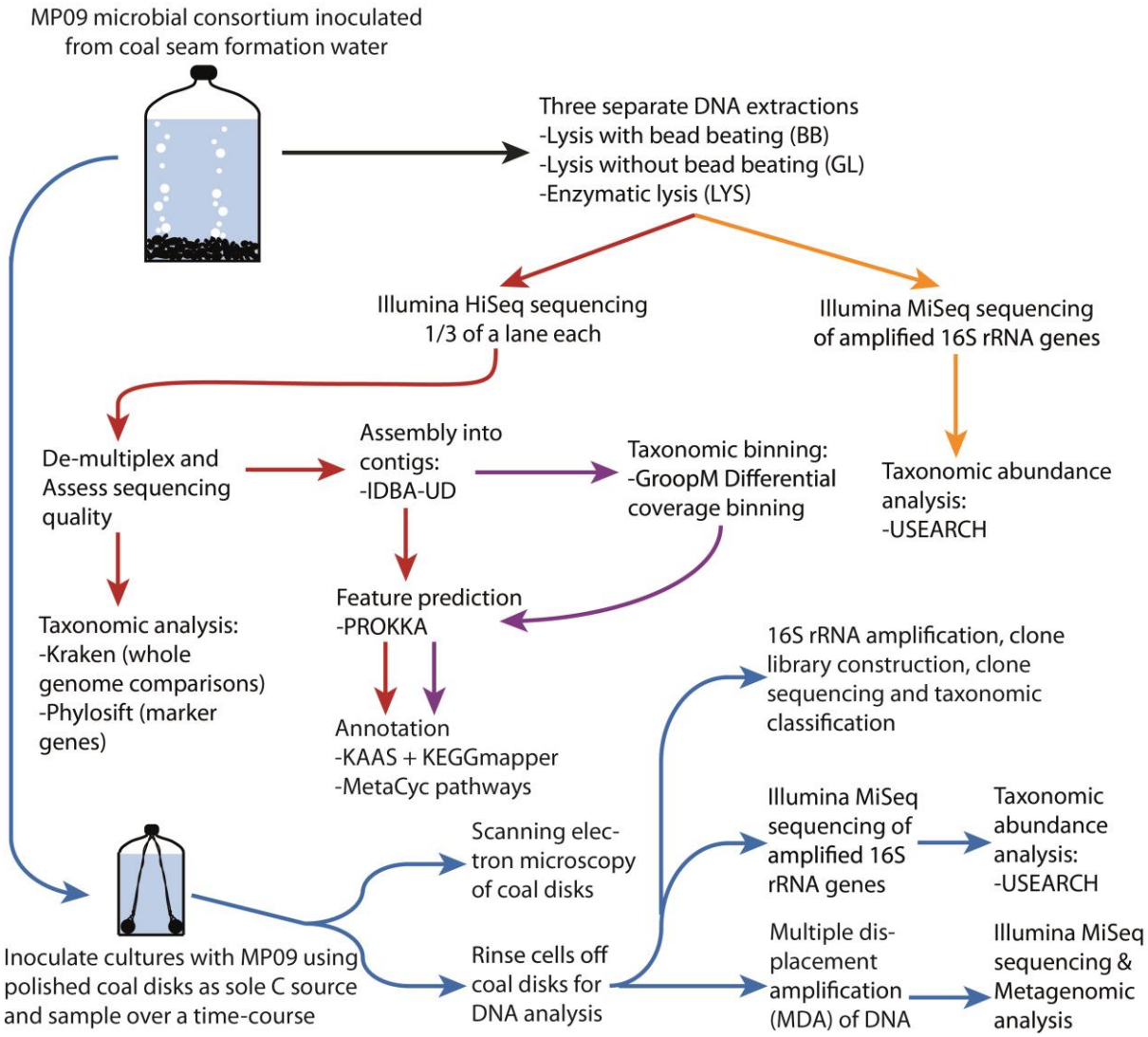


Figure 1.4 Experimental flow diagram of the present study. Arrow colours indicate the chapters in which the component experiments are presented. Chapter 2 (yellow), chapter 3 (red), chapter 4 (purple) and chapter 5 (blue).

Chapter 2. 16S rRNA amplicon diversity analysis of a microbial coal degrading community

2.1 Introduction

Contemporary microbial community studies routinely use deep sequencing of variable regions of the 16S rRNA gene to look at community membership and abundances. This involves the design of universal primers in conserved regions flanking the variable region, PCR amplification of this region from total genomic DNA extracted from the community and next gen sequencing of the amplification products. In this study 16S rRNA gene amplicon sequencing and analysis has been used as a starting point for looking at the coal degrading methanogenic MP09 microbial consortium to provide an overview of the taxonomic groups present and their relative abundances.

As previous studies looking at coal seam microbial communities in different coal seam environments across the planet (discussed in 1.2.4) have employed similar 16S rRNA gene amplicon or clone library sequencing this technique will allow me to closely compare the MP09 community with those reported in the literature. The taxonomic identification and relative abundance information generated with this approach can also be compared to those generated by metagenomic approaches on the same community to help assess the reliability and accuracy of this method for studying coal seam communities.

2.2 Methods

2.2.1 Sample preparation

Samples were taken from a coal degrading, methane generating microbial consortium named MP09. The consortium was grown in a minimal salts medium (100 mg/L NH_4Cl, 400 mg/L K_2HPO_4, 100 mg/L $MgCl_2$, 0.0001% resazurin, 1ml/L SL-10 trace element solution, 250 mg/L Na_2S, 200 mg/L Cysteine HCl and 1ml/L Wolins vitamin solution) with coal as the sole carbon source inoculated with coal seam formation water (chemical analysis of formation water is described in Supplementary Table 2) and incubated anaerobically without shaking at 30 °C for several months (Wolin et al, 1963).

600 ml of MP09 culture was filtered through three 0.1 μm filter disks and the disks equally cut into six sections for individual DNA extractions. Three protocols were used for DNA extraction of the microbes on the filter disks: the PowerSoil® DNA Isolation Kit (MO BIO Laboratories, Inc.) with 5 minutes of beat beating (BB), PowerSoil® DNA Isolation Kit (MO BIO Laboratories, Inc.) without bead beating but initial incubation in sample buffer increased to one hour (GL) and an enzymatic lysis protocol modified from Pope et al (2010) (LYS). The

LYS sample was purified again by addition of 96.5 µl 7.5 M NH_4OAc, 965 µl of 100 % ethanol and 2.5 µg of glycogen to 193 µl of LYS sample, vortexing and incubation at -20 °C overnight. This was followed by three steps of centrifugation at 12,000 x g (20 min, then 5 min) and pellet washing with 80% ethanol. The purified pellet was re-suspended in RNAse free water. Amplification and sequencing was carried out using 63 ng of extracted BB sample dsDNA and 54 ng of extracted GL sample dsDNA in ultrapure water.

2.2.2 DNA sequencing

16S rRNA amplicon sequencing was performed on DNA from BB and GL extractions using the Illumina MiSeq platform with 300 bp paired-end (Molecular Research LP). The library preparation, amplification and sequencing was performed by Mr DNA (Molecular Research LP) using the 515F (5'- GTGCCAGCMGCCGCGGTAA -'3) and 806R (5'- GGACTACHVGGGTWTCTAAT -3') universal bacterial and archaeal primer sequences which target amplification of the V4 hyper-variable region of the 16S rRNA gene (Caporaso et al, 2012). Amplicon sequencing was not performed on the LYS sample due to low DNA yield from extraction and purification.

2.2.3 Data processing and analysis

Quality filtering was performed on .fastq reads from 16S rRNA amplicon sequencing outlined above (2.2.2) using the usearch.exe -fastq_filter command with options --fastq minlen 269 to remove reads shorter than 269 bp and -fastq_truncqual to truncate reads at the first quality score below Q=5. Barcodes and primers were removed from the filtered reads using the python script Fastq_strip_barcode_relabel2.py, designating GTGCCAGCMGCCGCGGTAA as the primer sequence for trimming and barcodes specified in a separate file. Dereplication was performed on the trimmed sequences using the usearch.exe –derep_fulllength command. usearch.exe –sortbysize was used with the -minsize 2 option to sort the sequences and remove singletons. usearch -cluster_otus was used to cluster sequences into OTUs at a 97 % sequence similarity level. usearch.exe -uchime_ref was used for chimera removal using the Gold database. Filtered and debarcoded reads were mapped back to this set of OTUs with the usearch.exe -usearch_global command and converted to an OTU table using the usearch.exe -uc2otutab.py command. Finally OTUs were taxonomically identified using the RDP Classifier (Edgar, 2010, Edgar, 2013, Edgar et al, 2011, Wang et al, 2007). Rarefactions of the OTU table were generated with a custom python script.

2.3 Results

2.3.1 Sample preparation

Genomic DNA extracted from the mature MP09 consortium was determined to have 486 ng, 476 ng and 376 ng of dsDNA in the BB, GL and LYS extractions respectively using the Qubit HS DNA quantitation assay (LifeTechnologies). The DNA was also demonstrated to be intact by visualisation on an agarose gel (Figure 2.1).

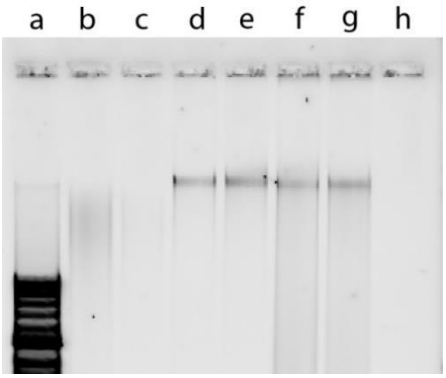


Figure 2.1 Visualisation of extracted genomic DNA. (a) HyperLadder™ 50 bp (highest band 2000 bp) (Bioline), (b&c) BB extraction, (d&e) GL extraction, (f&g) LYS extraction and (h) negative control.

2.3.2 Data processing and analysis

16S rRNA amplicon sequences for the BB and GL extractions were run through the usearch pipeline to perform quality filtering, generate OTUs and taxonomically assign OTU sequences. 99.5 % of the 511,868 reads in total passed the initial usearch quality filtering step resulting in the retention of 509,087 sequences. Of these sequences 497,143 (97.7 %) had barcode and primer sequences recognised and trimmed resulting in reads with a minimum length of 242 bp, average length of 271 bp and a maximum length of 274 bp. After removal of singletons and chimeras 145 OTUs were identified at 97 % similarity and 444,379 reads (86.8 % of total generated) were mapped back to these OTUs for taxonomic identification. OTU sequences were run through the rdp classifier (http://rdp.cme.msu.edu/classifier/classifier.jsp) to obtain taxonomic identifications (Wang et al, 2007). Taxonomic identifications and abundances are displayed in Figure 2.3. Rarefaction curves were generated with a custom python script for each extraction sample to assess species richness and whether sufficient sequencing had been performed to adequately sample the community (Figure 2.2).

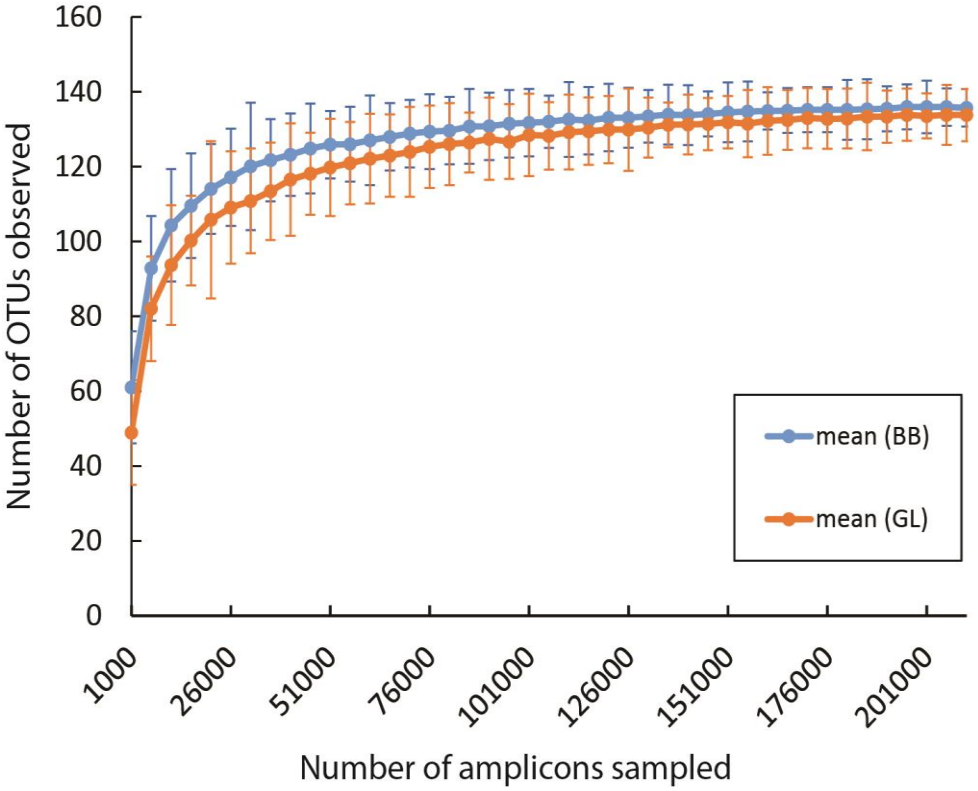


Figure 2.2. Rarefaction curves for 16S rRNA amplicons with OTUs defined at a 97 % sequence similarity level. 50 rarefactions were performed for each subsampling point and error bars indicate range of OTU numbers observed. The blue line indicates the bead beaten (BB) extraction sample and the orange line indicates the gentle lysis (GL) extraction sample.

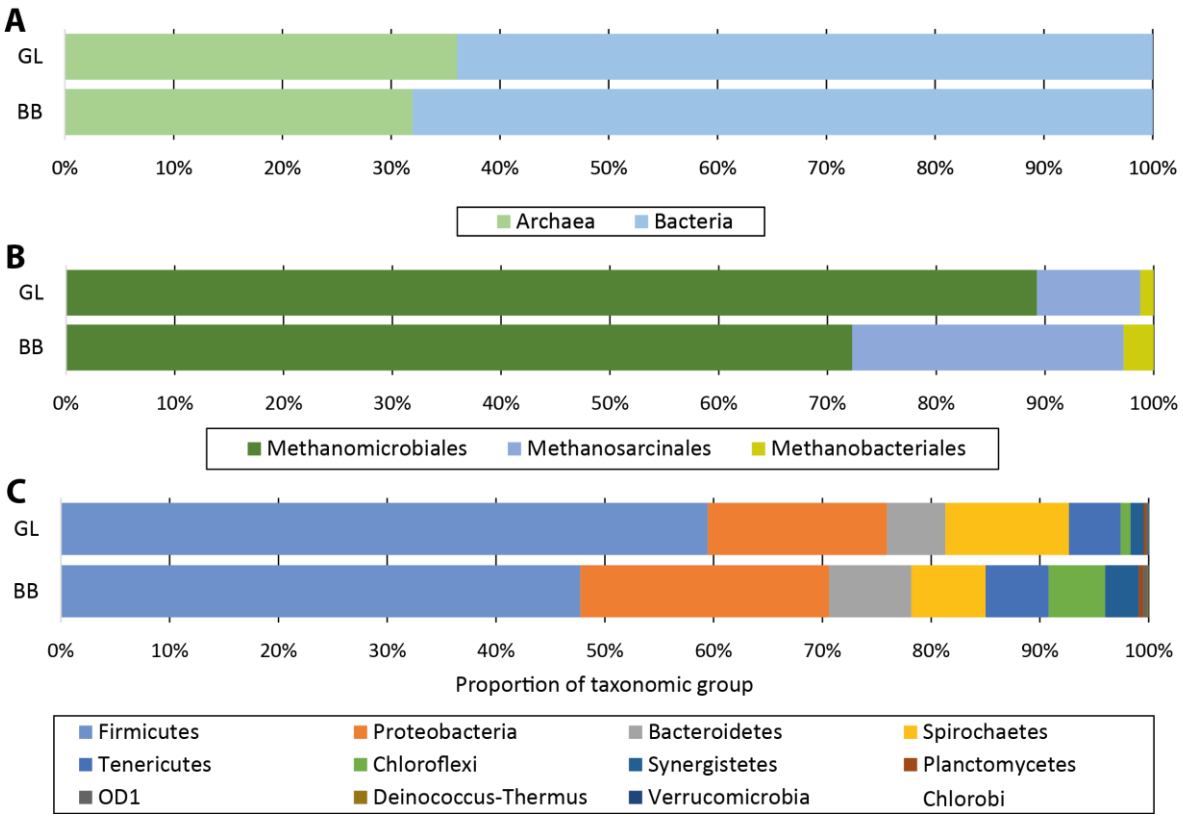


Figure 2.3 Taxonomic abundances of the mature MP09 consortium from bead beating (BB) and gentle lysis (GL) DNA extractions as determined by 16S rRNA analysis. (**A**) Shows the taxonomic abundances of organisms at the domain level, (**B**) shows archaeal abundances at the order level and (**C**) shows bacterial abundances at the phylum level.

2.4 Discussion

2.4.1 Quality assessment of sampling and sequences

DNA extractions appear to have yielded DNA of primarily high molecular weight although some reduction in DNA length was seen in the BB sample, probably due to the mechanical lysis mechanism. A high proportion of the reads generated (86.8 %) were retained through the stages of quality filtering in the USEARCH pipeline and included in the abundance analysis.

The rarefaction curve shown in Figure 2.2 shows a distinct flattening of the curve for both samples indicating that sampling depth was sufficient for the complexity of this community. It should be mentioned though that some of the levelling of the curve could be due to the pipeline used to generate the OTU table from which the rarefaction was generated. The USEARCH pipeline involves steps including the removal of singletons and predicted chimeras to reduce the number of false positive identifications generated through sequencing errors but may also remove some of the very low abundance OTUs (Edgar, 2013, Edgar et al, 2011).

Some differences in abundance were seen between the BB and GL samples and are likely caused by differences in extraction efficiencies for different taxonomic groups between the two extraction methods. However, while there were some small scale differences the overall community structure shows a high level of similarity between the two samples. The LYS extraction sample was not included in the amplicon survey due to the low levels of DNA in that extraction resulting in only enough DNA for metagenomic sequencing.

2.4.2 Analysis of taxonomic abundances

Comparing the relative abundances of organisms from this study to those in previously published taxonomic abundance analyses (Figure 1.3) reveals several similarities. Firstly the archaeal component of the community is largely dominated by one methanogenic order with other orders present in low abundances. This appears typical of the archaeal coal seam communities where one order typically dominates, presumably because niche physicochemical characteristics of the environment have a strong impact on the energetic favourability of a specific methanogenic pathway. In this case the Methanomicrobiales constitute the dominant archaeal group suggesting that hydrogenotrophic methanogenesis is dominant in the MP09 consortium.

