universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Taxonomic analysis and metagenomic binning of a community present in an agricultural biogas fermenter"

verfasst von / submitted by

Britta Nöbauer, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Danksagung

An erster Stelle möchte ich mich recht herzlich bei meinem Betreuer Thomas Rattei bedanken, der mir die Möglichkeit gegeben hat an diesem interessanten Projekt zu forschen, für seine Unterstützung aber auch die hilfreiche Kritik und Diskussion und nicht zuletzt für die Grillabende und Ausflüge mit unserem gesamten Department.

Ein großes Dankeschön gilt auch Dmitrij, der mit mir an dieser Studie gearbeitet hat, mir bei Problemen immer unterstützend zur Seite gestanden ist und dafür, dass er mich vom ersten Moment an wirklich nett im CUBE empfangen hat.

Dank gebührt auch unseren Kooperationspartnern der Universität Hamburg, Wolfgang Streit und Simon Güllert, für die Bereitstellung der Daten und für ihre Anregungen die Ergebnisse zu interpretieren.

Natürlich bedanke ich mich aber auch bei allen anderen Kollegen im CUBE und bei Gabi, für die freundschaftliche Atmosphäre und ihre Hilfe falls ich an manchen Punkten nicht mehr weitergekommen bin.

Ein Dankeschön aus tiefstem Herzen gebührt meinen Eltern, meinem Bruder und meiner Familie die mir mein Studium überhaupt erst ermöglicht haben und die mich in jeder Situation unterstützen, ein offenes Ohr für mich haben und mir immer eine Anlaufstelle bieten. Und nicht zu vergessen Ephraim, Maria, Harry, Vera, Daniela und alle anderen die mich auf meinem bisherigen Studienweg begleitet haben, möchte ich für die vielen lustigen Abende und die schönen Momente danken, die wir hier in Wien zusammen erlebt haben.

Abstract

Biogas fermenters harbour very complex microbial communities and the individual members have to fulfil distinct tasks on the way from substrate degradation to the final products methane and CO2. Only a minor proportion of the present species has been characterized by now and their genome sequences and functional assignments are often not available. Even if the representatives seem to be perfectly organised in the hydrolytic breakdown of plant material, they compare unfavourably to populations present in the digestive tracts of herbivores. Explanations for this observation are obscure so far.

In this thesis we analysed a one-stage agricultural biogas fermenter with respect to microbial community structure and functional genomic equipment of the underlying taxa. Metagenomic analyses resulted in 1.16 Gb of assembled DNA and binning of 104 high quality genome reconstructions. Our results display that *Firmicutes* are far more prominent in biogas plants whereas in natural systems, *Bacteroidetes* seem to be equally abundant. This observation was reflected by the lower prevalence of glycoside hydrolase family genes in the metagenomic bins assigned to members of *Bacteroidetes*. A deficiency in genes encoding presumable GH enzymes may be associated with the limited potential of biogas fermenters regarding hydrolysis. These findings tempt to speculate that increasing the proportion of *Bacteroidetes* in agricultural biogas plants, will presumably lead to increased hydrolysis of plant biomass.

Kurzzusammenfassung

Biogasfermenter beherbergen äußerst komplexe mikrobielle Gemeinschaften und die einzelnen Mitglieder erfüllen ganz spezielle Aufgaben auf dem Weg vom Substrat-Abbau zur Herstellung der Endprodukte Methan und CO2. Nur ein winziger Anteil der tatsächlich vorkommenden Spezies wurden bis jetzt näher beschrieben; oftmals sind die Genomsequenzen und funktionellen Zuordnungen nicht vorhanden. Auch wenn es so scheint als wären die einzelnen Vertreter perfekt organisiert im hydrolytischen Abbau von pflanzlicher Biomasse, so schneiden sie im direkten Vergleich zu Populationen in Verdauungstrakten von Pflanzenfressern doch wesentlich schlechter ab. Erklärungen für diese Feststellung liegen bis jetzt noch im Dunkeln. In dieser Thesis haben wir einen typischen landwirtschaftlichen Ein-Phasen Biogasfermenter in Bezug auf dessen mikrobielle Zusammensetzung und die funktionelle Genomausstattung der zugrundeliegenden Taxa untersucht. Metagenomische Analysen haben die Assemblierung von 1,16 Gb an DNA und die Rekonstruktion von 104 qualitativ hochwertigen Genomrekonstruktionen ermöglicht. Unsere Resultate zeigen, dass *Firmicutes* in Biogasanlagen weitaus häufiger vorkommen wohingegen in natürlichen Systemen, *Bacteroidetes* ziemlich gleich häufig zu sein scheinen. Diese Beobachtung spiegelt sich auch im selteneren Vorhandensein an Glycosid-Hydrolasen Genen in metagenomischen Bins wider, welche den *Bacteroidetes* zugeordnet sind. Ein Defizit an Genen, für GH Enzyme codierend, könnte mit dem limitierten Hydrolysepotential von Biogasfermentern zusammenhängen. Diese Entdeckungen animieren zu Spekulationen, dass eine Erhöhung an Bacteroidetes zu einer erhöhten Hydrolyserate von Pflanzenmaterial führen könnte.

Table of Contents

Abbreviations

AD	anaerobic digestion/degradation
BLASTn	nucleotide BLAST
bp	base-pairs
CAZy	carbohydrate-active enzymes
CDS	coding sequences
CE	carbohydrate esterase
COG	cluster of orthologous groups
DBG	de Bruijn graph
Gbp	Giga base-pairs
GH	glycoside hydrolase
GT	glycosyl transferase
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LCA	lowest common ancestor
Mbp	Mega base-pairs
nc-RNA	non-coding RNA
NGS	next-generation sequencing
OTU	operational taxonomic unit
RDP	ribosomal database project
rRNA	ribosomal RNA
SSU	small subunit
SVM	support vector machine
tRNA	transfer RNA

1. Introduction

1.1 Biological background and aim of the study

Our modern society's global energy demand is constantly increasing and the major part of it is covered by the use of fossil fuels. Addressing the problems of climatic changes and greenhouse gas emissions, many European countries, in particular Germany, Austria, Denmark and Sweden, have augmented biogas production, because it is an environmentally friendly renewable source of energy [1]. Biogas, which is mainly composed of methane and carbon dioxide, is produced in a process called *anaerobic digestion* (AD) which is carried out by complex microbial communities [2]-[5]. Applying this technology has several advantages as it couples waste disposal to the production of a highly valuable renewable fuel and additionally, nutrient recovery can replace mineral fertilizers [1]. The originating biomethane can completely substitute fossil fuels as it can be used in the generation of heat and electricity, as well as a vehicle fuel. The final stage of methane production is well understood, contrary to the microbial communities involved in the other stages of biogas production, their role and the underlying dynamics are not well characterized so far [2]-[5]. It is known that the way from anaerobic substrate degradation to the final products, methane and carbon dioxide, requires a close interaction of several hundreds to several thousand phylogenetically different microbial species [3]; [4]; [6]; [7]. The first step, being the hydrolysis of polysaccharides, is regarded as one of the rate-limiting steps in the entire process and therefore it is one determinant of the overall efficiency [8]. Cellulose is a rich source of organic carbon compounds however, the actual process from cellulose

degradation to biogas production leaves room for further investigation. Bacteria that are able to degrade cellulose are rare and the majority of them are belonging to the bacterial phyla *Firmicutes* or *Actinobacteria* [8]. *Clostridia* are reported to be the dominant class of hydrolytic Bacteria in biogas fermenters and therefore one can assume that they play a major role in the initial step [8]. In natural digestive tracts of studied herbivore organisms, the *Clostridia* appear to be less dominant and are outcompeted by the *Bacteroidetes*. Different studies indicate that *Bacteroidetes* can be found in all samples of digestive organs as well as feces of herbivores. In those natural biogas producing systems, they usually represent the main bacterial group and seem to be more dominant than *Firmicutes* [9]-[12]. This is one of the huge differences in the bacterial composition of biogas fermenters and natural cellulolytic systems.

The main aim of this thesis is to investigate the taxonomic composition of the underlying biogas fermenter sample, as well as analysing the possible effects that the differences, compared to microbial communities in natural digestive organs, may have in the effective degradation of plant biomass. The use of deep DNA-sequencing, RNA-sequencing and metagenomic analysis should enable gaining further knowledge about the relative taxonomic community composition and the occurrence of *glycoside hydrolase* (GH) enzymes. The comparison of the taxonomic composition, as well as the abundance of GH family genes, in artificial anaerobic digestion systems to natural cellulolytic systems may provide deeper insight for enhancing process efficiency and final biogas yield.

1.2 Biogas – production, usage and advantages

The term *"biogas"* generally refers to gas produced by anaerobic digestion units [13], is a major player in the category of renewable energies and a promising candidate addressing the global need for energy as well as having multiple environmental benefits [14]-[16]. Today, bioenergy production in general is estimated to be the fourth largest source of energy in the world [17]. *Table 1* gives an overview of biogas usage, its benefits and shows a comparison to other sources of energy.

Table 1: Overview of possible biogas usage, potential substrates and overall advantages comparing conventional energy sources. According to Mao et al., 2015, modified.

Advantages of biogas usage		References
Green energy production	Electricity Heat Vehicle fuel	[16]
Organic waste disposal	Agricultural residues Industrial wastes Municipal solid wastes Household wastes Organic waste mixtures	[1]
Environmental protection	Pathogen reduction through sanitation Less nuisance from insect flies Air & water pollution reduction; eutrophication and acidification reduction Forest vegetation conservation Replacing inorganic fertilizer	[14]; [15]
GHG emission reduction	Substituting conventional energy resources	[14]; [18]

Biogas is a very versatile energy carrier, depending on the specific requirements of the final processing techniques. It can be used in the generation of electricity, heat and as vehicle fuel, after post-processing e.g. desulfurization and water removal [1]. Among various factors, the final gas yield is strongly dependent on the injected

substrates. However, the range of possible fed in substrates is broad and frequently used source materials are animal manures, industrial wastes, commercial or municipal wastes and different grains or grasses [1]; [18].

Increasing the final yield in biogas does not solely depend on optimization of process parameters (e.g. temperature, pH, retention time); it is even more important that the different substrate conversion steps proceed in precise coordination. It is general knowledge that the slowest step of a reaction determines the overall reaction speed and performance but for being able to optimize these single parameters, all elements of the reaction chain, as well as their output, interactions and demands have to be known. Even if the microorganisms carrying out the conversion of plant material and the mechanisms, that are involved on the way to methane production, are well understood, the overall process and the microbial biogas producing community structure needs further investigation [2]-[5].

1.3 Anaerobic degradation process stages

Breaking it down, anaerobic degradation is a cooperation of a very complex microbial community but mainly three types of Bacteria work together in the most relevant stages, being *hydrolysis, acidogenesis, acetogenesis*, and *methanogenesis* [8]. The first step in the digestion process is the hydrolysis of complex molecules including carbohydrates, lipids, and proteins which are depolymerized by the help of a wide range of enzymes that are produced and secreted by hydrolytic Bacteria together with saccharolytic Bacteria. These Bacteria are either obligate anaerobic as *Bacteroides* and *Clostridia*, or they are facultative anaerobic, as for example *Streptococci* [19].

The products of this initial step and the following acidogenesis are organic acids, alcohols, CO_2 and H_2. The acetogenesis, performed by acetogenic Bacteria (syntrophic Bacteria) involves conversion of the products from the preceding steps into acetate. Acetotrophic and hydrogenotrophic Archaea then convert the acetate, CO_2 and H_2 to methane and CO_2. This last step, and possibly also the best understood, is the methanogenesis [7]; [8]. Representatives, which perform this reaction are for instance *Methanosarcina*, *Methanothrix*, *Methanobacterium* and *Methanococcus* [19]. *Figure 1* gives a brief overview of the most important chemical reactions as well as most abundant taxonomic groups. As indicated there, bacterial representatives are the dominating ones carrying out the first three steps of the process, the final conversion is mainly performed by hydrogenotrophic or acetotrophic Archaea, producing CH_4 and CO_2.

The initial step in the AD process, the hydrolysis, is known to be crucial for the overall production efficiency. Hydrolysis of the biomass is regarded as the rate limiting step as all downstream reactions depend on the initial hydrolysis production rate – the more substrate is used, the greater is the final methane yield [8]. This fact is one of the main reasons why a better understanding of the community involved during hydrolysis is crucial for improving overall process efficiency.

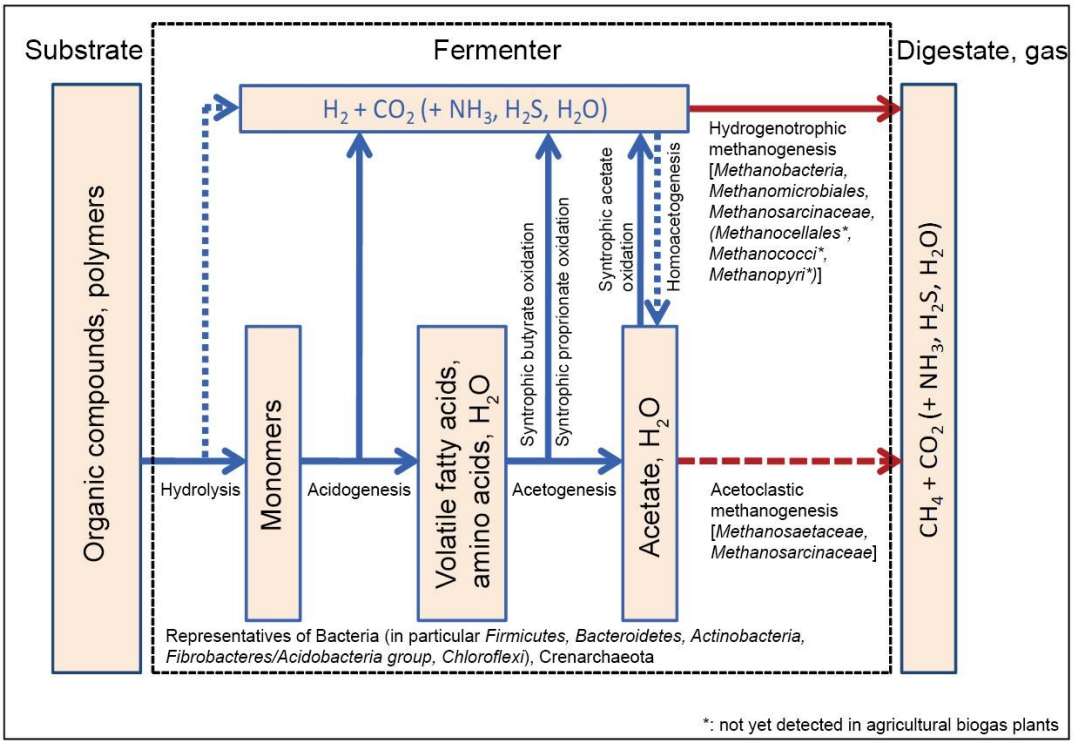

Figure 1: Major chemical reactions in the anaerobic degradation of organic compounds, substrates and products as well as most important microbial community members involved in the single steps. According to Lebuhn and Gronauer, 2009, modified.

1.4 Microbial community analysis – natural vs. artificial systems

The production of biogas by anaerobic degradation is not a novel technical invention. Utilising organic material as substrate for biogas production is widely found in nature, for instance, the intestine of plant-feeding animals or insects, swamps or marshes, composts or debris of lake grounds are all habitats for the microorganisms cited above [8]. *Bacteroidetes* are regarded to be the most competitive microbial group, together with the *Firmicutes*, as multiple studies indicate that they are present in nearly all herbivore rumen, gut and fecal samples. This is in contrast to the technical replica like biogas fermenters; here the *Clostridia* appear to be the predominant bacterial class and literature indicates that they play a major role during hydrolysis. However, cow rumen, gut or other organs involved in natural digestive processes of herbivorous animals are regarded as not being dominated by clostridial organisms [8]-[12]; [20]-[23].

These publications suggest that there is a vital contrast between natural and artificially generated microbial communities needed in the anaerobic digestion process and gaining detailed insight into possible reasons for this observation, might improve final product yield.

1.5 Glycoside hydrolase families and CAZy database

Carbohydrate-active enzymes (CAZy) are mandatory for the breakdown of complex carbohydrates. The CAZy enzyme database project (http://www.cazy.org/) [24] has collected CAZy enzyme families which are involved in the synthesis, modification and breakdown of oligo- and polysaccharides. CAZy genes are regarded to make up around 1-5% of the coding regions of an organism. As these CAZy enzymes are vital in the processing of cellulose, they are regarded as an essential key to success in the production of biogas [24]. Depolymerisation of plant-derived cellulose, which is recalcitrant to hydrolysis and is often found in crystalline form, needs key enzymes mainly belonging to different glycoside hydrolase families. Glycoside hydrolases break up the glycosidic bonds between carbohydrates, are capable of hydrolysing even crystalline cellulose and evolved in many different microorganisms, mainly filamentous Fungi and Bacteria [25]. Generally there are several types of GHs and their coefficient action is known to be necessary for the complete hydrolysis of cellulose to glucose [26]-[28]. The first type are *endoglucanases*, hydrolysing the internal bonds in cellulose chains randomly and therefore releasing products of variable length. *Exoglucanases* constitute the second type, releasing cellobiose through action either on the reducing or the non-reducing end of the cellulose polymer. The last type are the *β-glucosidases*, they affect the degradation of cellubiose, finally yielding glucose [28]; [29]. These cellulolytic enzymes can occur as free and independent enzymes, or they are packed together in a formation called *cellulosomes*. Whereas numerous cellulolytic organisms belonging to the class of *Clostridia* produce cellulosomes, Bacteria belonging to the phylum of *Bacteroidetes* are considered to lack these multi-enzyme complexes. However, literature indicates that for an efficient degradation of cellulose, they use so-called *polysaccharide*

utilisation loci (PULs) [30]; these sets of genes might be of importance in the process of degrading cellulose and were recently proposed to be an alternative approach for cellulose breakdown [31].

Differences in the abundance of potential GH genes between natural digestive systems as well as anaerobic degradation systems, established for biogas production, can undergird the findings of a vital underrepresentation of typical rumen or gut bacteria as well as provide more details about the mechanisms involved in the initial biomass hydrolysis.

1.6 Methodical background

Synergistic reactions of microorganisms are crucial for bioenergy production, in both natural or artificially created cellulose degrading systems. Whole genome shotgun sequencing, which is based on the invention of various *next-generation sequencing* (NGS) technologies, has dramatically improved our understanding of community structures and dynamics in the most diverse environments. In metagenomic community analysis, culturing of microorganisms for successful investigation is not necessary anymore. Even if the cultivation-based approaches have helped to gain important elementary knowledge about many key microorganisms, for which cultivation was possible, for the majority of them, found in more complex environments, cultivation has not been possible yet because their essential requirements are unknown. Aside from that, each of the organisms studied individually and isolated might exhibit different characteristics than when the whole complex microbial network is examined [32]-[34]. Due to the application of culture-independent techniques, a prior hardly conceivable diversity was observed, phylogenetically as well as metabolically [3]; [4]; [6]; [7]; [35]; [36]. Additionally, understanding community composition, interactions and reactions has helped to improve reactor set-ups and therefore positively influenced efficiency and stability [13]; [37]. Metagenomic *shotgun sequencing* refers to the random sequencing of all DNA material and accessing the genetic content of entire communities in a certain environment. It provides information about the gene composition of the underlying communities and therefore gives a more exact description than phylogenetic analysis solely. The application of various bioinformatic software tools can reveal potentially novel enzymes, information about genomic linkages between function and phylogeny and also create evolutionary profiles [38].

1.7 Typical metagenomic workflow

The ultimate goal in metagenomics is the reconstruction of all genomes found in a specific environment but due to several problems, e.g. lack in sequencing coverage or difficulties in assembly, this is hardly ever possible. Still, there are two different approaches that are applied instead and therefore getting at least an approximation to entire reconstructions; the first one is a read-based analysis of the taxonomic as well as functional components of the metagenomes, the other is the assembly of reads into longer, continuous stretches of genomic sequences, referred to as *contigs*, prior to taxonomic classifications and functional assignments [39]. Each of these strategies has several inherent limitations; especially the assembly of single sequencing reads into *contigs* can cause inconvenience. Algorithms that were created specifically for short-read assembly are computationally demanding in terms of memory costs, due to the high numbers of genomes found in these communities. Though, not only the number of genomes is challenging, also the wide range of abundance for each single genome in a sample is very complicated and sometimes it is not possible at all to assembly genomes with low abundance. Chimeric assemblies may result from the assembly of very closely related lineages similar to heterogeneity within certain lineages, which can lead to fragmented sequences. Even the analysis on single reads has several bottlenecks, because the multitudinous number of reads that have to be analysed cause long runtimes and the short reads that are produced in NGS experiments can lead to high error rates. *Figure 2* gives an overview about the single steps and processes that are generally executed in metagenomic analyses.

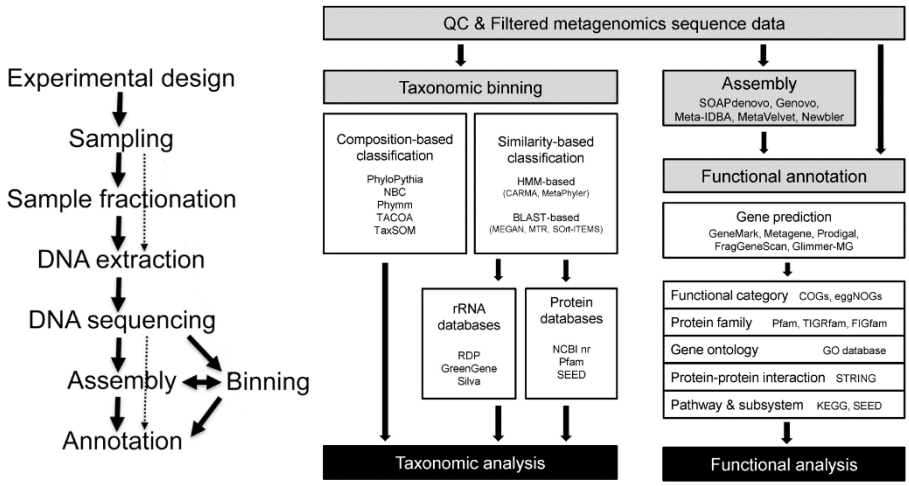

Figure 2: Overview of the procedures in typical metagenomic experiments. According to Thomas et al. 2012 and Kim et al. 2013, modified.

It is not invariably the case that each of these single steps is absolutely necessary and the leading question of the actual project will determine the exact procedure and the required tasks. More precise description of the single processes and software used in this thesis, including possible alternatives, is given below.

