Development of computational methods for the analysis of metagenome and metatranscriptome data

Ph. D. Thesis

submitted to the

Faculty of Technology
Bielefeld University, Germany

by

Martha Zakrzewski

June, 2012

Advisors: Prof. Dr. Jens Stoye
Prof. Dr. Alfred Pühler

Printed on non-aging paper according to DIN-ISO 9706

Summary

The fields of metagenomics and metatranscriptomics have evolved as helpful disciplines to unlock the taxonomic composition and functional diversity of heterogeneous microbial communities in their natural habitats. Both fields are mainly facilitated by advances in sequencing technologies that enabled the study of microorganisms in a high-throughput manner. At the same time, the sequencing technologies posed challenges on the storage, computational processing and analysis of high-throughput datasets.

In the scope of this thesis, methods were designed and developed that allow the interpretation of metagenome and metatranscriptome data in terms of taxonomic and functional information hidden in natural microbial communities. At first, the system MetaSAMS has been designed, developed and applied, which facilitates the automated storage, processing and analysis of whole metagenome shotgun datasets. MetaSAMS is accessible over a web-based user interface, which supplies the functional and taxonomic annotations for specific metagenome projects in graphical and tabular representations. Furthermore, the pipeline AMPLA for the analysis of the phylogenetic marker gene encoding 16S rRNA was designed and implemented, which generates an elaborate taxonomic profile of an underlying community. The workflow consists of several consecutive steps, namely the processing, clustering and taxonomic characterization of the data. Finally, the metatranscriptome pipeline MeTra was designed and implemented, which captures central RNA types for the taxonomic and functional profiling of the microorganisms in a community.

This thesis demonstrates the functionalities of the three pipelines on respective datasets obtained from a biogas plant. Knowledge of the microorganisms residing in a biogas fermenter is highly important, as biogas is a renewable and environmentally-friendly energy source. Analyses of the metagenome deduced in MetaSAMS confirmed previous findings that *Firmicutes* and *Euryarchaeota* dominate the biogas-producing community. Moreover, analyses of 16S rRNA gene sequences provided detailed insights into the diversity of species and highlighted that still the origin of some sequences is not well described, which is due to the absence of appropriate reference sequences in databases. The metatranscriptome pipeline unveiled that the most abundant species dominating the community also contributed the majority of the transcripts. The analysis shed light on the central processes of the anaerobic biogas digestion and the associated bacteria.

Finally, a method for the discovery of industrially relevant enzymes was designed. The method was applied for the identification of novel laccase genes in metagenomes obtained from marine habitats. Laccases are important in many industrial processes. Therefore, novel laccases with improved functionalities are required. The analysis demonstrated that laccases are widely distributed in bacterial species. Moreover, only 34% of metagenome sequences encoding fragments of putative laccases could be assigned to a genus indicating potentially novel enzymes.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Preface

Single-celled microorganisms (*microbes*) were the first form of life around 3.5 billion years ago [Altermann and Kazmierczak, 2003]. Today it is estimated that around 5×10^{30} microbes are on Earth [Whitman et al., 1998], constituting the most abundant and diverse form of life. Although microbes played a crucial role in vinegar production around 5000 BC [Lück and Jager, 1997] and in cheese production around 3000 BC [Loessner et al., 2006], it was not until the late 1670s that Antonie van Leeuwenhoek observed microbes through a microscope [Atlas and Bartha, 1998]. The study of microscopic organisms, termed *microbiology*, was born, which allowed for unveiling of important roles of microbes in many beneficial or harmful processes. Microscopic organisms were found in extreme environments like deep sea vents [Xie et al., 2011], the arctic [Varin et al., 2012] or acid mine drainage [Inskeep et al., 2010]. Due to their structure and metabolic capabilities, they are well adapted to live in different habitats.

Microbes are important in a range of fields such as agriculture, medicine and biotechnology. They support all life on Earth including the humans [Qin et al., 2010]. Microbes reside on and in the human body, and according to estimates the number of microbes outnumbers the number of human cells by tenfold [Ley et al., 2006]. Additionally, the genome of the microbial communities, the human microbiome, contains 100-fold more genes than the human genome. Many processes rely on the microbes that colonize the human body. They are involved in digestion of food, detoxification of harmful chemicals and defending the body against human pathogens, but they are also associated with obesity, cancer and allergies [Flint, 2011, Ly et al., 2011]. Examining how microbial communities affect human life could lead to advances in human health.

Microbial communities are crucial participants in agriculture. Some microorganisms support plants with nutrients that they need in order to grow. Atmospheric nitrogen, for example, is not accessible by plants, however, specific bacteria convert atmospheric nitrogen into ammonia, which can be utilized by plants and functions as a fertilizer [Desbrosses and Stougaard, 2011]. Other bacteria support plants by protecting them from infections caused by pathogens [Chen et al., 2007]. Finally, microbes play an immense role in remediation of natural and human-made waste. To illustrate, they are involved in decomposing biowaste composts [Partanen et al., 2010], cleaning up oil spills in oceans [Valentine et al., 2012] and removing contaminants from sewage [Evans and Furlong, 2011].

The role of microorganisms is not only important for natural processes. The variety of functions of microorganisms is utilized in a range of useful biotechnological applications, such as ethanol production. Since fossil fuels are finite and the global use of energy is increasing steadily, renewable energy sources have attracted considerable attention. In addition, using renewable energy sources helps to mitigate carbon dioxide, which is associated with global warming [Matthews et al., 2009]. An environmentally friendly, biologically based alternative energy source is ethanol, which has been used as biofuel [Vertes et al., 2011]. It is produced by microbes during fermentation of corn, sugarcane or other agricultural sources. Several microbes are necessary in this process, each carrying out different steps. The first group of microbes transforms cellulose contained in the agricultural wastes into sugar. This product is then fermented by other microbes, which produce ethanol. Deciphering how microbial communities are involved in this process might help to increase the yield of ethanol.

Despite the importance of microbes, our knowledge about their diversity and functions in these processes is limited. In the 20th century, genomics has proved to be successful in studying microbes and their genetic material. Traditional microbial genomics is an organism-centered approach, where cultures containing microorganisms of one species are grown in the laboratory, followed by the sequencing and annotation of their genomes. However, a theoretical analysis has shown that one milliliter of gut fluid contains 10^{11} bacteria [Whitman et al., 1998], while one gram soil harbors approximately 10^9 bacteria [Sait et al., 2002]. The enormous number of microorganisms present in these samples cannot be analyzed in the laboratory in appropriate time using traditional genomics approaches. In addition, some of the microbes require growth conditions that are so far unknown or difficult to obtain for cultivation in the laboratory.

Advances posed by novel abilities of accessing microbial genetic material without laboratorial cultivation allowed for the development of *metagenomics*. In metagenomics [Handelsman et al., 1998], a set of genomes, termed *metagenome*, from an environmental community is studied rather than one genome of an individual species. Metagenomics is a rapidly growing field, which is having a broad impact on the traditional microbiological and genomic research. For the first time, it is possible to get a comprehensive view of a microbial community in its natural environment and to study the entire genetic make-up of a community as a whole in terms of its taxonomic composition and metabolic potential [Bertin et al., 2008]. Above all, metagenome data are suitable to identify novel enzymes with potential industrial, biotechnological or medical applications. Moreover, an analysis of marker genes allows for unveiling a deeper view of the taxonomic compositions by using gene-centric approaches. Analogous to metagenomics, *metatranscriptome* refers to a set of transcripts expressed by a microbial community under specific conditions or at different time points. More precisely, metatranscriptome experiments give insights into the active members of a community and their functional importance within a habitat.

On one hand, metagenomics, gene-centric approaches and metatranscriptomics are more effective and less time-consuming means to get a comprehensive view of the community and to discover novel enzymes than conventional genomics and transcriptomics. On the other hand, metagenome, gene-centric and metatranscriptome projects produce, as a consequence of significant improvements in sequencing techniques, a high amount of data, which makes the computational analysis of the microbial community more challenging. In this regard, computational approaches are required that enable the interpretation of the vast amount of data as well as the identification of potential genes encoding enzymes, which are of interest in the agricultural, medical or environmental fields. In this thesis, methods are designed and implemented that tackle the high amount of data obtained by the three different approaches, namely metagenome, gene-centric and metatranscriptome sequencing, and unveil the taxonomic and functional potential of complex microbial communities.

1.2 Overview

Chapter 2 provides a broad overview of the biological background necessary for understanding the present work. After that in Chapter 3, existing approaches are listed and explained that are commonly used for the interpretation of metagenome and metatranscriptome data. In Chapter 4 the methods and implementations undertaken in this work are described. Thereafter, the methods are performed on application examples, and the outcomes are presented in Chapter 5. Then, Chapter 6 summarizes the work and discusses aspects associated with the results. Finally, an outlook in Chapter 7 closes the thesis.

CHAPTER 2

Background

This chapter gives background information into the biology to allow a better understanding of the context and questions that are addressed in this thesis.

2.1 DNA and sequencing techniques

Deoxyribose nucleic acid (DNA) is the genetic material of all known living organisms [Avery et al., 1944, Hershey and Chase, 1952]. It consists of repeating bases bound to a sugar-phosphate backbone. The unit of a base, sugar and phosphate group is also named "nucleotide" (Fig. 2.1). The four bases are adenine (A), cytosine (C), guanine (G) and thymine (T) [Levene, 1919]. DNA forms a double helix, which is held together by hydrogen bonds between the bases: adenine pairs with thymine and guanine with cytosine [Watson and Crick, 1953a, Watson and Crick, 1953b]. The double-stranded DNA is built during DNA replication, in which a single strand of DNA, also termed "template", is duplicated by adding nucleotides in a manner that complementary bases are opposite to each other [Bessman et al., 1958, Lehman et al., 1958].

The order of the bases, the DNA "sequence", in genomes varies between different species. Knowledge of the order of the bases is essential for the interpretation of genomes. The sequence of a genome consists of coding and non-coding stretches of DNA. The coding regions encode "genes". DNA sequences of genes can be transcribed into ribonucleic acid (RNA) sequences. There are different types of RNA molecules according to their roles in the cell [Lodish, 2004]. During translation, the information carried by a messenger RNA (mRNA) is decoded into a specific sequence of amino acids

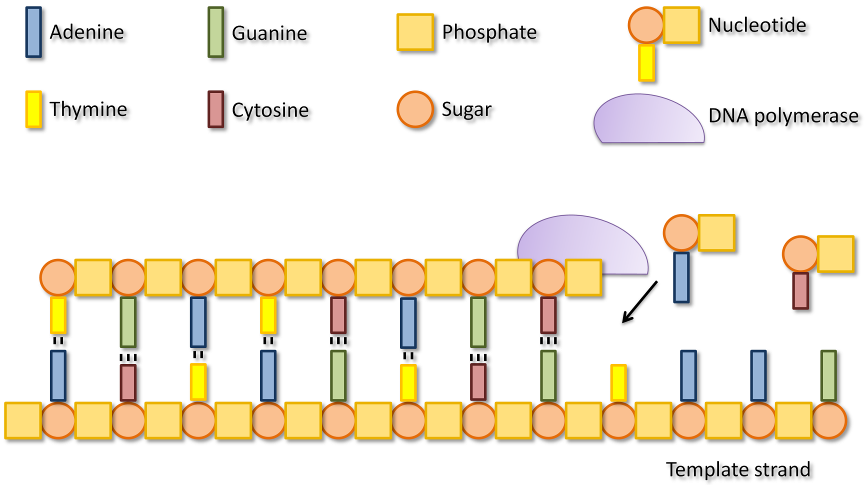


Figure 2.1: Structure of double-stranded DNA: DNA is made up from nucleotides that are joined together by sugar-phosphate linkages. During replication, the DNA polymerase runs across a template single strand and builds a new strand by adding suitable nucleotides in a manner that adenine binds to thymine and guanine to cytosine.

that form a "protein". Proteins can function as enzymes by catalyzing specific chemical reactions. Further RNA types are, for example, transfer RNAs (tRNAs) and ribosome RNAs (rRNAs), which are non-protein coding but indispensable in the translational process.

Different sequencing methods are used to determine the order of the four bases in DNA strand fragments of unknown sequences. The results of the sequencing procedure are "reads" that are contiguous sequences containing the order of the nucleotide letters in the fragments. Overlapping reads form a contiguous sequence, named "contig".

In the following, a brief overview of the most widely used sequencing techniques will be given. Since only the fundamental aspects and properties will be presented, the reader is referred to the reviews [Mardis, 2009, Metzker, 2010] for a detailed description. It is to be noted that the overview reflects only the current state of sequencing technologies as they are rapidly evolving.

2.1.1 Traditional sequencing

Initial studies used Sanger sequencing techniques [Sanger et al., 1977], also referred to as dideoxynucleotide sequencing or chain-termination sequencing, to determine the sequence in a DNA strand. Briefly, copies of DNA strands are generated using clone libraries or polymerase chain reaction (PCR) [Saiki et al., 1988] to obtain enough genomic material for sequencing. Subsequently, the amplified templates are replicated in reactions using oligonucleotide primer, DNA polymerase, unlabeled deoxynucleotides (dNTP: dATP, dCTP, dGTP, dTTP) and fluorescently labeled dideoxynucleotides (ddNTPs: ddATP, ddCTP, ddGTP, ddTTP). The amplification is carried out in four separate reaction sets, which differ by the contained ddNTPs. The primer initiates the replication of the template fragment by supporting the binding of the enzyme DNA polymerase, which extends the complementary chain by adding dNTPs or ddNTPs. Herein, the reaction terminates by the random incorporation of a labeled ddNTP. As multiple copies of the templates are present in each of the four reactions sets, complementary fragments of different sizes are generated. Sanger sequencing is based on the electrophoretic separation of these fragments and the detection of the fluorescence for each ddNTP using a laser.

Although Sanger sequencing was introduced in 1977, an improved and optimized version is nowadays still in use. The main reason for its continued use is that longer reads are achievable compared to the next-generation sequencing techniques. For example, the Applied Biosystems 3730xl DNA Analyzer produces on average 700 bases reads using Sanger sequencing and can generate 1.6 Mb of sequence data within one day.

2.1.2 Next-generation sequencing

The introduction of next-generation sequencing (NGS) techniques had a big impact on genomics and metagenomics studies. They allowed the direct sequencing of DNA molecules, bypassing the cloning step that is required for Sanger sequencing (Fig. 2.2). NGS consists of two steps. First, DNA templates are amplified. Second, the amplified fragments are *sequenced-by-synthesis*, where each read is produced in real-time during replication of template DNA. The methods applied by different sequencing techniques will be outlined in the next paragraphs.

Roche/454 pyrosequencing

The Genome Sequencer (GS) was introduced in 2005 by 454 Life Science[1] and was the first commercially available NGS platform. The applied technique is based on emulsion PCR [Williams et al., 2006] and 454 pyrosequencing [Hyman, 1988]. First, DNA is randomly broken into fragments [Margulies et al., 2005]. Two different adapters are attached to the fragments, which are required for subsequent purification, quantitation, amplification and sequencing. Next, the templates are amplified using emulsion PCR. Herein, each template DNA binds with the attached adapter to a primer-coated bead inside a droplet that is formed within an oil emulsion. Each droplet contains PCR reagents for the amplification of the template. Finally, the beads are placed into wells

[1] http://www.454.com/

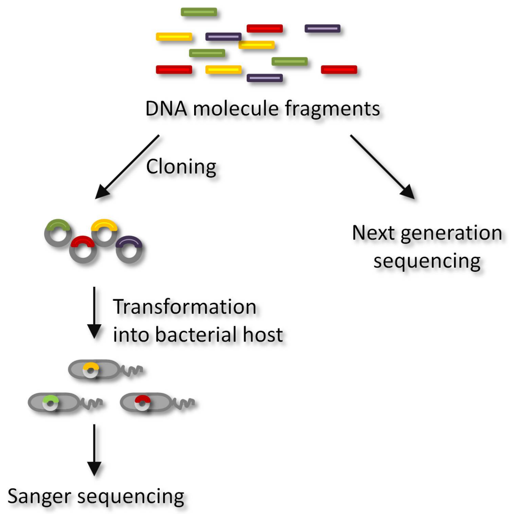

DNA molecule fragments
Cloning
Next generation sequencing
Transformation into bacterial host
Sanger sequencing

Figure 2.2: Traditional (left) and next-generation sequencing strategies (right): In the traditional Sanger approach, the template sequence is amplified by the replication system of a host. Therefore, the fragment is cloned into a vector and transformed into a host cell. Replication is necessary to achieve sufficient copies needed for the Sanger sequencing. Unfortunately, the host may not replicate all fragments due to incompatibility with the host metabolisms. Next-generation sequencing prevents cloning bias by avoiding the cloning step. Instead, the templates are directly utilized by NGS techniques.

located on an optical array of fibers. The wells are constructed in a way that only a single bead fits into it.

The sequencing is carried out using the pyrosequencing technique. A primer binds to the template at the beginning of the sequencing procedure. A DNA polymerase then adds complementary nucleotides. Pyrosequencing is based on pyrophosphate release during the incorporation of a nucleotide. Thereafter, pyrophosphate is involved in an enzymatic reaction, in which light is emitted [Nyren and Lundin, 1985]. The amplitude of each emission can be detected with a CCD camera and is approximately proportional to the number of incorporated nucleotides. After signal detection, the nucleotides that were not incorporated are washed off, and new nucleotides are added cyclically in a fixed order.

The pyrosequencing technique has a drawback with homopolymeric regions, which is a stretch on the template containing consecutive identical nucleotides. For homopolymers with a length of at least six bases, the detection can be inaccurate [De Schrijver et al., 2010]. The technique can produce deletions and insertions in homopolymeric stretches, while substitutions are less common. The Genome Sequencer

(GS) FLX+ device has recently reached Sanger-like read lengths of 700 bases (Tab. 2.1) and the typical throughput is 700 Mb per day.

Illumina (Solexa)

Solexa launched the 1G Genetic Analyzer in 2006. One year later Solexa was acquired by Illumina[2]. The Illumina technique uses bridge amplification for the replication of the template DNA molecules and reversible terminator chemistry for the sequencing step [Turcatti et al., 2008].

For bridge amplification, adapters are ligated to the templates allowing the binding of both ends to primers that are attached on a glass slide. Subsequently, the DNA templates are amplified and bridges of replicated DNA fragments are formed on the slide. The sequencing is based on dye-nucleotides that function as reversible terminators. Each dye-nucleotide is labeled with a base-specific color. If a nucleotide is incorporated during the amplification step, the synthesis terminates, and the dyes can be detected. After dye detection and nucleotide assignment, the dye and terminator moieties are removed such that the next nucleotide can bind.

The read length is shorter compared to 454 pyrosequencing, but Illumina produces a higher throughput and the overall sequencing costs are lower. The Illumina HiSeq2000 device can produce up to 25 Gb data with 2 × 100 bp long reads per day (Tab. 2.1).

Applied Biosystems SOLiD

Since 2007, Applied Biosystems SOLiD™ technology[3] employs sequencing by ligation [Shendure et al., 2005]. Similar to the 454 platform, the target fragments are amplified by emulsion PCR. After that, ligations are carried out using a mixture containing primers, ligases and dye-labeled oligonucleotides of the length 8 bases. These octamers consist of six degenerate nucleotides supporting the binding and two variable nucleotides mediating the binding specificity. Four dinucleotides of the 16 possible octamers are associated with fluorescent dyes of the same color. Ligases join the octamers and the template strand such that the dinucleotides within the octamer are complementary to the nucleotides in the template strand. After the ligation step, fluorescence is detected, three degenerate nucleotides carrying the dye are removed, and the ligation steps are repeated. The replication of a template using ligations is performed several times with varying starting positions for the primer. In total, each base of the template is read twice as a consequence of the primer shift and the dinucleotide-specific octamers. Finally, a two-base encoding color scheme is used to decipher the sequence of a template.

The ligase-based approach prevents those sequencing errors that are induced by the DNA polymerase. Substitution is the most common error [Metzker, 2010], since in-

[2] http://www.illumina.com/

[3] http://www.appliedbiosystems.com

sertions and deletion are unlikely due to the color-coded dinucleotide approach. The first device was launched in 2007, and at present time, SOLiD - 5500xl generates read lengths ranging from 35 to 75 bases and up to 15 Gb per day (Tab. 2.1).

Ion Torrent

The company Ion Torrent[4], which is a subsidiary of Life Technology, released the Personal Genome Machine (PGM) in 2010. Amplification is accomplished by emulsion PCR. Compared to the aforementioned procedures, PGM uses a pH- and not fluorescence-mediated sequencing method. The Ion Torrent technology is based on a semiconductor chip [Rothberg et al., 2011], which is capable of converting a chemical signal into digital information. After incorporation of a nucleotide by a polymerase, a hydrogen ion is released as a by-product and changes the pH of the solution. This event can be detected by the semiconductor chip. Similar to 454 pyrosequencing, homopolymeric regions are problematic, as the measurement of the pH change is inaccurate with a high number of released hydrogen ions. Currently, the 318 chip is able to sequence up to 250 bases, but a read length of 400 bases is expected in 2012.

Table 2.1: Comparison of next-generation sequencing platforms[1]

Device	GS FLX Titanium+	Illumina HiSeq 2000	5500xl System	Ion Torrent 318 chip
Platform	Roche/454	Illumina GA	ABI SOLid	Ion Torrent
Amplification method	Emulsion PCR	Bridge amplification	Emulsion PCR	Emulsion PCR
Sequencing method	Pyro-sequencing	Reversible dye terminators	Sequencing by ligation	Ion Semiconductor Sequencing
Read length [bp]	700	2 × 100	35-75	250
Throughput per day	700 Mb	25 Gb	10 - 15 Gb	1 Gb
Million of reads per run	1	2,000	2,000	4-8
Run time	23 h	8 days	2-8 days	< 2 h

[1]based on [Glenn, 2011], partly updated according to the corresponding websites of the companies as available in March 2012.

With the advances in NGS technologies, costs and manual efforts were dramatically reduced. Recently, NGS techniques have reached read lengths comparable to that obtained by Sanger sequencing. A characteristic of NGS is that many DNA fragments are processed in parallel, producing a high amount of sequences per run. Because of this, NGS outperforms Sanger sequencing in terms of the throughput. The sequencing performances of the different NGS techniques are listed in Table 2.1.

[4] http://www.iontorrent.com/

Currently, Roche 454 pyrosequencing and Illumina technology are the most widely used NGS methods. As illustrated in Table 2.1, the 454 platform provides longer reads facilitating better interpretation of the sequences, whereas Illumina allows a higher coverage with shorter reads. 454 pyrosequencing can introduce homopolymeric errors leading to the generation of artificial reads.

2.1.3 Single molecule sequencing

In NGS techniques, amplification is accomplished either by using emulsion PCR or bridge amplification prior to the sequencing step. Single molecule sequencing techniques avoid the costly amplification procedures by integrating sensitive detection techniques or by circumventing sequencing-by-synthesis.

Helicos

The first commercial single molecule sequencing device, HeliScope, was released in 2008 by Helicos BioSciences Corporation[5]. It is based on the true single-molecule sequencing technology (tSMS™) [Braslavsky et al., 2003], which works as follows: Fragments are randomly immobilized on a glass slide. After that, fluorescently labeled bases, one of the four types in each cycle, are added. Similarly as in the Illumina approach, the nucleotides are reversible terminators. Thus, an incorporation of a nucleotide terminates the extension of the reverse strand of the template. After an image is taken with a high-resolution optical microscope, the terminator is removed and the process is repeated for the next base. The read length is rather short with on average 35 bases. Main error types are deletions [Pushkarev et al., 2009] since it is likely that the emitted signal may not be detected. Nevertheless, Helicos has already been applied to sequence a human [Pushkarev et al., 2009] and a viral genome [Harris et al., 2008].

Pacific Bioscience

Pacific Biosciences[6] developed the single molecule real time (SMRT™) DNA sequencing technology. SMRT uses the zero mode waveguide (ZMV) [Levene et al., 2003], which is a nanoscale well. A polymerase is fixed at the bottom of the surface within each ZMV, which is illuminated by a laser. The nucleotides are fluorescently labeled with different colors. Consequently, when a nucleotide is incorporated to the DNA target, the dye within the polymerase can be detected for tens of milliseconds, a measurable magnitude longer than diffusion events of bypassing nucleotides. The ZMVs are constructed in a way that only the fluorescence occurring close to the DNA polymerase is detected. The dye is cleaved off as part of the template extension reaction, and the next nucleotide can be incorporated.
The first commercial device is PacBio *RS*, which was delivered in mid 2011. The

[5] http://www.helicosbio.com/

[6] http://www.pacificbiosciences.com/

generated reads reach a length of 2.2 kb. However, the error rate is very high at approximately 15%.

Oxford nanopore

The nanopore sequencing concept is based on the measurement of an electronic signal [Branton et al., 2008]. A nanopore is made of a protein, usually α-hemolysin, which has a nanoscale hole. When a nanopore is located in a membrane and an electrical current is applied to it, a passing nucleotide strand would partly block the current on account of its shape and charge. Nanopores are suitable to distinguish the four bases and also modified bases in a strand based on the change in the current.

Oxford Nanopore Technologies Ltd.[7] was founded 2005 to develop a system that uses nanopores for an electronics-based sequencing technology. In the sequencing method, named "strand sequencing", a single-stranded DNA polymer passes through a protein nanopore. At the same time, individual DNA bases in the strand are deciphered. In the approach of Oxford Nanopore, a chip is used that allows the processing of sequences by several nanopores in parallel. Oxford Nanopore intends to commercialize their systems, GridION and MinION, to customers within 2012. Oxford Nanopore claims to be able to sequence a read length of tens of kbs (http://www.nanoporetech.com).

2.2 Metagenomics

Classical genomics and microbiology rely upon cultivation and study of a single microorganism. Using classical approaches, only a small fraction of the microbes in an environment can be grown in monoculture [Amann et al., 1995]. Metagenomics uses culture-independent methods to analyze a collection of genomes from different microbes referred to as the "metagenome" [Handelsman et al., 1998]. Although the idea to examine a whole community was already described in 1985 [Pace, 1985], the term metagenomics was coined nearly a decade later by Jo Handelsman, who used this term in connection with the analysis of collective genomes obtained from soil [Handelsman et al., 1998]. Nowadays, metagenomics is a rapidly growing field of research that aims at studying a heterogeneous microbial community in terms of its taxonomic structure and metabolic pathways.

The increased number of metagenomics projects today is mainly facilitated by the development of NGS techniques, which allow for sequencing environmental samples at low costs and without the cloning process inherent in the traditional methods. The following sections will summarize the development and improvements of metagenomics.

[7] http://www.nanoporetech.com

2.2.1 The gene-centric strategy

Early metagenomics used conserved and universally existing marker genes to study the microbial community structure. In particular, the prokaryotic 16S ribosomal RNA (rRNA) genes have been widely used. In *Bacteria* and *Archaea*, the 16S rRNA gene encodes a functional rRNA, which is part of the small subunit (SSU) of the ribosome. As ribosomes play a fundamental role during translation, genes representing ribosomal subunits are present in all cellular organisms. Moreover, regions of the 16S rRNA genes are subjected to selective pressure accordant to their immense importance in the translation process. Hence, sequences of 16S rRNA genes are conserved among different bacterial and archaeal species. The conserved regions within prokaryotic 16S rRNA genes are interspersed with nine hypervariable stretches [Neefs et al., 1991], which have changed at a constant rate over time. Based on the hypervariable regions, phylogenetic analyses can be accomplished. 16S rRNA genes have been sequenced extensively such that large databases of characterized reference sequences exist today [Cole et al., 2003, DeSantis et al., 2006, Pruesse et al., 2007].

A cultivation independent survey studying 16S rRNA genes was reported by Pace *et al.* in 1985. The concept was realized for the first time in 1991 using λ phage libraries to examine a marine community [Schmidt et al., 1991]. The first step in this approach was the lysis of the microbial cells, followed by the extraction and fragmentation of environmental genomic DNA. The fragments were inserted into bacteriophage λ clone vectors and transformed into host cells. Since each clone carried a random genomic fragment, clones containing 16S rRNA genes were identified using screenings by hybridization with 16S rRNA gene-specific probes. The detected 16S rRNA genes were amplified by PCR and sequenced using the Sanger technique. Sixteen sequences similar to *Cyanobacteria*, *Proteobacteria* and *Eukaryota* were identified. It was not possible to find a closely related reference for two of the obtained sequences giving evidence for novel phylogenetic groups.

To speed up the laborious procedure by avoiding the hybridization step, Giovannoni *et al.* used a PCR-mediated amplification of 16S rRNA genes from environmental DNA using 16S rRNA gene-specific primers [Giovannoni et al., 1990]. For segregation, the PCR products were cloned into clone libraries. Finally, the fragments were sequenced by Sanger technology, where full-length 16S rRNA gene sequences can be generated.

Using NGS sequencing techniques, 16S rRNA genes are sequenced after DNA extraction in a targeted 16S rDNA amplicon approach (Fig. 2.3). Similar to the method proposed by Giovannoni [Giovannoni et al., 1990], the selected regions within 16S rRNA genes are amplified by PCR with a universal or a group-specific primer set. The primers bind to highly conserved regions flanking hypervariable regions, which provide species-specific information. Next, the generated 16S rDNA amplicons are sequenced using NGS. Due to the short read lengths generated by the NGS technique, only partial 16S rRNA genes can be sequenced.

Sequencing of 16S rDNA amplicons using NGS is a fast and cheap technique to assess the taxonomic composition in a sample. Moreover, the barcoded pyrosequencing [Hamady et al., 2008] technique allows for sequencing of 16S rDNA amplicons obtained from several samples in parallel. Knowledge of the microbial diversity is especially important in estimating the number of sequences needed to get a comprehensive overview of the microbial community structure and metabolic pathways by whole metagenome sequencing (Section 2.2.2). In addition, 16S rDNA amplicon sequencing enables the detection of organisms that probably produce enzymes of interest. The presence of such organisms might help in the decision whether to perform whole metagenome sequencing of the same sample.

Since 16S rRNA genes are in some cases too conserved between closely related organisms to conduct species-specific assignments, less conserved phylogenetic marker genes, such as housekeeping genes encoding recombinase A (RecA), heat-shock protein (Hsp70), RNA-polymerase subunit B (RpoB) and elongation factor Tu (EF-Tu), are used to determine the taxonomic composition [Wu and Eisen, 2008]. For the analysis of methanogenic *Archaea*, a gene encoding the α subunit of the methyl-coenzyme M reductase (McrA) [Friedrich, 2005] proved to be a valuable phylogenetic marker. A further limitation in using a 16S rRNA gene-based approach for abundance estimation is that the gene occurs in multiple copies in many bacteria leading to false conclusions. It has been estimated that the mean number of bacterial ribosomal operons per genome is approximately 4 [DeSantis et al., 2006]. A recent study has shown that a typical bacterial genome in the GenBank database contains 1 to 15 copies of 16S rRNA genes [Pei et al., 2010]. A solution for this limitation is to use single-copy phylogenetic markers. For instance, the primase gene *dnaG*, translation initiation factor gene *inf*C and ribosomal protein L1 gene *rplA* are single-copy genes in most genomes and universally distributed in bacteria [Wu and Eisen, 2008].

A disadvantage of protein encoding or 16S rRNA marker genes is that horizontal gene transfer among unrelated taxa and gene duplication events might have an impact on the taxonomic assumptions. A further drawback of the gene-centric sequencing approach is the bias introduced by PCR amplification of the phylogenetic marker gene. Although "universal" primers covering different groups are usually employed [Baker et al., 2003], not all marker genes may be amplified equally due to primer bias. Additionally, detection of the targeted gene is an indication of the ability to encode only this single function, but this approach gives no direct information about the whole functional capabilities of the community. Some of the mentioned limitations are circumvented by the environmental whole metagenome shotgun sequencing approach [Venter et al., 2004], which can provide insights to functional characteristics of a microbial community by sequencing all genomic fragments without using specific primer sequences.

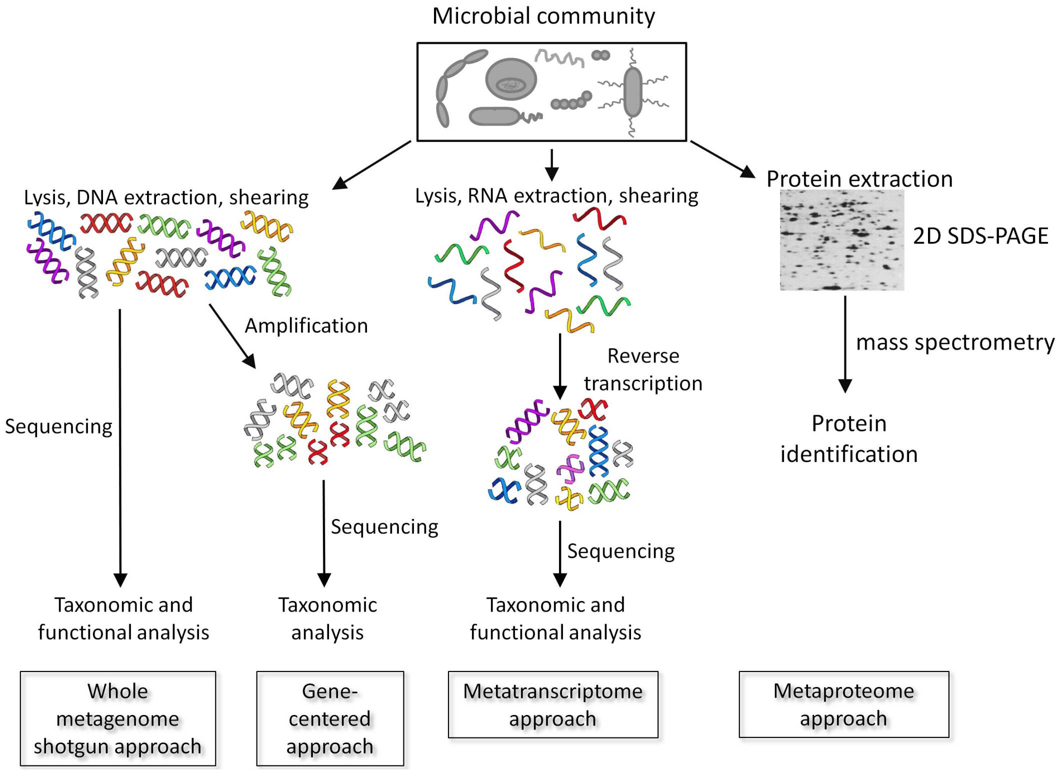


Figure 2.3: Schematic overview of different approaches for the analysis of microbial communities: In whole metagenome shotgun and gene-centered approaches microbial communities are analyzed based on their genomes or genes. Metatranscriptomics and metaproteomics explore the transcriptome and proteome of microorganisms, respectively.

2.2.2 The whole metagenome shotgun strategy

For the detection of taxa and biological functions present in an environmental sample, random shotgun sequencing of DNA extracted from an environmental community is carried out [Venter et al., 2004, Tyson et al., 2004]. In initial approaches, cell lysis and fragmentation of the isolated DNA were performed. The environmental fragments were used for the construction of clone libraries. For this purpose, the fragments were inserted into vectors, which were introduced into a suitable host cell, most commonly *Escherichia coli*. The host cells amplified the recombinant vectors in the course of cell division, and the amplified fragments were sequenced using the Sanger approach. The first projects mainly used contigs assembled from metagenome reads as the basis for the taxonomic and functional profiling. Thereby, the number of assembled contigs and the average contig length depend strongly on the diversity of the community, the size of the genomes, the relative abundance of species in the sampled community as well as the sequencing depth.

In 2004, two pioneering projects applying the whole metagenome shotgun strategy were published [Venter et al., 2004, Tyson et al., 2004]. One project was carried out by the J. Craig Venter Institute in 2004 [Venter et al., 2004]. The researchers used Sanger sequencing of a clone library to generate around two million randomly sequenced DNA fragments obtained from the Sargasso Sea. Approximately 1,800 species were identified with 148 previously unknown bacterial groups. In addition, 1.2 million unknown genes were discovered.
Tyson and colleagues sequenced the microbial community from a natural biofilm from an acid mine drainage [Tyson et al., 2004]. As the diversity in the community was low, the group was able to almost completely reconstruct the genomes of the dominant species.

NGS techniques enable amplification and sequencing of the templates subsequently after fragmentation, bypassing the need for the construction of a clone library. In particular, the cloning step may introduce bias into the results, as some inserts encoding for instance toxins are incompatible with the host's metabolism [Forns et al., 1997]. Additionally, using high-throughput sequencing techniques, the sequence coverage is increased and potential cloning biases can be avoided. The first pyrosequenced whole metagenome approach was applied to describe the environmental sample of two sites of the Soudan Mines that in spite of being adjacent to each other, differed in chemistry and hydrogeology [Edwards et al., 2006]. Comparative analyses revealed significant differences in the metabolic potential of the microbes within each site, which could be separated by metabolic processes like carbon utilization, iron acquisition mechanisms, nitrogen assimilation and respiratory pathway.

Latest developments in the area of high-throughput analytics have greatly increased the number of metagenome projects. Using the environmental genome shotgun strategy, taxa can be identified that are not captured with the 16S rRNA gene approach due to primer bias. Metagenomics provides the possibility to identify the taxonomic as well as the metabolic potential of a microbial community. Simultaneously, the interpretation of the functional repertoire allows access to novel key enzymes with potential biotechnological applications. However, the high-throughput feature of the NGS techniques and the short read length complicate the storage and interpretation of the data.

2.2.3 Biotechnological applications

Enzymes showing optimal activity and stability at different parameters (pH, temperature, pressure, salinity) are required in a broad range of industrial applications. In the conventional method, genes representing a desired function are isolated from known organisms and modified in a row of mutagenesis experiments [Kaur and Sharma, 2006]. The different mutant genes, "variant genes", are inserted into an expression system and

screened for a high activity under selective conditions. Unfortunately, discovering an optimal enzyme using the conventional method can be laborious and time-consuming [Fernández-Arrojo et al., 2010].
Nature has already engineered enzymes encoded by microbes, which have tremendously adapted to survive in all kinds of conditions. Thus, metagenomes are a promising source for discovering enzymes that efficiently function under conditions matching industrial requirements [Chistoserdova, 2010]. Two major types of strategies are successfully pursued: function-based and sequence-based metagenomics.

Function-based metagenomics

Using function-based metagenomics, novel enzymes catalyzing functions of interest can be identified in laboratories [Craig et al., 2010]. First, DNA from environmental samples is extracted. Next, DNA-fragments are inserted into vectors and transformed into host organisms, typically *E. coli*. Moreover, species of *Streptomyces* and *Pseudomonas* are used, in case where the transcription-translation machinery of *E. coli* is not compatible with the expression of the vector insert. Finally, the metagenomic library is screened for novel metabolic genes. The screening is based on the visual detection of growth of the host cell on selective media or the production of a colored metabolite.

Function-based metagenomics was firstly applied in 1995 [Healy et al., 1995]. Healy *et al.* constructed metagenome libraries, termed "zoolibraries", of a culture of environmental organisms obtained from a thermophilic, anaerobic digester. Genes encoding cellulases and other hydrolases were detected. One positive clone was sequenced by the Sanger technique to gain deeper knowledge of the phylogenetic origin of the novel enzyme.

The function-based screening approach has been used successfully for the identification of novel enzymes such as alcohol dehydrogenases, esterases and lipases [Ferrer et al., 2009, Rashamuse et al., 2009]. A disadvantage is that many screens are necessary in order to identify a positive clone in a metagenome library. In addition, a reaction may require several genes that encode different subunits or proteins acting together. Such functions might not be detected with metagenomic clones carrying only small inserts that might encode a partial, non-functional gene cluster. Finally, the expression of a gene and the subsequent detection rely upon the correct folding, availability of cofactors and the capability of the host organism to express this gene.

Sequence-based metagenomics

In sequence-based metagenomics, environmental sequences are screened for genes or gene fragments with homology to those encoding already described enzymes or conserved protein motifs of interest. Thus, compared to function-based metagenomics, a prior knowledge of the target DNA or protein sequence is required. Screenings are carried out in clone libraries using PCR amplification or hybridization techniques with target-specific probes or primers [Daniel, 2005]. A further sequence-based approach

utilizes *in silico* screenings for sequences in metagenome data that have similarity to a target gene.
The identified sequences can be custom-synthesized using a "synthetic metagenome" approach [Bayer et al., 2009]. Bayer *et al.* identified a putative methyl halide transferase by similarity searches in the NCBI database. The functionality of the synthesized fragment was then verified in expression libraries.

Sequence-based searches in metagenome data are shown to be a valuable approach to explore pathways, which might be important for understanding the conversion of renewable sources into biofuels [Warnecke et al., 2007, Pope et al., 2010, Hess et al., 2011]. Research projects aim to characterize efficient enzymes for the degradation of lignocellulosic biomass into biofuels for industrial-scale production. In nature, several enzymes and associated proteins are involved in the lignocellulose degradation. The enzymes are encoded by different microorganisms that convert biomass into energy. Enzymes important in this degradation process are glycoside hydrolases, a diverse family of carbohydrate active enzymes, and oxidoreductases [Evans and Furlong, 2011].
Enzymes involved in the lignocellulose degradation process were studied in metagenomes obtained from different habitats such as the termite gut [Warnecke et al., 2007], cow rumen [Hess et al., 2011] and foregut of the Tammar wallaby [Pope et al., 2010]. For detection, coding sequences predicted on assembled contigs were screened against specific glycoside hydrolases as classified by the databases Carbohydrate-Active enZYmes (CAZy) [Cantarel et al., 2009] and protein family (Pfam) [Finn et al., 2006]. The searches were performed using the Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1990] and profile hidden Markov models (HMMs) [Durbin et al., 2006]. Some of the identified glycoside hydrolases were confirmed by proteome analysis and *in vitro* activity tests.

2.3 Latest development of Meta-"omics"

Recently, other types of molecules from microbial communities have been studied, namely RNAs (transcripts) and proteins. The corresponding fields metatranscriptomics and metaproteomics together with metagenomics allow researchers to elucidate the composition and functions of a microbial community from a general perspective. This section gives an overview of further Meta-"omics" approaches.

2.3.1 Metatranscriptomics

Metagenomics gives insights into the taxonomic and functional potential of organisms in a selected habitat. However, metagenomics fails to separate expressed and non-expressed genes in an environmental sample. In metatranscriptomics, a collection of all expressed genetic information, the metatranscriptome, is analyzed. Metatranscriptomics addresses questions about active members and transcribed functions. Moreover, metatranscriptomics allows for studying the transcriptional responses to environmental changes.

Typically, DNA microarrays are used for RNA expression profiling. DNA microarrays are a suitable technique for studying the transcriptomic response of a single organism according to changes in environmental conditions or over different time points [Schena et al., 1998]. DNA microarrays have also been used to analyze several organisms at once [You et al., 2008, Bulow et al., 2008]. Two microarray types exist that are appropriate for the functional and transcriptomic analysis of the transcripts expressed by a whole community. The Geochip uncovers genes involved in various central processes [He et al., 2010b]. As the probe construction relies on the knowledge of the gene or genome sequences, DNA microarrays are not appropriate to discover novel enzymes. Moreover, the PhyloChip allows the comprehensive detection of bacterial and archaeal organisms residing in a microbial community [Brodie et al., 2007].

In metatranscriptomics, a collection of RNA molecules of an environmental sample is isolated (Fig. 2.3). The extracted RNA molecules are converted to double-stranded copy DNA (cDNA) using random primers for reverse transcription. Originally, cDNA libraries were constructed and randomly selected clones were sequenced in metatranscriptomics studies [Poretsky et al., 2005]. With the continuous advances of NGS methods, direct sequencing of the cDNA library is possible [Leininger et al., 2006, Frias-Lopez et al., 2008, Gilbert et al., 2008]. The analysis steps involved in metatranscriptomics are similar to the one in metagenomics.

So far, only a limited number of metatranscriptome experiments based on NGS have been performed. Most metatranscriptome projects comprise sequences without significant hits to any known gene sequence in the databases [Gilbert et al., 2008]. The first NGS-based metatranscriptome approach was performed on a sample obtained from soil [Leininger et al., 2006]. This analysis revealed that a key enzyme in the ammonia oxidation pathway is more abundant among archaeal than bacterial transcripts. Urich *et al.* analyzed the same soil sample and identified that only 8% of the metatranscriptome reads were assigned to known mRNA tags [Urich et al., 2008]. For the interpretation of the metatranscriptome data, Urich *et al.* developed a pipeline that identified rRNAs and mRNAs in two steps. The first molecules were utilized to deduce a taxonomic profile, while the latter molecules provide functional information.

A metatranscriptome approach has some limitations, which are absent in metagenomic-based studies. Ribosomal RNA (rRNA) and transfer RNA (tRNA) molecules are at relatively high levels in active cells [Kemp et al., 1993, Wagner, 1994], whereas messenger RNA (mRNA) molecules contribute only a small fraction of the transcripts. If the main aim of a metatranscriptome survey is to study the active functions, then either the enrichment of mRNA or the depletion of rRNA fragments is required. To overcome this, several strategies are available, for example, rRNA molecules are removed by selective hybridization or by digestion using exonucleases [Warnecke and Hess, 2009]. RNA is a highly unstable molecule with a rapid rate of turnover and a short cellular lifetime ranging from seconds to minutes [Poretsky et al., 2005] compared to the DNA molecule. Accordingly, expression profiles may be influenced by the stability of the RNA molecules [Velculescu et al., 1995]. A further limitation in metatranscriptome analysis is the retrieval of a low sample amount that complicates the extraction of enough RNA molecules [Amann et al., 1995].

2.3.2 Metaproteomics

In metaproteomics, the complete proteome of an environmental sample is studied. Metaproteomics is an emerging research field that aims at assessing the catalytic potential of a given microbial community [Simon and Daniel, 2011]. The metaproteome at a given time point is studied by two-dimensional polyacrylamide gel electrophoresis and mass spectrometry [Wilmes and Bond, 2004]. The drawback of metaproteomics is the low extraction yield and the lack of reference sequences in databases for functional assignments of protein fragments.

CHAPTER 3

State-of-the-art analysis of metagenome data

A growing interest in metagenomics resulted in the development of novel algorithms to accomplish tasks and challenges facing this research field. This chapter introduces the challenges and software tools for the interpretation of 16S rDNA amplicon sequences and whole metagenome shotgun data. As this thesis mainly deals with 454 pyrosequenced datasets, the main focus is on analysis methods for sequences obtained by the 454 technology. Computational methods and requirements for the analysis of data obtained by further next generation techniques are outlined in Chapter 8.

3.1 Methods for the analysis of 16S rDNA amplicon sequence data

16S rDNA amplicon sequencing is carried out to gain insights into the taxonomic composition and complexity of a microbial community. For this purpose, reads of 16S rDNA amplicons are clustered into phylotypes or operational taxonomic units (OTUs) [McCaig et al., 1999, Skirnisdottir et al., 2000] and taxonomically classified. Unfortunately, the analysis of 16S rDNA amplicon sequences is not straightforward, as problems are caused during 16S rRNA gene amplification and sequencing. Artifacts are generated that may lead to an inaccurate or misleading interpretation of the data, overestimation of the diversity or missing assignments of taxa. Different methods exist that identify artifacts and, hence, aid in the analysis of an underlying community.

3.1.1 Identification of artificial 16S rDNA amplicon sequences

Artificial sequences are generated as a consequence of PCR/sequencing nucleotide errors or co-amplification of 16S rRNA genes of different bacterial species. PCR errors are introduced by the polymerase that substitutes 1 base per $10^5 - 10^6$ bases [Cline et al., 1996]. Apart from this, sequencing errors generated by 454 pyrosequencing are estimated to be about 0.5% [Huse et al., 2007]. Consequently, both error rates would inflate the diversity estimation [Kunin et al., 2010]. In bacterial genome sequencing projects, errors are identified during the assembly step and eventually corrected. This is not possible in environmental analysis, since each identified read might originate from an individual organism. Accordingly, undetected errors would result in an overestimation of the diversity in a sample.

The single-linkage preclustering (SLP) algorithm [Huse et al., 2010] removes sequences that likely include pyrosequencing errors. The SLP algorithm presumes that unique sequences with a high occurrence in the dataset are accurate. Therefore, sequences are first ordered by the frequency of their uniqueness. The most abundant unique sequence initiates the first cluster. If any sequence of this cluster and a subsequent sequence in the ordered list have a pairwise distance less than 0.02, the sequence is added to the cluster. This is repeated for every unique sequence in the list. In a second step, less abundant clusters are compared to the larger clusters and merged together if the sequences differ by less than 0.02.

AmpliconNoise [Quince et al., 2009, Quince et al., 2011] identifies 454 sequencing errors by clustering original flow signal intensities. In addition, the sequences are clustered for removing sequences with PCR errors. As the analysis relies on calculations of pairwise alignments for each sequence pair, AmpliconNoise requires high computational resources [Schloss et al., 2011]. Because of this, SLP is more frequently applied in 16S rDNA amplicon studies [Mattila et al., 2012, Zhao et al., 2012, Biesbroek et al., 2012].

Another problem leading to an overestimation of the species diversity is the formation of chimeric sequences during PCR. Chimeric sequences are comprised of two or more phylogenetically distinct species [Lahr and Katz, 2009]. The chimeric fragments are then sequenced and interpreted as reads originating from an individual organism. The rate of chimeric sequences in 16S rDNA data is assumed to range from 5% to 45% [Schloss et al., 2011]. In general, it is estimated that at least one in twenty 16S rRNA gene sequences in public databases contains such anomalies [Ashelford et al., 2005].

Various tools were implemented to recognize 16S rDNA chimeras [Maidak et al., 2001, Huber et al., 2004, Ashelford et al., 2005, Gonzalez et al., 2005, Haas et al., 2011], but they were initially developed for the identification of chimeras in full-length 16S rRNA genes. Recently, the tools Perseus [Quince et al., 2011], DECIPHER [Wright et al., 2012] and UCHIME [Edgar et al., 2011] were developed for the detection 16S rDNA chimeras in short sequences.

Perseus [Quince et al., 2011] exploits the abundances of unique sequences. The amplicon query is pairwise compared to all sequences that have a higher abundance. The closest pair is selected and an alignment is calculated. Finally, chimeras are removed using supervised learning.

DECIPHER [Wright et al., 2012] first applies the RDP Classifier [Wang et al., 2007] (Section 3.1.3), which assigns the query to a taxonomic group. A query is classified as a chimera, if it has uncommon segments compared to the sequences within the taxonomic group but which are common for another taxonomic group.

UCHIME [Edgar et al., 2011] divides a query sequence into four non-overlapping segments and searches for each of the segments a matching reference (parent) in a database that is assumed to contain no chimeric sequences. If no reference database can be provided, UCHIME can be used to detect chimeras *de novo*. In this case, the 16S rDNA amplicon reads constitute the database, and it is assumed that a chimera has undergone fewer rounds of amplification than its parents. The best two hits of the four segments are determined and subsequently aligned with the query. Based on the alignment, a score is calculated for discrimination of whether the two hits are candidate parents of the query or not. The parents are only valid, if they have a higher abundance than the query. UCHIME yields results comparable to Perseus [Edgar et al., 2011] and outperforms DECIPHER in detecting chimeras in sequences ranging from 100 to 600 bases [Wright et al., 2012].

3.1.2 Clustering of 16S rDNA sequences to operational taxonomic units for diversity analysis

16S rDNA amplicon sequences are usually clustered into OTUs that are formed based on similarities of the sequences to each other. OTUs are associated with taxonomic levels according to the applied identity thresholds. Typically, genera and species are equated with an identity of 95% and 97% in 16S rRNA gene analysis, respectively [Schloss and Handelsman, 2005]. However, there are no universal definitions for the value reflecting the rank of species, as in some studies other thresholds are selected [Bonnet et al., 2002, Lin et al., 2012]. A further problem in OTU determination is the choice of the clustering method. Different OTU clustering methods or parameterizations can lead to different OTU estimations of the same analyzed sample [White et al., 2010].

MOTHUR [Schloss et al., 2009] has been developed to calculate clusters based on the nearest, the furthest and the average neighbor clustering algorithms. The clustering methods require distance matrices that are calculated for the aligned 16S rDNA amplicon sequences. However, recent concerns have arisen regarding the choice of the alignment methods, as different alignment methods result in distinct distance matrices and by that to a misestimation of the diversity [Schloss, 2010].

Unfortunately, the distance-based clustering approach is time and memory consuming. Therefore, fast sequence clustering algorithms have been developed to cluster 16S rRNA genes into OTUs without an initial multiple sequence alignment. In UCLUST [Edgar, 2010], sequences are sorted by their decreasing length. The algorithm works as follows: Initially, the UCLUST database for the storage of seed sequences is empty. If a sequence matches a seed sequence in the database, it is added to the cluster represented by the seed, otherwise a new cluster is established in the database with the sequence as a seed.

The high-throughput feature of NGS techniques provides access to the microbial "rare biosphere" [Sogin et al., 2006], which is constituted by low-abundant species. The existence of the rare biosphere has been exhaustively discussed in the light of potential errors (sequencing errors, chimeras) introduced by the 454 sequencing technique or PCR [Reeder and Knight, 2009, Kunin et al., 2010, Agogué et al., 2011]. Overall, 50% of obtained OTUs are represented only by a few or one single sequence [Zinger et al., 2011]. Therefore, low-abundant OTUs are suspected to be artifacts and are suggested to be removed from the downstream analysis [Reeder and Knight, 2009, Zhou et al., 2011]. Other researchers have successfully assigned low abundant OTUs to taxa and illustrated the importance of the rare biosphere in analyses of a microbial community [Galand et al., 2009, Agogué et al., 2011].