In terms of the bacterial component of this community there are some significant differences to many of the published 16s rRNA based abundance studies. Firstly the dominant phylum in the MP09 community appears to be the Firmicutes which is in contrast to most previous studies where the Proteobacteria typically dominate. However, there are two coal seams where Firmicutes are reported as being in high abundance, one in Illinois (Strapoc et al, 2008) and one in Wyoming (Green et al, 2008). Another interesting feature of the bacterial component of the MP09 community is the presence of a relatively high number of Spirochaetes, a taxon which is only reported in significant abundance in the same Illinois coal seam in which the Firmicutes were also a dominant group (Strapoc et al, 2008). Additionally, these results indicate the presence of the Tenericutes in the community, another phylum which is not commonly observed with the exceptions of in the Wyoming coal seam mentioned previously (Green et al, 2008) and in another Australian coal seam (Midgley et al, 2010). These observations of similarity between the Australian community presented here to those reported in Wyoming and Illinois are interesting and suggest that similarities in microbial coal communities are not due to geographical connections between seams but more likely due to physicochemical similarities in the environments such as temperature or the presence of similar chemical compositions driving convergent community structuring (Green et al, 2008, Strapoc et al, 2008). From the available metadata supplied with the previous

studies and those presented in Supplementary Table 2, no physicochemical characteristic could be distinguished as specific to the communities presented in (Green et al, 2008, Strapoc et al, 2008) and this study, compared to others in the literature (Fry et al, 2009, Green et al, 2008, Guo et al, 2012, Strapoc et al, 2008, Tang et al, 2012). However, the characteristics measured by the various microbial coal seam studies were varied and more comparable metadata sets may be required to draw any specific conclusions about physicochemical effects on community structure.

This MP09 consortium also shared many similarities with previously published coal seam communities. The Proteobacteria and Bacteroidetes which were almost ubiquitous throughout the published communities were also seen in high abundance in the MP09 community suggesting key roles being played by these organisms in degradation of organic matter in coal to methane.

While these results have allowed for a comparison of the MP09 community with those published in the literature, the abundances observed may not represent the true organismal abundances but rather the abundances of the 16S rRNA gene. As the 16S rRNA gene is known to be present in variable multiple copy number in many microbial species, organisms with high copy numbers will show inflated abundance measures in comparison to low copy number organisms. For example the three most abundant bacterial phyla observed in the amplicon survey of the MP09 consortium presented here: Firmicutes, Proteobacteria and Bacteroidetes have mean 16S rRNA gene copy numbers of 6.29, 4.11 and 3.09 and standard deviations of 3.04, 2.51 and 1.62 respectively (Lee et al, 2009). To take this variation into account attempts have been made to create tools for normalising 16S rRNA copy number based on 16S copy number databases (Kembel et al, 2012, Klappenbach et al, 2001, Lee et al, 2009). Such normalisation has not been performed for this data however, firstly because many of the species identified in this study do not have well annotated genome sequences in the recorded databases but also because 16S rRNA copy number has shown to correlate with growth rate. This means that where sequenced representatives of taxa seen in this study such as the Firmicutes can be found they are likely from much faster growing strains which are more amenable to culture (Kembel et al, 2012, Klappenbach et al, 2000, Lee et al, 2009).

Chapter 3. Functional and taxonomic diversity of a coal degrading microbial consortium based on metagenomic sequencing

3.1 Introduction

Metagenomic sequencing is emerging as a powerful and useful tool in the field of molecular ecology. Its benefits over traditional isolation, culturing and genome sequencing of individual community members are that it allows for the inclusion of unculturable microorganisms in microbial community studies as well as taking far less time and resources than culture based studies. It can also provide community abundance data for taxonomic groups or functional gene groups. It also has advantages over the marker gene amplicon studies of microbial communities as it bypasses the copy number variation and primer biases which affect these studies and provides a large amount of functional information on the community and its members not obtained from marker gene amplicon studies.

At this point there has only been one very recent publication of a metagenome from a coal associated microbial community (Ghosh et al, 2014) and pure cultures of organisms in these communities have generally not been possible due to the obligate syntrophy within these communities (Stams, 1994). The metagenomic study presented in this work aims to provide a unique insight into a coal degrading methanogenic community through examination of the functional genes present in the community from a metabolic pathway perspective as well as providing valuable validation for taxonomic abundance measurements generated by 16S rRNA amplicon based measures of this community generated in chapter 2.

3.2 Methods

3.2.1 Sample preparation and DNA sequencing

Metagenomic shotgun sequencing was performed on all three DNA extractions from the MP09 consortium (BB, GL & LYS) (outlined in 2.2.1) using the Illumina HiSeq2000 platform (1 lane, 100 bp paired end). 350 ng of both BB and GL purified dsDNA and 376 ng of extracted LYS dsDNA was used for sequencing. The TruSeq Nano (Illumina) sample library preparation kit was used for library preparation with a target insert size of 550 bp. Library preparation and sequencing were performed by The Ramaciotti Centre for Genomics (University of New South Wales).

3.2.2 Data Processing

Taxonomic abundances were calculated for reads from the three extraction samples separately in addition to together in a concatenated file, using the MetaPhlAn pipeline software (Segata et al, 2012). The metaphlan.py python script was run with the –rel_ab option to calculate

relative abundances of taxonomic groups and the –bowTie2_exe option to perform mapping against the MetaPhlAn BowTie2 marker set. The alignment stringency for BowTie2 alignment was left as default (e-value $1x10^{-6}$) along with all other options (Segata et al, 2012). Taxonomic abundances were also calculated for reads using the Kraken taxonomic sequence classification system software. Sequence reads from the three samples were concatenated into a single file for analysis with this tool and the full size Kraken database was used for taxonomic identification and abundance measurement (Wood and Salzberg, 2014).

The IDBA-UD sequence assembler was used for co-assembly of the BB, GL and LYS sample reads using default parameters (Peng et al, 2012). The co-assembled scaffolds were analysed using the software pipeline PROKKA to provide identification and annotation of putative coding sequences (CDSs). The RNA predictor used was RNAmmer, all options were left as default with the exception of --metagenome which optimises gene predictions for fragmented assemblies (Lagesen et al, 2007, Seemann, 2014).

CDSs generated from PROKKA was uploaded to the KEGG automatic annotation server (KAAS) (Moriya et al, 2007). KEGG orthology assignments were generated for the protein sequences using the bi-directional best hit assignment method. Generated KEGG orthology numbers were submitted to the KEGG mapper search pathway tool at (http://www.genome.jp/kegG/tool/map_pathway1.html) and the results visualised in KEGG pathway maps.

3.3 Results

3.3.1 Quality assessment of metagenomic data

DNA used for metagenomic sequencing came from the same extractions as outlined in Chapter 2 above. They were of a high molecular weight and appeared relatively intact for all three extractions (figure 2.1). Illumina HiSeq sequencing generated metagenomic reads which showed good sequence quality when assessed with FastQC (Supplementary Figure 1). Reads co-assembled from the three samples with IDBA-UD provided 197,494 scaffolds (83,530 over 500 bp) with a total length of 314,141,129 bp (279,396,050 bp in scaffolds over 500 bp) and had an N50 of 14,871 bp.

3.3.2 Taxonomic identification and abundance analysis of the MP09 community

For concatenated reads from all three extractions, Kraken provided taxonomic classification based on 4.13 % of the total reads while MetaPhlAn provided taxonomic classification based on matches of 1.74 % of the reads to the marker set. When MetaPhlAn was run on metagenomic reads of individual samples the following percentage of reads matched to the marker set: BB 2.43 %, GL 1.76 % & LYS 1.32 %.

Taxonomic identifications and their relative abundances from both the MetaPhlAn and Kraken analysis of the combined metagenomic reads were calculated and represented in Figure 3.1. Relative taxonomic identifications and their abundances generated from the three extractions individually were also generated by MetaPhlAn and are summarised at the phylum level in figure 3.2 to illustrate the level of inter-sample variation in relative abundance for the different taxa.

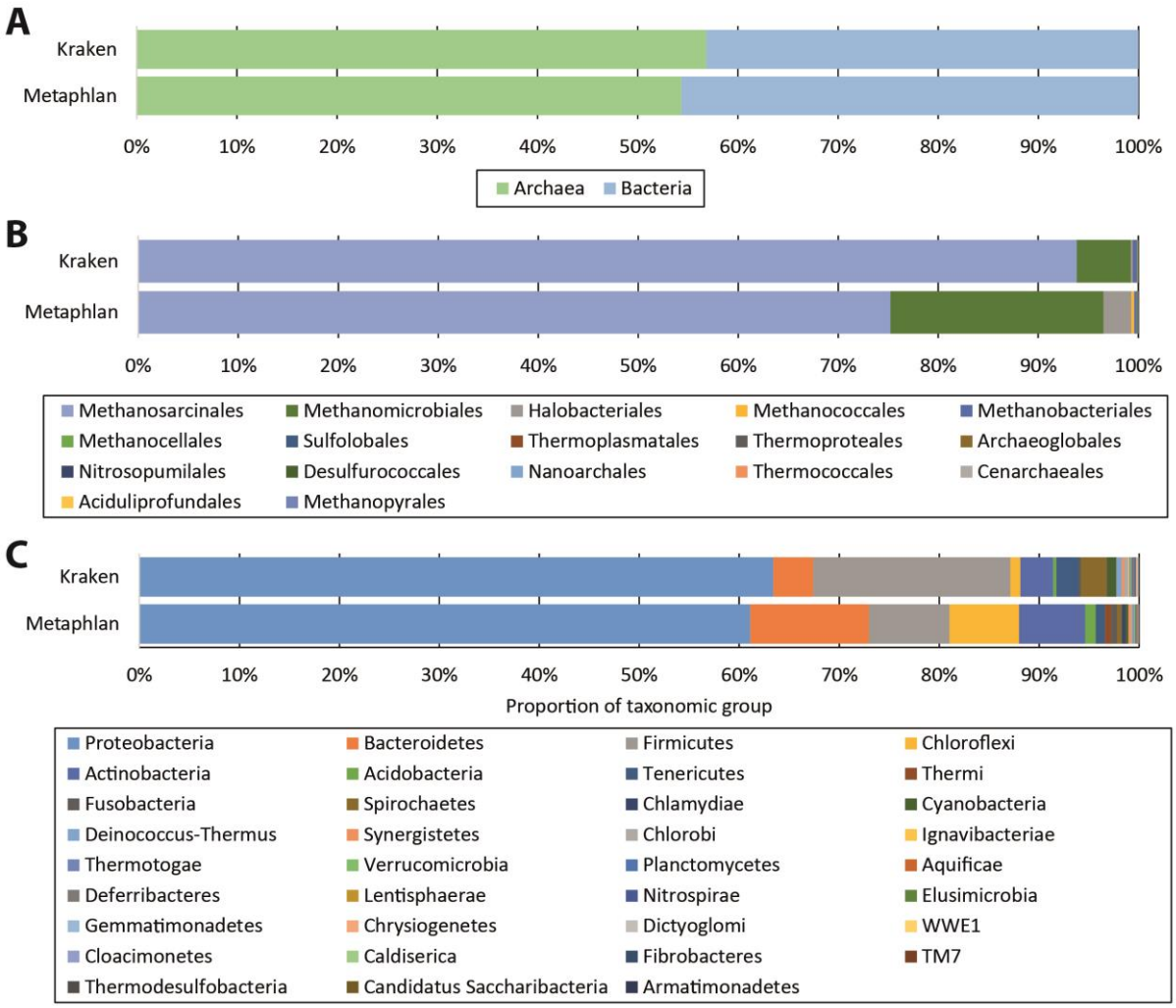


Figure 3.1 Taxonomic abundances of the mature MP09 consortium from pooled bead beating (BB), gentle lysis (GL) and enzymatic lysis (LYS) of DNA extractions as determined by analysis of DNA reads by Kraken and MetaPhlAn software. (**A**) Shows the taxonomic abundances of organisms at the domain level, (**B**) shows archaeal abundances at the order level and (**C**) shows bacterial abundances at the phylum level.

Table 3.1 Percent taxonomic abundances for the three extraction samples at the phylum level as determined by analysis with MetaPhlAn software. Depth of shading indicates the relative abundances of the taxonomic groups in each extraction.

	LYS	GL	BB		LYS	GL	BB
Acidobacteria	0.779	0.120	0.171	Fusobacteria	0.386	0.404	0.275
Actinobacteria	5.977	1.815	1.352	Gemmatimonadetes	0.029	0.005	0.003
Aquificae	0.047	0.036	0.015	Lentisphaerae	0.037	0.012	0.010
Bacteroidetes	7.954	3.962	2.801	Nanoarchaeota	0.003	0.004	0.003
Chlamydiae	0.665	0.146	0.083	Nitrospirae	0.030	0.013	0.010
Chlorobi	0.141	0.032	0.027	Planctomycetes	0.082	0.017	0.015
Chloroflexi	2.565	1.952	1.160	Proteobacteria	46.271	23.183	14.014
Chrysiogenetes	0.017	0.006	0.003	Spirochaetes	0.420	0.280	0.192
Crenarchaeota	0.017	0.000	0.000	Synergistetes	0.188	0.103	0.070
Cyanobacteria	0.213	0.170	0.148	Tenericutes	0.650	0.709	0.391
Deferribacteres	0.021	0.044	0.025	Thaumarchaeota	0.010	0.004	0.003
Dictyoglomi	0.012	0.010	0.007	Thermi	0.581	0.180	0.155
Elusimicrobia	0.027	0.017	0.011	Thermotogae	0.075	0.063	0.038
Euryarchaeota	28.328	61.643	76.572	TM7	0.000	0.000	0.000
Fibrobacteres	0.003	0.000	0.000	Verrucomicrobia	0.099	0.032	0.020
Firmicutes	4.369	5.035	2.424	WWE1	0.005	0.001	0.001

3.3.3 Community functional pathway analysis

To look at the functional capabilities of the community as a whole, KEGG functional annotation was performed for the co-assembled scaffolds mapped against the KEGG pathway maps. For this analysis we looked at the KEGG methane metabolism map (Figure 3.2) to visualise genes present in the community involved in methane metabolism. There were 105 matches to methanogenesis pathway genes identified from the KAAS ortholog annotation of total predicted CDSs. From the KAAS ortholog annotation of total predicted CDSs there was also 61 matches to sulphur metabolism pathway genes, 38 matches to degradation of aromatic compounds pathway genes and 5 matches to polycyclic aromatic hydrocarbon degradation pathway genes.

3.4 Discussion

3.4.1 Quality assessment of reads, assemblies and taxonomic assignment tools

The assembly of reads with the IDBA-UD assembler provided a large number of long contigs with a high level of total sequence covered. This suggests that identification and functional analysis of genes can be made on full length gene sequences in a high number of cases and in many instances gene sets or full operons are likely to be covered on single scaffolds.

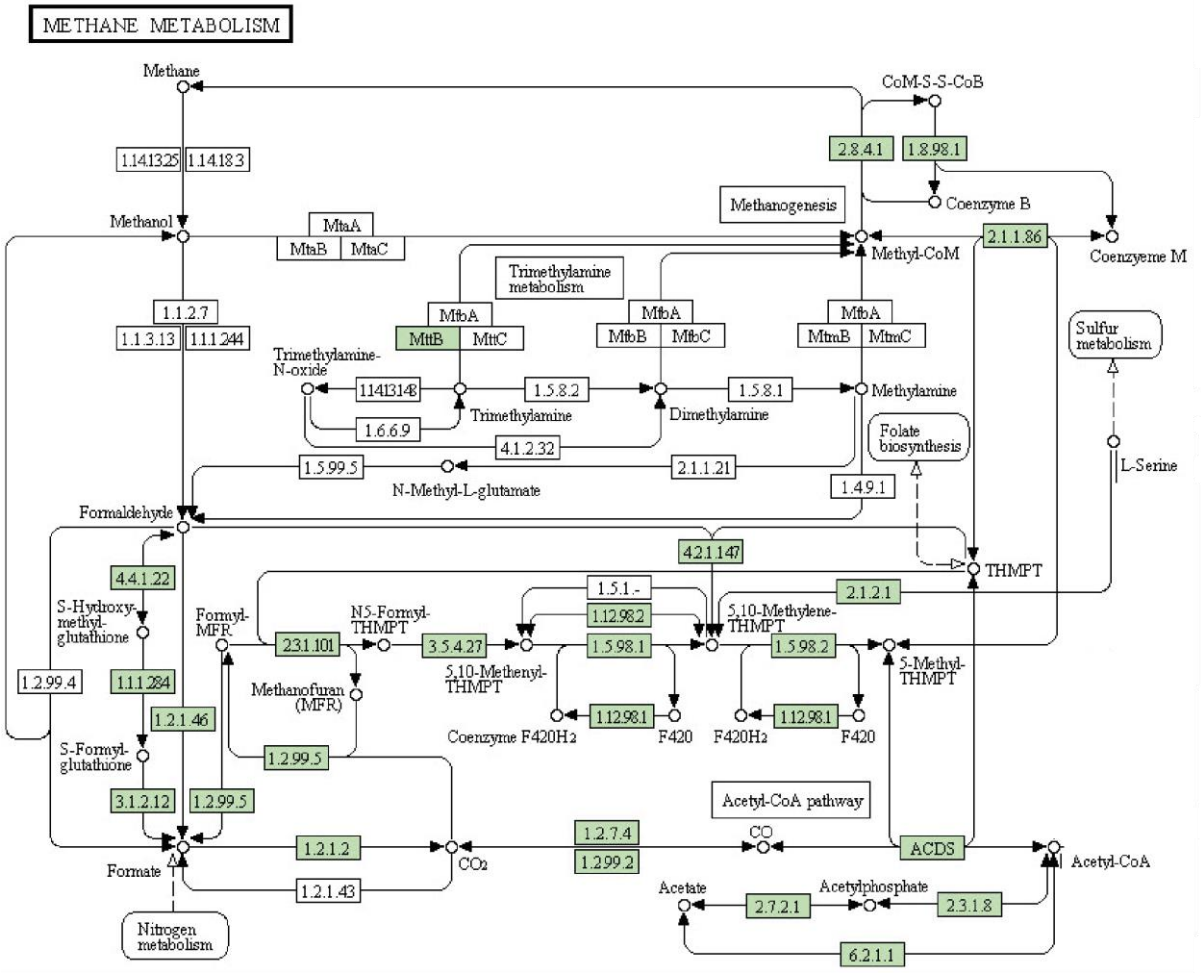


Figure 3.2 Pathway map of methanogenesis reactions modified from KeggMapper output. Arrows indicate enzymatic reactions and boxes indicate enzymes catalysing the reactions. Boxes coloured green indicate enzymes for which predicted CDSs have been identified in the assembled metagenomic dataset.

The quality of read sequences was also determined to be of high quality as determined by quality analysis with FastQC (Supplementary Figure 1) so erroneous taxonomic classification based on sequencing errors is unlikely in most cases. However, the proportion of reads which were utilised for taxonomic classification by Kraken was quite low compared to levels reported in the paper presenting the Kraken tool (4.13 % of reads versus 31.8 % of reads respectively) (Wood and Salzberg, 2014). A similar observation was made for results from the MetaPhlAn analysis where an average of 1.74 % of reads across the three samples had alignments to the maker set in comparison to the 7.7 % and 8.4 % of reads matching the markers in the paper presenting the tool (Segata et al, 2012). This is likely due to the low number of sequenced genomes from environmental samples such as those in coal seams

which are uncommon and difficult to culture as these tools rely on homology matches with sequences from annotated genomes.