1.8 Sequencing and data preprocessing

After sample taking and DNA-extraction, the genomic material has to be sequenced. In the last decade, metagenomic sequencing has gradually evolved from classical Sanger sequencing to next-generation sequencing (NGS). Sanger sequencing has for long been the method of choice due to lacking alternatives but the very time consuming cloning processes and the resulting high cost per gigabase has led to the application of NGS [38]-[41]. Clearly, there is no “holy grail” among them and each of them has pros and cons that have to be balanced to choose the suitable one. Sequencing always results in the acquisition of short nucleotide sequences, referred to as *reads*, which represent the amplified copies of the same genomic fragment that

has been randomly sheared into small pieces beforehand. This process is called *shotgun sequencing* [41]. In the last years, metagenomic analyses have mainly applied *454/Roche* pyrosequencing or *Illumina sequencing-by-synthesis* approach. Advantages of the *454/Roche* sequencing is that it produces reads about 600-800 bp in length, and therefore the greatest length of all second generation NGS technologies and substitution errors are very unlikely. Due to the inherent features of this technology, homopolymer stretches are prone to insertion/deletion errors, the yield of a single run is only about 500 Mbp and it is very costly compared to *Illumina* sequencing. Therefore, *Illumina* sequencing technologies are the ones most frequently used. The major disadvantage is the read length, only reaching up to 150 bp for now. The advantages are not only the lower costs, but also the higher accuracy compared to *454/Roche* and the higher yield of about 60 Gbp in a single run [38]; [41]-[43]. In the present study an *Illumina HiSeq 2500* instrument was used in paired-end mode with read lengths of 2 x 101 bp. The output of these experiments consist of a text file, containing millions of such reads in *FASTA* or *FASTQ* format which are then analysed. One important step in the avoidance of biased data is the preprocessing. Tools have been implemented which are measuring the probabilities of wrong base calls and based on that, provide quality scores for the overall sequencing procedure [41]; [44]-[46]. They are used for filtering, trimming and reformatting as well. *PRINSEQ-lite* [47] was the method of choice in this project, other possibilities would have been for example *FastQC* [48] and *Trimmomatic* [49].

1.9 Assembly and mapping

The next step in a typical metagenomic workflow is the assembly of reads into either *contigs* (longer, contiguous sequence) or *scaffolds* (multiple contigs and gaps together representing a longer stretch of the genomic sequence) for obtaining larger coherent genomic sequences. *Figure 3* is an illustration of these two basic constructs originating from the assembly process.

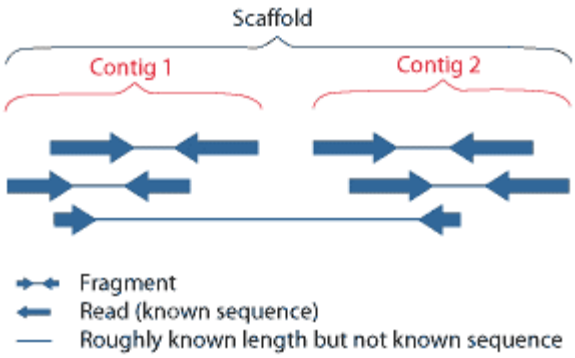

Figure 3: Basic illustration for the two possible products of the assembly of short reads. From http://genome.jgi.doe.gov.

As already discussed, this process is the bottleneck in terms of computational memory load of metagenomic analyses so far. A lot of algorithms have been developed in the last years that are addressing this problem. They can be basically divided into two categories, the *de novo* assemblers and the mapping assemblers. A template comprised of known reference genomes are needed for the assembly by mapping approach [40]; [50]. The results are rather reasonable as sequence repeats, short read lengths and low coverages are not that much of an issue compared to *de novo* approaches. Contrary, the *de novo* assembly method is one of the computationally

most expensive tasks in metagenomic analyses [41]. The assemblers fulfilling this task are divided into three classes, all relying on graph reduction algorithms [51]. The first one is the *overlay-layout-consensus* (OLC) method, which basically searches for path overlaps in three steps and is used for very long reads [52]; [53]. Examples are the *Celera* assembler [54], *Arachne* [55] or *Newbler* [56]. Another possibility are the *de Bruijn graph* (DBG) assemblers which use short *k-mer* subgraphs for memory reduction [57]. Examples would be *Velvet* [58], *IDBA* [59] or *Ray* [52]. The last class are the *greedy algorithms* which are the most intuitive form. They search for the best overlaps and then continue growing contigs iteratively. For example, *SSAKE* [60], *VCAKE* [61] and *SHARCGS* [62] are using this approach. Some assemblers have been designed especially for metagenomic experiments as those reads are more complex, due to the number of different species, strain heterogeneity and the uneven coverage across the genome or between different genomes. These algorithms have to adapt their reconstruction method, based on graphs, to handle variabilities in genome copy numbers and sequences that are conserved across several genomes [63]. As the genomic diversity in metagenomic analyses is known to be immense and reference genomes are lacking, the de-novo assembly is the method of choice, at least until knowledge about the underlying community members exists. In the assembly step of this project, two of those metagenomic short-read assemblers were applied, *Ray Meta* [64] and *IDBA-UD* [65].

Determination of the contig coverages is the next step in many metagenomic analyses, as this knowledge is required for many tasks in downstream analysis. For this purpose, short-read alignment algorithms exist for mapping the reads against the previously assembled contigs. These aligners are based on different algorithms for indexing the reads as well as the references.

The best known strategies are the Burrows-Wheeler transformation or the Smith-Waterman algorithm, but also short *k-mers* are used in the indexing process. Famous short-read aligners are for example *BBMap* [66], *Bowtie2* [67] and *BWA* [68], all of them were used at different stages in this study.

1.10 Taxonomic community analysis

"Who are they?" - this is one of the two fundamental questions in metagenomics and asks for a taxonomic analysis of all community members. There are basically two different approaches how the taxonomic composition can be analysed in NGS experiments. The first approach does not need any assembly or alignment before as it directly analyses the sequencing reads after trimming and filtering. The reads are used in similarity searches against databases that contain reference sequences of interest. Depending on the database used, protein or nucleotide searches are carried out. *BLAST* [69] is the most famous algorithm for detecting sequence similarities to known reference sequences and several variants exists that can browse various databases. However, BLAST is known to be limited in terms of computational speed, so many other tools were specifically designed to speed up the process as the amount of data that has to be handled in metagenomic projects is huge and runtimes have to be considered clearly beforehand. *Rapsearch2* [70] was used here for taxonomic classification of single-reads, as it achieves a fast protein similarity search with only minimal loss in accuracy compared to BLAST. Afterwards search results have then to be analysed for their taxonomic composition. *MEGAN5* [71] is a software tool capable of taxonomically placing reads based on their homology to a given taxon. A characteristic feature of this tool is that it places reads on the *lowest common ancestor* (LCA) of all organisms that contain the gene present in the read as well,

therefore it is a more conservative approach and minimizes the chance for false-positive assignments.

The other possible method for taxonomic community profiling is the analysis of assembled contigs; assigning contigs to different taxa can be done by searching for conserved marker gene sequences that have to be universally distributed across Bacteria or Archaea and are only present in a single copy in all genomes. *AMPHORA* [72] is a program that works by identifying 31 distinct bacterial marker genes from the input sequences for phylotyping (i.e. assigning sequences to taxa). *AMPHORA2* [73] has been used in the present study for phylogenetic marker analysis, as it is also capable of identifying marker genes in archaeal sequences. The underlying idea is the assumption that, if the marker is part of a larger assembled contig, then this contig can be classified into a specific taxonomic level.

1.11 Taxonomic analysis based on rRNA gene search

As each method has certain drawbacks and limitations, it is advisable to verify the results by the application of other approaches. In the present study, taxonomic community analysis results were matched with an analysis of sequencing reads that contain *ribosomal RNA* (rRNA) gene fragments. 16S rRNA gene profiling has been the method of choice for phylogenetic investigation and diversity analysis for the last 20 years now and marks one of the first steps in many metagenomic projects [74]. If 16S rRNA genes are used as phylogenetic marker genes in metagenomic shotgun experiments, then reads containing putative rRNA genes have to be sorted out and classified by similarity searches against specialised databases. Several public available databases exist that contain well-curated rRNA gene sequences, such as *SILVA* [75], *RDP* (ribosomal database project) [76] and *Greengenes* [77].

There are several tools for filtering rRNA gene containing reads out of NGS data such as *SortMeRNA* [78]. The underlying algorithm works with a seeding strategy which is based on the search for many short similarity regions between the read and a respective rRNA sequence database. Another option is a nucleotide BLAST (BLASTn) similarity search against one of the databases mentioned above. Advantages of specialized tools as SortMeRNA are a great speed-up of the process as well as increased sensitivity and selectivity compared to BLASTn analyses [78]. In the present study, both approaches have been applied to the underlying dataset for maximising the number of detected and classified 16S rRNA reads.

1.12 OTU clustering in metagenomic experiments

Operational taxonomic unit (OTU) is a term in microbial analysis, referring to the clustering of sequences with a varying amount of sequence-identity that they have to share at least [79]. It is used in metagenomic analyses for the sorting of microbial sequences according to their sequence similarity. Many metagenomic studies are using clustering approaches for sequence variants of the small subunit (SSU) rRNA marker gene, as it is highly conserved among bacterial and archaeal species. Clustering these variants according to a chosen percentage of similarity threshold can be indicative of the underlying population richness. Nevertheless, the 16S rRNA approach is limited in the resolution at species-level, as differences are often not sufficient for distinguishing at this taxonomic rank. Another problem is that genes might be similar on the nucleotide level, even when they belong to evolutionary distant species [72]; [80]. In general, deriving phylogenetic classification based on a single gene is always risky and has to be corroborated by the use of other markers. This is why microbial research has shifted the focus more towards the use of protein-

coding genes for phylogenetic analysis [72]. As protein-coding sequences are conserved at the amino acid level, the results are less biased by nucleotide composition [72]; [81]. As indicated before, AMPHORA2 uses universally conserved, single-copy marker genes for bacterial and archaeal taxonomic classification. In this study, these markers are identified by AMPHORA2 analysis and the results can then be clustered via *CD-HIT* [82], according to different levels of minimal sequence similarity. CD-HIT in general is used for a so-called clustering analysis, a method for searching for specific sequences and grouping them according to their similarity [83]. There are many other possible clustering programs available that are used for the grouping of protein sequences, for example *ProtoMap* [84], *ProClust* [85], *Blastclust* [86] and *UniqueProt* [87]. The drawback of these methods is that the underlying algorithm performs all-against-all comparisons and therefore CD-HIT was used, an algorithm that circumvents this problem, leading to a great acceleration [83]. CD-HIT uses a greedy algorithm as the first step is an ordering of sequences by decreasing length and the longest serves as seed for the first cluster. All sequences that are remaining are then compared to the existing seeds in a cluster. If the similarity threshold to a certain seed is met, then the sequence is grouped into the respective cluster [83]. Calculating the average number of markers present at a certain similarity threshold gives an indication of the number of OTUs in the underlying community.

1.13 Binning of metagenomic contigs

Binning refers to the process where single genomic fragments that have been shotgun sequenced and assembled, are clustered together into so-called bins, to ideally reproduce entire genomes. Even if the advances in NGS technologies today are able to provide sufficient sequencing depth for assemblers especially designed for metagenomic experiments, the binning of assembled contigs into clusters on strain- or species-level is still a tough challenge [88]. In principle, there are two different approaches that have been implemented to overcome this obstacle; the first one uses distinct genomic signatures, e.g. k-mer frequencies, as those are characteristic to each genome [89]-[91]. It has been shown that the frequency of oligonucleotide occurrence is conserved over the genome within a certain species, whereas noticeable differences exist between distinct species [92]-[94]. This approach has limitations as in very complex communities, not all organisms exhibit genomes with extreme base compositions needed for separation of related microorganisms whose tetranucleotide frequencies are very similar [95].

On the other hand, the second method deals with the creation of coverage profiles and comparing them across multiple samples [95]-[97]. The idea behind these techniques is that contigs with similar coverage profiles, are likely to be derived from the same organismal population [96]. Some attempts also use a combination of the two approaches [98]-[100]. In this study *CONCOCT* [98] is used, an algorithm which uses a combination of sequence-composition- and coverage-dependent analysis for the automatic binning of contigs or scaffolds into distinct species-level clusters. CONCOCT is useful as the analysis of coverage values does not have to be executed manually, ensuring better reproducibility between different studies.

However, CONCOCT also has the same inherent limitations as all of these techniques, being dependent on a high-quality assembly. If the number of contigs that are restricted to a certain species is low, then the software will not form a distinct cluster out of them.

1.14 Annotation of metagenomic bins

"What are they doing?" - this is the second fundamental question in metagenomics that aims to gain insights into the community's physiology by determining the collective functions encoded in all genomes in a community. Annotating functions to metagenomic sequences can be a quantitative measurement of the functional diversity of a community [101].

Generally, a functional annotation of metagenomes involves two distinct steps, the gene prediction and the functional annotation of the predicted gene [38]; [101]. *Figure 4* illustrates a typical functional annotation workflow in a metagenomic study.

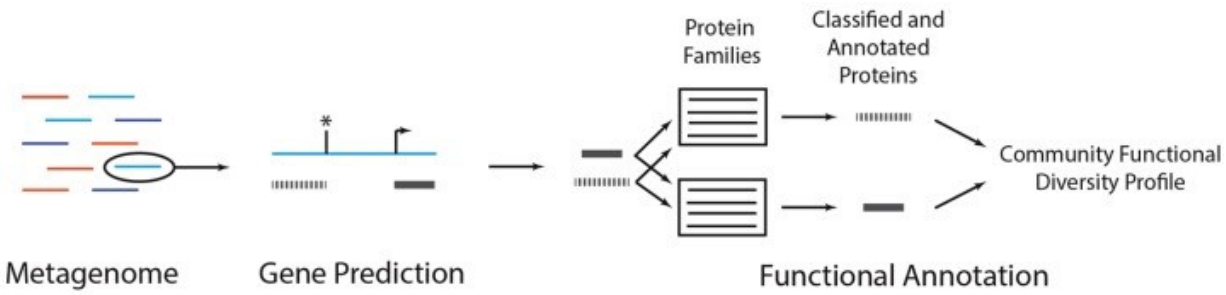

Figure 4: Typical steps involved in a metagenomic functional annotation analysis. According to Sharpton, 2014, modified.

Gene prediction refers to the process of identifying non-coding and coding sequences (CDS) within a read. These predictions do not need assembled reads as the prediction is also possible on unassembled metagenomic sequencing reads. Prediction on unassembled or poorly assembled reads is mainly a more challenging attempt as the read lengths are rather short in NGS experiments and therefore it involves the finding of partial coding sequences too [101]. Gene prediction is important in the functional annotation of metagenomic sequences, but it is truly crucial in identifying completely novel genes.

In general, there are three different approaches for gene prediction prior to functional annotation. The first one uses databases of known genetic sequences and maps metagenomic reads or contigs to the database entries. This method is often referred to “fragment recruitment” in literature. If the gene-entry in the database possesses a functional annotation, then this can be used to functionally label the metagenomic sequence of interest. The problem of this attempt is that it is not possible to identify novel genes or more diverse homologs of a known gene [101]. The second approach is similar to the first, but it additionally includes the translation of each read into all six protein-coding frames. The resulting peptides are then aligned to a database containing sequences of known proteins. As this method relies on database comparison too, novel genes are not identifiable as well. Still, it is possible to identify more diverged homologs of known protein sequences [101]. The third approach is referred to as *“de novo”* or *“ab initio”* gene prediction and the algorithms tracing this strategy use gene prediction models which are trained by distinct features of microbial genes. These properties are for example gene length, the codon usage or a GC-bias. There are numerous of those software tools using this approach, for example *MetaGeneMark* [102], *Orphelia* [103], *Glimmer-MG* [104], *GeneScan* [105], *Prodigal* [106] and *MetaGeneAnnotator*

[107]. They differ in the quality of the used training sets and their capability of handling short or error-prone sequences. The key advantage of software tools that rely on this idea is that it is completely independent of sequence similarity to certain reference databases and therefore this technique is the only one capable of identifying completely novel genes [101].

When the CDS in a metagenomic read or contig has been predicted it can be functionally annotated. Annotation of a genomic sequence is not done *de novo* as it comprises mapping of the CDS of interest to gene or protein databases of known sequences [38]. This is why the annotation of metagenomic sequences is a computationally very intense task. Commonly used reference libraries include the *Kyoto Encyclopaedia of Genes and Genomes* (KEGG) [108] which contains metabolic pathway modules; the *SEED* [109] system, which links functions to higher-level functional subsystems; *COG* [110] and *EggNOG* [111], databases of orthologous protein groups; *MetaCyc* [112], contains metabolic pathways and HMM databases like *Pfam* [113], which comprises models for protein domains [38]; [40]; [101].

CDS-prediction is not sufficient for annotation of metagenomic sequences. Other elements that have to be identified include *CRISPR* elements and *non-coding RNAs* (ncRNAs) like *tRNAs* or *rRNAs*. Complete frameworks have been developed that include all of the steps introduced above, CDS prediction as well as CRISPR element and ncRNA identification. In this study, the metagenomic bins were annotated by the use of *ConsPred* [114], a fully automatic, integrative and comprehensive annotation-framework for prokaryotic genomes.

1.15 Phenotype prediction in metagenomic bins

The persistent progress in NGS technologies led to a rapid increase in protein sequences and their respective databases. Assigning functions to these sequences is challenging and therefore many of them still remain unclassified. Computational studies have shown that prokaryotic proteins are highly conserved to a great extent. This fact is very important in the analysis of poorly investigated microorganisms as it enables functional deductions from well characterized homologous proteins. For deriving reasonable assertions for the function of two related proteins, their respective genes have to meet the criterion of *orthology* [110]. Orthologous genes have originated from a common ancestor gene, through a speciation event, and they perform either the same or very similar function in the two descending species. The protein *Clusters of Orthologous Groups* (COGs) database has evolved through the classification of proteins, based on the concepts of orthology in entirely sequenced genomes [115]. The concept of COGs is especially useful in functional or evolutionary genome studies as the information acquired of a single member can be transferred to the entire COG as orthology implicates functional similarity. A phenotype is a distinct trait of an organism that is noticeable when a certain genotype is expressed depending on the environment and conditions [116]. Those phenotypic traits can be highly versatile in microorganisms. Literature indicates that one of the basic methods for phenotype prediction in microbial genomes are computational methods which are based on databases for COGs [117]. As the expression of specific traits can be dependent of multiple genes and their specific combinations, the analysis of single marker genes is not sufficient for the prediction of many phenotypes. The phenotype prediction in the metagenomic bins generated in this study was performed by the use of the extended PICA framework implemented by

Feldbauer et al. [117]. They have investigated a method for prokaryotic phenotype prediction that is completely *support vector machine* (SVM) based [117]. Their work is based on the initial software framework PICA, developed by MacDonald and Beiko, which is designed to compare different phenotype prediction approaches [118]. *Feldbauer et al.* edited parts of the SVM plug-in of PICA to allow its application also for novel and incomplete genomes, which is a prerequisite in metagenomic experiments. The limiting factor of predictive power for phenotypes in metagenomes is genome completeness but they have shown that trait prediction is also possible for incomplete genomes to a great extent.

2. Material and Methods

2.1 Obtaining the data

The agricultural biogas fermenter samples were taken from a biogas plant which is located near Cologne (Germany) in March and May 2013. The studied biogas fermenter was running under steady conditions at the time of sampling, with fermentation conditions of 40°C and pH-value of 8. At this time, it produced 536 kW output. The biogas reactor was fed mostly with maize silage (69%), but also cow manure (19%) and chicken manure (12%). Total DNA was then isolated, sheared and libraries were generated for *Illumina*® sequencing as recommended by the manufacturer. The diluted libraries were then multiplex-sequenced on an *Illumina*® *HiSeq 2500* instrument. At first, only one lane was sequenced for each of the DNA isolations in a paired-end mode (2 x 101 bases). But then, two additional lanes (lanes 7 and 8) were sequenced for the second DNA isolation. These steps were done by our cooperation partners at the University of Hamburg. The number of generated reads is indicated in the *Results* section.

2.2 Metagenome sequencing, de novo assembly and mapping

The first step was to look for possible remaining sequencing adapters and removing them. This was done with a Python script generated in our department, the adapter sequences which were checked for, are given in *Supplementary Material* section.

The raw samples were then quality checked with *PRINSEQ-lite* [47] version 0.20.4. Our input data, different lanes of *Illumina® HiSeq* with paired-end reads, were in *FASTQ* format. The sequences were trimmed with a threshold of 30 from the 3' end and sequences below quality score mean of 30 and sequences that were shorter than 70, were filtered out. The *polyA/T-tail* was trimmed with a minimum length of 6 at the 3' and the 5'end. Quality was checked again after performing quality adjustments.

Different assemblies of the quality checked sequencing reads were created using either *Ray Meta* [64] version 2.3.1 or *IDBA-UD* [65] version 1.1.1. This was done in different sets for all sequencing samples, sample 1 was the first DNA isolation, sample 2 the second and also a combination of both was created for the assembly process. Ray Meta was run in parallel on 24 nodes, with a maximum *k-mer* number of 31 and each of the paired-end reads was provided for the assembly process. Quality statistics were checked for both assemblies with PRINSEQ-lite *"stats_all"* command and comparison is given in the *Results* section. As there were significantly better results for the Ray Meta assembly, it was the method of choice and the according results were used for all further steps.

There were also data available for a DNA sequencing of two additional lanes (7 and 8) of sample 2 which was included in the assembly process and all following steps

were conducted with these data. As quality characteristics of contigs and scaffolds were very similar, contigs were used in all further steps. Determination of contig coverage was conducted by using the short-read mapper *BBMap* [119] version 34.86, by mapping reads to generated contigs. Prior to the actual mapping step, it was necessary to split up the sequence read files into chunks with a maximum size of 2 million. Then the Ray Meta assembled contigs of sample 1, sample 2 and the combined sample 12 were indexed by BBmap. The actual mapping step was submitted to the grid engine system of our department giving the chunks in a paired-end mode. The *SAM* files of all three samples were converted, sorted and merged by using *Samtools* [120]. The average contig-wise coverage was calculated by *BEDtools* [121] version 2.20.1. The number, respectively percentage of reads in the assembly and mean coverage of contigs was calculated by an in-house script (given in *Supplementary Material*) as control. Data are given in the *Results* section.

As it has been shown that low-coverage and short contigs are error-prone [122], contigs that were less than 1 kb and had an average coverage below 3 were discarded from the assembly. This was done with a custom written Python script, given in *Supplementary Material*. *Bowtie2* [67] was used for assessing contig coverage, prior to the actual binning step. The *"bowtie2-build"* tool was used for creating Bowtie2 indices. Mapping the reads of each sample back to the assembly was done with the *"map-bowtie2-markduplicates.sh"* script contained in *CONCOCT* [98] version 0.4.0 software package. Parameters used here were *"-c"* for computing coverage histogram with *genomeCoverageBed*, *"-t"* for the number of threads (here 8), *"-p"* for extra parameters given to Bowtie2 (here *–q*).

2.3 Taxonomic community profiling

Taxonomic read profiling was conducted by a sequence similarity search of the raw samples using *Rapsearch2* [70] version 2.23, against an in-house generated database of universally conserved proteins. This database contains universally conserved sequences from the *NCBI non-redundant* database, occurring in 98% of all eukaryotes, bacteria and archaea and it was clustered to a level of 97% sequence similarity with the purpose of removing redundancy.