A further task in 16S rDNA analysis is to ascertain how well the sequences reflect the richness of an underlying community. A common method is to estimate the number of observed new OTUs with increased sampling [Tringe et al., 2005]. This accumulation can be projected in a rarefaction curve. A gentle rarefaction curve illustrates that the sample is well covered by the number of sequences, whereas a steep slope indicates that more sequences are required to cover all taxa.

3.1.3 Taxonomic assignments of unknown 16S rDNA amplicon sequences

Most commonly, the Ribosomal Database Project (RDP) Classifier [Wang et al., 2007] is used to assign unknown 16S rDNA or rRNA sequences into taxonomies. The classifier works well on partial or full-length sequences and does not require alignments. Instead it is a composition-based method that uses reference sequences to characterize unknown sequences to taxa from domain to genus. The reference sequences are acquired from the RDP database [Cole et al., 2003], which includes the data based on Bergey's Taxonomic Outline of the Prokaryotes [Garrity and Lilburn, 2004].

Briefly, all k-mers, by default 8-mers, in a training set of known taxa are calculated. The k-mers are used to train a Naïve Bayesian Classifier (NBC). Afterwards, the NBC is used to assign an unknown sequence based on its 8-mers to the closest matching genus. For a proper assignment of the queries, the RDP Classifier requires a query length of at least 50 bases. Bootstrap confidence estimates are provided for each assignment to evaluate the predictions. Therefore, randomly 1/8 of the k-mers of the query are chosen and classified *via* the NBC. This procedure is iterated 100 times. The number of times that the same classification is calculated is assigned as the confidence value. Typically classifications are selected that exceeded the RDP Classifier confidence threshold of 0.8.

3.1.4 Full pipelines for the analysis of 16S rDNA amplicon sequences

The RDP Classifier is embedded within the Ribosomal Database Project's (RDP) pyrosequencing pipeline. In addition, the RDP's pyrosequencing pipeline includes primer/MID trimming, chimeric sequence detection and automated alignment generation of the query sequences [Cole et al., 2003]. Finally, RDP offers a database of aligned 16S rRNA genes, which is regularly curated. Still, the database contains sequences that are not well-defined on lower taxonomic ranks because of the difficulty to culture the corresponding strains.

Table 3.1: Tools for the analysis of 16S rDNA amplicon data

Tool	Reference	Description
SLP	[Huse et al., 2010]	454 pyrosequencing error correction
AmpliconNoise	[Quince et al., 2009, Quince et al., 2011]	PCR/sequencing error identification and correction
UCHIME	[Edgar et al., 2011]	Chimera detection
DECIPHER	[Wright et al., 2012]	Chimera detection
Perseus	[Quince et al., 2011]	Chimera detection
UCLUST	[Edgar, 2010]	OTU clustering
RDP Classifier	[Wang et al., 2007]	taxonomic analysis of 16S rRNA genes
QIIME	[Caporaso et al., 2010]	Software package including several tools for sequence processing
RDP's pyrosequencing pipeline	[Cole et al., 2003]	Online available pipeline for the analysis of 16S rRNA genes
MOTHUR	[Schloss et al., 2009]	Software package for the processing of 16S rDNA genes
ESPRIT	[Sun et al., 2009]	Sequence processing and diversity assessment

Further pipelines including some of the aforementioned steps are provided by MOTHUR [Schloss et al., 2009], QIIME [Caporaso et al., 2010] and ESPRIT [Sun et al., 2009]. The software MOTHUR supports trimming of the sequences, chimera detection using various methods including UCHIME, an algorithm similar to SLP, classification by the RDP Classifier and clustering methods based on alignments and distance matrices. QIIME [Caporaso et al., 2010] is able to perform some downstream analyses of the data including trimming, chimera detection using Perseus, clustering with UCLUST and taxonomic assignments based on the RDP Classifier. ESPRIT [Sun et al., 2009] provides sequence processing and clustering based on pairwise sequence alignments. The introduced tools are listed in Table 3.1.

3.2 Methods for the annotation of whole metagenome shotgun data

Metagenomics is aimed at unveiling taxonomic compositions by assigning taxonomies to the metagenome reads and deducing the functions encoded on the fragments or assembled contigs. With the establishment of the metagenomics field, microbiologists were faced with novel tasks for the interpretation of the data. Next-generation sequencing techniques have boosted both the number of metagenome projects and throughput making the analysis of the data more challenging. Even though metagenomics is a relatively new scientific field, several computational methods exist and novel tools are regularly published. The algorithms' focus is to tackle such voluminous data, to assemble the sequences into contigs and to unveil the taxonomic structure as well as functional pathways of a heterogeneous microbial community.

Some software tools are provided for download only and, hence, users have to perform the computational tasks with their own compute resources. In some cases, a web interface is provided allowing the execution of time-consuming calculations on remote compute resources. In the following section, tools for the analysis of whole metagenome shotgun data obtained by 454 pyrosequencing will be introduced. Moreover, an overview of the tools is given in Table 3.2.

3.2.1 Methods for the assembly of short sequences

In genomics, the genome of a single organism is sequenced. Thus, each read and assembled contig belongs to the same genome. After reconstruction of the genome by ordering the reads and contigs, genomic elements are identified and functionally characterized. Overall, the genomic approach likely gives a complete picture of the capabilities of a single organism.

Initial metagenomic studies [Venter et al., 2004] used tools that were developed for the analysis of single genomes. Slightly modified versions of the genome assemblers Celera Assembler [Myers et al., 2000] and JAZZ [Aparicio et al., 2002] were applied for the reconstruction of contigs from a set of metagenome reads obtained by Sanger sequencing. Thereafter, gene prediction was carried out on the contigs prior to the functional annotations of the coding sequences.

With the introduction of NGS technologies, novel assemblers for short reads, such as Roche's GS *De Novo* Assembler, were developed [Pop, 2009]. The assemblers were mainly designed for single genome annotations, but have been also utilized for metagenome assembly [Eloe et al., 2011, Delmont et al., 2012]. Due to the high species diversity, simultaneous assembly of several genomes is more challenging than the assembly of a single genome. The danger of assembling sequences from a microbial community is the generation of interspecies chimeras [Mavromatis et al., 2007]. One requirement for a proper assembly is a sufficient read coverage of the genomes. Thus, regions with low coverage are returned as singleton reads, which in turn are not useful for assembly and excluded from the subsequent analysis. To deduce the taxonomic origin and encoded functions from all short sequences, read-based analysis methods were implemented, which are listed in the next sections.

3.2.2 Taxonomic classification of metagenomic data

RDP (Section 3.1.3) can be used for classification of 16S rRNA gene fragments extracted by similarity searches from metagenome datasets. However, metagenomes contain only a small amount (0.071% - 0.17%) of 16S rRNA genes [McHardy and Rigoutsos, 2007] and therefore, the results offer only limited insights into the microbial community. The set of 16S rRNA gene sequences may neither fully represent the taxonomic composition nor give functional insights into the microbial community. Further approaches are focusing on taxonomic classification of fragments encoding genes, also termed environmental gene tags (EGTs).

MEGAN [Huson et al., 2007, Mitra et al., 2009, Mitra et al., 2011, Huson et al., 2011], a stand-alone metagenome analysis tool, is based on sequence homology, which is determined by searching in reference databases of known genes or proteins using the Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1990]. Since MEGAN requires a BLAST output file, the user has to perform the compute-intensive BLAST searches. MEGAN then calculates a lowest common ancestor (LCA) for reads with multiple BLAST hits. The LCA approach assumes that sequence similarities are the result of evolutionary developments of genes in different species over time. If a read fragment matches sequences of different origins, an LCA of the taxonomies is computed in MEGAN and finally assigned to the read. To restrict the number of hits considered for LCA calculation, hits are only used that have a bit score larger or equal than 90% of the bit score of the best BLAST hit. After LCA calculation, the results can be viewed in the graphical interface in MEGAN.

A further tool, CARMA [Krause et al., 2008b], assigns environmental sequences to taxonomic groups based on similarities to conserved protein families and domains included in the protein family databases (Pfam) [Finn et al., 2006]. The classification into a higher-order taxonomy is based on the reconstruction of a phylogenetic tree for each matching Pfam family. CARMA3 [Gerlach and Stoye, 2011] is the succeeding application of

CARMA and is based on reciprocal BLAST searches against the non-redundant protein database (nr). For CARMA3, the web interface WebCARMA [Gerlach et al., 2009] is available, which allows for computing the taxonomic and functional profiles for metagenome data up to 100 MB. In a comparative analysis, CARMA3 shows better results than CARMA and MEGAN [Gerlach and Stoye, 2011]. Both, CARMA and MEGAN belong to the similarity-based approaches that are restricted to knowledge of known reference sequences in databases. Similarity-based approaches assign taxa only to sequences that have a homologue in the databases, whereas sequences without homology remain unclassified.

Contrary to similarity-based approaches, composition-based approaches utilize intrinsic sequence features, such as GC content, codon usage or oligonucleotide frequencies, which vary among organisms [Bentley and Parkhill, 2004]. Taking this characteristic features into account, metagenome sequences can be clustered into different bins. Tools employing composition-based approaches are for example PhyloPythia [McHardy et al., 2007, Patil et al., 2011] and TACOA [Diaz et al., 2009] (Tab. 3.2). PhyloPythia uses support vector machines (SVMs) [Hastie et al., 2003] for the assignment of sequences to taxa. The classifier TACOA combines a k-nearest neighbor approach [Cover and Hart, 1967] with kernel-based learning [Hastie et al., 2003] to assign genomic fragments based on their oligonucleotide frequencies to taxa. However, a requirement for a reliable assignment to a taxon using composition-based approaches is the availability of long reads or contigs (at least 800 bases).

3.2.3 Functional characterization of metagenomic reads

Compared to the taxonomic tools, less functional tools for metagenome short sequences are published, and it is still challenging to assign functions to a read because many protein families and functions are unknown [Gilbert et al., 2008]. Functional annotation relies on similarity searches of metagenome reads against annotated sequences in currently available databases. Databases containing functionally characterized sequences obtained from genome-based microbial studies are consequently biased towards cultivable organisms. Therefore, the databases represent only a partial picture of microbial genomes and their biological functions.

The short sequence reads are interpreted as genes encoding fragments of proteins, folds or domains. A metagenome read may carry genes encoding highly variable regions making the functional assignments even more challenging. However, the reliable functional assignment of short reads is a main step in the interpretation of metagenomes, since it is the basis to discover genes of interests and address specific biotechnological questions [Chistoserdova, 2010].

A common approach is to perform BLAST searches against gene or protein databases to predict COG categories [Tatusov et al., 2001], FIG families [Overbeek et al., 2005], KEGG numbers [Kanehisa and Goto, 2000], Pfams [Finn et al., 2006] and other functional categories. A limitation of this approach is that the majority of the proteins have not been experimentally characterized [Baxevanis and Ouellette, 2004], instead the annotations are transferred from homologous sequences based on similarity searches. Thus, wrong annotations can be derived from the databases.

3.2.4 Full pipelines for the analysis of metagenome data

Web-based annotation platforms such as the metagenomics RAST (MG-RAST) server [Meyer et al., 2008], the IMG/M server [Markowitz et al., 2006] and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) [Sun et al., 2011] have been developed to store and analyze metagenomic data (Tab. 3.2).

In 2008, MG-RAST was released, which allows taxonomic and functional analysis based on BLAST against different databases such as SEED [Overbeek et al., 2005], RDP [Cole et al., 2003] and Greengenes [DeSantis et al., 2006]. Additionally, the metabolic abilities of a community are provided by KEGG pathways. Initially, the MG-RAST platform calculated taxonomic profiles by assigning the taxonomy of the best hit obtained by BLAST searches against the SEED and 16S rRNA gene databases. The best BLAST hit approach provides no information of the phylogenetic distance of the query to the reference sequence. Thus, the assignments should be interpreted with caution, in particular on lower taxonomic levels such as genus or species. Meanwhile, MG-RAST has been improved vastly. The best BLAST hit approach is complemented by the LCA algorithm. The BLAST searches are performed against numerous databases including GenBank [Benson et al., 2011], RefSeq [Pruitt et al., 2009], UniProt [Apweiler et al., 2011] and eggNOG [Muller et al., 2010]. MG-RAST features a pipeline for the prediction of genes on contigs and their subsequent annotation. The protein sequences are available for download. However, the genes encoding the regions are not visualized on the contigs. It is also not possible to perform user-specific searches for target genes using BLAST or profile HMMs, as it has been applied in screenings for sequences with possible industrial applications (Section 2.2.3). Finally, MG-RAST is a static system, as it does not support an easy integration of novel tools. MG-RAST does not allow users to analyze function in the context of taxonomy and *vice versa*, which is possible by combining the results obtained from CARMA.

IMG/M uses BLAST searches to determine KEGG pathways, COG functional categories and Gene Ontology (GO) [Ashburner et al., 2000] assignments. The taxonomic characterization is deduced from sequence comparisons to individual genomes based on the best BLAST hit approach.

CAMERA has been initially implemented to store data and results of the Global Ocean Sampling obtained the by J. Craig Venter Institute [Venter et al., 2004]. The software provides analysis tools, which can be linked together into a user specific workflow. The functional annotation workflow of the data is based on BLAST searches against the Pfam, Tigrfam [Haft et al., 2001] and COG databases. A workflow for taxonomic assignments is composed of the RDP Classifier or BLAST against ribosomal RNA databases. In CAMERA, RAMMCAP [Li, 2009] is embedded, which uses clustering algorithms to cluster translated open reading frames (ORFs) by high sequence similarity. The clustering step reduces the data complexity and subsequent computational efforts. For functional annotations, a representative of each cluster is compared to sequences in Pfam, Tigrfam and COG databases. RAMMCAP is also available as a stand-alone tool.

To conclude, several automated pipelines exist that provide either taxonomic or functional profiling of the microbial community. However, metagenomic projects occasionally aim to associate functions with taxonomic groups. Therefore, functional profiles for specific taxa and *vice versa* are desirable. The systems lack the possibility to integrate novel algorithms fast and easily. This is very important, as novel tools are continuously published that improve the taxonomic predictions of reads. In addition, the pipelines perform gene predictions on contigs and functional annotations of the identified genes. The annotations are only listed in functional profiles and the protein sequences are available for download. For improving annotations or identification of target gene clusters, a view of the complete contig with the encoded genes is required. Moreover, user-specific searches using BLAST or profile HMM are necessary to identify reads encoding enzymes or domains with industrially relevant functions.

Table 3.2: Tools for the analysis of whole metagenome shotgun data

Tool	Reference	Description
MEGAN	[Huson et al., 2007, Mitra et al., 2009, Huson et al., 2011]	Similarity-based approach for taxonomic classification of metagenome reads based on LCA assignments; functional assignments based on the KEGG, COG, SEED databases
CARMA	[Gerlach and Stoye, 2011]	Similarity-based approach for taxonomic and functional classification of metagenome reads based on a reciprocal BLAST approach and homologies to the Pfam database
PhyloPythia	[McHardy et al., 2007, Patil et al., 2011]	Composition-based approach for taxonomic classification of large sequences
TACOA	[Diaz et al., 2009]	Composition-based approach for taxonomic classification of large sequences
MG-RAST	[Meyer et al., 2008]	Metagenome annotation software for the storage and analysis of metagenome data
CAMERA	[Sun et al., 2011]	Data repository and bioinformatics tool resource for metagenomic analysis
IMG/M	[Markowitz et al., 2006]	Storage and functional analysis of metagenome data

CHAPTER 4

Motivation and aims of the thesis

Microorganisms are relevant in biotechnological, medical and agricultural processes. Knowledge of taxonomic and functional characteristics of the natural microbial communities would improve the understanding and controlling of these processes. Unfortunately, the majority of microbes cannot be accessed and analyzed using conventional methods. Advances in sequencing technologies provide the opportunity to study the entire genetic make-up of microbial communities in terms of their taxonomic and metabolic potential. At the same time, the high-throughput feature of the sequencing technologies makes the storage and management of the data challenging. Moreover, tools for the interpretation of the data are continuously published. Therefore, novel methods are required that automatically apply the existing tools in order to deduce information relevant for understanding the functioning of complex communities in their natural habitats.

In this thesis, the design and development of methods are demonstrated that allow the interpretation of whole metagenome shotgun, 16S rDNA amplicon and metatranscriptome data. These novel methods should complement the computational methods for PolyOmics data analysis, which are provided at the Center for Biotechnology (CeBiTec). The first method should allow the interpretation of the huge amounts of metagenome data. In this regard, a metagenome platform is required that enables simple and automated processing as well as analyses of metagenome data in terms of the functional and taxonomic assignments. Next, a method for the analysis of 16S rDNA amplicon sequences is required, which deduces the taxonomy as well as the diversity in a complex microbial community and solves the challenges in a 16S rDNA amplicon study accurately and efficiently. Finally, a workflow for the analysis of metatranscriptome data is demanded that captures all relevant RNA types within

metatranscriptome data in order to assess the active taxa and expressed functions encoded by a microbial community.
After the successful realization of the methods, their capabilities and results should be proven. In this regard, metagenome, 16S rDNA amplicon and metatranscriptome data obtained from a production-scale biogas plant are studied. The knowledge of organisms and their functions in the biogas production process is of fundamental importance, as methane, a component of biogas, can be converted into electricity or heat. The aim of the analysis is to identify taxa and pathways that are relevant in the biogas plant by using the novel methods.

A final goal is the screening for industrially relevant enzymes in metagenome data. In the focus of this thesis are laccases, which are important in the pulp processing and bleaching industry due to their ability to degrade lignin. For this purpose, a method should be developed that facilitates searches for genes encoding putative laccases in metagenome data.

To summarize, the major aims within this thesis are as follows:

1. Design and implementation of a metagenome platform that unveils the taxonomic and functional potential of a heterogeneous community.
2. Design and implementation of a pipeline for the analysis of 16S rDNA amplicon data that provides deeper information about the taxonomic composition of a natural microbial community.
3. Design and implementation of a pipeline for the analysis of metatranscriptome data that allows identification of the active members and their transcripts in a microbial community.
4. Application of the methods to examine the taxonomic and functional characteristics of a biogas-producing microbial community.
5. Identification of genes encoding industrially important enzymes, for example laccases, in metagenome data.

CHAPTER 5

Methods and implementation

This chapter is divided into four sections. In the first section, the requirements and implementation of a novel metagenome analysis platform are presented. In the second section, steps involved in the analysis of 16S rDNA amplicon data are described in detail. A workflow for the interpretation of metatranscriptome data is introduced in the third section. Finally, the fourth section describes a method for the discovery of enzymes that are potentially applicable in the biotechnological field.

5.1 The novel platform MetaSAMS for the analysis of metagenome data

The overview of the available analysis platforms for whole metagenome shotgun data (Section 3.2) indicates that the most common platforms generate taxonomic profiles based on the best BLAST [Altschul et al., 1990] hit approach or LCA calculations [Huson et al., 2007] of multiple BLAST hits. Nevertheless, promising taxonomic tools are published continuously. As they produce more accurate taxonomic predictions, the assignments should also be considered in the taxonomic analysis. However, the system design of the available platforms does not allow a straightforward integration of novel tools. In addition, the available platforms do not provide a profile combining functional and taxonomic assignments. Above all, no system exists yet that captures genes encoding a desired function. However, optimal enzymes are required to accomplish specific industrial processes under extreme conditions.

Because of these limitations in existing platforms, a novel system was required. The basic requirements of a metagenome platform include:

- the import and storage of metagenome reads and contigs,
- the storage of functional and taxonomic results,
- the availability of different projects with secured and authenticated user access,
- flexible pipelines for taxonomic and functional analysis,
- the support for visualizations of the taxonomic and functional results,
- the integration of comparative analyses and their visualizations.

Metagenonomics produces data in a high-throughput manner. Consequently, robust data processing and fast evaluation strategies need to be realized to cope with the increasing data amount. The Sequence Analysis and Management System (SAMS) [Bekel et al., 2009] has been originally developed for quality control in whole genome shotgun projects, for the automated analysis of expressed sequence tags (ESTs) and copy DNA (cDNA) data generated by Sanger sequencing. In general, SAMS has been applied to analyze EST projects with $45,000 - 235,000$ ESTs [Bekel et al., 2009]. In contrast, metagenome data produced in one run on the Genome Sequencer FLX+ system contain 1,000,000 reads (Tab. 2.1). Clearly, the original version of SAMS is not suitable to examine such amounts of data. In addition, SAMS lacks tools that address metagenome-specific tasks. However, an advantage of the SAMS system is that basic tools, such as BLAST, are already available. The modular design of SAMS allows the integration of further tools that may be relevant for the interpretation of metagenome data. Because of these advantages of the SAMS platform, it is adapted for the analysis of metagenome data in collaboration with Thomas Bekel. In the following, the metagenome platform, termed MetaSAMS, will be introduced in regard to the realizations of the mentioned requirements. The structure of this section follows the workflow from processing and analysis of raw metagenome data to visualizations of taxonomic and functional annotations.

5.1.1 Design

Since MetaSAMS has been implemented as an extension of SAMS, the general architecture of MetaSAMS is basically the same. It is based on a three-tier approach that embeds the presentation layer, database layer and business logic layer (Fig. 5.1). Similar to SAMS, MetaSAMS is available *via* a web interface, which is based on Perl Computer-Generated Imagery (CGI) scripts running on an Apache server. The CGI scripts dynamically generate Hypertext Markup Language (HTML) content. The interactivity is provided through JavaScript and 'asynchronous JavaScript and XML' (AJAX). The user interface enables the access to the results of the analysis.

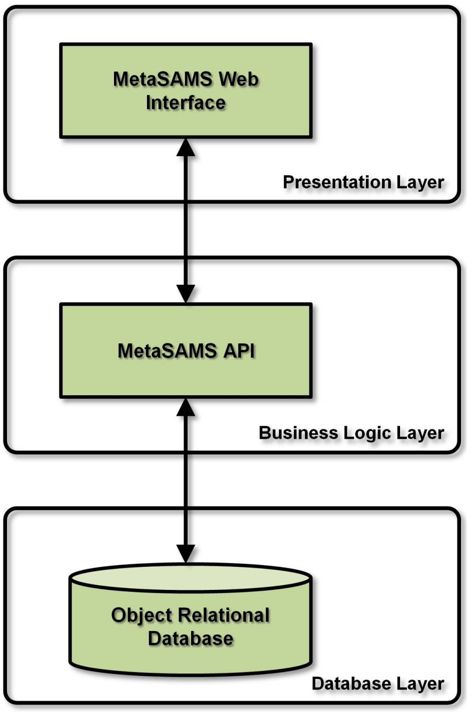


Figure 5.1: The three-tier architecture of the MetaSAMS platform: The general architecture of MetaSAMS covers the web frontend, business logic and data backend.

For the data storage, MySQL is used as a relational database management system (RDBMS). Access to the data is implemented by using the O2DBI software (unpublished) that provides an automatic object relational mapping. The user interface and the database are connected *via* the documented application programming interface (API). The API is the basis of the business logic layer and supplies the core functionalities of the MetaSAMS system.

The security of the data is of high importance for data privacy and acceptance of the software. As MetaSAMS is available through a web interface, access to the application and stored data are controlled by a security module. The business layer enables access to MetaSAMS projects by the generalized project management system (GPMS), which is commonly used in various bioinformatics software packages at Bielefeld University [Meyer et al., 2003, Neuweger et al., 2008, Dondrup et al., 2009]. The GPMS follows a role-based approach. For each user a role can be assigned within a project. This role defines the access rights and, hence, restricts the actions that can be performed by a user within a specific project. As an example, the role "Guest" has the right to view the results of the metagenome data but is not allowed to rerun pipelines or delete data.

The authentication to a project is facilitated on the login page over the web frontend of the MetaSAMS system.

5.1.2 Data model

The data model for the storage of metagenome reads and their functional and taxonomic results was adapted by Thomas Bekel. Briefly, all reads obtained from a sequencing procedure are joined together to a single entity termed "ReadSet". Contrary to SAMS, the sequences are stored in a file and not individually in the database. The next modification deals with the storage of the results for each read. Within the original SAMS system, tool results are modeled as individual "Observations" consisting of many attributes like species names or protein functions. This approach is suitable for smaller datasets, such as ESTs, with less abundant results but has its limits in the analysis of large metagenome datasets, which generate highly redundant results. MetaSAMS solves the redundancy by storing each observed feature, like a specific taxonomic classification, only once. At the end the individual reads are linked to the corresponding results.

In this thesis, new classes are designed for the storage of regional and functional annotations of metagenome contigs. The data model for the analysis of metagenome contigs is basically derived from GenDB [Meyer et al., 2003], which is an annotation pipeline for bacterial genomes. In GenDB, the class "Region" represents arbitrary sequences. The class "Contig", which inherits the class Region, reflects a bacterial genome, genome part or replicon. Predicted coding sequences are stored as "CDS" objects in reference to the contig. The concept is not capable to analyze metagenome contigs, as an analysis tool, e.g., a gene prediction tool, would have to be called for each contig. In addition, partial genes located at the end of a contig would not be recognized by gene prediction tools due to the lack of a start or stop codon. To avoid the generation of multiple jobs and allow the identification of terminal genes, a new concept for the characterization of metagenome contigs was designed.

A schematic overview of the designed classes and their interactions in MetaSAMS is given in Figure 5.2. An artificial contig, herein termed "supercontig", is generated by alternately concatenating a sequence of an assembled contig and the linker sequence "CATAGCATAGCATAGCTATGCTATGCTATG", which consists of start and stop codons in all possible six reading frames. An advantage of the concatenation is that subsequent analysis tools are only executed once on a supercontig instead on each assembled contig. A supercontig belongs to the class Contig. An object Contig stores the artificial sequence of the supercontig. To distinguish between the artificial contig and assembled metagenome contig in this thesis, the first will be referred to as supercontig, while the latter will be termed contig or, after import into MetaSAMS, metatig. To identify the assembled contigs and the linker in the supercontig sequence, objects of "Metatigs" representing assembled contigs and objects of "Linkers" describing linkers are stored in relation to the supercontig sequence in the database.

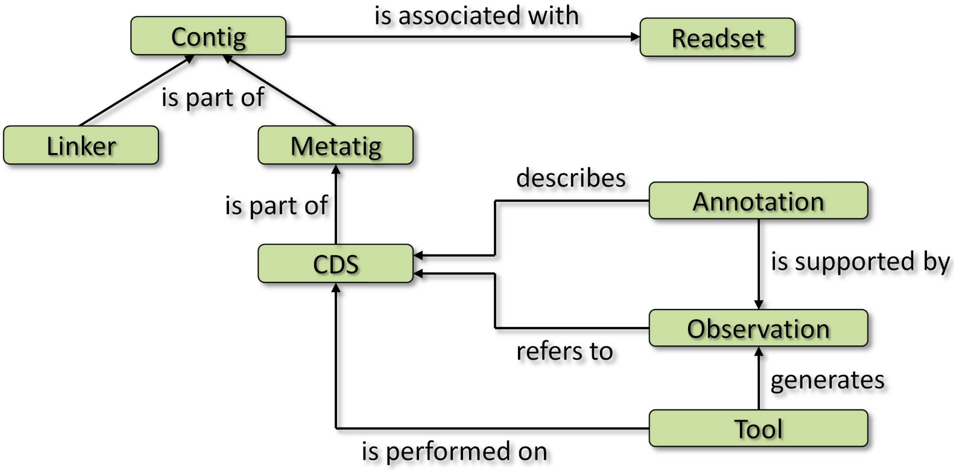


Figure 5.2: A schematic representation of the main data objects for the storage of contigs and their annotations: The data model in MetaSAMS is basically derived from GenDB. To allow the analysis of contigs, the novel classes Metatig and Linker are integrated.

5.1.3 Importer

Two importers, one for reads and the other for metagenome contigs, are implemented to load the metagenome data into MetaSAMS. Each importer is provided by a Perl command-line script. The project name of the database is supplied in the command as a parameter, in order to access the MetaSAMS database using the O2DBI API. The importer for metagenome reads has been implemented by Thomas Bekel. Basically, it imports except of the sequences all read information present in a fasta file into the database and connects each read to the corresponding readset. The sequences are then stored in a fasta file.

For the import of contigs, metagenome reads are first assembled by using assembly tools for short reads such as the GS *De Novo* Assembler [Pop, 2009]. Since prokaryotic genes are on average approximately 1000 bases long [Xu et al., 2006], the importer removes contigs smaller than 500 bases by default. The remaining contigs are then grouped by their GC content to ease gene prediction (Section 5.1.6). An artificial supercontig is generated by alternatively concatenating a contig sequence from the sorted list and a linker sequence containing start and stop codons. Finally, the supercontig is stored in the database and linked to at least one readset (Fig. 5.2). The last information enables the user to generate several assemblies based on different combinations of readsets.

The GS *De Novo* Assembler generates an ace file storing information relating to the assembly. The importer of MetaSAMS utilizes the file to associate the reads that were used for an assembly with the corresponding metatig.

5.1.4 Tool concept

As described in Section 3.2, different software tools are available for taxonomic and functional analysis. Still, novel methods are steadily published that improve the taxonomic classification. Similar to SAMS, MetaSAMS has a modular tool concept that allows integrating novel tools easily. A "Tool" class represents a software tool with its executable path, parameters and type of input data. Each tool is defined by parameters, which are usually entered by command-line. MetaSAMS allows flexibility in the parameter settings by storing several parameterizations. Thus, results generated by different tool settings can be obtained.

A tool in MetaSAMS is restricted to the input data, for example, a tool for taxonomic assignments of reads is only applicable on readsets, whereas a tool for gene predictions can be executed on objects representing supercontigs. Therewith, MetaSAMS avoids false applications of tools and supports the generation of pipelines. Several tools combined form a pipeline. New analysis pipelines have been implemented to allow the annotation of metagenomic data. MetaSAMS provides two major pipelines, one for the analysis of metagenome reads and another for metatigs (Fig. 5.3). The first pipeline computes taxonomic assignments and functional characterizations for each read (Section 5.1.5). It requires a readset object as input data. In contrast, the second pipeline is applied on a supercontig object and generates regional predictions of the supercontig or functional assignments to the genes (Section 5.1.6).

Due to the high amount of data, large computational requirements with respect to runtime and memory consumption can be expected. Therefore, the submission of the computational jobs to a compute cluster allows a scalable and efficient analysis. MetaSAMS utilizes the Distributed Resource Management Application API (DRMAA) for the submission and control of jobs to Distributed Resource Management Systems (DRMS). This approach allows the analysis of high-throughput metagenome data in appropriate time. After the submission, a "Job" represents a combination of a specific tool and applied region. The Job status summarizes the computational progress. Moreover, a Job stores possible error and warning messages.

5.1.5 The pipeline for the analysis of metagenome reads

In MetaSAMS, three taxonomic classifiers are integrated to compute taxonomic assignments of metagenome sequence reads: the Lowest Common Ancestor (LCA) approach based on multiple BLAST [Altschul et al., 1990] hits of a read, the RDP Classifier

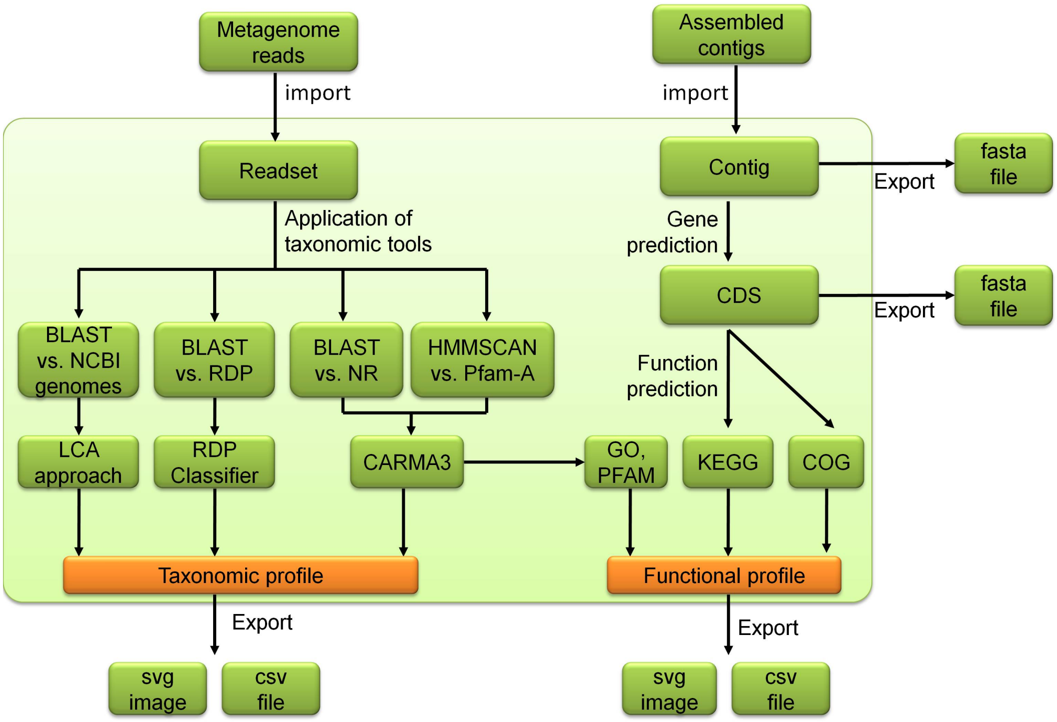


Figure 5.3: Workflows for the analysis of metagenome reads (left) and contigs (right): The pipeline based on reads basically consists of three tools that generate taxonomic assignments. CARMA3 also supports functional predictions in terms of GO and PFAM information. Additional functional annotations are provided by the pipeline based on assembled contigs that includes a gene and function prediction module.

[Wang et al., 2007] and CARMA3 [Gerlach and Stoye, 2011]. Hence, MetaSAMS provides taxonomic profiling based on environmental gene tags (EGTs) as well as on phylogenetic marker genes (16S rRNA). Since the taxonomic results are linked to the tool and read, they are reproducible and transparent. On account of the modular implementation of MetaSAMS, novel tools can be easily added to the pipeline.

The procedure for the classification of metagenome 16S rRNA genes is basically composed of two steps, namely the detection of 16S rRNA gene fragments and their taxonomic classification. In the first step, reads carrying fragments of 16S rRNA genes are identified using BLAST searches against the RDP database [Cole et al., 2003] with an E-value threshold of 10^{-10} and by disabling the low complexity filter. The module "IterateFasta" is used for the BLAST search to improve the performance. It retrieves the executable BLAST command and the sequences, which are split into several subsets. After that, it submits jobs for each subset to the compute cluster. The results are merged

and returned to the tool, which determines the best BLAST hit and stores it in the database.
In the second step, sub-regions of reads with significant BLAST hits to the RDP database are extracted. Only reads larger than 50 bases are collected and the RDP Classifier is executed, which supplies taxonomic assignments with confidence estimates from superkingdom to genus for each read. Since MetaSAMS stores each confidence value, individual profiles with user-defined thresholds for the confidence value are retrievable *via* the web interface.

As the fraction of fragments that carry a 16S rRNA gene in the whole metagenome shotgun data typically is very low [McHardy and Rigoutsos, 2007], further tools based on genes encoding protein sequences are integrated into MetaSAMS. CARMA supports taxonomic assignments and functional characterizations of environmental gene tags (EGTs) in metagenome sequences. In MetaSAMS, CARMA3 is executed, and the taxonomic results and gene functions based on Pfam accession numbers [Finn et al., 2006] and GO terms [Ashburner et al., 2000] are stored in the database.

The LCA-based approach is composed of two relevant steps. First, a BLAST search of the reads against genomes obtained from the NCBI bacterial database is performed using the aforementioned IterateFasta module. Second, the taxonomy of reads with multiple hits is determined by calculating the lowest common ancestor (LCA) of the taxonomies of multiple hits. For the LCA approach, only hits with a bit score equal or higher than 90% of the bit score of the best hit are considered. The percentage value is variable in the tool and influences the sensitivity and specificity [Huson et al., 2007]. The LCA and best BLAST hit are stored in the database for each read. The latter information is used for a mapping of reads against reference genomes (5.1.8). The LCA module was provided by Thomas Bekel.

5.1.6 The Pipeline for the analysis of metagenome contigs

The short sequence length produced by NGS may prevent significant matches to proteins in databases. Because of this, the functional annotations of metagenome contigs contribute to the functional profile of the metagenome in MetaSAMS. For handling and exploring the functional or metabolic potential of microbial communities, a pipeline has been implemented that allows gene calling and annotation of metagenome contigs assembled from short reads. The contigs imported to MetaSAMS are called "Metatigs". Accordingly, the pipeline is termed "Metatig pipeline".

First, the Metatig pipeline starts a gene prediction pipeline for the identification of coding sequences (CDSs) on the supercontig (Fig. 5.4). The gene prediction relies on existing tools initially implemented for the identification of CDSs in genomes of isolated organisms. Gene prediction of the data is achieved by the prokaryotic gene prediction tools Glimmer3 [Delcher et al., 2007] or Prodigal [Hyatt et al., 2010]. Due to the modular implementation of MetaSAMS, further gene prediction tools can be easily

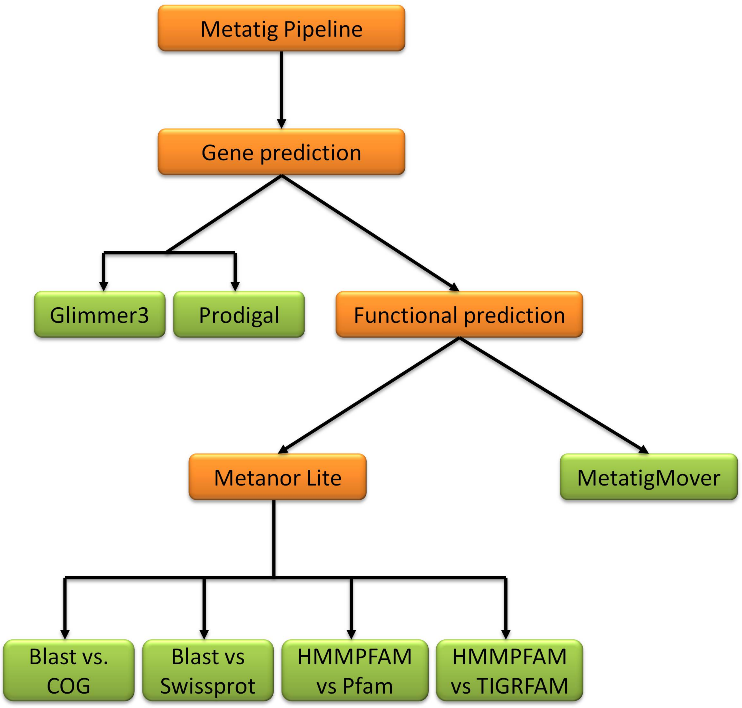


Figure 5.4: A schematic representation of the Metatig pipeline: Green boxes indicate the applied tools, orange boxes the involved pipelines. The Metatig pipeline basically consists of a gene and functional prediction module. The latter executes Metanor-Lite, which includes tools for the functional annotation of the metagenome data.

added. The selection of the gene prediction tool depends on the GC content of the given supercontig. Since Prodigal operates well in GC rich regions [Hyatt et al., 2010], it is automatically applied for the regional interpretation of a supercontig with a GC content above 40%. Otherwise, the supercontig is annotated with Glimmer3, which shows better accuracy in genomic regions with a low GC content. Parameters of the gene prediction tools as well as the pipeline, e.g., the threshold of the GC content for the application of Prodigal instead of Glimmer3, can be changed.

After gene predictions, the functional annotation of the CDSs is carried out (Fig. 5.4). The functional analysis is based on similarity searches against various sequence databases. Therefore, the novel functional prediction pipeline named Metanor-Lite is executed. Metanor-Lite is a reduced version of the GenDB Metanor pipeline [Meyer et al., 2003], which is also used in SAMS for functional predictions. The Metanor-Lite pipeline applies BLAST comparisons of the predicted CDSs to the SwissProt [Boutet et al., 2007] and COG [Tatusov et al., 2001] databases. Moreover, hidden Markov model (HMM) [Durbin et al., 2006] based searches against the databases Tigrfam [Haft et al., 2001] and Pfam [Finn et al., 2006] using the HMMER3 package [Eddy, 2011] are applied to assign protein functions. In comparison to the Metanor pipeline in SAMS, Metanor-Lite excludes huge databases such as the non-redundant protein sequence database (nr) and nucleotide sequence database (nt) to decrease the number of comparisons and computing time. The results of the Metanor-Lite pipeline are stored as Observations in the object-relational database of MetaSAMS. Based thereon, the automatic function prediction is performed. It generates annotations, which provide the functional interpretation such as KEGG numbers [Kanehisa and Goto, 2000] and COG accessions [Muller et al., 2010], of the genes and gene products.

As the gene prediction tools are executed on the sequence of the supercontig, the start and stop position of each CDS refer to the supercontig sequence. To obtain the gene coordinates on the metatig sequence, the tool "MetatigMover" is applied (Fig. 5.4). MetatigMover screens each metatig object on the supercontig and checks whether the start and stop positions of the metatig overlap with those of a predicted gene. If such an overlap is identified, the coordinates of the CDS are changed in respect of the metatig sequence in the last step of the Metatig pipeline.

5.1.7 Statistical tools

The R Project for Statistical Computing[8] provides packages for statistical analyses. In MetaSAMS, access to the R functionalities is realized using the RSPerl module. The data matrices are converted into R data objects, which are used as inputs for R functions. The PNG images representing the results are generated in R and are accessible *via* the MetaSAMS web frontend.

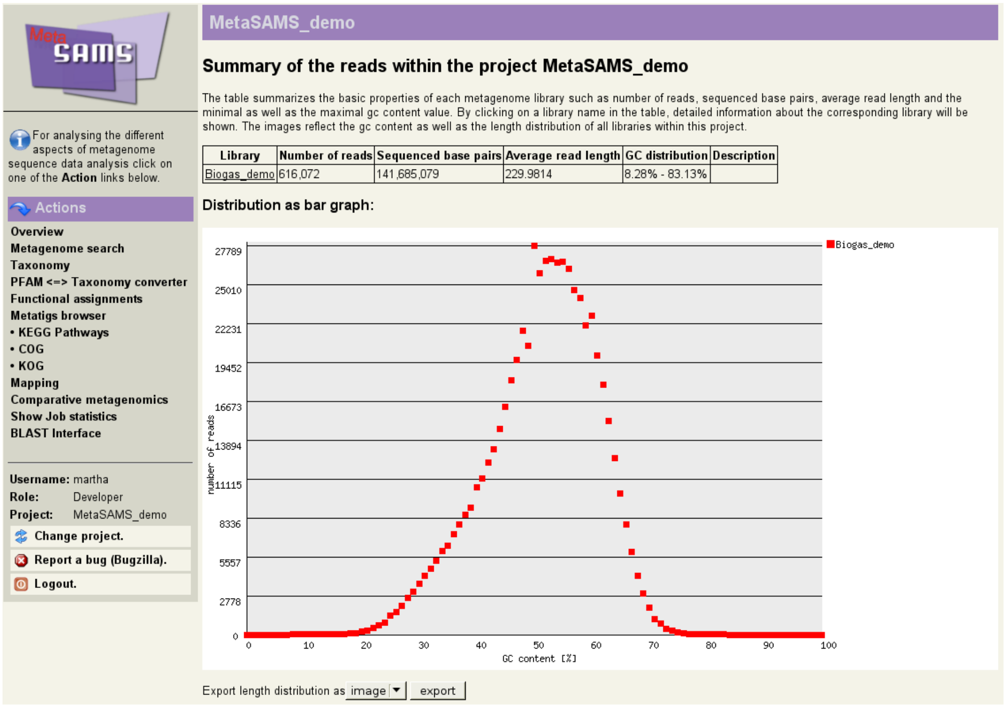
Rarefaction analysis

The taxonomic profile deduced from a metagenome can be used to estimate the coverage of an environmental sample by performing rarefaction analysis. Rarefaction curves are computed by plotting the number of estimated taxa on a rank versus the size of subsamples. Rarefaction analysis is carried out using the Vegan package available in R. In MetaSAMS, rarefaction analysis is calculated based on the taxonomic profiles for each classifier on a selected rank. The results of the rarefaction analysis are illustrated in images and tables.

Comparative analysis

MetaSAMS provides various tools and visualization features for the comparative analysis of different metagenomic sequence data. The comparisons are based on either

[8] http://www.r-project.org/



For analysing the different aspects of metagenome sequence data analysis click on one of the **Action** links below.

Username: martha
Role: Developer
Project: MetaSAMS_demo

Summary of the reads within the project MetaSAMS_demo

The table summarizes the basic properties of each metagenome library such as number of reads, sequenced base pairs, average read length and the minimal as well as the maximal gc content value. By clicking on a library name in the table, detailed information about the corresponding library will be shown. The images reflect the gc content as well as the length distribution of all libraries within this project.

Library	Number of reads	Sequenced base pairs	Average read length	GC distribution	Description
Biogas_demo	616,072	141,685,079	229.9814	8.28% - 83.13%	

Distribution as bar graph:

Export length distribution as image export

Figure 5.5: Representation of raw data in MetaSAMS: The overview, which is the starting site in MetaSAMS, gives first insights into the raw data in terms of GC content and read length distribution. The information is provided in tables and images. A navigation bar allows a fast retrieval of specific visualizations in MetaSAMS.

functional results or taxonomic classifications. It incorporates Venn diagrams and tables making it well-suited for gaining first insights into similarities and differences between metagenomes. Furthermore, MetaSAMS enables the generation of user-specific histograms for comparative visualizations. Methods such as principal component analysis (PCA) and hierarchical clustering analysis (HCA) are supplied using the R modules.

5.1.8 Frontend

The functionalities of MetaSAMS are accessible *via* a web-based frontend. A navigation bar has been designed, which categorizes the visualizations of the results computed by the taxonomic and functional pipelines (Fig. 5.5). This allows the user to access specific information very fast and easily. In the following the different visualizations are described.

Overview - raw data representation

To allow researchers to assess the quality of the sequenced reads, visualizations of the raw data are provided (Fig 5.5). During the import of raw reads, GC content and length of the sequences are stored. This information is utilized to generate GC and length plots for reads contained in a readset, library or project.
In addition, a table summarizes the sequencing results, namely the number of sequenced bases, the average read length and the number of reads in a project (Fig. 5.5). Deviations of expected parameters can be easily identified in the tabular or graphical representations.

Visualizations of the taxonomic and functional annotations

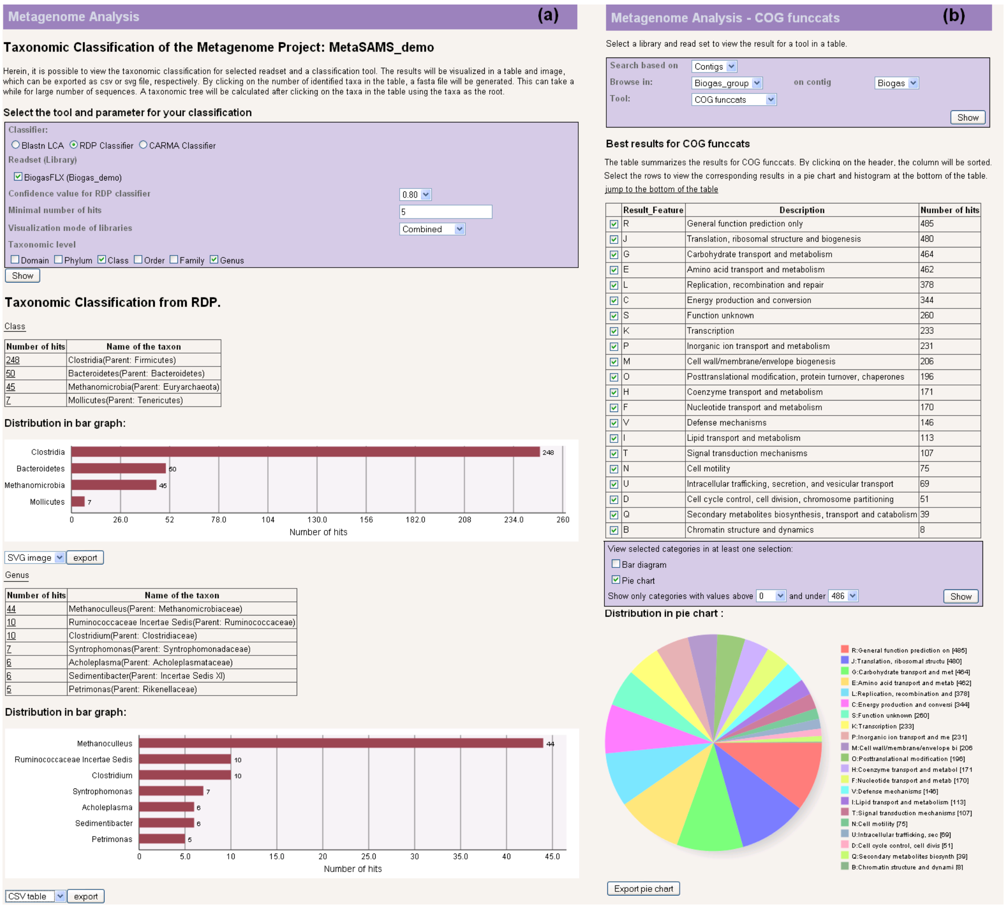
The results of the pipelines are stored in the new data schema and can be viewed and downloaded over the web interface (Fig. 5.6). In addition, the taxonomic and functional profiles are cached to improve the performance. The taxonomic profiles can be accessed in tables or bar charts (Fig. 5.6a). An important feature of MetaSAMS is the ability to compare the taxonomic profiles generated with different classifiers.

The functional assignment browser lists the functional results such as KEGG pathways, EC numbers and COG accessions with the counts of identified genes (Fig. 5.6b). It is also possible to view Pfam accessions and GO terms with the counts of identified reads. The user can select specific functional categories for the subsequent visualization consisting of either exportable bar or pie charts in SVG formats. MetaSAMS provides a mapping of annotations to functional categories, for example EC numbers or COG accessions are mapped on KEGG pathways or COG functional categories, respectively.

Based on the CARMA results, MetaSAMS allows the generation of taxonomic profiles for all metagenome reads or only for reads that are functionally assigned to specific Pfam accessions. Conversely, a Pfam-based profile for a specific taxon can be determined. The combination of functional and taxonomic results allows the user to explore a metagenome in such a way that organisms encoding specific functions or functional roles of important organisms can be identified. Furthermore, sequences for each assigned taxon can be retrieved in fasta format.

Metatig representation

All metatigs are listed in the web frontend (Fig. 5.7a). The list can be filtered based on the length, GC content and number of predicted CDSs. If a metatig has been selected, it is visualized in the "Metatig Viewer" (Fig. 5.7b). CDSs are represented by green arrows with gene names, if provided, above the arrow. The position of the arrow represents the position of the CDS in the metatig. By moving the mouse over a gene, regional and functional annotations are summarized in a tooltip. Additional information, e.g., observations, DNA or amino acid sequence, are accessible by clicking on a gene (Fig. 5.7c). In this case, the red arrow represents the active CDS in the Metatig Viewer, and a



Taxonomic Classification of the Metagenome Project: MetaSAMS_demo

Herein, it is possible to view the taxonomic classification for selected readset and a classification tool. The results will be visualized in a table and image, which can be exported as csv or svg file, respectively. By clicking on the number of identified taxa in the table, a fasta file will be generated. This can take a while for large number of sequences. A taxonomic tree will be calculated after clicking on the taxa in the table using the taxa as the root.

Select the tool and parameter for your classification

Classifier:
○ Blastn LCA ● RDP Classifier ○ CARMA Classifier
Readset (Library)
☑ BiogasFLX (Biogas_demo)
Confidence value for RDP classifier: 0.80
Minimal number of hits: 5
Visualization mode of libraries: Combined
Taxonomic level
☐ Domain ☐ Phylum ☑ Class ☐ Order ☐ Family ☑ Genus

Taxonomic Classification from RDP.

Class

Number of hits	Name of the taxon
248	Clostridia(Parent: Firmicutes)
50	Bacteroidetes(Parent: Bacteroidetes)
45	Methanomicrobia(Parent: Euryarchaeota)
7	Mollicutes(Parent: Tenericutes)

Distribution in bar graph:

Genus

Number of hits	Name of the taxon
44	Methanoculleus(Parent: Methanomicrobiaceae)
10	Ruminococcaceae Incertae Sedis(Parent: Ruminococcaceae)
10	Clostridium(Parent: Clostridiaceae)
7	Syntrophomonas(Parent: Syntrophomonadaceae)
6	Acholeplasma(Parent: Acholeplasmataceae)
6	Sedimentibacter(Parent: Incertae Sedis XI)
5	Petrimonas(Parent: Rikenellaceae)

Distribution in bar graph:

Select a library and read set to view the result for a tool in a table.