3.4.2 Inter sample variation in taxonomic abundance

The inter sample variation, (Table 3.1) based on MetaPhlAn analysis, showing the differences in taxonomic abundances between the three extraction protocols indicates that while the majority of the phyla are extracted with broadly similar efficiencies a few abundant phyla show high levels of variation in terms of number of reads attributed to them, particularly the Euryarchaeota and Proteobacteria. This indicates that the type of extraction protocol can have a significant effect on the measured taxonomic abundances, an observation which has previously been recorded for some methanogenic Euryarchaeota due to their cell walls resistance to proteinases which some lysis protocols rely on more heavily than others (Dridi et al, 2009). However, because these are relative abundances it should be noted that the effect could be due to extraction efficiency of one of the taxonomic groups rather than both as highly abundant sequences will take up more of the fixed number of reads in each sample. This could explain the observation that the two highest abundant groups show inverse relative abundance changes (Table 3.1). In addition to the large relative abundance differences in the high abundance taxa there were also several changes observed for many of the lower abundance taxa such as the Chlamydiae. While these were small differences in terms of numbers of reads they represented quite large fold changes in abundance between the samples, an example being the Chlamydiae which showed an approximately 10 fold difference in abundance between the LYS and BB extractions (Table 3.1). These findings illustrate the importance of extraction method in sampling a complex microbial community and show that whilst the true abundances of the microorganisms in the community may be impossible to determine a multifaceted extraction approach is recommended to catch organisms which could be missed by a single extraction method.

3.4.3 Taxonomic abundance analysis

Whilst the LYS sample which showed the lowest proportion of Euryarchaeota was not included in the 16S rRNA amplicon analysis, both metagenomic based taxonomic analyses showed a much higher proportion of Euryarchaeota in comparison to the 16S rRNA amplicon analysis for the BB and GL samples (Figure 3.1). This increase in archaeal abundances could be due to an overestimation of the bacterial groups in the 16S rRNA survey due to typically higher 16s rRNA gene copy number in bacteria compared to archaea, or due to a potential primer amplification bias towards bacterial 16S rRNA sequences (Klappenbach et al, 2001, Lee et al, 2009). Within the Euryarchaeota at the order level there was also significant

differences between the 16S rRNA amplicon and metagenomic abundance analyses. While the Methanosarcinales and Methanomicrobiales comprised the most abundant archaea in both analyses, the abundances relative to each other were very different between the amplicon survey where the Methanomicrobiales dominated and the metagenomic analysis where the Methanosarcinales dominated. One reason for this change could be differences in genome size increasing the predicted abundance in the metagenomic analysis due to a higher number of marker genes being observed in the larger genome. This could explain the observation as Methanosarcinales contains species with the largest archaeal genomes recorded (Deppenmeier et al, 2002, Galagan et al, 2002, Maeder et al, 2006).

The bacteria also showed quite different abundances between the amplicon and metagenomic analyses at the phylum level where the Firmicutes appeared far less abundant and the Proteobacteria appeared more abundant in the metagenomics data (Figure 3.1). A result which is likely due to a combination of known high 16S rRNA copy number in the Firmicutes phylum and the fact that the Proteobacteria are a very highly represented phylum in the genomic sequence databases (Klappenbach et al, 2001). Other phylum level groupings appeared to be broadly in accordance between the amplicon and metagenomics analyses with both reporting similar taxonomic groups although the metagenomic approach identified a much greater number of taxa at low abundances.

3.4.4 KEGG functional pathway annotation

A majority of the methanogenic metabolism pathway could be assigned from KAAS annotation of CDSs identified by PROKKA (Figure 3.2). One noticeable absence from the methanogenic pathway map was genes associated with degradation of the various methylamines to methane. This is surprising given the presence of members of the order Methanosarcinales which are known to be able to utilise these molecules for methanogenesis (Liu and Whitman, 2008). Otherwise the genes involved in hydrogenotrophic methanogenesis from CO_2 and acetogenic methanogenesis are both quite well covered in the pathway map and genes involved in methane utilisation were absent which supports the accurate annotation of these CDSs as this process requires aerobic conditions to function which are not present in the coal seam environment.

For other pathways such as the aromatics degradation pathway and polycyclic aromatic hydrocarbon degradation pathway there were very few matches to predicted CDSs. This is surprising since the MP09 community survives on coal which is largely composed of these aromatic and polyaromatic hydrocarbon types of compounds (Strąpoć et al, 2011). This could indicate that the mature MP09 community contains the organisms responsible for these

processes at such low abundances that they are not well covered in the metagenome. Another hypothesis could be that the organisms degrading coal are utilising novel gene pathways for the degradation of coal. By studying the community at early stages of coal colonisation we should be able to gain some insight why such low numbers of genes in these pathways were observed.

Chapter 4. Inferred functional analysis of taxonomically-binned metagenomic data

4.1 Introduction

4.1.1 Taxonomic binning of metagenomes

Analysis of a metagenome can provide information about the community as a whole in terms of its functional capabilities and taxonomic members but generally does not allow for determination of which members are responsible for individual functions within the group, except through inference from previously described functions of taxonomic groups. Taxonomic binning aims to address this limitation by grouping or "binning" scaffolds to a taxonomic group to which there is bioinformatic evidence that the sequences belong. Several methods are available for this task based on either grouping by alignment of contig sequences to annotated single genomes or grouping based on shared diagnostic sequence characteristics of the contig. Here a newly developed tool, named GroopM, was utilised which employs differential coverage information of individual scaffolds from three extraction methods in addition to shared sequence characteristics to base assignment of scaffolds to taxonomic bins (Albertsen et al, 2013, Imelfort et al, 2014).

This technique relies on the assumption that the DNA from individual taxonomic groups will extract with different efficiency using different extraction procedures and so the coverage profiles across each contig from the three methods should be very similar for all the scaffolds originating from the same taxonomic group. This assumption is supported by the observations made in the previous two chapters that the extraction procedure has caused taxonomy specific changes to the extraction efficiencies as can be seen in the differing relative abundance measures (Figure 2.3 & Table 3.1) (Albertsen et al, 2013).

4.1.2 Assessment of taxonomic bins

This component of work aimed to group assembled scaffolds into taxonomic bins, assess the accuracy of the binning, classify the taxonomy of the individual bins and analyse their functional capability in terms of methanogenic pathway functions. For evaluating the accuracy of the binning results and classifying the taxonomy of each bin a taxonomic identification tool which utilises a database of taxonomically conserved marker genes was used to check if taxonomy is conserved on scaffolds within bins. For functional analysis of the bins two approaches were used to analyse CDSs identified on the scaffolds within each bin. The first method uses KAAS based assignment of bins as described in 3.3.3 and the second is based on hidden markov modelling (HMM) searches of the CDSs against a set of methanogenesis functional marker gene HMM models. By using these two different

approaches a better idea of the bins containing elements of methanogenic function as well as the likely methanogenic pathways present in the bins can be achieved.

4.2 Methods

4.2.1 Taxonomic binning of scaffolds

Coverage information was mapped to the scaffolds for each extraction method by first aligning reads to the scaffold sequences using the Burrows-Wheeler Aligner (BWA) -aln command (Li and Durbin, 2009), then creating indexed and sorted .bam files for each of the three alignments using Samtools (Li et al, 2009). These were then used to create a GroopM database file using the -parse command in GroopM. The -core command was used to create a set of core bins with minimum of 10 scaffolds larger than 1,500 bp unless the cumulative length of scaffolds exceeds 1,000,000 bp. The "recruit" command with default settings was used to recruit unbinned scaffolds larger than 500 bp into the core bins. The bins were refined automatically using the "refine" command with the –auto argument (Imelfort et al, 2014).

4.2.2 Taxonomic assignment of bins

Individual files containing all scaffolds assigned to respective bins were analysed with the Phylosift pipeline against the core marker set using the Phylosift –all command (Darling et al, 2014). Taxonomies were assigned at the class level by the following criteria; single taxonomic assignment when ≥50 % of the markers indicated a single class, double assignment when ≥60 % of the markers indicated two classes where each made up ≥30 % and mixed assignment where these two criteria were not met. A no hits assignment was given to bins where no marker gene was detected on any of the contained scaffolds.

4.2.3 Functional analysis of taxonomic bins with KEGG

A file containing all of the predicted translated protein CDSs generated from PROKKA found on binned scaffolds was uploaded to the KEGG automatic annotation server (KAAS). KEGG orthology assignments were generated for the protein sequences using the bi-directional best hit assignment method. A list of KEGG orthology numbers for the genes in the methanogenesis pathway was searched against the assigned KEGG orthology numbers generated and numbers of matches within each bin recorded.

4.2.4 Functional analysis of taxonomic bins with Pfam HMM searches

To look for functional pathways within individual bins a file containing scaffolds with relevant bin numbers included into the header lines was generated. The HMMER3 (Eddy, 1998) hmmsearch tool was used to search for the presence of Pfams matching key methanogenesis pathway marker genes *mcrα* (PF02745), *mrcβ* (PF02241), *MttB* (PF06253), *MtmB* (PF05369), *MtbC* (PF09505), *MtbB* (PF12176), *CdhD* (PF03599) and *MTmer*

(PF00296). A BLAST search was performed on the scaffold sequences which had matches to the *Mrcα* gene to assign taxonomy (www.ncbi.nlm.nih.gov/BLAST) (Altschul et al, 1990).

4.3 Results

4.3.1 Taxonomic binning of scaffolds

Taxonomic binning of scaffolds from the combined metagenomic dataset with the GroopM software resulted in 117 bins containing 41.65 % of the total scaffolds and representing 88.38 % of the total sequence length covered by the scaffolds. Taxonomic analysis using Phylosift identified 57 bins with single taxa classification, 8 with two dominant taxa, 35 with mixed taxa assignments and 14 with no matches to the marker database. A table listing all the bins generated by GroopM, the number of scaffolds, total sequence length covered, mean GC content and assigned taxonomy of each individual bin can be found in Supplementary Table 1. A three-dimensional coverage plot of all the binned scaffolds and their individual coverages in the three extractions was also generated by GroopM (Figure 4.1).

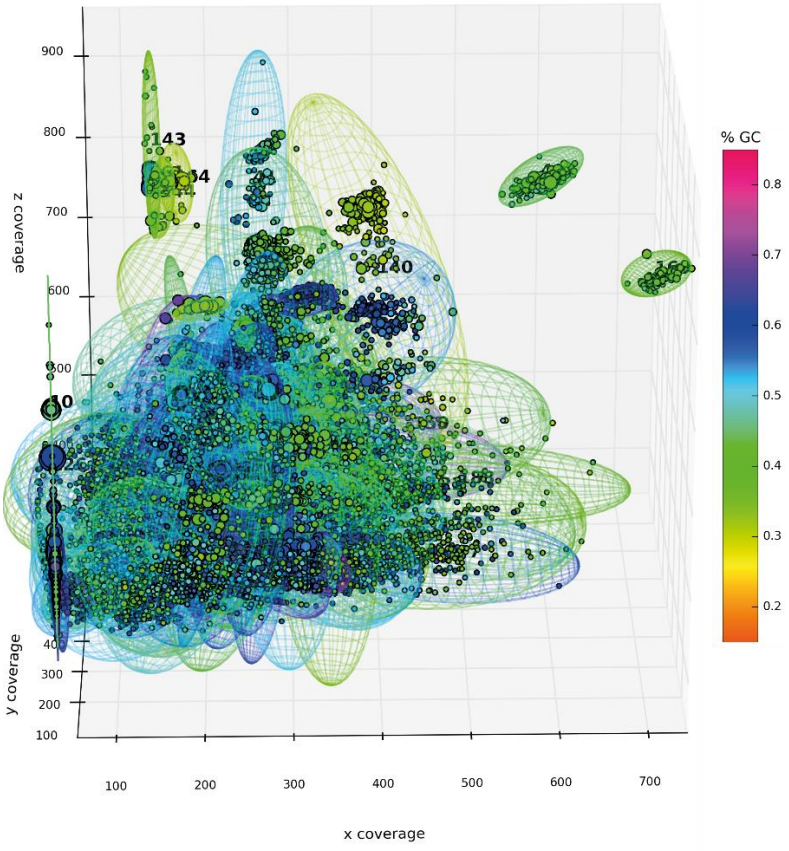


Figure 4.1 Three-dimensional plot from GroopM showing scaffolds (coloured circles with circle size representing scaffold length) from a co-assembly of reads from the three different extractions and the bins they have been assigned to (coloured elliptical meshes) coloured by average GC content. X, y and z axes represent read coverage of the scaffolds from each of the three extractions. The point at which the x and y axes converge in the bottom left of the figure is the origin of the three axes where coverage from all three extractions is zero.

4.3.1 Functional analysis of taxonomic bins with KEGG

A functional analysis of the binned sequences was performed with KEGG pathway mapper to identify bins containing functional methanogenesis genes and to quantitate the number of associated genes found for each bin. A chart showing the number of predicted CDSs matching genes listed as part of the methanogenesis pathway in KEGG was generated to visualise the distribution of functional methanogenesis genes within the bins (Figure 4.2).

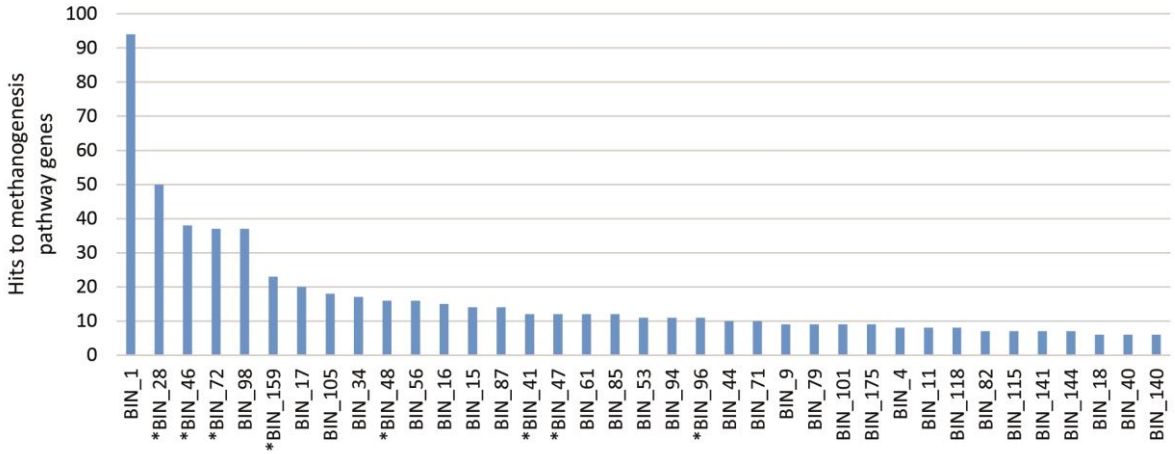


Figure 4.2 Number of matches to methanogenic pathway genes for predicted CDSs in labelled taxonomic bins. Bins marked with a (*) were determined to contain the methanogenesis marker gene *mcrα* or *mcrβ* by HMM searching of Pfams against the bins CDSs (Figure 4.3). Bins with less than six matches were not included in the figure.

4.2.4 Functional analysis of taxonomic bins with Pfam HMM searches

To provide further evidence of the methanogenic function of particular bins and to try and distinguish the mode of methanogenesis for identified bins, a second approach was performed utilising HMM searching of Pfam HMM models. Pfam models for key methanogenesis marker genes were searched against the predicted coding sequences contained on all binned scaffolds (Figure 4.3).

Table 4.1 Matches to characteristic substrate specific methanogenesis marker genes in taxonomic bins which showed matches to one or both of the general methanogenesis marker genes *mcrα* and *mcrβ*. Matches were included at an e-value cut-off of 4.0×10^{-7}. Taxonomic assignment of whole bins was performed by Phylosift as described in 4.2.2. Taxonomic assignment of scaffolds was performed by a nucleotide BLAST search of the contig containing the *mcrα* (or *mcrβ* when no mcrα identified) gene against the NCBI non-redundant nucleotide collection.

			Methanogenesis		Methylamines			methanol	acetate	CO2
BIN	Taxonomic assignment of whole bin	Taxonomic assignment of contig containing mcrα gene	mcrα (PF02745)	mcrβ (PF02241)	MttB (PF06253)	MtmB (PF05369)	MtbC (PF09505)	MtbB (PF12176)	CdhD (PF03599)	MTmer (PF00296)
28	Methanobacteria	Methanobacterium	2	2	0	0	0	0	2	0
41	Methanomicrobia	Methanospirillum/ Methanoculleus	2	2	0	0	0	0	1	1
46	Methanomicrobia	Methanosarcina	1	1	4	3	5	3	2	3
47	Methanomicrobia	Thermoplasmatales sp.	1	1	2	7	2	3	2	1
48	mixed	Methanoculleus	1	1	0	0	0	0	3	4
49	Methanomicrobia*	Methanoculleus	0	1	0	0	0	0	0	0
72	Methanomicrobia	Methanolobus	1	1	2	2	0	1	3	2
96	Bacteroidia	Uncultured bacterium clone	1	1	0	0	0	0	2	0
103	no hits	Methanoculleus	1	0	0	0	0	0	0	0
159	mixed	Methanobacterium/ Methanolobus	2	1	2	1	2	1	0	1
165	Bacilli/Methanomicrobia	Methanolobus	0	1	0	0	0	0	0	0

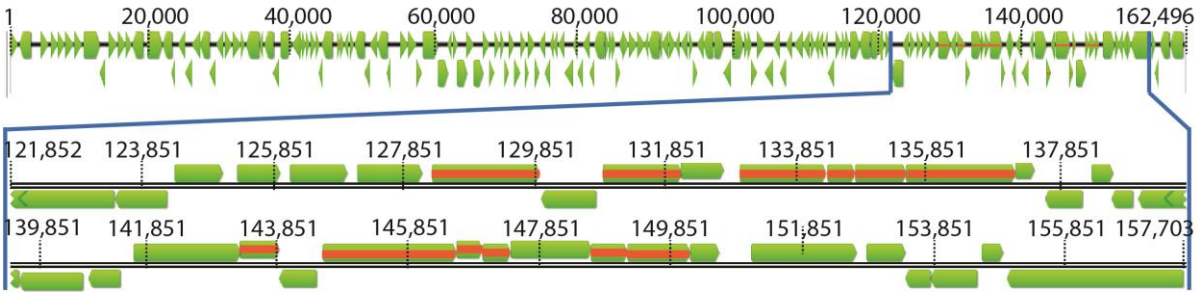


Figure 4.3 Annotation of genetic elements on the scaffold containing the *mcrα* gene identified in bin 47. Green arrows indicate CDSs identified by PROKKA. Arrows marked with an orange line represent genes associated with methanogenesis according to their CDS PROKKA annotation. Black numbers represent length in base pairs. A list of predicted CDS protein product names generated by PROKKA are listed in Supplementary Table 3.

4.4 Discussion

4.4.1 Assessment of the GroopM taxonomic binning of coal metagenome data

Taxonomic binning of the metagenome was performed with the GroopM differential coverage binning tool to add an additional dimension of information to the assignment of scaffolds to bins.