The lanes of sample 2 were converted from *FASTQ* to *FASTA* format and split into chunks of length 500,000. Rapsearch2 analysis was run with the database of universally conserved protein sequences mentioned before. Results of chunks for the certain lanes were merged afterwards. Taxonomic assignment of the Rapsearch2 results was conducted by using *MEGAN5* [71]. To speed up the analysing process and reducing the amount of data loaded into MEGAN5, Rapsearch2 results were pre-filtered to a minimal *bitscore* of 60. This was done with a custom made Python script which is given in *Supplementary Material*. MEGAN5 was run for each lane separately and results were combined later. MEGAN5 outputs were graphically viewed with *KRONA* [123] version 2.5.

Taxonomic profiling of assembled contigs was done by the use of *AMPHORA2* [73] and the set of 31 universal marker genes. Identification of bacterial and archaeal marker sequences was done with the *"MarkerScanner.pl"* script of the AMPHORA2 software package. This program identifies the marker sequences in the input sequences and generates a protein *FASTA* file for each marker gene in the working directory. Parameters used were *"-DNA"*, as input sequences are DNA sequences. Next, the *"MarkerAlignTrim.pl"* script was run for aligning, masking and trimming of the marker protein sequences. Options given here were *"-WithReference"*, for

keeping reference sequences in the alignment and as outputformat *"phylip"* was chosen. AMPHORA2 results were filtered for a minimum length of 1 kb and a minimum coverage of 3. The NCBI taxonomy IDs were mapped to phylogenetic lineages given by AMPHORA2.

Comparison of read-based and assembly-based taxonomic community profiling was done for verifying if there is compliance between the plain sequencing reads in the sample and the assembly data. KRONA charts were compiled for visual comparison, as for taxonomic read-analysis, and are given in *Results* section.

2.4 Filtering and taxonomic profiling of rRNA sequencing reads

SortMeRNA [78], a local sequence alignment tool capable of filtering, mapping as well as OTU-picking, was used to filter out rRNA fragments from the metagenomic sequencing reads. For this purpose, the *FASTQ* files of the sequenced reads were split up into chunks with a size of 1 million. SortMeRNA version 2.0 was then used to merge paired-end reads, as the sequencing data were in two separate files, one for the forward and one for the reverse paired-end reads, and the software only accepts one input file. This was done with the script *"merge-paired-reads.sh"*, included in the software package. The desired rRNA sequences were filtered out against the 8 indexed and prepacked databases, provided by the software. The chunks, that had been merged in the previous step, were given as input files. As output, it was defined to just report the first alignment per read reaching the E-value, in *SAM* alignment format for the aligned sequences. The rejected sequences were chosen to be reported in *FASTQ/FASTA* format.

Furthermore, overall statistics were chosen to be reported and the *verbose* function was used. The reads were loaded into memory by the use of one thread and a maximum of 5000 Mb. All results of the chunks were merged again afterwards.

The *orphan reads*, which are paired-end sequences where only one of them had mapped to the reference databases, were afterwards used for a sequence homology search against the *SILVA* database [75], release number 119. For reducing the amount of data, prior of running a BLAST job, the sequence files were split up into length of 1,000. These chunk files were used for running a *BLASTn* [69] homology search with default parameters. As stated above, the SILVA database number 119 was used as BLAST database. Taxonomic assignment of the BLASTn results was conducted by using *MEGAN5* [71]. Again, to speed up the analysing process and reducing the amount of data loaded into MEGAN5, results were pre-filtered to a minimal *bitscore* of 60. MEGAN5 was run for each lane separately and results were combined later. MEGAN5 outputs were graphically viewed with *KRONA* [123] version 2.5

2.5 CD-HIT OTU assessment

CD-HIT [82] was used for assessing knowledge about the number of underlying OTUs (operational taxonomic units). For this purpose, the previously generated AMPHORA2 marker files, in *PEP* format, were used as input and the threshold of clustering identity was varied from 90%-98% in even steps and additionally one run with 99% identity was performed. For every run the word size was set to "5", the sequence name in the *FASTA*-header was used until first white space and 8GBs of *RAM* and 4 threads were used.

The number of clusters for each marker was used for calculating an average distribution of OTUs over the different clustering identity values and was graphically illustrated in a boxplot. As there were huge differences in the average number of OTUs between Bacteria and Archaea, the respective evaluations have been done separately.

2.6 Binning of metagenomic contigs based on composition and differential coverage data

The visualisation tool *Elviz* [124] was used for a graphic illustration of contig coverage, length, GC content, taxonomy and in general for determination of possible binning strategies. Binning was performed with the software *CONCOCT* [98] version 0.4.0, which is using the composition and differential coverage data of contigs. CONCOCT was run with default parameters as suggested in the detailed CONCOCT tutorial. As described there, long contigs were cut up to a final length of 10 kb and previously generated contig coverage information was used here. Then the *"gen_input_table.py"* and the *"bam_to_linkage.py*" scripts, both part of

CONCOCT software, were used for generating a coverage and a linkage table. This was done exactly like explained in CONCOCT tutorial. The input table was parsed to just contain mean coverage for each contig in each sample. CONCOCT binning was run with standard settings and with the *"-c"* parameter, the maximal number of clusters was set to 400. As CONCOCT gave some strange bugs, format of the contig names of all needed samples had to be adjusted first. *CheckM* [125] version 1.0.3 was used for assessing completeness and contamination of the bins and also for creating plots as a graphical representation. The CheckM documentation suggests the lineage specific workflow for determining completeness and contamination of genome bins, which uses lineage-specific marker sets. The workflow generally consists of four steps that are mandatory and one step that is recommended and can be executed in one single run with the *"checkm lineage_wf"* command, which was used here with 8 threads. For creating plots that are depicting genome bin quality, the *"bin_qa_plot"* command of CheckM was used. This plot gives a visual illustration of *completeness*, *contamination* and *strain heterogeneity* within each of the genome bins.

2.7 Refinement of the binning process and second round of CONCOCT binning

Bins that showed a high degree of contamination were inspected by *VizBin* [91]. This allowed a further separation of bins, which seemed to form two or more distinct clusters. Contrary, bins that showed a very low degree of contamination but were incomplete to a certain extent were used for merging via the *"checkm merge"* function. For this purpose, the taxonomic-specific workflow of CheckM was used as suggested in the manual. This workflow analyses all genome bins with the same marker set and consists of three steps that are mandatory and one recommended step. The *"checkm taxon_list"* command produces a table that indicates all taxa for which custom marker sets can be created. The *"checkm taxon_set"* command was used for creating a marker set for the domain of Bacteria and one for the archaeal domain. The markers in the genome bins were analyses with the *"checkm analyse"* command and a use of 4 threads. Quality was checked afterwards with the *"checkm qa"* command, by the use of 4 threads. Genome bins were merged with the *"checkm merge"* command, for each of the two domains separately, and the use of 4 threads.

After merging of genome bins, the quality statistics were checked and the resulting bins were analysed in coverage plots for checking quality improvements. These plots were created by the use of some custom written Python scripts and *R script*. Some of them are given in *Supplementary Material.*

After all refinement steps, bins that showed completeness higher than 80%, contamination lower than 10% and a heterogeneity value higher than 50%, were filtered out of the resulting bins. These bins were considered as *"high quality bins"* and their underlying contigs were filtered out of the assembly file. This was done by putting all contigs of the high quality bins into one file that served as a *"blacklist"*.

The actual filtering step was carried out with the command line tool *"grep"*, giving the assembly file and the blacklist file as input as well as the *"-w"*, *"-F"* and *"-f"* options.

The CONCOCT coverage table was also filtered for contigs belonging to the high quality bins. The remaining contigs, that belong to bins with lower quality, were also cut up to a final length of 10 kb and another round of CONCOCT binning was performed exactly as described above. This procedure had the underlying idea, that binning could be refined by filtering out the good bins, which could hinder further binning of lower quality bins. Quality criteria as completeness, contamination and heterogeneity of bins were assessed and quality plots created by CheckM version 1.0.3, exactly as described above. All final bins were categorized into four different quality classes based on completeness, contamination and heterogeneity values. Values are given in *Results* section. Only bins that were grouped into these four classes were used for all further steps.

Taxonomy of the bins resulting from the second CONCOCT binning was assessed by determining consensus lineage of all bin-specific marker genes employing AMPHORA2. The cut-off confident scores were higher than 0.8.

2.8 Genome bin annotation

Annotation of genome bins was done with the annotation framework *ConsPred* [114] version 1.21. It was run exactly as described in the documentation. The *"conspred_input_specification.txt"* file was modified in a way that the parameters *"taxon exclude"*, *"minimal number rrna"* and *"minimal number trna"* were all set to *"0"*. Since metagenomic bins are no representation of complete genomes, there cannot be made any solid estimations about the minimal RNA numbers and therefore these parameters are set to zero, because otherwise the annotation process would stop.

As part of the ConsPred workflow, a sequence similarity search of protein coding genes against the *KEGG* database [108], version of March 2014, was performed. This information was needed later on for the prediction of cellulose processing enzymes.

2.9 Phenotype prediction in metagenomic bins

Phenotype prediction was based on the *PICA framework* [118], which was extended with various machine learning techniques for a reliable prediction of phenotypic traits based on comparative genomics and was performed as described in *Feldbauer et al.* [117]. The first step was the assignment of *COGs* to the metagenomic bins. This was done by *PRODIGAL* [106] v2.60 gene calling procedure, using the default translation table. *NCBI cognitor* [110] software was used to map these genes to an in-department generated reference of sequences which represents all *eggNOG* [111] version 4.0 COG proteins. These jobs are carried out by *PSI-BLAST* [86] computations. The COG profile was finally created by combining all resulting genotype files to a single file. Testing for and prediction of the desired phenotypes, given in *Table 2*, was done by the extended PICA framework, as described above, using models and scripts written in our department.

Table 2: Phenotypic models used for trait prediction.

aerobe	*bacterial ammonium oxidizers*
anaerobe	*thermophilic*
fakultative anaerobic	*methanotrophs*
gram-negative	*nitrite oxidizers*
halophilic	*nitrifiers*
motile	*intracellular microorganisms*
phototroph	*obligate intracellular*
ammonium oxidizers	*facultative intracellular*
archaeal ammonium oxidizers	

2.10 CAZy database and prediction of carbohydrate active enzymes

Annotation of genes that encode presumable carbohydrate active enzymes was conducted via a sequence similarity search against sequences contained in *CAZy* database [24], version of May 2015. For this purpose, the whole database was downloaded via a custom written Python script. *Getorf*, part of the *EMBOSS* suite [126], was used for extracting the open reading frames out of the assembly of sample 2 with a minimal size of 75. Then a *BLASTp* [69] search was run with the ORFs against the downloaded database. For the BLASTp search, the ORF files that were created via Getorf, were split up into chunks of size 2,000 and the BLAST jobs were submitted to the grid engine system with default parameters. The results of the chunks were combined again afterwards and filtered for a minimum *bitscore* value of $1e^{-20}$. For the enzyme family profiling, only the best matching BLASTp hit was used and queries were assigned to CAZy families. The taxonomic assignment of carbohydrate-active gene candidates was created by a *Rapsearch2* [70] similarity search against the *NCBI non-redundant* [127] database of universally conserved proteins occurring in 98% of Eukaryotes, Bacteria and Archaea and clustered to a level of 97% sequence similarity for removing redundancy. The maximal number of target sequences was set to 20 and an e-value cutoff of $1e^{-2}$ was utilized for this search. The results were filtered for a minimal *bitscore* of 50, for decreasing the amount of data prior to loading into MEGAN5. With the LCA algorithm and default settings, the sequences were classified phylogenetically and exported manually.

MEGAN5 exports were loaded into KRONA version 2.5, for creating a graphic representation of the taxonomic assignment by AMPHORA2 and heatmaps were compiled in R. This was done by the *heatmap.2 function* of the *gplots package* [128].

2.11 RNA-Seq mapping and expression of CAZy enzymes in the metagenomic bins

In March 2015, a sample for extraction and sequencing of RNA, was taken at the same biogas fermenter. The sample was processed, RNA extracted and sequenced by our cooperation partners at the University of Hamburg, Germany.

Illumina raw sequence reads of two lanes were quality checked via *FastQC* [48] version 0.11.4. *PRINSEQ*-lite [47] version 0.20.4 was used to trim and filter the sequencing reads, for a quality improvement. For this purpose, 10 bases were trimmed starting from the 5' end, bases that had a quality score < 5 were trimmed at the 3' end and sequences that had a mean quality below 20, or were less than 70 bp in length, were discarded. The RNA sequencing read files were split up into chunks of size 2 million for handling the huge amount of data. With the use of *Bowtie2* [67] version 2.2.6, the RNA sequencing reads were mapped to the assembled contigs of sample 2. The *"very sensitive"* pre-set of Bowtie was used here. *SAM* records for unaligned reads, discordant alignments for paired reads as well as unpaired alignments for paired reads were all suppressed. The resulting alignment files were converted with the *"view"* command of *Samtools* [120] software package, version 1.3, in *BAM* format. Samtools *"sort"* and *"index"* functions were used and results of the chunks were merged again afterwards.

Determination of the transcriptional activity within a certain bin was conducted by evaluating marker gene expression, identified by *AMPHORA2* [73]. The function *"multicov"* of *Bedtools* [121] version 2.24.0 was used for obtaining coverage values of potential CAZy glycoside hydrolase genes, as described above.

In brief, the gene coordinates of all putative CAZy enzymes were evaluated via a Blastp sequence similarity search against the CAZy database that had been downloaded before.

By the use of a custom-written Python script, *RPKM* values were calculated out of the coverage data. *Figure 5* illustrates the formula for calculation of RPKM values; these are marker-specific and the average of two lanes. Respective square roots of RPKM values were plotted against the bin taxonomy. This was done by the heatmap.2 function of the gplots package [128] in R.

RPKM = reads per kilobase transcript per million reads

$$\boldsymbol{RPKM(X) = \frac{10^9 \times C}{N \times L}}$$

C … number of mappable reads that fell onto the genes exons

N … total number of mappable reads in the experiment

L … sum of the exons in base-pairs

Figure 5: RPKM value calculation, needed for assessing transcription rates of CAZy enzyme clusters.

3. Results

3.1 Conditions and parameters of the agricultural biogas plant

Biogas fermenters constitute a complex habitat of various microbial communities which are crucial for the different steps in the production of hydrogen and methane. It has been shown that the final yield in biogas, which is limited, depends on the initial hydrolysis step of the plant biomass that is fed in [2]; [8]; [129]. A key in optimising the overall biogas production is the identification of limitations that each single process is facing on the way to the final products hydrogen and methane. For this purpose, we took samples of a biogas plant located near Cologne (Germany) and analysed the underlying phylogenetic community structure and their genome contents. This typical one-stage agricultural biogas fermenter was kept under steady fermentation conditions with a temperature of 40°C and a pH-value of 8. The produced output was 536 kW for this 2,800 m³ plant. Main source materials that were fed in, were maize silage (69%), cow manure (19%) and chicken manure (12%). A brief overview of the process conditions and parameters is given in *Table 3*.

Table 3: General parameters and fermenter conditions characterising the agricultural biogas plant, running under steady conditions.

Fermenter characteristics	
Location	*Near Cologne (Germany)*
Volume	*2,800 m³*
Temperature	*40°C*
pH-value	*8*
Produced output	*536 kW*
Source material	*Maize silage (69%)* *Cow manure (19%)* *Chicken manure (12%)*
DNA sequencing platform	*Illumina® HiSeq 2500*

3.2 DNA sequencing and metagenomic assembly

Two different samples for DNA extraction were taken in March (sample 1) and May (sample 2) 2013, as well as another sample in May 2015 for RNA extraction, described in detail in the *Material and Methods* section. Additional lanes 7 and 8 were taken for the May sample. These samples were multiplex-sequenced on an *Illumina® HiSeq 2500* sequencer, generating a large metagenomic dataset, with a total of 897 million reads that were used in two different assembly methods. As indicated in *Table 4*, the additionally sequenced lanes of sample 2 increased the amount of generated reads nearly two-fold.

Table 4: Number of sequencing reads after filtering and trimming of low quality reads. Note that the read number of the combined sample 1 + 2 only slightly increased compared to sample 2 solely.

sample №	№ of reads (101 bp) used for assembly
sample 1	*159,458,382*
sample 2 w/o L7 + 8	*421,986,642*
sample 1 + 2 w/o L 7 + 8	*581,445,024*
sample 2	*737,631,618*
sample 1 + 2	*897,090,000*

Different metagenomic assemblies were generated using either IDBA-UD or Ray Meta. This was done for all samples indicated in *Table 5*. The assemblies generated using Ray Meta showed a more than two-fold increase in generated contigs, note that the additional lanes of sample 2 are not included in this analysis. Hence, the Ray Meta assembled contigs were used for all subsequent steps in analysis.

Table 5: Comparison of the different assembly methods. The number of contigs assembled by Ray Meta is more than twice the number of contigs created by IDBA-UD.

	Total № of contigs	
sample №	**IDBA-UD**	**Ray Meta**
sample 1	*556,160*	*1,201,371*
sample 2 w/o L 7 + 8	*947,772*	*2,003,618*
sample 1 + 2 w/o L 7 + 8	*1,142,608*	*2,319,807*

For comparison purposes, all common quality criteria were calculated and are indicated in *Table 6*. As there was only a slight quantitative increase in the combined sample 1 + 2, but a decrease in the *N50-value*, only reads and assembly data belonging to sample 2 were used for subsequent analysis. The 737 million already filtered and trimmed reads of sample 2 resulted in generating 123,435 contigs > 1 kb.

Table 6: Quality characteristics for the assembly created by Ray Meta. Note that the number of contigs and Mb in contigs for the combined sample 1 + 2 only slightly increases, compared to sample 2. However, the N50 value is higher for sample 2.

	Ray Meta assembly				
sample №	**№ of contigs**	**№ of contigs > 1000**	**Mb in contigs > 1000**	**N50 for contigs > 1000**	**Mb total**
sample 1	*1,201,371*	*57,009*	*209.68*	*7,183.0*	*486.0*
sample 2 w/o L7+8	*2,003,618*	*94,702*	*425.90*	*11,536.0*	*876.6*
sample 2	*2,593,366*	*123,435*	*581.31*	*12,418.0*	*1,161.7*
sample 1 + 2 w/o L7+8	*2,319,807*	*112,571*	*512.03*	*10,871.0*	*1,035.8*
sample 1 + 2	*2,826,937*	*140,535*	*653.73*	*11,784.0*	*1,292.2*

By comparing number of contigs and scaffolds, and their corresponding Mb content, decision was made for continuing work with contig data as there were no striking quantitative differences, numbers are given in *Table 7*.

Table 7: Comparison of quality characteristics for contigs and scaffolds in the assembly of sample 2. There is no notable difference regarding their absolute quantity.

Ray Meta assembly - contigs vs. scaffolds	
№ of contigs	*2,593,366*
№ of scaffolds	*2,568,676*
Mb in contigs	*1161.7*
Mb in scaffolds	*1166.9*
№ of contigs ≥ 500	*293,792*
№ of scaffolds ≥ 500	*269,102*

Determination of contig coverage, via the short-read mapper *BBMap*, revealed 324 million mapped reads in total, indicated in *Table 8*. The additionally sequenced lanes were not solely very important for the assembly process, but they also contributed greatly to the mapping procedure, as lanes 7 and 8 were the ones providing the highest amount of reads in this step.

Table 8: Number of reads mapped to contigs by BBMap, only different lanes of sample 2 are given.

Mapping sample 2	
Sequencing lane	**№ mapped reads**
Lane 1	*46,502,117*
Lane 2	*47,118,419*
Lane 3	*47,648,665*
Lane 4	*47,344,343*
Lane 7	*68,667,968*
Lane 8	*67,419,711*
Total	***324,701,223***

3.3 Taxonomic community profiling

3.3.1 Taxonomic read profiling

Taxonomic read profiling, performing a *Rapsearch2* analysis against the *NCBI non-redundant database*, revealed a total of 7.9 million classified sequencing reads. As there were no obvious differences in relative species classification amongst the different lanes, *Figure 6* shows sequencing lane 8 as representative.

Sequences were classified as Bacteria in 89% of all cases and only 4% were categorised as Archaea. *MEGAN5* placed 5% of all reads as unassigned reads.

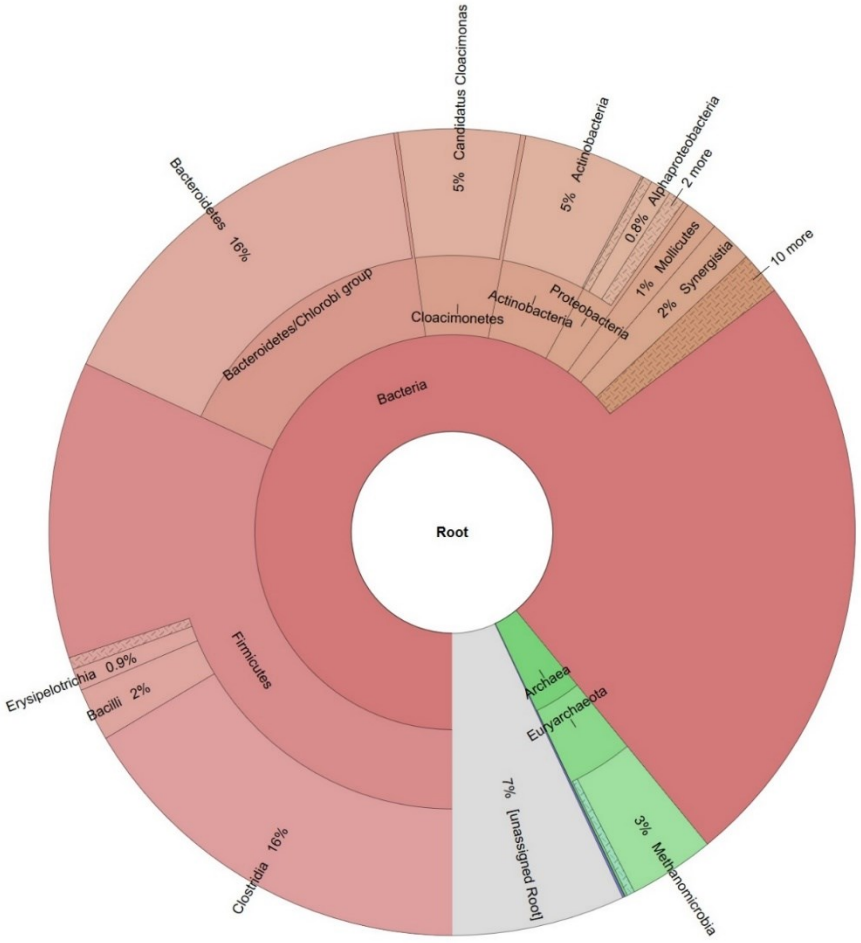

Figure 6: KRONA chart of the taxonomic read profiling for sequencing lane 8. Percentages refer to the relative number of reads assigned to a certain taxonomic level in proportion to the root.