Search based on: Contigs
Browse in: Biogas_group on contig: Biogas
Tool: COG funccats

Best results for COG funccats

The table summarizes the results for COG funccats. By clicking on the header, the column will be sorted. Select the rows to view the corresponding results in a pie chart and histogram at the bottom of the table.
jump to the bottom of the table

	Result_Feature	Description	Number of hits
☑	R	General function prediction only	485
☑	J	Translation, ribosomal structure and biogenesis	480
☑	G	Carbohydrate transport and metabolism	464
☑	E	Amino acid transport and metabolism	462
☑	L	Replication, recombination and repair	378
☑	C	Energy production and conversion	344
☑	S	Function unknown	260
☑	K	Transcription	233
☑	P	Inorganic ion transport and metabolism	231
☑	M	Cell wall/membrane/envelope biogenesis	206
☑	O	Posttranslational modification, protein turnover, chaperones	196
☑	H	Coenzyme transport and metabolism	171
☑	F	Nucleotide transport and metabolism	170
☑	V	Defense mechanisms	146
☑	I	Lipid transport and metabolism	113
☑	T	Signal transduction mechanisms	107
☑	N	Cell motility	75
☑	U	Intracellular trafficking, secretion, and vesicular transport	69
☑	D	Cell cycle control, cell division, chromosome partitioning	51
☑	Q	Secondary metabolites biosynthesis, transport and catabolism	39
☑	B	Chromatin structure and dynamics	8

View selected categories in at least one selection:
☐ Bar diagram
☑ Pie chart
Show only categories with values above 0 and under 486

Distribution in pie chart :

R:General function prediction on [485]
J:Translation, ribosomal structu [480]
G:Carbohydrate transport and met [464]
E:Amino acid transport and metab [462]
L:Replication, recombination and [378]
C:Energy production and conversi [344]
S:Function unknown [260]
K:Transcription [233]
P:Inorganic ion transport and me [231]
M:Cell wall/membrane/envelope bi [206
O:Posttranslational modification [196]
H:Coenzyme transport and metabol [171
F:Nucleotide transport and metab [170]
V:Defense mechanisms [146]
I:Lipid transport and metabolism [113]
T:Signal transduction mechanisms [107]
N:Cell motility [75]
U:Intracellular trafficking, sec [69]
D:Cell cycle control, cell divis [51]
Q:Secondary metabolites biosynth [39]
B:Chromatin structure and dynami [8]
Export pie chart

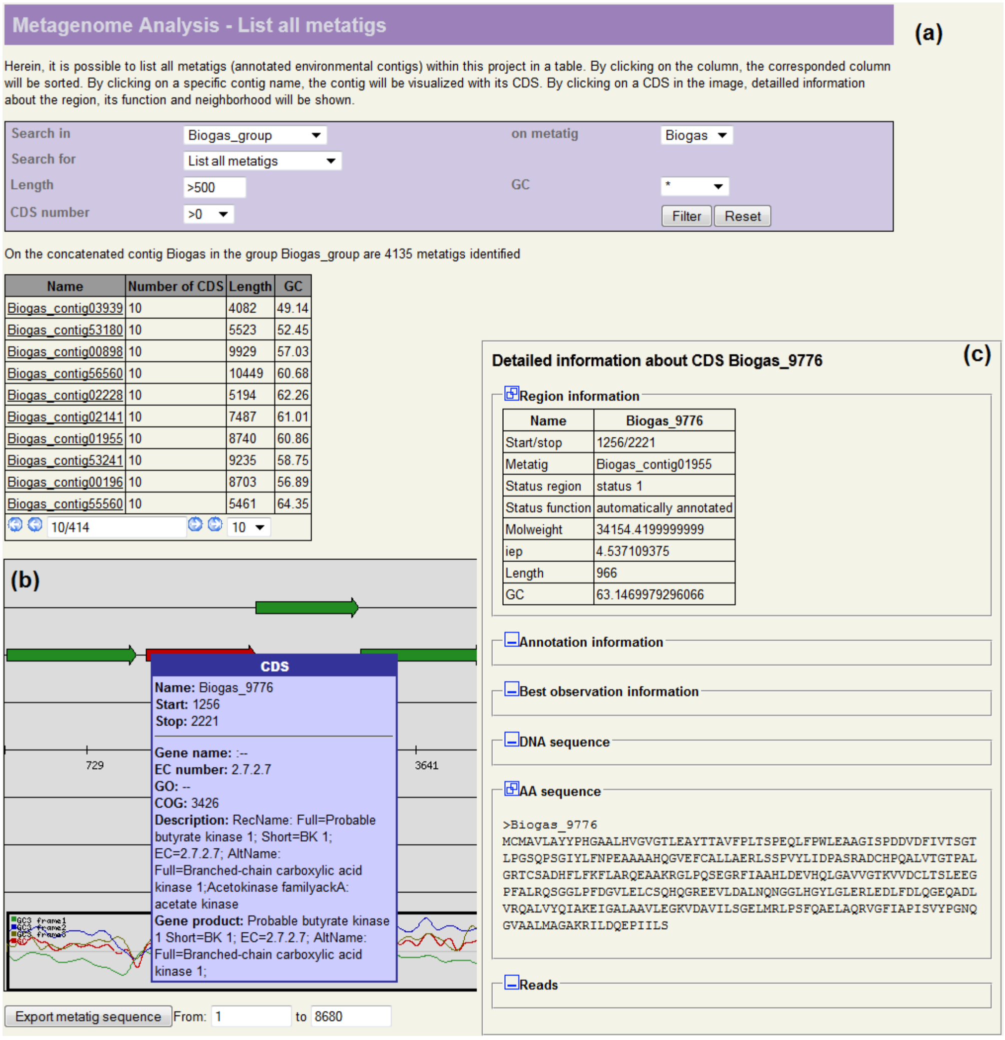
Figure 5.6: Visualization of the results obtained by the read- or contig-based analysis pipelines: (a) Taxonomic profiles are illustrated in tables and bar charts, (b) whereas functional profiles are shown in tables, bar and pie charts.

table provides detailed regional and functional information, the DNA sequence, the amino acid sequence and a mapping of read sequences that assemble the selected CDS.

Search for metagenome reads and contigs

MetaSAMS has the capability to specifically explore the sequence data by a metagenome context search. The search is based on descriptions of reads or genes. By entering a taxon name, Pfam accession or GO term, reads with the corresponding features will be listed. Using keywords describing gene names or functions, the respective gene or metatig sequences can be obtained and downloaded. This search allows the identification and retrieving of full-length genes of interest.

The BLAST and HMM interface allow to search with custom queries in databases composed of metagenome reads and CDSs. BLAST searches against reads, predicted



Metagenome Analysis - List all metatigs

Herein, it is possible to list all metatigs (annotated environmental contigs) within this project in a table. By clicking on the column, the corresponded column will be sorted. By clicking on a specific contig name, the contig will be visualized with its CDS. By clicking on a CDS in the image, detailled information about the region, its function and neighborhood will be shown.

Search in: Biogas_group
on metatig: Biogas
Search for: List all metatigs
Length: >500
GC: *
CDS number: >0

On the concatenated contig Biogas in the group Biogas_group are 4135 metatigs identified

Name	Number of CDS	Length	GC
Biogas_contig03939	10	4082	49.14
Biogas_contig53180	10	5523	52.45
Biogas_contig00898	10	9929	57.03
Biogas_contig56560	10	10449	60.68
Biogas_contig02228	10	5194	62.26
Biogas_contig02141	10	7487	61.01
Biogas_contig01955	10	8740	60.86
Biogas_contig53241	10	9235	58.75
Biogas_contig00196	10	8703	56.89
Biogas_contig55560	10	5461	64.35

10/414 10

Name: Biogas_9776
Start: 1256
Stop: 2221

Gene name: :--
EC number: 2.7.2.7
GO: --
COG: 3426
Description: RecName: Full=Probable butyrate kinase 1; Short=BK 1; EC=2.7.2.7; AltName: Full=Branched-chain carboxylic acid kinase 1;Acetokinase familyackA: acetate kinase
Gene product: Probable butyrate kinase 1 Short=BK 1; EC=2.7.2.7; AltName: Full=Branched-chain carboxylic acid kinase 1;

From: 1 to 8680

Detailed information about CDS Biogas_9776

Region information

Name	Biogas_9776
Start/stop	1256/2221
Metatig	Biogas_contig01955
Status region	status 1
Status function	automatically annotated
Molweight	34154.4199999999
iep	4.537109375
Length	966
GC	63.1469979296066

Annotation information

Best observation information

DNA sequence

AA sequence

```
>Biogas_9776
MCMAVLAYYPHGAALHVGVGTLEAYTTAVFPLTSPEQLFPWLEAAGISPDDVDFIVTSGT
LPGSQPSGIYLFNPEAAAAHQGVEFCALLAERLSSPVYLIDPASRADCHPQALVTGTPAL
GRTCSADHFLFKFLARQEAAKRGLPQSEGRFIAAHLDEVHQLGAVVGTKVVDCLTSLEEG
PFALRQSGGLPFDGVLELCSQHQGREEVLDALNQNGGLHGYLGLERLEDLFDLQGEQADL
VRQALVYQIAKEIGALAAVLEGKVDAVILSGELMRLPSFQAELAQRVGFIAPISVYPGNQ
GVAALMAGAKRILDQEPIILS
```

Reads

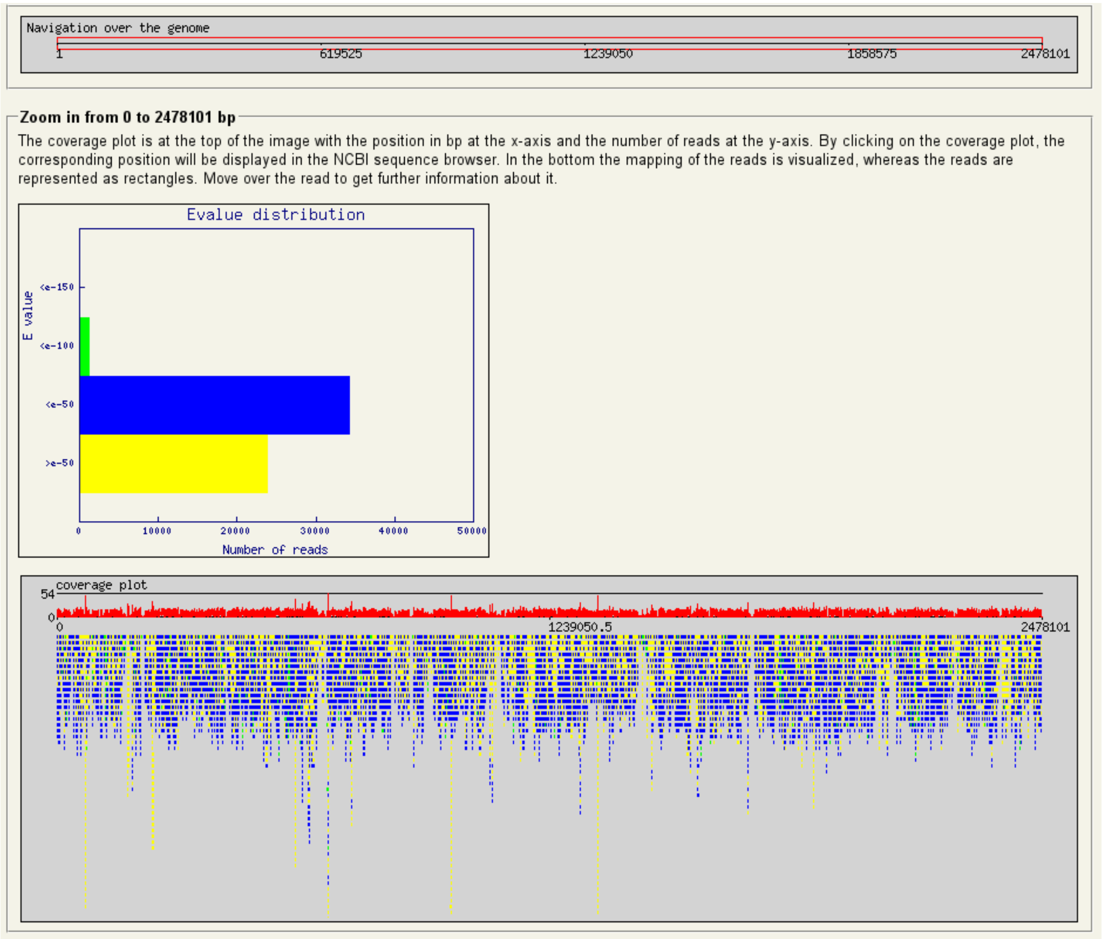
Figure 5.7: Visualization of metatigs and CDSs in MetaSAMS: (a) Metatigs are presented in a table and can be filtered in terms of gene number, GC content and length. (b) A metatig is visualized in the Metatig Viewer, where each CDS is illustrated as an arrow. After clicking on a CDS in the Metatig Viewer, the CDS is highlighted in red and (c) regional and functional annotations are provided in a table.

genes and deduced proteins can be performed. MetaSAMS also provides a more sensitive approach based on profile hidden Markov models (HMMs) for the search for reads or CDSs with an E-value cutoff of 10^{-10}. It is possible to upload either a custom DNA or protein alignment. MetaSAMS utilizes the alignment for building a profile HMM. The profile HMM based on DNA sequences is used for HMM searches against the metagenome reads, whereas the profile HMM built from protein sequences is utilized for searches against the translated CDSs. Finally, the sequences of the matching reads or genes are aligned to the target profile HMM. The alignment is available in a fasta format over the web interface. Using the BLAST and HMM interface, variants of genes can be identified, which may show biotechnologically relevant properties.

Mapping of metagenome reads to references

Aligning metagenome reads to reference genomes gives insights into species presence, genetic variation of the species and the coverage of genomes or genes. Reference genome mapping is in particular efficient in metagenomes with predominant species. For the visualization, reads are mapped on a reference genome using the best BLAST hit approach. The advantage of this approach is the low computational effort compared to exact mapping algorithms of short reads. On the other hand, it is not very accurate. A disadvantage of similarity-based mappings is that horizontal gene transfer (HGT) cannot be handled. However, since it allows mapping of metagenome reads to different reference genomes in a fast time, the best BLAST hit approach has been used to map reads to a reference genome.

The best BLAST hits, which have been already generated for the LCA approach by applying BLAST searches against bacterial genomes (Section 5.1.5), are retrieved from the MetaSAMS database. The database stores the best hits including the hit names, E-values and start as well as stop positions on the query and hit. This information is extracted from the database and utilized for a visualization in the "GenomeMapper" of MetaSAMS (Figure 5.8). The GenomeMapper displays a bacterial genome and the distribution of mapped reads. The user can select a reference genome, on which the sequencing reads are arranged according to their matching positions. Each read is also visualized in a specific color reflecting the quality of the hit based on the E-value. The user can either view the distribution of the reads over the whole or part of the genome. Each position of the reference genome in the GenomeMapper is linked to the NCBI sequence viewer of the corresponding genome to enable detailed regional or functional exploration of specific regions.



Zoom in from 0 to 2478101 bp

The coverage plot is at the top of the image with the position in bp at the x-axis and the number of reads at the y-axis. By clicking on the coverage plot, the corresponding position will be displayed in the NCBI sequence browser. In the bottom the mapping of the reads is visualized, whereas the reads are represented as rectangles. Move over the read to get further information about it.

Figure 5.8: GenomeMapper illustrating the location of mapped reads on a complete genome of *Methanoculleus marisnigri* JR1: A navigation over the genome is provided at the top of the GenomeMapper view. The black line represents the complete genome, whereas the red box describes the region that is shown in the plot at the bottom of the image. The mapped reads are color-coded according to their E-value.

5.2 The AMPLA pipeline for the analysis of 16S rDNA amplicon sequences

The focus of this section is on the analysis of 16S rDNA amplicon sequence data. 16S rRNA genes are phylogenetic markers suitable to provide insight into the diversity and composition of a microbial community. As described in Section 3.1, various tasks exist in the analysis of 16S rDNA amplicon sequence data. To accomplish these tasks, a workflow, termed 16S rDNA **ampl**icon **a**nalyzer (AMPLA), has been designed for the analysis of 16S rDNA amplicon sequence data that includes existing methods and tools. Briefly, the reads are processed in AMPLA by trimming specific oligonucleotides that were required for the construction of the 16S rDNA amplicon sequences. To avoid overestimation of operational taxonomic units (OTUs), sequences with low quality or artificial reads (e.g., chimera) are excluded prior to the cluster analysis. Overall, the AMPLA pipeline consists of a quality control procedure, clustering step and taxonomic assignments calculations (Fig. 5.9). In the following, the involved steps are described in detail.

5.2.1 Processing of raw 16S rDNA amplicon sequences

For the filtering, several files are utilized in AMPLA:

- a sequence file in fasta format,
- a quality file with the quality values for each nucleotide position of the reads,
- a file containing the specification of the used multiplex identifier (MID) tags and primer sequences.

Using the sample-specific MIDs, several communities can be pyrosequenced simultaneously. A typical 454 pyrosequencing read contains adapters from the library preparation, MID tags and primer bases from the PCR amplification. As the sequences of the adapters, MIDs and primers are known, they can be utilized for a quality control of the reads. 16S rDNA amplicon reads are removed that contain ambiguous bases (Ns), mismatches in the MID tags or an average quality score less than 20. Subsequently, the forward and reverse PCR amplification primers are searched allowing 2 mismatches. Sequences that have no recognizable forward or reverse primer are discarded. Finally, sequences shorter than 50 bases are removed from the dataset. The processing is performed by using the trimming script implemented in QIIME, as it is fast and accurate [Caporaso et al., 2010]. After that, unique sequences are identified in the collection and utilized for the subsequent analysis. In addition, the number of times each unique reference sequence is observed is tracked in a file. The advantage of this reduction step is that the computation time of the time-consuming downstream analysis such as calculation of alignments and clusters is decreased. The unique 16S rDNA reads are analyzed using a single-linkage preclustering (SLP) approach [Huse et al., 2010], which discards reads that likely contain pyrosequencing errors. For this purpose, the sequences are aligned using the Needleman-Wunsch pairwise alignment algorithm [Needleman and Wunsch, 1970] in MOTHUR [Schloss et al., 2009]. An alignment obtained from the Greengenes database [DeSantis et al., 2006] serves as a reference. Finally, the pre.cluster command is executed using default settings in MOTHUR, which applies a pseudo-single linkage algorithm. SLP is widely applied for the analysis of sequencing errors, as it is fast and easy to use [Schloss et al., 2011]. Potential chimeras are detected using the *de novo* implementation of the UCHIME algorithm [Edgar et al., 2011]. The

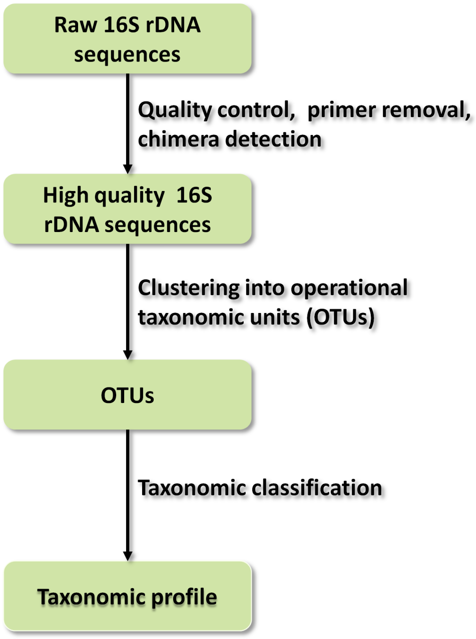

Raw 16S rDNA sequences
Quality control, primer removal, chimera detection
High quality 16S rDNA sequences
Clustering into operational taxonomic units (OTUs)
OTUs
Taxonomic classification
Taxonomic profile

Figure 5.9: Overview of the pipeline for the analysis of 16S rDNA amplicon sequences: The AMPLA pipeline consists of three basic steps. First, the raw reads are processed by removing the primer and MID sequences. Moreover, raw sequences are excluded that reach a low mean quality score or are likely artificial due to PCR amplifications and sequencing errors. Second, the filtered sequences are clustered into operational taxonomic units, which are used for diversity assessment. Finally, a taxonomic profile was generated.

detection of chimeric sequences is a challenging task. So far, the *de novo* mode of UCHIME showed to be a valuable tool for removing potential chimeras compared to other developed methods [Edgar et al., 2011, Wright et al., 2012].

5.2.2 Clustering of 16S rDNA amplicon sequences for diversity estimations

The remaining non-chimeric, high-quality sequences are clustered into OTUs using UCLUST version 3.0 [Edgar, 2010] and an identity threshold value of 97%, which is usually regarded as representing the species level [Schloss and Handelsman, 2005]. UCLUST can process huge datasets accurately [Edgar, 2010, Barriuso et al., 2011]. Singleton OTUs, which contain only one sequence, are excluded, since they are probably associated with PCR and sequencing errors [Reeder and Knight, 2009]. The longest sequence within each OTU is picked as a representative sequence for the cluster. For assessing the completeness of a sequencing effort, rarefaction curves are calculated using the Vegan package in R.

5.2.3 Taxonomic classification of 16S rDNA amplicon sequences

The sequences are taxonomically classified using the RDP Classifier (version 2.3) [Wang et al., 2007]. For the interpretation of the data, only classifications with at least 0.80 assignment confidence are considered. The taxonomic profile is visualized using Krona [Ondov et al., 2011].

5.2.4 Phylogenetic characterization of 16S rDNA sequences

16S rDNA amplicon sequences are suitable for phylogenetic tree reconstructions, since they comprise at least one variable region known to differ between species. Thus, multiple alignments and subsequent calculations of evolutionary distances between reads covering the same region segment are possible. For phylogenetic characterization, only representative OTU sequences obtained by using UCLUST are considered that were assigned to a specific taxon by the RDP Classifier in the previous step. The collected sequences are aligned by MUSCLE [Edgar, 2004a, Edgar, 2004b]. Based thereon, the tree reconstruction is performed in MEGA 5 [Tamura et al., 2007] using the neighbor-joining method [Saitou and Nei, 1987] with genetic distances as defined by Jukes Cantor [Jukes and Cantor, 1969] and a bootstrap value of 1,000 [Tamura et al., 2007].

5.3 The MeTra pipeline for the characterization of metatranscriptome data

This section concentrates on the study of metatranscriptome data. Metatranscriptome data reveal the active members and transcribed functions within microbial communities. For this purpose, the Perl-based **Me**ta**Tra**nscriptome (MeTra) pipeline has been developed, which extracts the ribosomal, messenger and non-coding RNA sequences from a metatranscriptome dataset. Ribosomal and messenger RNAs can be exploited to generate a taxonomic profile of active members, whereas only the messenger RNA can contribute to the functional profile. In the following sections, the databases necessary for the identification of relevant RNA types and the procedure for the study of metatranscriptome data will be provided in detail.

5.3.1 Database construction

For the extraction of ribosomal RNA (rRNA) sequences from the metatranscriptome dataset, three databases comprising the small subunit sequences (SSUdb), large subunit sequences (LSUdb) and further RNAs (profile hidden Markov model RNAdb, pHMM-RNAdb), e.g., tRNAs and 5S rRNAs, are constructed (Tab. 5.1). Sequences for the large subunit (LSU), which is composed of prokaryotic 23S and eukaryotic

Table 5.1: Databases constructed for the identification of different RNA types

Database	RNA-type	Data source
LSUdb	prokaryotic 23S, eukaryotic 28S rRNAs	SILVA
SSUdb	prokaryotic 16S, eukaryotic 18S rRNAs	Greengenes, SILVA, RDP
pHMM-RNAdb	functional RNAs	RFAM, NCBI database

28S rRNAs, are retrieved from SILVA (Release 102) [Pruesse et al., 2007] as an arb file [Ludwig et al., 2004]. The arb file contains LSU as well as adjacent gene sequences (e.g., SSU rDNA) from *Archaea*, *Bacteria* and *Eukaryota* in an alignment. To retrieve only LSU sequences, they are exported as a fasta file including the aligned sequences from position 66,155 to 129,061[9] (according to the 23S rDNA gene in *E. coli*) using the software package ARB [Wang et al., 2007].

Sequences for the SSUdb are obtained from several public ribosomal databases including Greengenes [DeSantis et al., 2006], RDP-II (Release 10.21) [Cole et al., 2003] and SILVA (Release 102). SILVA provides 16S/18S rRNAs for all three domains of life, whereas RDP and Greengenes store bacterial and archaeal 16S rRNAs. The small subunit sequences from SILVA and Greengenes are retrieved in an arb file format. Likewise the LSU database in SILVA, the SSU sequences include more than the small ribosomal RNAs such as incorrectly annotated large subunit ribosomal RNAs. Therefore, only positions between 986 and 43,332 (according to 16S rDNA gene in *E. coli*) of the aligned SILVA sequences are considered for the setup of the SSUdb. The trimmed sequences are exported as a fasta file using the ARB software.

For the search of additional functional RNAs, a profile hidden Markov model (HMM) [Durbin et al., 2006] database is built (pHMM-RNAdb). The RFAM 10.0 database [Griffiths-Jones et al., 2005] is utilized for the generation of the pHMM-RNAdb. The alignments for all available RNA families are exported. Using the HMMER3 package [Eddy, 2011], a profile HMM is built for each alignment. Additionally, tRNA sequences from all available NCBI genomes are extracted and filtered for the length ranging from 30 to 180 bp in order to remove non-tRNA fragments. Subsequently, the sequences are separated according to their anticodons and bacterial or archaeal origin and aligned using MUSCLE [Edgar, 2004a, Edgar, 2004b]. Profile HMMs are generated for each alignment and added to the pHMM-RNAdb.

[9] The positions refer to the aligned LSU sequences and not to the raw LSU sequence

5.3.2 Pipeline for the identification of different RNA types

For the identification of RNA tags within the dataset a three-step analysis pipeline has been developed in MeTra (Fig. 5.10). First, a BLAST [Altschul et al., 1990] search using an E-value cutoff of 10^{-5} against the custom SSUdb and LSUdb is performed to identify small and large subunit ribosomal sequences in the metatranscriptome dataset (Fig. 5.10, step 1). The sequence complexity filter is explicitly disabled to include regions with low sequence complexity. As the SSUdb comprises both archaeal and bacterial 16S rRNA as well as eukaryotic 18S rRNA gene sequences, BLAST results for eukaryotic sequences are discarded. For this purpose, the BLAST result headers containing the accession number as identifier are examined. Sequences with matches to the bacterial and archaeal entries in the SSUdb databases are classified by means of the RDP Classifier [Cole et al., 2003]. Only classifications with at least 0.80 assignment confidence are considered for the taxonomic profile.

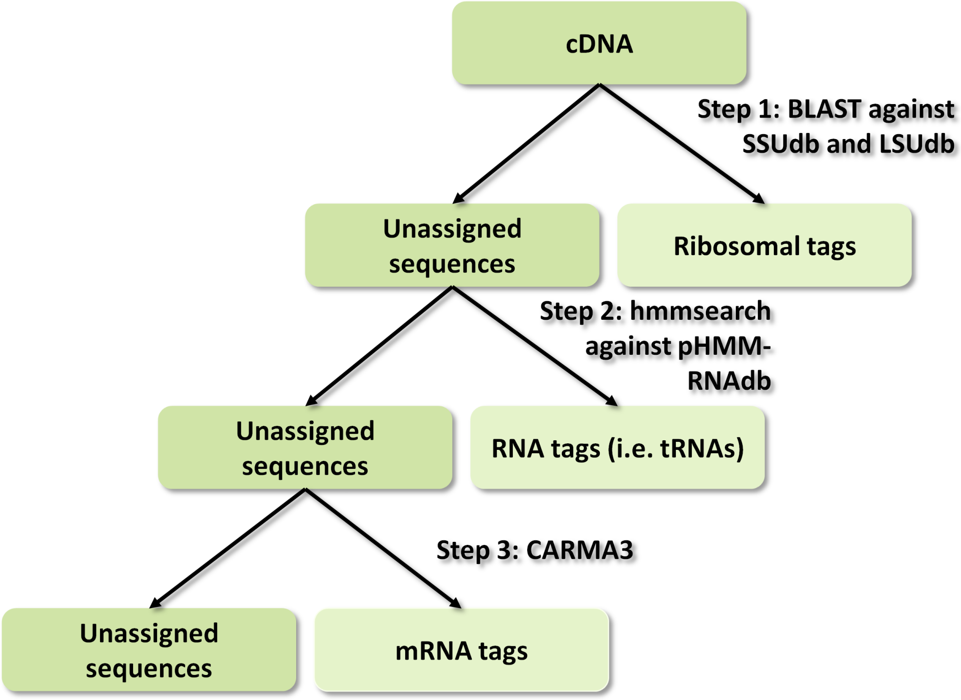


Figure 5.10: Steps involved in the metatranscriptomic MeTra pipeline: The pipeline identifies rRNA tags, remaining RNA tags and mRNA tags in three different steps.

Second, a profile HMM-based approach is applied to identify additional RNA types such as tRNAs and 5S rRNAs in the pHMM-RNAdb database (Fig. 5.10, step 2). Reads without matches to SSUdb and LSUdb are compared to the pHMM-RNAdb using the HMMER3 package (E-value cutoff 10^{-5}).

In the third step, mRNA tags are identified using CARMA3 [Gerlach and Stoye, 2011] (Fig. 5.10, step 3). All RNA sequences that have neither a BLAST hit to the ribosomal databases SSUdb and LSUdb nor a hit to the pHMM-RNAdb are further functionally and taxonomically characterized using CARMA3, which simultaneously performs BLASTX searches against the non-redundant GenBank protein database and HMM-based searches against the Pfam database [Finn et al., 2006].

Thereafter, the active biological processes within the underlying community are determined. For this purpose, sequences classified as mRNA tags by CARMA3 are compared against the eggNOG [Muller et al., 2010] database using BlastX (E-value cutoff of 10^{-5}, disabled low complexity filter). Finally, the sequences are annotated with COG or NOG accessions according to their best hit.

5.4 A method for the identification of industrially relevant enzymes

In this section, a method will be outlined that is applicable for the discovery of novel genes encoding target enzymes in metagenome datasets. Reads encoding functions of industrial interest can be identified by exploiting the knowledge of so far described enzymes. Recently, hydrolase genes were identified in metagenome data, which were confirmed in subsequent activity tests (Section 2.2.3). The surveys clearly demonstrated that profile HMMs representing enzymes of interest are suitable to capture specific sequences from metagenome databases. The profile HMMs were obtained from the Pfam database, which is a large collection of protein families [Finn et al., 2006]. Since the Pfam database has a limited number of HMMs, the development of new models is needed in order to search for reads encoding desired enzymes. In the following, an approach for the construction of a novel profile HMM will be described. A requirement for this approach is the knowledge of described enzyme sequences and a conserved domain that is specific for the enzyme.

5.4.1 Construction of a profile hidden Markov model (HMM)

A two-step approach is carried out to construct a novel profile HMM representing an enzyme of industrial interest (Fig. 5.11). First, an initial profile HMM is built. For this purpose, a BLAST search is applied to the NCBI nr protein database using known queries of the target enzyme. The sequences of the hits are collected and aligned using MUSCLE [Edgar, 2004a, Edgar, 2004b]. Sequences that did not cover the required conserved domain are excluded from subsequent analysis. In addition, duplicates are removed to avoid bias in the profile HMM. The remaining aligned sequences are used

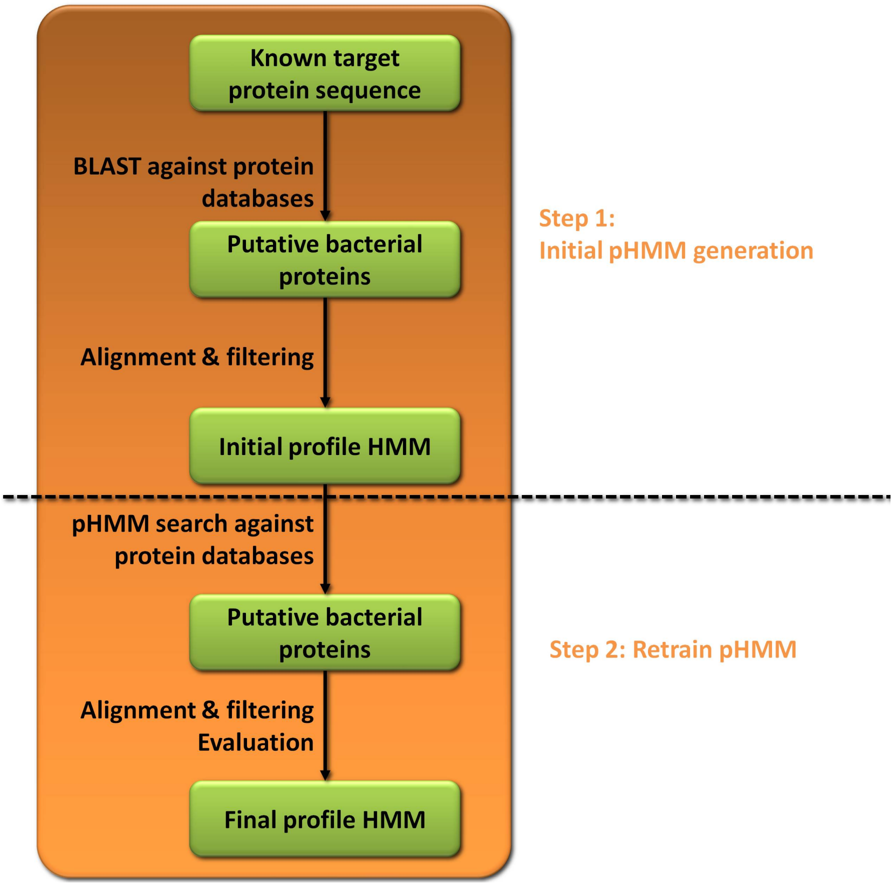

Known target protein sequence
BLAST against protein databases
Putative bacterial proteins
Alignment & filtering
Initial profile HMM
Step 1: Initial pHMM generation
pHMM search against protein databases
Putative bacterial proteins
Step 2: Retrain pHMM
Alignment & filtering Evaluation
Final profile HMM

Figure 5.11: Approach for the construction of a profile hidden Markov model (HMM) representing a target enzymes: An initial set of described protein sequences is used to search for similar proteins. The hits serve for building an initial profile HMM, which is refined with further sequences and rebuilt from improved alignments.

as a basis for building an initial profile HMM by applying the HMMER3 package [Eddy, 2011].

Second, the initial profile HMM modeling the target enzyme is retrained. Therefore, an HMM-based search is applied to identify additional enzymes in a public protein database that match the initial profile HMM. The sequences with a hit are extracted and aligned using HMMER3. The alignment is verified by removing duplicates or invalid sequences. The remaining sequences are aligned to the initial profile HMM with HMMER3, and a final profile HMM is constructed.

CHAPTER 6

Application examples

This chapter describes the outcomes of the analyses of whole metagenome shotgun, 16S rDNA amplicon and metatranscriptome tags obtained from a biogas plant. For a better understanding of the results, the process of biogas production is introduced in the first section. Thereafter, the taxonomic and functional results deduced by means of MetaSAMS are reported for the metagenome of the biogas-producing community. Next, a deeper resolution of the taxonomic composition of the underlying community is described based on 16S rDNA amplicon analysis in AMPLA. Additionally, active members and functions of the biogas-producing community are presented by applying the developed MeTra pipeline on the corresponding metatranscriptome dataset. Finally, the results of a search for laccases encoded in genomes and metagenomes are reported.

6.1 Introduction to biogas

The dwindling of fossil fuel supplies and the worldwide growing demand for energy are the driving forces to find sustainable energy sources. Simultaneously, burning of fossil fuels is associated with the increase in atmospheric carbon dioxide, which is considered to affect global warming. In this context, biogas from renewable resources or organic waste is a promising carrier of bioenergy [Weiland, 2010]. Methane, the energy-rich molecule in biogas, can be converted to electric energy or heat in an ecologically-friendly way. The conversion of organic material to biogas is a complex process, which is composed of four successive steps, namely hydrolysis, acidogenesis, acetogenesis and methanogenesis (Fig. 6.1) [Deublein and Steinhauser, 2008]. The processes are

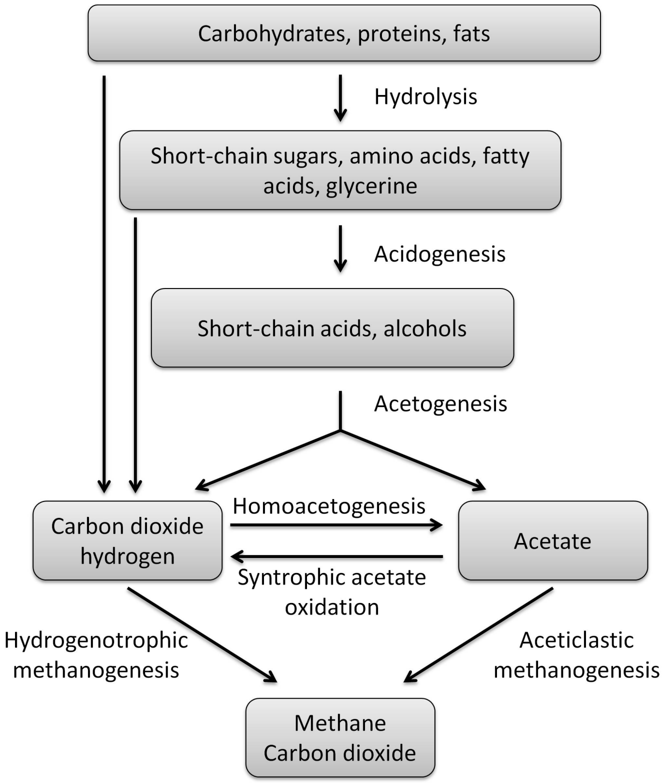

Carbohydrates, proteins, fats
Hydrolysis
Short-chain sugars, amino acids, fatty acids, glycerine
Acidogenesis
Short-chain acids, alcohols
Acetogenesis
Carbon dioxide hydrogen
Homoacetogenesis
Syntrophic acetate oxidation
Acetate
Hydrogenotrophic methanogenesis
Aceticlastic methanogenesis
Methane Carbon dioxide

Figure 6.1: The process of anaerobic digestion in biogas plants (modified from Deublein and Steinhauser, 2008): The anaerobic decomposition of biomass consists basically of four stages, namely hydrolysis, acidogenesis, acetogenesis and methanogenesis. In addition, homoacetogenesis and syntrophic acetate oxidation are coupled to the basic stages.

accomplished by certain groups of microorganisms under anaerobic conditions. Some of them partly stand in syntrophic associations, i.e. their growth relies on the metabolisms of other microorganisms.

In the first step, the hydrolysis, anaerobic bacteria cleave polymers, like polysaccharides, proteins and lipids, into monomers by hydrolytic enzymes [Cirne et al., 2007]. The products of this reaction are short compounds such as short-chain sugars, amino acids, fatty acids and glycerin. Frequently, species being relevant for the hydrolytic cleavage belong to the class *Bacteroides* and *Clostridia* [Jaenicke et al., 2011].
Under anaerobic conditions, the short compounds are taken up by bacteria and transformed into short-chain organic acids (e.g., butyric acid, propionic acid, acetic acid), alcohols, carbon dioxide and hydrogen (Fig. 6.1) [Deublein and Steinhauser, 2008]. This step of microbial degradation is called acidogenesis. The degradation of sugars can

be catalyzed in different pathways. In the succinate and acrylic pathways, sugars are converted to propionic acid. Another pathway is the butyric acid pathway, in which butyric acid is formed from sugar by acidogenic *Clostridium* species. A degradation pathway for fatty acids is β-oxidation.
The products of the acidogenic step are converted into acetate during acetogenesis (Fig. 6.1). In addition, acetate production is carried out by homoacetogenic microorganisms, which transform carbon dioxide and hydrogen to acetic acid.
The final step is methanogenesis (Fig. 6.1), which is conducted by methanogenic *Archaea* under strictly anaerobic conditions. Aceticlastic methanogens convert acetate to methane, whereas hydrogenotrophic methanogens use carbon dioxide and hydrogen to produce methane [Demirel and Scherer, 2008]. Aceticlastic methanogens include the genera *Methanosaeta* and *Methanosarcina*. Hydrogenotrophic methanogens are a diverse group. Species of *Methanobacterium*, *Methanospirillum*, *Methanomicrobium* and *Methanothermobacter* are capable of carrying out hydrogenotrophic methanogenesis [Demirel and Scherer, 2008]. Some of them are described to be in association with syntrophic acetate-oxidizing bacteria, which convert acetate into carbon dioxide and hydrogen [Ahring, 2003, Hattori, 2008].

The process of anaerobic degradation is well described but the knowledge about the taxonomic composition is still limited. To optimize the yield and efficiency of the biogas production process, a better understanding of the underlying taxonomic structure and metabolic properties of the biogas-producing microbes is essential. The first metagenome of a microbial community of a continuously stirred tank reactor (CSTR) fed with maize silage provided information about the taxonomic composition and the functional capabilities [Krause et al., 2008b, Schlüter et al., 2008, Jaenicke et al., 2011]. The analyses revealed *Clostridia* from the phylum *Firmicutes* as the most prevalent bacterial class, whereas species of the order *Methanomicrobiales* were shown to be dominant among *Archaea*.

In this thesis, the metagenome of a microbial community from a biogas reactor analyzed previously [Jaenicke et al., 2011] was used to illustrate the capabilities of the novel system MetaSAMS. Moreover, 16S rDNA and metatranscriptome datasets were generated to deepen the knowledge about the biogas-producing microbial community. The first 16S rDNA amplicon and metatranscriptome approaches were carried out for a microbial community residing in a biogas plant. For this purpose, a fermentation sample was taken from the standard sampling device installed at the main fermenter of the same biogas plant for which the metagenome sequencing project was carried out. Sequences were generated by means of the 454 pyrosequencing technique. Both datasets were processed and analyzed by performing the pipelines MeTra and AMPLA. The aim of this analysis is to shed more light on the processes and organisms relevant for the anaerobic digestion in the studied biogas plant.

6.2 Analysis of a metagenome obtained from a biogas plant by means of MetaSAMS

The platform MetaSAMS provides a broad spectrum of different features for the characterization of metagenome datasets generated by high-throughput sequencing techniques. To demonstrate the capabilities of MetaSAMS, the metagenome of a biogas plant generated on the GS FLX platform using Titanium chemistry was analyzed [Jaenicke et al., 2011]. The sequencing procedure yielded 1,347,644 reads with an average read length of 368 bases. As previously described [Jaenicke et al., 2011], the data was normalized with respect to the observed GC bias of the sequencing technique and duplicates were removed, as they are considered as artificial sequences generated during emulsion PCR that likely influence the abundances of taxa [Gomez-Alvarez et al., 2009]. The filtered metagenome dataset consists of 1,019,333 reads, which were imported into MetaSAMS and taxonomically annotated using the automated taxonomic pipeline. Moreover, an assembly of the unfiltered metagenome dataset was carried out using the GS *De Novo* Assembler version 2.6 with standard parameters resulting in 21,843 contigs of at least 500 bases in length. To increase the number of large contigs, reads that were generated on the GS FLX using the standard chemistry for the same fermeter sample [Krause et al., 2008b, Schlüter et al., 2008] were used for the assembly. In total, the GS *De Novo* Assembler generated 43,745 contigs, whereas 27,576 contigs reached a length above 500 bases. The large contigs were imported into MetaSAMS and annotated using the functional annotation pipeline.

6.2.1 Taxonomic profiling of a biogas-producing community

MetaSAMS provides three methods for the taxonomic characterization of a metagenome dataset: 1) an analysis of 16S rRNA gene fragments extracted from the metagenome dataset using the RDP Classifier [Wang et al., 2007], 2) a Lowest Common Ancestor (LCA) analysis based on multiple BLAST [Altschul et al., 1990] hits, and 3) an analysis based on reciprocal BLAST searches by means of CARMA3 [Gerlach and Stoye, 2011]. MetaSAMS allows a comparative analysis of different taxonomic profiles generated by the available classifiers. This feature was utilized to examine the taxonomic assignments on superkingdom level. All three approaches consistently disclosed that *Bacteria* are the most dominant superkingdom followed by *Archaea* (Fig. 6.2). The taxonomic profiles based on the LCA and CARMA approaches exemplify the advantage of using environmental gene tags (EGTs) for classification. LCA and CARMA3 utilized 30% and 60% of the metagenome sequences for taxonomic profiling, respectively. Compared to the EGT-based approaches, 16S rRNA gene sequences only constituted a small amount of the biogas metagenome (approximately 0.3%) and thus might poorly represent the underlying organisms. Because CARMA3 performs better than the LCA approach [Gerlach and Stoye, 2011] and generates more results than the 16S rRNA gene-based approach (Fig. 6.2), the assignments predicted by CARMA3 were used for the taxonomic analysis on lower ranks. Therefore, the CARMA3 profile was exported and explored using the visualization tool Krona [Ondov et al., 2011] (Fig. 6.3).

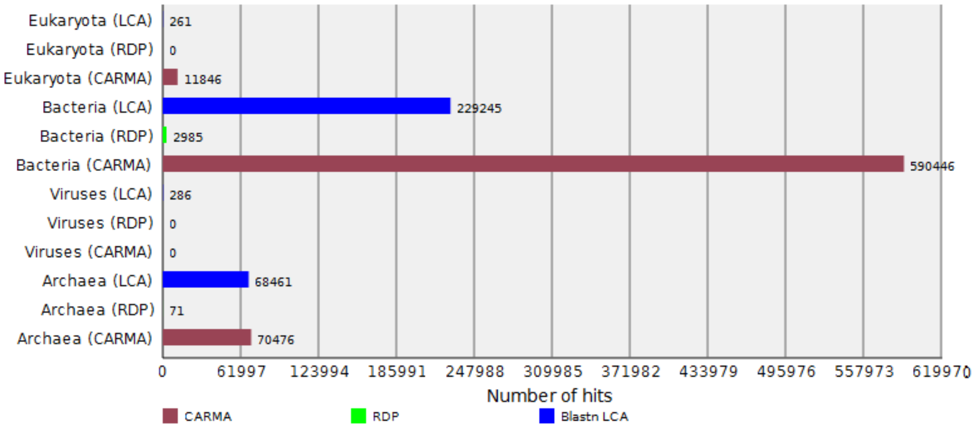


Figure 6.2: A comparative visualization of different taxonomic profiles in MetaSAMS: MetaSAMS allows comparing the taxonomic distributions obtained by different classifiers. As an example, the classifications for the rank superkingdom obtained by CARMA3, the RDP Classifier and LCA for a metagenome from a biogas-producing community are presented.

The bacterial phyla *Firmicutes* (28% of all metagenome reads) and *Bacteroidetes* (7%) as well as the archaeal phylum *Euryarchaeota* (7%) dominate the biogas community. The phyla *Proteobacteria*, *Tenericutes* and *Spirochaetes* contribute only a small amount to the metagenome EGTs (1-2%). Species of *Bacteroidetes* and *Proteobacteria* are associated with the hydrolysis step [Jaenicke et al., 2011]. *Spirochaetes* use carbohydrate and amino acids for their energy metabolism [Johnson, 1977]. Most of the *Firmicutes* sequences belong to the class *Clostridia* and *Bacilli* (55 and 6% of *Firmicutes* reads, respectively) with *Clostridiales* and *Bacillales* being the most represented orders within these classes. Many *Clostridia* are capable of anaerobic digestion of complex carbohydrates such as cellulosic material [Guedon et al., 2000]. Hence, they play a major role for the hydrolytic step of plant biomass. All of these phyla were previously observed in a taxonomic profile based on the same metagenome data of the biogas plant [Jaenicke et al., 2011]. They were also described in other biogas fermentation samples, for example in biogas reactors that were fed with rye silage and winter barley straw [Rademacher et al., 2012].

Additional abundant classes are *Methanomicrobia* (6%), *Bacilli* (2%) and *Spirochaetia* (0.4%). On family and genus level, *Methanomicrobiaceae* (5%) and *Methanoculleus* (1%) are dominant, respectively. Members of the family *Methanomicrobiaceae* are described to produce methane using the hydrogenotrophic pathway [Demirel and Scherer, 2008]. Only 4% of the metagenome reads have an assignment to a known taxon on genus

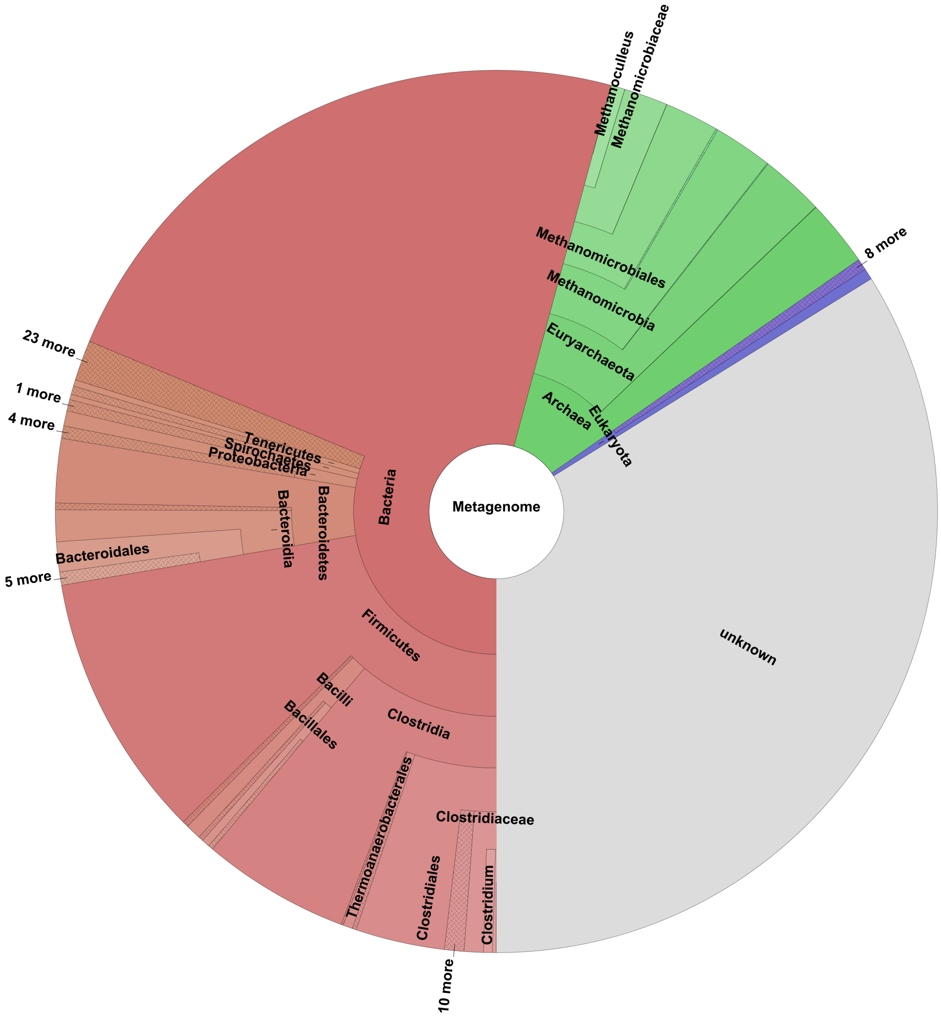


Figure 6.3: A taxonomic profile of a metagenome based on CARMA3 and exported from MetaSAMS: The taxonomic profile for a biogas-producing community was determined by performing CARMA3 in MetaSAMS. The profile was exported and visualized using Krona [Ondov et al., 2011].

level. Apart from *Methanoculles*, *Clostridium* and *Bacteroides* occur with 31% and 4% of all classifiable reads on genus level. In addition, 200 reads feature a high degree of sequence similarity to *Methanoculleus marisnigri* and 198 to *Clostridium thermocellum* on species level.

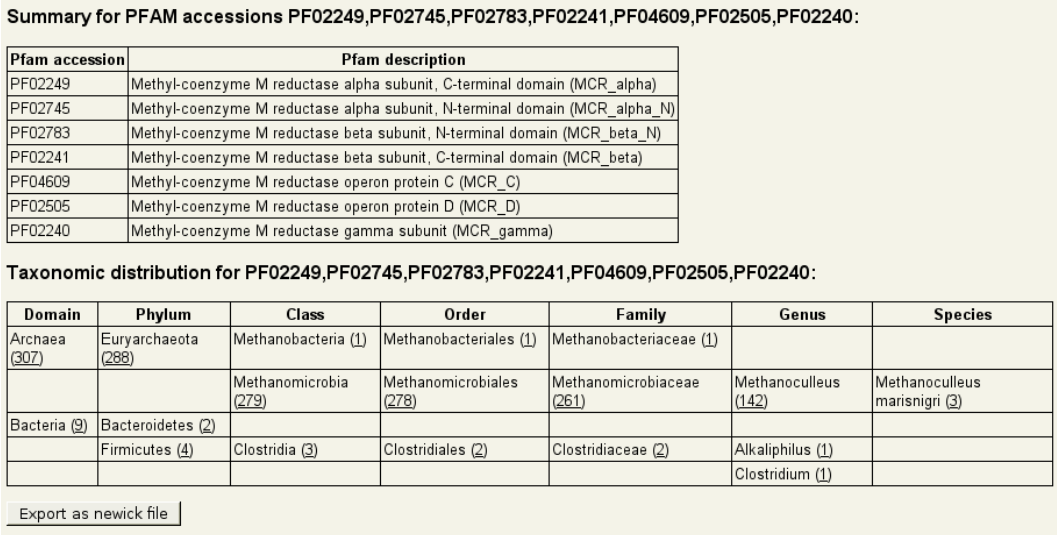
However, most of the taxa residing in a biogas plant are so far not described. In total, 34% of the reads have no significant assignments to known reference sequences on superkingdom level. The lack of references continues even more on the lower ranks, as only 29% and 11% of the metagenome reads could be classified to a taxon on class and family level, respectively.