As can be seen in Figure 4.1 many high coverage scaffolds can be distinguished as clearly independent clusters with consistent GC content levels. However, as the coverage level drops there is less clearly distinguished groupings due to coverage overlaps of the many low abundance organisms. This lack of visually distinguishable groupings at lower coverages does not eliminate successful binning based on sequence composition clustering. The high level variation in GC content across the scaffolds between 30 % and 70 % in these low coverage groups suggests many bins could be assigned from sequence composition characteristics.

Based on the criteria outlined in 4.2.2, just over half the bins appeared to contain a single taxonomic group and the taxonomic identifications were common to those seen in the community surveys presented in Chapters 2 and 3 (Supplementary Table 1). Whilst this lends good support to the hypothesis that binning of taxonomic groups in the metagenome was successful, the assignment cut-off limits for these identifications had to be set arbitrarily based on a best estimate of how a good single taxonomic bin would appear. The fact that these identifications are commonly of microbial taxa which are not well represented in the nucleotide databases had to be taken into account as well. This problem of how to assess the quality of taxonomic bins is common and as of yet there does not appear to be a consensus on how this should be done. The issue is commented upon in the paper presenting the GroopM tool where the authors used the identification of essential genes and the observed copy number of single copy number genes in a bin to assess bin quality (Albertsen et al, 2013). However, this method has potential flaws as it relies on the identification of a very small subset of genes. This requires a high level of coverage of the genome in the bin as well as relying on the assumption that these genes are truly essential or single copy number across the entire prokaryotic domain, which may not be the case in some uncharacterised taxonomic groups upon which this binning tool is most useful (Albertsen et al, 2013).

4.4.2 Pathway analysis with KEGG

Mapping of the CDSs associated with methanogenesis identified by KAAS to their relative bins provided a list of bins putatively containing organisms with methanogenic function (Figure 4.2). This identified many more bins associated with methanogenesis than would be expected from the taxonomic abundance surveys presented in Chapters 2 and 3. In particular bin 1 showed a very high number of genes associated methanogenesis according to KEGG pathways. However, it did not show the presence of the core methanogenesis marker genes *mcrα* or *mcrβ* and was identified as representing predominantly Proteobacterial sequences by taxonomic classification of the bin scaffolds (Supplementary Table 1). When the

methanogenic genes from this bin were mapped to the KEGG methanogenesis pathway the majority appeared in the peripheral regions of the map linking to acetate conversion and central carbon metabolism suggesting that the organism in this bin is likely not involved in methanogenesis but has enzymes which are shared with the methanogenic pathways (Supplementary Figure 2). This suggests that this method of identifying the functional capability of bins with KAAS and KEGG may provide a high degree of false positives for some functionalities such as methanogenesis. This is supported by the observation that only 8 out of the 37 bins identified by this method were also identified as containing either one or both of the *mcrα* and *mcrβ* genes by identification with an HMM search.

4.4.2 Pathway analysis with Pfam HMM searches

Results from the HMM search for key methanogenesis genes indicated a number of methanogenic bins more in line with abundance observations from Chapters 2 and 3 and were also more in accordance with the taxonomic classifications assigned to the bins (Table 4.1). However, some bins such as 103 and 165 contain only one copy of either the *mcrα* or *mcrβ* genes and no other methanogenic markers and could represent contamination of a non-methanogen bin. This analysis indicates that several bins contain multiple copies of the genes responsible for methylotrophic, hydrogenotrophic and acetoclastic methanogenesis, a characteristic only reported to be seen within the order Methanosarcinales. This suggests that bins 46 and 47 belong to the Methanosarcinales order within the class Methanomicrobia and can perform all three types of methanogenesis. The high copy number of methylotrophic genes is in line with what has been reported in the literature for members of the Methanosarcinales which have been shown to contain genes for multiple methylotrophic pathways in multiple copy number (Pritchett and Metcalf, 2005). In other bins such as bins 28, 41 & 48 the *mcrα/β* genes are found as well as genes involved in acetate and/or CO_2 utilisation but not methylamine utilisation. This is expected for members of the Methanomicrobiales which are within the class Methanomicrobia and are not known to perform methyl group removing methanogenesis (Liu and Whitman, 2008). This means that the MP09 consortium contains members capable of methanogenesis from all three of the methanogenic pathways, the acetoclastic, hydrogenotrophic and methylotrophic pathways. It should also be noted that the scaffolds on which these genes were identified were often quite large and appeared to contain whole operons associated with methanogenesis as well as multiple other functional genes (Figure 4.4). This means that there is the potential to examine entire operon structures if not entire genomes with this dataset to gain insights into the potential regulation and structuring of pathway genes within organisms in this community.

Chapter 5. Imaging and taxonomic analysis of microbial coal colonisation over a time-series

5.1 Introduction

5.1.1 Culturing of a closed consortium system

While many studies have looked at coal seam communities directly as environmental samples or as established consortia nothing is known about how microbial communities colonise a coal surface and establish into a mature community. In addition, as the initial stages of coal degradation have been shown to be the rate limiting step in methanogenesis from coal, it is possible that the organisms responsible for the initial coal degradation process are present in low abundances relative to the whole mature community (Strapoc et al, 2008). This makes it difficult to study the initial degraders through metagenomic or amplicon studies of mature communities. By having a microbial consortium containing all the members necessary to convert coal to methane we can inoculate and culture the community on a coal surface reproducibly and so study the colonisation process of this community. Here highly polished coal disks were inoculated with the mature MP09 community to examine the colonisation of the coal surface over time using both microscopy and DNA sequencing techniques.

5.1.2 Microscopic analysis of microbial coal disk colonisation

Previously coal associated microbes have been difficult to image due to the highly uneven porous surface of coal. Here this issue was overcome by culturing the MP09 community on highly polished coal disks used for petrological studies of vitrinite reflectance. This novel approach enabled imaging the coal surface using SEM over a time-course of colonisation and maturation. The aim of this research was to identify the point at which colonisation starts to occur, how many and what morphotypes are present at early colonisation and later stages of community maturation and how the different morphotypes physically interact with each other and the coal surface at different stages of colonisation. This information directed subsequent DNA sequencing efforts to target the first colonisers as well as determining additional time points of interest where the community appears to undergo change. The use of microscopy also makes it possible to look for examples of microbial interactions which may play roles in colonisation, adhesion, biofilm formation or cell-cell communication and can be investigated more thoroughly via DNA-based analysis. This is of importance as coal degrading methanogenic communities are known to rely on syntrophic interactions to survive and grow under these conditions although the exact interactions remain unknown (Stams, 1994).

5.1.2 DNA analysis of MP09 members from different stages of coal disk colonisation

All previous DNA based investigations of coal seam communities have focussed on the mature communities and have predominantly utilised studies of taxonomic marker genes. This has meant that very little is known about the taxonomic groups or functional genes present at different stages of coal colonisation. Here two types of DNA based analyses were performed on samples representing the first colonisers and at subsequent stages where community structure appeared to undergo change. The first is 16S rRNA amplicon sequencing aimed at providing taxonomic classifications and relative abundances at the various stages. This provides information on what organisms are responsible for initial colonisation and utilisation of coal in the community as well as identifying what organisms change in abundance at the different stages of the coal colonisation and maturation process. The second type of DNA analysis is multiple displacement amplification (MDA) of DNA from samples at different time points, sequencing of the amplified DNA and metagenomic analysis of the generated sequence data to identify the functional genes responsible for processes involved in initial coal colonisation and degradation to gain insight into how to these organisms colonise and degrade coal as a carbon and energy source.

5.2 Methods

Figure 5.1 Glass serum vial used for coal disk culturing with polished coal disks held in place with nickel wire prior to addition of medium.

5.2.1 Microbial consortium culturing and sampling on polished coal disks

Small disks of coal ~1cm diameter by 3 mm thickness with one highly polished surface were used for anaerobic culturing of the MP09 microbial consortium with borosilicate glass disks of equivalent dimensions used as controls. Disks were held two at a time vertically in 200 ml glass serum vials with fine nickel wire, as shown in Figure 5.1, so that the polished surface did not accumulate settling sediment. 20 ml of (M9) minimal salts medium containing 100 mg/L NH_4Cl, 400 mg/L K_2HPO_4, 100 mg/L $MgCl_2$, 0.0001% resazurin, 1ml/L SL-10 trace element solution, 250 mg/L Na_2S, 200 mg/L Cysteine HCl and 1ml/L Wolins vitamin solution was added to the serum vials to submerge the coal disks under anaerobic conditions (Wolin et al, 1963). Vials were sealed and incubated in the dark at 30 °C without shaking. Vials were removed for sampling at 0 (pre inoculation), 4, 8, 16, 24, 34 and 48 days after inoculation and

again in a replicate experiment at 8, 27 and 49 days. A single coal disk from each time point selected for microbial cell harvesting was dipped in DNA-free water then washed 5 times with 200 µl of DNA-free water using a micropipette and this elution water stored at -20 °C for DNA analysis. One coal disk from each time point was dipped in sterile water before preparation for scanning electron microscopy (SEM). In addition 1 coal disk from day 8 and day 27 (second replicate) were also rinsed using the same technique as for DNA analysis but the rinse solution was filtered through a 0.2 µm filter which was prepared for SEM. Single borosilicate glass disks were sampled at days 16 and 48 (first replicate) and prepared for scanning electron microscopy (SEM) using the same method as for coal disks.

5.2.2 Scanning electron microscopy of culture samples

Coal disk samples, glass disk samples and coal disk filter samples were treated with 3% gluteraldehyde for 24 to 48 hours at room temperature before being washed for 10 minutes three times with either 0.1 M phosphate buffer or filter sterilized coal seam formation water. Samples were then given sequential 10 minute washes with 20%, 50%, 70%, 80% and 90% ethanol, twice with 100% ethanol, once with a solution of 50% ethanol 50% hexamethyldisilazane (HMDS) and three times with a 100% HMDS solution before air drying for 24 to 48 hours. Samples were mounted on metal stubs then gold plated to a depth of 20 µm with an Emitech K550 gold sputter coater unit. Sample imaging was carried out using a JEOL JSM- 6480 LA scanning electron microscope.

5.2.3 16S rRNA Clone Library

The rinse solution from the cultured day 8 coal disks was concentrated 10 fold by centrifugation of 200 µl of rinse solution at 4 °C and 13,000 rpm for 10 min, removal of 180 µl of the supernatant and resuspension of the pellet in the remaining liquid. 2 µl of 0.1 mm silica beads were added to the sample, then vortexed for 5 min to lyse cells. 2 µl of this concentrated and lysed sample solution was used as the template for PCR amplification with the universal 16S ribosomal RNA marker gene primer sequences 515F & 806R described in 2.2.2. The reaction was performed in a 25 µl volume which contained 200 µM dNTPs, 0.2 µM of each primer, 5.5 mM $MgCl_2$ and 0.625 U of GoTaq® DNA polymerase enzyme (Promega). PCR reactions were carried out with the following thermal cycling conditions: denaturation at 94 °C for 3 min, 35 cycles of denaturation at 94 °C for 45 s, reannealing at 50 °C for 1 min, extension at 72 °C for 90 s and final extension at 72 °C for 10 min. Reaction products were visualized by gel electrophoresis with a 1% agarose gel pre-stained with SYBR safe DNA stain (Life Technologies).

The PCR product generated above was ligated into the pGEM®-T cloning vector by incubation of 50 ng pGEM®-T vector with 15 ng PCR product (insert to vector ratio of 1:3), 5 µl Rapid Ligation Buffer and 3 U of T4 DNA ligase for 1 hour at room temperature as per manufacturer's instructions (Promega). The ligated plasmid was transformed into α-Select Gold Competent Cells (Bioline) following the pGEM®-T Vector Systems (Promega) manufacturer's instructions with the exception that Luria-Bertani (LB) medium was used instead of super optimal broth with catabolite repression (SOC) medium. Transformed cells were plated onto LB amp-xGal-IPTG agar plates made according to the pGEM®-T Vector System (Promega) instructions and incubated at 37 °C overnight.

94 white colonies were spotted off the plate with sterile pipette tips and used directly as the template material for PCR amplification of the plasmid insert sequence. The primers used were M13F (5'-GTAAAACGACGGCCAG-3') and M13R (5'-CAGGAAACAGCTATGAC-3') and are specific for regions either side of the pGEM®-T cloning insert site. The reactions were performed in 25 µl volumes with 200 µM dNTPs, 0.2 µM of each primer, 3.5 mM $MgCl_2$ and 0.625 U of GoTaq® DNA polymerase enzyme (Promega). PCR thermal cycling reactions were carried out as above with the exception that 30 cycles of denaturation and reannealing were used. Reaction products were visualized by gel electrophoresis with a 1% agarose gel pre-stained with SYBR safe DNA stain (Life Technologies).

10 PCR reaction products which showed a band size of ~500 bp were chosen for DNA sequencing and purified using the Wizard® SV Gel and PCR Clean-Up System using the purification by centrifugation protocol (Promega). The concentration of the purified DNA was measured using a NanoDrop spectrophotometer (Thermo Scientific) and products diluted with ultrapure water to within the range of 1 – 1.5 ng/µl, M13F primer was added to a concentration of 0.83 µM. 12 µl of each of these samples were then sent to the Australian Genome Research Facility (AGRF) for Sanger sequencing.

5.2.4 Data processing and analysis of 16S rRNA clone sequences

Sequence matching the pGEM®-T vector and the forward primer were manually trimmed. A Basic Local Alignment Search Tool (BLAST) search of the trimmed sequences against the NCBI ribosomal refseq database was performed and top hits recorded (Zhang et al, 2000). The RDP Naïve Bayesian rRNA Classifier was also used to identify the taxonomy of the trimmed sequences (Wang et al, 2007).

5.2.5 16S rRNA Amplicon Assay

The rinse solution from the day 8 and 24 cultured coal disks was used directly as the template for PCR amplification using the universal 16S ribosomal RNA primers 515F & 806R listed in

2.2.2. PCR amplifications were performed in six replicate 25 µl reactions containing 200 µM dNTPs, 0.2 µM of each primer, 1.25 U of FastStart Taq DNA polymerase enzyme and proofreading enzyme mix (Roche). PCR thermal cycling reactions and visualization were carried out as described in section 5.2.4.

The six replicate PCR products were pooled and purified using the Wizard SV gel and PCR clean up system (Promega) following the manufacturers' instructions. Purified PCR product was sent for 16S rRNA amplicon sequencing with the Illumina MiSeq platform with 300 bp paired-end reads and 20,000 reads/sample. The library preparation and sequencing was performed by Mr DNA (Molecular Research LP) using the 515F/806R primer sequences (Caporaso et al, 2012).

5.2.6 Multiple Displacement Amplification (MDA)

Multiple displacement amplification (MDA) using the GenomiPhi V2 DNA amplification kit (GE Healthcare) was used to amplify the coal disk rinse samples. Briefly, 3 µl of the day 8 coal disk rinse solution was added to 27 µl of the sample buffer with 2 µl of 0.1 mm silica beads and vortexed for 5 min, denatured at 95 °C for 3 min and cooled to 4 °C. Enzyme and reaction buffer were added to the sample solution to a reaction volume of 10 µl in six replicates and incubated at 30 °C for 8 hours, 65 °C for 10 min and stored at -20 °C. Reaction products were visualized on a 1% agarose gel post-stained with GelRed (Biotum). The products from the six replicate reactions were pooled and purified using the Genomic DNA Clean and Concentrator kit (Zymo Research). Purified DNA was sent to the Ramaciotti Centre for Genomics (UNSW) for Nextera library preparation (Illumina) and sequenced using 1 lane of a MiSeq DNA sequencer (Illumina) to generate 250 bp paired-end reads.

5.3 Results

5.2.1 Microbial consortium culturing and sampling on polished coal disks

To examine the colonisation process of coal seam degrading communities on coal, MP09 consortia cultures were used to inoculate discs, which were then incubated over a time-course of 48 days. During this time the culture media appeared to become slightly more opaque and more sediment appeared to accumulate on the bottom of the culture vials.

5.3.2 SEM microscopy of a coal degrading microbial community

To visualise the colonisation process coal disks from the inoculated consortia were sampled at six time points across the culturing period. One coal disk from each time point was imaged with SEM to observe the presence of different morphotypes, their abundances, fine scale localisations and interactions, including development of biofilms, on the polished coal disk surfaces. Representative SEM images from the sampled time points are shown in Figure 5.2.

Images of the control borosilicate glass disks from days 16 and 48 (first replicate) are shown in Figure 5.3 and coal disks from the second replicate experiment for days 8 and 27 are displayed in Supplementary Figure 4.

Rinsing of the coal disks to dislodge microbes for DNA analysis was an important aim of this culturing experiment. Filtrations of coal disk rinses from days 8 and 24 (second replicate) were performed with a 0.2 μm filter and the filter surface and corresponding un-rinsed coal disk imaged to confirm that a representative collection of the community was being removed by the rinsing procedure. Images of the filter surfaces and coal disks from the corresponding time points are shown in Figure 5.3 with common morphotypes identified on each.

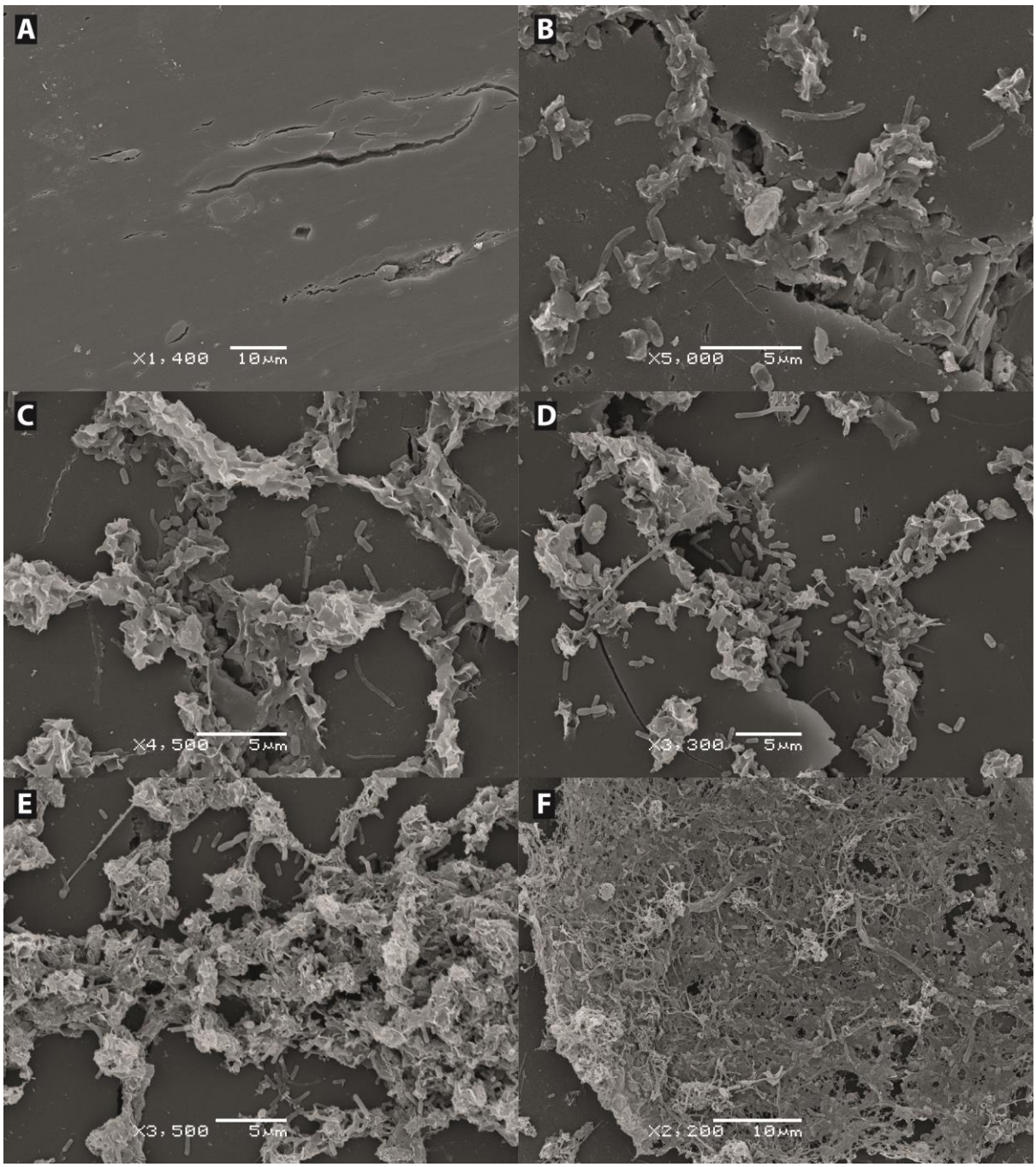


Figure 5.2 Scanning electron microscopy (SEM) images of polished coal disks after inoculation with the MP09 consortium and incubation for 0 days (A), 8 days (B), 16 days (C), 24 days (D), 36 days (E) and 48 days (F) (first replicate).