The most prominent phylum within Bacteria was *Firmicutes* with 36%, followed by *Bacteroidetes* (18%), *Cloacimonetes* (6%), *Actinobacteria* (6%), *Proteobacteria* (2%) and other less abundant phyla. The major amount of reads that belong to *Firmicutes* were subdivided into the class of *Clostridia* (52%) and only small sections belonged to *Bacilli* (7%) and *Erysipelotrichia* (3%). The distribution among *Bacteroidetes* was dominated by *Bacteroidia* (62%), and only minor parts were classified as *Flavobacteria* (1%), *Cytophagia* (0.9%) and *Sphingobacteria* (0.7%).

The *Firmicutes*/*Bacteroidetes* ratio was 2:1 for the single read profiling. A detailed overview of the bacterial classification is given in *Table 9*.

Within the Archaea, nearly all reads (99%) were identified as *Euryarchaeota*, among this phylum, 90% were classified as *Methanomicrobia* and only minor parts were assigned to *Methanobacteria* (6%) and *Thermoplasmata* (2%).

Table 9: Detailed overview of the bacterial taxonomic read classification, evaluated by Rapsearch2 search against NCBI non-redundant database.

Single read taxonomy profiling – Bacteria	
phylum	**class**
36% Firmicutes	*52% Clostridia*
	7% Bacilli
	3% Erysipelotrichia
	1% Negativicutes
	37% unassigned Firmicutes
18% Bacteroidetes	*62% Bacteroidia*
	1% Flavobacteria
	0.9% Cytophagia
	0.7% Sphingobacteria
	35% unassigned Bacteroidetes
6% Cloacimonetes	*96% Candidatus Cloacimonas*
	4% unassigned Cloacimonetes
6% Actinobacteria	*98% Actinobacteria*
	2% Cariobacteriia
2% Proteobacteria	*41% Alphaproteobacteria*
	20% delta/epsilon subdivisions
	18% Gammaproteobacteria
	11% Betaproteobacteria
	10% unassigned Proteobacteria
2% Mollicutes	
2% Synerigista	
0.9% Spirochaetia	

3.3.2 Taxonomic profiling of assembled contigs

Taxonomic community characterisation of assembled contigs by *AMPHORA2*, revealed a total of 21,602 contigs that were sorted into different taxonomic levels. *Figure 7* shows the relative taxonomic community profiling for the assembly of sample 2. AMPHORA2 analysed contigs were categorised as Bacteria in 89% of all assembled sequences and only 10% were grouped as Archaea. 0.3% could not be assigned to a specific taxon which corresponds to a total of 10 contigs.

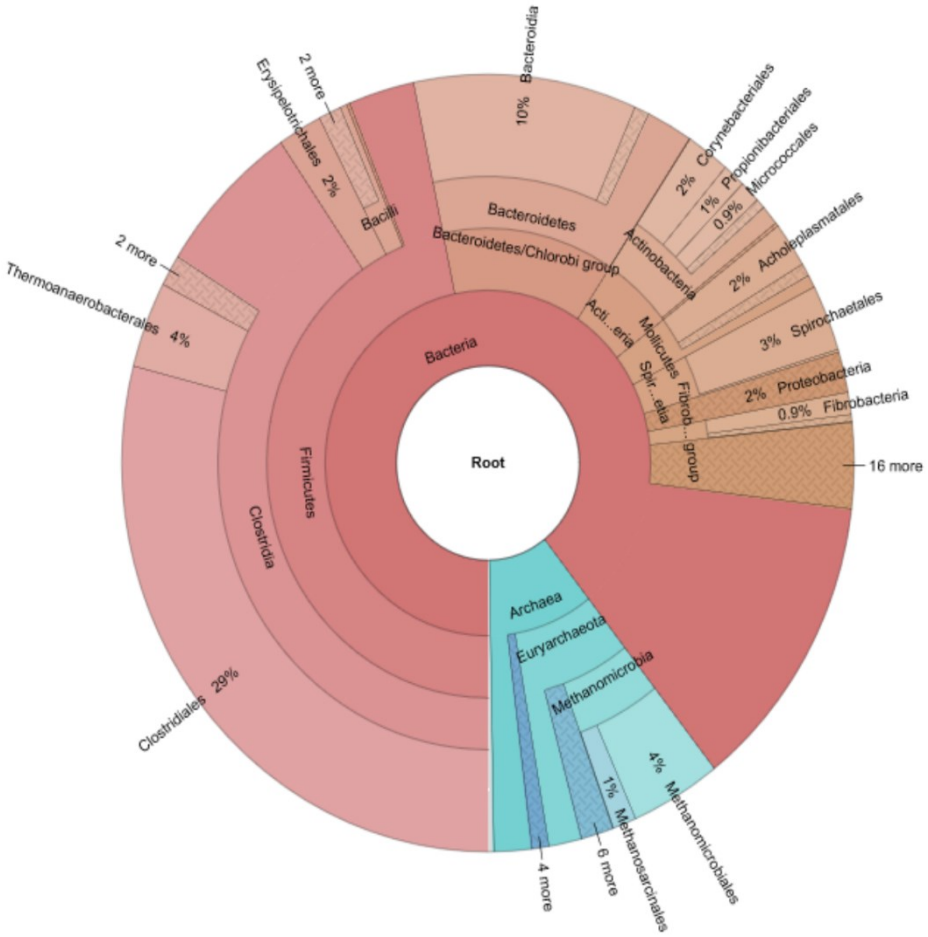

Figure 7: KRONA chart illustrating the taxonomic community characterisation of assembled contigs. Note that percentages refer to the different taxonomic levels in proportion to all assembled contigs that were classified by AMPHORA2 in the taxonomic analysis.

Firmicutes was the most occurring phylum with 52% of all bacterial sequences, followed by *Bacteroidetes* (14%), *Actinobacteria* (6%), *Mollicutes* (3%) and many other less abundant phyla. 14% were unassigned bacterial contigs. *Fimicutes* were subdivided into *Clostridia* (86%) and only minor amounts of contigs were grouped as *Erysipelotrichia* (4%) and *Bacilli* (1%). However, 6% of all assembled sequences were categorised as unassigned.

The dominating class within *Bacteroidetes* was *Bacteroidia* (78%) and only tiny fractions belonged to *Sphingobacteriia* (2%) and *Flavobacteriia* (2%). The proportion of unassigned reads was higher and added up to 16%.

A detailed overview of the bacterial classification is given in *Table 10*. The *Firmicutes*/*Bacteroidetes* ratio for the analysis on assembled contigs was 3.7:1.

Table 10: Detailed overview of the taxonomic composition of assembled contigs belonging to sample 2, evaluated by AMPHORA2. Only bacterial contigs are depicted.

AMPHORA2 taxonomy profiling of assembled contigs - Bacteria	
phylum	**class**
52% Firmicutes	*86% Clostridia*
	4% Erysipelotrichales
	3% Bacilli
	6% unassigned Firmicutes
14% Bacteroidetes	*78% Bacteroidia*
	2% Flavobacteriia
	2% Sphingobacteriia
	16% unassigned Bacteroidetes
6% Actinobacteria	
3% Mollicutes	
3% Spirochaetia	
2% Proteobacteria	
1% Fibrobacteres/Acidobacteria group	
0.9% Synergistales	
0.8% Planctomycetales	
0.7% Chloroflexi	
0.4% Verrucomicrobia	
0.3% Thermotogae	
0.2% Fusobacteriales	
0.2% Deinococci	

Archaeal contigs were classified in 77% of all cases as *Euryarchaeota*, 4% as *Thermoprotei* and 2% were classified belonging to the species of *Nanoarchaeum equitans*. AMPHORA2 also classified 16% as unassigned Archaea. The phylum *Euryarchaeota* was dominated by 64% *Methanomicrobia* and only smaller fractions accounted for *Methanobacteriales* (11%), *Thermoplasmatales* (2%), *Aciduliprofundum boonei* (2%) and *Methanococcales* (2%).

3.4 Filtering of rRNA sequencing reads and determination of their taxonomic origin

Corroborating the taxonomic community profiling of the analysed biogas plant, another attempt was made targeting sequencing reads containing ribosomal RNA genes. Analysis of the filtered ribosomal RNA fragments and their taxonomic origin in the sequenced biogas fermenter sample revealed that, according to the initial expectations, only about 1.6 million rRNA reads were successfully screened and taxonomically assigned. *Figure 8* illustrates a KRONA chart for the taxonomic rRNA profiling for sequencing lane 1 of sample 2, obtained by scouring the *SILVA* database.

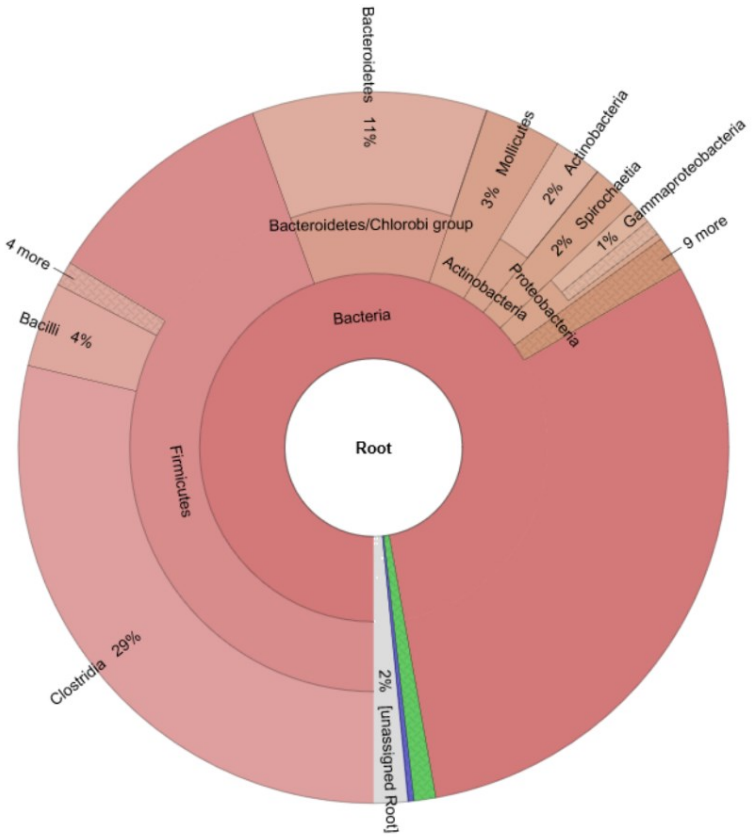

Figure 8: KRONA chart illustrating the taxonomic origin of filtered rRNA sequences. Note that given percentages are relative to the root.

97% of the encountered ribosomal RNA sequences were of bacterial origin and only 1% were assigned to the archaeal domain. Elaborating on bacterial rRNA sequences, 46% of them were assigned to the *Firmicutes* phylum and only 11% were classified as *Bacteroidetes*. This resulted in a calculated *Firmicutes*/*Bacteroidetes* ratio of 4.2:1, being in accordance with the taxonomic profiling on assembled contigs. Notably, 31% of all detected rRNA reads were classified as unassigned Bacteria. *Table 11* is a brief summary of the relative taxonomic assignment and observed bacterial diversity for all reads that obtained a ribosomal RNA tag during *SortMeRNA* analysis

Table 11: Overview of the taxonomic assignment achieved by SortMeRNA filtering of rRNA sequences and BLASTn homology search against the SILVA ribosomal database.

SILVA rRNA taxonomy profiling of rRNA sequences - Bacteria	
phylum	**class**
46% Firmicutes	*65% Clostridia*
	4% Bacilli
	2% Erysipelotrichales
	24% unassigned Firmicutes
11% Bacteroidetes	*53% Bacteroidia*
	4% Flavobacteriia
	4% Sphingobacteriia
	39% unassigned Bacteroidetes
31% unassigned Bacteria	
4% Mollicutes	
2% Actinobacteria	
2% Spirochaetia	
2% Proteobacteria	
0.6% Synergistales	
0.3% Fibrobacteres/Acidobacteria group	
0.2% Chlamydiae/Verrucomicrobia group	
0.1% Atribacteria	

3.5 CD-HIT OTU assessment

The OTU evaluation, for the AMPHORA2 marker protein search on assembled contigs, conducted by *CD-HIT*, revealed huge differences in the number of bacterial and archaeal marker protein clusters. This goes in line with our expectations, as there are known to be considerably less different archaeal phyla in agricultural biogas fermenters comparing bacterial ones. Hence, they were evaluated separately with the intention not to bias the statistical evaluation. On average 195 distinct bacterial OTUs were present in the fermenter sample, depending on the clustering threshold, exact numbers are given in *Table 12*.

Table 12: Overview of the quantitative bacterial OTU distribution at varying percentages of shared identity.

cluster threshold	average № of OTUs
90%	180
92%	187
94%	194
96%	199
98%	203
99%	204

Figure 9 illustrates a boxplot compiled of all obtained bacterial OTU numbers at the different percentages of shared identity in the CD-HIT grouping. The archaeal analysis yielded on average only 6 distinct OTUs.

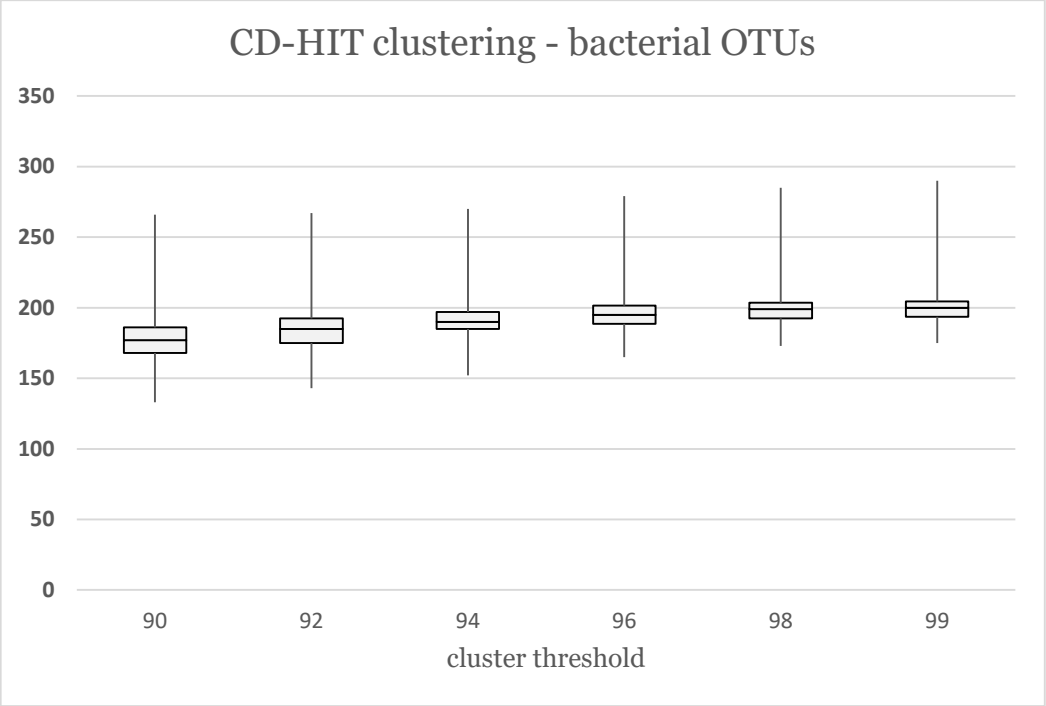

Figure 9: Boxplot figuring the OTU evaluation conducted by a CD-HIT clustering of the AMPHORA2 bacterial marker protein search. Cluster threshold corresponds to the percentage of shared identity for the CD-HIT grouping.

3.6 CONCOCT binning and manual refinement

The first round of *CONCOCT* binning, without any manual modifications, resulted in the grouping of 251 distinct bins with varying quality characteristics, as observed via *CheckM* analysis. Some of the low quality bins were then checked via *VizBin* for possible separation. This attempt resulted in the generation of 15 additional bins. However, CheckM was not solely used for assessing bin quality, but also for possible merging of high quality bins that showed a lower degree of completeness. CheckM default merging function automatically suggested 14 bins for uniting. After checking these bins manually, by compiling coverage plots, 3 of them were free to merge and for all others, the decision was against fusion. *Figure 10 (A-C)* illustrates two bins, for which the default merging by CheckM was approved after checking their corresponding coverage plots. Whereas *Figure 11 (A-C)* highlights one case were merging was not accepted after verification.

As the merging process did not eventuate in significant qualitative improvements, the regarding bins could not be included into the final set of high quality metagenomic bins, as none of them fulfilled the stringent quality criteria.

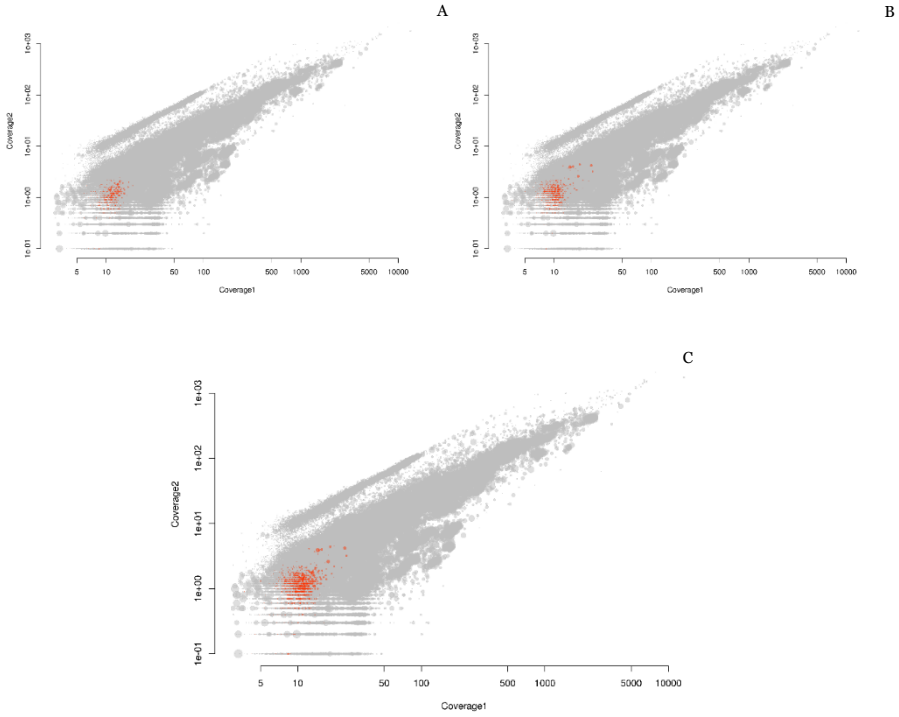

Figure 10: Coverage plots illustrating approved fusion candidates. Bins prior (A,B) and post (C) default CheckM merging.

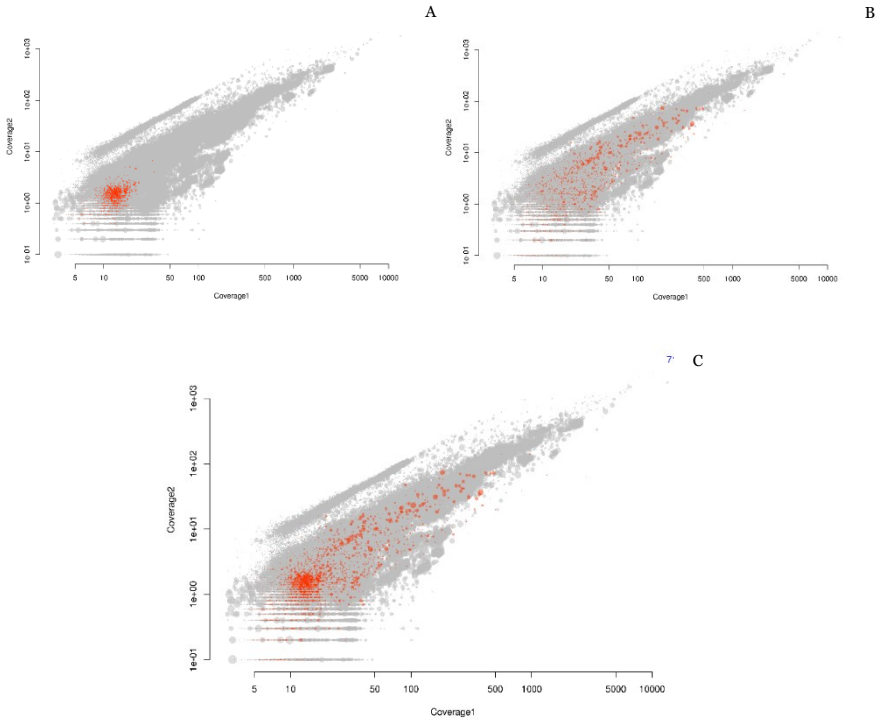

Figure 11: Coverage plots illustrating disapproved fusion candidates. Bins prior (A,B) and post (C) default CheckM merging.

Filtering out all high quality bins and performing a second round of CONCOCT binning resulted in a total of 104 bins that meet very high quality criteria. These bins were sorted in different categories according to the filtering criteria indicated in *Table 13*. They represent the final set that was used in all further analyses.

Table 13: Classes of different quality criteria for the 104 high quality bins.

Category name	%age of completeness, contamination and heterogeneity	№ of bins in category
good bins	>95% compl., <5% cont., or >95% compl., <10% cont., >90% het.	**20**
nearly complete genome drafts	>90% compl., <5% cont.	**20**
nearly complete pangenome drafts	>90% compl., >5% cont.	**37**
incomplete genome drafts	60-90% compl., <7% cont.	**27**

The *"good bins"* category refers to bins, which fulfil the most stringent quality criteria. These bins are nearly complete (95%) and have a very low amount of contamination, 5-10% depending on the level of heterogeneity. Heterogeneity in this context indicates if the present contamination traces back to a closely or distantly related species. The *"nearly complete genome drafts"* show a little lower completeness and the difference to the *"nearly complete pangenome drafts"* is that the pangenome drafts have a higher level of contamination. The terminus pangenome in this context signifies that these bins consist of a mixture of very closely related species. *"Incomplete genome drafts"* are low in contamination but only show 60-90% of completeness, however these bins are also important for further analyses. As valid for all metagenomic experiments, the amount of information that can be derived from the analyses, are always strongly dependent on the missing parts of the recovered metagenomes.

Figure 12 illustrates the quality fulfilment of the different members of the *"good bins"* class. Green bars indicate that a certain marker gene is solely present as a single copy. Bars in different shades of blue display that the marker is present more than once, but the contamination comes from a closely related species. Whereas different shades ranging from yellow to red show that the underlying multi-copy markers are deriving from distant species. *Figure 13* gives a brief qualitative overview about the three other bin categories.