6.2.2 Functional analysis of processes central in the anaerobic digestion

Next, the functions encoded by the microbes residing in the biogas fermenter were examined. MetaSAMS computes functional profiles by using two methods, namely based on metagenome reads and contigs. First, a functional profile is provided based on Pfam [Finn et al., 2006] and GO [Ashburner et al., 2000] assignments of the reads. Thereby, the advantage of CARMA3, which unveils both taxonomic as well as functional characterizations of metagenome reads, is exploited. Second, functional annotations in terms of Clusters of Orthologous Groups (COGs) [Tatusov et al., 2001] and Enzyme Commission (EC) [Kanehisa and Goto, 2000] numbers predicted on contigs by the Metatig pipeline are utilized for functional profiling. In the following sections, results generated by both approaches are illustrated.

Functional profiling based on reads

The feature of CARMA3 to combine taxonomic and functional assignments was utilized to identify microbes that are responsible for the methanogenesis in the investigated biogas plant. A central enzyme in the aceticlastic and hydrogenotrophic methanogenesis pathway is methyl-coenzyme M reductase (Mcr), which is composed of several subunits [Rastogi et al., 2008]. In gene-centric approaches, genes encoding the McrA subunit are employed as molecular markers for profiling methanogenesis relevant members, as the gene appears to be unique for methanogens [Rastogi et al., 2008]. In this thesis, a taxonomic profile is generated for reads encoding the *mcr* gene. Therefore, five Mcr subunits were manually categorized according to Pfam families. The taxonomic profile (Fig. 6.4) based on the selected Pfams representing the Mcr subunits in MetaSAMS confirmed former observations that species of the order *Methanomicrobiales* play a pivotal role in methane production by utilizing the hydrogenotrophic pathway [Jaenicke et al., 2011]. No organisms performing the aceticlastic methanogenesis, such as *Methanosarcina*, are present in the profile. These results infer a dominance of hydrogenotrophic methanogens in the studied biogas plant. In addition, three reads carrying fragments of the *mcr* gene were assigned to *Methanoculleus marisnigri*, which is also the most abundant archaeal species according to CARMA3 predictions (Fig. 6.3). Surprisingly, reads were identified that were allocated to *Bacteria* by means of CARMA3. These results are controversial, as *mcr* genes are only present in specific *Archaea* [Steinberg and Regan, 2008]. Hence, the functional or taxonomic results might be false assignments by CARMA3.



Summary for PFAM accessions PF02249,PF02745,PF02783,PF02241,PF04609,PF02505,PF02240:

Pfam accession	Pfam description
PF02249	Methyl-coenzyme M reductase alpha subunit, C-terminal domain (MCR_alpha)
PF02745	Methyl-coenzyme M reductase alpha subunit, N-terminal domain (MCR_alpha_N)
PF02783	Methyl-coenzyme M reductase beta subunit, N-terminal domain (MCR_beta_N)
PF02241	Methyl-coenzyme M reductase beta subunit, C-terminal domain (MCR_beta)
PF04609	Methyl-coenzyme M reductase operon protein C (MCR_C)
PF02505	Methyl-coenzyme M reductase operon protein D (MCR_D)
PF02240	Methyl-coenzyme M reductase gamma subunit (MCR_gamma)

Taxonomic distribution for PF02249,PF02745,PF02783,PF02241,PF04609,PF02505,PF02240:

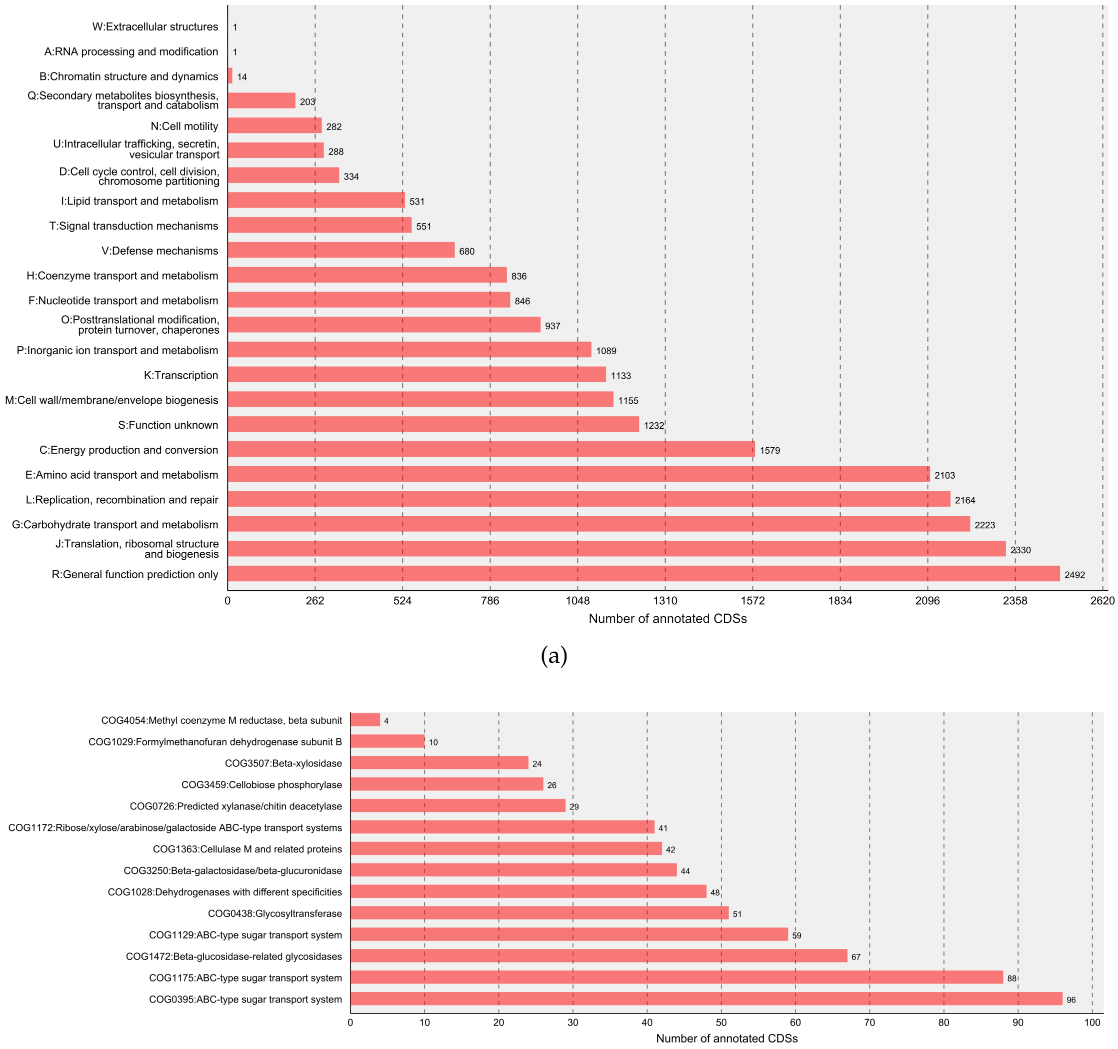
Domain	Phylum	Class	Order	Family	Genus	Species
Archaea (307)	Euryarchaeota (288)	Methanobacteria (1)	Methanobacteriales (1)	Methanobacteriaceae (1)		
		Methanomicrobia (279)	Methanomicrobiales (278)	Methanomicrobiaceae (261)	Methanoculleus (142)	Methanoculleus marisnigri (3)
Bacteria (9)	Bacteroidetes (2)					
	Firmicutes (4)	Clostridia (3)	Clostridiales (2)	Clostridiaceae (2)	Alkaliphilus (1)	
					Clostridium (1)	

Figure 6.4: A taxonomic profile of EGTs assigned to Mcr subunits in MetaSAMS: The subunits of Mcr were manually categorized according to Pfam families. A profile was created for the selected Pfam families in MetaSAMS. The functional and taxonomic results were deduced from CARMA3 assignments stored in the MetaSAMS database. For each rank, the taxa are listed with the number of identified EGTs in parentheses.

Functional profiling based on contigs

The Metatig pipeline performed gene prediction on 27,576 metagenome contigs of the biogas-producing microbes yielding 72,373 coding sequences (CDSs). After that, the identified CDSs were functionally annotated. Based thereon, functional profiles in terms of EC and COG numbers can be created in MetaSAMS. Moreover, KEGG pathways [Kanehisa and Goto, 2000] and functional COG categories can be deduced from the annotations.

In order to reveal a comprehensive view of the functions encoded by the biogas-producing community, a profile based on COG functional categories was built in MetaSAMS (Fig. 6.5a). Complex sugar polymers are degraded in the hydrolysis step in the biogas production process. In this context, the COG category "carbohydrate transport and metabolism" (G) is important. Indeed, CDSs were identified that encode relevant enzymes categorized into the functional group G. Moreover, "energy production and conversion" (C) is highly covered by annotated CDSs. This category includes enzymes that are essential during the acetogenesis step in the biogas production process. A fundamental COG category is "coenzyme transport and metabolism" (H), as it represents enzymes relevant for methanogenesis. CDSs assigned to this COG category were identified in the biogas community.



(a)

(b)

Figure 6.5: Functional profiles based on annotated CDSs: The Metatig pipeline creates functional annotations in terms of (a) COG functional categories and (b) COG annotations, which are predicted for translated CDSs in metagenome contigs. Only those COG annotations are shown that are associated with the stages of the anaerobic digestion process.

In the next step, the profile based on individual COGs was investigated in detail. MetaSAMS allows exporting selected COGs that might be of interest in SVG format (Fig 6.5b). The Metatig pipeline predicted 2,232 different COG numbers on translated CDSs with the largest COG being "ABC-type multidrug transport system" (COG1132) including 137 proteins. In particular, COGs related to sugar transport and degradation of carbohydrates are among the most abundant COGs (Fig 6.5b). As an example, 96 translated CDSs were assigned to the permease component of an "ABC-type sugar transport system" (COG0395), which is the third largest COG in the functional profile. The subunits of formylmethanofuran dehydrogenase (COG1029) and methyl coenzyme M reductase (COG4054), which are fundamental in the methanogenesis pathway, are also encoded by genes in the metagenome contigs.

6.2.3 Mapping of metagenome reads to the genome of *M. marisnigri* JR1 via the GenomeMapper

To examine the coverage of reference genomes or genes, metagenome sequences can be mapped against selected reference genomes based on similarity criteria. Therefore, a 'GenomeMapper' has been implemented in MetaSAMS, which represents the distribution of metagenome reads on a selected region. Moreover, each position in the reference region is linked to the NCBI sequence viewer of the corresponding genome. Thereby, it is possible to retrieve detailed regional or functional annotations of interesting regions.

Herein, the GenomeMapper was utilized to investigate the distribution of metagenome reads obtained from the studied fermentation sample on the genome of *M. marisnigri* JR1. As this species was identified in the taxonomic profile, it is expected that the genome is well covered. Overall, *M. marisnigri* JR1 is the best BLAST hit for 83,834 reads. In particular, the region encoding a central methanogenesis gene cluster is of interest. A visualization of the corresponding region (position from 577,137 to 589,552) in the GenomeMapper illustrates that the relevant genes are well covered by the metagenome reads (Fig. 6.6). Only one region, which is located in an intergenic region between Memar_0617 and Memar_0618, is sparsely covered by metagenome reads. These observations lead to the suggestion that the dominant methanogens of the studied biogas plant possess methanogenesis genes that are highly related to *M. marisnigri* JR1.

6.2.4 Identification of variant genes encoding the B subunit of methyl-coenzyme M reductase

As indicated in Figure 6.4, two archaeal families, namely *Methanobacteriaceae* and *Methanomicrobiaceae*, carry genes encoding Mcr subunits, which are important in the methanogenesis process. In this regard, the gene encoding the subunit McrB was used as a reference to identify variants. For this purpose, a hidden Markov model (HMM) [Durbin et al., 2006] was built from an alignment modeling the *mcrB* gene on a metatig.

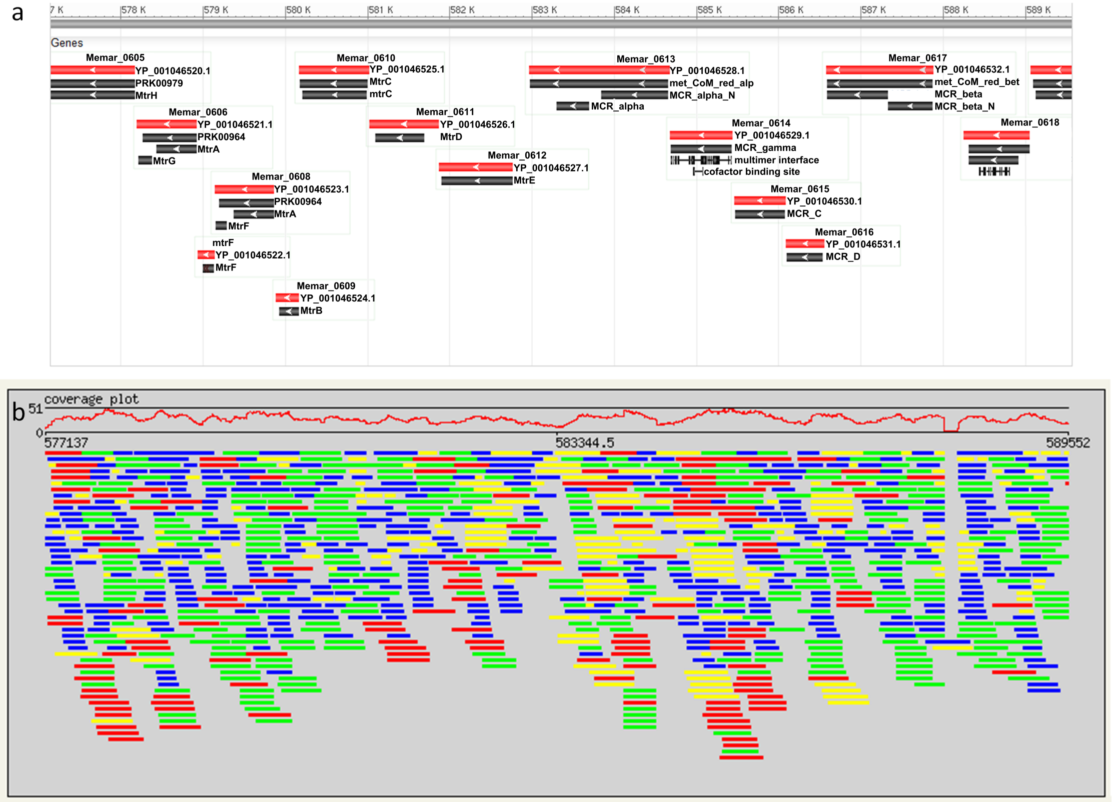


Figure 6.6: Visualization of the GenomeMapper showing a methanogenesis gene cluster of *M. marisnigri* JR1: A central methanogenesis region on the genome of *Methanoculleus marisnigri* JR1 (NC_009051.1, from nucleotide position 577,137 to 589,552) is presented (a) in the NCBI sequence viewer and (b) in the MetaSAMS GenomeMapper. The NCBI sequence viewer shows the annotations of the genes, whereas the GenomeMapper displays the arrangement of reads on the genome region. All genes encoding methanogenic enzymes are covered by metagenome sequences. E-value: red $< 10^{-150}$, green $< 10^{-100}$, blue $< 10^{-50}$, yellow $\geq 10^{-50}$, only reads with an E-value of $\leq 10^{-10}$ are shown.

MetaSAMS supplies the alignment, as it stores the aligned reads that assemble each metatig. The system captured only the reads from the metatig alignment that overlap with the *mcrB* gene. The alignment was used as an input for the HMM-interface in MetaSAMS, which automatically carried out three successive steps. First, a profile HMM of the mcrB gene was built based on the provided alignment. Next, the model was applied on the biogas metagenome reads to extract further *mcrB* gene fragments. Finally, the matching reads were aligned to the model in MetaSAMS by using the HMMER3 package. Finally, the alignment of the identified reads was retrieved and the corresponding genes of *M. marisnigri* JR1 were manually added. The alignment revealed five variants encoding McrB (Fig. 6.7). In *Methanoculleus marisnigri* JR1, the *mcr* cluster is partly duplicated. Two of the variant genes are similar to Memar_0375 (Fig. 6.7 a-b), while the remaining three genes exhibit a high similarity to Memar_0617 (Fig. 6.7 c-e).

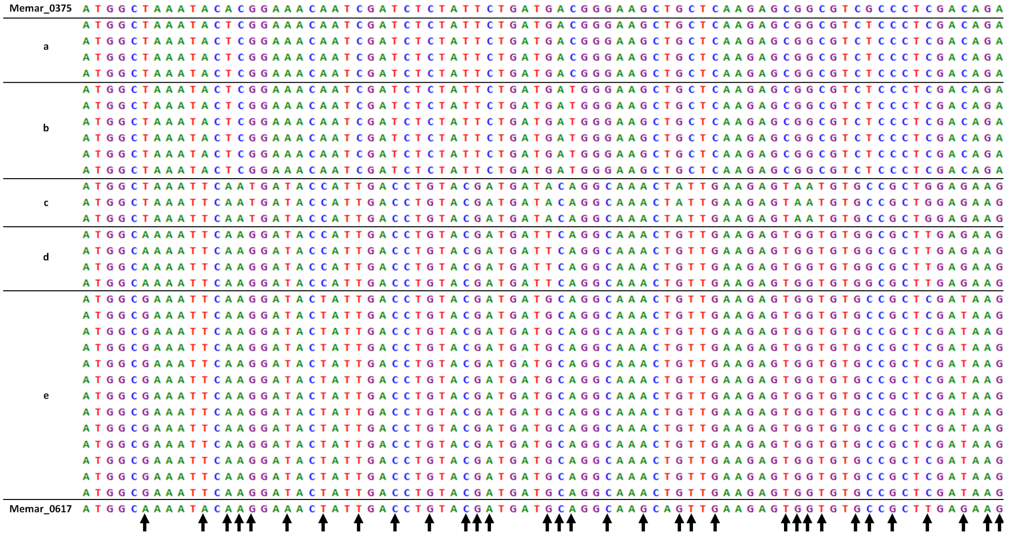

Memar_0375 ATGGCTAAATACACGGAAACAATCGATCTCTATTCTGATGACGGGAAGCTGCTCAAGAGCGGCGTCGCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGACGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
a
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGACGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGACGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGACGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
b
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATACTCGGAAACAATCGATCTCTATTCTGATGATGGGAAGCTGCTCAAGAGCGGCGTCTCCCTCGACAGA
ATGGCTAAATTCAATGATACCATTGACCTGTACGATGATACAGGCAAACTATTGAAGAGTAATGTGCCGCTGGAGAAG
c
ATGGCTAAATTCAATGATACCATTGACCTGTACGATGATACAGGCAAACTATTGAAGAGTAATGTGCCGCTGGAGAAG
ATGGCTAAATTCAATGATACCATTGACCTGTACGATGATACAGGCAAACTATTGAAGAGTAATGTGCCGCTGGAGAAG
ATGGCAAAATTCAAGGATACCATTGACCTGTACGATGATTCAGGCAAACTGTTGAAGAGTGGTGTGGCGCTTGAGAAG
d
ATGGCAAAATTCAAGGATACCATTGACCTGTACGATGATTCAGGCAAACTGTTGAAGAGTGGTGTGGCGCTTGAGAAG
ATGGCAAAATTCAAGGATACCATTGACCTGTACGATGATTCAGGCAAACTGTTGAAGAGTGGTGTGGCGCTTGAGAAG
ATGGCAAAATTCAAGGATACCATTGACCTGTACGATGATTCAGGCAAACTGTTGAAGAGTGGTGTGGCGCTTGAGAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
e
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
ATGGCGAAATTCAAGGATACTATTGACCTGTACGATGATGCAGGCAAACTGTTGAAGAGTGGTGTGCCGCTCGATAAG
Memar_0617 ATGGCAAAATACAAGGAAACTATTGACCTGTACGATGATGCAGGCAAGCAGTTGAAGAGTGGTGTGCCGCTTGAGAAG

Figure 6.7: Alignment of reads representing the *mcrB* gene fragment: An HMM-based search was applied in MetaSAMS using an HMM modeling the *mcrB* gene. The pipeline generated an alignment of matching reads. Reference sequences of *M. marisnigri* JR1 were added to the alignment. The arrows indicate single nucleotide variations in the alignment of the reads. In total, 5 different variants can be deduced from the alignment. Only the first 88 bases of the *mcrB* gene are shown.

6.3 Analysis of 16S rDNA amplicon sequences from a community of a biogas plant

As approximately only 0.3% of the metagenome data represents 16S rRNA genes, a gene-centric approach was carried out to achieve a deeper resolution of the taxonomic composition of the biogas-producing community. Based thereon, the most dominant species can be unveiled that might be relevant for the anaerobic digestion process. In particular, phylogenetic analyses based on 16S rRNA gene fragments can give insight into the relation of different taxa represented by the amplicon and related known reference species.

16S rDNA amplicon sequences were generated from microbes obtained from the biogas plant that was analyzed by the metagenome approach described above. Sampling, DNA extraction and PCR amplification were previously described [Zakrzewski et al., 2012]. Amplicons spanning the third and fourth variable (V3, V4) regions of the 16S rRNA gene were sequenced on the GS FLX system using Multiplex Identifiers (MIDs) and the Titanium chemistry. For the amplicon sequences generated from the biogas-producing community, the MID tag 'CAGTAGACGT' was used. The analysis of the data was carried out using AMPLA.

6.3.1 Processing of raw amplicon sequences obtained from a biogas-producing community

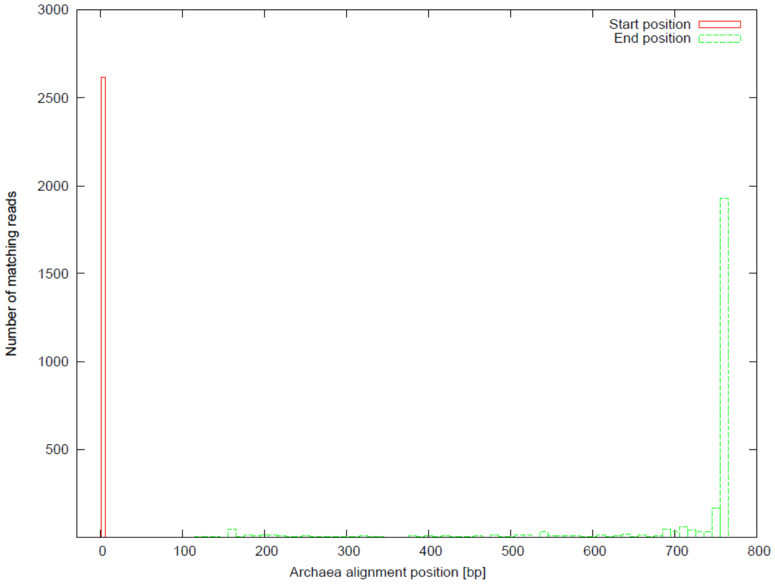
Extracting the sequences according to the MID tag sequence 'CAGTAGACGT' yielded 25,805 reads with an average length of 328 bp and 8,451,545 sequenced bases. Reads with an average quality below 20 and ambiguous bases (including N) were discarded resulting in 23,654 reads (Tab. 6.1). The subsequent trimming procedure was carried out using QIIME [Caporaso et al., 2010]. Thereby, MID sequences were removed from the reads. Moreover, sequences of the forward primer 341F_35 (Sequence: CCTAYGGGRBG-CASCAG) and reverse primer 806R (Sequence: GGACTACNNGGGTATCTAAT) were trimmed allowing two mismatches. Reads without recognizable primer sequences were discarded from the downstream analysis.

Table 6.1: Overview of the filtered sequences during the amplicon processing using the AMPLA pipeline

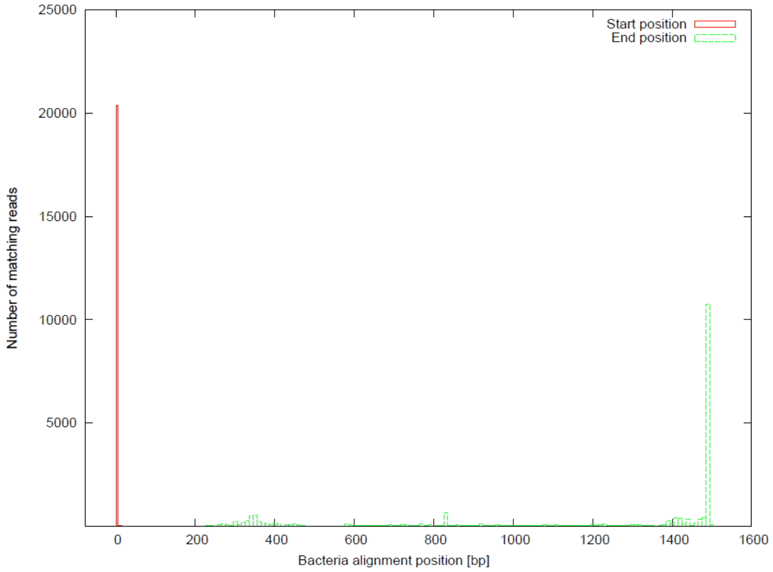
Processing step	Number of removed reads[1]	Number of remaining reads
Raw data		25,805
Length below 50 bp	1,329	24,476
Mean quality score below 20	39	24,437
Ambiguous bases	783	23,654
Primer removal	10,920	12,734
SLP	2,170	10,564
UCHIME	928	9,636

[1] The number refers to the amount of reads remaining in the preceding step

Since a 16S rRNA gene fragment spanning the hypervariable V3 and V4 regions in *Escherichia coli* is 466 bases long [Neefs et al., 1991], the 454 pyrosequencing procedure might not reach the reverse primer sequence. The location of the 16S rDNA amplicon sequence on the corresponding gene was determined by searching for hits in a database containing profile hidden Markov models (HMMs) for archaeal and bacterial 16S rRNA gene fragments. Therefore, aligned bacterial and archaeal 16S rRNA gene sequences were downloaded from the RDP database (release 10.28) [Cole et al., 2003] and trimmed



(a) alignment to archaeal profile HMM



(b) alignment to bacterial profile HMM

Figure 6.8: Mapping of 16S rDNA sequences against archaeal and bacterial alignments: The 16S rDNA amplicon sequences were aligned to the (a) archaeal or (b) bacterial reference alignments, which were obtained from the RDP database and trimmed for the variable V3 and V4 region. The start and stop positions for each read were collected and mapped in a histogram. Most of the sequences do not reach the reverse primer. As the RDP database contains 16S rRNA gene sequences in an aligned version, the scale of the x-axis is longer than the corresponding unaligned sequence region.

for the region spanning V3 and V4. The modified bacterial and archaeal alignments were used as a basis for building profile HMMs using the HMMER package [Eddy, 2011]. Finally, the amplicon reads were searched for matches to one of the available models using an E-value cutoff of 10^{-10}. In total, the search separated the amplicon sequences to 2,770 archaeal and 20,882 bacterial sets. Each read set was then aligned to the corresponding model. Thereafter, the start and stop positions for each read were deduced from the alignments and visualized in a histogram (Fig 6.8). Indeed, mapping the sequences against 16S rRNA gene references reveals that most of the sequences end before the reverse primer sequence. Around 13,000 sequences were excluded due to the absence of a reverse primer sequence. However, the reverse primer sequence is an important indicator to determine the quality of the reads and was therefore taken into account in the quality control step.

To avoid an overestimation of the number of operational taxonomic units (OTUs) in the data, reads containing sequencing errors were identified using a single linkage preclustering (SLP) [Huse et al., 2010] implementation in MOTHUR [Schloss et al., 2009]. This step yielded 10,564 amplicon sequences (Tab. 6.1), which were then examined for chimeric features using UCHIME [Edgar et al., 2011]. Approximately 9% chimeric sequences were identified, which is in accordance with estimations [Schloss et al., 2011]. After quality control, 9,636 16S rDNA sequences remained that were used for subsequent analysis.

6.3.2 OTU-based analysis of the biogas-producing community

To investigate the number of taxonomic groups in the biogas-producing microbial community, 16S rDNA sequences were clustered into operational taxonomic units (OTUs) using UCLUST [Edgar, 2010]. From the collection of 9,636 quality-filtered sequences, 2,546 OTUs were estimated with an identity value of 97%, which is in accordance with species level [Schloss and Handelsman, 2005]. Of these estimated OTUs, 1,782 are OTUs including only one sequence (singletons). After singleton removal, 764 OTUs remained, which represent 82% of the 9,636 studied amplicon sequences. Rarefaction analysis based on the observed OTUs was carried out to estimate the coverage of the sequenced fermentation sample (Fig. 6.9). As the sequences of singletons are assumed to contain pyrosequencing errors or to be composed of several organisms as a result of chimera formation during PCR [Reeder and Knight, 2009], an additional rarefaction curve was calculated based on OTUs without singletons.
The rarefaction curve including singleton OTUs has a steep slope indicating either the presence of artifacts in the data or that more sequencing is required to reach a full coverage of the community. However, the rarefaction curve based on the OTU counts after singleton removal nearly reaches an asymptotic trend suggesting a complete coverage of the microbial community.

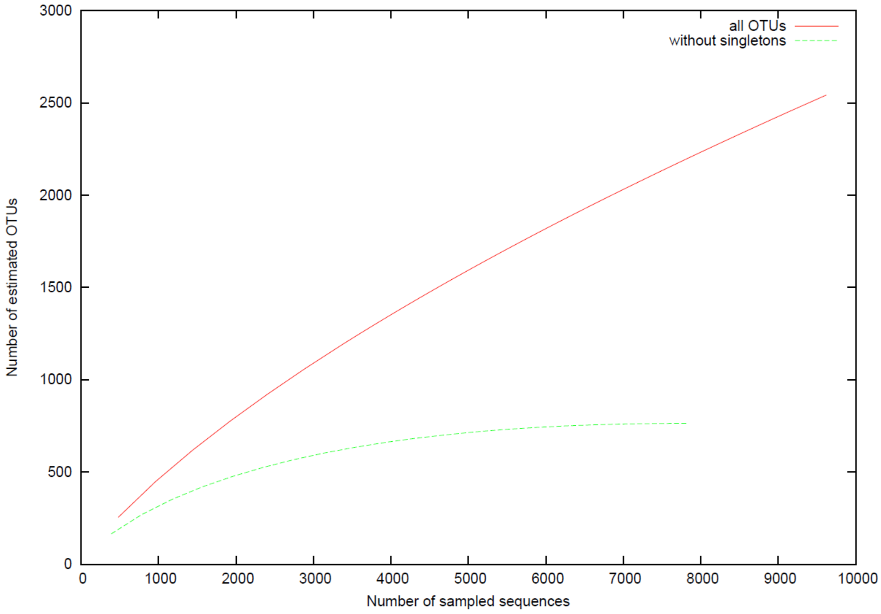


Figure 6.9: Rarefaction analysis of operational taxonomic units (OTUs) clustered from 16S rDNA amplicons: Rarefaction curves describe the dependence of observing novel OTUs as a function of sampling efforts. The rarefaction analysis was carried out for OTU estimations based on an identity of 97% with and without singletons.

As the ten largest OTUs cover approximately one fourth of the studied dataset, the longest sequence for each of the ten largest OTUs was selected as a representative for the corresponding OTU. Thereafter, the representative sequences were classified using the RDP Classifier. At the same time, reference sequences were searched in the NCBI nucleotide database (nt) using BLAST with default settings. Only few representative sequences could be assigned to lower taxonomic ranks by the RDP Classifier suggesting the presence of so far unknown bacterial species in the biogas plant (Tab. 6.2). The largest 16S rDNA OTU, which accounts for 10% of the sequences within the analyzed dataset, was predicted to stem from close relatives of species belonging to *Methanoculleus* (Tab. 6.2, no. 1). This analysis shows the dominance of the genus *Methanoculleus* and signifies that methane may mainly be produced by *Methanoculleus* species. The BLAST search retrieved a hit to *Methanoculleus bourgensis* MS2 with a high similarity (98%). Moreover, an identical sequence was identified in the 16S rRNA clone library constructed from the same biogas plant [Kröber et al., 2009].

Many representative sequences could only be classified to the class *Clostridia* or the order *Clostridiales* by the RDP Classifier with a confidence value above 0.8 showing that a majority of bacterial species residing in the biogas reactor are still not characterized. This observation was confirmed by the BLAST approach, as for most of the OTUs no

Table 6.2: The ten largest OTUs and their taxonomic characterization based on the RDP Classifier and BLAST

OTU number	OTU member	RDP classification	Best BLAST hit[1] Identity	Coverage	Description	Best BLAST hit[2]
1	1036	*Methanoculleus* (genus)	98%	100%	*Methanoculleus bourgensis*	FJ205773 [Kröber et al., 2009]
2	493	*Clostridia* (class)	88%	99%	*Natranaerobaculum magadiensis*	FJ205808 [Kröber et al., 2009], HQ156187 (unpublished)
3	488	*Clostridia* (class)	88%	100%	*Clostridium sp.* PPf35E10	FJ205846 [Kröber et al., 2009]
4	438	*Clostridium* (genus)	95%	100%	*Clostridium aciditolerans* JW/YJL-B3	FJ205850 [Kröber et al., 2009]
5	364	*Clostridiales* (order)	90%	100%	*Garciella nitratireducens*	HQ155155, HQ156167 (unpublished)
6	199	*Porphyromonadaceae* (family)	90%	100%	*Proteiniphilum acetatigenes* TB107	CU919517 [Riviére et al., 2009]
7	157	*Bacteroidetes* (phylum)	79%	100%	*Adhaeribacter aerophilus*	EF559054 (unpublished)
8	142	*Bacteria* (superkingdom)	88%	99%	*Gelria glutamica* TGO	FJ205823 [Kröber et al., 2009]
9	137	*Clostridia* (class)	88%	99%	*Tissierella* sp. LBN 292	HQ155127 (unpublished)
10	114	*Alkaliflexus* (genus)	99%	100%	*Ruminofilibacter xylanolyticum*	FJ205818 [Kröber et al., 2009]

[1] excluding environmental sequences

[2] including environmental sequences, with 100% identity and read coverage

close reference sequences from culturable microbes were available in the nt database. For example, representatives for OTU number 2 and 3 (Tab. 6.2, no. 2, no. 3), which cover 493 and 488 amplicon sequences, respectively, exhibit only low similarities (88%) to reference sequences obtained from cultured species. Nevertheless, identical fragments from environmental samples were identified for both representatives in the nt database. The matching sequences originate from 16S rRNA clones obtained from the same biogas plant (accession: FJ205808 and FJ205846) [Kröber et al., 2009]. The presence of identical sequences confirms that the representative sequences of these OTUs are no artifacts but rather originate from an organism.
Similarly, OTU number 4 is identical to a subsequence of a 16S rRNA clone (accession FJ205850) from the same biogas plant (Tab. 6.2, no. 4). In addition, the representative sequence for OTU number 4 was assigned to *Clostridium* on the rank genus by the RDP Classifier. The best BLAST hit to a culturable species is to *Clostridium aciditolerans* [Lee et al., 2007], which is an obligately anaerobic, moderately acid-tolerant bacterium and produces acetate, butyrate and ethanol as end products from glucose [Lee et al., 2007].

The sequence representing OTU number 5 (Tab. 6.2, no. 5) covering 364 reads could only be classified to *Clostridiales* on the rank order and presented a low identity percentage (90%) with a culturable species, namely *Garciella nitratireducens*, which ferments several sugars and organic acids [Miranda-Tello et al., 2003]. The sequence is identical to 16S rRNA clone sequences obtained from biogas plants treating pig manure (accession HQ155155) and chopped rice straw (accession HQ156167). OTU number 6 (Tab. 6.2, no. 6), which includes 185 sequences, was assigned to the family *Porphyromonadaceae* by the RDP Classifier and forms the largest OTU within the phylum *Bacteroidetes*. BLAST results suggested a similarity to *Proteiniphilum acetatigenes*, which was identified in a methanogenic propionate-degrading mixture obtained from an upflow anaerobic sludge blanket reactor and was associated with the degradation of amino acids [Chen and Dong, 2005].

It was not possible to assign the representative sequence of OTU number 7 (Tab. 6.2, no. 7), which includes 157 reads, to low taxonomic ranks. The RDP Classifier affiliated the sequence to the phylum *Bacteroidetes*, while BLAST presented a low similarity (79%) to *Adhaeribacter aerophilus*. Nevertheless, identical 16S rRNA clone sequences occurred in other environmental samples. A matching clone sequence (accession EF559054) originates from an anaerobic digester treating municipal solid water.
For the representative amplicon of OTU number 8, no low taxonomic assignment based on the RDP Classifier and no similar reference sequence of a known species based on BLAST (Tab. 6.2, no. 8) were available. However, the sequence is identical to uncultured 16S rRNA clones obtained from the same biogas plant (accession FJ205823). In addition, the sequence representing OTU number 9 (Tab. 6.2, no. 9) was only classified to the rank class by the RDP Classifier. Identical reference sequences were detected in uncultured clones obtained from a biogas digester treating pig measure and rice straw (accession HQ155127).

Finally, OTU number 10 (Tab. 6.2, no. 10) represents according to the RDP Classifier the genus *Alkaliflexus* and was assigned to a defined species, namely *Ruminofilibacter xylanolyticum*, using the BLAST-based approach. The matching species is a rumen bacterium involved in the digestion of xylan [Weiss et al., 2011] and was also identified among 16S rRNA clones of the same biogas plant (accession FJ205818).

In total, 4 of the 10 representative sequences are not covered by 16S rRNA clones from the same biogas plant showing the advantages of high-throughput sequencing. As identical sequences occur in other biogas plants, the organisms may play an important role in the anaerobic fermentation process.

6.3.3 Taxonomic profiling of the biogas-producing community based on 16S rDNA amplicon sequences

Next, the taxonomic structure of the whole community was analyzed by applying the RDP Classifier on the quality-filtered 16S rDNA amplicon sequences. The classifier assigned 100% and 94% of the sequences to the rank superkingdom and phylum, respectively. Figure 6.10 illustrates that *Bacteria* dominate within the biogas reactor with 88% of all reads, while 12% of the sequences were classified to *Archaea*.

16S rDNA amplicon sequences were predominantly classified to the phylum *Firmicutes* (73%). Most of these belong to the class *Clostridia* (91%) and *Bacilli* (3%). The phylum *Euryarchaeota* is represented with 12% of the analyzed dataset. Approximately, 8% of the amplicon sequences were assigned to the phylum *Bacteroidetes* with *Porphyromonadaceae* as the dominant family. *Synergistetes*, *Proteobacteria* and *Actinobacteria* are present each with 1% of the 16S rDNA sequences.

87% of the sequences could be assigned to a taxon on the rank class with *Clostridia* (66%) and *Methanomicrobia* (12%) being the most dominant taxa. Moreover, 36% of the 16S rRNA gene fragments could be assigned to a taxon on the level family. *Methanomicrobiaceae* (12%), *Clostridiaceae* (8%) and *Porphyromonadaceae* (3%) provide the largest number of sequences on this rank. These families are also present among the ten largest OTUs observed in the same dataset (Section 6.3.2).

Finally, only 28% of the 16S rDNA amplicon sequences could be classified at the taxonomic rank genus. Hence, many genera residing in the biogas reactor are still unknown. As observed in the OTU analysis, *Methanoculleus* is the most prevalent archaeal genus. In total, 1,144 sequences were assigned to this genus accounting for 12% of the filtered sequences and 42% of all reads classified on rank genus. With 7% of the analyzed 16S rDNA amplicon sequences, *Clostridium* is the second largest genus followed by *Alkaliflexus* and *Acetivibrio*, each with 2% of all sequences. *Alkaliflexus* is also among the 10 largest OTUs observed in microbial community from the biogas plant (Tab. 6.2, no. 10). The species *Acetivibrio cellulolyticus*, which belongs to the latter genus,

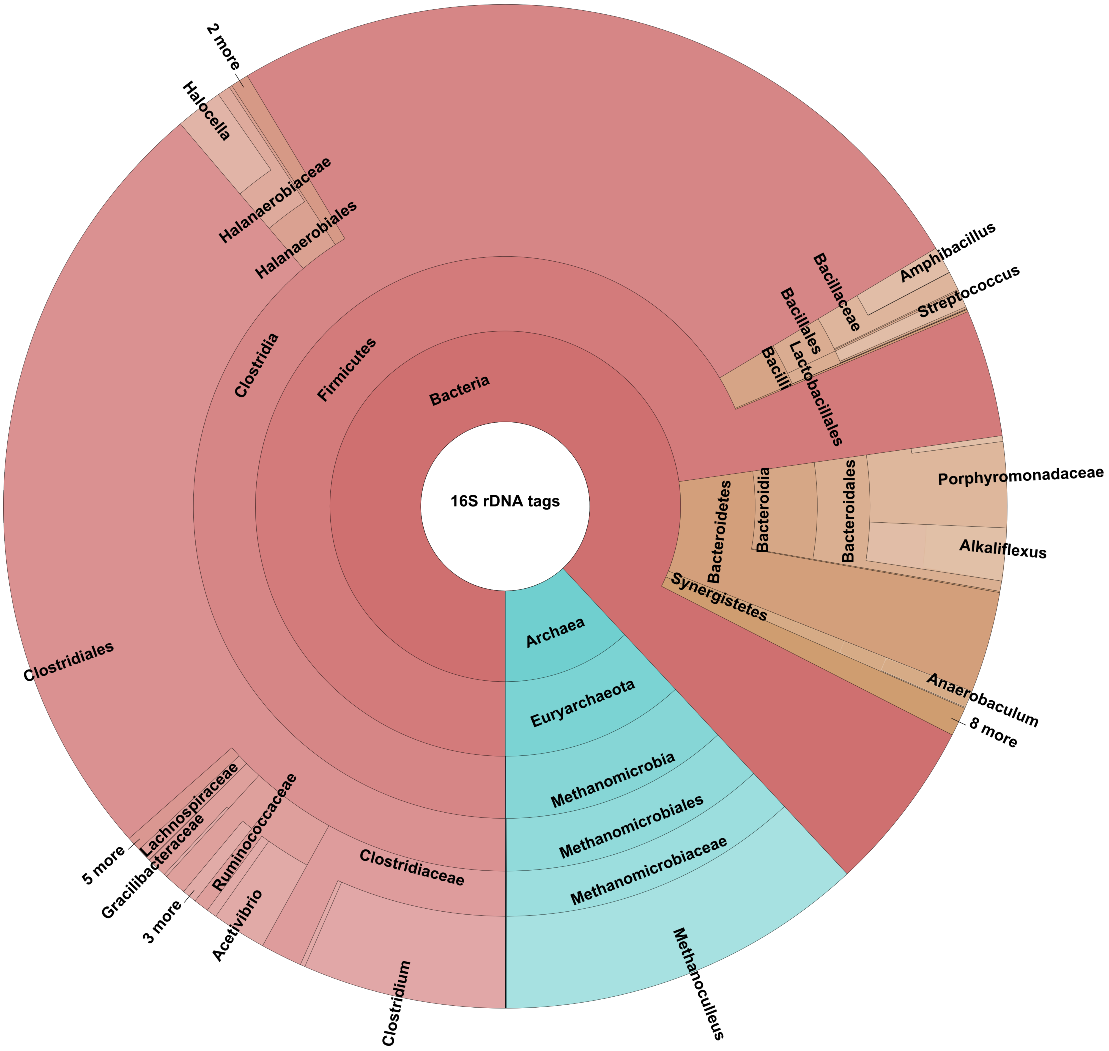


Figure 6.10: Taxonomic profiling of the biogas plant community based on 16S rDNA amplicon sequences: 16S rDNA sequences were classified using the RDP Classifier. For the profile, only assignments with a confidence of 0.8 were utilized. The profile was visualized using Krona [Ondov et al., 2011].

was firstly isolated from a sewage sludge culture and is described as a mesophilic, cellulolytic and anaerobe bacterium [Saddler and Khan, 1981].

6.3.4 Comparative taxonomic analysis of DNA-based profiles created for the biogas-producing community

The profiles based on identified and characterized metagenome 16S rRNA gene fragments as well as mRNA tags were compared to the taxonomic composition deduced from the 16S rDNA amplicon dataset. The taxonomic characterization of the metagenome 16S rRNA gene fragments and EGTs were exported from the MetaSAMS system, which assigned taxonomies by applying the RDP Classifier and CARMA3, respectively. All profiles confirm a dominance of members belonging to *Firmicutes* followed by *Euryarchaeota* and *Bacteroidetes* (Fig. 6.11). The classification based on 16S rDNA revealed only a low number of archaeal sequences (2%) as compared to the other approaches (7%-12%). As only about 0.3% of the sequences were identified coding for 16S rRNA genes, the taxonomic profile may be biased. The majority of the metagenome sequences in both profiles have no references on the level phylum, whereas only 6% of the 16S rDNA amplicon sequences are unknown demonstrating the advantage of using hypervariable regions to get extensive knowledge about the taxonomic structure of a microbial community.

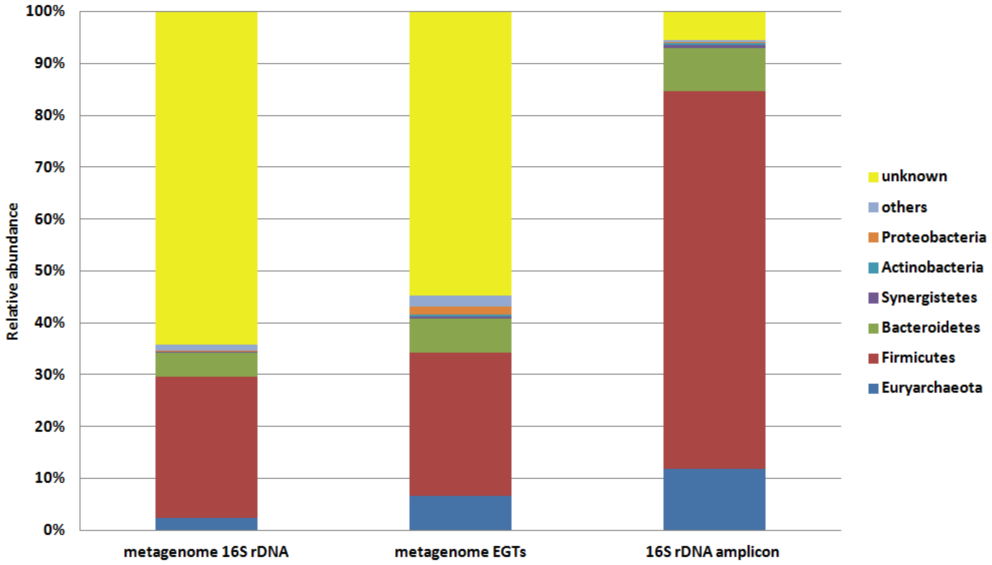


Figure 6.11: Comparison of taxonomic profiles generated from the DNA-based approaches: The taxonomic composition on phylum level was calculated based on metagenome 16S rRNA gene fragments, metagenome EGTs and 16S rDNA amplicon sequences. The metagenome 16S rRNA and amplicon sequences were analyzed using the RDP Classifier, whereas CARMA3 was applied for taxonomic characterizations of metagenome EGTs.

6.3.5 Phylogenetic analysis of 16S rDNA amplicon sequences classified as *Archaea* and *Synergistetes*

Phylogenetic trees represent evolutionary relationships between sequences. The most widely used sequences for phylogenetic tree reconstruction are 16S rRNA genes. The amplicon sequences, which were analyzed in this thesis, cover the V3 and V4 hypervariable regions of 16S rRNA genes. Therefore, they are suitable for the generation of phylogenetic trees. The aim of the following sections is a phylogenetic characterization of interesting taxa by an analysis integrating the amplicon sequences.
Phylogenetic analyses of OTUs belonging to *Firmicutes* and *Bacteroidetes* were exhaustively described for different anaerobic environments and settings [Li et al., 2009, Cardinali-Rezende et al., 2009, Weiss et al., 2009]. Because of this, phylogenetic examination within this thesis was focused on the methanogenic *Archaea* and the relatively new phylum *Synergistetes*. For this purpose, representative sequences of OTUs, which were assigned to *Archaea* or *Synergistetes* with a confidence of at least 0.8 by the RDP Classifier, were extracted and phylogenetically analyzed by performing bootstrap analysis [Tamura et al., 2007] with the neighbor-joining method [Saitou and Nei, 1987].

Phylogenetic analysis of archaeal 16S rDNA amplicon sequences

Since the taxonomic analysis inferred that *Methanoculleus* is the most abundant archaeal genus in the biogas-producing microbial community (Section 6.3.3), representative OTU sequences assigned to the superkingdom *Archaea* by the RDP Classifier were extracted and phylogenetically characterized. In total, 11 representative sequences accounting for 1,159 reads were assigned to *Archaea* with a confidence value of at least 0.8.

The archaeal 16S rDNA sequences have a limited diversity distribution (Fig 6.12). The third largest archaeal OTU contains five sequences and is represented by the 16S rDNA amplicon ARCH01, which is in a phylogentic cluster with *Methanoculleus bourgensis* and *Methanoculleus olentangyi* (Fig 6.12). Both *Methanoculleus* species are synonyms on the basis of their genotypic and phylogenetic features [Asakawa and Nagaoka, 2003]. The 16S rDNA sequence of ARCH01 has a high similarity (98%) to the sequence of the species *M. bourgensis* MS2. The presence of *Methanoculleus bourgensis* or a related species in 16S rDNA amplicon sequences is in agreement with previous analyses based on 16S rRNA clones obtained from the same biogas plant [Kröber et al., 2009]. The representative sequence of ARCH01 is highly covered by 16S rRNA clone sequences affiliated to *M. bourgensis*.

A further phylogenetic cluster is formed by the amplicons ARCH03 and ARCH04, which are in close proximity to the reference sequences of *Methanoculleus* species. The amplicons represent OTUs that cover each 1,038 and 88 sequences. There is so far no closely related reference species known for this phylogenetic cluster. Nevertheless, it was annotated as an unknown *Methanoculleus* group since it is in close proximity to described *Methanoculleus* species. The representative sequences of ARCH03 and ARCH04 are almost identical (99%-100%) to the 16S rRNA clones A52 (accession: FJ205773) and

A12 (accession: FJ205758) obtained from the same biogas plant [Kröber et al., 2009], which were likewise related to an unknown *Methanoculleus* group by the authors.

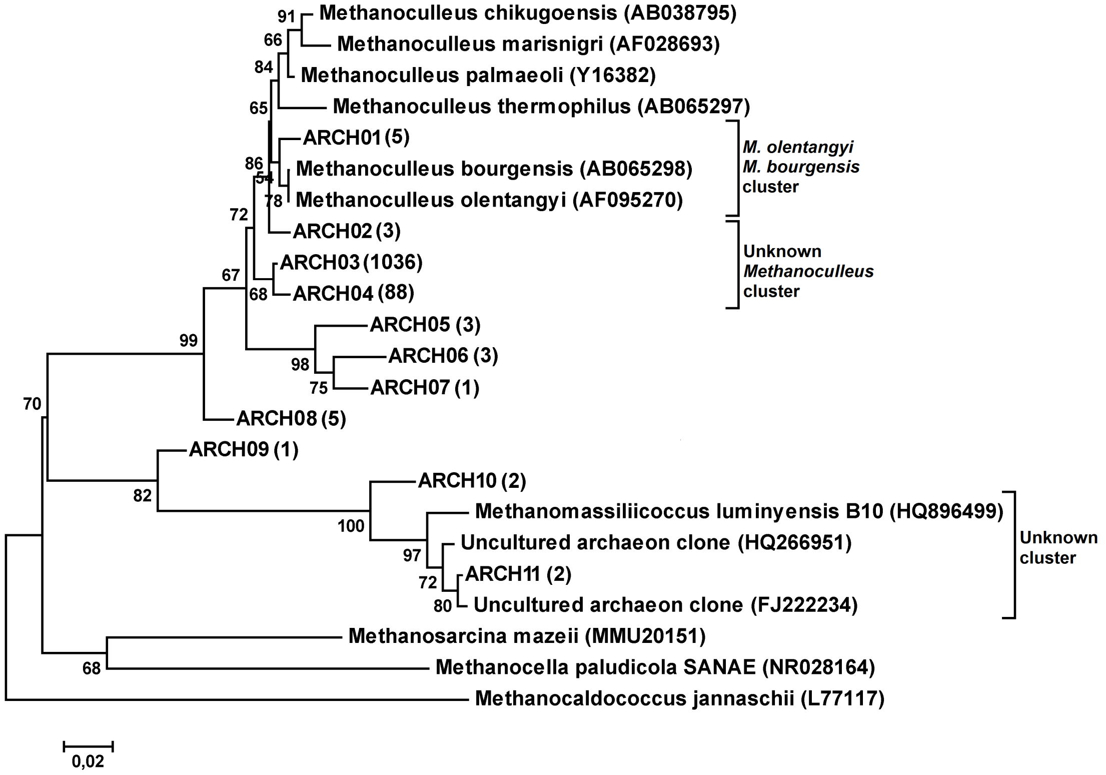

91
66
84
65
86
54
78
72
67
68
99
98
75
70
82
100
97
72
80
68
Methanoculleus chikugoensis (AB038795)
Methanoculleus marisnigri (AF028693)
Methanoculleus palmaeoli (Y16382)
Methanoculleus thermophilus (AB065297)
ARCH01 (5)
Methanoculleus bourgensis (AB065298)
Methanoculleus olentangyi (AF095270)
ARCH02 (3)
ARCH03 (1036)
ARCH04 (88)
ARCH05 (3)
ARCH06 (3)
ARCH07 (1)
ARCH08 (5)
ARCH09 (1)
ARCH10 (2)
Methanomassiliicoccus luminyensis B10 (HQ896499)
Uncultured archaeon clone (HQ266951)
ARCH11 (2)
Uncultured archaeon clone (FJ222234)
Methanosarcina mazeii (MMU20151)
Methanocella paludicola SANAE (NR028164)
Methanocaldococcus jannaschii (L77117)
M. olentangyi
M. bourgensis
cluster
Unknown
Methanoculleus
cluster
Unknown
cluster
0,02

Figure 6.12: Phylogenetic tree for representative 16S rDNA sequences assigned to the superkingdom *Archaea* based on neighbor-joining analysis: Sequences representing OTUs previously assigned to *Archaea* were used for phylogenetic tree reconstruction using neighbor-joining analysis. The evolutionary distances were computed using the Jukes-Cantor method. Bootstrap values of 1,000 replications are noted at each branch. *Methanocaldococcus jannaschii* was used as an outgroup. The representative read name is noted for each archaeal OTU. The number in parentheses indicates the amount of sequences assigned to the corresponding OTU. The accession numbers for reference strains and clones are shown in parentheses. The scale bar represents 2% nucleotide substitution.