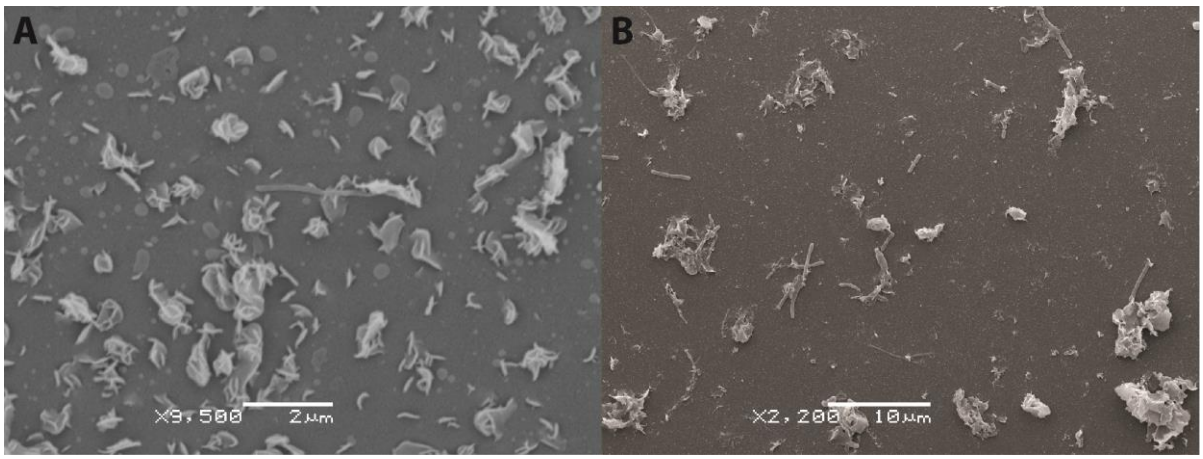


Figure 5.3 Scanning electron microscopy (SEM) images of borosilicate glass disks after inoculation with the MP09 consortium and incubation for 16 days (A) and 48 days (B).

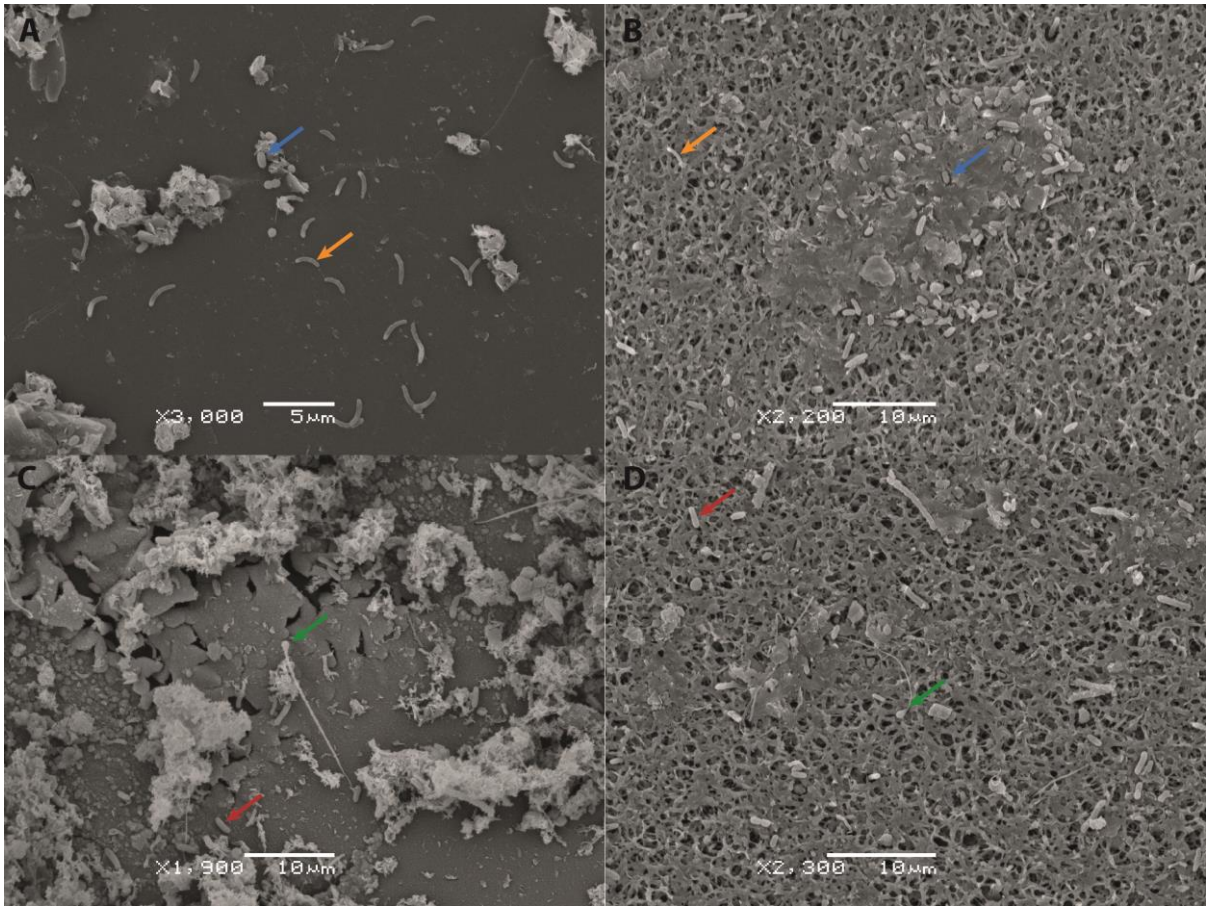


Figure 5.4 Scanning electron microscopy (SEM) images of polished coal disks and 0.2 µm filters from filtration of water after rinsing of the coal disk surface. Images are from coal disks after inoculation with the MP09 consortium and incubation for 8 days (A) and 24 days (C) and filters from their corresponding time points (B) and (C) (respectively) (second replicate). Coloured arrows indicate morphotypes which could be seen on both the coal disk surface and corresponding filter surface.

5.3.3 Observation of characteristic morphotypes

Several morphotypes were observed on the coal disk surfaces whose diversity and abundances changed over time. One morphotype which showed a significant increase in abundance at the final time point could be putatively identified as a spirochaete due to its unique cellular morphology, which was highly similar to a previously described free living spirochaete isolated from an anaerobic lake sediment (Holt and Canale-Parola, 1968). A comparison of the structures observed on the day 48 coal disk sample with those from (Holt and Canale-Parola, 1968) is shown in Figure 5.5.

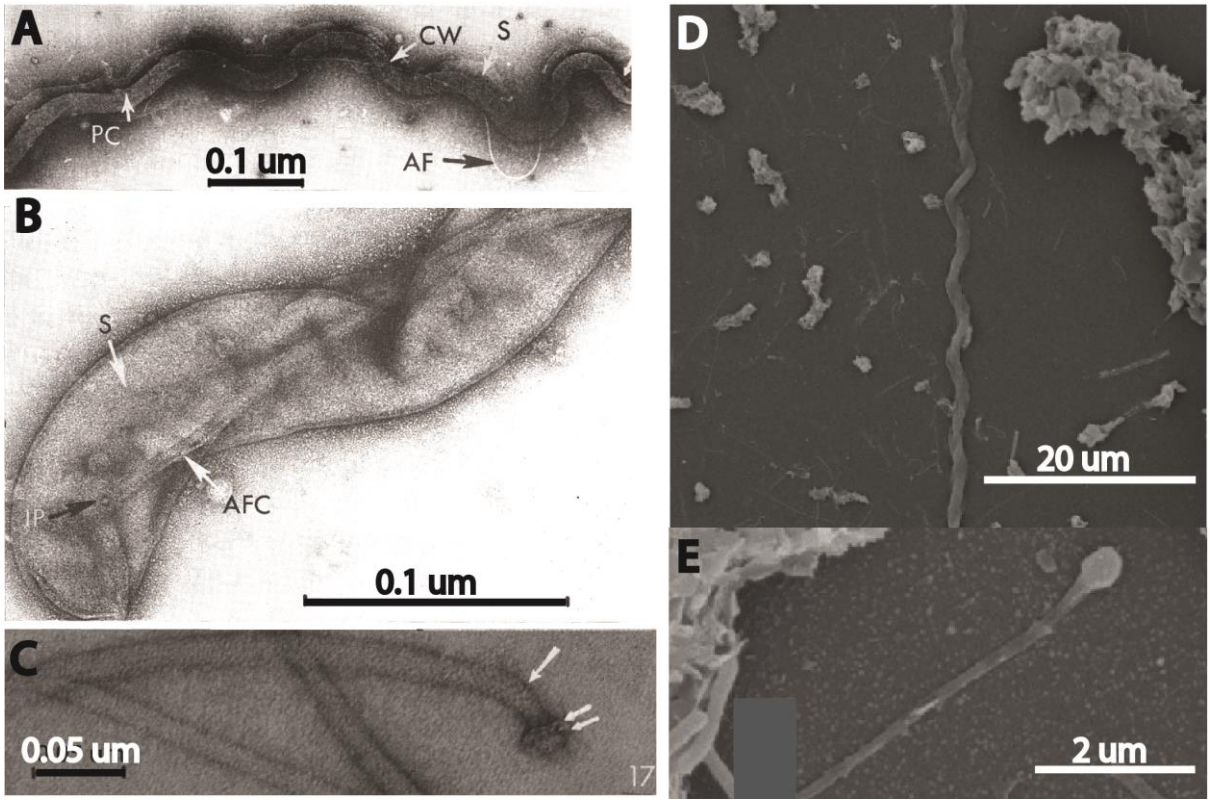


Figure 5.5 Transmission electron microscopy (TEM) images of the cell morphology of a free living spirochaete (A & B) and the attachment end of the spirochaete flagellum separated from the cell and outer sheath (C) (adapted from (Holt and Canale-Parola, 1968)). Scanning electron microscopy (SEM) images of a putative spirochaete (D) and a putative spirochaete flagellum (E) observed on a polished coal disk after 48 days of incubation.

5.3.4 Taxonomic analysis of 16S rRNA clone library sequences

To provide an initial insight into the organisms present at the early stages of coal colonisation PCR amplification of the 515F/806R 16S rRNA region was performed on the coal disk rinse from day 8 (Figure 5.5) and a clone library constructed from the products. Ten clones were sequenced and identified based on BLAST searches against the NCBI ribosomal refseq database as well as classification using the RDP classifier (Table 5.1). The 10 sequences showed a low level of taxonomic diversity with seven showing closest identity to the order

Desulfuromonadales, most closely related to *Geobacter grbiceae*. PCR amplification of the 16S RNA region was performed again for day 8 and 24 (Figure 5.7) using the 515F/806R primers and sent for sequencing however, data was not received in time for inclusion in this thesis. Similarly MDA products of the day 8 coal disk rinse were generated (Figure 5.8) and sent for sequencing on the Illumina MiSeq platform but this data was also not received in time for inclusion.

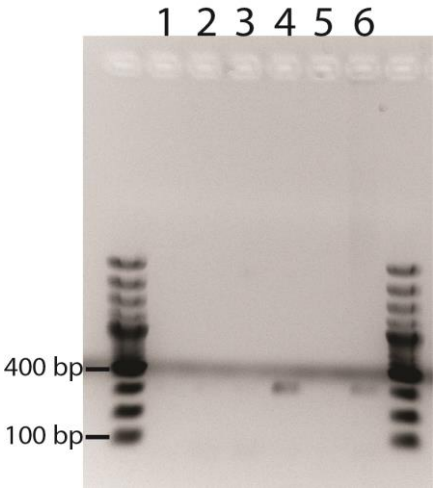


Figure 5.6 Visualisation of a PCR product from 16S rRNA amplification of day 8 rinse used for construction of clone library (lane 4), positive control lane 6, negative control lane 5.

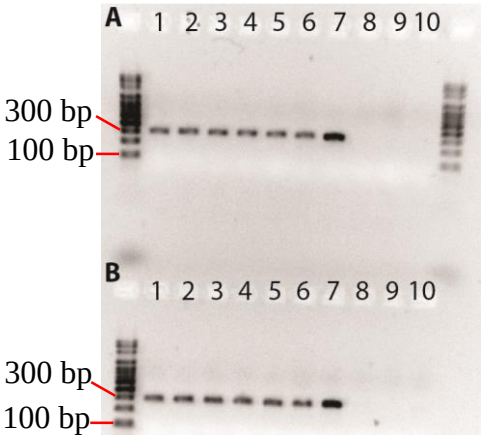


Figure 5.7 16S rRNA PCR amplification products from day 8 (A, lanes 1 to 6) and 24 (B, lanes 1 to 6) rinses.

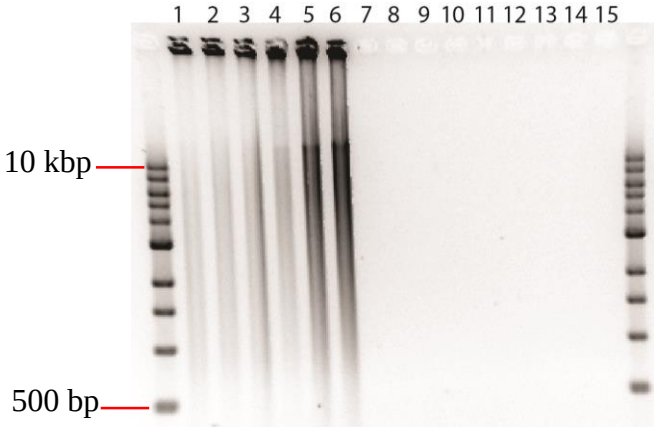


Figure 5.8 Day 8 MDA product lanes 1 to 6, failed MDA lanes 7 to 12 and negative controls lanes 13 to 15.

Table 5.1 Best match taxonomic identification of 16S rRNA clone library sequences derived from day 8 coal disk cells. Basic Local Alignment Search Tool (BLAST) assignments indicate best matches at the species level and RDP Naïve rRNA classifier assignments indicate closest taxonomic assignment at the order level.

	BLAST results against NCBI RNA refseq sequence database						RDP Naive Bayesian rRNA Classifier assignments	
sample	Best Match	Max score	Total score	Query cover	E value	Ident	Order	Confidence level
c1	*Geobacter grbiciae*	412	412	68%	3.00E-112	94%	Desulfuromonadales	94%
c2	*Campylobacter lari*	424	424	68%	1.00E-115	95%	Campylobacterales	89%
c3	*Geobacter grbiciae*	424	424	67%	1.00E-115	95%	Desulfuromonadales	91%
c4	*Geobacter grbiciae*	418	418	67%	6.00E-114	94%	Desulfuromonadales	91%
c5	*Desulfovibrio piezophilus*	379	379	68%	3.00E-102	92%	Desulfovibrionales	100%
c6	*Geobacter grbiciae*	407	407	67%	1.00E-110	93%	Desulfuromonadales	84%
c7	*Geobacter grbiciae*	412	412	68%	3.00E-112	94%	Desulfuromonadales	94%
c8	*Geobacter grbiciae*	418	418	67%	6.00E-114	94%	Desulfuromonadales	92%
c9	*Campylobacter lari*	435	435	67%	6.00E-119	95%	Campylobacterales	82%
c10	*Geobacter grbiciae*	412	412	67%	3.00E-112	94%	Desulfuromonadales	91%

5.4 Discussion

5.4.1 Scanning electron microscopy of polished coal disk cultures

Results from imaging of the cultured coal disks (Figure 5.2) showed that at day 0 (prior to inoculation) there were no cells observed on the surface of the coal disk. This indicates that there were no native microbial cells on the coal disks and cells seen later in the incubation originate from the MP09 community. By day 8, single cells and microbial biofilm-like structures had appeared on the coal surface, however a limited number of morphotypes were observed, being primarily 2 – 4 µm rods. The low level of morphotype diversity at day 8 when cells are first observed suggests that only a small number of organisms are playing an important role in initial colonisation (Figure 5.2).

By day 16 a number of new morphotypes had appeared and the cell and biofilm density had increased. Morphotypes which had appeared at this time point include longer narrower rods as well as small 0.5-1 µm cocci most commonly associating with the biofilm and crevices in the coal surface. This indicates that the initial colonisation has taken place and the additional morphotypes may represent other primary or secondary fermenters, which are starting to utilise the degradation products (Figure 5.2).

Samples imaged from days 24 and 36 show little change in the observable morphotypes and their relative abundances but a steady increase in total cellular abundance and biofilm level could be observed. At day 48 the biofilm and cell numbers had increased again and new

morphotypes were observed on the coal surface (Figure 5.2). From these results it appears that the community structure undergoes changes at multiple points in time suggesting distinct stages of colonisation and maturation starting at some point in the first 8 days. The observation of changes at day 48 is interesting and could mean that the community maturation process may not have been completed by the end of sampling. This means that sampling over a longer time period should be performed in any future repeat experiments.

Imaging of the coal disk rinse filtrations at days 8 and 24 shows the same morphotypes that were observed on un-rinsed coal disks at the corresponding time points. This indicates that the rinsing process is an appropriate method for sampling the cells on the coal disk and that there does not appear to any major biases in the morphotypes which are dislodged for DNA analysis by the rinsing (Figure 5.4). This gives increased confidence in the representative nature of taxonomic identifications which are made by DNA analysis of the coal disk rinses.