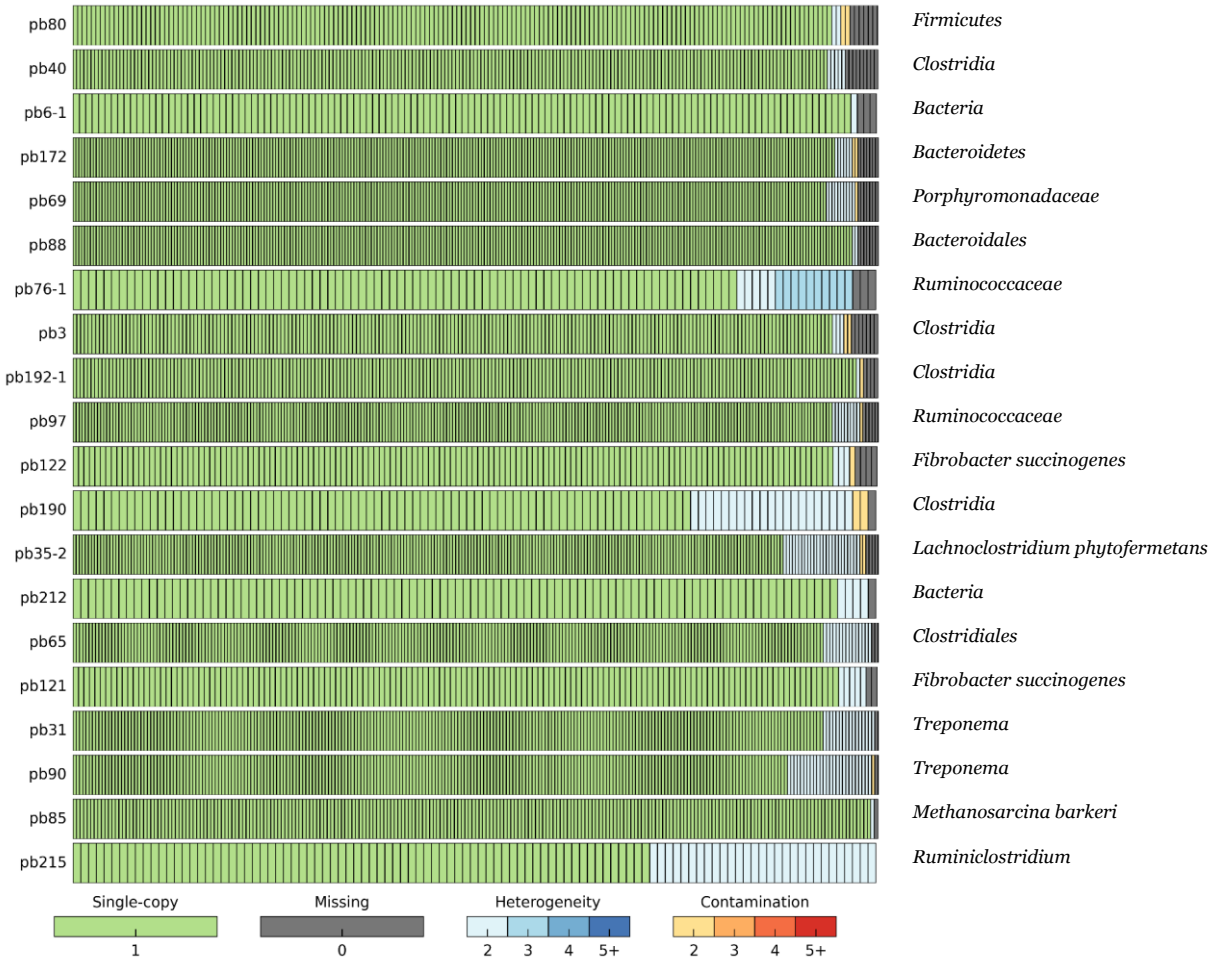

Figure 12: CheckM quality plot illustrating all bins within the "good bins" category. AMPHORA2 taxonomy (consensus score > 0.8) is given for each bin on the right.

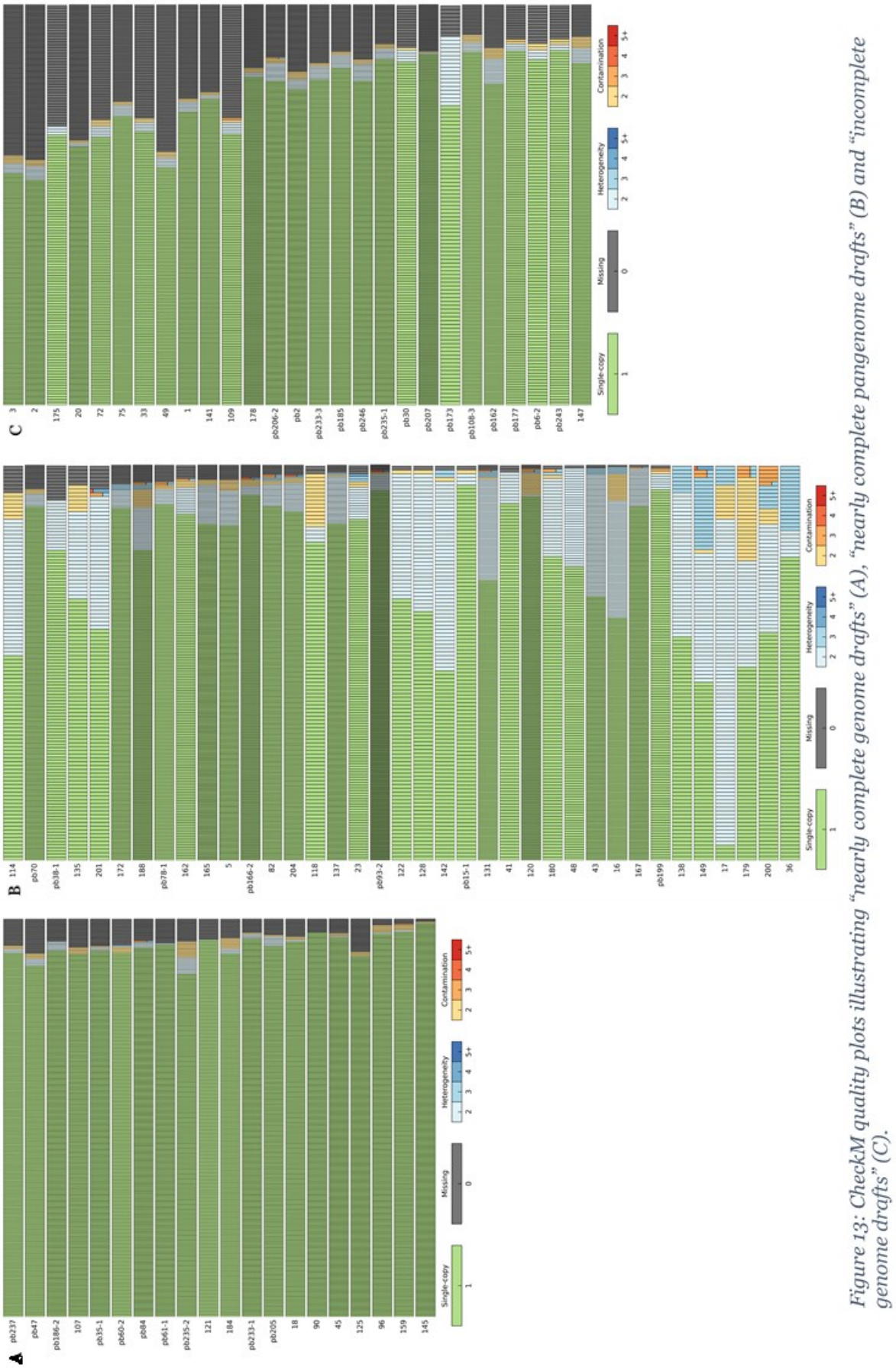

Figure 13: CheckM quality plots illustrating "nearly complete genome drafts" (A), "nearly complete pangenome drafts" (B) and "incomplete genome drafts" (C).

3.7 Taxonomic profiling of metagenomic bins

In terms of taxonomic community profiling, the metagenomic binning goes in line with the single read and assembly analyses. 57 of the 104 high quality bins were assigned to *Firmicutes* and the major amount within was attributed to the class of *Clostridia* (51). Whereas 21 of the bins were marked as *Bacteroidetes*, and the main class within was *Bacteroidia* with 16 representatives. Therefore, a *Firmicutes*/*Bacteroidetes* ratio of 2,7:1 was observed amongst the different metagenomic bins. A table with a detailed overview of the taxonomic classification, the estimated completeness and contamination rates, as well as the size of each of the high quality bins is given in *Supplementary Material*.

The other prominent phyla were *Spirochaetes* (4), *Fibrobacteres* (3), *Euryarchaeota* (3), *Verrumicrobia* (2) and *Actinobacteria* (2). For some other phyla, only one representative bin was accordingly classified, for example *Tenericutes*, *Proteobacteria* and *Planctomycetes*. Some of the bins were taxonomically classified at species level with high confidence scores by AMPHORA2, and therefore their genomes could be reconstructed at a very high level of completeness. Confidence threshold for AMPHORA2 taxonomy classification was set to 0.8. In the phylum *Firmicutes* some bins could be taxonomically assigned down to species level, for example *Lachnoclostridium phytofermentans* (binIDs pb35-2, pb186-2, pb35-1 and pb235-1) as well as *Ruminiclostridium thermocellum* (binID 96), *Mageeibacillus indolicus* (binIDs 18, 104, pb185 and pb233-3), *Oceanobacillus iheyensis* (binID pb84) and *Pelotomaculum thermopropionicum* (binID 165).

In the *Bacteroidetes* phylum, the species *Paludibacter propionicigenes* (binIDs 145 and 201), *Alkaliphilus oremlandii* (binID pb70), *Erysipelothrix rhusiopathiae* (binID109), showed the deepest taxonomic classification.

Nine of the 104 bins were only classified as Bacteria where no further assignment was possible (8%). As expected, only a small fraction of all assigned bins were classified as Archaea. All of them belong to the phylum *Euryarchaeota* and it was possible to assign one of them to the species *Methanosarcina barkeri* (binID pb85), a methanogenic archaeon. *Table 14* gives a brief overview of the taxonomic assignment, the number as well as *N50 values* of contigs for all bins belonging to the *"good bins"* category.

Table 14: Overview of the taxonomic classification by AMPHORA2 (consensus score > 0.8), the corresponding binIDs, the number of contigs belonging to each bin as well as their N50 values.

Good Bins			
binID	**Taxonomy**	**№ contigs**	**N50 contigs**
pb121	*Bacteria|Fibrobacteres|Fibrobacter succinogenes*	*217*	*19,300*
pb122	*Bacteria|Fibrobacteres|Fibrobacter succinogenes*	*259*	*11,855*
pb172	*Bacteria|Bacteroidetes|Bacteroidetes*	*233*	*19,077*
pb190	*Bacteria|Firmicutes|Clostridia*	*234*	*16,089*
pb192-1	*Bacteria|Firmicutes|Clostridia*	*36*	*129,429*
pb212	*Bacteria*	*210*	*22,750*
pb215	*Bacteria|Firmicutes|Clostridia|Ruminiclostridium*	*144*	*45,291*
pb31	*Bacteria|Spirochaetes|Treponema*	*213*	*16,062*
pb35-2	*Bacteria|Firmicutes|Clostridia|Lachnoclostridium phytofermentans*	*143*	*28,325*
pb3	*Bacteria|Firmicutes|Clostridia*	*139*	*33,655*
pb40	*Bacteria|Firmicutes|Clostridia*	*160*	*12,735*
pb6-1	*Bacteria*	*37*	*105,819*
pb65	*Bacteria|Firmicutes|Clostridia|Clostridiales*	*77*	*66,410*
pb69	*Bacteria|Bacteroidetes|Porphyromonadaceae*	*181*	*17,355*
pb76-1	*Bacteria|Firmicutes|Clostridia|Ruminococcaceae*	*134*	*21,374*
pb80	*Bacteria|Firmicutes*	*95*	*77,072*
pb85	*Archaea|Euryarchaeota|Methanomicrobia|Methanosarcina barkeri*	*111*	*70,311*
pb88	*Bacteria|Bacteroidetes|Bacteroidales*	*336*	*17,475*
pb90	*Bacteria|Spirochaetes|Treponema*	*215*	*22,018*
pb97	*Bacteria|Firmicutes|Clostridia|Ruminococcaceae*	*91*	*36,708*

3.8 Phenotype predictions in metagenomic bins

Prediction of 17 selected phenotypic traits was conducted by the use of an extended PICA framework, as described in the *Material and Methods* section. *Table 15* gives a detailed overview of the characteristics chosen for investigation. The majority of bins (88%) were predicted to be anaerobic and only one of them was classified as aerobic. Considering the fact that biogas production is an anaerobic process, this result seems to be rather plausible. Oxygen inflow during sample taking might be responsible for the detection of one aerobic organism. 39% of the analysed bins were predicted to be gram-negative, 53% were treated as motile organisms, only 2% were categorised as halophilic, 19% were labelled as thermophilic and none of the 104 bins was classified as a phototrophic prokaryote. These findings do not contradict with the initial expectations. Notably, a quite high number of bins were predicted to have intracellular traits. This was rather unexpected for a community present in a biogas plant. Though, intracellularity is a trait which is predicted by the lack of certain genes. This is problematic in metagenomic experiments as most of the reconstructed genomes are not complete and that might influence prediction. Nevertheless, if those bins are complete to a high extent, then further analyses with novel prediction models would be interesting to clarify if those bins are truly symbiotic organisms. No bins were labelled as ammonium oxidizers, nitrite oxidizers, nitrifiers and methanotrophs. Methanotrophic organisms are very important in this context because of the fact that these species could affect the energy gain of the agricultural biogas plant.

Table 15: List of 17 phenotypic traits that were searched for in all 104 metagenomic bins, which of them were predicted and how often.

Phenotype	№ of bins	Phenotype	№ of bins
aerobe	1/104	**bacterial ammonium oxidizers**	0/104
anaerobe	91/104	**thermophilic**	20/104
facultative anaerobe	1/104	**methanotrophs**	0/104
gram-negative	41/104	**nitrite oxidizers**	0/104
halophilic	2/104	**nitrifiers**	0/104
motile	55/104	**intracellular**	17/104
phototrophs	0/104	**obligate intracellular**	18/104
ammonium oxidizers	0/104	**facultative intracellular**	0/104
archaeal ammonium oxidizers	0/104		

A table containing all trait predictions for each bin can be found in *Supplementary Material* section. However, all the assertions made above are only predictions and should be clearly treated as such. These results should give a first impression about the underlying community, but they cannot be taken as concrete evidence.

3.9 RNA-Seq mapping and evaluation of CAZy enzymes expression

The CAZy database is a collection of enzyme families that modify, create or degrade glycosidic bonds for example glycoside hydrolases (GHs), carbohydrate esterases (CEs), glycosyl transferases (GTs) and others. Mapping of the processed and filtered RNA sequencing reads to the assembled contigs of sample 2 and the subsequent coverage value evaluation of potential GH families identified in the metagenomic bins was conducted as described in the *Material and Methods* section. This exploration resulted in a visual representation of all 104 high quality bins and their respective taxonomic classification plotted against the calculated expression values of the most important glycoside hydrolase enzyme clusters. Those 15 highest expressed glycoside hydrolase families in the biogas plant samples were *GH1*, *GH3*, *GH5*, *GH6*, *GH8*, *GH9*, *GH12*, *GH14*, *GH30*, *GH44*, *GH45*, *GH48*, *GH51*, *GH74* and *GH94*. All these enzyme families cluster together different enzymes that are needed in the hydrolytic breakdown of carbohydrates via cleavage of glycosidic bonds. The corresponding heatmap is illustrated in *Figure 14*.

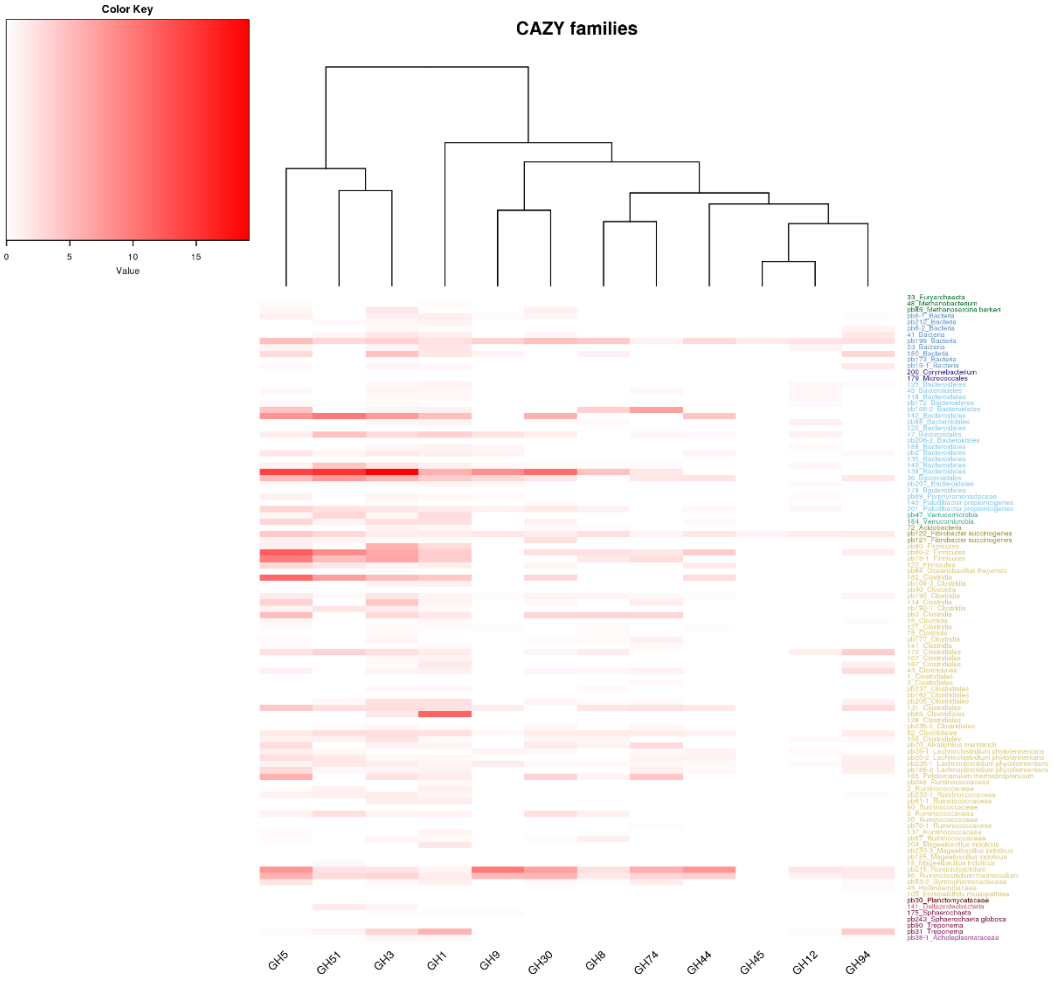

Figure 14: Heatmap illustrates the transcriptional activity of GH families and their expression levels in all of the 104 clustered bins. Expression values are represented as the square root of RPKM values and taxonomic assignment was made by AMPHORA2 with a consensus score > 0.8. Yellow highlighted bins were assigned to Firmicutes and light blue ones to Bacteroidetes.

3.10 Prediction and taxonomic assignment of carbohydrate-active gene candidates

The taxonomic investigation of CAZy gene candidates, present in the assembly of sample 2, revealed the same trend towards a greater proportion of *Firmicutes* versus *Bacteriodetes*. Regarding the origin of carbohydrate-active gene candidates, the *Firmicutes*/*Bacteroidetes* ratio calculated to 2.1:1.

Figure 15 gives an overview about the general taxonomic origin of the CAZy family enzymes in the generated assembly of sample 2. 95% of detected CAZy enzyme sequences were of bacterial origin and only 5% were assigned to Archaea. Nearly all of the archaeal sequences (4%) were further assigned to the class of *Methanomicrobia*. The major part of enzyme hits that were classified as Bacteria, were further divided into the phylum *Firmicutes* (53%), most of them subdivided into the class *Clostridia*. The *Bacteroidetes* phylum accounted for 25% of all bacterial sequences and only minor parts were assigned to the phyla of *Fibrobacteres* (8%), *Actinobacteres* (3%), *Cloacimonetes* (2%) and others.

Figure 15: KRONA chart representing the taxonomic origin of all CAZy enzyme gene candidates present in the assembly of sample 2. Note that the percentages refer to the proportion of sequences relative to the root.

The class of *Clostridia* showed the most hits for gene sequences that are classified into these enzyme families. However, also the *Bacteroidetes* phylum seems to contribute greatly to the overall abundance of these GH groups. Representatives of the class *Bacilli* or phyla *Fibrobacteres* and *Actinobacteria* seem to play a role in the hydrolytic breakdown of plant biomass as well, although the examined GH families are predicted to be less common in those taxonomic groups. *Figure 16* illustrates a heatmap plotting the taxonomic origin versus the number of identified potential glycoside hydrolase family hits in the assembly of sample 2.

As stated above, GH enzyme families are one of the main players in the breakdown of plant material. Their value of occurrence was computed via a Blastp similarity search against the NCBI non-redundant database combined with the LCA algorithm of MEGAN5 for taxonomic assignment. The most predominant glycoside hydrolase families in both, the *Firmicutes* as well as *Bacteroidetes* phyla, seem to be GH3 and GH5. Notably, the clusters of GH51, GH12 and GH14 seem to be more common within the *Bacteroidetes* compared to *Firmicutes* phyla.

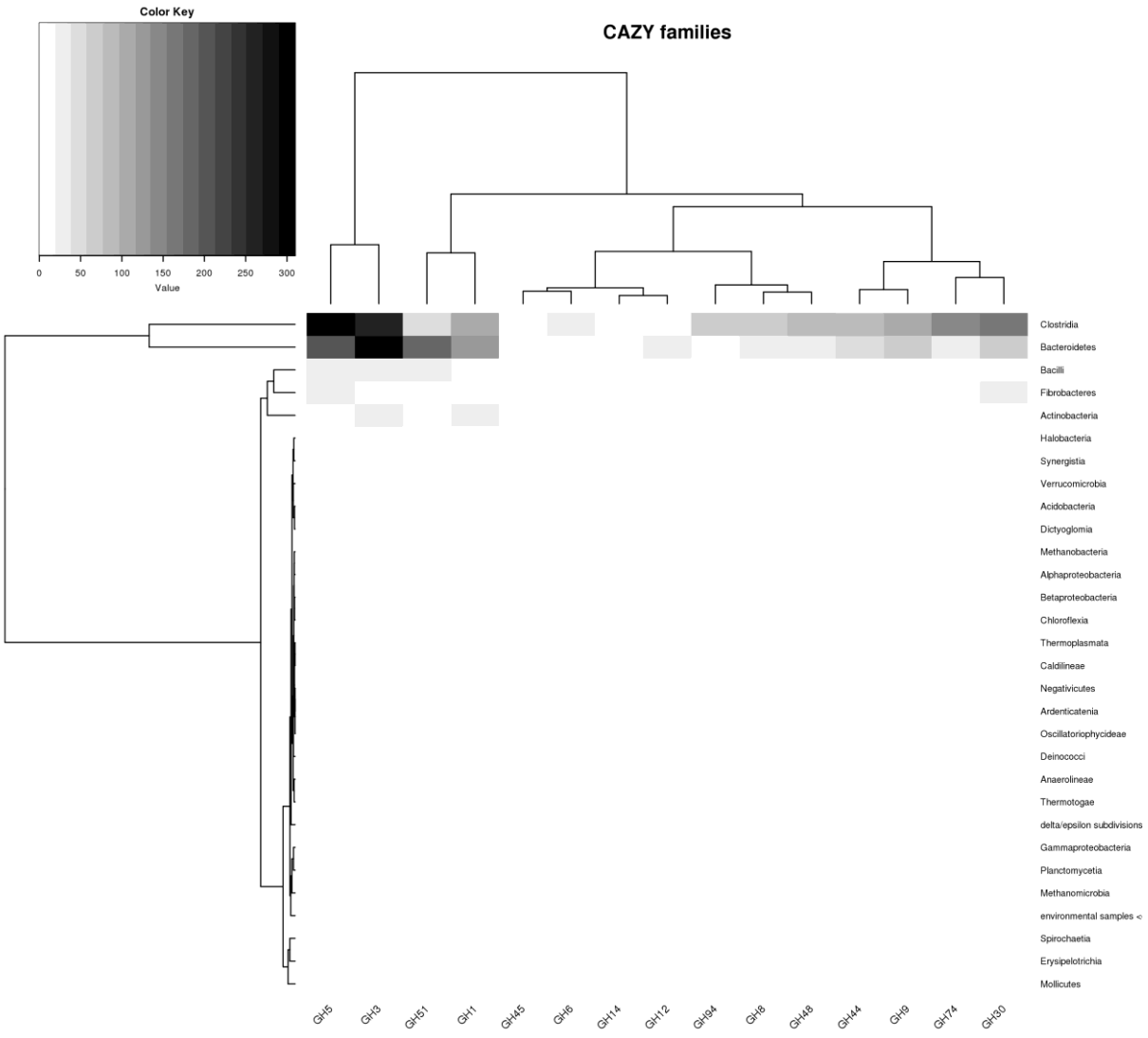

Figure 16: Heatmap representing the potential level of occurrence of different GH family genes in the most represented taxonomic taxa of the assembled sequences of sample 2. Plotted values are actual counts for enzyme-encoding sequences belonging to the respective GH classes. Taxonomic assignment is based on the NCBI non-redundant database.