One phylogenetic cluster, which includes the two representative amplicons ARCH10 and ARCH11, is located outside the phylogenetic cluster formed by *Methanoculleus* species. The phylogenetic tree distinctly affiliated the sequence to the recently identified species *Methanomassiliicoccus luminyensis* B10 [Dridi et al., 2012]. *M. luminyensis* is present in the human gut microbiome, but the prevalence of this species is unknown. Unfortunately, a closely related archaeal reference with species assignments is not described. However, the representative sequence clustered with an unknown archaeal clone (accession: FJ222234), which originates from an agricultural biogas plant supplied with water, maize silage and barley grains [Nettmann et al., 2010]. The representative amplicon ARCH11, which has a length of 384 bases, is completely covered by the sequence of the uncultured archaeon (accession: FJ222234) and differs only by two deletions. In addition, the fragment shows a high similarity (96%) to sequences (accessions: HQ266951, HQ266939, HQ266925) obtained from an Italian rice field soil [Liu and Conrad, 2011]. The distribution of similar sequences in various habitats suggests a wide occurrence of species related to *M. luminyensis*. No 16S rRNA clones carrying a similar sequence were discovered in the clone library created from the same biogas plant.

Finally, for most of the representative sequences, no known archaeal references were available. A search for references in the NCBI nucleotide database revealed that substrings of some representative reads (ARCH05, ARCH06, ARCH07, ARCH09) matched microbial genomes of different species. This observation infers that some sequences might be chimeric, which were not detected using UCHIME.

Phylogenetic analysis of 16S rDNA amplicon sequences classified as *Synergistetes*

A phylogenetic tree was generated for the fourth largest taxonomic phylum (1 % of the 16S rDNA amplicon sequences) representing the novel phylum *Synergistetes*. Species classified as *Synergistetes* were not present in previously deduced taxonomic profiles for the analyzed biogas-producing community, because many 'Synergistes' taxa have been misallocated to other phyla, mainly *Firmicutes*, in prior studies [Vartoukian et al., 2007].

Based on 14 representative OTU sequences assigned to the phylum *Synergistetes*, a phylogenetic tree was constructed (Fig. 6.13). The sequences represent 56 amplicon reads. The phylogenetic tree is composed of two defined *Synergistetes* clusters. One phylogenetic cluster includes two *Anaerobaculum* reference strains and the amplicon SYN01, which represents most of the identified *Synergistetes* sequences. Species of this genus ferment a range of organic acids, amino acids and a limited number of carbohydrates [Rees et al., 1997]. Utilization of glucose and malate by species of *Anaerobaculum* was enhanced in the presence of the methanogen *Methanothermobacter thermoautotrophicus* [Menes and Muxí, 2002].
The second phylogenetic cluster with one amplicon sequence is characterized by *Aminobacterium colombiense*, which was firstly identified in an anaerobic lagoon of a dairy wastewater treatment plant [Baena et al., 1998]. *Aminobacterium colombiense* ferments pyruvate, amino acids but is not able to use carbohydrates. In a mixed community including *Methanobacterium formicium* several other amino acids were utilized by *Aminobacterium colombiense*. For the remaining sequences no references were recognized suggesting either so far unknown species or artifacts.

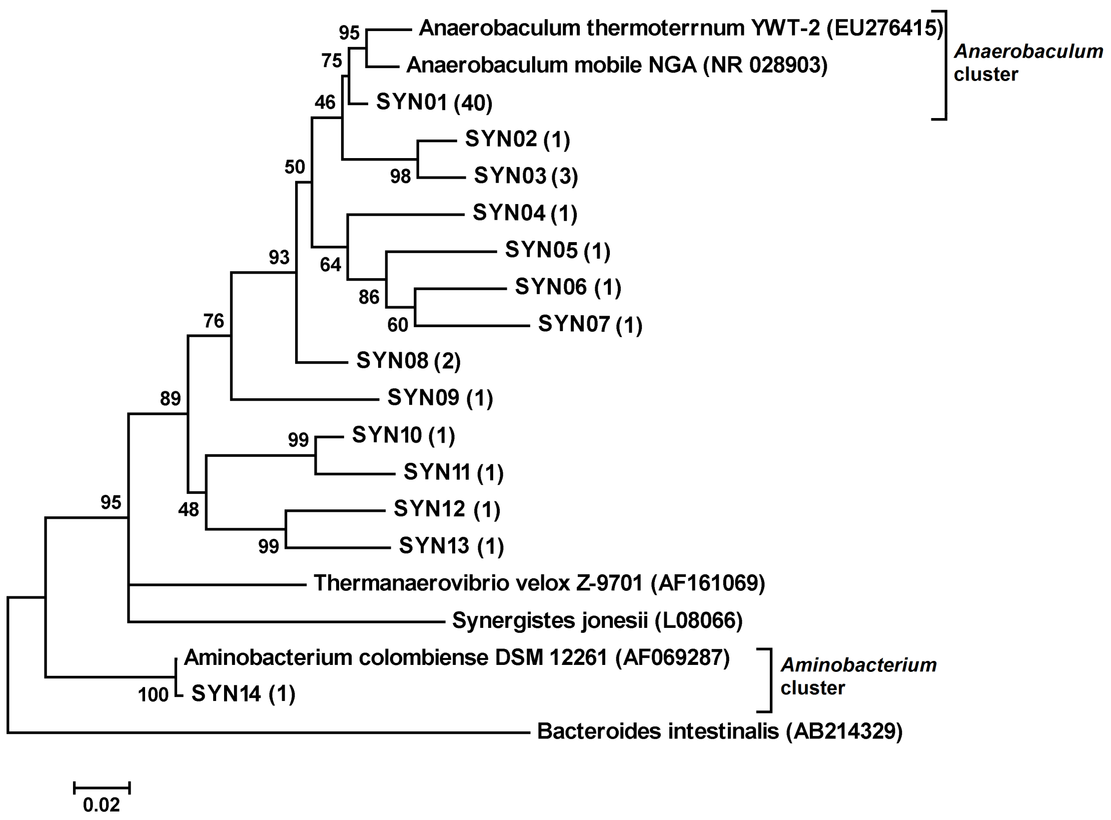

95
75
46
Anaerobaculum thermoterrnum YWT-2 (EU276415)
Anaerobaculum mobile NGA (NR 028903)
SYN01 (40)
Anaerobaculum cluster
SYN02 (1)
98
SYN03 (3)
50
SYN04 (1)
93
64
SYN05 (1)
86
SYN06 (1)
76
60
SYN07 (1)
SYN08 (2)
89
SYN09 (1)
99
SYN10 (1)
SYN11 (1)
95
48
SYN12 (1)
99
SYN13 (1)
Thermanaerovibrio velox Z-9701 (AF161069)
Synergistes jonesii (L08066)
Aminobacterium colombiense DSM 12261 (AF069287)
100
SYN14 (1)
Aminobacterium cluster
Bacteroides intestinalis (AB214329)
0.02

Figure 6.13: Phylogenetic tree for 16 rDNA amplicon sequences assigned to the phylum *Synergistetes*. The tree was constructed by means of the neighbor-joining method using genetic distances as defined by Jukes Cantor. *Bacteroides intestinalis* was used as an outgroup. Bootstrap values of 1,000 replications are represented at each branch. Each reference is annotated with its accession number in parentheses. Representative amplicons are noted with the number of sequences included in the OTUs. Two *Synergistetes* clusters were identified and could be affiliated to two defined genera, namely *Aminobacterium* and *Anaerobaculum*.

6.4 The metatranscriptome of a biogas-producing microbial community

A metatranscriptome approach was applied to elucidate the transcriptionally active organisms and biological processes within the biogas-producing community. The main aim of this analysis is to unveil the species that are important in the anaerobic digestion process. Knowledge of the key organisms would aid in improving the yield of biogas production.

In this thesis, the previously analyzed 16S rDNA amplicon sequences served as a reference to evaluate the active functions and taxa in the biogas-producing microbial community. As the amount and type of transcripts may be influenced by the conditions of the sampling time, the metatranscriptome approach was performed on the same sample that was used for the analysis of the 16S rDNA amplicon sequences (Section 6.3). The isolated total community RNA was not depleted for ribosomal RNA, since these RNA types should enable taxonomic profiling of the active community. More details concerning the RNA extraction and cDNA generation are given in [Zakrzewski et al., 2012].

6.4.1 Identification of different RNA types in the metatranscriptome data obtained from a biogas fermenter

The biogas metatranscriptome generated on the Genome Sequencer (GS) FLX platform using FLX chemistry yielded 484,920 reads with an average length of 114 bases accounting for 55,164,919 bases. Metatranscriptome sequences were screened for ambiguous nucleotides and internal poly-T or poly-A regions in order to include only valid total RNA tags in the downstream analysis pipeline in MeTra. This approach resulted in 421,387 RNA-derived reads with an average read length of 108 bases.

The downstream analysis pipeline presented in chapter 5.3 uncovered 321,544 (76.3%) large subunit ribosomal RNA-derived sequences and 67,906 (16.1%) small subunit ribosomal RNA-derived sequences (Table 6.3). The subsequent hidden-Markov-model (HMM) based search for functional, non-protein-coding RNAs identified 1,053 non-coding RNA (ncRNA) tags. CARMA3 detected 12,301 mRNA tags, whereas 8,881 were retrieved from the BLASTx analysis against the NCBI protein database and 9,090 from the HMM-based search in the Pfam database. The remaining 18,583 sequences could not be classified with this approach. The sequence lengths of the unassigned reads range from 30 to 411 bases with an average length of 79 bases.

Table 6.3: Identified RNA types in the metatranscriptome data obtained from a biogas fermenter

RNA tags	Used database	Number of reads (Percent of complete data)
large subunit rRNA	LSUdb	321,544 (76.3%)
small subunit rRNA	LSUdb	67,906 (16.1%)
further ncRNA	Rfam, custom tRNA database	1,053 (0.2%)
mRNA	nt, Pfam (CARMA3)	12,301 (3%)
unknown		18,583 (4.4%)

6.4.2 Profiling of the transcriptionally active community based on ribosomal sequence tags

Regulation of rRNA synthesis is of key importance for ribosome formation, metabolic activity and cell growth [Kemp et al., 1993, Wagner, 1994]. Accordingly, 16S rRNA gene sequences are a valuable marker for taxonomic profiling of transcriptionally active organisms. In this context, the taxonomic profile deduced from 16S ribosomal sequence tags of the metatranscriptome dataset was examined to get insights into the transcriptionally active members of the biogas-producing microbial community. Each metatranscriptome sequence previously assigned to an SSU transcript was extracted from the metatranscriptome dataset yielding 67,906 small subunit rRNA sequences. Sequences of at least 50 bases in length (66,128) were taxonomically classified using the RDP Classifier, which could assign 99.5% of the 16S rRNA gene fragments on the rank superkingdom. Looking at the relative abundances of the 16S rRNA tags (Fig. 6.14), it can be observed that 76% of the sequences are represented by *Bacteria* and 24% by *Archaea*.

Only 49% of the metatranscriptome 16S rRNA gene fragments could be assigned to taxa on phylum level. Two phyla, *Euryarchaeota* and *Firmicutes*, contribute with, respectively, 48% and 45% of the sequences classified on rank phylum the largest number of ribosomal tags. *Bacteroidetes*, *Actinobacteria* and *Synergistetes* were identified among the active phyla, though they accounted for fewer ribosomal tags (1 to 2% of the classified sequences on rank phylum).

Most of the *Firmicutes* sequences belong to the class *Clostridia* (48%) and *Bacilli* (10%) with *Clostridiales* (55%) and *Lactobacillales* (67%) being the most represented orders for these classes. On the ranks class, order and family, 38%, 33% and 26% of the reads were classifiable. *Methanomicrobia*, *Methanomicrobiales* and *Methanomicrobiaceae* dominate with 60%, 67% and 76% of all the reads that were allocated to the ranks class, order and family, respectively. In total, only 18% of the ribosomal tags were classified at taxonomic rank genus. Moreover, 15% of the 16S rRNA tags were assigned to the genus *Methanoculleus*.

6.4.3 Comparison of the taxonomic profiles obtained by DNA- and RNA-based approaches

To compare the relative fractions between the 16S rDNA amplicon and 16S rRNA metatranscriptome datasets, the rank phylum (Fig. 6.15) was considered for detailed analysis, as the number of classifiable reads on lower levels strongly decreased.
Compared to the profile based on 16S rDNA amplicon sequences, the number of archaeal reads clearly increased in the transcriptome-based profiles (Fig. 6.15). In the 16S rDNA amplicon dataset approximately 12% of the reads were assigned to *Euryarchaeota*, whereas 24% and 21% of the metatranscriptome mRNA and 16S rRNA

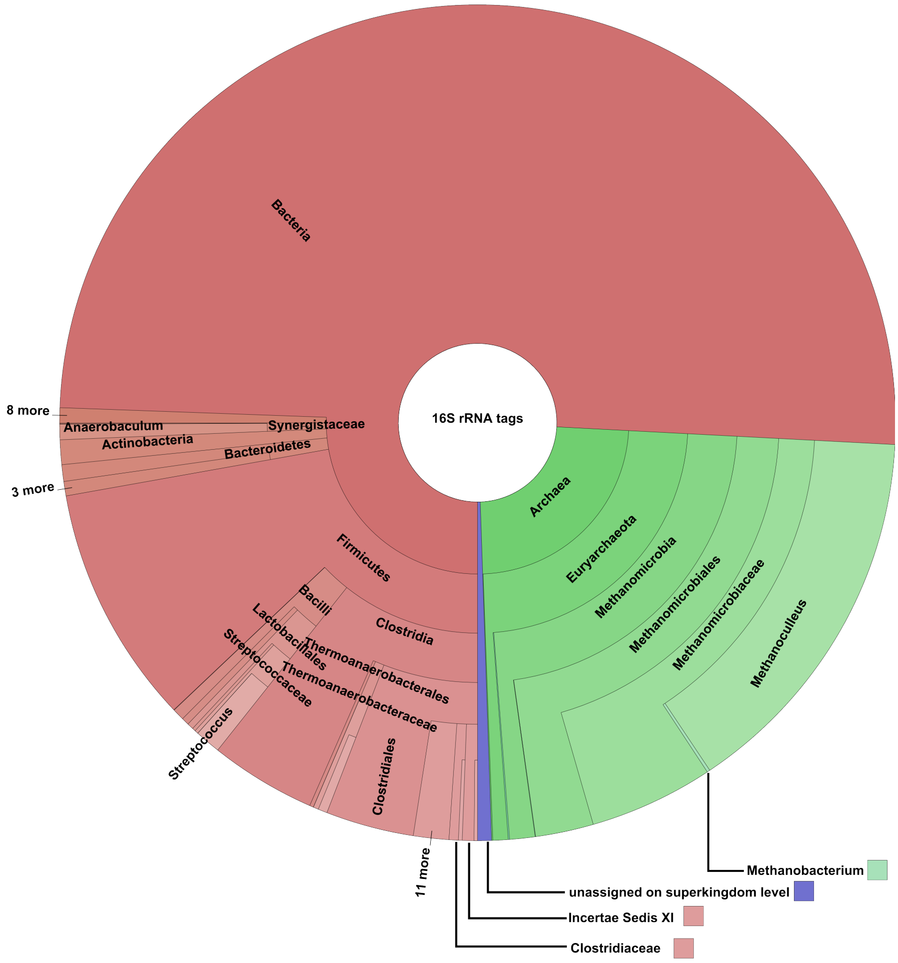


Figure 6.14: Taxonomic profile based on metatranscriptome 16S rRNA tags: Ribosomal tags discovered by similarity-based searches were classified by the RDP Classifier. The visualization was carried out using Krona [Ondov et al., 2011]. Only assignments with a confidence value of at least 0.8 were considered for the visualization.

tags were affiliated to *Euryarchaeota*, respectively. The 16S rDNA amplicon sequences were predominantly classified to belong to the phylum *Firmicutes* (73%). This phylum was present with 22% and 37% in the metatranscripome-based profiles. It is to be noted that still a large fraction of metatranscriptome sequences was not classified to this level. Approximately, 51% of the 16S rRNA tags and 23% of the mRNA tags in the metatranscriptome could not be classified at the taxonomic rank phylum. Thus, the fraction of the phyla was not completely determined due to the lack of references in the existing databases. However, the relative abundances of the metatranscriptome datasets clearly show that archaeal species have a higher transcriptional activity in the community compared to other taxa (Fig. 6.15). Finally, both metatranscriptome profiles indicate a low transcription of the phylum *Bacteroidetes*.

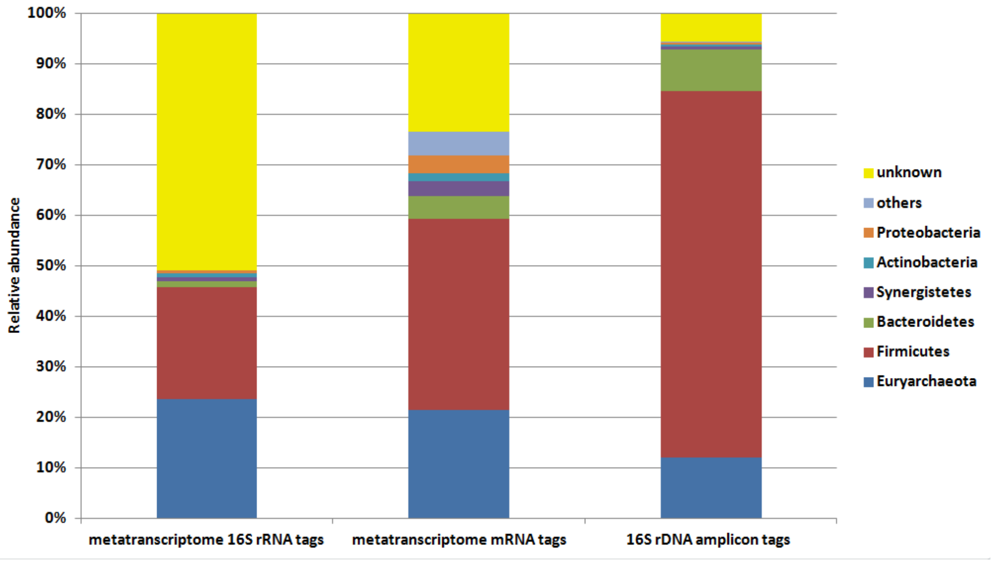


Figure 6.15: Fraction of the taxonomic assignments on the rank phylum based on metatranscriptome and 16S rDNA amplicon sequences: The taxonomic predictions based on CARMA3 were utilized for the classifications of the metatranscriptome environmental gene tags (EGTs). The metatranscriptome 16S ribosomal sequence tags and the 16S rDNA amplicons were classified by the RDP Classifier.

6.4.4 Functional characterizations of mRNA tags identified in the metatranscriptome of the biogas-producing community

Next, the functions transcribed by the biogas-producing community were investigated. In particular, transcripts for proteins that are fundamental for the anaerobic digestion were examined for their taxonomic origin. To assess the potential functions of the transcript sequences, reads neither matching the ribosomal RNA databases nor the non-coding, functional RNA database were searched for similarities to proteins. Therefore, a CARMA3 analysis was performed, which resulted in 12,301 mRNA tags. More precisely, the BLAST-based search in CARMA3 yielded 8,881 EGTs, while the Pfam-based approach uncovered 9,040 EGTs.

Thereafter, active functional processes operating in the microbial community were deduced. For this purpose, the 12,301 mRNA sequences were compared to the "evolutionary genealogy of genes: Non-supervised Orthologous Groups"′ (eggNOG) database [Muller et al., 2010] using BLASTx. The best hits were determined and used to categorize mRNA tags according to "Clusters of Orthologous Groups" (COGs) and "Non-supervised Orthologous Groups" (NOGs).

Functional characterization of mRNA-derived tags based on classification according to eggNOG

This section deals with the functional annotation of the metatranscriptome mRNA tags in order to get a comprehensive picture of the processes Overall, 4,791 mRNA tags (39% of all identified mRNA tags) were assigned to COGs and NOGs, which were annotated to functional categories (Fig. 6.16). Some categories, such as "energy production and conversion" (C) and "amino acid transport and metabolism" (E), are well covered by transcripts. Other categories, for example "extracellular structures" (W) and "secondary metabolites biosynthesis, transport and catabolism" (Q) are poorly represented or even missing in the metatranscriptome data. In the following, functional categories and COGs relevant for the biogas production are explored in detail.

During the conversion of biomass into methane, polysaccharide components of plant cell material such as cellulose, xylan and pectin are broken down into monosaccharides. Accordingly, the COG category "carbohydrate transport and metabolism" (G) and its associated COGs are important in the biogas production process. The functional category is well represented among the mRNA tags (Fig. 6.16). Assignments to cellulose M (COG1363), beta-glucosidases and related enzymes (COG1472, COG3250) and cellobiose phosphorylases (COG3459) indicate the degradation of cellulose by the microorganisms (Tab. 6.4). Xylanase/chitin deacetylase (COG0726), xylose isomerase (COG2115), beta-xylosidase (COG3507), the ABC-type xylose transport system (COG 4213) and pectin methylesterase (COG4677) represent enzymes involved in the degradation of xylan and pectin, which are both components of the plant cell wall.

Acetate, hydrogen and carbon dioxide are produced in the acetogenesis step of anaerobic degradation of biomass. In this regard, acetyl-CoA synthase, phosphotransacetylase and acetate kinase are central enzymes. The COG category "energy production and conversion" (C) includes enzymes required for the acetogenesis. The functional profile (Fig. 6.16) infers that most of the transcripts belong to the COG category C in this analysis. In total, 24 environmental gene tags (EGTs) were detected in the metatranscriptome encoding acetyl-CoA synthase (COG1614, COG1456), acetate kinase (COG0282) and phosphotransacetylase (COG0280) (Tab. 6.4). As the selected COGs also represent enzymes in the syntrophic acetate oxidation or aceticlastic methanogenesis, proteins encoded by the identified EGTs may function in these processes. A detailed analysis of the three enzymes and their potential role is supplied in the next section.

The functional contributions of phyla based on CARMA3 classifications for each COG category are illustrated in Figure 6.17. *Firmicutes* and *Euryarchaeota* appear in almost all functional categories. Overall, the distribution of the phyla along the categories is with some exceptions similar. The functional category "Chromatin structure and dynamics" (B) consists of 5 EGTs, whereas four of them were classified as *Euryarchaeota*. Due to

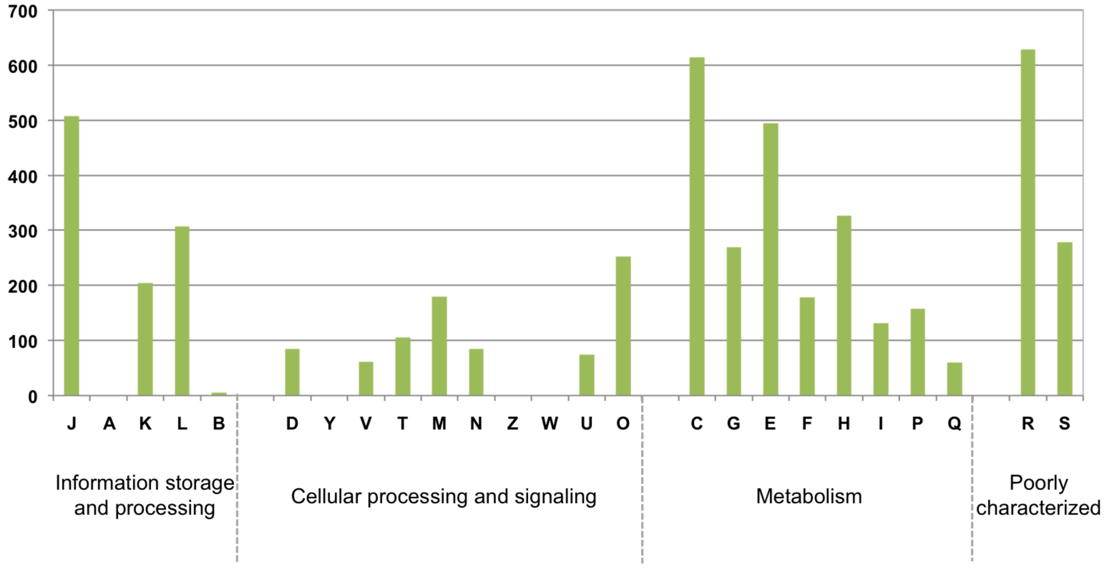


Figure 6.16: Fraction of the taxonomic assignments on the rank phylum based on metatranscriptome 16S rDNA amplicon sequences: The amounts of detected EGTs in the reads in terms of their assigned COG categories were visualized. The COG category grouping is as follows: J, translation, ribosomal structure and biogenesis; A, RNA processing and modification; K, transcription; L, replication, recombination and repair; B, chromatin structure and dynamics; D, cell cycle control, cell division, chromosome partitioning; Y, nuclear structure; V, defense mechanisms; T, signal transduction mechanisms; M, cell wall/membrane/envelope biogenesis; N, cell motility; W, extracellular structures; U, intracellular trafficking, secretion, and vesicular transport; O, posttranslational modification, protein turnover, chaperones; C, energy production and conversion; G, carbohydrate transport and metabolism; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown.

the underrepresentation of EGTs in the functional category B, the taxonomic profile may be biased.

Euryarchaeota is a major phylum in the functional category "coenzyme transport and metabolism" (H) and "inorganic ion transport and metabolism" (P), while it is less common in "cell cycle control, cell division, chromosome partitioning" (D) and "carbohydrate transport and metabolism" (G). The functional category H includes key enzymes of the methanogenesis pathway (Tab. 6.4). Indeed, the most abundant COGs contributing to the category H represent archaeal subunits of methyl coenzyme M

Table 6.4: Frequencies of metatranscriptomic mRNA tags matching to selected COG numbers

COG/NOG number	COG category	Description	mRNA tags
COG1614	C	CO dehydrogenase/acetyl-CoA synthase beta subunit	18
COG1456	C	CO dehydrogenase/acetyl-CoA synthase gamma subunit	1
COG2141	C	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases	18
COG0282	C	Acetate kinase	2
COG0280	C	Phosphotransacetylase	3
COG1363	G	Cellulase M and related proteins	9
COG1472	G	Beta-glucosidase-related glycosidases	6
COG3250	G	Beta-galactosidase/beta-glucuronidase	5
COG3459	G	Cellobiose phosphorylase	3
COG2115	G	Xylose isomerase	2
COG4213	G	ABC-type xylose transport system, periplasmic component	6
COG3507	G	Beta-xylosidase	4
COG4677	G	Pectin methylesterase	2

reductase (Mcr) (data not shown). The category P includes the ABC transport systems for CO_2 (COG0310), which is a component for hydrogenotrophic methanogenesis. No EGTs assigned to COG0310 were identified among bacterial reads.

On the other hand, *Firmicutes* dominate the functional categories D and G. The first category covers bacterial specific COGs (COG0772: Bacterial cell division membrane protein, COG02385: Sporulation protein and related proteins), which are typically absent in *Archaea*. The latter category includes COGs representing enzymes and transport systems required for the hydrolysis step. Among the most abundant COGs within this category are cellulases (COG1363) as well as ABC-type sugar and xylose transport systems (COG1175, COG1082, COG4213).

Compared to the other functional categories, the amount of transcripts from *Spirochaetes* is increased in the functional category G. Some EGTs assigned to this phylum were annotated to COGs representing sugar transport systems, galactosidases and sugar kinases. One of the EGTs was classified as *Treponema* on genus level. *Treponema primitia* was isolated from termite hindguts and was characterized to ferment homoacetogenically hexoses, pentoses and disaccharides as energy sources [Graber and Breznak, 2004]. Finally, the phylum *Bacteroidetes* is increased in the same functional category. EGTs af-

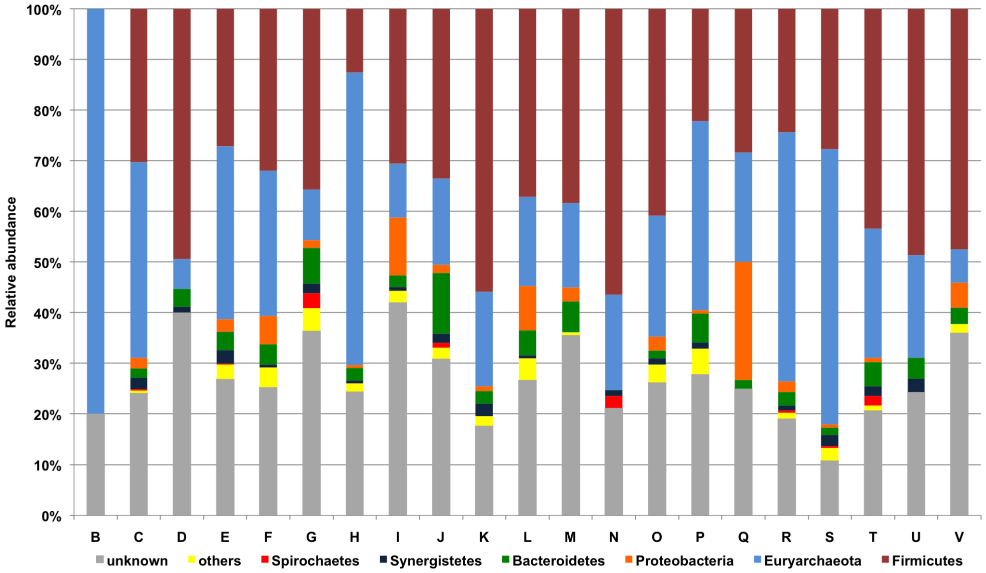


Figure 6.17: Distribution of taxa in functional COG categories: The EGTs are taxonomically characterized within each functional COG category by performing CARMA3. The assignments on rank phylum are only considered to deduce changes between the functional processes on higher ranks.

filiated to *Bacteroidetes* encode, e.g., glucosidases (COG1472) and α-L-fucosidase, which breaks down fucose, a component in plant cell walls.

Functional characterization of mRNA-derived tags based on classification according to CARMA3

The functional profile based on CARMA3 was studied for the presence of Pfam families involved in the anaerobic digestion process. In this thesis, Pfam families were in the focus that cover hydrolysis, acidogenesis, acetogenesis and methanogenesis, which are the four steps in the anaerobic digestion. Moreover, syntrophic associations relevant during the whole process were of interest. This analysis addresses the question of what organisms might be involved in the anaerobic digestion process.

Regarding the digestion of biomass material, the presence of enzymes participating in the degradation of cellulose (PF00331), pectin (PF01095), arabinose and xylan (PF04616, PF01261) is interesting. Furthermore, the cellulose binding domain (PF00553), carbohydrate binding domain (PF02837) and TIM barrel domain (PF02836) of glycosyl hydrolase family 2 (PF00703) are important for the hydrolysis step. The glycosyl hydrolase family 2 consists of enzymes that hydrolyze the glycosidic bond between two or more carbohydrates. The selected Pfams were discovered in the annotations of the mRNA tags (Tab. 6.5).

Table 6.5: EGTs that were assigned to Pfam families representing enzymes involved in the anaerobic digestion process

Enzyme/Domain	Pfam accessions	EGTs
Glycoside hydrolase family 10 (e.g. endoglucanases)	PF00331	4
Pectinesterase	PF01095	3
Glycosyl hydrolases family 43 (e.g. xylanase, beta-xylosidase), xylose isomerase-like TIM barrel	PF04616, PF01261	4
Cellulose binding domain	PF00553	1
Glycoside hydrolase family 2 (e.g. beta-galactosidase)	PF00703	2
Glycosyl hydrolases family 2, sugar binding domain	PF02837	1
Glycosyl hydrolases family 2, TIM barrel domain	PF02836	3
3-hydroxyacyl-CoA dehydrogenase	PF02737, PF00725	10
Enoyl-CoA hydratase/isomerase, Acyl-CoA dehydrogenase,	PF00378, PF02770, PF02771, PF00441	19
Alcohol dehydrogenase	PF08240	21
Methylmalonyl-CoA mutase	PF01642	2
Phosphotransacetylase	PF01515	4
Acetate kinase	PF00871	8
Acetyl-CoA synthase	PF03598, PF03599	46
Formylmethanofuran dehydrogenase	PF07969, PF00384, PF01493, PF01568, PF02663	34
Formylmethanofuran-tetrahydromethanopterin N-formyltransferase	PF02741	13
N5N10-methenyl-tetrahydromethanopterin cyclohydrolase	PF02289	6
Coenzyme F420-dependent N5,N10-methylene-tetrahydromethanopterin dehydrogenase	PF01993	23
Coenzyme F420-dependent N5,N10-methylene-tetrahydromethanopterin reductase	PF00296	19
N5-Methyl-tetrahydromethanopterin: methyltransferase	PF04208, PF05440, PF04211, PF04207, PF04206, PF09472, PF04210, PF02007	35
Methyl coenzyme M reductase	PF02249, PF02745, PF02783, PF02241, PF04609, PF02505, PF02240	105

The acidogenesis process, the second step of biogas production, was represented by Pfam families involved in fatty acid metabolism (PF02737, PF00725), butyrate synthesis (PF00378, PF02770, PF02771, PF00441), alcohol synthesis (PF08240) and propionate synthesis (PF01642) (Tab. 6.5). The enzymes phosphotransacetylase (PF01515), acetate kinase (PF00871) and acetyl-CoA synthase (PF03598, PF03599) were identified in the Pfam profile and were associated with the acetogenesis step (Tab. 6.5).

The acetate produced in the acetogenesis can be subsequently used as a substrate for the aceticlastic methanogenesis. *Archaea* capable of aceticlastic methanogenesis use the reverse Wood-Ljungdahl pathway to convert acetate into methane and carbon dioxide [Pierce et al., 2008, Ragsdale, 2008, Ragsdale and Pierce, 2008]. Enzymes involved in this pathway are also acetyl-CoA synthase, phosphotransacetylase and acetate kinase. In an alternative process, syntrophic acetate-oxidizing bacteria convert acetate to hydrogen. For known acetate-oxidizing bacteria it was shown that specific enzymes involved in the CO dehydrogenase/acetyl-CoA pathway operate in both, acetate oxidation and acetate formation [Lee and Zinder, 1988, Schnürer et al., 1997, Hattori et al., 2005]. Depending on the hydrogen concentration in the medium, acetate is either produced or oxidized by syntrophic acetate-oxidizing bacteria [Schnürer et al., 1997].
To distinguish whether the identified Wood-Ljungdahl pathway EGTs are active in acetogenesis, syntrophic acetate oxidation or in aceticlastic methanogenesis, the taxonomic profile obtained by CARMA3 was studied in detail. For this purpose, the MetaCyc pathway "reductive acetyl coenzyme A" was utilized to annotate corresponding enzymes according to Pfam families. All expected Pfams were detected in the functional profile calculated by CARMA3 (Fig. 6.18a). In total, 166 EGTs were identified. The taxonomic profile of those EGTs indicates a dominance of bacterial transcripts (59%). Further 3% of the identified EGTs, were assigned to *Archaea*. The archaeal sequences belong to the order *Methanomicrobiales*. However, for 37% of all EGTs representing the MetaCyc "reductive acetyl coenzyme A" pathway, no references were available that could enable taxonomic classification. Only 10 EGTs could be assigned to a family rank belonging mainly to *Thermoanaerobacteraceae* and *Peptococcaceae*. Species of the family *Thermoanaerobacteraceae* are known to produce acetate under extreme conditions [Bao et al., 2002, Onyenwoke et al., 2007, Feng et al., 2009], whereas *Peptococcaceae* species are capable to ferment proteins or carbohydrates to mainly lower fatty acids [Rogosa, 1971]. In conclusion, the detected EGTs in this analysis encode enzymes that participate either in acetogenesis or syntrophic acetate oxidation rather than in aceticlastic methanogenesis.

Finally, Pfam families involved in methanogenesis such as formylmethanofuran dehydrogenase (PF07969, PF00384, PF01493, PF01568, PF02663) and methyl coenzyme M reductase (PF02249, PF02745, PF02783, PF02241, PF04609, PF02505, PF02240) were discovered (Tab. 6.5). As methyl coenzyme M reductase (Mcr) plays a central role in both, hydrogenotrophic and aceticlastic methanogenesis, it was used as a marker to deduce whether the hydrogenotrophic or aceticlastic methanogenesis is preferred in the analyzed biogas plant. Therefore, the taxonomic classification of EGTs assigned

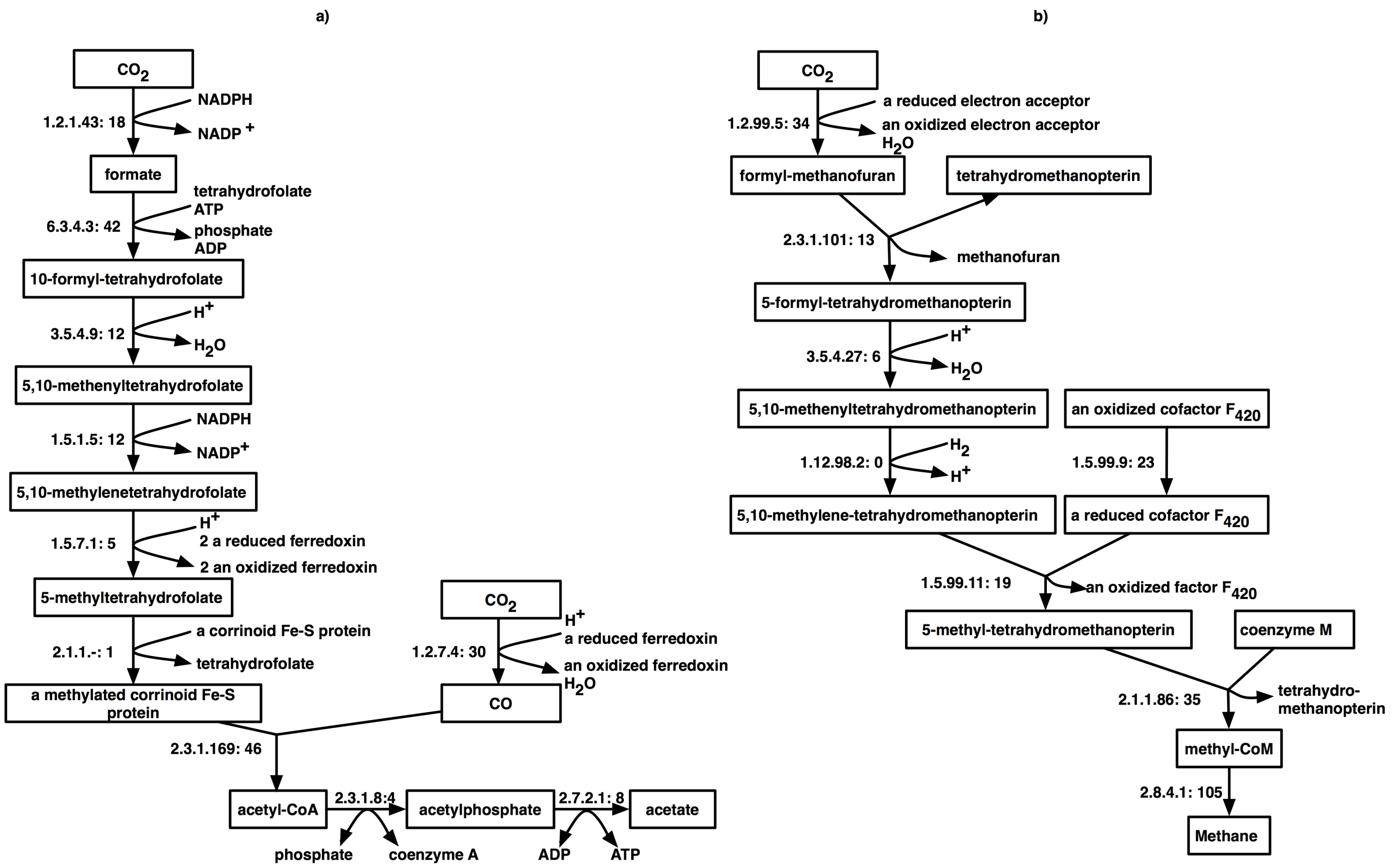


Figure 6.18: Reconstruction of the (a) "reductive acetyl coenzyme A" and (b) "Methanogenesis from CO$_2$" pathway as described in MetaCyc. The sequences of the involved enzymes are Pfam categorized. Thereafter, EGTs representing corresponding Pfam accession numbers were searched in the mRNA tags based on CARMA3 results. The counts of identified EGTs are denoted after the EC number of the particular enzyme.

to Pfam families representing Mcr subunits was determined. In total, 105 EGTs representing Mcr subunits were identified. Of these, 75 EGTs were classified as *Archaea* (71%). The remaining reads were of unknown origin. However, according to CARMA3 assignments, 75% and 3% of the archaeal EGTs belong to the classes *Methanomicrobia* and *Methanobacteria*, respectively. A similar taxonomic composition was observed in the metagenome data (Fig. 6.4). A percentage of 67% of the archaeal EGTs could not be classified to a genus. *Methanoculleus* is the only characterized genus in the profile and constitutes 33% of all archaeal *mcr* transcripts. Though some reads were unclassified, obtained results infer that methane is dominantly produced in the hydrogenotrophic methanogenesis pathway.
Sequences of methanogens that are known to conduct aceticlastic methanogenesis were rarely identified in the metatranscriptome mRNA tags. Only five EGTs assigned to the genus *Methanosaeta* were detected in the functional profile based on all identified EGTs in the metatranscriptome, whereas *Methanosarcina* is present with one EGT.
To analyze the coverage of the methanogenesis pathway by mRNA tags, the MetaCyc pathway "Methanogenesis from CO_2" was examined for the presence of the involved enzymes by investigating the CARMA3 results (Fig. 6.18b). In total, 235 EGTs represent the methanogenesis pathway. All enzymes except for one (EC number 1.12.98.2) was identified in the metatranscriptome. The taxonomic profiling of EGTs for methanogenesis revealed that most of them were assigned to *Methanomicrobiales* (62%) followed by *Methanobacteriales* (0.9%) on the rank order. On family rank, *Methanomicrobiaceae* (45%), *Methanobacteriaceae* (0.9%) and *Methanospirillaceae* (0.4%) were predicted.

Hydrogenotrophic methanogenesis frequently is accomplished in a syntrophic association with acetate-oxidizing bacteria. In this association, acetate oxidizers produce hydrogen that is scavenged by hydrogenotrophic methanogens for biogas production. Syntrophic bacteria known to oxidize acetate to hydrogen and carbon dioxide in association with hydrogenotrophic methanogens are *Thermacetogenium phaeum* [Hattori et al., 2005], *Thermotoga lettingae* [Balk et al., 2002], *Clostridium ultunense* [Schnürer et al., 1996], the acetate-oxidizing rod-shaped bacterium AOR [Lee and Zinder, 1988] and *Tepidanaerobacter acetatoxydans* [Westerholm et al., 2011].
Thermacetogenium phaeum belongs to the family *Thermoanaerobacteraceae* and oxidizes acetate in association with *Methanothermobacter thermautotrophicus*. The genus *Thermacetogenium* was not found in the taxonomic profile created by CARMA3, but 64 EGTs were assigned to the family *Thermoanaerobacteraceae*. Of these, one EGT encodes an acetyl-CoA synthase (PF03598). Moreover, one EGT is similar to the gene encoding methyl-coenzyme M reductase (PF04609) in *Methanothermobacter* suggesting a syntrophic association between a hydrogenotrophic methanogen and a related but so far unknown *Thermoanaerobacteraceae* species. Unknown species similar to the family *Thermoanaerobacteraceae* were recently reported to be responsible for syntrophic oxidation of acetate with hydrogenotrophic *Methanocellales* species in thermophilic soils [Rui et al., 2011].
The acetate-oxidizing *Thermotoga lettingae* strain TMO degrades acetate in the presence of the methanogen *M. thermautotrophicus* [Balk et al., 2002]. Four sequences belonging

to the genus *Thermotoga* were identified among the metatranscriptome mRNA tags. EGTs assigned to the species *Clostridium ultunense* were absent. As the genus *Clostridium* was present in the taxonomic profile of the biogas plant, syntrophic conversion of acetate into methane by acetate-oxidizing *Clostridium* species and hydrogenotrophic *Archaea* may occur. Finally, three EGTs were assigned to *Tepidanaerobacter*. Species belonging to this genus are capable of syntrophic acetate-, alcohol- or lactate-degradation [Sekiguchi et al., 2006, Westerholm et al., 2011].

Species of *Syntrophomonas* are capable of degrading fatty acids by β-oxidation in coculture with methanogens such as *Methanospirillum hungatei* [McInerney et al., 1981, Zhang et al., 2004]. From the five EGTs that were classified as *Syntrophomonas*, one EGT was assigned to enoyl-CoA hydratase/isomerase family (PF00378) and another one to alcohol dehydrogenase (PF08240). Each of these Pfams represents a key enzyme in the β-oxidation process. Moreover, one EGT encoding a methanogenesis enzyme (PF01993) was taxonomically assigned to *Methanospirillum*. Further genera involved in syntrophic oxidation of fatty acids in association with methanogens are *Pelotomaculum*, *Smithella*, *Syntrophus* and *Syntrophobacter* [McInerney et al., 2009]. The genus *Pelotomaculum* is represented by 8 EGTs with 2 EGTs belonging to the species *Pelotomaculum thermopropionicum*, which is a syntrophic propionate-oxidizing bacterium growing in coculture with *M. thermautotrophicus*. The corresponding genus of this archaeal species was identified with one EGT. Finally, no EGTs classified to the genera *Smithella*, *Syntrophus* and *Syntrophobacter* were detected by CARMA3. However, the latter genus belongs to the order *Syntrophobacterales*, to which two mRNA tags were assigned. In summary, these results imply that syntrophic acetate/propionate/fatty acids oxidizing bacteria are likely to interact with H_2-scavenging methanogens in the biogas plant.

Overall, the functional profile in Table 6.5 infers that the source for energy production is obtained from fermentation of polysaccharides with subsequent production of short-chain fatty acids. Finally, methane is likely produced in hydrogenotrophic methanogenesis in association with syntrophic bacteria.

Functional characterization of mRNA-derived tags assigned to specific phyla

Since most of the transcripts were classified as originating from the taxa *Archaea* and *Firmicutes* (Section 6.4.2), these sequences were analyzed for their functional assignments in terms of Pfam families based on CARMA3 predictions. The 2,072 reads assigned to the phylum *Firmicutes* cover 854 different Pfam families. The CO dehydrogenase/acetyl-CoA synthase complex (PF03598) is present among the most abundant Pfam families and is supported by 16 EGTs. This Pfam family represents a key enzyme of the Wood-Ljungdahl pathway. CARMA3 classified ten of these EGTs to the order *Clostridiales* and one to *Thermoanaerobacterales*. The remaining reads are unknown. Furthermore, functions were identified that are related to hydrolytic reactions in the first step of the anaerobic digestion. For example, pectinesterase (PF01098) and xylose isomerase (PF01261) are in the functional profile based on Pfam families. These outcomes confirmed that *Firmicutes* play a central role in acetogenesis, syntrophic acetate oxidation and hydrolysis.

CARMA3 yielded 1,158 archaeal sequences covering 463 different Pfam families. Looking at the most abundant Pfam families within the archaeal EGTs, a high representation of transcripts encoding methanogenesis-relevant enzymes can be noticed (Tab. 6.6). Corresponding Pfam families are well represented among the archaeal transcripts and constitute around 14% of the total Pfam families assigned to *Archaea*. This result indicates that the archaeal transcriptome is predominantly composed of methanogenesis transcripts.

Table 6.6: The most abundant Pfam assignments of EGTs classified to *Archaea*[1]

Number of EGTs	Pfam accession	Description
43	PF00107	Zinc-binding dehydrogenase
28	PF00037	4Fe-4 binding domain-terminal domain
21	PF02241	Methyl-coenzyme M reductase beta subunit, C-terminal domain
17	PF00296	Luciferase-like monooxygenase
16	PF01993	methylene-5,6,7,8-tetrahydromethanopterin dehydrogenase
14	PF02915	Rubrerythrin
14	PF02745	Methyl-coenzyme M reductase alpha subunit, N-terminal domain
13	PF01243	Pyridoxamine 5'-phosphate oxidase
13	PF01493	GXGXG motif
12	PF02249	Methyl-coenzyme M reductase alpha subunit, C-terminal domain
12	PF10050	Predicted metal-binding protein (UF2284)
11	PF02505	Methyl-coenzyme M reductase operon protein D
11	PF01176	Translation initiation factor 1A / IF-1
11	PF02741	FTR, proximal lobe
10	PF03130	PBS lyase HEAT-like repeat

[1] Gray-colored rows represent Pfams associated with the methanogenesis pathway

6.4.5 Non-coding RNAs identified in the metatranscriptome of microorganisms residing in a biogas fermenter

Non-coding RNAs (ncRNAs) are transcripts that are not translated into proteins but have key roles in regulating important biological processes [Storz and Haas, 2007].

Typically, the length of ncRNAs ranges from 50 to 650 bases making their discovery in genomes challenging. Recently, ncRNAs were detected in metatranscriptomes of the human gut [Gosalbes et al., 2011] and ocean water [Shi et al., 2009] illustrating the importance of detailed studies of ncRNAs in metatranscriptomics.

Approximately 1,000 ncRNA tags were discovered with transfer-messenger RNA (tmRNA, RF00023) [Ray and Apirion, 1979, Keiler, 2008], Ribonuclease P (RNase P, RF00373, RF00010, RF00011) [Guerrier-Takada et al., 1983] and 'ornate large extremophilic RNA' (OLE RNA, RF01071) [Puerta-Fernandez et al., 2006] assignments as the most abundant characterized ncRNAs. In a metatranscriptome study of an ocean sample [Shi et al., 2009], transcripts for tmRNA and RNase P RNA were also among the most abundant Rfam families, while no OLE RNAs were identified. Further ncRNAs detected in the metatranscriptome analyzed within this work are signal recognition particle RNAs [Bernstein and Hyndman, 2001, Zwieb and Eichler, 2002], which enable the secretion of proteins with respect of the presence of signal peptides, riboswitches [Tucker and Breaker, 2005] and 6S RNAs [Wassarman and Storz, 2000], which inhibit the activity of the RNA polymerase. Non-coding RNAs perform important gene control and protein sensing tasks, which are critical for the organisms. The existence of small, non-coding RNAs implies that the biogas community is a suitable source for studying fundamental regulation of cellular processes. The next sections briefly describe the three most abundant Rfam families.

The transfer-messenger RNA

The molecule tmRNA is involved in the quality control pathway and is one of the most abundant RNAs in bacteria [Keiler, 2008]. It is a component of the ribonucleoprotein complex, which is responsible for resetting ribosomes in case an erroneously transcribed mRNA stalls the translation at a ribosome. In total, 202 sequences were assigned to tmRNAs using the metatranscriptomic MeTra pipeline introduced in chapter 5.3.
For taxonomic profiling, a lowest common ancestor (LCA) approach based on BLAST hits against the NCBI nucleotide (nt) database was carried out. Only 63 sequences had no BLAST hits against NCBI nt database (E-value cutoff of 10^{-5}). An LCA approach was performed for the remaining 139 sequences to elucidate the taxonomic assignments for the putative tmRNA fragments. No archaeal tmRNAs were discovered proving former suggestions that tmRNAs are absent in archaeal organisms [Keiler, 2008]. On the rank phylum, 166 sequences were classified to *Firmicutes*, 3 to *Thermotogae*, 2 to *Spirochaetes* and *Fusobacteria* and 1 to *Synergistetes*. These phyla also occurred in the taxonomic profile based on metatranscriptome mRNA tags. Moreover, 69 reads could be assigned to the level genus with *Desulfotomaculum* (21 reads) and *Clostridium* (10 reads) as the most abundant genera. Sequences affiliated to *Desulfotomaculum* exist in the metatranscriptome mRNA tags. Species of the genus *Desulfotomaculum* were recently identified in biogas plants and belong to the sulfate-reducing bacteria [Deublein and Steinhauser, 2008]. As sulfate-reducing bacteria require hydrogen and acetate, they compete with methane-producing bacteria.