Similar morphotypes and trends were observed in the SEM images of the day 8 and 27 replicate coal disks with the exception that the community appeared to be at a more mature stage at day 8 in comparison to the original replicate (Supplementary Figure 3). This difference in colonisation or maturation rate could be due to stochastic events associated with inoculation with a complex community or due to some differences in the coal structure and chemistry between the disks allowing for more rapid development. Further replication of this experiment would help to determine the level of variability seen in the colonisation rates and precise petrological analysis of the coal disks prior to culturing may help to explain if coal structure and chemistry is having an observable effect on colonisation rate.

5.4.2 Putative identification of a free living spirochaete on coal disk cultures

What appears morphologically to be several components of a spirochaete cell were observed on a coal disk incubated for 48 days (Figure 5.5). The morphological features closely match those made by (Holt and Canale-Parola, 1968) of a free living spirochaete cultured from an anaerobic lake sediment, another methanogenic environment which shares several physicochemical similarities with a coal mine. The fine strands with bulbous ends appear to be spirochaete flagella which sit outside the cell but within an outer sheath. The observation of these strands independent of the rest of the cell can be explained by the very easy degradation of the outer sheath reported in (Holt and Canale-Parola, 1968). The observation of this organism is supported by the 16S rRNA amplicon survey presented in Chapter 2 and the metagenomic taxonomic identifications presented in Chapter 3, both of which identified Spirochaetes as an abundant taxa in the mature MP09 community. The microscopic observation that these organisms appear to become abundant only at a late time point in

cultivation suggest that analysis of the 16S rRNA amplicon sequencing, when possible, may enable taxonomic identification of the putative spirochaete and confirm it as one of those seen in the mature community. In addition, the late appearance of these cells in the development of this community suggests this taxonomic group play a role in the later stages of maturation of this community. This would be an interesting finding as the Spirochaetes do not appear as community members in the majority of previously described coal seam communities (Figure 1.3). However, they have been demonstrated to be responsible for homoacetogenesis in methanogenic termite guts, which utilises the final degradation products before methane (Figure 1.1) (Graber and Breznak, 2004, Leadbetter et al, 1999, Pester and Brune, 2006). This could suggest that the spirochaetes are playing a role in shuttling carbon between the CO_2 and acetate carbon pools before their utilisation by the methanogenic archaea. Normally the homoacetogens are outcompeted by hydrogenotrophic methanogens but under some conditions such as low temperatures or high acidity the homoacetogenic reaction becomes favourable and outcompetes the methanogenic reaction, shifting the community to acetoclastic methanogenesis (Lee and Zinder, 1988, Schink, 1997).

5.4.3 Analysis of DNA from cultured coal disk samples

While clone library generation and sequencing from day 8, 16S rRNA amplicon sequencing from days 8 and 24 and metagenomic sequencing from MDA amplified day 8 DNA were performed only results from the clone library sequencing were available for inclusion in this thesis. The analysis of the 16S rRNA sequences generated from the clone library provide a preliminary insight into the taxonomic groups which likely predominate in the early stages of coal colonisation (Table 5.1). The observation of a majority of sequences classifying as part of the Desulfuromonadales order, and most closely relating to the family Geobacter within that, is interesting as this group is reported to be capable of utilising sulfur, nitrate, organic molecules and metal ions as electron receptors for anaerobic respiration, as well as having the capability to perform several types of fermentation (Butler et al, 2009). The Campylobacterales also identified by this analysis have been shown to be similarly capable of using sulfur compounds as an energy source (Nakagawa and Takaki, 2001). This suggests that the first colonisers may be utilising sulfur, nitrate or metal ion compounds present in the coal as terminal electron donors with anaerobic respiration switching to methanogenesis once these preferential electron accepters are depleted. The Geobacteraceae are also known for their ability to degrade a range of hydrocarbons in the absence of O_2, which in conjunction with their appearance at early colonisation stages could indicate that they are also responsible for degradation of coal components (Butler et al, 2009). Analysis of the MDA amplified

metagenome from day 8 should provide further insight into the capabilities of these organisms through identification of functional genes associated with these various metabolic processes in the metagenome. One interesting point to note was that taxonomic classification only had a high confidence level to the order level after which the classification confidence level from the RDP classifier dropped significantly, indicating that these organisms may not be best described as Geobacter but may form a new clade within the Desulfuromonadales order (Table 5.1).

Chapter 6. Final discussion and conclusions

6.1 Discussion

Understanding how microbial communities grow and produce methane in coal seams is of interest from both an academic and industrial perspective, as methane is both a potent greenhouse gas and valuable fuel source. In addition, coal seam communities present novel methanogenic communities as they inhabit an ecological niche in which the carbon source is a complex poly-aromatic matrix in contrast to the biopolymer based carbon sources seen in animal gut and sediment environments. This study aimed to examine a microbial consortium, sourced from an Australian coal seam, capable of methane generation from coal as the sole carbon source. The community was examined both in its mature state and over a period of colonisation and maturation on a fresh coal source, in terms of its taxonomic and functional gene composition and the colonisation and maturation steps it undergoes.

Investigations into the taxonomic identities and relative abundances of organisms in the MP09 community were performed using a 16S rRNA amplicon survey and two types of metagenomic analysis. This showed that the bacterial and archaeal taxa in the MP09 consortium were broadly similar to those previously presented in the literature and showed most similarity to coal seam communities from Illinois (Strapoc et al, 2008) and Wyoming (Green et al, 2008) in the USA and the Gippsland basin in Australia (Midgley et al, 2010) with which it shared the dominance of the Firmicutes and presence of Spirochaetae and Tenericutes phlya. This similarity with distant geographically isolated coal seams has previously been observed in the literature (figure 1.1) and is interesting in terms of how the communities have originated and adapted to different coal seams. It is hypothesised that there may be a core group of organisms which are necessary for methanogenesis in all coal seams and other groups which can participate in functional roles only when specific conditions are met.

The level of variation in relative abundances of different taxonomic groups observed between the different DNA extraction protocols was considerable (Figure 2.2 & Table 3.1). This highlights the importance of taking taxa specific DNA extraction efficiencies into account for community abundance studies, particularly where a diverse range of organisms are being sampled which will possess a range of different cell membrane structures. There was also a certain amount of variation seen between the taxonomic identifications and relative abundance measures generated by metagenomics tools compared to the 16S rRNA amplicon survey (Figures 2.2 and 3.1). One main difference observed was an increase in the abundance

of archaea reported in the metagenomic survey, suggesting a 16S rRNA gene copy number bias and/or that the 515F/806R primer set has an amplification bias towards the bacterial 16S rRNA sequences. Another observation supporting the effect of a 16S rRNA gene copy number bias was the Firmicutes appearing less abundant and the Proteobacteria increasing in abundance in the metagenomic survey, as the Firmicutes are known to commonly exhibit high 16S rRNA copy number (Větrovský and Baldrian, 2013).

Until very recently, studies of coal seam communities had focussed primarily on the taxonomic components of the communities and the functional genetic components were inferred through taxonomy. A recently published study (Ghosh et al, 2014) looking at a coal seam formation water sample from India includes a metagenomic component comprised of a preliminary taxonomic and functional analysis of this community, based on the automated MG-RAST pipeline (Meyer et al, 2008). This is the first example of application of metagenomic techniques to a coal seam community, however the investigation into function is limited to examination of the prevalence of aromatic hydrocarbon degradation pathway genes in the community. Findings indicated that the microbial community was typical of many coal seams previously examined (Figure 1.3) with Proteobacteria dominating and other common community members such as the Bacteroidetes, Planctomycetes and Firmicutes present at lower levels. A number of genes were also identified as part of the peripheral and central aromatic compound degradation pathways, an observation which was not shared with the results presented here, where very few genes associated with these pathways were observed (Ghosh et al, 2014). This difference could be due in part to the different methodology utilised by (Ghosh et al, 2014) which was based upon the automated MG-RAST pipeline.

In this work an attempt was made to go further than identifying functions belonging to the community by connecting the functions identified to particular taxonomic groups in the community. This was done by using a novel taxonomic-binning method based on differential coverage between multiple DNA extractions from the same community (Albertsen et al, 2013, Imelfort et al, 2014). This functional analysis of the mature MP09 metagenome focuses on methanogenesis as it represents the best studied functional component of coal seam communities and so presented a good target to demonstrate that the taxonomic binning process presented here can be used in conjunction with taxonomic identification tools (Chapters 2 and 3) and functional gene analysis (Chapters 3 and 4) to assign specific functions to taxonomy. We were able to show successful taxonomic binning of several members of the Methanomicrobiales and Methanosarcinales, the two most abundant archaeal

orders with meaningful examinations of the functional capabilities demonstrated through knowledge of the taxonomic specificity of the methanogenic pathways which are highly conserved within the archaeal orders (Table 4.1) (Garcia et al, 2000, Liu and Whitman, 2008).

Here investigation into the methanogenesis pathways has been performed. However, other pathways could also be investigated using the same approach and dataset for other pathways thought to be important in methanogenic communities such as homoacetogenesis, sulphate or nitrate reduction or various carbon utilisation and degradation pathways, to get a more complete picture of which organisms are responsible for different community functions (Berdugo-Clavijo et al, 2012, Elferink Oude et al, 1994, Fuchs et al, 2011, Hirschler-Réa et al, 2012, Stams, 1994).

Another major component of this work has been the microscopic and DNA-based examination of the MP09 community as it colonises and matures on a coal surface. The previously described work is performed on a "mature" consortium which has presumably reached a state of equilibrium in terms of microbial abundances and the flow of metabolites. By inoculating fresh coal disks with this community and examining it over a time-series, this work provides the first observations on how coal seam community members start to colonise coal as a new carbon and energy source. This is valuable as it is these stages that researchers must alter if attempting to stimulate the generation of CSG generation through amendments with chemical additions or exogenous microbial inoculations (Huang, 2013, Jin et al, 2010, Jones et al, 2010, Strąpoć et al, 2011).

6.2 Conclusions

Taxonomic analysis with a 16S rRNA amplicon survey and metagenomic sequencing shows that the coal degrading, methanogenic consortium MP09 represents a complex microbial community comprised of both archaeal and bacterial representatives. This community shares much in common with other previously described coal seam communities from geographically isolated coal seams. The community contains the core taxonomic groups identified in the majority of coal seam communities including the Proteobacteria, Bacteroidetes and Firmicutes although it also contains some phyla such as the Spirochaetae and Tenericutes which were less commonly found in other coal seam communities (Green et al, 2008, Midgley et al, 2010, Strapoc et al, 2008). Metagenomic analysis of the functional genes in the community identified the presence of all three methanogenic pathways and placed them in taxonomic bins within the archaeal orders Methanomicrobiales and Methanosarcinales, suggesting that the community is capable of hydrogenotrophic, acetoclastic and methylotrophic methanogenesis by a small number of archaeal groups.

Culturing experiments on polished coal disks provide the first reported insight into the coal colonisation process by a coal degrading methanogenic community. Several distinct stages of coal colonisation by the MP09 community were identified. A low level of morphotype diversity at the earliest stages, based on preliminary DNA analysis, suggested the dominance of a group within the Desulfuromonadales order. This was followed by an increase in abundance and diversity over several weeks and finally a late change in the abundances of various morphotypes after 48 days of incubation. These experiments also highlighted multiple interactions between the coal, microbes and biofilm that can be investigated by analysis of DNA sequences from those time points when this sequence data becomes available. Results and data from this study provide valuable information regarding the taxonomic, genetic and functional aspects of coal degrading methanogenic communities, including one of the first metagenomes and the first investigation into the coal colonisation process. This is an important step in understanding the microbial ecology underpinning biogenic generation of CSG, an emerging fuel source with widespread environmental and industrial importance.

6.3 Future directions

The generation of metagenomes from an increasing number of environmental and cultured coal seam communities, paired with collection of metadata such as functional capabilities and physicochemical characteristics of the environment, will enable future investigations to focus on comparative metagenomic approaches. This has great potential for discovering the genetic components behind community functions as the presence or absence of sequences shared between communities can be examined to explain any functional differences between the communities. Comparison of communities can also be based on the presence and abundance of functional genes or gene pathways rather than taxonomy, which has previously been used as a proxy for function in studies performing 16S rRNA surveys. This approach will begin to answer the question of which genes, and by inference enzymes, are common in all coal degrading methanogenic communities and which are necessary only under specific conditions and how this compares to taxonomic observations of the community.

The time-course culturing experiment presented in this work could be used to produce additional metagenomes which represent the community at particular points in its colonisation and maturation. As the early stages of colonisation are likely to be dominated by organisms performing initial solubalisation of the coal, comparisons between the metagenome of this time point and a mature community could provide a good method to search for enzymes involved in the solubilisation of coal, the first step in the coal degradation pathway. This could be performed by looking for genes or gene classes which appear overrepresented in the

metagenomes of the earliest stages of culturing in comparison to the later or mature stages. This could be performed without any specific target genes by looking at all overrepresented genes or it could be directed towards genes which have been previously suggested to play a role such as those identified by (Strachan et al, 2014) who screened a fosmid library built from a coal seam metagenome for lignin degradation phenotypes (Strachan et al, 2014). These culturing experiments would also complement additional metagenomic studies and be informative in seeing if the same stages of coal colonisation and maturation are undertaken by all consortia. As these communities are reliant on syntrophic interactions it would also be of particular interest to see if the same interactions occur between morphotypes, biofilm and the coal surface in different communities (Stams, 1994).

Another tool which could be employed in culture based experiments on coal seam communities is stable isotope probing (SIP). This can be used for investigating the utilisation of various compounds by different members of the community and has previously been used in the discovery of aromatic carbon utilisers from other non-coal seam methanogenic environments (Zhang et al, 2012). However, it should be noted that this tool requires isotopically labelled target substrates for use and so would be limited to commercially available labelled substrates, making this method useful for identifying the members performing the intermediate steps in coal degradation to methane, but not the initial coal degraders.

In addition to the metagenomics approach taken in this study further '–omics' experiments such as metatranscriptomics and metaproteomics would be useful in gaining a greater understanding of how these communities are functioning. While metagenomics provides information only on the genetic capabilities of a community metatranscriptomics can give information on the functions which are actually being performed at a certain point in time. This approach would be especially useful performed in conjunction with the coal disk culturing experiments as it would identify which genes are being actively expressed during the periods of colonisation and maturation observed in Chapter 5, without prior knowledge of specific gene sequences (Warnecke and Hess, 2009). While metaproteomics can also be used to answer these questions, it provides less data than DNA based techniques and is technically more demanding. However, it may be useful for identifying the enzymes responsible for coal solubilisation, the initial step in coal utilisation, through looking at the excreted proteins during coal colonisation (Wilmes et al, 2008).

Reference List

Ahmed M, Smith JW (2001). Biogenic methane generation in the degradation of eastern Australian Permian coals. *Organic Geochemistry* 32**:** 809-816.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 31**:** 533-538.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* 215**:** 403-410.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19**:** 455-477.

Beckmann S, Lueders T, Krüger M, Von Netzer F, Engelen B, Cypionka H (2011). Acetogens and acetoclastic Methanosarcinales govern methane formation in abandoned coal mines. *Appl Environ Microbiol* 77**:** 3749-3756.

Beeston J (1986). Coal rank variation in the Bowen Basin, Queensland. *International journal of coal geology* 6**:** 163-179.

Berdugo-Clavijo C, Dong X, Soh J, Sensen CW, Gieg LM (2012). Methanogenic biodegradation of two-ringed polycyclic aromatic hydrocarbons. *FEMS Microbiol Ecol* 81**:** 124-133.

Butler JE, Young ND, Lovley DR (2009). Evolution from a respiratory ancestor to fill syntrophic and fermentative niches: comparative genomics of six *Geobacteraceae* species. *BMC Genomics* 10**:** 103.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal* 6**:** 1621-1624.

Chang W, Um Y, Holoman TRP (2006). Polycyclic aromatic hydrocarbon (PAH) degradation coupled to methanogenesis. *Biotechnol Lett* 28**:** 425-430.

Conrad R, Wetter B (1990). Influence of temperature on energetics of hydrogen metabolism in homoacetogenic, methanogenic, and other anaerobic bacteria. *Arch Microbiol* 155**:** 94-98.

Daniels L, Fuchs G, Thauer R, Zeikus J (1977). Carbon monoxide oxidation by methanogenic bacteria. *J Bacteriol* 132**:** 118-126.

Darling AE, Jospin G, Lowe E, Matsen Iv FA, Bik HM, Eisen JA (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2**:** e243.

Dawson KS, Strapoc D, Huizinga B, Lidstrom U, Ashby M, Macalady JL (2012). Quantitative Fluorescence In Situ Hybridization Analysis of Microbial Consortia from a Biogenic Gas Field in Alaska's Cook Inlet Basin. *Appl Environ Microbiol* 78**:** 3599-3605.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* 99**:** 5261-5266.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, et al (2002). The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *Journal of molecular microbiology and biotechnology* 4**:** 453-61.

Dridi B, Henry M, El Khéchine A, Raoult D, Drancourt M (2009). High prevalence of *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* detected in the human gut using an improved DNA detection protocol. *PLoS One* 4**:** e7063.

Eddy SR (1998). Profile hidden Markov models. *Bioinformatics* 14**:** 755-763.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26**:** 2460-2461.