CAZy enzyme-predictions were also coupled with the sequences clustered into the 104 high quality metagenomic bins. *Figure 17* illustrates a heatmap indicating the predictions of the most important glycoside hydrolase families and the taxonomic assignment for the respective bin. A table containing the actual hits for each GH family found in the 104 high quality metagenomic bins is given in *Supplementary Material* section.

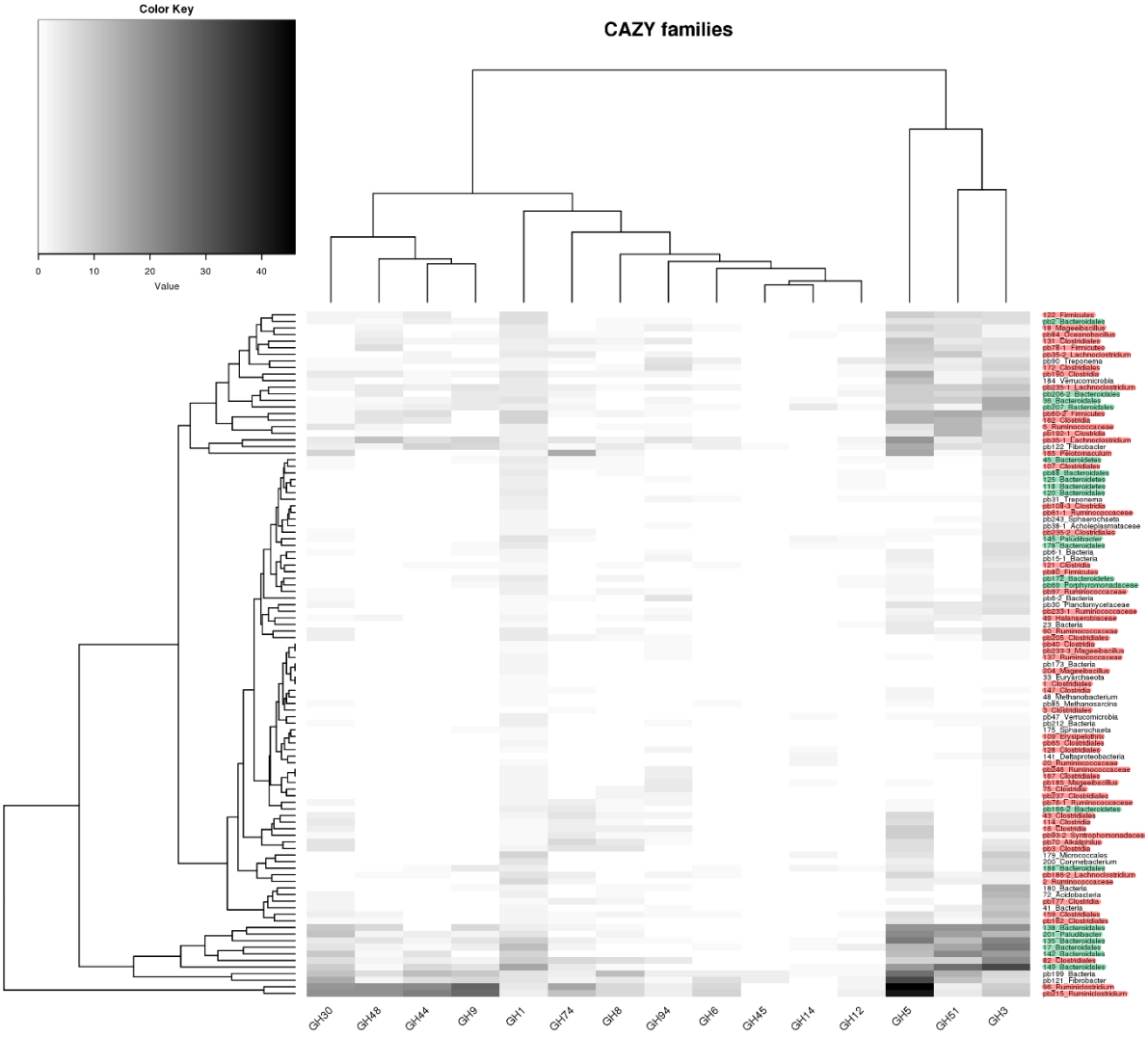

Figure 17: GH family predictions in the 104 high quality bins. AMPHORA2 taxonomic classification is given on the right (consensus score > 0.8). Red highlighted taxa belong to the Firmicutes phylum, whereas a green highlight indicates that these bins were assigned to the Bacteroidetes phylum. Values are actual counts for GH family sequences in the respective bins.

4. Discussion

Biogas fermenters comprise a habitat for microbial communities with highly complex population structures. The members act together in concerted action to conduct the reactions involved in the process of degrading the initial substrate, yielding the final products methane and carbon dioxide. The general procedure and main reactions, being *hydrolysis*, *acidogenesis*, *acetogenesis* and *methanogenesis*, have extensively been studied in the past. Past research focussed predominantly on the methanogenesis because this step results in the generation of the end-products. In contrast, knowledge about the initial hydrolysis, which is the rate limiting reaction of the entire process, is rather limited and needs better understanding for affecting the final yield positively. Therefore, the main aim of this thesis was the better characterization of the taxonomic community composition concentrating on the key players in cellulose degradation and the responsible enzymes.

Sequencing of our biogas plant sample generated nearly 900 million sequencing reads which were assembled into 123,435 contigs (> 1000 bp) and constitute about 1.16 Gb of assembled DNA. These benchmark data are the reason why it belongs to the biggest assembled data sets currently published. This substantial metagenomic assembly was used for the taxonomic and phylogenetic examination of the underlying community composition. Similarity search of single reads against the *NCBI non-redundant database* and phylogenetic placement via *MEGAN5* resulted in the classification of nearly eight million sequencing reads. About 5% of them were placed as unassigned reads and those reads might represent sequences that are not included in the database or comprise sequences of completely novel species. The NCBI non-redundant database represents a vast collection of various reference

sequences and this is the reason for the considerably low proportion of unclassified sequencing reads. Coupling the analysis with the *lowest common ancestor* (LCA) approach of MEGAN5 makes the evaluation more conservative but also prevents false-positive assignments. Taxonomic characterisation of assembled contigs by *AMPHORA2* marker-protein search enabled the classification of about 21,600 contigs. The reliance on marker-protein sequences limits assignment capacity because only contigs which contain those gene sequences can be classified accordingly. The very low proportion of unassigned contigs (0.3%) results partially from the assembly process as short and qualitatively poor reads are sorted out before and long continuous stretches increase the chance for obtaining full-length gene sequences. Informative content of a rRNA sequence analysis in metagenomic shotgun sequencing experiments is constrained as only a small fraction of all obtained reads contain fragments of ribosomal RNA genes by chance. Therefore, similarity search against the ribosomal RNA database of *SortMeRNA* resulted in dedication of roughly 1.6 million reads and about 2% of them remained unclassified. This percentage may refer to novel sequences that are not represented in the database so far. The rRNA-dependent approach was mainly a verification for the other two assignment methods. In general, the taxonomic composition analysis revealed that on average, the community consists of about 89% bacterial and only 4-10% archaeal members, depending on the approach used. For the rRNA marker-gene containing reads the percentage of classified bacterial sequences was even higher and added up to 97% and only 1% were classified as Archaea. Nevertheless, this results could emerge to some extent from the overrepresentation of bacterial reference sequences in databases as many of them originate from cultivation-based experiments and culture conditions are even more complex for Archaea, or in general because of a lack of endeavour for examination.

However, the severe underrepresentation of Archaea in this artificially generated environment was expected and goes in line with published literature [3]; [130]; [131]. By comparing different studies of other investigated biogas plants, the taxonomic structure appeared to be rather consistent and the composition of this sample confirms the impression, as no obvious differences were observed. *Firmicutes* and *Bacteroidetes* are bacterial phyla that comprise various species that were reported to be involved in the breakdown of cellulose and proteins, the acidogenesis and homoacetogenesis [3]; [132]. *Firmicutes* is the dominant bacterial phylum in manure-based systems, far more prevalent than *Bacteroidetes* [5]; [7]; [36]; [131]; [133];. Regardless of how the taxonomic community structure of our sample was investigated, on average a *Firmicutes*/*Bacteroidetes* ratio of 3.3:1 was observed. Similar values were calculated for the samples of searched published literature.

CD-HIT clustering of the AMPHORA2 marker-protein search on assembled contigs resulted in a distinct grouping of 195 bacterial and 5 archaeal OTUs on average, this is a major step forwards compared to previously published studies based on clustering of 16S rDNA sequencing [132]; [134]. However, the CD-HIT grouping solely outlines the number of different taxa present in the sample, but is not an indicator of quality or quantity for the comprised sequences. The composition and coverage based metagenomic binning of our biogas plant sample allowed the generation of 251 distinct bins with varying quality characteristics. 104 of them were extracted as high quality genome reconstructions where most of the bins are more than 90% complete, according to *CheckM* analysis. Some of the bins where taxonomically assigned down to species level with high confidence scores. This result is highly satisfactory as, to my knowledge, there is no other published study

available so far that showed this high degree of deep reconstruction potential for the microbial community in biogas fermenters.

The taxonomic assignment of these metagenomic bins basically reflects the impression of a considerable overrepresentation of *Firmicutes*, reasoned by a *Firmicutes*/*Bacteroidetes* ratio of 2.7:1. These observations led to a search for published literature about metagenomic studies dealing with bacterial composition in the guts of herbivores, which represent natural cellulose degrading systems. By calculating an average ratio of abundance for *Firmicutes* and *Bacteroidetes* in those studies, we observed that in digestion systems of herbivores the mean *Firmicutes*/*Bacteroidetes* ratio is almost 1:1 [9]; [11]; [12]; [21]; [22]; [135]. As most agricultural biogas plants are run with animal manures, it was expected that the *Bacteroidetes* are highly abundant in our sample because it was mainly fed with cow and chicken manure, besides maize-silage. It is possible that various species belonging to the phylum of *Bacteroidetes* are present in biogas fermenters at high levels initially but they may be outcompeted over time due to the operating conditions or a lack in essential factors that are usually present in their natural habitats. Summarising these discoveries, it is likely that in agricultural biogas fermenters, representatives belonging to the phylum of *Bacteroidetes* are not that abundant and do not compete equally as in their natural habitats or compared to *Firmicutes*.

To determine the transcription of distinct cellulolytic glycoside hydrolase families in the high quality bins, which represent individual organisms in the sampled biogas fermenter, RNA-Seq reads were mapped on binned contigs. The most observed GH families in the metagenomic bins can be subdivided into those with known main cellulolytic activity (*GH1*, *GH3*, *GH5*, *GH6*, *GH8*, *GH9*, *GH12*, *GH45*, *GH48*, *GH51* and *GH74*) and those with mainly hemicelluolytic activity (*GH30*) [136]. Across

genome bins, the GHs that showed the highest expression values belong to the groups of GH3, GH5 and GH51. All GH families seem to be higher expressed in *Firmicutes* compared to *Bacteroidetes*, except the cluster of GH51 shows higher transcription in species belonging to *Bacteroidetes*. The taxonomic assignment of glycoside hydrolase families in the individual genome bins confirms the trend towards a greater proportion of *Firmicutes* compared to *Bacteroidetes*; the ratio of selected GHs derived from *Firmicutes*/*Bacteroidetes* was 2.1:1. Furthermore, representatives of *Bacilli*, *Fibrobacteres* and *Actinobacteria* seem to contribute greatly to the overall abundance of selected GH groups. This investigation provides insight into the expression of glycoside hydrolase genes of distinct members in a biogas producing community and confirms the predominance of various *Firmicutes* in cellulolytic breakdown. As it was expected, different species belonging to the class of *Clostridia* showed the highest transcriptional levels of GHs.

This study provides evidence for the dominance of *Firmicutes* in agricultural biogas plants compared to various samples of guts or feces of herbivores where the *Bacteroidetes* seem to be equally abundant. Observation is corroborated by the finding that selected GH families are twice as often affiliated with *Firmicutes* than with *Bacteroidetes*. That indicates that *Firmicutes*, and especially *Clostridia*, are the ones mainly responsible for the initial degradation of plant biomass in agricultural biogas plants. Therefore, one can speculate that a shift in population from dominating abundance of *Firmicutes to Bacteroidetes* might increase the overall hydrolytic performance of biogas fermenters. This would require altered process parameters and conditions which favour the growth or survival of *Bacteroidetes*. For this, further research is needed to get a better understanding about the specific requirements of this phylum on its environment or surrounding community for efficient adjustment of running parameters in agricultural biogas fermenters.

Additionally, the reconstruction of the 104 high quality genomes will provide new possibilities for studying functional diversity and capacities in agricultural one-stage biogas fermenters. As knowledge about individual genomes participating in the degradation of cellulose and other plant material is rather limited so far, this data will provide a good starting point for future projects.

5. Supplementary Material

Table S1: Completeness, contamination, strain heterogeneity, Mb, N50 and predicted numbers of distinct rRNAs for the 20 representatives belonging to the "good bins" category. Quality measurements were obtained via CheckM analysis and rRNAs were predicted by the use of ConsPred. Compl = completeness; cont = contamination; het = heterogeneity

		good bins							
Bin-ID	Taxonomy (consensus score > 0.8)	compl	cont	het	5S rRNAs	16S rRNAs	23S rRNAs	Mb	N50 for contigs
pb121	Fibrobacter succinogenes	98.84	4.40	100.00	1	1	1	2.645	19,300
pb122	Fibrobacter succinogenes	97.74	2.85	75.00	0	0	0	2.109	11,855
pb172	Bacteroidetes	96.24	2.07	75.00	1	0	0	2.501	19,077
pb190	Clostridia	98.28	6.74	91.30	0	0	0	2.230	16,089
pb192-1	Clostridia	96.77	1.61	50.00	2	1	1	2.250	129,429
pb212	Bacteria	98.31	6.78	100.00	2	1	1	2.485	22,750
pb215	Ruminiclostridium	100.00	8.77	100.00	3	1	0	3.177	45,291
pb31	Treponema	99.30	7.23	100.00	1	1	1	2.150	16,062
pb35-2	Lachnoclostridium phytofermentans	98.30	8.58	92.59	0	0	0	2.710	28,325
pb3	Clostridia	96.61	2.97	60.00	3	1	1	2.357	33,655
pb40	Clostridia	95.11	0.13	100.00	1	0	0	1.469	12,735
pb6-1	Bacteria	96.00	1.33	100.00	2	0	0	1.750	105,819
pb65	Clostridiales	98.60	7.69	100.00	0	0	0	2.314	66,410
pb69	Porphyromonadaceae	96.24	3.73	90.91	1	0	0	1.741	17,355
pb76-1	Ruminococcaceae	96.55	8.62	100.00	0	0	0	1.767	21,374
pb80	Firmicutes	95.10	2.97	50.00	4	0	0	2.648	77,072
pb85	Methanosarcina barkeri	99.84	0.03	100.00	0	1	1	3.431	70,311
pb88	Bacteroidales	96.24	1.08	100.00	2	0	0	3.549	17,475
pb90	Treponema	99.30	9.91	96.30	0	0	0	2.444	22,018
pb97	Ruminococcaceae	97.09	1.43	90.00	3	0	0	2.143	36,708

Table S2: Completeness, contamination, strain heterogeneity, Mb content, N50 and predicted numbers of distinct rRNAs for the 20 representatives belonging to the "nearly complete genome drafts" category. Quality measurements were obtained via CheckM analysis and rRNAs were predicted by the use of ConsPred. Compl = completeness; cont = contamination; het = heterogeneity

nearly complete genome drafts									
Bin-ID	Taxonomy (consensus score > 0.8)	compl	cont	het	5S rRNAs	16S rRNAs	23S rRNAs	Mb	N50 for contigs
107	Clostridiales	91.03	1.34	0.00	2	0	0	2.012	67,725
121	Clostridia	93.22	0.00	0.00	1	0	0	2.043	51,306
125	Bacteroidetes	96.24	1.34	0.00	1	0	0	2.340	20,117
145	Paludibacter propionicigenes	98.92	0.54	0.00	2	1	0	2.932	88,660
159	Clostridiales	97.90	3.85	14.29	2	0	0	2.319	56,632
184	Verrucomicrobia	93.92	3.58	33.33	1	0	0	4.152	13,165
18	Mageeibacillus indolicus	94.44	1.77	33.33	2	0	0	2.121	15,834
45	Bacteroidetes	95.59	0.81	33.33	0	0	0	1.986	17,036
90	Ruminococcaceae	95.45	0.00	0.00	2	1	1	2.004	166,983
96	Ruminiclostridium thermocellum	97.65	2.57	16.67	1	0	0	3.255	16,777
pb186-2	Lachnoclostridium phytofermentans	90.72	2.06	100.00	0	0	0	2.106	39,480
pb205	Clostridiales	94.38	4.20	85.71	1	0	0	2.087	55,007
pb233-1	Ruminococcaceae	93.96	1.68	75.00	1	0	1	2.000	46,118
pb235-2	Clostridiales	93.17	4.79	50.00	2	0	0	2.018	17,760
pb237	Clostridiales	90.03	2.82	50.00	1	0	0	1.655	8,072
pb35-1	Lachnoclostridium phytofermentans	91.05	1.34	66.67	0	0	0	2.288	51,148
pb47	Verrucomicrobia	90.03	3.41	57.14	1	1	1	2.785	22,427
pb60-2	Firmicutes	92.58	2.54	50.00	6	0	0	2.729	106,815
pb61-1	Ruminococcaceae	93.10	0.67	100.00	0	0	0	1.733	9,928
pb84	Oceanobacillus iheyensis	92.88	1.90	71.43	1	0	0	2.351	14,552

Table S3: Completeness, contamination, strain heterogeneity, Mb content, N50 and predicted numbers of distinct rRNAs for the 37 representatives belonging to the"nearly complete pangenome" category. Quality measurements were obtained via CheckM analysis and rRNAs were predicted by the use of ConsPred. Compl = completeness; cont = contamination; het = heterogeneity

nearly complete pangenome drafts									
Bin-ID	**Taxonomy (Consensus Score > 0.8)**	**compl**	**cont**	**het**	**5S rRNAs**	**16S rRNAs**	**23S rRNAs**	**Mb**	**N50 for contigs**
114	Clostridia	90.80	19.75	83.72	1	0	0	2.456	14,476
118	Bacteroidetes	96.55	8.62	22.22	0	0	0	1.717	7,498
120	Bacteroidales	98.57	5.86	20.00	2	0	0	3.371	107,165
122	Firmicutes	98.28	15.52	97.06	4	0	0	2.756	98,798
128	Clostridiales	98.28	38.09	97.30	3	1	1	2.304	47,314
131	Clostridiales	98.39	27.90	95.71	2	0	1	2.771	42,529
135	Bacteroidales	92.24	28.62	76.67	2	0	0	3.713	8,219
137	Ruminococcaceae	96.64	17.24	97.22	2	1	1	2.021	32,847
138	Bacteroidales	100.00	50.71	100.00	1	0	0	3.221	14,313
142	Bacteroidales	98.28	36.39	98.25	2	1	1	3.603	19,789
149	Bacteroidales	100.00	91.44	94.23	0	0	0	5.381	11,252
162	Clostridia	94.23	13.46	72.22	2	0	0	2.647	74,664
165	Pelotomaculum thermopropionicum	94.94	11.18	87.88	2	1	1	2.864	56,541
167	Clostridiales	99.30	16.08	92.59	2	1	1	2.243	50,875
16	Clostridia	99.19	42.97	83.87	1	0	0	3.191	8,807
172	Clostridiales	93.71	11.89	100.00	0	0	0	2.446	58,832
179	Micrococcales	100.00	61.99	52.54	1	0	0	5.944	12,829
17	Bacteroidales	100.00	98.78	91.82	4	1	1	6.802	50,520
180	Bacteria	98.67	31.39	90.62	0	1	0	2.400	45,126
188	Bacteroidales	94.05	19.03	72.73	3	0	1	2.278	40,218
200	Corynebacterium	100.00	56.62	72.31	3	1	1	4.101	17,663
201	Paludibacter propionicigenes	93.10	23.28	93.18	3	1	0	4.195	17,714
204	Mageeibacillus indolicus	96.34	13.65	76.67	3	1	1	2.387	33,234
23	Bacteria	96.70	23.08	92.31	3	0	0	3.171	74,083
36	Bacteroidales	100.00	25.86	100.00	1	1	1	3.075	48,908
41	Bacteria	98.56	12.00	100.00	2	0	0	1.999	64,174
43	Clostridiales	99.07	38.58	100.00	4	0	0	3.220	95,854
48	Methanobacterium	99.07	19.63	100.00	0	1	1	2.241	47,585

Table S4: Completeness, contamination, strain heterogeneity, Mb content, N50 and predicted numbers of distinct rRNAs for the 27 representatives belonging to the "incomplete genome drafts" category. Quality measurements were obtained via CheckM analysis and rRNAs were predicted by the use of ConsPred. Compl = completeness; cont = contamination; het = heterogeneity

incomplete genome drafts									
Bin-ID	**Taxonomy (Consensus Score > 0.8)**	**compl**	**cont**	**het**	**5S rRNAs**	**16S rRNAs**	**23S rRNAs**	**Mb**	**N50 for contigs**
109	Erysipelothrix rhusiopathiae	79.90	4.04	55.56	0	0	0	1.120	5,151
141	Deltaproteobacteria	73.99	2.58	50.00	1	0	0	2.545	5,013
147	Clostridia	90.25	6.79	57.14	0	1	0	2.117	42,492
175	Sphaerochaeta	64.12	2.84	100.00	0	0	0	1.015	2,321
178	Bacteroidales	80.46	1.40	42.86	0	0	0	1.499	8,230
1	Clostridiales	73.48	2.62	75.00	0	0	0	1.283	4,246
20	Ruminococcaceae	64.22	2.35	50.00	0	0	0	1.547	2,173
2	Ruminococcaceae	63.64	4.80	69.23	0	0	0	2.007	1,864
33	Euryarchaeota	71.64	2.26	66.67	0	0	0	1.172	2,027
3	Clostridiales	61.88	5.44	54.55	0	0	0	1.007	2,137
49	Halanaerobiaceae	72.10	5.26	62.50	0	0	0	1.233	1,959
72	Acidobacteria	68.80	5.41	62.50	0	0	0	2.193	2,569
75	Clostridia	71.08	3.04	75.00	0	0	0	1.380	2,916
pb108-3	Clostridia	88.35	5.23	55.56	2	0	0	2.207	37,524
pb162	Clostridiales	88.36	7.07	70.00	0	0	0	1.811	3,318
pb173	Bacteria	87.77	4.39	100.00	1	0	0	0.826	12,005
pb177	Clostridia	88.45	3.06	60.00	0	0	0	2.110	5,461
pb185	Mageeibacillus indolicus	84.97	4.18	80.00	0	0	0	1.728	4,443
pb206-2	Bacteroidales	81.97	5.41	72.22	1	0	0	2.138	11,327
pb207	Bacteroidales	87.50	0.50	50.00	0	0	1	2.513	11,906
pb233-3	Mageeibacillus indolicus	84.14	4.49	80.00	1	0	0	1.413	11,939
pb235-1	Lachnoclostridium phytofermentans	86.35	4.36	70.00	0	0	0	2.336	71,225
pb243	Sphaerochaeta globosa	89.14	3.51	50.00	1	1	0	2.397	12,472
pb246	Ruminococcaceae	85.81	5.09	78.57	1	0	0	1.580	3,920
pb2	Bacteroidales	82.22	5.20	58.33	0	0	0	2.148	3,435
pb30	Planctomycetaceae	87.12	0.99	80.00	1	0	1	3.035	7,797
pb6-2	Bacteria	88.55	1.82	60.00	0	0	0	1.390	8,194

Table S5: Phenotype predictions for bins 1 – 26.