The ribonuclease P RNA

The ribonuclease (RNase) P complex cleaves RNA during the tRNA maturation process in species across all kingdoms [Guerrier-Takada et al., 1983]. 293 sequences were assigned to the Rfam accessions RF00373, RF00010, RF00011, which represent an RNA chain of bacterial or archaeal RNase P. Of these, 261 reads had a BLAST hit to reference sequences in the nt database. The LCA approach revealed 34 archaeal and 225 bacterial sequences. In bacteria, the RNase P complex consists of the RNA subunit being 350-400 bases in length and a protein subunit [Sidote et al., 2004]. The protein subunit is represented by PF00825, which occurs in the metatranscriptome mRNA tags with one EGT of unknown origin according to CARMA3 classifications. The archaeal RNase P complex contains an RNA component with a length of 227 to 400 bases and at least four protein subunits. Also one archaeal RNase P protein subunit (PF04032) was discovered in the metatranscriptome mRNA tags.

In the Rfam database, RNase P RNA is annotated from the position 1,271,599 to 1,271,940 on the genome in *Methanoculleus marisnigri* JR1. In total, 34 reads were assigned to the species by the LCA approach. All reads overlapped with the RNase P gene in the genome of *M. marisnigri* JR1. One metatranscriptome read of the length 83 bases matched the genome (1,271,856 - 1,271,938) with an identity of 98.8%. Because of the high similarity of the reads to the genome of *M. marisnigri*, the archaeal sequences were aligned to the corresponding RNase P alignment obtained from the Rfam database using the HMMER3 package. The 34 reads together covered the whole RNase P RNA component of *M. marisnigri* (data not shown).

The ornate large extremophilic RNA

The non-coding RNA family OLE [Puerta-Fernandez et al., 2006] is a conserved molecule occurring predominantly in extremophilic, anaerobic species of the phylum *Firmicutes*. This ncRNA family was firstly identified in a comparative sequence analysis in 2005 [Corbino et al., 2005] and named ornate large extremophilic (OLE) RNA due to its particularly ornate secondary structure, large size of approximately 650 bases and occurrence in almost exclusively extremophilic species. In *Bacillus halodurans* C-125, this RNA family was observed to be highly expressed [Block et al., 2011]. Still, the function of OLE RNAs is unknown, but they are suggested to form a ribonucleoprotein complex with proteins in microorganisms residing in extreme environments [Puerta-Fernandez et al., 2006]. This ribonucleoprotein complex is likely localized near the cell membranes, where OLE RNA might be catalytic, structural or required for global regulation of gene expression [Block et al., 2011].

Species in the biogas plant are anaerobic, a characteristic of all known OLE RNA-carrying organisms. As OLE RNA genes are found primarily in *Firmicutes*, a phylogenetic tree for the identified reads was generated. For this purpose, aligned reference sequences representing OLE RNAs were firstly obtained from the Rfam database (RF01071). Duplicated or highly similar sequences were manually discarded resulting in 31 aligned reference sequences that were utilized as a basis for a phylogenetic tree reconstruction using the neighbor-joining method [Saitou and Nei, 1987] with genetic distances corrected by Jukes and Cantor [Jukes and Cantor, 1969] and a bootstrap value of 1,000 [Tamura et al., 2007]. Finally, RAxML [Berger and Stamatakis, 2011] was applied, which placed the reads assigned to OLE RNA into the neighbor-joining tree of reference sequences. Similar metatranscriptome reads were joined together in order to reduce the tree complexity. The phylogenetic tree indicates that OLE RNAs are widely distributed among species of anaerobic *Firmicutes* encountered within the biogas plant (Fig. 6.19).

Most of the OLE RNA tags clustered close to species of the order *Clostridiales*. Genera belonging to this order are *Alkaliphilus* and *Clostridium*. Representative sequences of these genera are close to metatranscriptome reads in the phylogenetic tree. The genera are among the 10 largest OTUs formed by 16S rDNA amplicons (Section 6.3.2). Moreover, sequences are in close proximity to sulfate-reducing bacteria [Imachi et al., 2002] of the family *Peptococcaceae*, which belongs to the order *Clostridiales* (Fig. 6.19, OLE08, OLE23, OLE24, OLE25, OLE26, OLE27). Further 13 RNA tags are similar to a sequence of *Syntrophomonas wolfei*, which beta-oxidizes propionate to acetate and occurs in syntrophic associations with H_2 oxidizing organisms [McInerney et al., 1981].

In particular, many sequences (Fig. 6.19, OLE04, OLE22, OLE31) are similar to species related to the order *Thermoanaerobacterales*. The reference *Thermoanaerobacter* sp. X514 was not identified in the metatranscriptome-based profiles, but it was present in the taxonomic profile of the metagenome of the same biogas plant. Several species of the genus *Thermoanaerobacter* are of industrial interest due to their capability to ferment sugars to ethanol or acetate under thermophilic conditions [Bao et al., 2002, Onyenwoke et al., 2007, Feng et al., 2009].

Six metatranscriptome reads clustered within a *Bacillus* clade (Fig. 6.19). Species of *Bacillus* were also observed in taxonomic profiles based on metatranscriptome mRNAs and 16S rRNAs. Additionally, a read representing 7 OLE RNA tags (OLE20) is related to a reference sequence of the order *Natranaerobiales*. This order is neither present in the metagenome- nor in the metatranscriptome-based profiles. However, a clone similar to an unknown *Natranaerobius* species is described in a lower-surface of a cattle manure compost [Maeda et al., 2010]. Hence, a species related to *Natranaerobiales* might be present in the biogas plant. Finally, some OLE RNA tags are similar to metagenome fragments obtained from the human or mouse gut. Consequently, OLE RNAs encoded by related species may play an important role in other habitats. Unfortunately, so far no defined species are described for these OLE RNAs sequences.

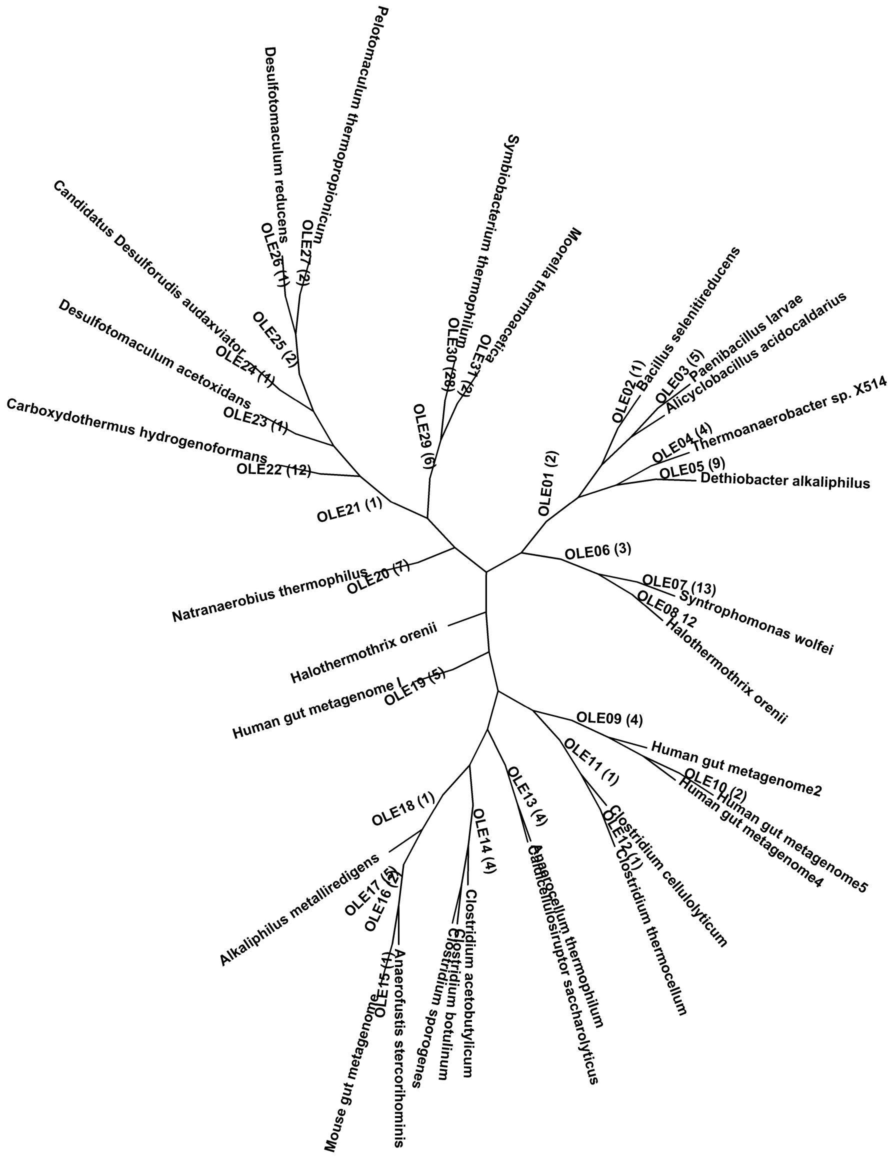


Figure 6.19: Unrooted tree for OLE RNA tags: Aligned reference sequences representing OLE RNAs were obtained from the Rfam database. A phylogenetic tree was built based on selected reference sequences. The references are annotated either according to their originating species name or source. Metatranscriptome reads assigned to OLE RNA genes were placed in the tree using the tool RAxML. Similar metatranscriptome reads were grouped together and a representative sequence was selected for each group. The representative sequences are denoted with the number of reads that were joined together.

6.5 Characterization and identification of laccases

6.5.1 Introduction to laccases

Laccases belong to the copper metalloenzymes and function as oxidoreductases in all domains of life [Hoegger et al., 2006]. The substrate range for laccases is broadened and includes dye pollutants [Singh Arora and Kumar Sharma, 2010], nonphenolic compounds [Bourbonnais and Paice, 1990] and polycyclic aromatic hydrocarbons [Pickard et al., 1999]. Because of their broad substrate specificity and wide reaction capabilities, laccases possess a considerable industrial potential. Promising applications of laccases are for example textile dye bleaching [Claus et al., 2002], pulp processing [Murugesan, 2003] and bioremediation of soils as well as of water [Fang et al., 2011, Palanisami et al., 2010].

Laccases are well known to degrade lignin, which comprises 10-30% of plant lignocellulose [Sarkanen and H., 1972]. The process of lignin degradation is well studied in fungi [Wesenberg et al., 2003]. So far, fungal laccases were used for industrial applications [Rodríguez Couto and Toca Herrera, 2006]. Unfortunately, most fungal laccases lose their activities under alkaline conditions [Murugesan, 2003] and are sensitive to chloride [Jimenez-Juarez et al., 2005]. There are several reports about laccases that have been found widely distributed among *Bacteria* [Alexandre and Zhulin, 2000]. Bacterial organisms are in particular important sources for the identification of laccases, as they are well adapted to environments matching industrial conditions.

Two groups of laccases exist, which differ in their structure [Komori et al., 2009]. Three-domain laccases are intensively studied, as they occur in fungi and most of the known bacterial species. They consist of two conserved domains (domain 1 and 3), which are dispersed by an additional domain (domain 2) (Fig. 6.20). The conserved domains contain each two copper-binding regions (cbr), which are typical for laccase molecules. A characteristic of each copper-binding region is the HXH motif, where H is an abbreviation for the amino acid residue histidine and X represents any amino acid residue. Contrary to the three-domain laccases, the two-domain laccases lack the second domain. Three subtypes of two-domain laccases are characterized with respect to the location of the motif HCH, where C is the amino acid residue cysteine [Nakamura et al., 2003]. Type-A possesses the motif HCH in the first cbr of domain 1 (cbr1) and the second cbr (cbr4) of domain 2. Type-B has only one HCH motif in cbr1, whereas the HCH motif in type-C is present in the second cbr of domain 1 (cbr2). So far, two-domain laccases have only been discovered in bacterial species [Komori et al., 2009].

Because of their relevance in biotechnological applications, laccases were exhaustively investigated in this thesis. For this purpose, a method based on profile hidden Markov models (HMMs) was applied that captures putative laccases in genome and metagenome data. The aim of the analysis was to get a deeper knowledge into the diversity and functions of laccases in bacteria.

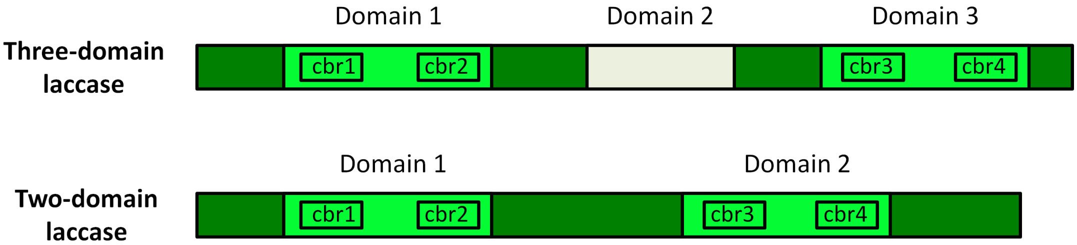


Figure 6.20: Structure of two-domain and three-domain laccases: In three-domain laccases, domain 2 is embedded between domain 1 and domain 3, which contain together four copper-binding regions (cbr). In two-domain laccases, the middle domain is missing.

6.5.2 Database construction

A database containing 3,602,197 proteins from annotated draft genomes in the NCBI database was obtained (September, 2010) to search for putative laccases. Altogether, the proteins originated from 995 microbial genomes. Excluding one viral and six archaeal, all sequences are of bacterial origin. Moreover, a database of proteins from complete NCBI genomes was included into the analysis, which contains 3,819,638 proteins from 1,216 genomes (September, 2010). A joint database of the proteins from the draft and complete genomes was constructed, which was used to examine the diversity of laccases in bacterial genomes. Overall, the joint database stores proteins from 2,211 microbial genomes.

6.5.3 Building of profile hidden Markov models representing laccase proteins

A two-step approach was applied to build a profile HMM for the identification of laccase-like proteins encoded in microbial genomes or metagenomes. In the first step, an initial profile HMM was generated. For this purpose, a BLAST search was performed against the NCBI non-redundant protein database (nr). Four characterized queries were chosen for the search for two-domain laccases, namely the type-B laccases

- SLAC from *Streptomyces coelicolor* [Machczynski et al., 2004],
- EpoA from *Streptomyces griseus* [Endo et al., 2003]

as well as the type-C laccases from

- *Nitrosomonas europaea* [Lawton et al., 2009],
- a metagenome [Komori et al., 2009].

For each query, the first 100 hits were selected. In total, 400 sequences were obtained by the BLAST searches.

The same procedure was repeated with queries that represent three-domain laccases. Known bacterial three-domain laccases are encoded by

- *Bacillus licheniformis* ATCC 14580 [Koschorreck et al., 2008],
- *Bacillus subtilis* subsp. subtilis str. 168 [Martins et al., 2002],
- *Escherichia coli* str. K-12 substr. MG1655 [Roberts et al., 2003],
- *Bacillus halodurans* C-125 [Ruijssenaars and Hartmans, 2004],
- *Thermus thermophilus* HB27 [Miyazaki, 2005],
- *Streptomyces cyaneus* strain CECT 3335 [Arias et al., 2003],
- *Streptomyces lavendulae* REN-7[Suzuki et al., 2003],
- *Marinomonas mediterranea* [Sanchez-Amat et al., 2001].

Similarly, the best 100 hits for each query were selected. Overall, 800 sequences were collected.

Since the results of the BLAST searches for three-domain laccases contained sequences representing two-domain laccases, criteria were defined to distinguish both groups. Therefore, sequences obtained by all BLAST searches were joined. Moreover, duplicates and sequences without the four conserved copper-binding regions were removed. An analysis of the lengths of the whole proteins as well as of the region between the copper-binding region 1 (cbr1) and copper-binding region 4 (cbr4) was carried out. The length of the whole protein sequence of two-domain and three-domain laccases did not separate both groups (Fig. 6.21a), but the distance between cbr1 and cbr4 is surprisingly constant within each group (Fig. 6.21b). The fragment between cbr1 and cbr4, herein noted as cbr14, has in two-domain and three-domain laccases an average length of approximately 200 and 390 amino acids, respectively. The Figure 6.21 indicates that the two-domain and three-domain laccases can be distinguished on the basis of the length of the cbr14 fragment rather than the complete length.

Based on this observation, 160 sequences representing putative two-domain laccases remained, which were separated into three different types as proposed by Nakamura [Nakamura et al., 2003]. Type-A has a conserved copper-binding domain including the HCH motif in cbr1 and cbr4, whereas type-B and type-C have a HCH structure in cbr1 and cbr2, respectively. Based on these characteristics, the sequences were grouped into the three different types. In order to reduce a potential bias of the models, similar sequences derived from related species were manually deleted. Each group was aligned using MUSCLE [Edgar, 2004a, Edgar, 2004b] and the alignments were modified by deleting all columns prior to cbr1 and after cbr4. Based on the alignments, profile HMMs were built for the type-B and type-C laccases. Since type-A laccases represent

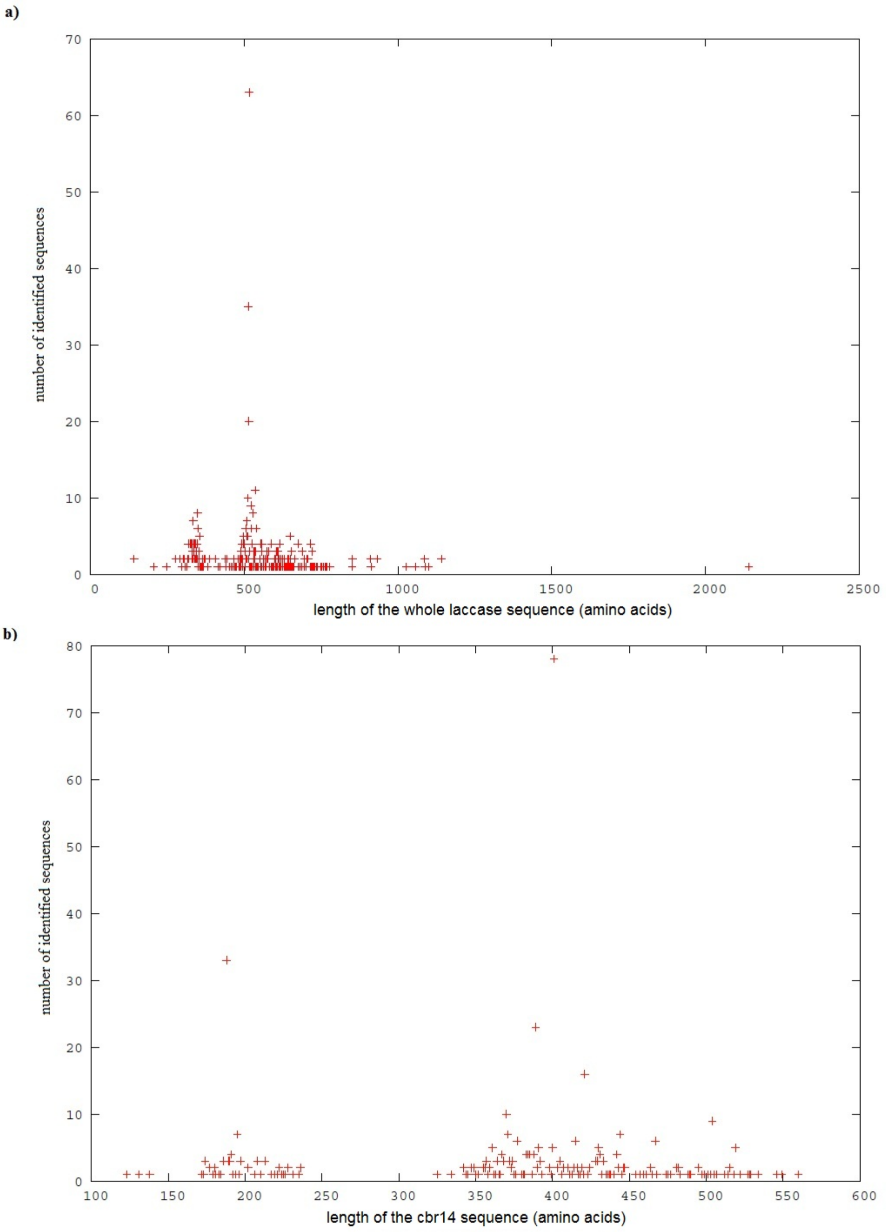


Figure 6.21: Length distributions of (a) the whole sequences of the laccase proteins and (b) the fragment between copper-binding region 1 and copper-binding region 4: It is not possible to separate two-domain and three-domain laccases based on the whole protein sequence, but there is a clear difference in the fragment size covering the four copper-binding regions (cbr) in both groups. Based on this observation, two-domain and three-domain sequences can be distinguished.

archaeal sequences and the focus is on studying bacterial sequences, the sequences of type-A laccases were excluded from further analysis. The initial profile HMM representing type-B laccases is based on 97 sequences, while 19 sequences were used to build the type-C initial profile HMM.

In case of three-domain laccases, 580 sequences were extracted. More difficulties arose when addressing the categorization of three-domain laccases since many highly diverse sequences were retrieved. Nevertheless, the length plot of the protein fragments covering cbr1 to cbr4 (Fig. 6.21b) indicates a dominance of sequences with a length below 450 amino acids, which are partly similar in sequence (data not shown). To avoid bias in the model introduced by the short sequences, three models were generated. 324 sequences have a cbr14 length below 450 amino acids and are represented by the group "small3d". The large sequences that reached a length of at least 450 were assigned to the group "big3d". In addition, a third laccase model, termed "cotA", was established based on 41 sequences that differ in sequence from the previously described laccase sets.

The four initial profile HMMs were applied to specifically search for more laccase sequences. For this purpose, the NCBI protein database based on annotated complete genomes was utilized for the search. After filtering duplicates, separating the sequences based on the length of cbr14 and locating the triplet HCH, 49 and 96 sequences were extracted for the type-C and type-B two-domain laccases, respectively. For building the final model for small3d and big3d, 324 and 215 laccase fragments were utilized, respectively. The model cotA still represents 41 laccase sequences. In summary, five profile HMMs were constructed based on laccase proteins available from the NCBI database.

6.5.4 Distribution and functions of bacterial laccase-like proteins in NCBI database entries

The five profile HMMs were applied for exhaustive searches for laccase-like proteins in the database of proteins that were annotated in NCBI complete and draft genomes. Overall, 221 two-domain laccases and 1,019 three-domain laccases were identified with the corresponding models (Tab. 6.7). In total, 1,240 unique protein sequences for laccase-like enzymes were discovered in 807 different microorganisms, which are 36% of 2,211 organisms included in the study. 252 organisms encode more than one laccase protein: 58 organisms have 3 proteins, 18 have 4 protein, 16 have 5 proteins and 7 harbor more than 5 genes for laccase-like proteins. The species *Xanthobacter autotrophicus* Py2 contains with 10 genes the highest number of putative laccase genes. Both *Sulfitobacter* sp. NAS-14.1 and *Sorangium cellulosum* So ce56 have eight proteins encoded on their genomes.

Next, the genes in *X. autotrophicus* Py2 are examined in order to deduce the putative functions of laccases. Therefore, annotations of the genes upstream and downstream of the identified laccase genes were obtained from the NCBI database. In the genome of

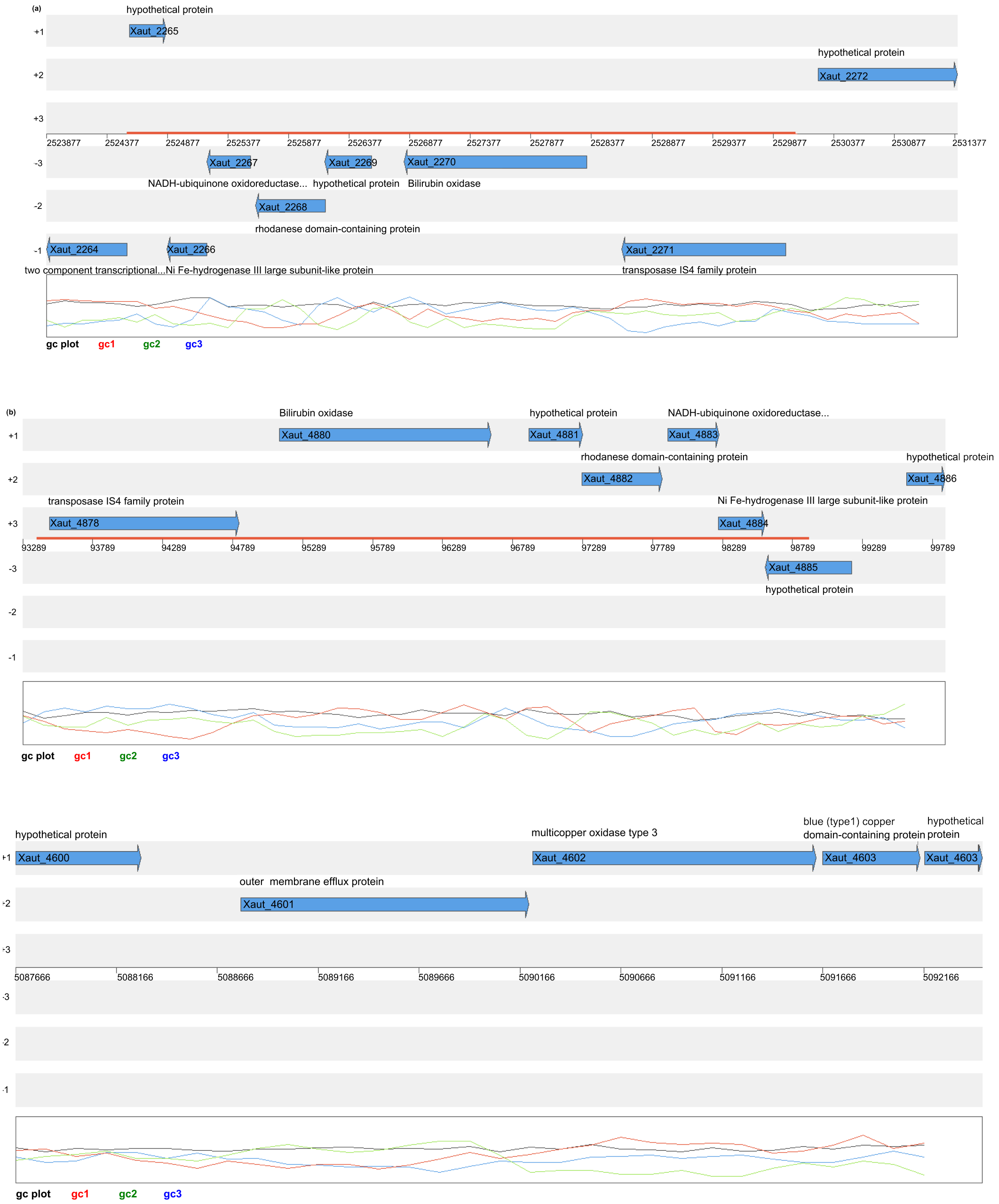


Figure 6.22: Genomic fragments of *Xanthobacter autotrophicus* Py2 encoding laccase-like genes: A fragment of the length 5,528 (red bar on the scales) encoding a laccase is identical on the (a) chromosome and (b) plasmid pXAUT01 of *X. autotrophicus*. (c) An operon encoding an outer membrane efflux protein, laccase and copper domain-containing protein, which might be relevant for mediating resistance for metals in *X. autotrophicus* Py2.

Table 6.7: Summary of the identified laccase-like proteins identified in draft and complete genomes

laccase type	model name	total[1]	unique entries[2]	No (%) of signal peptides
two-domain	typeC2D	63	63	40 (63.5)
	typeB2D	158	158	127 (80.4)
three-domain	small3D	822	355	303 (85.4)
	big3D	308	159	118 (74.2)
	cot3D	200	38	26 (68.4)
		sum	1240[3]	943 (76.0)

[1] Total no. of proteins retrieved with the model

[2] No. of proteins not retrieved with any other model

[3] No. of unique protein entries identified with the five models. Additionally, 467 entries are retrieved by more than one three-domain model.

X. autotrophicus Py2, four of the laccase-like genes are in close proximity to transposases. Two genomic fragments each of the length 5,528 bases are completely identical and carry genes for a putative laccase as well as a transposase. The copied fragments are located on the plasmid pXAUT01 (accession: CP000782, position: 93,391 - 98,918 bp) and the chromosome of *Xanthobacter autotrophicus* Py2 (accession: CP000781, position: 2,524,546 - 2,530,073 bp) (Fig. 6.22). This example shows that a duplication of a fragment carrying a putative laccase has occurred in *X. autotrophicus*.

The actual functions of laccases are not completely understood. Multicopper oxidases were already reported to be involved in sporulation, utilization of plant phenolic compounds and mediating resistance to copper and iodide [Arnesano et al., 2003]. Copper functions as bactericides as a high copper concentration causes damages in molecules. Therefore, a regulation of copper concentration in the cytoplasm of a cell is important. In *X. autotrophicus*, the putative laccase gene (Xaut_4602) is located adjacent to a gene encoding a blue copper domain-containing protein (Xaut_4603) (Fig. 6.22c). The latter protein shows a high similarity to copper resistance proteins. A link between this protein and a laccase was reported in mediating copper resistance [Grass and Rensing, 2001, Arnesano et al., 2003]. In surveys, copper domain-containing proteins and multicopper oxidases are described as periplasmic copper-binding proteins, which regulate copper concentration in the cytoplasm [Grass and Rensing, 2001, Arnesano et al., 2003]. A further gene associated with copper resistance is an outer membrane efflux protein, which is likely involved in copper transportation through the outer membrane [Espariz et al., 2007]. Upstream of the gene encoding the putative laccase (Xaut_4602) is a gene representing an outer membrane efflux protein (Xaut_4601) in the genome of *X. autotrophicus* (Fig. 6.22c). The genes are predicted to be co-regulated in an operon in the OperonDB [Pertea et al., 2009], which stores potential operon structures in bacterial genomes. Proteins encoded in the same operon typically have closely related biological functions. This observation indicates that *X. autotrophicus* Py2 possesses a system that contributes resistance to copper or a similar metal. Moreover, genes encoding multicopper oxidases (Xaut_3196, Xaut_3408, Xaut_3971) are located closely to iron-associated genes (Xaut_3197, Xaut_3198) or cobalt-zinc-cadium efflux genes (Xaut_3406, Xaut_3972) in *X. autotrophicus*. As multicopper oxidases are linked with iron acquisition [Huston et al., 2002] and cobalt, zinc, cadium resistance [Tamás and Martinoia, 2006], the proteins may mediate iron and zinc homeostasis in *X. autotrophicus*. Overall, the putative laccases in *X. autotrophicus* might mainly maintain resistance towards metals or are duplicates by transposase-mediated gene transfers.

In *Pseudomonas stutzeri* A1501, a gene encoding CopRS (PST_2711, PST_2712), which is a regulator for copper tolerance, was identified close to an operon containing genes for a multicopper oxidase (PST_2715) and a copper resistance protein (PST_2717). The gene *copRS* encodes a two-component signal transduction system, which is required for sensing copper ion concentrations and induction of the expression of genes regulating copper homeostasis [Hu et al., 2009, Schelder et al., 2011]. A further regulator involved in copper stress is a member of the MerR family, which is broadly distributed in bacteria. MerR belongs to the metal-responsive regulators and was characterized to regulate the expression of a multicopper oxidase in *E. coli* [Rensing and Grass, 2003]. A gene encoding MerR (Bcep1808_3973) was identified close to a gene for a multicopper oxidase (Bcep1808_3977) on the chromosome of *Burkholderia vietnamiensis* G4. The latter gene is additionally surrounded by genes encoding copper resistance proteins (Bcep1808_3974, Bcep1808_3975) and an outer membrane efflux protein (Bcep1808_3978).

Moreover, the distribution of laccases in microbial genomes was studied. Several phyla were represented with very few sequences, while in other groups many laccase genes were retrieved (Fig. 6.23). As an example, 852 sequences belonged to *Proteobacteria*, which is 69% of the total number of identified laccases. A reason for the unbalanced coverage of the microbial organisms might be the bias of sequencing efforts towards *Proteobacteria*. Within this phylum, 368 sequences belonged to *Gammaproteobacteria* with only 14 of these encoding two-domain laccases. *Alphaproteobacteria* and *Betaproteobacteria* are two further classes that carry 63 and 76 genes, respectively, encoding two-domain laccases, which are completely absent in the classes *Deltaproteobacteria* and *Epsilonproteobacteria*.

In total, 172 proteins were identified in the phylum *Actinobacteria*. Ten of these (6%) are two-domain laccases, which are common among *Streptomyces*. The phyla *Acidobacteria* and *Bacteroidetes* seem to lack two-domain laccases, whereas no three-domain laccases were found in *Planctomycetes*. Finally, 34 laccases were discovered in *Cyanobacteria*. Notably, all eight two-domain laccases within this class were assigned to type-C.

In the phylum *Firmicutes*, 98 laccase-like proteins were uncovered by the model. In some *Firmicutes* species, including *Streptomcyes* and *Bacillus*, multicopper oxidases were suggested to participate in the biosynthesis of a brown spore pigment during sporulation [Hullo et al., 2001]. Indeed, a multicopper oxidase gene (Btus_1147) is

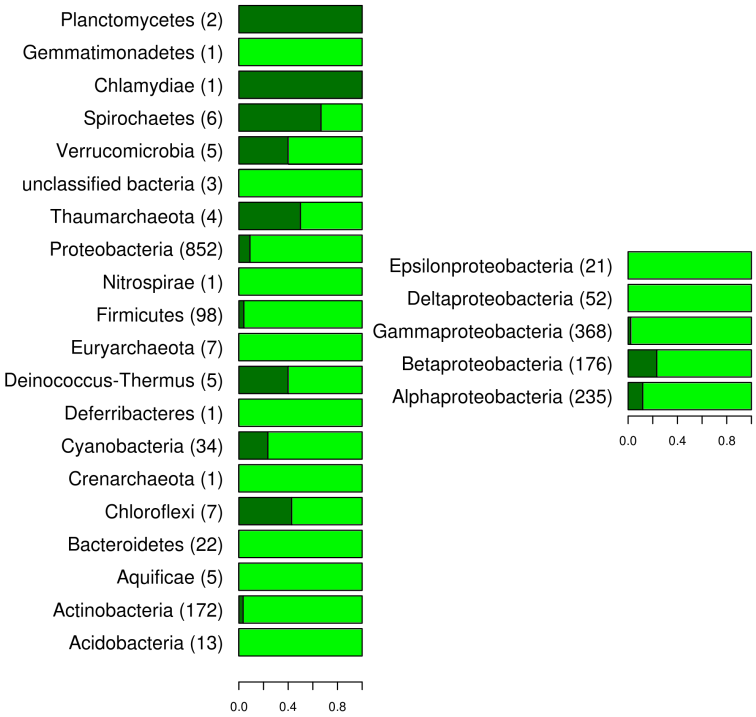


Figure 6.23: Proportions of two-domain (dark green) and three-domain (light green) laccases in phyla and classes of *Proteobacteria*: The relative amount of two-domain laccases is shown for bacterial genomes in different phyla (left) and classes of *Proteobacteria* (right). The numbers in brackets represent the total number of laccase genes found in each taxon.

annotated next to a gene for a spore germination protein (Btus_1146) in *Bacillus tusciae* DSM 2912.

In the next step, the location of the genes for laccases was examined on the genome. The information whether a gene is located in a chromosome or plasmid is only provided by the protein database based on complete genomes. Overall, 749 genes for putative laccases were identified in the genomes, whereas 76 genes were encoded on plasmids originating from 46 different microbes (Fig. 6.24). Some organisms, e.g. species of *Mycobacterium*, *Ralstonia* and *Leuconostoc* carry laccase-like genes only on plasmids (Fig. 6.24). One third of these (34%) are associated with various *Rhizobiales* species, which usually have multiple genes for laccases in their genomes. In the *Rhizobiales* species *Sinorhizobium fredii* NGR234, a laccase-like gene (NGR_b14380) is annotated next to a manganese transport regulator (NGR_b14390). *Rhizobia* establish symbiosis with plants to fix nitrogen [Weidner et al., 2003]. Mutations in a gene for manganese uptake caused a symbiotic defect [Davies and Walker, 2007]. It has been demonstrated that a high manganese level influenced the laccase production in fungi [Stajic et al., 2006]. Therefore, the putative laccase gene might be essential for manganese oxidation in order to establish a functional symbiosis.

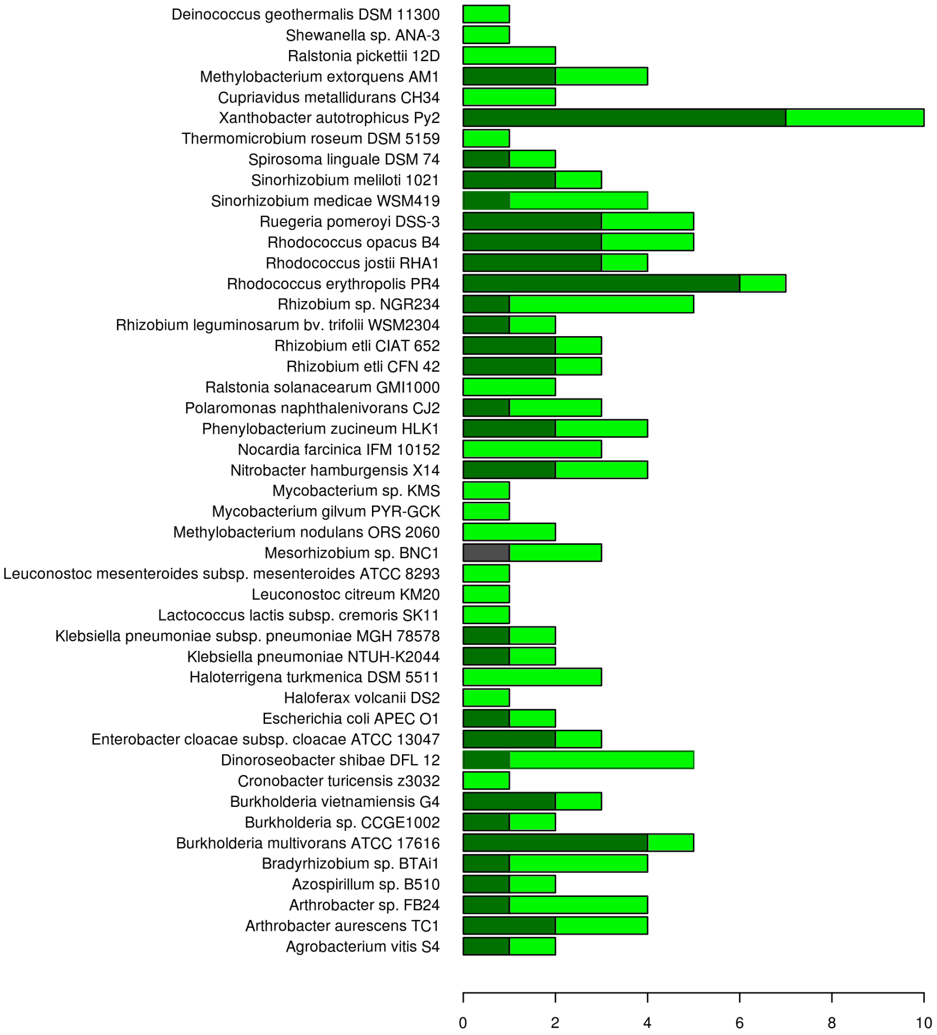
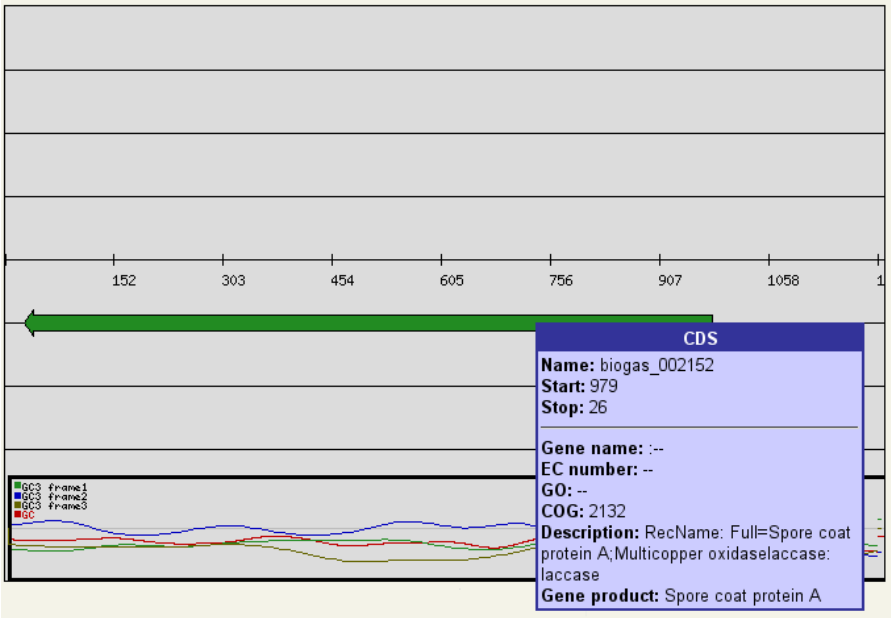


Figure 6.24: List of species that harbor laccases in their plasmids: The bars represent the number of laccase genes in genomes (dark green) and the number of laccase genes in plasmids (light grey). The length of the bar shows the total number of genes identified in each organism.

The identified laccase-like proteins were further analyzed for the presence of signal peptides using SignalP [Petersen et al., 2011]. Surprisingly, three quarters of the enzymes contain putative signal peptides at the N-terminal end indicating that the majority of the bacterial laccases may be exported out of the cytoplasm. So far, this observation is contrary to the current knowledge [Sharma et al., 2007].

6.5.5 Bacterial laccase-like sequences in metagenomic datasets

MetaSAMS stores 27,576 contigs assembled from metagenome reads that were obtained from a biogas plant (Section 6.2). The annotated proteins on the contigs were compared to the five laccase models in order to identify putative laccase genes. Therefore, the HMM interface of MetaSAMS was utilized. Only one protein matched to the model of the three-domain laccases (Fig. 6.25). A subsequent BLAST search against the NCBI protein database using standard settings revealed that the identified laccase protein exhibits a similarity of 59% to a spore coat protein from *Clostridium* sp. 7_2_43FAA (accession: ZP05132033).



Name: biogas_002152
Start: 979
Stop: 26

Gene name: :--
EC number: --
GO: --
COG: 2132
Description: RecName: Full=Spore coat protein A;Multicopper oxidaselaccase: laccase
Gene product: Spore coat protein A

Figure 6.25: Metatig view of a metatig encoding a laccase-like protein: Using the HMM-interface, a putative three-domain laccase was identified on one metatig, which was assembled from reads obtained from a biogas plant.

Recently, it has been reported that marine bacterial species have the potential to degrade lignin [Palanisami et al., 2010]. Laccases were identified in marine metagenomes using functional screenings [Fang et al., 2011]. The discovered laccase has properties of alkalescence-dependent activity, high chloride tolerance and the ability to decolorize several industrial dyes under alkalescent conditions. These characteristics are interesting with respect to potential industrial applications. Because of this, the marine metagenome of the Global Ocean Survey [Venter et al., 2004] was used to search for sequences matching the five laccase models. For this purpose, metagenome reads of the Global Ocean Survey were obtained from CAMERA [Sun et al., 2011]. The metagenome consists of approximately 12 million reads, which were translated into the six reading frames.

The profile HMM-based search retrieved numerous hits for prokaryotic laccases. In total, 277 and 847 translated sequences exhibit similarities to two-domain and three-domain laccase-like sequences, respectively, which properly align to the copper-binding regions. As the average length of the gene fragment encoding the cbr14 of laccases is approximately 600 bases and 1,170 bases in two-domain and three-domain laccases, respectively, the Sanger-reads may not cover all copper-binding regions. However, for the two-domain laccases, some of the translated reads included all four copper-binding regions (Appendix, Fig. A.1).

Finally, a taxonomic profile of the matching 1,124 reads was generated by executing CARMA3 [Gerlach and Stoye, 2011]. In total, 1,045 (93%) sequences could be affiliated to a superkingdom (Fig. 6.26). With 89% of the reads encoding laccases, *Bacteria* are the largest superkingdom. *Archaea* and *Eukaryota* are represented each with 2% of the sequences, and only 7% of the sequences are unknown.

Two of the 22 identified archaeal sequences were assigned to the family *Nitrosopumilaceae*. Both sequences encode three-domain laccases. *Nitrosopumilus maritimus* SCM1 belongs to the same family and is identified in the previous genome-based study with both two- and three-domain laccases. The remaining 20 archaeal sequences are of unknown origin. Surprisingly, 22 sequences were assigned to the eukaryotic superkingdom indicating that the profile HMMs are capable to capture not only laccase-like proteins of bacterial but also of eukaryotic and archaeal origin. Six reads encoding only one or two conserved copper-binding regions were further classified to belong to the green algae from the phylum *Chlorophyta*. Soil algae were recently reported to encode laccase genes [Otto et al., 2010]. Only one of the six sequences was affiliated to an order, namely *Mamiellales*, which contains several widespread marine taxa [McDonald et al., 2010]. In addition, one read was assigned to the class *Dothideomycetes*, which belongs to the *Fungi* kingdom. Laccases were previously described in species of the class *Dothideomycetes* [Luis et al., 2004].

The majority of the metagenome sequences encoding laccases were assigned to the superkingdom *Bacteria* (89%). Only, 5 of the bacterial sequences were not classified on rank phylum. The most abundant phyla belonging to *Bacteria* are *Proteobacteria* (86%) and *Cyanobacteria* (2%). Similar to the previously described genome-based analysis, many laccases were identified among species of *Alphaproteobacteria*, *Betaproteobacteria* and *Gammaproteobacteria*, whereas *Deltaproteobacteria* are with 5 reads encoding only three-domain laccases less common. Within the phylum *Alphaproteobacteria*, the genera *Citromicrobium*, *Sphingomonas*, *Roseobacter* and *Erythrobacter* were identified with more than two EGTs encoding laccases. These genera were also present in the laccase database

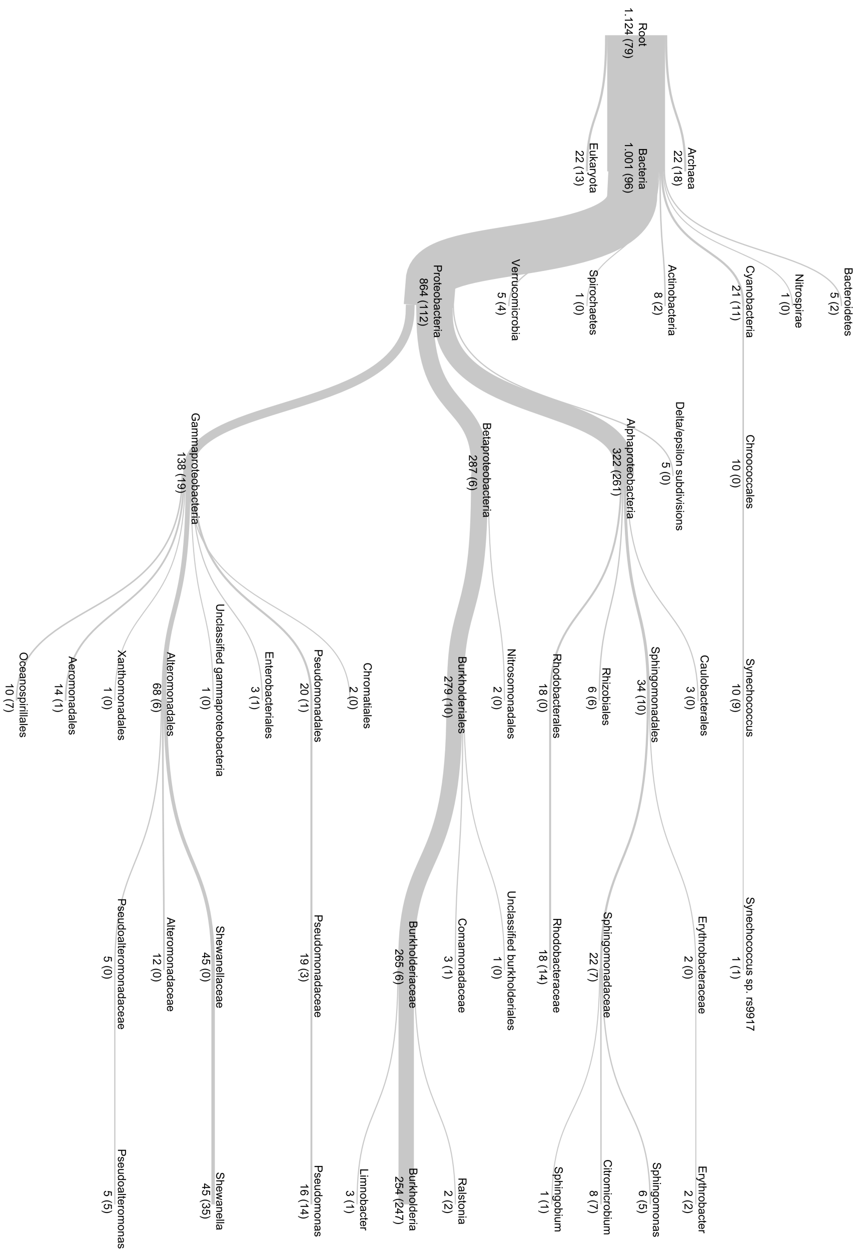


Figure 6.26: Taxonomic tree of metagenome reads encoding putative laccases: Metagenome reads that were matching the laccase models were taxonomically assigned using CARMA3. The classification was visualized by means of an unpublished in-house tool.

obtained from bacterial genomes. *Proteobacteria* is represented by the genera *Ralstonia*, *Limnobacter* and *Burkholderia*. The latter is the most abundant genus that carries laccases-like genes. The most abundant genera belonging to the phylum *Gammaproteobacteria* are *Shewanella*, *Pseudomonas*, *Alteromonas* and *Pseudoalteromonas*. Except of *Glaciecola* and *Rheinheimera*, all genera of *Betaproteobacteria* and *Gammaproteobacteria* are represented by species in the laccase database obtained from bacterial genomes. Two reads with taxonomic assignments to *Glaciecola* and *Rheinheimera* were identified in the metagenome encoding putative laccases. A detailed analysis of the composition of the NCBI protein databases revealed that the two genera were missing in the genome database in September 2010. 10 of the 21 cyanobacterial sequences were further classified to the genus *Synechococcus*, which is present in the genome-based profile. Only 8 laccase genes (0.8%) were assigned to the phylum *Actinobacteria* with *Actinomycetales* as the only order.

The phylum *Bacteroidetes* was identified with five reads, which were deeper classified to the order *Flavobacteriales*. The phylum *Verrucomicrobia* is present with five reads. Both phyla were identified in the taxonomic profile based on genomic laccases (Fig. 6.23). Finally, *Spirochaetes* and *Nitrospirae* were predicted each with one read encoding a two-domain and three-domain laccase, respectively. One read of the phylum *Spirochaetes* was classified to the family *Leptospiraceae*. In the genome database, species of this family possess three-domain laccases. Similarly, *Candidatus Nitrospira defluvii*, the only species of the phylum *Nitrospirae*, is detected to encode laccases in the complete genomes database.

CARMA3 could assign 386 sequences (34%) to a genus. Approximately 66% of the classifiable EGTs encoding laccases were affiliated to *Burkholderia* followed by *Shewanella* and *Pseudomonas* with 11% and 4%, respectively. Finally, 29 sequences were classified on species level. *Burkholderia* sp. TJI49 is represented by 7 reads, whereas *Shewanella* sp. ANA-3 was assigned to 8 EGTs. The latter species was identified in the laccase genome database. The translated metagenome sequences, which were affiliated to *Shewanella* sp. ANA-3, are identical to the proteins encoded by the respective genome.

Overall, the identified bacterial taxa are also abundant in the database of complete and draft genomes carrying laccase genes, except for the organisms that were absent in the NCBI database. This analysis indicates that the HMM-based search is a powerful tool for the identification of bacterial laccase proteins in translated metagenome sequences. Moreover, 34% of the laccase-like sequences were affiliated to a genus indicating the presence of so far unknown taxonomic groups and novel laccase.

CHAPTER 7

Discussion

This chapter highlights and discusses the novelties and major outcomes of the designed and implemented methods for the analysis of whole-shotgun metagenome, 16S rDNA amplicon and metatranscriptome data. Moreover, new insights into the biogas-producing community are emphasized. In this regard, main aspects of the whole-shotgun metagenome, 16S rDNA amplicon and metatranscriptome approach are critically examined. Finally, this chapter focuses on the interpretation of the outcomes of the search for laccase-like genes in genomes and metagenomes.

7.1 MetaSAMS - Advantages and limitations

MetaSAMS is a web-based system for performing automated taxonomic and functional analysis of metagenome data. It complements the existing genomics, transcriptomics, proteomics and metabolomics software platforms at Bielefeld University. The system proved to be a useful platform for exploring metagenome datasets of a complex biogas-producing microbial community as presented in Chapter 6. Apart from this, MetaSAMS has been applied in several collaborations to describe the taxonomic and functional characteristics of metagenomes obtained from various habitats, such as watt sediments, silage and biogas batch systems [Rademacher et al., 2012]. Moreover, results generated by MetaSAMS have been recently published [Rademacher et al., 2012]. This illustrates that the requirements for the storage and interpretation of metagenome data have been successfully addressed.