Edgar RC (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods* 10**:** 996-998.
Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27**:** 2194-2200.
Elferink Oude SJ, Visser A, Hulshoff Pol LW, Stams AJ (1994). Sulfate reduction in methanogenic bioreactors. *FEMS Microbiol Rev* 15**:** 119-136.
Faiz M, Saghafi A, Sherwood N, Wang I (2007). The influence of petrological properties and burial history on coal seam methane reservoir characterisation, Sydney Basin, Australia. *International Journal of Coal Geology* 70**:** 193-208.
Falkowski P, Scholes R, Boyle EEA, Canadell J, Canfield D, Elser J, et al (2000). The global carbon cycle: a test of our knowledge of earth as a system. *Science* 290**:** 291-296.
Forster P, Ramaswamy V, Artaxo P, Berntsen T, Betts R, Fahey DW, et al 2007. Changes in atmospheric constituents and in radiative forcing. Chapter 2. *Climate Change 2007. The Physical Science Basis.*
Franzmann P, Springer N, Ludwig W, Conway De Macario E, Rohde M (1992). A Methanogenic Archaeon from Ace Lake, Antarctica: *Methanococcoides burtonii sp.* nov. *Syst Appl Microbiol* 15**:** 573-581.
Franzmann PD, Liu Y, Balkwill DL, Aldrich HC, De Macario EC, Boone DR (1997). *Methanogenium frigidum sp*. nov., a psychrophilic, H_2-using methanogen from Ace Lake, Antarctica. *Int J Syst Bacteriol* 47**:** 1068-1072.
Fry JC, Horsfield B, Sykes R, Cragg BA, Heywood C, Kim GT, et al (2009). Prokaryotic populations and activities in an interbedded coal deposit, including a previously deeply buried section (1.6–2.3 km) above~ 150 Ma basement rock. *Geomicrobiology Journal* 26**:** 163-178.
Fuchs G, Boll M, Heider J (2011). Microbial degradation of aromatic compounds—from one strategy to four. *Nature Reviews Microbiology* 9**:** 803-816.
Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, Fitzhugh W, et al (2002). The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res* 12**:** 532-542.
Garcia J-L, Patel BK, Ollivier B (2000). Taxonomic, Phylogenetic, and Ecological Diversity of Methanogenic *Archaea*. *Anaerobe* 6**:** 205-226.
Ghosh S, Jha P, Vidyarthi AS (2014). Unraveling the microbial interactions in coal organic fermentation for generation of methane—A classical to metagenomic approach. *International Journal of Coal Geology* 125**:** 36-44.
Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols* 2010**:** pdb. prot5368.
Gorby YA, Yanina S, Mclean JS, Rosso KM, Moyles D, Dohnalkova A, et al (2006). Electrically conductive bacterial nanowires produced by *Shewanella oneidensis* strain MR-1 and other microorganisms. *Proc Natl Acad Sci U S A* 103**:** 11358-11363.
Graber JR, Breznak JA (2004). Physiology and nutrition of *Treponema primitia*, an H_2/CO_2-acetogenic spirochete from termite hindguts. *Appl Environ Microbiol* 70**:** 1307-1314.
Green MS, Flanegan KC, Gilcrease PC (2008). Characterization of a methanogenic consortium enriched from a coalbed methane well in the Powder River Basin, USA. *International Journal of Coal Geology* 76**:** 34-45.
Guo H, Liu R, Yu Z, Zhang H, Yun J, Li Y, et al (2012). Pyrosequencing reveals the dominance of methylotrophic methanogenesis in a coal bed methane reservoir associated with Eastern Ordos Basin in China. *International Journal of Coal Geology* 93**:** 56-61.
Haack SK, Breznak JA (1993). Cytophaga xylanolytica sp. nov., a xylan-degrading, anaerobic gliding bacterium. *Arch Microbiol* 159**:** 6-15.

Hedderich R, Whitman WB 2006. Physiology and biochemistry of the methane-producing *Archaea*. *The prokaryotes.* Springer.
Hedderich R, Whitman WB 2013. Physiology and Biochemistry of the Methane-Producing *Archaea*. *The Prokaryotes.* Springer.
Hirschler-Réa A, Cravo-Laureau C, Casalot L, Matheron R (2012). Methanogenic octadecene degradation by syntrophic enrichment culture from brackish sediments. *Curr Microbiol* 65**:** 561-567.
Holt S, Canale-Parola E (1968). Fine structure of *Spirochaeta stenostrepta*, a free-living, anaerobic spirochete. *J Bacteriol* 96**:** 822.
Huang Z. (2013). *An evaluation of pretreatment agents for the stimulation of secondary biogenic coalbed natural gas.* Ph.D., University of Wyoming.
Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW (2014). GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2:e603.
Jin S, Bland AE, Price HS (2010). Biogenic methane production enhancement systems. Google Patents.
Jones EJ, Voytek MA, Corum MD, Orem WH (2010). Stimulation of methane generation from nonproductive coal by addition of nutrients or a microbial consortium. *Appl Environ Microbiol* 76**:** 7013-7022.
Jones W, Leigh J, Mayer F, Woese C, Wolfe R (1983). *Methanococcus jannaschii sp*. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Arch Microbiol* 136**:** 254-261.
Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32**:** D277-D280.
Kembel SW, Wu M, Eisen JA, Green JL (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* 8**:** e1002743.
Klappenbach JA, Dunbar JM, Schmidt TM (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* 66**:** 1328-1333.
Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001). rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res* 29**:** 181-184.
Kotsyurbenko OR, Glagolev MV, Nozhevnikova AN, Conrad R (2001). Competition between homoacetogenic bacteria and methanogenic archaea for hydrogen at low temperature. *FEMS Microbiol Ecol* 38**:** 153-159.
Kurr M, Huber R, König H, Jannasch HW, Fricke H, Trincone A, et al (1991). *Methanopyrus kandleri, gen*. and sp. nov. represents a novel group of hyperthermophilic methanogens, growing at 110 C. *Arch Microbiol* 156**:** 239-247.
Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 35**:** 3100-3108.
Leadbetter J, Schmidt T, Graber J, Breznak J (1999). Acetogenesis from H_2 plus CO_2 by *spirochetes* from termite guts. *Science* 283**:** 686-689.
Lee MJ, Zinder SH (1988). Carbon monoxide pathway enzyme activities in a thermophilic anaerobic bacterium grown acetogenically and in a syntrophic acetate-oxidizing coculture. *Arch Microbiol* 150**:** 513-518.
Lee ZM-P, Bussema C, Schmidt TM (2009). rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res* 37**:** D489-D493.
Li D, Hendry P, Faiz M (2008). A survey of the microbial populations in some Australian coalbed methane reservoirs. *International Journal of Coal Geology* 76**:** 14-24.
Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25**:** 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25**:** 2078-2079.

Liu Y, Whitman WB (2008). Metabolic, phylogenetic, and ecological diversity of the methanogenic *archaea*. *Ann N Y Acad Sci* 1125**:** 171-189.

Lueders T, Chin KJ, Conrad R, Friedrich M (2001). Molecular analyses of methyl-coenzyme M reductase α-subunit (mcrA) genes in rice field soil and enrichment cultures reveal the methanogenic phenotype of a novel archaeal lineage. *Environ Microbiol* 3**:** 194-204.

Maeder DL, Anderson I, Brettin TS, Bruce DC, Gilna P, Han CS, et al (2006). The *Methanosarcina barkeri* genome: comparative analysis with *Methanosarcina acetivorans* and *Methanosarcina mazei* reveals extensive rearrangement within *methanosarcinal* genomes. *J Bacteriol* 188**:** 7922-7931.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Grechkin Y, et al (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* 40**:** D123-D129.

Meyer F, Paarmann D, D'souza M, Olson R, Glass EM, Kubal M, et al (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9**:** 386.

Midgley DJ, Hendry P, Pinetown KL, Fuentes D, Gong S, Mitchell DL, et al (2010). Characterisation of a microbial community associated with a deep, coal seam methane reservoir in the Gippsland Basin, Australia. *International Journal of Coal Geology* 82**:** 232-239.

Morgavi D, Forano E, Martin C, Newbold C (2010). Microbial ecosystem and methanogenesis in ruminants. *Animal* 4**:** 1024-1036.

Morita M, Malvankar NS, Franks AE, Summers ZM, Giloteaux L, Rotaru AE, et al (2011). Potential for direct interspecies electron transfer in methanogenic wastewater digester aggregates. *MBio* 2**:** e00159-11.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35**:** W182-W185.

Nakagawa S, Takaki Y (2001). Nonpathogenic Epsilonproteobacteria. *Encyclopedia of Life Sciences.* Chichester: John Wiley & Sons, Ltd.

Nicholson MJ, Evans PN, Joblin KN (2007). Analysis of methanogen diversity in the rumen using temporal temperature gradient gel electrophoresis: identification of uncultured methanogens. *Microb Ecol* 54**:** 141-150.

O'brien JM, Wolkin R, Moench T, Morgan J, Zeikus J (1984). Association of hydrogen metabolism with unitrophic or mixotrophic growth of *Methanosarcina barkeri* on carbon monoxide. *J Bacteriol* 158**:** 373-375.

Papendick SL, Downs KR, Vo KD, Hamilton SK, Dawson GK, Golding SD, et al (2011). Biogenic methane potential for Surat Basin, Queensland coal seams. *International Journal of Coal Geology* 88**:** 123-134.

Paster B, Canale-Parola E (1982). Physiological diversity of rumen spirochetes. *Appl Environ Microbiol* 43**:** 686-693.

Peng Y, Leung HC, Yiu S-M, Chin FY (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28**:** 1420-1428.

Penner TJ, Foght JM, Budwill K (2010). Microbial diversity of western Canadian subsurface coal beds and methanogenic coal enrichment cultures. *International Journal of Coal Geology* 82**:** 81-93.

Pester M, Brune A (2006). Expression profiles of fhs (FTHFS) genes support the hypothesis that spirochaetes dominate reductive acetogenesis in the hindgut of lower termites. *Environ Microbiol* 8**:** 1261-1270.

Phelps T, Zeikus J (1984). Influence of pH on terminal carbon metabolism in anoxic sediments from a mildly acidic lake. *Appl Environ Microbiol* 48**:** 1088-1095.
Pritchett MA, Metcalf WW (2005). Genetic, physiological and biochemical characterization of multiple methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A. *Mol Microbiol* 56**:** 1183-1194.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al (2012). The Pfam protein families database. *Nucleic Acids Res* 40**:** D290-D301.
Reguera G, Mccarthy KD, Mehta T, Nicoll JS, Tuominen MT, Lovley DR (2005). Extracellular electron transfer via microbial nanowires. *Nature* 435**:** 1098-1101.
Rogosa M (1969). Acidaminococcus gen. n., Acidaminococcus fermentans sp. n., anaerobic gram-negative diplococci using amino acids as the sole energy source for growth. *J Bacteriol* 98**:** 756-766.
Schink B (1997). Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* 61**:** 262-280.
Seemann T (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics***:** btu153.
Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9**:** 811-814.
Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, et al (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* 35**:** D260-D264.
Shimizu S, Akiyama M, Naganuma T, Fujioka M, Nako M, Ishijima Y (2007). Molecular characterization of microbial communities in deep coal seam groundwater of northern Japan. *Geobiology* 5**:** 423-433.
Singh DN, Kumar A, Sarbhai MP, Tripathi AK (2011). Cultivation-independent analysis of *archaeal* and *bacterial* communities of the formation water in an Indian coal bed to enhance biotransformation of coal into methane. *Appl Microbiol Biotechnol* 93**:** 1337-1350.
Stams AJ (1994). Metabolic interactions between anaerobic bacteria in methanogenic environments. *Antonie Van Leeuwenhoek* 66**:** 271-294.
Stams AJ, Plugge CM (2009). Electron transfer in syntrophic communities of anaerobic *bacteria* and *archaea*. *Nature Reviews Microbiology* 7**:** 568-577.
Strachan CR, Singh R, Vaninsberghe D, Ievdokymenko K, Budwill K, Mohn WW, et al (2014). Metagenomic scaffolds enable combinatorial lignin transformation. *Proc Natl Acad Sci U S A* 111**:** 10143-10148.
Strąpoć D, Mastalerz M, Dawson K, Macalady J, Callaghan AV, Wawrik B, et al (2011). Biogeochemistry of microbial coal-bed methane. *Annual Review of Earth and Planetary Sciences* 39**:** 617-656.
Strapoc D, Picardal FW, Turich C, Schaperdoth I, Macalady JL, Lipp JS, et al (2008). Methane-Producing Microbial Community in a Coal Bed of the Illinois Basin. *Appl Environ Microbiol* 74**:** 2424-2432.
Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* 39**:** D546-D551.
Tang Y-Q, Ji P, Lai G-L, Chi C-Q, Liu Z-S, Wu X-L (2012). Diverse microbial community from the coalbeds of the Ordos Basin, China. *International Journal of Coal Geology* 90-91**:** 21-33.
Teeling H, Glöckner FO (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Briefings in bioinformatics* 13**:** 728-742.
Thomas T, Gilbert J, Meyer F (2012). Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2**:** 3.

Větrovský T, Baldrian P (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 8**:** e57923.
Wang Q, Garrity GM, Tiedje JM, Cole JR (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73**:** 5261-5267.
Warnecke F, Hess M (2009). A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J Biotechnol* 142**:** 91-95.
Wei M, Yu Z, Zhang H (2013). Microbial Diversity and Abundance in a Representative Small-Production Coal Mine of Central China. *Energy Fuels* 27**:** 3821-3829.
Wilmes P, Wexler M, Bond PL (2008). Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS One* 3**:** e1778.
Wolin E, Wolin MJ, Wolfe R (1963). Formation of methane by bacterial extracts. *J Biol Chem* 238**:** 2882-2886.
Wood D, Salzberg S (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15**:** R46.
Yilmaz S, Allgaier M, Hugenholtz P (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature methods* 7**:** 943-944.
Zengler K, Richnow HH, Rosselló-Mora R, Michaelis W, Widdel F (1999). Methane formation from long-chain alkanes by anaerobic microorganisms. *Nature* 401**:** 266-269.
Zhang S, Wang Q, Xie S (2012). Stable isotope probing identifies anthracene degraders under methanogenic conditions. *Biodegradation* 23**:** 221-230.
Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7**:** 203-214.

Supplementary material

Supplementary Table 1. Summary of all taxonomic bins generated by GroopM. Taxonomic assignment is indicated at the class level where >50% of the identified marker sequences were assigned to a single class. Where multiple assignments are given >30% of the identified marker sequences were assigned to each class. Mixed indicates no clear taxonomic identity at the class level and no hits indicates where no phylogenetic markers were found in that bin. Taxonomies marked with (*) had a high viral marker gene load.

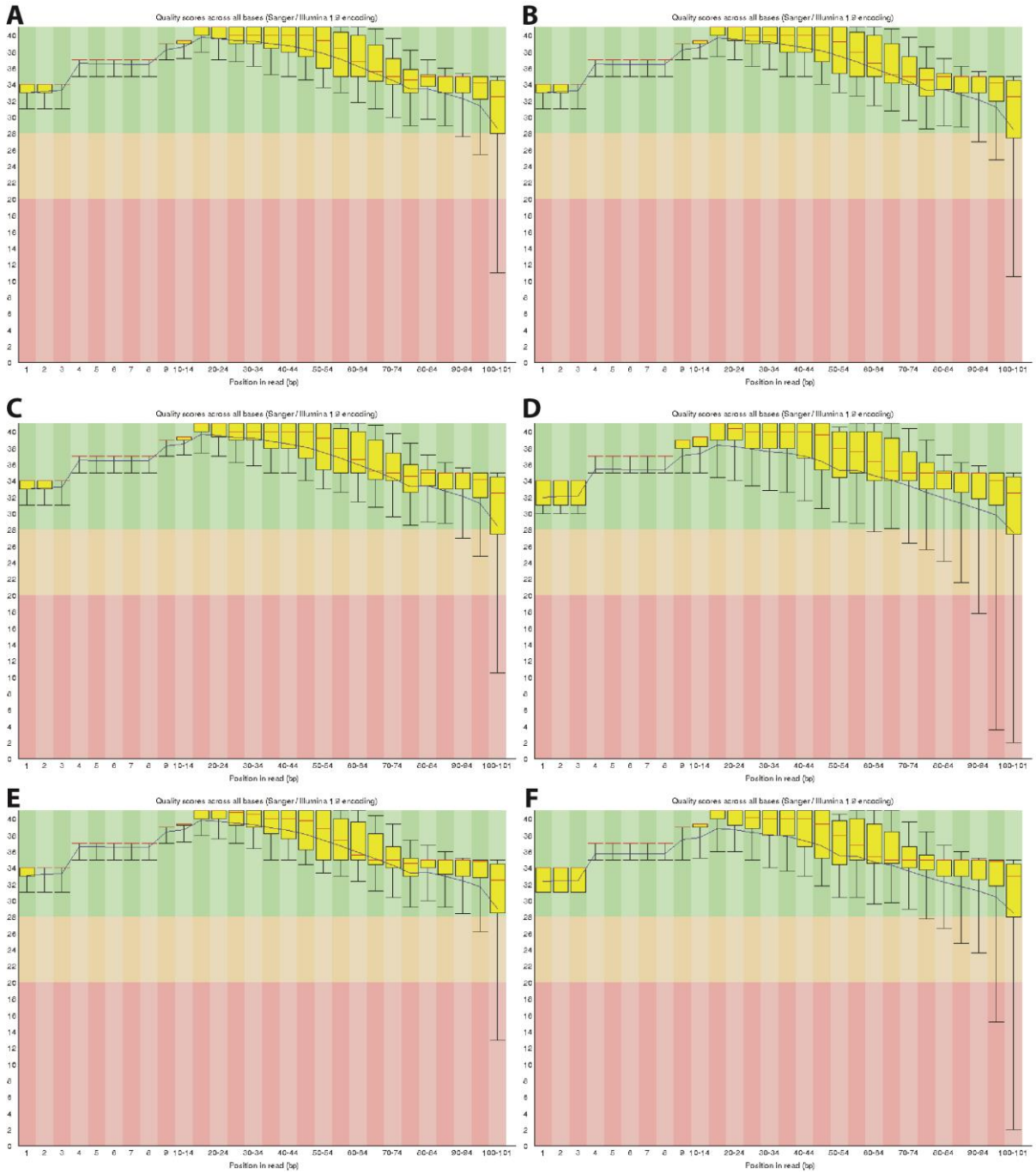
Bin	total length (bp)	# contigs	mean GC	Taxonomic assignment
1	30447357	7912	0.6679	Alphaproteobacteria*
4	5967526	945	0.4446	Sphingobacteria
9	8148365	401	0.3691	Sphingobacteria/Flavobacteria
10	911456	24	0.3993	Sphingobacteria
11	7089968	6151	0.6423	mixed
15	10090770	1550	0.5511	Chloroflexi
17	14385436	1809	0.5704	mixed
18	2956517	817	0.589	Chloroflexi
24	1695150	603	0.4511	Microgenomates
28	1924099	129	0.3991	Methanobacteria
29	915370	355	0.4137	Microgenomates
33	5100300	1478	0.4462	Bacteroidia
34	8513869	1010	0.5989	Synergistia/unknown Synergistetes
35	151690	37	0.5053	no hits
36	56429	2	0.5218	Chloroflexi
37	158993	58	0.442	no hits
38	284383	82	0.4387	Negativicutes
39	2787100	252	0.4475	Chloroflexi
41	2326929	399	0.5317	Methanomicrobia
42	1521548	416	0.6145	Deltaproteobacteria
43	191766	88	0.6999	no hits
44	7000738	1681	0.532	mixed
46	5210425	638	0.4174	Methanomicrobia
47	2766962	91	0.5742	Methanomicrobia
48	2485343	48	0.6188	mixed
49	710964	251	0.6318	Methanomicrobia*
50	3439096	406	0.6445	mixed
53	7222832	600	0.4071	mixed
54	4950398	81	0.6217	mixed
56	5274578	82	0.4609	Clostridia
57	3732817	1285	0.4177	Bacteroidia/Clostridia
60	1140622	333	0.5608	Deltaproteobacteria*
61	4664375	1143	0.559	mixed
63	109639	61	0.394	no hits