Bin ID	Taxa	aerobe	anaerobe	facultative anaerobe	gram-negative	haloophilic	motile	phototrophs	thermophilic	intracelluar	facultative intracellular	obligate intracellular	methanotrophs	ammonium oxidizers	archaeal ammonium ox.	bacterial ammonium ox.	nitrite oxidizers	nitrifiers
1	Clostridiales	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	Ruminococcaceae	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	Clostridiales	-	+	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
5	Ruminococcaceae	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
16	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
17	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
18	Mageeibacillus indolicus	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	Ruminococcaceae	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	Bacteria	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
33	Euryarchaeota	-	+	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-
36	Bacteroidales	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
41	Bacteria	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
43	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
45	Bacteroidetes	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
48	Methanobacterium	-	+	-	+	-	+	-	-	+	-	+	-	-	-	-	-	-
49	Halanaerobiaceae	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
72	Acidobacteria	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
75	Clostridia	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
82	Clostridiales	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
90	Ruminococcaceae	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
96	Ruminiclostridium thermocellum	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
107	Clostridiales	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
109	Erysipelothrix rhusiopathiae	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
114	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
118	Bacteroidetes	-	+	-	+	-	-	-	-	+	-	+	-	-	-	-	-	-
120	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S6: Phenotype predictions for bins 27 – 52.

Bin ID	Taxa	aerobe	anaerobe	fac. an.	gram-neg.	halo.	mot.	phototr.	thermoph.	intracelluar	fac. intracell.	ob. intracell.	methanotr.	am. ox.	arch. am. ox.	bact. am. ox.	nitrite ox.	nitrif.
121	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
122	Firmicutes	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
125	Bacteroidetes	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
128	Clostridiales	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
131	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
135	Bacteroidales	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
137	Ruminococcaceae	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
138	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
141	Deltaproteobacteria	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
142	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
145	Paludibacter propionicigenes	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
147	Clostridia	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
149	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
159	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
162	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
165	Pelotomaculum thermopropionicum	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
167	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
172	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
175	Sphaerochaeta	-	-	-	+	-	+	-	-	+	-	+	-	-	-	-	-	-
178	Bacteroidales	-	+	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-
179	Micrococcales	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
180	Bacteria	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
184	Verrucomicrobia	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
188	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
200	Corynebacterium	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
201	Paludibacter propionicigenes	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S7: Phenotype predictions for bins 53 – 78.

Bin ID	Taxa	aerobe	anaerobe	fac. an.	gram-neg.	halo.	mot.	phototr.	thermoph.	intracell.	fac. intracell.	ob. intracell.	methanotr.	am. ox.	arch. am. ox.	bact.am. ox.	nitrite ox.	nitrif.
204	Mageeibacillus indolicus	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
pb108-3	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb121	Fibrobacter succinogenes	-	-	-	+	-	+	-	-	+	-	+	-	-	-	-	-	-
pb122	Fibrobacter succinogenes	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb15-1	Bacteria	-	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-
pb162	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb166-2	Bacteroidetes	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb172	Bacteroidetes	-	+	-	+	-	+	-	-	+	-	+	-	-	-	-	-	-
pb173	Bacteria	-	-	-	+	-	-	-	-	+	-	+	-	-	-	-	-	-
pb177	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb185	Mageeibacillus indolicus	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
pb186-2	Lachnoclostridium phytofermentans	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb190	Clostridia	-	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-
pb192-1	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb199	Bacteria	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
pb2	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb205	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb206-2	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb207	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb212	Bacteria	-	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-
pb215	Ruminiclostridium	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb233-1	Ruminococcaceae	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb233-3	Mageeibacillus indolicus	-	+	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
pb235-1	Lachnoclostridium phytofermentans	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb235-2	Clostridiales	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb237	Clostridiales	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S8: Phenotype predictions for bins 79 - 104.

Bin ID	Taxa	aerobe	anaerobe	fac. an.	gram-neg.	halo.	mot.	phototr.	thermoph.	intracell.	fac. intracell.	ob. intracell.	methanotr.	am. ox.	arch. am. ox.	bact. am. ox.	nitrite ox.	nitrif.
pb243	Sphaerochaeta globosa	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb246	Ruminococcaceae	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb3	Clostridia	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb30	Planctomycetaceae	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
pb31	Treponema	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
pb35-1	Lachnoclostridium phytofermentans	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb35-2	Lachnoclostridium phytofermentans	-	+	-	-	-	+	-	-	+	-	+	-	-	-	-	-	-
pb38-1	Acholeplasmataceae	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
pb40	Clostridia	-	+	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
pb47	Verrucomicrobia	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb60-2	Firmicutes	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb6-1	Bacteria	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
pb61-1	Ruminococcaceae	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
pb6-2	Bacteria	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
pb65	Clostridiales	-	+	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
pb69	Porphyromonadaceae	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb70	Alkaliphilus oremlandii	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
pb76-1	Ruminococcaceae	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb78-1	Bacteria|Firmicutes	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb80	Firmicutes	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb84	Oceanobacillus iheyensis	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
pb85	Methanosarcina barkeri	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-
pb88	Bacteroidales	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
pb90	Treponema	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
pb93-2	Syntrophomonadaceae	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
pb97	Ruminococcaceae	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S9: Taxonomic assignment and predicted number of occurrence for the 15 most important GH enzyme families in bins 1 – 52.

Taxon	GH1	GH3	GH5	GH6	GH8	GH9	GH12	GH14	GH30	GH44	GH45	GH48	GH51	GH74	GH94
107_Clostridiales	3	3	1	0	1	0	0	0	1	0	0	0	0	1	0
109_Erysipelothrix	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0
114_Clostridia	1	4	6	0	1	0	0	0	4	0	0	0	0	3	0
118_Bacteroidetes	3	3	0	0	0	0	1	0	0	0	0	0	0	0	0
120_Bacteroidales	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0
121_Clostridia	1	4	2	1	1	1	0	0	0	1	0	0	0	0	0
122_Firmicutes	6	6	9	0	0	0	0	0	2	4	0	2	7	0	0
125_Bacteroidetes	3	3	0	0	0	0	1	0	0	0	0	0	0	0	1
128_Clostridiales	1	2	0	0	0	0	0	1	0	0	0	0	0	0	2
131_Clostridiales	3	5	11	0	2	2	0	0	0	2	0	4	4	3	4
135_Bacteroidales	9	22	24	0	1	5	0	1	4	2	0	5	12	3	1
137_Ruminococcaceae	2	1	1	0	0	0	0	0	0	0	0	0	0	0	1
138_Bacteroidales	3	20	19	0	1	7	0	0	10	0	0	7	19	2	0
141_Deltaproteobacteria	0	3	0	0	0	0	0	2	0	0	0	0	1	0	0
142_Bacteroidales	8	18	12	0	0	0	0	0	10	5	0	1	21	0	0
145_Paludibacter	5	4	1	0	0	0	1	2	1	0	0	0	0	1	0
147_Clostridia	1	1	2	0	0	0	0	0	0	0	0	0	0	1	0
149_Bacteroidales	16	33	18	0	1	5	1	1	9	7	0	1	26	4	0
159_Clostridiales	2	12	6	1	2	0	1	0	3	0	0	2	2	1	1
162_Clostridia	8	7	14	0	0	0	0	0	1	4	0	5	13	0	0
165_Pelotomaculum	2	5	16	0	2	0	0	0	8	0	0	0	1	15	0
167_Clostridiales	2	1	0	0	0	0	0	0	0	0	0	0	0	0	3
16_Clostridia	1	5	9	0	3	0	0	0	3	0	0	0	0	3	3
172_Clostridiales	2	8	5	1	1	0	1	0	1	0	0	1	4	0	6
175_Sphaerochaeta	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0
178_Bacteroidales	4	6	0	0	0	1	1	1	0	0	0	0	0	0	0
179_Micrococcales	8	9	5	0	0	0	0	2	0	0	0	0	0	0	0
17_Bacteroidales	12	24	13	1	0	3	3	1	5	1	0	2	16	2	0
180_Bacteria	2	14	3	0	1	1	0	0	0	0	0	0	0	0	1
184_Verrucomicrobia	4	8	12	0	0	0	0	0	2	1	0	0	2	0	0
188_Bacteroidales	3	8	4	1	0	4	1	2	0	1	0	0	0	0	0
18_Mageeibacillus	4	2	8	1	1	0	1	0	0	0	0	2	6	0	3
1_Clostridiales	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
200_Corynebacterium	5	7	3	0	0	0	0	0	0	0	0	0	0	0	0
201_Paludibacter	5	13	21	0	2	2	1	0	11	3	0	2	17	0	0
204_Mageeibacillus	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20_Ruminococcaceae	1	2	0	0	0	0	0	2	0	0	0	0	0	0	0
23_Bacteria	1	2	4	0	0	0	1	0	0	0	0	0	0	0	0
2_Ruminococcaceae	7	3	3	0	0	0	0	0	0	0	0	0	3	2	0
33_Euryarchaeota	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36_Bacteroidales	5	15	9	0	0	4	0	1	1	2	0	1	9	1	2
3_Clostridiales	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
41_Bacteria	2	10	0	0	0	0	0	0	2	0	0	0	3	0	1
43_Clostridiales	2	4	7	0	2	0	0	0	3	0	0	0	0	5	2
45_Bacteroidetes	4	3	1	0	1	1	1	0	1	0	0	0	0	1	0
48_Methanobacterium	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
49_Halanaerobiaceae	2	2	3	0	1	0	0	0	1	0	0	2	3	0	2
5_Ruminococcaceae	3	6	5	0	1	0	0	0	5	0	0	2	13	0	1
72_Acidobacteria	3	13	0	0	1	0	0	0	2	0	0	0	0	0	0
75_Clostridia	2	1	0	0	2	0	0	0	0	0	0	0	0	0	4
82_Clostridiales	9	21	8	0	6	2	0	0	6	3	0	2	8	7	4
90_Ruminococcaceae	4	6	4	0	2	0	0	0	3	0	0	0	3	0	0

Table S10: Taxonomic assignment and predicted number of occurrence for the 15 most important GH enzyme families in bins 53 – 104.

Taxon	GH1	GH3	GH5	GH6	GH8	GH9	GH12	GH14	GH30	GH44	GH45	GH48	GH51	GH74	GH94
96_Ruminiclostridium	4	9	46	7	7	28	2	0	19	21	0	20	3	16	3
pb108-3_Clostridia	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0
pb121_Fibrobacter	5	6	33	6	4	7	2	0	15	6	3	2	11	2	1
pb122_Fibrobacter	4	5	15	3	4	7	1	0	3	8	1	2	6	1	1
pb15-1_Bacteria	2	5	3	0	0	0	0	0	0	0	0	0	0	0	2
pb162_Clostridiales	2	9	5	0	0	0	0	0	1	0	0	1	1	1	0
pb166-2_Bacteroidetes	3	2	1	0	1	0	0	0	0	0	0	0	0	4	0
pb172_Bacteroidetes	4	5	2	0	2	2	1	0	0	0	0	0	0	0	0
pb173_Bacteria	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pb177_Clostridia	1	11	2	0	2	0	0	0	2	0	0	0	0	3	0
pb185_Mageeibacillus	2	1	1	0	0	0	0	1	0	0	0	0	0	0	4
pb186-2_Lachnoclostridium	5	3	7	1	0	0	0	0	0	1	0	1	2	3	2
pb190_Clostridia	5	5	15	0	2	2	1	0	4	1	0	4	2	2	2
pb192-1_Clostridia	3	7	4	0	0	0	0	0	0	0	0	1	13	0	1
pb199_Bacteria	5	7	26	3	12	10	1	1	11	12	3	6	14	3	3
pb205_Clostridiales	4	6	0	0	1	0	0	0	3	0	0	0	1	2	1
pb206-2_Bacteroidales	4	9	9	0	0	4	2	0	0	0	0	4	7	2	0
pb207_Bacteroidales	3	15	7	1	0	4	1	4	1	3	0	2	2	2	1
pb212_Bacteria	3	1	0	0	0	0	0	0	1	0	0	0	1	0	1
pb215_Ruminiclostridium	3	10	42	8	7	27	3	0	20	23	0	20	4	11	3
pb233-1_Ruminococcaceae	1	6	2	0	0	0	0	0	0	0	0	0	4	0	1
pb233-3_Mageeibacillus	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
pb235-1_Lachnoclostridium	4	11	9	2	1	5	1	0	2	3	0	4	9	3	3
pb235-2_Clostridiales	2	4	1	0	0	0	0	0	1	0	0	0	1	2	0
pb237_Clostridiales	2	1	0	0	2	0	0	0	0	0	0	0	0	0	2
pb243_Sphaerochaeta	2	3	0	0	0	0	0	0	0	0	0	0	1	0	0
pb246_Ruminococcaceae	2	1	0	0	0	0	0	0	0	0	0	0	0	0	3
pb2_Bacteroidales	6	6	5	0	1	2	0	0	2	3	0	1	5	0	1
pb30_Planctomycetaceae	1	5	5	0	0	0	0	0	2	0	0	0	4	0	0
pb31_Treponema	3	3	1	1	0	0	1	0	0	0	0	0	1	0	2
pb35-1_Lachnoclostridium	3	7	19	3	2	8	2	0	4	6	0	11	4	5	5
pb35-2_Lachnoclostridium	4	4	9	0	0	0	0	0	0	2	0	0	10	4	2
pb38-1_Acholeplasmataceae	2	4	0	0	0	0	0	0	0	0	0	0	0	0	1
pb3_Clostridia	2	5	6	0	5	0	0	0	5	0	0	0	0	5	0
pb40_Clostridia	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
pb47_Verrucomicrobia	3	1	1	0	0	0	0	1	0	0	0	0	1	0	0
pb6-1_Bacteria	2	6	3	0	0	1	0	0	1	0	0	0	0	0	1
pb6-2_Bacteria	1	6	1	0	0	0	0	0	1	0	0	0	0	1	5
pb60-2_Firmicutes	8	12	14	0	2	0	0	0	1	6	0	5	15	1	1
pb61-1_Ruminococcaceae	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0
pb65_Clostridiales	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0
pb69_Porphyromonadaceae	3	4	2	0	0	1	1	0	0	0	0	0	0	0	0
pb70_Alkaliphilus	1	4	6	0	4	0	0	0	5	0	0	0	0	7	1
pb76-1_Ruminococcaceae	2	2	1	0	1	0	0	0	2	0	0	0	0	3	2
pb78-1_Firmicutes	3	5	10	0	1	0	0	0	0	0	0	6	6	1	0
pb80_Firmicutes	1	5	1	0	0	0	0	0	0	0	0	0	0	0	0
pb84_Oceanobacillus	4	2	5	0	1	1	0	1	0	0	0	3	6	1	0
pb85_Methanosarcina	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0
pb88_Bacteroidales	4	4	0	0	1	0	1	0	0	0	0	0	0	0	0
pb90_Treponema	2	7	8	3	0	2	4	0	1	2	0	1	3	1	5
pb93-2_Syntrophomonadaceae	1	1	9	0	1	0	0	0	1	0	0	0	0	4	1
pb97_Ruminococcaceae	3	4	2	0	2	0	0	0	1	0	0	0	1	0	1

Script 1: Removing adapter sequnces from fastq file

```
#!/usr/bin/python

import sys, gzip
import collections
from Bio.SeqIO.QualityIO import FastqGeneralIterator

if len(sys.argv) != 3:
  print >> sys.stderr, "script <adapters.fa> <sequences.fa>"
  sys.exit(1)

infile1 = sys.argv[1]
infile2 = sys.argv[2]

from Bio import SeqIO
adapters = collections.OrderedDict()
for seq_record in SeqIO.parse(infile1, "fasta"):
  adapters[seq_record] = 0

myseqs = ()
counter = 0
for title, seq, qual in FastqGeneralIterator(gzip.open(infile2)):
  counter += 1
  if counter == 500000: break
  for adapter in adapters:
    if str(adapter.seq) in seq:
      adapters[adapter] += 1
      #print >> sys.stdout, "%s\n%s" % (myseq.id, myseq.seq)

for key, item in adapters.items():
  print >> sys.stderr, "%s: %s" % (key.id, item)
```

Adapter sequences checked

>Universal

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

>FirstPartOfIndexed

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

>Indexed1

GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG

>Indexed2

GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG

>Indexed3

GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG

>Indexed4

GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG

>Indexed5

GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG

>Indexed6

GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG

>Indexed7

GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG

>Indexed8

GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG

>Indexed9

GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG

>Indexed10

GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG

>Indexed11

GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG

>Indexed12

GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG

>FirstPartOfUniversalR

AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

>UniversalR

AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

>FirstPartOfReverseIndexed

CAAGCAGAAGACGGCATACGAGAT

>Indexed1R

CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

```
>Indexed2R
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed3R
CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed4R
CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed5R
CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed6R
CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed7R
CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed8R
CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed9R
CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed10R
CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed11R
CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
>Indexed12R
CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
```

Script 2: Calculating mean coverage by contig

```
#!/usr/bin/env python

import sys

sum_by_contig={}
num_by_contig={}

for line in sys.stdin:
  parts=line.strip().split("\t")
  contig = parts[0]
  cov = int(parts[1])
  num = int(parts[2])
  #print contig, cov, num; sys.exit(1)
```

```
  if not sum_by_contig.has_key(contig):
    sum_by_contig[contig]=0
    num_by_contig[contig]=0

  sum_by_contig[contig]=sum_by_contig[contig]+num*cov
  num_by_contig[contig]=num_by_contig[contig]+num

contignames=sum_by_contig.keys()
contignames.sort()

for contig in contignames:
  if contig == "genome":
    continue

  average_cov=float(sum_by_contig[contig])/float(num_by_contig[contig])
  sys.stdout.write("%s\t%1.1f\n" % (contig, average_cov))
```

Script 3: Filter single assembly

```
#!/usr/bin/env python

import sys
from Bio import SeqIO

assemblyfilename=sys.argv[1]
coveragefilename=sys.argv[2]
minlength=int(sys.argv[3])
mincoverage=float(sys.argv[4])
assemblyoutfilename=sys.argv[5]
coverageoutfilename=sys.argv[6]

coverage={}
with open(coveragefilename) as infile:
  for line in infile:
    parts = line[:-1].split("\t")
    name = parts[0].split()[0]
    c = float(parts[1])
```

```
    if c >= mincoverage:
      coverage[name]=c

lengths=[]
entries={}
with open(assemblyfilename) as infile:
  for entry in SeqIO.parse(infile, "fasta"):
    #import pdb; pdb.set_trace()
    if coverage.has_key(entry.id):
      if len(entry.seq) >=minlength:
        entries[entry.id]=entry
        lengths.append((len(entry.seq), entry.id))

lengths.sort()
lengths.reverse()

id=0
with open(assemblyoutfilename, "w") as seqoutfile:
  with open(coverageoutfilename, "w") as coutfile:
      coutfile.write('"Name","Average coverage","Reference length"\n')

      for (length, name) in lengths:

        print >> sys.stdout, name

        id += 1
        entry=entries[name]
        c=coverage[name]

        entry.id = str(id)
        SeqIO.write(entry, seqoutfile, "fasta")

        coutfile.write('"%i","%1.1f","%i"\n' % (id, c, length))
```

Script 4: Filter Combined Assembly

```
#!/usr/bin/env python

import sys
from Bio import SeqIO

assemblyfilename=sys.argv[1]
coverage1filename=sys.argv[2]
coverage2filename=sys.argv[3]
minlength=int(sys.argv[4])
mincoverage=float(sys.argv[5])
assemblyoutfilename=sys.argv[6]
coverage1outfilename=sys.argv[7]
coverage2outfilename=sys.argv[8]

coverage1={}
with open(coverage1filename) as infile:
  for line in infile:
    parts = line[:-1].split("\t")
    name = parts[0].split()[0]
    coverage = float(parts[1])
    #if coverage >= mincoverage:
    coverage1[name]=coverage

coverage2={}
with open(coverage2filename) as infile:
  for line in infile:
    parts = line[:-1].split("\t")
    name = parts[0].split()[0]
    coverage = float(parts[1])
    #if coverage1.has_key(name):
    #if coverage >= mincoverage:
    coverage2[name]=coverage

lengths=[]
entries={}
with open(assemblyfilename) as infile:
```

```
  for entry in SeqIO.parse(infile, "fasta"):
    #if coverage1.has_key(entry.id) or coverage2.has_key(entry.id):
    try:
      coverage1[entry.id]
    except KeyError:
      coverage1[entry.id] = 0
    try:
      coverage2[entry.id]
    except KeyError:
      coverage2[entry.id] = 0
    if len(entry.seq) >=minlength and (coverage1[entry.id] >= mincoverage
or coverage2[entry.id] >= mincoverage) :
      entries[entry.id]=entry
      lengths.append((len(entry.seq), entry.id))


lengths.sort()
lengths.reverse()


id=0
with open(assemblyoutfilename, "w") as seqoutfile:
  with open(coverage1outfilename, "w") as c1outfile:
    with open(coverage2outfilename, "w") as c2outfile:
      c1outfile.write('"Name","Average coverage","Reference length"\n')
      c2outfile.write('"Name","Average coverage","Reference length"\n')


      for (length, name) in lengths:


        print >> sys.stdout, name


        id += 1


        entry=entries[name]
        c1=coverage1[name]
        c2=coverage2[name]


        entry.id = str(id)
        SeqIO.write(entry, seqoutfile, "fasta")
```

```
        c1outfile.write('"%i","%1.1f","%i"\n' % (id, c1, length))
        c2outfile.write('"%i","%1.1f","%i"\n' % (id, c2, length))
```