In this thesis, MetaSAMS has been applied to a metagenome obtained from a biogas plant. The results calculated by MetaSAMS are in accordance to previously described observations of the same biogas plant [Schlüter et al., 2008, Krause et al., 2008a, Jaenicke et al., 2011]. Moreover, the results are similar to the outcomes obtained by other analysis methods such as 16S rRNA clone libraries and polymerase chain reaction single strand conformation polymorphism (PCR-SSCP), though the samples were obtained from different biogas plants [Chachkhiani et al., 2004, Klocke et al., 2007].

The application case within this thesis demonstrated that MetaSAMS is capable of performing various tasks encountered in the metagenomics field. MetaSAMS accomplishes taxonomic profiling, functional assignment, a mapping of reads and various searches for genes of interest. The platform offers parameterized tools that can be submitted to a compute cluster allowing for performing the required analysis in an appropriate time. Moreover, MetaSAMS provides flexible parameter settings for the visualization of the results, for instance, the confidence value obtained from the RDP Classifier is adjustable over the web-interface. Such flexible settings are often not provided in the most commonly used metagenome platforms.
A further advantage of MetaSAMS is the availability of different taxonomic classifiers. The profiles can easily be compared to each other. A comparative analysis of taxonomic analysis is not facilitated in existing metagenome analysis platforms, where the taxonomic analysis relies on the classifications based on best BLAST hits or LCA assignments [Meyer et al., 2008, Huson et al., 2011]. The MetaSAMS system applies state-of-the-art methods and supports the possibility for a rapid integration of novel analysis algorithms. This is in particular an advantage, as taxonomic classification methods are constantly published.
However, the results of the taxonomic tools still have a significant degree of uncertainty. Therefore, the generated profiles should be interpreted with caution. The taxonomic classification relies on the content of used databases, which are often biased towards cultivable organisms. Hence, many so far unknown species remain unassigned or are not classified on lower ranks. To illustrate this point, in MetaSAMS, only 11% of metagenome reads obtained from a biogas-producing community have an assignment on lower ranks using the EGT-based approach implemented in CARMA3 [Gerlach and Stoye, 2011].

In particular, the Metatig pipeline, a full-featured gene calling and annotation pipeline, is a valuable tool for functional analysis. For regional annotation, heterogeneous gene prediction tools are performed. Due to the modular design of MetaSAMS, additional gene prediction tools can easily be integrated into the system. Similarly, functional tools for the annotation of potential coding regions can be added rapidly. The advantage of the Metatig pipeline is that it reduces the number of sequences that have to be functionally analyzed. Instead of comparing each read against entries in databases, genes are annotated. Consequently, the computing time is decreased. The BLAST- and HMM interfaces can be used to search for reads that were not used for the assembly of the contigs. Thereby, MetaSAMS ensures that genes of interest, which are not annotated

on contigs, can still be identified in metagenome reads.
More importantly, the Metatig pipeline enables the discovery of full-length or parts of genes with functions of interest by using targeted searches. This feature is unique for MetaSAMS, as methods for the identification of specific genes are missing in existing metagenome platforms.
Similar to taxonomic assignments, the functional annotation depends on the use of databases, which contain DNA and protein sequences of known organisms. Moreover, the predicted functions should be considered carefully. The identification of a closely homologous gene encoding a specific enzyme does not necessarily imply that the metabolic pathway is present in the metagenome, since some enzymes can occur in multiple pathways, where they catalyze different reactions. A further drawback of the Metatig pipeline is the presence of potential chimeric contigs in the data. Hence, the identified genes might originate from different species. In addition, the abundances of the genes should be critically examined, as highly conserved genes might consist of a high number of reads originating from different organisms.

A bottleneck of MetaSAMS is the storage of the results for large metagenomic datasets, for example, obtained by the Illumina technique. A solution for this problem would be a data reduction step prior to the import of the results to the MetaSAMS platform. Moreover, a proper normalization is required for the interpretation of the results. MetaSAMS estimates the relative abundances in relation to the total number of sequenced reads in an analyzed dataset. However, species with a large genome or long genes are more likely sampled and sequenced. Hence, species and functions will be overrepresented in a taxonomic and functional profile, respectively. Therefore, normalization towards the genome size of the closest known relative would improve the community analysis.

In addition, there are some intrinsic biases that might influence the taxonomic and functional profiling. Metagenomics has still drawbacks regarding the biological procedure. Different extraction methods might influence the genomic material and consequently shift the taxonomic abundances. The variation of taxonomic abundances in relation to DNA isolation methods are well described in metagenomic studies [Delmont et al., 2011]. Despite these disadvantages, metagenomic surveys give valuable and considerable insights into the taxonomic structure and functional characteristics of a microbial community. As illustrated in the application case in this thesis, MetaSAMS is suitable to acquire knowledge about a microbial community of interest and to identify specific genes of potential biotechnological relevance.

7.2 Challenges in 16S rDNA amplicon sequencing

The pipeline AMPLA has been designed that uses scripts of existing pipelines and state-of-the-art methods for the analysis of 16S rDNA amplicon data obtained by high-throughput sequencing. AMPLA includes a quality control step using QIIME [Caporaso et al., 2010], SLP [Huse et al., 2010] and UCHIME [Edgar et al., 2011]. Moreover, OTU clustering is performed using UCLUST [Edgar, 2010] and taxonomic profiling by means of the RDP Classifier [Wang et al., 2007].

An advantage of using high-throughput amplicon sequencing for taxonomic profiling is the higher amount of classifiable reads. In the analyzed biogas sample, 6% of the 16S rDNA amplicon sequences were not assigned to a phylum, whereas 64% and 55% of metagenome 16S rRNA genes and mRNA tags were not classified on this rank, demonstrating that the hypervariable regions covered by the amplicons improve the classification compared to the whole shotgun metagenome sequencing strategy. The amplicon analysis reveals that *Methanoculleus* is among the most abundant genera in the biogas-producing community. In addition, the analysis of the ten most abundant OTUs indicates that for some of the OTUs no corresponding reference sequence to known, well-characterized species are available suggesting that still many species residing in the biogas plant are unknown. However, the sequences are representatives of valid species, as identical sequences are found in digestion systems fed with other substrates [Riviére et al., 2009]. The distribution of identical sequences in further habitats highlights the relevance of the species in the anaerobic digestion process. The deeper resolution of high-throughput sequences became noticeable when comparing the representative sequences for some dominant OTUs in 16S rRNA clone libraries of the same biogas sample [Kröber et al., 2009]. The amplicon survey detected highly represented OTUs that were missing in 16S rRNA clone libraries of the analyzed fermentation sample of the same biogas plant [Kröber et al., 2009].

The subsequent phylogenetic analysis of archaeal sequences confirms previous results that species of the genus *Methanoculleus* are highly represented within the methanogenic archaeal community [Kröber et al., 2009]. The identification of phylogenetic clusters without known species references highlights that the archaeal community is so far incompletely described. Identical 16S rDNA amplicon sequences occur in other anaerobic habitats suggesting the importance of these archaeal species. Notably, archaeal OTUs were detected that are similar to sequences obtained from other habitats. The phylogenetic analysis affiliates the representative sequence of an unknown *Archaea* to a species related to *Methanomassiliicoccus luminyensis* B10 [Dridi et al., 2012]. Although this OTU is common in other samples obtained from Italian rice field soil [Liu and Conrad, 2011] or the human gut, a taxonomic characterization for this OTU on species level is until now not available.

Still, the processing of 16S rDNA amplicon sequences is a challenging procedure, which became obvious during the interpretation of the data. The analysis deals with chimeric sequences and pyrosequencing errors that may overestimate the number of OTUs in a sample. The existing tools identify such artifacts not accurately. Some representative archaeal 16S rDNA sequences are likely chimeric, since a manual BLAST search for selected 16S rDNA amplicon sequences produced hits to different organisms for each end of the reads. Because of this problem, researchers have suggested to remove singletons from the post-processing analysis [Reeder and Knight, 2009]. This step would eliminate rare species such as a related *Aminobacterium* species, which is represented with one read in the 16S rDNA amplicon dataset.

Further issues in 16S rDNA analysis occur in steps prior to the sequencing. The usage of primers for the amplification may induce a bias towards the replicated fragments. Comparing the taxonomic classifications of 16S rDNA amplicon sequences with the profile of the metatranscriptome 16S rRNA fragments shows the absence of *Methanobacterium*. So far, the reason for the absence of sequences in the amplicon dataset is not clear. A missing taxon may shift the relative abundances. Another issue is the phylogenetic marker itself that may influence the relative abundances. In some microbes, multiple 16S rRNA genes occur. Thus, the amplicon sequences would overestimate the abundances of these species. A normalization of the taxonomic results would be necessary to obtain reliable abundances of the corresponding species. Although there are several pipelines and tools available, manual inspections were required during the performed analysis, as the tools have missed artifact sequences in the quality control step. Nevertheless, 16S rDNA amplicon sequencing is a valuable and cheap method to get a global view of the microbes that reside in a habitat of interest.

7.3 Analysis of a metatranscriptome of a biogas-producing community by means of MeTra

The results of the first metatranscriptome approach of a biogas-producing community have been demonstrated in this thesis. In this respect, the pipeline MeTra has been designed that allows the annotation of different RNA types including rRNA, mRNA and non-coding RNA. Most of these were assigned to the phyla *Firmicutes* and *Euryarchaeota*. This tendency was confirmed by a profile based on expressed mRNA tags indicating that these phyla contribute most of the transcripts in the biogas plant. Transcripts for enzymes functioning in methanogenesis are among the most abundant mRNA tags indicating that the corresponding pathway is very active in the methanogenic subcommunity. As the metatranscriptome was not enriched for mRNA tags, the number of sequences encoding proteins is very low. Nevertheless, genes for enzymes participating in major steps of anaerobic digestion were identified among the mRNA tags. To obtain a deeper functional profile, mRNA enrichment or rRNA depletion would be required. Recently, techniques for removal of ribosomal RNA in a metatranscriptome RNA preparation were outlined [He et al., 2010a].

In addition, a detailed study of non-coding RNA was carried out. The identification of non-coding RNAs enables to broaden the research field of metatranscriptomics. Studies of non-coding RNAs in metatranscriptomes are very rare [Shi et al., 2009, Gosalbes et al., 2011]. However, the most abundant non-coding RNA families are also highly abundant in the transcriptome of microbial communities from different habitats [Shi et al., 2009, Gosalbes et al., 2011] as well as of single microorganisms [Block et al., 2011] indicating that the pipeline produces valuable results. Unfortunately, functional, non-coding RNAs are so far not well described making it challenging to decipher the regulation process based on non-coding RNAs. The taxonomic analysis based on non-coding RNAs showed that they are highly transcribed by species participating in the anaerobic digestion process. Therefore, understanding the functions and regulations of non-coding RNAs in the biogas plant might help to reconstruct the biogas formation process and control the methane yield.

Several factors have to be considered as intrinsic biases during the metatranscriptome analysis implicating that the data might not represent the whole complexity of transcripts synthesized by a microbial community [Velculescu et al., 1995]. Transcript-based analysis may be influenced by the instability, rapid turnover and short cellular half-life of the RNA [Poretsky et al., 2005]. In addition, biases may be introduced during RNA extraction and enzymatic conversion of RNA into cDNA. As a consequence, functional and taxonomic information might remain unexplored in the data analysis. Similarly to the metagenome-based analysis, the taxonomic and functional profiles are dependent on existing databases, which are likely biased towards cultured species. Despite these pitfalls, metatranscriptomics is a potential approach to address questions regarding active organisms and functions of the biogas-producing community.

Overall, three approaches, namely whole metagenome shotgun, 16S rDNA amplicon and metatranscriptomics, were performed to study a biogas-producing community. Each of these approaches has advantages and disadvantages. However, a combination of the outcomes gives a comprehensive understanding of the organisms residing in a biogas plant and their metabolic functions.

7.4 Identification of laccases using hidden Markov models

Nature has invented a variety of enzymes, which are potentially useful for biotechnological applications. Instead of engineering industrially optimal enzymes, it is possible to search for genes of interest encoded by microorganisms that live in environments matching industrial conditions. Herein, a method based on profile hidden Markov models (HMM) [Eddy, 2011] has been designed and applied to identify genes encoding laccases-like enzymes in metagenomes obtained from the biogas-producing community as well as an ocean sampling project. Such probabilistic models of protein families are commonly used in the analysis of high-throughput sequencing data [Krause et al., 2008a, Pope et al., 2010]. The main advantage of a profile HMM-based approach is the high accuracy in detecting conserved domains compared to other methods such as BLAST. As laccase proteins are conserved in the four copper binding regions, the usage of profile HMMs is suitable for a sequence-based search.

Since salt- and pH-tolerant laccases are desired for industrial applications, marine metagenomes are promising to identify laccases with desired characteristics. Using *in silico* screenings, novel putative laccase genes were discovered that might be relevant for industrial applications. Moreover, reads were identified that covered all central regions of the small bacterial laccases (two-domain laccase). In the metagenome from a biogas plant only one gene has been found. A probable reason for the low number of laccase-like genes is the anaerobic environment in the biogas plant. For activity, laccases require aerobic conditions. However, it is likely that the organisms residing in the biogas fermenter use other enzymes, such as peroxidases, for breakdown of phenolic plant material.

The method presented in this thesis allows the identification of sequences matching a specific model. Hence, still it is not known that such proteins function as laccases in biotechnological applications. However, it was already demonstrated that reads identified in a metagenome by sequence-based screenings harbored *in vitro* the function of interest [Warnecke et al., 2007, Pope et al., 2010, Hess et al., 2011].

The generated models were applied to gain detailed knowledge about the diversity and functions of bacterial laccases. So far, not so much has been described about bacterial laccases, as their discovery is relatively new [Alexandre and Zhulin, 2000]. Because of this, the models were used to capture laccase-like proteins encoded in published bacterial genomes. This study clearly illustrated a broad distribution of laccases in the bacterial world. Laccase-like genes are also diverse within a single species. An explanation for numerous laccases in a single species is that the enzymes function in different pathways such as pigment formation and stress resistance. The analysis provides evidence that multiple laccases are results of duplication events mediated by transposases. Notably, signal peptides are identified in approximately 76% of the putative proteins suggesting that they may be secreted from the cytoplasm. As laccase-like enzymes with signal peptides are present in anaerobic organisms, they might be active in a more aerobic environment away from the cytoplasm.

CHAPTER 8

Conclusion and outlook

Researchers examine the microbial life from different angles using whole metagenome shotgun, gene-centric, metatranscriptome and metaproteome approaches. The work in this thesis contributes analysis methods for the emerging fields of whole metagenomics, 16S rDNA amplicon research and metatranscriptomics. The methods have been successfully applied on respective data obtained from a biogas fermenter. Finally, a method has been proposed for the discovery of genes for industrially relevant enzymes. Based thereon, novel laccase genes could be identified in metagenomes and genomes. Hence, all objectives introduced in Chapter 4 are realized within this thesis.

New sequencing technologies enable the accomplishment of metagenomics and metatranscriptomics projects at affordable costs and appropriate time. Simultaneously, they have boosted the number and size of sequencing projects. Because of this, methods and concepts for the analysis of metagenome and metatranscriptome data are continuously evolving. In the context of this work, MetaSAMS has been developed that tackles the large data volumes and characterizes the short reads in terms of their origin and function. MetaSAMS performs taxonomic characterizations based on three different classifiers, but the modular design allows the integration of novel taxonomic tools. To illustrate the features of MetaSAMS, it has been applied for the automated analysis of 454 pyrosequencing reads. Recently, a metagenome obtained by sequencing using Ion Torrent Technology with the chip 316 has been imported and analyzed in the system for an external collaboration partner. Metagenome projects are increasingly carried out by means of Illumina sequencing. The platform needs to be extended by further functionalities in order to ensure the analysis of the large data volumes. For the analysis of Illumina reads, a data reduction step would be required in the first place, e.g., by clustering similar sequences using UCLUST. Another possibility to manage the high-throughput data is to explore the metagenome based on contigs assembled from Illumina reads by means of the Metatig pipeline implemented in MetaSAMS.

The next milestone that will influence the microbial research field is the third-generation sequencing technology. The novel sequencing methods claim to produce longer reads, which reach sizes of current assembled contigs. Hence, some short-read tools may be outdated. MetaSAMS attempts to functionally analyze the long reads using the Metatig pipeline. Moreover, due to the modular design of MetaSAMS, contig-based taxonomic classifiers such as the intrinsic tools TACOA [Diaz et al., 2009] and PhyloPythia [McHardy et al., 2007], can be integrated into MetaSAMS.
To ease the functional annotations, it is important to unveil the functions of so far unknown genes. For this purpose, traditional genomics will still have to complement metagenomics that can directly characterize a specific microbe. In particular, the novel single cell genome sequencing [Yilmaz and Singh, 2011] approach is promising, as it uncovers the genomes of uncultivable species within a community.

The 16S rDNA amplicon analysis in this work unveils major problems that complicate the final interpretation of the data. Chimeric sequences produced during PCR amplification inflate the number of estimated OTUs. A manual analysis of the processed sequences indicated that still some potential chimeric sequences remain undetected. Chimera formation might be likely when two hypervariable regions are amplified, as the conserved regions between the hypervariable ones from different tags might attach together during PCR. Therefore, an analysis of 16S rDNA amplicon sequences covering one hypervariable region would be desired. Based thereon, also bias introduced by primer sequences during the PCR can be studied.

Herein, the metatranscriptome of the biogas plant has been exhaustively examined. Approximately, 90% of the metatranscriptome sequences were identified as rRNA tags, which enabled analysis of the diversity of metabolically active microorganisms. Unfortunately, only 2.6% of the metatranscriptome reads represent mRNA tags. Therefore, an efficient mRNA enrichment method is needed to deepen the functional analysis derived from mRNA tags. Further sequencing efforts of cDNA obtained from mRNA-enriched RNA preparations would be required for a more precise picture of the active functional transcripts of the biogas-producing community.

In this work all methods have been applied on data obtained from a biogas community. Together, they provide a comprehensive view of the heterogeneous community and the biogas-formation process. Still, much remains to be learned regarding the microorganisms and their roles in the biogas plant.

The fields of metagenomics and metatranscriptomics give the potential for discovering novel enzymes. More and more projects arise that aim to identify enzymes using sequence- and functional-based methods. Therefore, an HMM-based method has been realized for the identification of laccases. The same method will be repeated for the detection of chitinases and chitin-binding proteins in order to get knowledge about the sequence and species diversity. Finally, the sequence information will support the design of oligonucleotide primers for screenings of chitinase genes in metagenomic libraries. The current primers for chitinase genes do not cover the whole diversity of chitinases. Instead they are biased to known species, e.g., of the genus *Streptomyces*. A collection of chitinases from soil or marine metagenomes would supplement the current sequence diversity. This knowledge could be elaborated to construct novel primers for screening of clone libraries.

Bibliography

[Agogué et al., 2011] Agogué, H., Lamy, D., Neal, P. R., Sogin, M. L., and Herndl, G. J. (2011). Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Mol. Ecol.*, 20(2):258–274.

[Ahring, 2003] Ahring, B. K. (2003). Perspectives for anaerobic digestion. *Adv. Biochem. Eng. Biotechnol.*, 81:1–30.

[Alexandre and Zhulin, 2000] Alexandre, G. and Zhulin, I. B. (2000). Laccases are widespread in bacteria. *Trends Biotechnol.*, 18(2):41–42.

[Altermann and Kazmierczak, 2003] Altermann, W. and Kazmierczak, J. (2003). Archean microfossils: a reappraisal of early life on Earth. *Res. Microbiol.*, 154:611–617.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

[Amann et al., 1995] Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59:143–169.

[Aparicio et al., 2002] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y. H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, 297(5585):1301–1310.

[Apweiler et al., 2011] Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., Bower, L., Browne, P., Chan, W. M., Dimmer, E., Eberhardt, R., Fazzini, F., Fedotov, A., Foulger, R., Garavelli, J., Castro, L. G., Huntley, R., Jacobsen, J., Kleen, M., Laiho, K., Legge, D., Lin, Q., Liu, W., Luo, J., Orchard, S., Patient, S., Pichler, K., Poggioli, D., Pontikos, N., Pruess, M., Rosanoff, S., Sawford, T., Sehra, H., Turner, E., Corbett, M., Donnelly, M., van Rensburg, P., Xenarios, I., Bougueleret, L., Auchincloss, A., Argoud-Puy, G., Axelsen, K., Bairoch, A., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Bollondi, L., Boutet, E., Quintaje, S. B., Breuza, L., Bridge, A., deCastro, E., Coudert, E., Cusin, I., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gehant, S., Ferro, S., Gasteiger, E., Gateau, A., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hulo, N., James, J., Jimenez, S., Jungo, F., Kappler, T., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Martin, X., Masson, P., Moinat, M., Morgat, A., Paesano, S., Pedruzzi, I., Pilbout, S., Poux, S., Pozzato, M., Redaschi, N., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stanley, E., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Barker, W. C., Chen, C., Chen, Y., Dubey, P., Huang, H., Mazumder, R., McGarvey, P., Natale, D. A., Natarajan, T. G., Nchoutmboube, J., Roberts, N. V., Suzek, B. E., Ugochukwu, U., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., and Zhang, J. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, 39(Database issue):D214–219.

[Arias et al., 2003] Arias, M. E., Arenas, M., Rodriguez, J., Soliveri, J., Ball, A. S., and Hernandez, M. (2003). Kraft pulp biobleaching and mediated oxidation of a non-phenolic substrate by laccase from *Streptomyces cyaneus* CECT 3335. *Appl. Environ. Microbiol.*, 69(4):1953–1958.

[Arnesano et al., 2003] Arnesano, F., Banci, L., Bertini, I., Mangani, S., and Thompsett, A. R. (2003). A redox switch in CopC: an intriguing copper trafficking protein that binds copper(I) and copper(II) at different sites. *Proc. Natl. Acad. Sci. U.S.A.*, 100(7):3814–3819.

[Asakawa and Nagaoka, 2003] Asakawa, S. and Nagaoka, K. (2003). *Methanoculleus bourgensis*, *Methanoculleus olentangyi* and *Methanoculleus oldenburgensis* are subjective synonyms. *Int. J. Syst. Evol. Microbiol.*, 53(Pt 5):1551–1552.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29.

[Ashelford et al., 2005] Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, 71:7724–7736.

[Atlas and Bartha, 1998] Atlas, R. M. and Bartha, R. (1998). *Microbial ecology: fundamentals and applications*. Benjamin Cummings series in the life sciences. Benjamin/Cummings.

[Avery et al., 1944] Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medecine*, 79(2):137–158.

[Baena et al., 1998] Baena, S., Fardeau, M. L., Labat, M., Ollivier, B., Thomas, P., Garcia, J. L., and Patel, B. K. (1998). *Aminobacterium colombiense*gen. nov. sp. nov., an amino acid-degrading anaerobe isolated from anaerobic sludge. *Anaerobe*, 4(5):241–250.

[Baker et al., 2003] Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods*, 55(3):541–555.

[Balk et al., 2002] Balk, M., Weijma, J., and Stams, A. J. (2002). *Thermotoga lettingae* sp. nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. *Int. J. Syst. Evol. Microbiol.*, 52(Pt 4):1361–1368.

[Bao et al., 2002] Bao, Q., Tian, Y., Li, W., Xu, Z., Xuan, Z., Hu, S., Dong, W., Yang, J., Chen, Y., Xue, Y., Xu, Y., Lai, X., Huang, L., Dong, X., Ma, Y., Ling, L., Tan, H., Chen, R., Wang, J., Yu, J., and Yang, H. (2002). A complete sequence of the *T. tengcongensis* genome. *Genome Res.*, 12(5):689–700.

[Barriuso et al., 2011] Barriuso, J., Valverde, J. R., and Mellado, R. P. (2011). Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics*, 12:473.

[Baxevanis and Ouellette, 2004] Baxevanis, A. and Ouellette, B. (2004). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Methods of Biochemical Analysis. John Wiley & Sons.

[Bayer et al., 2009] Bayer, T. S., Widmaier, D. M., Temme, K., Mirsky, E. A., Santi, D. V., and Voigt, C. A. (2009). Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc.*, 131(18):6508–6515.

[Bekel et al., 2009] Bekel, T., Henckel, K., Kuster, H., Meyer, F., Mittard Runte, V., Neuweger, H., Paarmann, D., Rupp, O., Zakrzewski, M., Pühler, A., Stoye, J., and Goesmann, A. (2009). The Sequence Analysis and Management System – SAMS-2.0: data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies. *J. Biotechnol.*, 140:3–12.

[Benson et al., 2011] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). Genbank. *Nucleic Acids Research*, 39(suppl 1):D32–D37.

[Bentley and Parkhill, 2004] Bentley, S. D. and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.*, 38:771–792.

[Berger and Stamatakis, 2011] Berger, S. A. and Stamatakis, A. (2011). Aligning short reads to reference alignments and trees. *Bioinformatics*, 27(15):2068–2075.

[Bernstein and Hyndman, 2001] Bernstein, H. D. and Hyndman, J. B. (2001). Physiological basis for conservation of the signal recognition particle targeting pathway in *Escherichia coli*. *J. Bacteriol.*, 183(7):2187–2197.

[Bertin et al., 2008] Bertin, P. N., Médigue, C., and Normand, P. (2008). Advances in environmental genomics: towards an integrated view of micro-organisms and ecosystems. *Microbiology*, 154:347–359.

[Bessman et al., 1958] Bessman, M. J., Lehman, I. R., Simms, E. S., and Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid. II. General properties of the reaction. *J. Biol. Chem.*, 233(1):171–177.

[Biesbroek et al., 2012] Biesbroek, G., Sanders, E. A., Roeselers, G., Wang, X., Caspers, M. P., Trzciński, K., Bogaert, D., and Keijser, B. J. (2012). Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PLoS ONE*, 7(3):e32942.

[Block et al., 2011] Block, K. F., Puerta-Fernandez, E., Wallace, J. G., and Breaker, R. R. (2011). Association of OLE RNA with bacterial membranes via an RNA-protein interaction. *Mol. Microbiol.*, 79(1):21–34.

[Bonnet et al., 2002] Bonnet, R., Suau, A., Doré, J., Gibson, G. R., and Collins, M. D. (2002). Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs. *Int. J. Syst. Evol. Microbiol.*, 52(Pt 3):757–763.

[Bourbonnais and Paice, 1990] Bourbonnais, R. and Paice, M. G. (1990). Oxidation of non-phenolic substrates. An expanded role for laccase in lignin biodegradation. *FEBS Lett.*, 267(1):99–102.

[Boutet et al., 2007] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, 406:89–112.

[Branton et al., 2008] Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, 26:1146–1153.

[Braslavsky et al., 2003] Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 100:3960–3964.

[Brodie et al., 2007] Brodie, E. L., DeSantis, T. Z., Parker, J. P., Zubietta, I. X., Piceno, Y. M., and Andersen, G. L. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. *Proc. Natl. Acad. Sci. U.S.A.*, 104(1):299–304.

[Bulow et al., 2008] Bulow, S. E., Francis, C. A., Jackson, G. A., and Ward, B. B. (2008). Sediment denitrifier community composition and *nirS* gene expression investigated with functional gene microarrays. *Environ. Microbiol.*, 10:3057–3069.

[Cantarel et al., 2009] Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.*, 37(Database issue):D233–238.

[Caporaso et al., 2010] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336.

[Cardinali-Rezende et al., 2009] Cardinali-Rezende, J., Debarry, R. B., Colturato, L. F., Carneiro, E. V., Chartone-Souza, E., and Nascimento, A. M. (2009). Molecular identification and dynamics of microbial communities in reactor treating organic household waste. *Appl. Microbiol. Biotechnol.*, 84(4):777–789.

[Chachkhiani et al., 2004] Chachkhiani, M., Dabert, P., Abzianidze, T., Partskhaladze, G., Tsiklauri, L., Dudauri, T., and Godon, J. J. (2004). 16S rDNA characterisation of bacterial and archaeal communities during start-up of anaerobic thermophilic digestion of cattle manure. *Bioresour. Technol.*, 93(3):227–232.

[Chen and Dong, 2005] Chen, S. and Dong, X. (2005). *Proteiniphilum acetatigenes* gen. nov., sp. nov., from a UASB reactor treating brewery wastewater. *Int. J. Syst. Evol. Microbiol.*, 55(Pt 6):2257–2261.

[Chen et al., 2007] Chen, X. H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Heinemeyer, I., Morgenstern, B., Voss, B., Hess, W. R., Reva, O., Junge, H., Voigt, B., Jungblut, P. R., Vater, J., Süssmuth, R., Liesegang, H., Strittmatter, A., Gottschalk, G., and Borriss, R. (2007). Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nat. Biotechnol.*, 25:1007–1014.

[Chistoserdova, 2010] Chistoserdova, L. (2010). Recent progress and new challenges in metagenomics for biotechnology. *Biotechnol. Lett.*, 32(10):1351–1359.

[Cirne et al., 2007] Cirne, D. G., Lehtomäki, A., Björnsson, L., and Blackall, L. L. (2007). Hydrolysis and microbial community analyses in two-stage anaerobic digestion of energy crops. *J. Appl. Microbiol.*, 103:516–527.

[Claus et al., 2002] Claus, H., Faber, G., and König, H. (2002). Redox-mediated decolorization of synthetic dyes by fungal laccases. *Appl. Microbiol. Biotechnol.*, 59(6):672–678.

[Cline et al., 1996] Cline, J., Braman, J. C., and Hogrefe, H. H. (1996). PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.*, 24(18):3546–3551.

[Cole et al., 2003] Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., Chandra, S., McGarrell, D. M., Schmidt, T. M., Garrity, G. M., and Tiedje, J. M. (2003). The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.*, 31:442–443.

[Corbino et al., 2005] Corbino, K. A., Barrick, J. E., Lim, J., Welz, R., Tucker, B. J., Puskarz, I., Mandal, M., Rudnick, N. D., and Breaker, R. R. (2005). Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol.*, 6(8):R70.

[Cover and Hart, 1967] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

[Craig et al., 2010] Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C., and Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl. Environ. Microbiol.*, 76:1633–1641.

[Daniel, 2005] Daniel, R. (2005). The metagenomics of soil. *Nat. Rev. Microbiol.*, 3(6):470–478.

[Davies and Walker, 2007] Davies, B. W. and Walker, G. C. (2007). Disruption of *sitA* compromises *Sinorhizobium meliloti* for manganese uptake required for protection against oxidative stress. *J. Bacteriol.*, 189(5):2101–2109.

[De Schrijver et al., 2010] De Schrijver, J. M., De Leeneer, K., Lefever, S., Sabbe, N., Pattyn, F., Van Nieuwerburgh, F., Coucke, P., Deforce, D., Vandesompele, J., Bekaert, S., Hellemans, J., and Van Criekinge, W. (2010). Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline. *BMC Bioinformatics*, 11:269.

[Delcher et al., 2007] Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23:673–679.

[Delmont et al., 2012] Delmont, T. O., Prestat, E., Keegan, K. P., Faubladier, M., Robe, P., Clark, I. M., Pelletier, E., Hirsch, P. R., Meyer, F., Gilbert, J. A., Le Paslier, D., Simonet, P., and Vogel, T. M. (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J.*

[Delmont et al., 2011] Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R., and Vogel, T. M. (2011). Accessing the soil metagenome for studies of microbial diversity. *Appl. Environ. Microbiol.*, 77(4):1315–1324.

[Demirel and Scherer, 2008] Demirel, B. and Scherer, P. (2008). The roles of acetotrophic and hydrogenotrophic methanogens during anaerobic conversion of biomass to methane: a review. *Reviews in Environmental Science and Biotechnology*, 7:173–190. 10.1007/s11157-008-9131-1.

[DeSantis et al., 2006] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72(7):5069–5072.

[Desbrosses and Stougaard, 2011] Desbrosses, G. J. and Stougaard, J. (2011). Root nodulation: a paradigm for how plant-microbe symbiosis influences host developmental pathways. *Cell Host Microbe*, 10:348–358.

[Deublein and Steinhauser, 2008] Deublein, D. and Steinhauser, A. (2008). *Biogas from waste and renewable resources: an introduction*. Wiley-VCH.

[Diaz et al., 2009] Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., and Nattkemper, T. W. (2009). TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, 10:56.

[Dondrup et al., 2009] Dondrup, M., Albaum, S. P., Griebel, T., Henckel, K., Jünemann, S., Kahlke, T., Kleindt, C. K., Küster, H., Linke, B., Mertens, D., Mittard-Runte, V., Neuweger, H., Runte, K. J., Tauch, A., Tille, F., Pühler, A., and Goesmann, A. (2009). EMMA 2–a MAGE-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics*, 10:50.

[Dridi et al., 2012] Dridi, B., Henry, M.and Richet, H., Raoult, D., and Drancourt, M. (2012). Age-related prevalence of *Methanomassiliicoccus luminyensis* in the human gut microbiome. *APMIS*, pages n/a–n/a.

[Durbin et al., 2006] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (2006). *Biological sequence analysis*. eleventh edition.

[Eddy, 2011] Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, 7(10):e1002195.

[Edgar, 2004a] Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

[Edgar, 2004b] Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

[Edgar, 2010] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

[Edgar et al., 2011] Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200.

[Edwards et al., 2006] Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., Alexander, E. C., and Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, 7:57.

[Eloe et al., 2011] Eloe, E. A., Fadrosh, D. W., Novotny, M., Zeigler Allen, L., Kim, M., Lombardo, M. J., Yee-Greenbaum, J., Yooseph, S., Allen, E. E., Lasken, R., Williamson, S. J., and Bartlett, D. H. (2011). Going deeper: metagenome of a hadopelagic microbial community. *PLoS ONE*, 6(5):e20388.

[Endo et al., 2003] Endo, K., Hayashi, Y., Hibi, T., Hosono, K., Beppu, T., and Ueda, K. (2003). Enzymological characterization of EpoA, a laccase-like phenol oxidase produced by *Streptomyces griseus*. *J. Biochem.*, 133(5):671–677.

[Espariz et al., 2007] Espariz, M., Checa, S. K., Audero, M. E., Pontel, L. B., and Soncini, F. C. (2007). Dissecting the *Salmonella* response to copper. *Microbiology*, 153(Pt 9):2989–2997.

[Evans and Furlong, 2011] Evans, G. and Furlong, J. C. (2011). *Environmental biotechnology*. Wiley-Blackwell, Oxford.

[Fang et al., 2011] Fang, Z., Li, T., Wang, Q., Zhang, X., Peng, H., Fang, W., Hong, Y., Ge, H., and Xiao, Y. (2011). A bacterial laccase from marine microbial metagenome exhibiting chloride tolerance and dye decolorization ability. *Appl. Microbiol. Biotechnol.*, 89(4):1103–1110.

[Feng et al., 2009] Feng, X., Mouttaki, H., Lin, L., Huang, R., Wu, B., Hemme, C. L., He, Z., Zhang, B., Hicks, L. M., Xu, J., Zhou, J., and Tang, Y. J. (2009). Characterization of the central metabolic pathways in *Thermoanaerobacter* sp. strain X514 via isotopomer-assisted metabolite analysis. *Appl. Environ. Microbiol.*, 75(15):5001–5008.

[Fernández-Arrojo et al., 2010] Fernández-Arrojo, L., Guazzaroni, M. E., López-Cortés, N., Beloqui, A., and Ferrer, M. (2010). Metagenomic era for biocatalyst identification. *Curr. Opin. Biotechnol.*, 21:725–733.

[Ferrer et al., 2009] Ferrer, M., Beloqui, A., Timmis, K. N., and Golyshin, P. N. (2009). Metagenomics for mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.*, 16:109–123.

[Finn et al., 2006] Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.*, 34:D247–251.

[Flint, 2011] Flint, H. J. (2011). Obesity and the gut microbiota. *J. Clin. Gastroenterol.*, 45 Suppl:S128–132.

[Forns et al., 1997] Forns, X., Bukh, J., Purcell, R. H., and Emerson, S. U. (1997). How *Escherichia coli* can bias the results of molecular cloning: preferential selection of defective genomes of hepatitis C virus during the cloning procedure. *Proc. Natl. Acad. Sci. U.S.A.*, 94(25):13909–13914.

[Frias-Lopez et al., 2008] Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and Delong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. U.S.A.*, 105:3805–3810.

[Friedrich, 2005] Friedrich, M. W. (2005). Methyl-coenzyme M reductase genes: unique functional markers for methanogenic and anaerobic methane-oxidizing *Archaea*. *Meth. Enzymol.*, 397:428–442.

[Galand et al., 2009] Galand, P. E., Casamayor, E. O., Kirchman, D. L., and Lovejoy, C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc. Natl. Acad. Sci. U.S.A.*, 106(52):22427–22432.

[Garrity and Lilburn, 2004] Garrity, G. M., B. J. and Lilburn, T. (2004). *Taxonomic outline of the procaryotes. Bergey's manual of systematic bacteriology*. Springer-Verlag.

[Gerlach et al., 2009] Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., and Stoye, J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10:430.

[Gerlach and Stoye, 2011] Gerlach, W. and Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.*, 39(14):e91.

[Gilbert et al., 2008] Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., and Joint, I. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE*, 3:e3042.

[Giovannoni et al., 1990] Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345:60–63.

[Glenn, 2011] Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 11:759–769.

[Gomez-Alvarez et al., 2009] Gomez-Alvarez, V., Teal, T. K., and Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J*, 3(11):1314–1317.

[Gonzalez et al., 2005] Gonzalez, J. M., Zimmermann, J., and Saiz-Jimenez, C. (2005). Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics*, 21(3):333–337.

[Gosalbes et al., 2011] Gosalbes, M. J., Durán, A., Pignatelli, M., Abellan, J. J., Jiménez-Hernández, N., Pérez-Cobas, A. E., Latorre, A., and Moya, A. (2011). Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE*, 6(3):e17447.

[Graber and Breznak, 2004] Graber, J. R. and Breznak, J. A. (2004). Physiology and nutrition of *Treponema primitia*, an H2/CO2-acetogenic spirochete from termite hindguts. *Appl. Environ. Microbiol.*, 70(3):1307–1314.

[Grass and Rensing, 2001] Grass, G. and Rensing, C. (2001). Genes involved in copper homeostasis in *Escherichia coli*. *J. Bacteriol.*, 183(6):2145–2147.

[Griffiths-Jones et al., 2005] Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33(Database issue):D121–124.

[Guedon et al., 2000] Guedon, E., Payot, S., Desvaux, M., and Petitdemange, H. (2000). Relationships between cellobiose catabolism, enzyme levels, and metabolic intermediates in *Clostridium cellulolyticum* grown in a synthetic medium. *Biotechnol. Bioeng.*, 67(3):327–335.

[Guerrier-Takada et al., 1983] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–857.

[Haas et al., 2011] Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., Petrosino, J. F., Knight, R., and Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.*, 21(3):494–504.

[Haft et al., 2001] Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., and White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, 29:41–43.

[Hamady et al., 2008] Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, 5:235–237.

[Handelsman et al., 1998] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.*, 5:R245–249.

[Harris et al., 2008] Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science*, 320:106–109.

[Hastie et al., 2003] Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer, New York.

[Hattori, 2008] Hattori, S. (2008). Syntrophic acetate-oxidizing microbes in methanogenic environments. *Microbes Environ.*, 23:118–127.

[Hattori et al., 2005] Hattori, S., Galushko, A. S., Kamagata, Y., and Schink, B. (2005). Operation of the CO dehydrogenase/acetyl coenzyme A pathway in both acetate oxidation and acetate formation by the syntrophically acetate-oxidizing bacterium *Thermacetogenium phaeum*. *J. Bacteriol.*, 187(10):3471–3476.

[He et al., 2010a] He, S., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., Wang, Z., Chen, F., Lindquist, E. A., Sorek, R., and Hugenholtz, P. (2010a). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods*, 7(10):807–812.

[He et al., 2010b] He, Z., Deng, Y., Van Nostrand, J. D., Tu, Q., Xu, M., Hemme, C. L., Li, X., Wu, L., Gentry, T. J., Yin, Y., Liebich, J., Hazen, T. C., and Zhou, J. (2010b). GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *ISME J*, 4(9):1167–1179.

[Healy et al., 1995] Healy, F. G., Ray, R. M., Aldrich, H. C., Wilkie, A. C., Ingram, L. O., and Shanmugam, K. T. (1995). Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Appl. Microbiol. Biotechnol.*, 43:667–674.

[Hershey and Chase, 1952] Hershey, A. D. and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, 36:39–56.

[Hess et al., 2011] Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennacchio, L. A., Tringe, S. G., Visel, A., Woyke, T., Wang, Z., and Rubin, E. M. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331:463–467.

[Hoegger et al., 2006] Hoegger, P. J., Kilaru, S., James, T. Y., Thacker, J. R., and Kües, U. (2006). Phylogenetic comparison and classification of laccase and related multicopper oxidase protein sequences. *FEBS J.*, 273(10):2308–2326.

[Hu et al., 2009] Hu, Y. H., Wang, H. L., Zhang, M., and Sun, L. (2009). Molecular analysis of the copper-responsive CopRSCD of a pathogenic *Pseudomonas fluorescens* strain. *J. Microbiol.*, 47(3):277–286.

[Huber et al., 2004] Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, 20(14):2317–2319.

[Hullo et al., 2001] Hullo, M. F., Moszer, I., Danchin, A., and Martin-Verstraete, I. (2001). CotA of *Bacillus subtilis* is a copper-dependent laccase. *J. Bacteriol.*, 183(18):5426–5430.

[Huse et al., 2007] Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, 8:R143.

[Huse et al., 2010] Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, 12:1889–1898.

[Huson et al., 2007] Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.*, 17:377–386.

[Huson et al., 2011] Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, 21(9):1552–1560.

[Huston et al., 2002] Huston, W. M., Jennings, M. P., and McEwan, A. G. (2002). The multicopper oxidase of *Pseudomonas aeruginosa* is a ferroxidase with a central role in iron acquisition. *Mol. Microbiol.*, 45(6):1741–1750.

[Hyatt et al., 2010] Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119.

[Hyman, 1988] Hyman, E. D. (1988). A new method of sequencing DNA. *Anal. Biochem.*, 174:423–436.

[Imachi et al., 2002] Imachi, H., Sekiguchi, Y., Kamagata, Y., Hanada, S., Ohashi, A., and Harada, H. (2002). *Pelotomaculum thermopropionicum* gen. nov., sp. nov., an anaerobic, thermophilic, syntrophic propionate-oxidizing bacterium. *Int. J. Syst. Evol. Microbiol.*, 52(Pt 5):1729–1735.

[Inskeep et al., 2010] Inskeep, W. P., Rusch, D. B., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Richardson, T. H., Macur, R. E., Hamamura, N., Jennings, R., Fouke, B. W., Reysenbach, A. L., Roberto, F., Young, M., Schwartz, A., Boyd, E. S., Badger, J. H., Mathur, E. J., Ortmann, A. C., Bateson, M., Geesey, G., and Frazier, M. (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS ONE*, 5:e9773.

[Jaenicke et al., 2011] Jaenicke, S., Ander, C., Bekel, T., Bisdorf, R., Dröge, M., Gartemann, K. H., Jünemann, S., Kaiser, O., Krause, L., Tille, F., Zakrzewski, M., Pühler, A., Schlüter, A., and Goesmann, A. (2011). Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. *PLoS ONE*, 6:e14519.

[Jimenez-Juarez et al., 2005] Jimenez-Juarez, N., Roman-Miranda, R., Baeza, A., Sánchez-Amat, A., Vazquez-Duhalt, R., and Valderrama, B. (2005). Alkali and halide-resistant catalysis by the multipotent oxidase from *Marinomonas mediterranea*. *J. Biotechnol.*, 117(1):73–82.

[Johnson, 1977] Johnson, R. C. (1977). The Spirochetes. *Annu. Rev. Microbiol.*, 31:89–106.

[Jukes and Cantor, 1969] Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*. Academy Press.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30.

[Kaur and Sharma, 2006] Kaur, J. and Sharma, R. (2006). Directed evolution: an approach to engineer enzymes. *Crit. Rev. Biotechnol.*, 26:165–199.

[Keiler, 2008] Keiler, K. C. (2008). Biology of trans-translation. *Annu. Rev. Microbiol.*, 62:133–151.

[Kemp et al., 1993] Kemp, P. F., Lee, S., and Laroche, J. (1993). Estimating the growth rate of slowly growing marine bacteria from RNA content. *Appl. Environ. Microbiol.*, 59(8):2594–2601.

[Klocke et al., 2007] Klocke, M., Mähnert, P., Mundt, K., Souidi, K., and Linke, B. (2007). Microbial community analysis of a biogas-producing completely stirred tank reactor fed continuously with fodder beet silage as mono-substrate. *Syst. Appl. Microbiol.*, 30(2):139–151.

[Komori et al., 2009] Komori, H., Miyazaki, K., and Higuchi, Y. (2009). Crystallization and preliminary X-ray diffraction analysis of a putative two-domain-type laccase from a metagenome. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, 65(Pt 3):264–266.

[Koschorreck et al., 2008] Koschorreck, K., Richter, S. M., Ene, A. B., Roduner, E., Schmid, R. D., and Urlacher, V. B. (2008). Cloning and characterization of a new laccase from *Bacillus licheniformis* catalyzing dimerization of phenolic acids. *Appl. Microbiol. Biotechnol.*, 79(2):217–224.

[Krause et al., 2008a] Krause, L., Diaz, N. N., Edwards, R. A., Gartemann, K. H., Krömeke, H., Neuweger, H., Pühler, A., Runte, K. J., Schlüter, A., Stoye, J., Szczepanowski, R., Tauch, A., and Goesmann, A. (2008a). Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. *J. Biotechnol.*, 136(1-2):91–101.

[Krause et al., 2008b] Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A., and Stoye, J. (2008b). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, 36(7):2230–2239.

[Kröber et al., 2009] Kröber, M., Bekel, T., Diaz, N. N., Goesmann, A., Jaenicke, S., Krause, L., Miller, D., Runte, K. J., Viehöver, P., Pühler, A., and Schlüter, A. (2009). Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *J. Biotechnol.*, 142(1):38–49.

[Kunin et al., 2010] Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, 12:118–123.

[Lahr and Katz, 2009] Lahr, D. J. and Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, 47(4):857–866.

[Lawton et al., 2009] Lawton, T. J., Sayavedra-Soto, L. A., Arp, D. J., and Rosenzweig, A. C. (2009). Crystal structure of a two-domain multicopper oxidase: implications for the evolution of multicopper blue proteins. *J. Biol. Chem.*, 284(15):10174–10180.

[Lee and Zinder, 1988] Lee, M. J. and Zinder, S. H. (1988). Isolation and Characterization of a Thermophilic Bacterium Which Oxidizes Acetate in Syntrophic Association with a Methanogen and Which Grows Acetogenically on H(2)-CO(2). *Appl. Environ. Microbiol.*, 54(1):124–129.

[Lee et al., 2007] Lee, Y. J., Romanek, C. S., and Wiegel, J. (2007). *Clostridium aciditolerans* sp. nov., an acid-tolerant spore-forming anaerobic bacterium from constructed wetland sediment. *Int. J. Syst. Evol. Microbiol.*, 57(Pt 2):311–315.

[Lehman et al., 1958] Lehman, I. R., Bessman, M. J., Simms, E. S., and Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. *J. Biol. Chem.*, 233(1):163–170.

[Leininger et al., 2006] Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G. W., Prosser, J. I., Schuster, S. C., and Schleper, C. (2006). *Archaea* predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, 442:806–809.

[Levene et al., 2003] Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299:682–686.

[Levene, 1919] Levene, P. A. (1919). The structure of yeast nucleic acid. *Journal of Biological Chemistry*, 40(2):415–424.

[Ley et al., 2006] Ley, R. E., Peterson, D. A., and Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124:837–848.

[Li et al., 2009] Li, T., Mazéas, L., Sghir, A., Leblon, G., and Bouchez, T. (2009). Insights into networks of functional microbes catalysing methanization of cellulose under mesophilic conditions. *Environ. Microbiol.*, 11(4):889–904.

[Li, 2009] Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, 10:359.

[Lin et al., 2012] Lin, W., Wang, Y., Li, B., and Pan, Y. (2012). A biogeographic distribution of magnetotactic bacteria influenced by salinity. *ISME J*, 6(2):475–479.

[Liu and Conrad, 2011] Liu, F. and Conrad, R. (2011). Chemolithotrophic acetogenic H2/CO2 utilization in Italian rice field soil. *ISME J*, 5(9):1526–1539.

[Lodish, 2004] Lodish, H. (2004). *Molecular Cell Biology*. W.H. Freeman and Company.

[Loessner et al., 2006] Loessner, M., Golden, D., and Jay, J. (2006). Modern food microbiology. *Annals of Microbiology*, 56:81–81.

[Lück and Jager, 1997] Lück, E. and Jager, M. (1997). *Antimicrobial food additives*. Springer, Berlin [u.a.].

[Ludwig et al., 2004] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Kumar, Y., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., and Schleifer, K.-H. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research*, 32(4):1363–1371.

[Luis et al., 2004] Luis, P., Walther, G., Kellner, H., Martin, F., and Buscot, F. (2004). Diversity of laccase genes from basidiomycetes in a forest soil. *Soil Biology and Biochemistry*, 36(7):1025 – 1036.

[Ly et al., 2011] Ly, N. P., Litonjua, A., Gold, D. R., and Celedón, J. C. (2011). Gut microbiota, probiotics, and vitamin D: interrelated exposures influencing allergy, asthma, and obesity? *J. Allergy Clin. Immunol.*, 127:1087–1094.

[Machczynski et al., 2004] Machczynski, M. C., Vijgenboom, E., Samyn, B., and Canters, G. W. (2004). Characterization of SLAC: a small laccase from *Streptomyces coelicolor* with unprecedented activity. *Protein Sci.*, 13(9):2388–2397.

[Maeda et al., 2010] Maeda, K., Hanajima, D., Morioka, R., and Osada, T. (2010). Characterization and spatial distribution of bacterial communities within passively aerated cattle manure composting piles. *Bioresour. Technol.*, 101(24):9631–9637.

[Maidak et al., 2001] Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., and Tiedje, J. M. (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.*, 29(1):173–174.

[Mardis, 2009] Mardis, E. R. (2009). New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med*, 1:40.

[Margulies et al., 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

[Markowitz et al., 2006] Markowitz, V. M., Ivanova, N., Palaniappan, K., Szeto, E., Korzeniewski, F., Lykidis, A., Anderson, I., Mavromatis, K., Kunin, V., Garcia Martin, H., Dubchak, I., Hugenholtz, P., and Kyrpides, N. C. (2006). An experimental metagenome data management and analysis system. *Bioinformatics*, 22(14):e359–367.

[Martins et al., 2002] Martins, L. O., Soares, C. M., Pereira, M. M., Teixeira, M., Costa, T., Jones, G. H., and Henriques, A. O. (2002). Molecular and biochemical characterization of a highly stable bacterial laccase that occurs as a structural component of the Bacillus subtilis endospore coat. *J. Biol. Chem.*, 277(21):18849–18859.

[Matthews et al., 2009] Matthews, H. D., Gillett, N. P., Stott, P. A., and Zickfeld, K. (2009). The proportionality of global warming to cumulative carbon emissions. *Nature*, 459(7248):829–832.

[Mattila et al., 2012] Mattila, H. R., Rios, D., Walker-Sperling, V. E., Roeselers, G., and Newton, I. L. (2012). Characterization of the Active Microbiotas Associated with Honey Bees Reveals Healthier and Broader Communities when Colonies are Genetically Diverse. *PLoS ONE*, 7(3):e32962.

[Mavromatis et al., 2007] Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., and Kyrpides, N. C. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, 4(6):495–500.

[McCaig et al., 1999] McCaig, A. E., Glover, L. A., and Prosser, J. I. (1999). Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl. Environ. Microbiol.*, 65:1721–1730.

[McDonald et al., 2010] McDonald, S., Plant, J., and Worden, A. (2010). The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: a case study of micromonas. *Mol. Biol. Evol.*, 27(10):2268–2283.

[McHardy et al., 2007] McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, 4(1):63–72.

[McHardy and Rigoutsos, 2007] McHardy, A. C. and Rigoutsos, I. (2007). What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.*, 10(5):499–503.

[McInerney et al., 1981] McInerney, M. J., Bryant, M. P., Hespell, R. B., and Costerton, J. W. (1981). *Syntrophomonas wolfei* gen. nov. sp. nov., an Anaerobic, Syntrophic, Fatty Acid-Oxidizing Bacterium. *Appl. Environ. Microbiol.*, 41(4):1029–1039.

[McInerney et al., 2009] McInerney, M. J., Sieber, J. R., and Gunsalus, R. P. (2009). Syntrophy in anaerobic global carbon cycles. *Curr. Opin. Biotechnol.*, 20(6):623–632.

[Menes and Muxí, 2002] Menes, R. J. and Muxí, L. (2002). *Anaerobaculum mobile* sp. nov., a novel anaerobic, moderately thermophilic, peptide-fermenting bacterium that uses crotonate as an electron acceptor, and emended description of the genus *Anaerobaculum*. *Int. J. Syst. Evol. Microbiol.*, 52(Pt 1):157–164.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11:31–46.

[Meyer et al., 2003] Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., and Pühler, A. (2003). GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, 31:2187–2195.

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

[Miranda-Tello et al., 2003] Miranda-Tello, E., Fardeau, M. L., Sepúlveda, J., Fernández, L., Cayol, J. L., Thomas, P., and Ollivier, B. (2003). *Garciella nitratireducens* gen. nov., sp. nov., an anaerobic, thermophilic, nitrate- and thiosulfate-reducing bacterium isolated from an oilfield separator in the Gulf of Mexico. *Int. J. Syst. Evol. Microbiol.*, 53(Pt 5):1509–1514.