67	694622	111	0.3058	mixed
68	942374	424	0.5553	Clostridia
69	123400	15	0.5275	no hits
70	1041717	432	0.3961	Clostridia
71	5980609	1303	0.6381	mixed
72	2661718	107	0.4806	Methanomicrobia
73	740748	404	0.3892	Clostridia/Bacteroidia
78	1298466	553	0.3763	Bacteroidia
79	4382187	407	0.6947	Alphaproteobacteria*
82	3884714	92	0.4591	mixed
84	198645	34	0.5343	no hits
85	3905450	93	0.5457	Clostridia
86	814675	23	0.3258	mixed
87	6235105	798	0.4653	Negativicutes
90	538302	156	0.4727	Methanomicrobia
92	686114	513	0.677	mixed
94	1005941	141	0.4833	Methanomicrobia
96	3916712	363	0.4231	Bacteroidia
98	1512557	31	0.335	Methanobacteria
99	2615386	1480	0.5298	mixed
100	1532308	635	0.3663	Clostridia
101	4219363	469	0.5992	mixed
102	2562143	958	0.6179	Spirochaetes
105	3903719	41	0.4273	mixed
106	288138	165	0.4015	Deltaproteobacteria
108	395696	113	0.6875	Lentisphaerae
111	4830000	3539	0.5796	mixed
112	578312	397	0.6552	Chloroflexi
114	631049	529	0.449	Flavobacteria
115	3521813	1782	0.5164	mixed
116	1776492	1474	0.587	mixed
117	429548	386	0.6071	no hits
118	1338024	887	0.4477	Clostridia
119	308034	300	0.3577	no hits
120	527932	333	0.6842	Lentisphaerae
121	1074133	754	0.528	Spirochaetes
122	1743759	1262	0.4318	mixed
123	532485	396	0.579	Clostridia/Lentisphaerae
124	836556	582	0.4669	mixed
125	1198380	758	0.5158	mixed
126	94195	52	0.5702	no hits
127	1114434	961	0.6055	mixed
128	844343	613	0.5271	Methanomicrobia

129	942621	348	0.3704	Bacteroidia
130	2057992	1415	0.4732	mixed
131	883249	918	0.4661	Clostridia
136	413726	238	0.3802	Bacteroidia
137	1232899	758	0.6365	mixed
139	532237	393	0.4671	Negativicutes
140	2162543	480	0.3247	Mollicutes
141	1386338	846	0.5748	mixed
142	1378718	482	0.3859	Clostrodia
144	3191989	157	0.3404	Clostridia
145	1748735	1014	0.5276	mixed
147	954787	655	0.6214	Clostridia/Bacteroidia
148	308186	279	0.402	Clostridia
149	257391	214	0.621	Clostridia/Bacteroidia
150	640234	322	0.3895	no hits
151	533939	414	0.6057	Clostridia
152	1416496	789	0.5916	no hits
153	684134	564	0.6326	Deltaproteobacteria
154	1990033	2074	0.6032	Chloroflexi
155	1411134	1566	0.4032	mixed
156	1410181	926	0.5757	mixed
157	111045	65	0.3245	Clostridia
158	788774	884	0.5722	no hits
159	2218234	1279	0.4106	mixed
161	2810823	34	0.5596	Spirochaetes*
165	742780	414	0.4433	Bacilli/Methanomicrobia
167	1089696	851	0.4377	mixed
169	577746	306	0.3907	no hits
170	1293319	878	0.357	Bacteroidia
171	517650	473	0.6594	mixed
175	2942913	477	0.4674	Epsilonproteobacteria
176	1107033	16	0.375	mixed
177	1265785	1432	0.5839	mixed
178	1472526	1382	0.394	Clostridia
179	1127990	1096	0.4975	mixed
181	387519	331	0.3746	Clostridia
182	558190	327	0.6084	Alphaproteobacteria
184	152745	69	0.3423	no hits
188	1061424	901	0.385	Mollicutes
191	124968	108	0.3395	no hits

Supplementary Table 2. Coal seam well water chemistry of formation water used to inoculate the MP09 community.

Trace Elements	Iron Filtered (ug/L)	260
	Magnesium Total (mg/L)	5.9
	Phosphorus Totalm (g/L)	0.13
	Potassium Total (mg/L)	25
	Sodium Total (mg/L)	3800
	Sulphur Total (mg/L)	32
Miscellaneous	Chloride (mg/L)	560
	Bicarbonate (mg/L)	6800
	Carbonate (mg/L)	1300
	Hydroxide (mg/L)	<5
	Conductivity (uS/cm)	14000
	Nitrogen - Total (mg/L)	16
	Orthophosphate-P (mg/L)	0.13
	Sulphate (mg/L)	10
	Ferrous Iron (mg/L)	0.24
	Nitrate-N (mg/L)	<0.05
	Nitrite-N (mg/L)	<0.05
	pH	9.2



Supplementary Figure 1. Sequence quality scores for raw HiSeq reads. BB forward (A), BB reverse (B), GL forward (C), GL reverse (D), LYS forward (E), LYS reverse (F).

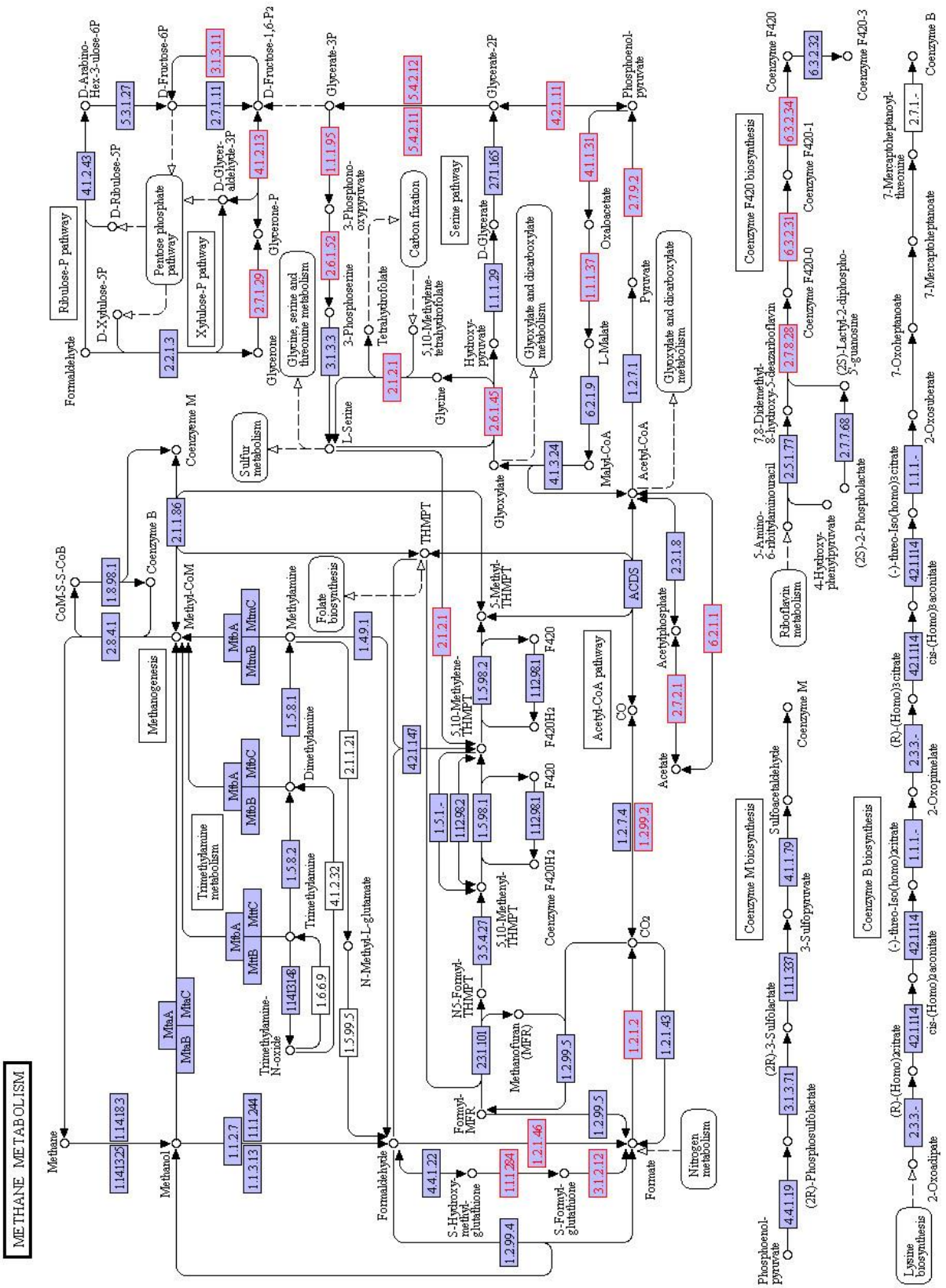
Supplementary Table 3. List of predicted protein product names for CDSs shown in Figure 4.4 listed from 5' to 3' or left to right on the contig sequence representation. Rows highlighted in green indicate proteins on the sequence shown in detail (Figure 4.4) and rows in orange indicate methanogenesis associated proteins.

CDS predicted protein product name	Length (bp)	Direction
dTDP-3-amino-3,6-dideoxy-alpha-D-galactopyranose transaminase	1,053	forward
Pectate lyase superfamily protein	1,539	reverse
Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase	984	forward
dTDP-glucose 4,6-dehydratase	945	forward
UDP-glucose 4-epimerase	996	forward
6)-phosphatidylinositol monomannoside mannosyltransferase	1,182	forward
Glycogen synthase	1,152	forward
putative membrane protein	2,196	forward
D-inositol 3-phosphate	786	reverse
colanic acid biosynthesis protein	1,218	forward
Glycosyl transferases group 1	1,020	forward
N-glycosyltransferase	924	forward
colanic acid exporter	1,512	reverse
putative membrane protein	2,157	forward
sugar transferase, PEP-CTERM/EpsH1 system associated	1,206	reverse
hypothetical protein	381	reverse
Putative glycosyltransferase EpsE	942	reverse
GDP-L-fucose synthase	942	reverse
GDP-mannose 4,6-dehydratase	1,098	reverse
Teichoic acids export ATP-binding protein TagH	1,251	reverse
Teichoic acid translocation permease protein TagG	834	reverse
hypothetical protein	972	reverse
hypothetical protein	957	forward
D-inositol 3-phosphate	1,146	forward
lactoylglutathione lyase	402	forward
hypothetical protein	1,782	reverse
hypothetical protein	447	reverse
FeS cluster assembly protein SufB	1,224	reverse
putative ABC transporter ATP-binding protein	750	reverse
Pectate lyase superfamily protein	1,521	forward
putative membrane protein	1,131	reverse
hypothetical protein	93	forward
Flavodoxin	456	forward
hypothetical protein	567	forward
hypothetical protein	684	forward
hypothetical protein	534	forward
hypothetical protein	327	reverse
MFS transporter, metabolite:H+ symporter (MHS) family protein	1,443	reverse
Coenzyme F420-reducing hydrogenase, delta subunit	276	forward
hypothetical protein	291	forward
hypothetical protein	480	reverse
putative heme utilization carrier protein HutX	516	forward
Cobyrinic acid A,C-diamide synthase	1,350	reverse
Nitrogenase iron protein 1	780	reverse
Nitrogenase molybdenum-iron protein alpha chain	1,074	reverse
UDP-N-acetylmuramoylalanyl-D-glutamate--2, 6-diaminopimelate ligase	1,293	reverse
Sirohydrochlorin ferrochelatase	447	reverse
Ferric-citrate-binding protein	1,041	forward
Hemin transport system permease	1,119	forward
Hemin import ATP-binding protein HmuV	807	forward
Nitrogenase iron protein 1	2,196	forward
Light-independent protochlorophyllide reductase	1,290	forward
hypothetical protein	213	reverse
Siroheme synthase	630	forward
Glutamyl-tRNA reductase	1,308	forward
Delta-aminolevulinic acid	972	forward
Glutamate-1-semialdehyde 2,1-aminomutase	1,278	forward
Porphobilinogen deaminase	891	forward
Uroporphyrinogen-III C-methyltransferase	750	forward
uroporphyrinogen-III synthase	783	forward
putative cobalt-precorrin-6Y C(15)-methyltransferase [decarboxylating]	495	forward
Cobalt-precorrin-4 C(11)-methyltransferase	606	forward
Cobalt-precorrin-4 C(11)-methyltransferase	714	forward
cobalamin biosynthesis protein CbiG	867	forward
Cobalt-precorrin-3B C(17)-methyltransferase	786	forward
precorrin-8X methylmutase	336	forward
cobalt-precorrin-6A synthase	1,005	forward
putative cobalt-precorrin-6Y C(5)-methyltransferase	576	forward
Energy-coupling factor transporter ATP-binding protein EcfA3	828	reverse
Energy-coupling factor transporter transmembrane protein NikQ	735	reverse
hypothetical protein	285	reverse

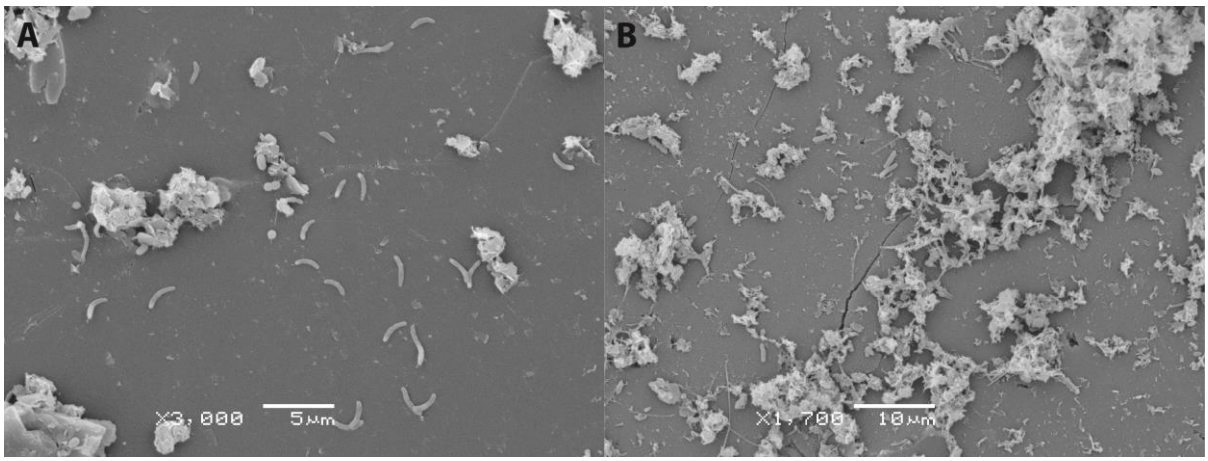
Energy-coupling factor transporter probable substrate-capture protein NikMN	672	reverse
Tryptophan synthase alpha chain	777	reverse
Tryptophan synthase beta chain	1,185	reverse
N-(5'-phosphoribosyl)anthranilate isomerase	624	reverse
synthase	768	reverse
Anthranilate phosphoribosyltransferase 2	984	reverse
Para-aminobenzoate synthase glutamine amidotransferase component II	582	reverse
Anthranilate synthase component 1	1,362	reverse
2-amino-4-deoxychorismate dehydrogenase	576	forward
hypothetical protein	156	forward
Acetyltransferase (GNAT) family protein	507	forward
adenosyltransferase	612	forward
hypothetical protein	330	reverse
zinc transporter ZupT	762	forward
Putative zinc metalloproteasec/MT2700	1,137	forward
hypothetical protein	1,050	forward
hypothetical protein	330	reverse
Polysaccharide biosynthesis protein	1,455	reverse
hypothetical protein	216	reverse
dTDP-glucose 4,6-dehydratase	936	reverse
UDP-glucose 4-epimerase	888	forward
hypothetical protein	567	reverse
carboxylate/amino acid/amine transporter	936	reverse
hypothetical protein	528	forward
hypothetical protein	615	forward
UTP--glucose-1-phosphate uridylyltransferase	867	reverse
UDP-glucose 6-dehydrogenase TuaD	1,257	reverse
Polysaccharide biosynthesis protein	1,305	forward
Sensor protein kinase WalK	1,044	reverse
1-pyrroline-5-carboxylate dehydrogenase 1	1,572	forward
PRC-barrel domain protein	654	forward
hypothetical protein	234	forward
Fluoroquinolones export ATP-binding proteinc/MT2762	924	forward
ABC-type transport system involved in multi-copper enzyme maturation, permease component	1,026	forward
methyltransferase	1,158	reverse

Putative aliphatic sulfonates-binding protein precursor	1,002	reverse
Aliphatic sulfonates import ATP-binding protein SsuB	846	reverse
Bicarbonate transport system permease protein CmpB	888	reverse
Inner membrane transport	1,089	reverse
Daunorubicin/doxorubicin resistance ATP-binding protein DrrA	945	reverse
Molybdopterin-synthase adenylyltransferase	735	forward
Transcriptional regulator YqjI	372	reverse
hypothetical protein	726	reverse
hypothetical protein	504	reverse
hypothetical protein	834	reverse
Small-conductance mechanosensitive channel	885	forward
putative RNA-binding protein	495	forward
hypothetical protein	384	forward
deoxyguanosinetriphosphate triphosphohydrolase-like protein	1,314	forward
RNA-splicing ligase RtcB	1,425	reverse
hypothetical protein	1,209	forward
tRNA-His	76	reverse
hypothetical protein	255	forward
Ornithine carbamoyltransferase	891	forward
ssDNA exonuclease RecJ	1,608	reverse
putative exonuclease	798	reverse
putative phosphoesterase	726	forward
Methionine synthase	642	forward
Putative DNA ligase-like protein/MT0965	870	forward
Uroporphyrinogen decarboxylase	999	forward
methylamine methyltransferase corrinoid protein reductive activase	1,641	forward
hypothetical protein	837	reverse
putative methanogenesis marker protein 1	1,200	forward
hypothetical protein	642	forward
methyl-coenzyme M reductase, beta subunit	1,320	forward
methyl-coenzyme M reductase operon protein D	420	forward
methyl-coenzyme M reductase, gamma subunit	774	forward
methyl-coenzyme M reductase, alpha subunit	1,665	forward
hypothetical protein	291	forward
hypothetical protein	561	reverse
alkylhydroperoxidase/carboxymuconolactone decarboxylase family	327	forward

hypothetical protein	330	reverse
hypothetical protein	852	reverse
hypothetical protein	978	reverse
hypothetical protein	495	reverse
Methionine import ATP-binding protein MetN	1,623	forward
methyl-coenzyme M reductase I	603	forward
hypothetical protein	570	reverse
putative methanogenesis marker protein 3	2,040	forward
putative methanogenesis marker protein 6	402	forward
putative methanogenesis marker protein 5	435	forward
2-hydroxyglutaryl-CoA dehydratase	1,233	forward
putative methanogenesis marker protein 17	558	forward
putative methanogenesis marker protein 7	951	forward
hypothetical protein	438	forward
putative FAD-linked oxidoreductase	1,605	forward
HTH-type transcriptional regulator LrpC	597	forward
Enamine/imine deaminase	381	reverse
Phosphoglycolate phosphatase, chromosomal	702	reverse
hypothetical protein	333	forward
transcriptional regulator MalT	2,691	reverse
hypothetical protein	543	forward
hypothetical protein	546	reverse
Tryptophan synthase beta chain	1,311	reverse
Ribonucleoside-diphosphate reductase NrdZ	1,779	reverse



Supplementary Figure 2. Pathway map of methanogenesis generated by KEGG. KAAS annotations mapping to bin 1 associated with methanogenesis highlighted in red.



Supplementary Figure 3. Scanning electron microscopy (SEM) images of polished coal disks after inoculation with the MP09 consortium and incubation for 8 days (A) and 27 days (B) in the replicate culturing experiment.