Script 5: Filter Blast results to minimal bitscore

```
#! /usr/bin/python

import sys, os, argparse

minscore = float(sys.argv[1])

mydict = {}
previous = None

for line in sys.stdin:

  elements = line.strip().split('\t')
  assert len(elements) == 12
  bitscore = float(elements[11])

  if bitscore >= minscore:
    sys.stdout.write(line)
```

Script 6: Contig2Bin

```
#!/usr/bin/python

import sys, os
from Bio import SeqIO

fastafiles = sys.argv[1:]

for fastafile in fastafiles:
  bin = fastafile.split('/')[-1].split('.')[0]

  for seq in SeqIO.parse(fastafile, "fasta"):
    id = seq.description.split()[1]
```

```
    print >> sys.stdout, '%s\t%s' % (id, bin)
```

Script 7: Combine Tables

```
#!/usr/bin/python

import sys, os

contig2taxon = {}
with open(sys.argv[1]) as fin:
  for line in fin:
    contig, taxon = line.strip().split('\t')
    contig2taxon[contig] = taxon

with open(sys.argv[2]) as fin:
  for i, line in enumerate(fin):
    if i == 0:
      print >> sys.stdout, line.strip()
      continue
    els = line.strip().split()
    try: taxon = contig2taxon[els[0]]
    except KeyError: taxon = '-'
    els[-2] = taxon
    print >> sys.stdout, '\t'.join(els)
```

6. List of Tables

7. List of Figures

8. References

1. Weiland P. Biomass Digestion in Agriculture: A Successful Pathway for the Energy Production and Waste Treatment in Germany. Engineering in Life Sciences. 2006;6: 302–309. doi:10.1002/elsc.200620128

2. Wirth R, Kovács E, Maróti G, Bagi Z, Rákhely G, Kovács KL. Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. Biotechnology for Biofuels. 2012;5: 41. doi:10.1186/1754-6834-5-41

3. Stolze Y, Zakrzewski M, Maus I, Eikmeyer F, Jaenicke S, Rottmann N, et al. Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation conditions. Biotechnology for Biofuels. 2015;8: 14. doi:10.1186/s13068-014-0193-8

4. Zakrzewski M, Goesmann A, Jaenicke S, Jünemann S, Eikmeyer F, Szczepanowski R, et al. Profiling of the metabolically active community from a production-scale biogas plant by means of high-throughput metatranscriptome sequencing. Journal of Biotechnology. 2012;158: 248–258. doi:10.1016/j.jbiotec.2012.01.020

5. Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann K-H, et al. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. Journal of Biotechnology. 2008;136: 77–90. doi:10.1016/j.jbiotec.2008.05.008

6. Li A, Chu Y, Wang X, Ren L, Yu J, Liu X, et al. A pyrosequencing-based metagenomic study of methane-producing microbial community in solid-state biogas reactor. Biotechnology for Biofuels. 2013;6: 1. doi:10.1186/1754-6834-6-3

7. Solli L, Håvelsrud OE, Horn SJ, Rike AG. A metagenomic study of the microbial communities in four parallel biogas reactors. Biotechnology for Biofuels. 2014;7. doi:10.1186/s13068-014-0146-2

8. Zverlov VV, Köck DE, Schwarz WH. The Role of Cellulose-Hydrolyzing Bacteria in the Production of Biogas from Plant Biomass [Internet]. Springer Berlin Heidelberg; 2015. Available: http://link.springer.com/chapter/10.1007/978-3-662-45209-7_12

9. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, et al. Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities. Bertilsson S, editor. PLoS ONE. 2013;8: e74787. doi:10.1371/journal.pone.0074787

10. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nature Reviews Microbiology. 2008;6: 121–131. doi:10.1038/nrmicro1817

11. Ilmberger N, Güllert S, Dannenberg J, Rabausch U, Torres J, Wemheuer B, et al. A Comparative Metagenome Survey of the Fecal Microbiota of a Breast- and a Plant-Fed Asian Elephant Reveals an Unexpectedly High Diversity of Glycoside Hydrolase Family Enzymes. Bereswill S, editor. PLoS ONE. 2014;9: e106707. doi:10.1371/journal.pone.0106707

12. Zeng B, Han S, Wang P, Wen B, Jian W, Guo W, et al. The bacterial communities associated with fecal types and body weight of rex rabbits. Scientific Reports. 2015;5: 9342. doi:10.1038/srep09342

13. Mao C, Feng Y, Wang X, Ren G. Review on research achievements of biogas from anaerobic digestion. Renewable and Sustainable Energy Reviews. 2015;45: 540–555. doi:10.1016/j.rser.2015.02.032

14. Cuéllar AD, Webber ME. Cow power: the energy and emissions benefits of converting manure to biogas. Environmental Research Letters. 2008;3: 34002. doi:10.1088/1748-9326/3/3/034002

15. Tambone F, Scaglia B, D'Imporzano G, Schievano A, Orzi V, Salati S, et al. Assessing amendment and fertilizing properties of digestates from anaerobic digestion through a comparative study with digested sludge and compost. Chemosphere. 2010;81: 577–583. doi:10.1016/j.chemosphere.2010.08.034

16. Rehl T, Müller J. Life cycle assessment of biogas digestate processing technologies. Resources, Conservation and Recycling. 2011;56: 92–104. doi:10.1016/j.resconrec.2011.08.007

17. Chen HH, Lee AHI. Comprehensive overview of renewable energy development in Taiwan. Renewable and Sustainable Energy Reviews. 2014;37: 215–228. doi:10.1016/j.rser.2014.04.055

18. Weiland P. Biogas production: current state and perspectives. Applied Microbiology and Biotechnology. 2010;85: 849–860. doi:10.1007/s00253-009-2246-7

19. Yadvika, Santosh, Sreekrishnan TR, Kohli S, Rana V. Enhancement of biogas production from solid substrates using different techniques––a review. Bioresource Technology. 2004;95: 1–10. doi:10.1016/j.biortech.2004.02.010

20. He S, Ivanova N, Kirton E, Allgaier M, Bergin C, Scheffrahn RH, et al. Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. Korb J, editor. PLoS ONE. 2013;8: e61126. doi:10.1371/journal.pone.0061126

21. Roggenbuck M, Sauer C, Poulsen M, Bertelsen MF, Sørensen SJ. The giraffe (*Giraffa camelopardalis*) rumen microbiome. FEMS Microbiology Ecology. 2014;90: 237–246. doi:10.1111/1574-6941.12402

22. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, et al. Metagenomics of the Svalbard Reindeer Rumen Microbiome Reveals Abundance of Polysaccharide Utilization Loci. Liles MR, editor. PLoS ONE. 2012;7: e38571. doi:10.1371/journal.pone.0038571

23. Morrison M, Pope PB, Denman SE, McSweeney CS. Plant biomass degradation by gut microbiomes: more of the same or something new? Current Opinion in Biotechnology. 2009;20: 358–363. doi:10.1016/j.copbio.2009.05.004

24. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42: D490-495. doi:10.1093/nar/gkt1178

25. Hiras J, Wu Y-W, Deng K, Nicora CD, Aldrich JT, Frey D, et al. Comparative Community Proteomics Demonstrates the Unexpected Importance of

Actinobacterial Glycoside Hydrolase Family 12 Protein for Crystalline Cellulose Hydrolysis. mBio. 2016;7: e01106-16. doi:10.1128/mBio.01106-16

26. Doi RH. Cellulases of Mesophilic Microorganisms: Cellulosome and Noncellulosome Producers. Annals of the New York Academy of Sciences. 2008;1125: 267–279. doi:10.1196/annals.1419.002

27. Gomez del Pulgar EM, Saadeddin A. The cellulolytic system of *Thermobifida fusca*. Critical Reviews in Microbiology. 2014;40: 236–247. doi:10.3109/1040841X.2013.776512

28. Torto-Alalibo T, Purwantini E, Lomax J, Setubal JC, Mukhopadhyay B, Tyler BM. Genetic resources for advanced biofuel production described with the Gene Ontology. Frontiers in Microbiology. 2014;5. doi:10.3389/fmicb.2014.00528

29. Yang C, Xia Y, Qu H, Li A-D, Liu R, Wang Y, et al. Discovery of new cellulases from the metagenome by a metagenomics-guided strategy. Biotechnology for Biofuels. 2016;9. doi:10.1186/s13068-016-0557-3

30. Mackenzie AK, Pope PB, Pedersen HL, Gupta R, Morrison M, Willats WGT, et al. Two SusD-Like Proteins Encoded within a Polysaccharide Utilization Locus of an Uncultured Ruminant Bacteroidetes Phylotype Bind Strongly to Cellulose. Applied and Environmental Microbiology. 2012;78: 5935–5937. doi:10.1128/AEM.01164-12

31. Naas AE, Mackenzie AK, Mravec J, Schuckel J, Willats WGT, Eijsink VGH, et al. Do Rumen Bacteroidetes Utilize an Alternative Mechanism for Cellulose Degradation? mBio. 2014;5: e01401-14-e01401-14. doi:10.1128/mBio.01401-14

32. Zarraonaindia I, Smith DP, Gilbert JA. Beyond the genome: community-level analysis of the microbial world. Biology & Philosophy. 2013;28: 261–282. doi:10.1007/s10539-012-9357-8

33. Kleerebezem R, van Loosdrecht MC. Mixed culture biotechnology for bioenergy production. Current Opinion in Biotechnology. 2007;18: 207–212. doi:10.1016/j.copbio.2007.05.001

34. Vanwonterghem I, Jensen PD, Ho DP, Batstone DJ, Tyson GW. Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. Curr Opin Biotech. 2014;27: 55–64.

35. Nelson MC, Morrison M, Yu Z. A meta-analysis of the microbial diversity observed in anaerobic digesters. Bioresource Technology. 2011;102: 3730–3739. doi:10.1016/j.biortech.2010.11.119

36. Sundberg C, Al-Soud WA, Larsson M, Alm E, Yekta SS, Svensson BH, et al. 454 pyrosequencing analyses of bacterial and archaeal richness in 21 full-scale biogas digesters. FEMS Microbiology Ecology. 2013;85: 612–626. doi:10.1111/1574-6941.12148

37. Lebuhn M, Gronauer A. Microorganisms in the biogas-process - the unknown beings. Landtechnik. 2009;64: 127–130.

38. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation. 2012;2: 3. doi:10.1186/2042-5783-2-3

39. Scholz MB, Lo C-C, Chain PS. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current Opinion in Biotechnology. 2012;23: 9–15. doi:10.1016/j.copbio.2011.11.013

40. Kim M, Lee K-H, Yoon S-W, Kim B-S, Chun J, Yi H. Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. Genomics & Informatics. 2013;11: 102. doi:10.5808/GI.2013.11.3.102

41. Ladoukakis E, Kolisis FN, Chatziioannou AA. Integrative workflows for metagenomic analysis. Frontiers in Cell and Developmental Biology. 2014;2. doi:10.3389/fcell.2014.00070

42. Mardis ER. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 2008;9: 387–402. doi:10.1146/annurev.genom.9.081307.164359

43. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008;26: 1135–1145. doi:10.1038/nbt1486

44. Patel RK, Jain M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. Liu Z, editor. PLoS ONE. 2012;7: e30619. doi:10.1371/journal.pone.0030619

45. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods. 2013;63: 41–49. doi:10.1016/j.ymeth.2013.06.027

46. Yang X, Liu D, Liu F, Wu J, Zou J, Xiao X, et al. HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinformatics. 2013;14: 33. doi:10.1186/1471-2105-14-33

47. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27: 863–864. doi:10.1093/bioinformatics/btr026

48. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. Babraham Bioinformatics; Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

49. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

50. Pop M. Comparative genome assembly. Briefings in Bioinformatics. 2004;5: 237–248. doi:10.1093/bib/5.3.237

51. De Filippo C, Ramazzotti M, Fontana P, Cavalieri D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Briefings in Bioinformatics. 2012;13: 696–710. doi:10.1093/bib/bbs070

52. Boisvert S, Laviolette F, Corbeil J. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. Journal of Computational Biology. 2010;17: 1519–1533. doi:10.1089/cmb.2009.0238

53. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95: 315–327. doi:10.1016/j.ygeno.2010.03.001

54. Myers EW. A Whole-Genome Assembly of Drosophila. Science. 2000;287: 2196–2204. doi:10.1126/science.287.5461.2196

55. Batzoglou S. ARACHNE: A Whole-Genome Shotgun Assembler. Genome Research. 2002;12: 177–189. doi:10.1101/gr.208902

56. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; doi:10.1038/nature03959

57. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences. 2001;98: 9748–9753. doi:10.1073/pnas.171285098

58. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008;18: 821–829. doi:10.1101/gr.074492.107

59. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In: Berger B, editor. Research in Computational Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. pp. 426–440. Available: http://link.springer.com/10.1007/978-3-642-12683-3_28

60. Warren RL, Sutton GG, Jones SJM, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2007;23: 500–501. doi:10.1093/bioinformatics/btl629

61. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, et al. Extending assembly of short DNA sequences to handle error. Bioinformatics. 2007;23: 2942–2944. doi:10.1093/bioinformatics/btm451

62. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Research. 2007;17: 1697–1706. doi:10.1101/gr.6435207

63. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. Molecular Systems Biology. 2014;9: 666–666. doi:10.1038/msb.2013.22

64. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology. 2012;13: R122. doi:10.1186/gb-2012-13-12-r122

65. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28: 1420–1428. doi:10.1093/bioinformatics/bts174

66. Bushnell B. BBMap - http://sourceforge.net/projects/bbmap/ [Internet]. Available: http://sourceforge.net/projects/bbmap/

67. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

68. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

70. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics. 2012;28: 125–126. doi:10.1093/bioinformatics/btr595

71. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. Genome Research. 2011;21: 1552–1560. doi:10.1101/gr.120618.111

72. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biology. 2008;9: R151. doi:10.1186/gb-2008-9-10-r151

73. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 2012;28: 1033–1034. doi:10.1093/bioinformatics/bts079

74. Shah N, Tang H, Doak TG, Ye Y. Comparing bacterial communities inferred from16S rRNA gene sequencing and shotgun metagenomics. Biocomputing 2011. WORLD SCIENTIFIC; 2010. pp. 165–176. Available: http://www.worldscientific.com/doi/abs/10.1142/9789814335058_0018

75. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 2013;41: D590–D596. doi:10.1093/nar/gks1219

76. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research. 2014;42: D633–D642. doi:10.1093/nar/gkt1244

77. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and Environmental Microbiology. 2006;72: 5069–5072. doi:10.1128/AEM.03006-05

78. Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28: 3211–3217. doi:10.1093/bioinformatics/bts611

79. Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, et al. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in Bioinformatics. 2012;13: 107–121. doi:10.1093/bib/bbr009

80. Woese CR, Achenbach L, Rouviere P, Mandelco L. Archaeal Phylogeny: Reexamination of the Phylogenetic Position of Archaeoglohus fulgidus in Light of Certain Composition-induced Artifacts. Systematic and Applied Microbiology. 1991;14: 364–371. doi:10.1016/S0723-2020(11)80311-5

81. Baldauf SL. A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. Science. 2000;290: 972–977. doi:10.1126/science.290.5493.972

82. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565

83. Li W, Fu L, Niu B, Wu S, Wooley J. Ultrafast clustering algorithms for metagenomic sequence analysis. Briefings in Bioinformatics. 2012;13: 656–668. doi:10.1093/bib/bbs035

84. Yona G. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Research. 2000;28: 49–55. doi:10.1093/nar/28.1.49

85. Pipenbacher P, Schliep A, Schneckener S, Schonhuth A, Schomburg D, Schrader R. ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics. 2002;18: S182–S191. doi:10.1093/bioinformatics/18.suppl_2.S182

86. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389

87. Mika S. UniqueProt: creating representative protein sequence sets. Nucleic Acids Research. 2003;31: 3789–3791. doi:10.1093/nar/gkg620

88. Lin H-H, Liao Y-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Scientific Reports. 2016;6: 24175. doi:10.1038/srep24175

89. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, et al. Community-wide analysis of microbial genome sequence signatures. Genome Biology. 2009;10: R85. doi:10.1186/gb-2009-10-8-r85

90. Laczny CC, Pinel N, Vlassis N, Wilmes P. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. Scientific Reports. 2014;4. doi:10.1038/srep04516

91. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian H, et al. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3: 1. doi:10.1186/s40168-014-0066-1

92. Sandberg R, Bränden C-I, Ernberg I, Cöster J. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene. 2003;311: 35–42.

93. Pride DT. Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. Genome Research. 2003;13: 145–158. doi:10.1101/gr.335003

94. Abe T. Informatics for Unveiling Hidden Genome Signatures. Genome Research. 2003;13: 693–702. doi:10.1101/gr.634603

95. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology. 2014;32: 822–828. doi:10.1038/nbt.2939

96. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2: e603. doi:10.7717/peerj.603

97. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage

binning of multiple metagenomes. Nature Biotechnology. 2013;31: 533–538. doi:10.1038/nbt.2579

98. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nature Methods. 2014;11: 1144–1146. doi:10.1038/nmeth.3103

99. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome. 2014;2: 26. doi:10.1186/2049-2618-2-26

100. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3: e1165. doi:10.7717/peerj.1165

101. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science. 2014;5. doi:10.3389/fpls.2014.00209

102. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Research. 2010;38: e132–e132. doi:10.1093/nar/gkq275

103. Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Research. 2009;37: W101–W105. doi:10.1093/nar/gkp327

104. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Research. 2012;40: e9–e9. doi:10.1093/nar/gkr1067

105. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology. 1997;268: 78–94. doi:10.1006/jmbi.1997.0951

106. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

107. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. DNA Research. 2008;15: 387–396. doi:10.1093/dnares/dsn027

108. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Research. 2007;36: D480–D484. doi:10.1093/nar/gkm882

109. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research. 2014;42: D206–D214. doi:10.1093/nar/gkt1226

110. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4: 41. doi:10.1186/1471-2105-4-41

111. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, et al. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised

orthologous groups, species and functional annotations. Nucleic Acids Research. 2010;38: D190–D195. doi:10.1093/nar/gkp951

112. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl Acids Res. 2014;42: D459–D471. doi:10.1093/nar/gkt1103

113. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research. 2014;42: D222–D230. doi:10.1093/nar/gkt1223

114. Weinmaier T, Platzer A, Frank J, Hellinger H-J, Tischler P, Rattei T. ConsPred: a rule-based (re-)annotation framework for prokaryotic genomes. Bioinformatics. 2016; btw393. doi:10.1093/bioinformatics/btw393

115. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278: 631–637.

116. Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, et al. An ontology for microbial phenotypes. BMC Microbiology. 2014;14. doi:10.1186/s12866-014-0294-3

117. Feldbauer R, Schulz F, Horn M, Rattei T. Prediction of microbial phenotypes based on comparative genomics. BMC Bioinformatics. 2015;16: S1. doi:10.1186/1471-2105-16-S14-S1

118. MacDonald NJ, Beiko RG. Efficient learning of microbial genotype-phenotype association rules. Bioinformatics. 2010;26: 1834–1840. doi:10.1093/bioinformatics/btq305

119. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner [Internet]. [cited 3 Oct 2016]. Available: http://jgi.doe.gov/wp-content/uploads/2013/11/BB_User-Meeting-2014-poster-FINAL.pdf

120. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

121. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

122. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, et al. Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. Parkinson J, editor. PLoS ONE. 2012;7: e31386. doi:10.1371/journal.pone.0031386

123. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12: 385. doi:10.1186/1471-2105-12-385

124. Cantor M, Nordberg H, Smirnova T, Hess M, Tringe S, Dubchak I. Elviz – exploration of metagenome assemblies with an interactive visualization tool. BMC Bioinformatics. 2015;16. doi:10.1186/s12859-015-0566-4

125. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015;25: 1043–1055. doi:10.1101/gr.186072.114

126. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

127. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2011;39: D38–D51. doi:10.1093/nar/gkq1172

128. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R programming tools for plotting data. 2009.

129. Keating C, Cysneiros D, Mahony T, O'Flaherty V. The hydrolysis and biogas production of complex cellulosic substrates using three anaerobic biomass sources. Water Science & Technology. 2012;67: 293. doi:10.2166/wst.2012.543

130. Wirth R, Kovács E, Maróti G, Bagi Z, Rákhely G, Kovács KL. Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. Biotechnology for Biofuels. 2012;5: 41. doi:10.1186/1754-6834-5-41

131. Jaenicke S, Ander C, Bekel T, Bisdorf R, Dröge M, Gartemann K-H, et al. Comparative and Joint Analysis of Two Metagenomic Datasets from a Biogas Fermenter Obtained by 454-Pyrosequencing. Aziz RK, editor. PLoS ONE. 2011;6: e14519. doi:10.1371/journal.pone.0014519

132. Liu FH, Wang SB, Zhang JS, Zhang J, Yan X, Zhou HK, et al. The structure of the bacterial and archaeal community in a biogas digester as revealed by denaturing gradient gel electrophoresis and 16S rDNA sequencing analysis. Journal of Applied Microbiology. 2009;106: 952–966. doi:10.1111/j.1365-2672.2008.04064.x

133. Krause L, Diaz NN, Edwards RA, Gartemann K-H, Krömeke H, Neuweger H, et al. Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. Journal of Biotechnology. 2008;136: 91–101. doi:10.1016/j.jbiotec.2008.06.003

134. Ziganshina EE, Bagmanova AR, Khilyas IV, Ziganshin AM. Assessment of a biogas-generating microbial community in a pilot-scale anaerobic reactor. Journal of Bioscience and Bioengineering. 2014;117: 730–736. doi:10.1016/j.jbiosc.2013.11.013

135. Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, et al. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. Science. 2011;331: 463–467. doi:10.1126/science.1200387

136. López-Mondéjar R, Zühlke D, Becher D, Riedel K, Baldrian P. Cellulose and hemicellulose decomposition by forest soil bacteria proceeds by the action of structurally variable enzymatic systems. Scientific Reports. 2016;6: 25279. doi:10.1038/srep25279