[Mitra et al., 2009] Mitra, S., Klar, B., and Huson, D. H. (2009). Visual and statistical comparison of metagenomes. *Bioinformatics*, 25:1849–1855.

[Mitra et al., 2011] Mitra, S., Rupek, P., Richter, D. C., Urich, T., Gilbert, J. A., Meyer, F., Wilke, A., and Huson, D. H. (2011). Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 Suppl 1:S21.

[Miyazaki, 2005] Miyazaki, K. (2005). A hyperthermophilic laccase from *Thermus thermophilus* HB27. *Extremophiles*, 9(6):415–425.

[Muller et al., 2010] Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L. J., and Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, 38(Database issue):D190–195.

[Murugesan, 2003] Murugesan, K. (2003). Bioremediation of paper and pulp mill effluents. *Indian J. Exp. Biol.*, 41(11):1239–1248.

[Myers et al., 2000] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern,

D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000). A whole-genome assembly of *Drosophila*. *Science*, 287:2196–2204.

[Nakamura et al., 2003] Nakamura, K., Kawabata, T., Yura, K., and Go, N. (2003). Novel types of two-domain multi-copper oxidases: possible missing links in the evolution. *FEBS Lett.*, 553(3):239–244.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453.

[Neefs et al., 1991] Neefs, J. M., Van de Peer, Y., De Rijk, P., Goris, A., and De Wachter, R. (1991). Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.*, 19 Suppl:1987–2015.

[Nettmann et al., 2010] Nettmann, E., Bergmann, I., Pramschüfer, S., Mundt, K., Plogsties, V., Herrmann, C., and Klocke, M. (2010). Polyphasic analyses of methanogenic archaeal communities in agricultural biogas plants. *Appl. Environ. Microbiol.*, 76(8):2540–2548.

[Neuweger et al., 2008] Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, 24:2726–2732.

[Nyren and Lundin, 1985] Nyren, P. and Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal. Biochem.*, 151:504–509.

[Ondov et al., 2011] Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12:385.

[Onyenwoke et al., 2007] Onyenwoke, R. U., Kevbrin, V. V., Lysenko, A. M., and Wiegel, J. (2007). *Thermoanaerobacter pseudethanolicus* sp. nov., a thermophilic heterotrophic anaerobe from Yellowstone National Park. *Int. J. Syst. Evol. Microbiol.*, 57(Pt 10):2191–2193.

[Otto et al., 2010] Otto, B., Schlosser, D., and Reisser, W. (2010). First description of a laccase-like enzyme in soil algae. *Arch. Microbiol.*, 192(9):759–768.

[Overbeek et al., 2005] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, 33:5691–5702.

[Pace, 1985] Pace, N. R. (1985). Analyzing natural microbial populations by rRNA sequences. *American Society for Microbiology News*, 51:4–12.

[Palanisami et al., 2010] Palanisami, S., Saha, S., and Lakshmanan, . (2010). Laccase and polyphenol oxidase activities of marine cyanobacteria: a study with Poly R-478 decolourization. *World Journal of Microbiology and Biotechnology*, 26:63–69.

[Partanen et al., 2010] Partanen, P., Hultman, J., Paulin, L., Auvinen, P., and Romantschuk, M. (2010). Bacterial diversity at different stages of the composting process. *BMC Microbiol.*, 10:94.

[Patil et al., 2011] Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T., and McHardy, A. C. (2011). Taxonomic metagenome sequence assignment with structured output models. *Nat. Methods*, 8:191–192.

[Pei et al., 2010] Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., Jin, Z., Lee, P., Yang, L., Poles, M., Brown, S. M., Sotero, S., Desantis, T., Brodie, E., Nelson, K., and Pei, Z. (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl. Environ. Microbiol.*, 76:3886–3897.

[Pertea et al., 2009] Pertea, M., Ayanbule, K., Smedinghoff, M., and Salzberg, S. L. (2009). OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.*, 37(Database issue):D479–482.

[Petersen et al., 2011] Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, 8(10):785–786.

[Pickard et al., 1999] Pickard, M. A., Roman, R., Tinoco, R., and Vazquez-Duhalt, R. (1999). Polycyclic aromatic hydrocarbon metabolism by white rot fungi and oxidation by *Coriolopsis gallica* UAMH 8260 laccase. *Appl. Environ. Microbiol.*, 65(9):3805–3809.

[Pierce et al., 2008] Pierce, E., Xie, G., Barabote, R. D., Saunders, E., Han, C. S., Detter, J. C., Richardson, P., Brettin, T. S., Das, A., Ljungdahl, L. G., and Ragsdale, S. W. (2008). The complete genome sequence of *Moorella thermoacetica* (f. Clostridium thermoaceticum). *Environ. Microbiol.*, 10(10):2550–2573.

[Pop, 2009] Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinformatics*, 10:354–366.

[Pope et al., 2010] Pope, P. B., Denman, S. E., Jones, M., Tringe, S. G., Barry, K., Malfatti, S. A., McHardy, A. C., Cheng, J. F., Hugenholtz, P., McSweeney, C. S., and Morrison, M. (2010). Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc. Natl. Acad. Sci. U.S.A.*, 107:14793–14798.

[Poretsky et al., 2005] Poretsky, R. S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M., Pate, W. M., Moran, M. A., and Hollibaugh, J. T. (2005). Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.*, 71:4121–4126.

[Pruesse et al., 2007] Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, 35:7188–7196.

[Pruitt et al., 2009] Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009). Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37(suppl 1):D32–D36.

[Puerta-Fernandez et al., 2006] Puerta-Fernandez, E., Barrick, J. E., Roth, A., and Breaker, R. R. (2006). Identification of a large noncoding RNA in extremophilic eubacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 103(51):19490–19495.

[Pushkarev et al., 2009] Pushkarev, D., Neff, N. F., and Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.*, 27:847–850.

[Qin et al., 2010] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., Wang, J., Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Minardi, R. M., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., and Zoetendal, E. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65.

[Quince et al., 2009] Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F., and Sloan, W. T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, 6(9):639–641.

[Quince et al., 2011] Quince, C., Lanzén, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, 12:38.

[Rademacher et al., 2012] Rademacher, A., Zakrzewski, M., Schlüter, A., Schonberg, M., Szczepanowski, R., Goesmann, A., Pühler, A., and Klocke, M. (2012). Characterization of microbial biofilms in a thermophilic biogas system by high-throughput metagenome sequencing. *FEMS Microbiol. Ecol.*, 79:785–799.

[Ragsdale, 2008] Ragsdale, S. W. (2008). Enzymology of the wood-Ljungdahl pathway of acetogenesis. *Ann. N. Y. Acad. Sci.*, 1125:129–136.

[Ragsdale and Pierce, 2008] Ragsdale, S. W. and Pierce, E. (2008). Acetogenesis and the Wood-Ljungdahl pathway of CO(2) fixation. *Biochim. Biophys. Acta*, 1784(12):1873–1898.

[Rashamuse et al., 2009] Rashamuse, K., Magomani, V., Ronneburg, T., and Brady, D. (2009). A novel family VIII carboxylesterase derived from a leachate metagenome library exhibits promiscuous beta-lactamase activity on nitrocefin. *Appl. Microbiol. Biotechnol.*, 83:491–500.

[Rastogi et al., 2008] Rastogi, G., Ranade, D. R., Yeole, T. Y., Patole, M. S., and Shouche, Y. S. (2008). Investigation of methanogen population structure in biogas reactor by molecular characterization of methyl-coenzyme M reductase A (mcrA) genes. *Bioresour. Technol.*, 99(13):5317–5326.

[Ray and Apirion, 1979] Ray, B. K. and Apirion, D. (1979). Characterization of 10S RNA: a new stable rna molecule from *Escherichia coli*. *Mol. Gen. Genet.*, 174(1):25–32.

[Reeder and Knight, 2009] Reeder, J. and Knight, R. (2009). The 'rare biosphere': a reality check. *Nat. Methods*, 6(9):636–637.

[Rees et al., 1997] Rees, G. N., Patel, B. K., Grassia, G. S., and Sheehy, A. J. (1997). Anaerobaculum thermoterrenum gen. nov., sp. nov., a novel, thermophilic bacterium which ferments citrate. *Int. J. Syst. Bacteriol.*, 47(1):150–154.

[Rensing and Grass, 2003] Rensing, C. and Grass, G. (2003). *Escherichia coli* mechanisms of copper homeostasis in a changing environment. *FEMS Microbiol. Rev.*, 27(2-3):197–213.

[Riviére et al., 2009] Riviére, D., Desvignes, V., Pelletier, E., Chaussonnerie, S., Guermazi, S., Weissenbach, J., Li, T., Camacho, P., and Sghir, A. (2009). Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *ISME J*, 3(6):700–714.

[Roberts et al., 2003] Roberts, S. A., Wildner, G. F., Grass, G., Weichsel, A., Ambrus, A., Rensing, C., and Montfort, W. R. (2003). A labile regulatory copper ion lies near the T1 copper site in the multicopper oxidase CueO. *J. Biol. Chem.*, 278(34):31958–31963.

[Rodríguez Couto and Toca Herrera, 2006] Rodríguez Couto, S. and Toca Herrera, J. L. (2006). Industrial and biotechnological applications of laccases: a review. *Biotechnol. Adv.*, 24(5):500–513.

[Rogosa, 1971] Rogosa, M. (1971). *Peptococcaceae*, a new family to include the gram-positive, anaerobic cocci of the genera *Peptococcus*, *Peptostreptococcus*, and *Ruminococcus*. *International Journal of Systematic Bacteriology*, 21(3):234–237.

[Rothberg et al., 2011] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T., and Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348–352.

[Rui et al., 2011] Rui, J., Qiu, Q., and Lu, Y. (2011). Syntrophic acetate oxidation under thermophilic methanogenic condition in Chinese paddy field soil. *FEMS Microbiol. Ecol.*, 77(2):264–273.

[Ruijssenaars and Hartmans, 2004] Ruijssenaars, H. J. and Hartmans, S. (2004). A cloned *Bacillus halodurans* multicopper oxidase exhibiting alkaline laccase activity. *Appl. Microbiol. Biotechnol.*, 65(2):177–182.

[Saddler and Khan, 1981] Saddler, J. N. and Khan, A. W. (1981). Cellulolytic enzyme system of *Acetivibrio cellulolyticus*. *Can. J. Microbiol.*, 27(3):288–294.

[Saiki et al., 1988] Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239:487–491.

[Sait et al., 2002] Sait, M., Hugenholtz, P., and Janssen, P. H. (2002). Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ. Microbiol.*, 4:654–666.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

[Sanchez-Amat et al., 2001] Sanchez-Amat, A., Lucas-Elío, P., Fernández, E., García-Borrón, J. C., and Solano, F. (2001). Molecular cloning and functional characterization of a unique multipotent polyphenol oxidase from *Marinomonas mediterranea*. *Biochim. Biophys. Acta*, 1547(1):104–116.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74:5463–5467.

[Sarkanen and H., 1972] Sarkanen, K. V. and H., L. C. (1972). Lignins: Occurrence, formation, structure and reactions. *Journal of Polymer Science Part B: Polymer Letters*, 10(3):228–230.

[Schelder et al., 2011] Schelder, S., Zaade, D., Litsanov, B., Bott, M., and Brocker, M. (2011). The two-component signal transduction system CopRS of *Corynebacterium glutamicum* is required for adaptation to copper-excess stress. *PLoS ONE*, 6(7):e22143.

[Schena et al., 1998] Schena, M., Heller, R. A., Theriault, T. P., Konrad, K., Lachenmeier, E., and Davis, R. W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.*, 16(7):301–306.

[Schloss, 2010] Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.*, 6(7):e1000844.

[Schloss et al., 2011] Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*, 6(12):e27310.

[Schloss and Handelsman, 2005] Schloss, P. D. and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, 71:1501–1506.

[Schloss et al., 2009] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75(23):7537–7541.

[Schlüter et al., 2008] Schlüter, A., Bekel, T., Diaz, N. N., Dondrup, M., Eichenlaub, R., Gartemann, K. H., Krahn, I., Krause, L., Krömeke, H., Kruse, O., Mussgnug, J. H., Neuweger, H., Niehaus, K., Pühler, A., Runte, K. J., Szczepanowski, R., Tauch, A., Tilker, A., Viehöver, P., and Goesmann, A. (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J. Biotechnol.*, 136(1-2):77–90.

[Schmidt et al., 1991] Schmidt, T. M., DeLong, E. F., and Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.*, 173:4371–4378.

[Schnürer et al., 1996] Schnürer, A., Schink, B., and Svensson, B. H. (1996). Clostridium ultunense sp. nov., a mesophilic bacterium oxidizing acetate in syntrophic association with a hydrogenotrophic methanogenic bacterium. *Int. J. Syst. Bacteriol.*, 46(4):1145–1152.

[Schnürer et al., 1997] Schnürer, A., Svensson, B., and Schink, B. (1997). Enzyme activities in and energetics of acetate metabolism by the mesophilic syntrophically acetate-oxidizing anaerobe *Clostridium ultunense*. *FEMS Microbiology Letters*, 154(2):331 – 336.

[Sekiguchi et al., 2006] Sekiguchi, Y., Imachi, H., Susilorukmi, A., Muramatsu, M., Ohashi, A., Harada, H., Hanada, S., and Kamagata, Y. (2006). Tepidanaerobacter syntrophicus gen. nov., sp. nov., an anaerobic, moderately thermophilic, syntrophic alcohol- and lactate-degrading bacterium isolated from thermophilic digested sludges. *Int. J. Syst. Evol. Microbiol.*, 56(Pt 7):1621–1629.

[Sharma et al., 2007] Sharma, P., Goel, R., and Capalash, N. (2007). Bacterial laccases. *World Journal of Microbiology and Biotechnology*, 23:823–832.

[Shendure et al., 2005] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–1732.

[Shi et al., 2009] Shi, Y., Tyson, G. W., and DeLong, E. F. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature*, 459(7244):266–269.

[Sidote et al., 2004] Sidote, D. J., Heideker, J., and Hoffman, D. W. (2004). Crystal structure of archaeal ribonuclease P protein aRpp29 from *Archaeoglobus fulgidus*. *Biochemistry*, 43(44):14128–14138.

[Simon and Daniel, 2011] Simon, C. and Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.*, 77:1153–1161.

[Singh Arora and Kumar Sharma, 2010] Singh Arora, D. and Kumar Sharma, R. (2010). Ligninolytic fungal laccases and their biotechnological applications. *Applied Biochemistry and Biotechnology*, 160:1760–1788.

[Skirnisdottir et al., 2000] Skirnisdottir, S., Hreggvidsson, G. O., Hjörleifsdottir, S., Marteinsson, V. T., Petursdottir, S. K., Holst, O., and Kristjansson, J. K. (2000). Influence of sulfide and temperature on species composition and community structure of hot spring microbial mats. *Appl. Environ. Microbiol.*, 66:2835–2841.

[Sogin et al., 2006] Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.*, 103(32):12115–12120.

[Stajic et al., 2006] Stajic, M., Persky, L., Hadar, Y., Friesem, D., Duletic-Lausevic, S., Wasser, S. P., and Nevo, E. (2006). Effect of copper and manganese ions on activities of laccase and peroxidases in three Pleurotus species grown on agricultural wastes. *Appl. Biochem. Biotechnol.*, 128(1):87–96.

[Steinberg and Regan, 2008] Steinberg, L. M. and Regan, J. M. (2008). Phylogenetic comparison of the methanogenic communities from an acidic, oligotrophic fen and an anaerobic digester treating municipal wastewater sludge. *Appl. Environ. Microbiol.*, 74(21):6663–6671.

[Storz and Haas, 2007] Storz, G. and Haas, D. (2007). A guide to small RNAs in microorganisms. *Current Opinion in Microbiology*, 10(2):93–95.

[Sun et al., 2011] Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E. E., Ellisman, M., Grethe, J., and Wooley, J. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.*, 39(Database issue):D546–551.

[Sun et al., 2009] Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., and Farmerie, W. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.*, 37:e76.

[Suzuki et al., 2003] Suzuki, T., Endo, K., Ito, M., Tsujibo, H., Miyamoto, K., and Inamori, Y. (2003). A thermostable laccase from *Streptomyces lavendulae* REN-7: purification, characterization, nucleotide sequence, and expression. *Biosci. Biotechnol. Biochem.*, 67(10):2167–2175.

[Tamás and Martinoia, 2006] Tamás, M. and Martinoia, E. (2006). *Molecular Biology of Metal Homeostasis And Detoxification: From Microbes to Man*. Topics in Current Genetics. Springer.

[Tamura et al., 2007] Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, 24(8):1596–1599.

[Tatusov et al., 2001] Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, 29:22–28.

[Tringe et al., 2005] Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P., and Rubin, E. M. (2005). Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557.

[Tucker and Breaker, 2005] Tucker, B. J. and Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15(3):342–348.

[Turcatti et al., 2008] Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.*, 36:e25.

[Tyson et al., 2004] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37–43.

[Urich et al., 2008] Urich, T., Lanzén, A., Qi, J., Huson, D. H., Schleper, C., and Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE*, 3(6):e2527.

[Valentine et al., 2012] Valentine, D. L., Mezic, I., Macesic, S., Crnjaric-Zic, N., Ivic, S., Hogan, P. J., Fonoberov, V. A., and Loire, S. (2012). Dynamic autoinoculation and the microbial ecology of a deep water hydrocarbon irruption. *Proc Natl Acad Sci U S A*.

[Varin et al., 2012] Varin, T., Lovejoy, C., Jungblut, A. D., Vincent, W. F., and Corbeil, J. (2012). Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. *Appl. Environ. Microbiol.*, 78:549–559.

[Vartoukian et al., 2007] Vartoukian, S. R., Palmer, R. M., and Wade, W. G. (2007). The division "*Synergistes*". *Anaerobe*, 13(3-4):99–106.

[Velculescu et al., 1995] Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270:484–487.

[Venter et al., 2004] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66–74.

[Vertes et al., 2011] Vertes, D., Qureshi, D., Yukawa, H., and Blaschek, H. (2011). *Biomass to Biofuels: Strategies for Global Industries*. John Wiley & Sons.

[Wagner, 1994] Wagner, R. (1994). The regulation of ribosomal RNA synthesis and bacterial cell growth. *Arch. Microbiol.*, 161(2):100–109.

[Wang et al., 2007] Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73:5261–5267.

[Warnecke and Hess, 2009] Warnecke, F. and Hess, M. (2009). A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J. Biotechnol.*, 142(1):91–95.

[Warnecke et al., 2007] Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S. G., Podar, M., Martin, H. G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N. C., Matson, E. G., Ottesen, E. A., Zhang, X., Hern'andez, M., Murillo, C., Acosta, L. G., Rigoutsos, I., Tamayo, G., Green, B. D., Chang, C., Rubin, E. M., Mathur, E. J., Robertson, D. E., Hugenholtz, P., and Leadbetter, J. R. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450:560–565.

[Wassarman and Storz, 2000] Wassarman, K. M. and Storz, G. (2000). 6S RNA regulates *E. coli* RNA polymerase activity. *Cell*, 101(6):613–623.

[Watson and Crick, 1953a] Watson, J. D. and Crick, F. H. (1953a). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967.

[Watson and Crick, 1953b] Watson, J. D. and Crick, F. H. (1953b). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[Weidner et al., 2003] Weidner, S., Pühler, A., and Küster, H. (2003). Genomics insights into symbiotic nitrogen fixation. *Curr. Opin. Biotechnol.*, 14(2):200–205.

[Weiland, 2010] Weiland, P. (2010). Biogas production: current state and perspectives. *Appl. Microbiol. Biotechnol.*, 85:849–860.

[Weiss et al., 2009] Weiss, A., Jérôme, V., Burghardt, D., Likke, L., Peiffer, S., Hofstetter, E. M., Gabler, R., and Freitag, R. (2009). Investigation of factors influencing biogas production in a large-scale thermophilic municipal biogas plant. *Appl. Microbiol. Biotechnol.*, 84(5):987–1001.

[Weiss et al., 2011] Weiss, S., Zankel, A., Lebuhn, M., Petrak, S., Somitsch, W., and Guebitz, G. M. (2011). Investigation of mircroorganisms colonising activated zeolites during anaerobic biogas production from grass silage. *Bioresour. Technol.*, 102(6):4353–4359.

[Wesenberg et al., 2003] Wesenberg, D., Kyriakides, I., and Agathos, S. N. (2003). White-rot fungi and their enzymes for the treatment of industrial dye effluents. *Biotechnol. Adv.*, 22(1-2):161–187.

[Westerholm et al., 2011] Westerholm, M., Roos, S., and Schnürer, A. (2011). *Tepidanaerobacter acetatoxydans* sp. nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from two ammonium-enriched mesophilic methanogenic processes. *Syst. Appl. Microbiol.*, 34(4):260–266.

[White et al., 2010] White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M. R., Kingsford, C., and Pop, M. (2010). Alignment and clustering of phylogenetic markers–implications for microbial diversity studies. *BMC Bioinformatics*, 11:152.

[Whitman et al., 1998] Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U.S.A.*, 95:6578–6583.

[Williams et al., 2006] Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S., and Griffiths, A. D. (2006). Amplification of complex gene libraries by emulsion PCR. *Nat. Methods*, 3(7):545–550.

[Wilmes and Bond, 2004] Wilmes, P. and Bond, P. L. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.*, 6:911–920.

[Wright et al., 2012] Wright, E. S., Yilmaz, L. S., and Noguera, D. R. (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl. Environ. Microbiol.*, 78(3):717–725.

[Wu and Eisen, 2008] Wu, M. and Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.*, 9(10):R151.

[Xie et al., 2011] Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., Huang, G., Li, Y., Yan, Q., Wu, S., Wang, X., Chen, S., He, G., Xiao, X., and Xu, A. (2011). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J*, 5:414–426.

[Xu et al., 2006] Xu, L., Chen, H., Hu, X., Zhang, R., Zhang, Z., and Luo, Z. W. (2006). Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular Biology and Evolution*, 23(6):1107–1108.

[Yilmaz and Singh, 2011] Yilmaz, S. and Singh, A. K. (2011). Single cell genome sequencing. *Curr Opin Biotechnol*.

[You et al., 2008] You, Y., Fu, C., Zeng, X., Fang, D., Yan, X., Sun, B., Xiao, D., and Zhang, J. (2008). A novel DNA microarray for rapid diagnosis of enteropathogenic bacteria in stool specimens of patients with diarrhea. *J. Microbiol. Methods*, 75:566–571.

[Zakrzewski et al., 2012] Zakrzewski, M., Goesmann, A., Jaenicke, S., Jünemann, S., Eikmeyer, F., Szczepanowski, R., Al-Soud, W. A., Sørensen, S., Pühler, A., and Schlüter, A. (2012). Profiling of the metabolically active community from a production-scale biogas plant by means of high-throughput metatranscriptome sequencing. *J. Biotechnol.*, 158(4):248–258.

[Zhang et al., 2004] Zhang, C., Liu, X., and Dong, X. (2004). *Syntrophomonas curvata* sp. nov., an anaerobe that degrades fatty acids in co-culture with methanogens. *Int. J. Syst. Evol. Microbiol.*, 54(Pt 3):969–973.

[Zhao et al., 2012] Zhao, J., Carmody, L. A., Kalikin, L. M., Li, J., Petrosino, J. F., Schloss, P. D., Young, V. B., and Lipuma, J. J. (2012). Impact of enhanced *Staphylococcus* DNA extraction on microbial community measures in cystic fibrosis sputum. *PLoS ONE*, 7(3):e33127.

[Zhou et al., 2011] Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y. H., Tu, Q., Xie, J., Van Nostrand, J. D., He, Z., and Yang, Y. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J*, 5(8):1303–1313.

[Zinger et al., 2011] Zinger, L., Gobet, A., and Pommier, T. (2011). Two decades of describing the unseen majority of aquatic microbial diversity. *Mol. Ecol.*

[Zwieb and Eichler, 2002] Zwieb, C. and Eichler, J. (2002). Getting on target: the archaeal signal recognition particle. *Archaea*, 1(1):27–34.

APPENDIX A

Appendix: Laccases in microbial genomes and metagenomes

Table A.1: Overview of identified laccases encoded in draft and complete genomes

	laccase tye		location on		
Species	2D	3D	chro-mosome	plas-mid	un-known
Acetobacter pasteurianus IFO 3283-01		1	1		
Achromobacter piechaudii ATCC 43553		2			2
Acidiphilium cryptum JF-5		1	1		
Acidobacterium capsulatum ATCC 51196		2	2		
Acidobacterium sp. MP5ACTX8		6			6
Acidobacterium sp. MP5ACTX9		2			2
Acidovorax avenae subsp. avenae ATCC 19860	1	1			2
Acidovorax avenae subsp. citrulli AAC00-1	1		1		
Acidovorax delafieldii 2AN	2				2
Acidovorax ebreus TPSY	2		2		
Acidovorax sp. JS42	3	1	4		
Acinetobacter baumannii AB900		2			2
Acinetobacter baumannii ACICU		1	1		
Acinetobacter baumannii ATCC 17978		2	2		
Acinetobacter haemolyticus ATCC 19194		2			2
Acinetobacter radioresistens SK82		2			2
Acinetobacter sp. 6013113		1			1
Acinetobacter sp. 6013150		1			1
Acinetobacter sp. 6014059		1			1
Acinetobacter sp. ADP1		1	1		
Acinetobacter sp. ATCC 27244		1			1
Actinobacillus minor 202		1			1
Actinobacillus minor NM305		1			1
Actinomyces viscosus C505		1			1
Afipia sp. 1NLS2	3	2			5
Agrobacterium radiobacter K84	1	1	2		
Agrobacterium tumefaciens str. C58	1		1		

Continued on next page

Agrobacterium vitis S4	1	1	1	1	
Alcanivorax borkumensis SK2		2	2		
Alicycliphilus denitrificans BC		2			2
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446	1	2	3		
Alkalilimnicola ehrlichii MLHE-1	1		1		
Alteromonas macleodii 'Deep ecotype'		1	1		
Amycolatopsis mediterranei U32		1	1		
Anabaena variabilis ATCC 29413	1		1		
Anaeromyxobacter dehalogenans 2CP-1		3	3		
Anaeromyxobacter dehalogenans 2CP-C		2	2		
Anaeromyxobacter sp. Fw109-5		5	5		
Anaeromyxobacter sp. K		2	2		
Aquifex aeolicus VF5		1	1		
Arthrobacter aurescens TC1		4	2	2	
Arthrobacter chlorophenolicus A6		1	1		
Arthrobacter sp. FB24	1	3	1	3	
Asticcacaulis excentricus CB 48		2			2
Aurantimonas manganoxydans SI85-9A1	2	1			3
Azoarcus sp. BH72	1		1		
Azorhizobium caulinodans ORS 571		2	2		
Azospirillum sp. B510		2	1	1	
Azotobacter vinelandii DJ		2	2		
Bacillus amyloliquefaciens FZB42		1	1		
Bacillus cereus AH1272		1			1
Bacillus cereus AH1273		1			1
Bacillus cereus AH603		1			1
Bacillus cereus AH621		1			1
Bacillus cereus Rock3-44		1			1
Bacillus clausii KSM-K16		1	1		
Bacillus coagulans 36D1		2			2
Bacillus licheniformis ATCC 14580		2	2		
Bacillus mycoides Rock1-4		1			1
Bacillus mycoides Rock3-17		1			1
Bacillus pseudomycoides DSM 12442		1			1
Bacillus pumilus ATCC 7061		1			1
Bacillus pumilus SAFR-032		1	1		
Bacillus sp. B14905	1	2			3
Bacillus subtilis subsp. spizizenii ATCC 6633		1			1
Bacillus subtilis subsp. subtilis str. 168		2	1		1
Bacillus subtilis subsp. subtilis str. JH642		1			1
Bacillus subtilis subsp. subtilis str. NCIB 3610		1			1
Bacillus subtilis subsp. subtilis str. SMY		1			1
Bacillus tusciae DSM 2912	1		1		
Beijerinckia indica subsp. indica ATCC 9039	1		1		
Beutenbergia cavernae DSM 12333		1	1		
Blastopirellula marina DSM 3645	1				1
Bordetella bronchiseptica RB50		1	1		
Bordetella parapertussis 12822		1	1		
Bordetella pertussis Tohama I		1	1		
Bordetella petrii DSM 12804		2	2		
Bradyrhizobium japonicum USDA 110	1		1		
Bradyrhizobium sp. BTAi1	3	1	1	3	
Bradyrhizobium sp. ORS278	1		1		
Brevundimonas subvibrioides ATCC 15264		1	1		
Brucella abortus S19		1	1		
Brucella abortus bv. 1 str. 9-941		1	1		
Brucella abortus str. 2308 A		1			1
Brucella canis ATCC 23365		1	1		
Brucella ceti str. Cudo		1			1
Brucella melitensis ATCC 23457		1	1		

Continued on next page

Brucella melitensis biovar Abortus 2308		1	1		
Brucella melitensis bv. 1 str. 16M		1	1		
Brucella melitensis bv. 2 str. 63/9		1			1
Brucella microti CCM 4915		1	1		
Brucella ovis ATCC 25840		1	1		
Brucella suis 1330		1	1		
Brucella suis ATCC 23445		1	1		
Burkholderia ambifaria AMMD	1	1	2		
Burkholderia ambifaria IOP40-10	1				1
Burkholderia ambifaria MC40-6		1	1		
Burkholderia ambifaria MEX-5	1				1
Burkholderia cenocepacia AU 1054	1	1	2		
Burkholderia cenocepacia HI2424		1	1		
Burkholderia cenocepacia J2315		1	1		
Burkholderia cenocepacia MC0-3	1	1	2		
Burkholderia glumae BGR1		1	1		
Burkholderia graminis C4D1M	1				1
Burkholderia mallei ATCC 23344		1	1		
Burkholderia mallei GB8 horse 4		1			1
Burkholderia mallei NCTC 10229		1	1		
Burkholderia mallei NCTC 10247		1	1		
Burkholderia mallei PRL-20		1			1
Burkholderia mallei SAVP1		1	1		
Burkholderia multivorans ATCC 17616	1	4	4	1	
Burkholderia multivorans CGD1		1			1
Burkholderia multivorans CGD2		1			1
Burkholderia multivorans CGD2M		1			1
Burkholderia oklahomensis C6786	1				1
Burkholderia oklahomensis EO147	1	1			2
Burkholderia phymatum STM815	1		1		
Burkholderia phytofirmans PsJN	1		1		
Burkholderia pseudomallei 1106a		1	1		
Burkholderia pseudomallei 112	1				1
Burkholderia pseudomallei 14	1				1
Burkholderia pseudomallei 1710b	1	1	2		
Burkholderia pseudomallei 305	1	1			2
Burkholderia pseudomallei 576	1	1			2
Burkholderia pseudomallei 668		1	1		
Burkholderia pseudomallei 7894	1				1
Burkholderia pseudomallei 9	1				1
Burkholderia pseudomallei 91	1				1
Burkholderia pseudomallei B7210	1	1			2
Burkholderia pseudomallei BCC215	1	1			2
Burkholderia pseudomallei DM98	1				1
Burkholderia pseudomallei K96243	1	1	2		
Burkholderia pseudomallei MSHR346	1	1	1		1
Burkholderia pseudomallei NCTC 13177	1				1
Burkholderia pseudomallei Pakistan 9	1	1			2
Burkholderia sp. 383	1	1	2		
Burkholderia sp. CCGE1001	1				1
Burkholderia sp. CCGE1002	1	1	1	1	
Burkholderia sp. CCGE1003	1				1
Burkholderia sp. Ch1-1	2				2
Burkholderia sp. H160		1			1
Burkholderia thailandensis Bt4	1				1
Burkholderia thailandensis E264	2	2	2		2
Burkholderia thailandensis MSMB43		1			1
Burkholderia thailandensis TXDOH	1	2			3
Burkholderia ubonensis Bu	1	1			2
Burkholderia vietnamiensis G4	1	2	2	1	

Continued on next page

Burkholderia xenovorans LB400	1	1	2	
Campylobacter coli JV20		1		1
Campylobacter coli RM2228		1		1
Campylobacter fetus subsp. fetus 82-40		1	1	
Campylobacter fetus subsp. venerealis str. Azul-94		1		1
Campylobacter jejuni RM1221		1	1	
Campylobacter jejuni subsp. jejuni 1336		1		1
Campylobacter jejuni subsp. jejuni 260.94		1		1
Campylobacter jejuni subsp. jejuni 414		1		1
Campylobacter jejuni subsp. jejuni 81-176		1	1	
Campylobacter jejuni subsp. jejuni 81116		1	1	
Campylobacter jejuni subsp. jejuni 84-25		1		1
Campylobacter jejuni subsp. jejuni CF93-6		1		1
Campylobacter jejuni subsp. jejuni CG8421		1		1
Campylobacter jejuni subsp. jejuni HB93-13		1		1
Campylobacter jejuni subsp. jejuni NCTC 11168		1	1	
Campylobacter lari RM2100		1	1	
Campylobacter upsaliensis RM3195		1		1
Candidatus Koribacter versatilis Ellin345		1	1	
Candidatus Methanosphaerula palustris E1-9c		1	1	
Candidatus Nitrospira defluvii		1	1	
Candidatus Poribacteria sp. WGA-A3	1			1
Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)		1	1	
Candidatus Vesicomyosocius okutanii HA		1	1	
Carnobacterium sp. AT7		1		1
Catenulispora acidiphila DSM 44928		1	1	
Caulobacter crescentus CB15		1	1	
Caulobacter crescentus NA1000		1	1	
Caulobacter segnis ATCC 21756		2	2	
Caulobacter sp. K31	1	2	3	
Cellvibrio japonicus Ueda107	1	1	2	
Chitinophaga pinensis DSM 2588		1	1	
Chloroflexus aurantiacus J-10-fl		1	1	
Chloroflexus sp. Y-400-fl		1	1	
Chromobacterium violaceum ATCC 12472		1	1	
Chryseobacterium gleum ATCC 35910		3		3
Chthoniobacter flavus Ellin428	1			1
Citrobacter koseri ATCC BAA-895		1	1	
Citrobacter rodentium ICC168		1	1	
Citrobacter youngae ATCC 29220		1		1
Citromicrobium bathyomarinum JL354		2		2
Clostridium beijerinckii NCIMB 8052	1		1	
Clostridium botulinum A str. ATCC 19397		1	1	
Clostridium botulinum A str. ATCC 3502		1	1	
Clostridium botulinum A str. Hall		1	1	
Clostridium botulinum D str. 1873		1		1
Clostridium carboxidivorans P7		1		1
Clostridium sporogenes ATCC 15579	1			1
Comamonas testosteroni CNB-2	1	1	2	
Comamonas testosteroni KF-1	1	2		3
Comamonas testosteroni S44	1	1		2
Conexibacter woesei DSM 14684	2	1	3	
Congregibacter litoralis KT71		3		3
Corynebacterium accolens ATCC 49725		1		1
Corynebacterium ammoniagenes DSM 20306		1		1
Corynebacterium aurimucosum ATCC 700975		2	1	1
Corynebacterium diphtheriae NCTC 13129		1	1	
Corynebacterium efficiens YS-314		2	1	1
Corynebacterium genitalium ATCC 33030		2		2
Corynebacterium glucuronolyticum ATCC 51866		3		3

Continued on next page

Corynebacterium glucuronolyticum ATCC 51867		3			3
Corynebacterium glutamicum ATCC 13032		4	4		
Corynebacterium glutamicum R		2	2		
Corynebacterium jeikeium ATCC 43734		1			1
Corynebacterium jeikeium K411		1	1		
Corynebacterium kroppenstedtii DSM 44385		1	1		
Corynebacterium lipophiloflavum DSM 44291		1			1
Corynebacterium matruchotii ATCC 14266		1			1
Corynebacterium matruchotii ATCC 33806		1			1
Corynebacterium pseudogenitalium ATCC 33035		1			1
Corynebacterium pseudotuberculosis FRC41		1	1		
Corynebacterium striatum ATCC 6940		1			1
Corynebacterium tuberculostearicum SK141		1			1
Coxiella burnetii Dugway 5J108-111		1	1		
Croceibacter atlanticus HTCC2559		1	1		
Cronobacter sakazakii ATCC BAA-894		2	2		
Cronobacter turicensis		1	1		
Cronobacter turicensis z3032		1		1	
Cupriavidus metallidurans CH34		2		2	
Cupriavidus taiwanensis		1	1		
Cyanothece sp. ATCC 51142	1		1		
Cyanothece sp. CCY0110	1	1			2
Cyanothece sp. PCC 7424	1	2	3		
Cyanothece sp. PCC 7425	1		1		
Cyanothece sp. PCC 7822	1	2			3
Cyanothece sp. PCC 8801	1	2	3		
Cyanothece sp. PCC 8802	1	2	3		
Cylindrospermopsis raciborskii CS-505	1				1
Deferribacter desulfuricans SSM1		1	1		
Deinococcus deserti VCD115	1		1		
Deinococcus geothermalis DSM 11300	1			1	
Delftia acidovorans SPH-1	1	2	3		
Desulfococcus oleovorans Hxd3		1	1		
Desulfohalobium retbaense DSM 5692		1	1		
Desulfomicrobium baculatum DSM 4028	1		1		
Desulfovibrio sp. FW1012B	1				1
Desulfovibrio vulgaris str. 'Miyazaki F'		1	1		
Desulfuromonas acetoxidans DSM 684		1			1
Dickeya dadantii Ech586		1	1		
Dickeya dadantii Ech703		1	1		
Dickeya zeae Ech1591		1	1		
Dinoroseobacter shibae DFL 12		5	1	4	
Enhydrobacter aerosaccus SK60		1			1
Enterobacter cancerogenus ATCC 35316		1			1
Enterobacter cloacae subsp. cloacae ATCC 13047		3	2	1	
Enterobacter sp. 638		1	1		
Enterococcus faecalis TUSoD Ef11		1			1
Enterococcus faecalis TX1322		1			1
Enterococcus faecium E1071		1			1
Enterococcus faecium PC4.1		1			1
Enterococcus faecium TX1330		1			1
Erwinia amylovora ATCC 49946		1	1		
Erwinia amylovora CFBP1430		1	1		
Erwinia billingiae Eb661		1	1		
Erwinia pyrifoliae Ep1/96		1	1		
Erwinia tasmaniensis Et1/99		1	1		
Erythrobacter litoralis HTCC2594		3	3		
Erythrobacter sp. NAP1		3			3
Erythrobacter sp. SD-21		3			3
Escherichia albertii TW07627		2			2

Continued on next page

Escherichia coli 101-1	1			1
Escherichia coli 536	1	1		
Escherichia coli 53638	1			1
Escherichia coli 55989	2	2		
Escherichia coli 83972	1			1
Escherichia coli APEC O1	2	1	1	
Escherichia coli ATCC 8739	2	2		
Escherichia coli B str. REL606	1	1		
Escherichia coli B171	1			1
Escherichia coli B7A	1			1
Escherichia coli BL21-Gold(DE3)pLysS AG	1	1		
Escherichia coli BW2952	1	1		
Escherichia coli CFT073	1	1		
Escherichia coli E110019	1			1
Escherichia coli E22	1			1
Escherichia coli E24377A	1	1		
Escherichia coli ED1a	1	1		
Escherichia coli F11	1			1
Escherichia coli HS	1	1		
Escherichia coli IAI1	1	1		
Escherichia coli IAI39	1	1		
Escherichia coli MS 107-1	1			1
Escherichia coli MS 115-1	2			2
Escherichia coli MS 116-1	1			1
Escherichia coli MS 119-7	1			1
Escherichia coli MS 124-1	1			1
Escherichia coli MS 146-1	1			1
Escherichia coli MS 175-1	1			1
Escherichia coli MS 182-1	1			1
Escherichia coli MS 185-1	1			1
Escherichia coli MS 187-1	1			1
Escherichia coli MS 196-1	1			1
Escherichia coli MS 198-1	1			1
Escherichia coli MS 200-1	1			1
Escherichia coli MS 21-1	1			1
Escherichia coli MS 69-1	1			1
Escherichia coli MS 78-1	1			1
Escherichia coli MS 84-1	1			1
Escherichia coli O103 H2 str. 12009	1	1		
Escherichia coli O111 H- str. 11128	1	1		
Escherichia coli O127 H6 str. E2348/69	1	1		
Escherichia coli O157 H7 EDL933	1	1		
Escherichia coli O157 H7 str. EC4024	1			1
Escherichia coli O157 H7 str. EC4042	1			1
Escherichia coli O157 H7 str. EC4045	1			1
Escherichia coli O157 H7 str. EC4076	1			1
Escherichia coli O157 H7 str. EC4113	1			1
Escherichia coli O157 H7 str. EC4115	1	1		
Escherichia coli O157 H7 str. EC4196	1			1
Escherichia coli O157 H7 str. EC4206	1			1
Escherichia coli O157 H7 str. EC4401	1			1
Escherichia coli O157 H7 str. EC4486	1			1
Escherichia coli O157 H7 str. EC4501	1			1
Escherichia coli O157 H7 str. EC508	1			1
Escherichia coli O157 H7 str. EC869	1			1
Escherichia coli O157 H7 str. FRIK2000	1			1
Escherichia coli O157 H7 str. FRIK966	1			1
Escherichia coli O157 H7 str. Sakai	1	1		
Escherichia coli O157 H7 str. TW14359	1	1		
Escherichia coli O157 H7 str. TW14588	1			1

Continued on next page

Escherichia coli O26 H11 str. 11368		1	1		
Escherichia coli O55 H7 str. CB9615		1	1		
Escherichia coli S88		1	1		
Escherichia coli SE11		1	1		
Escherichia coli SMS-3-5		1	1		
Escherichia coli UMN026		1	1		
Escherichia coli UTI89		1	1		
Escherichia coli str. K-12 substr. DH10B		1	1		
Escherichia coli str. K-12 substr. MG1655		1	1		
Escherichia coli str. K-12 substr. W3110		1	1		
Escherichia fergusonii ATCC 35469		1	1		
Escherichia sp. 4_1_40B					
Exiguobacterium sp. AT1b		1	1		
Flavobacterium johnsoniae UW101		1	1		
Francisella philomiragia subsp. philomiragia ATCC 25017		1	1		
Frankia alni ACN14a		1	1		
Frankia sp. CcI3	1		1		
Frankia sp. EAN1pec		1	1		
Frankia sp. EUN1f		2			2
Fulvimarina pelagi HTCC2506	2	5			7
Gemmatimonas aurantiaca T-27	1		1		
Geobacillus kaustophilus HTA426		1	1		
Geobacillus sp. C56-T3	1	1	2		
Geobacillus sp. G11MC16		1			1
Geobacillus sp. Y412MC52	1				1
Geobacillus sp. Y412MC61	1		1		
Geobacillus thermodenitrificans NG80-2		2	2		
Geobacter bemidjiensis Bem		3	3		
Geobacter lovleyi SZ		2	2		
Geobacter metallireducens GS-15		5	5		
Geobacter sp. FRC-32		3	3		
Geobacter sp. M18		1			1
Geobacter sp. M21		2	2		
Geobacter sulfurreducens PCA		2	2		
Geobacter uraniireducens Rf4		1	1		
Geodermatophilus obscurus DSM 43160		1	1		
Gloeobacter violaceus PCC 7421	1		1		
Gluconacetobacter diazotrophicus PAl 5		2	2		
Gluconacetobacter hansenii ATCC 23769		1			1
Gluconobacter oxydans 621H		1	1		
Gordonia bronchialis DSM 43247		1	1		
Gramella forsetii KT0803		2	2		
Granulibacter bethesdensis CGDNIH1		1	1		
Granulicatella adiacens ATCC 49175		1			1
Haemophilus somnus 2336		2	2		
Hahella chejuensis KCTC 2396		1	1		
Haliangium ochraceum DSM 14365		1	1		
Haloferax volcanii DS2		1		1	
Halorubrum lacusprofundi ATCC 49239		2	2		
Haloterrigena turkmenica DSM 5511	1	2		3	
Halothiobacillus neapolitanus c2		2	2		
Herbaspirillum seropedicae SmR1	1		1		
Herminiimonas arsenicoxydans	2	2	4		
Herpetosiphon aurantiacus ATCC 23779	1		1		
Hirschia baltica ATCC 49814		1	1		
Hydrogenivirga sp. 128-5-R1-1		1			1
Hyphomicrobium denitrificans ATCC 51888		1	1		
Hyphomonas neptunium ATCC 15444		2	2		
Idiomarina baltica OS145		2			2
Idiomarina loihiensis L2TR		1	1		

Continued on next page

Janibacter sp. HTCC2649		2			2
Janthinobacterium sp. Marseille		2	2		
Kangiella koreensis DSM 16069		1	1		
Klebsiella pneumoniae 342		1	1		
Klebsiella pneumoniae NTUH-K2044		2	1	1	
Klebsiella pneumoniae subsp. pneumoniae MGH 78578		2	1	1	
Klebsiella pneumoniae subsp. rhinoscleromatis ATCC 13884		1			1
Klebsiella variicola At-22		1	1		
Kribbella flavida DSM 17836		2	2		
Kytococcus sedentarius DSM 20547		1	1		
Lactobacillus brevis ATCC 367		1	1		
Lactobacillus brevis subsp. gravesensis ATCC 27305		1			1
Lactobacillus buchneri ATCC 11577		1			1
Lactobacillus casei ATCC 334		1	1		
Lactobacillus casei BL23		1	1		
Lactobacillus casei str. Zhang		1	1		
Lactobacillus crispatus 214-1		1			1
Lactobacillus crispatus JV-V01		1			1
Lactobacillus crispatus ST1		1	1		
Lactobacillus delbrueckii subsp. bulgaricus PB2003/044-T3-4		1			1
Lactobacillus fermentum ATCC 14931		1			1
Lactobacillus fermentum IFO 3956		1	1		
Lactobacillus hilgardii ATCC 8290		1			1
Lactobacillus jensenii 208-1		2			2
Lactobacillus jensenii 269-3		1			1
Lactobacillus jensenii JV-V16		1			1
Lactobacillus paracasei subsp. paracasei ATCC 25302		1			1
Lactobacillus plantarum JDM1		1	1		
Lactobacillus plantarum WCFS1		1	1		
Lactobacillus plantarum subsp. plantarum ATCC 14917		1			1
Lactobacillus rhamnosus GG		1	1		
Lactobacillus rhamnosus HN001		1			1
Lactobacillus rhamnosus LMS2-1		1			1
Lactobacillus rhamnosus Lc 705		1	1		
Lactobacillus vaginalis ATCC 49540		1			1
Lactococcus lactis subsp. cremoris SK11		1		1	
Laribacter hongkongensis HLHK9		2	2		
Leeuwenhoekiella blandensis MED217		1			1
Legionella drancourtii LLAP12		2			2
Legionella longbeachae NSW150		2	2		
Legionella pneumophila 2300/99 Alcoy		1	1		
Legionella pneumophila str. Corby		1	1		
Legionella pneumophila str. Lens		1	1		
Legionella pneumophila str. Paris		1	1		
Legionella pneumophila subsp. pneumophila str. Philadelphia 1		1	1		
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Ames)'	1		1		
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Paris)'	1		1		
Leptospira borgpetersenii serovar Hardjo-bovis JB197	1		1		
Leptospira borgpetersenii serovar Hardjo-bovis L550	1		1		
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	1		1		
Leptospira interrogans serovar Lai str. 56601	1		1		
Leptothrix cholodnii SP-6	1	1	2		
Leuconostoc citreum KM20		1		1	
Leuconostoc mesenteroides subsp. cremoris ATCC 19254		1			1
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293		1		1	
Limnobacter sp. MED105	2	1			3
Lutiella nitroferrum 2002		1			1
Lyngbya sp. PCC 8106	1	1			2
Lysinibacillus fusiformis ZC1		1			1
Lysinibacillus sphaericus C3-41	1	2	3		

Continued on next page

Magnetospirillum magnetotacticum MS-1	1				1
Maricaulis maris MCS10	1	1	2		
Marinobacter algicola DG893		3			3
Marinobacter aquaeolei VT8		1	1		
Marinobacter sp. ELB17		2			2
Marinomonas sp. MED121		1			1
Meiothermus ruber DSM 1279		1	1		
Meiothermus silvanus DSM 9946		1	1		
Mesorhizobium loti MAFF303099		2	2		
Mesorhizobium opportunistum WSM2075		3			3
Mesorhizobium sp. BNC1	1	2	1	2	
Methylacidiphilum infernorum V4	1	1	2		
Methylibium petroleiphilum PM1	2		2		
Methylobacillus flagellatus KT	2		2		
Methylobacterium chloromethanicum CM4	1		1		
Methylobacterium extorquens AM1	2	2	2	2	
Methylobacterium extorquens DM4	1	1	2		
Methylobacterium nodulans ORS 2060	1	1		2	
Methylobacterium populi BJ001	1		1		
Methylobacterium sp. 4-46		1	1		
Methylocella silvestris BL2		2	2		
Methylococcus capsulatus str. Bath	1	1	2		
Methylotenera sp. 301	1		1		
Methylovorus sp. SIP3-4	1		1		
Micrococcus luteus NCTC 2665		2	1		1
Micrococcus luteus SK58		1			1
Micromonospora sp. L5		1			1
Moritella sp. PE36		1			1
Mycobacterium abscessus ATCC 19977		2	2		
Mycobacterium avium 104		1	1		
Mycobacterium avium subsp. avium ATCC 25291		2			2
Mycobacterium avium subsp. paratuberculosis K-10		1	1		
Mycobacterium bovis AF2122/97		1	1		
Mycobacterium bovis BCG str. Pasteur 1173P2		1	1		
Mycobacterium bovis BCG str. Tokyo 172		1	1		
Mycobacterium gilvum PYR-GCK		1		1	
Mycobacterium intracellulare ATCC 13950		2			2
Mycobacterium kansasii ATCC 12478		2			2
Mycobacterium marinum M		2	2		
Mycobacterium parascrofulaceum ATCC BAA-614		5			5
Mycobacterium sp. JLS		3	3		
Mycobacterium sp. KMS		1		1	
Mycobacterium tuberculosis '98-R604 INH-RIF-EM'		1			1
Mycobacterium tuberculosis 210		1			1
Mycobacterium tuberculosis CDC1551		1	1		
Mycobacterium tuberculosis F11		1	1		
Mycobacterium tuberculosis H37Ra		2	1		1
Mycobacterium tuberculosis H37Rv		1	1		
Mycobacterium tuberculosis KZN 1435		1	1		
Mycobacterium tuberculosis KZN 4207		2			2
Mycobacterium tuberculosis KZN R506		1			1
Mycobacterium tuberculosis KZN V2475		1			1
Mycobacterium ulcerans Agy99		1	1		
Mycobacterium vanbaalenii PYR-1		1	1		
Myxococcus xanthus DK 1622	2		2		
Nakamurella multipartita DSM 44233		1	1		
Nitratiruptor sp. SB155-2		1	1		
Nitrobacter hamburgensis X14	3	1	2	2	
Nitrobacter sp. Nb-311A	1				1
Nitrobacter winogradskyi Nb-255	1	1	2		

Continued on next page

Nitrosococcus halophilus Nc4	3	2	5		
Nitrosococcus oceani ATCC 19707	2		2		
Nitrosococcus watsoni C-113	2	1	3		
Nitrosomonas europaea ATCC 19718	1	1	2		
Nitrosomonas eutropha C91	2	3	5		
Nitrosomonas sp. AL212		4			4
Nitrosopumilus maritimus SCM1	3	1	4		
Nitrosospira multiformis ATCC 25196		1	1		
Nocardia farcinica IFM 10152	1	2		3	
Nocardioides sp. JS614		2	2		
Nodularia spumigena CCY9414	1	1			2
Nostoc azollae 0708	1		1		
Nostoc sp. PCC 7120	1		1		
Novosphingobium aromaticivorans DSM 12444		2	2		
Oceanibulbus indolifex HEL-45	2	1			3
Oceanicaulis alexandrii HTCC2633		1			1
Oceanicola batsensis HTCC2597		4			4
Oceanicola granulosus HTCC2516		1			1
Oceanobacillus iheyensis HTE831		1	1		
Ochrobactrum anthropi ATCC 49188		2	2		
Oenococcus oeni ATCC BAA-1163		1			1
Oenococcus oeni AWRIB429		1			1
Oenococcus oeni PSU-1		1	1		
Oligotropha carboxidovorans OM5	3	1	4		
Opitutus terrae PB90-1		1	1		
Oscillatoria sp. PCC 6506	1				1
Paenibacillus curdlanolyticus YK9	1				1
Pantoea ananatis LMG 20103		1	1		
Pantoea sp. At-9b		1			1
Pantoea sp. aB		1			1
Parachlamydia acanthamoebae str. Halls coccus	1				1
Parvibaculum lavamentivorans DS-1		1	1		
Parvularcula bermudensis HTCC2503		2	2		
Pasteurella multocida subsp. multocida str. Pm70		1	1		
Pectobacterium wasabiae WPP163		1	1		
Pediococcus acidilactici DSM 20284		1			1
Pediococcus pentosaceus ATCC 25745		1	1		
Pedobacter sp. BAL39		1			1
Persephonella marina EX-H1		1	1		
Phaeobacter gallaeciensis 2.10		2			2
Phaeobacter gallaeciensis BS107		1			1
Phenylobacterium zucineum HLK1		4	2	2	
Photobacterium profundum 3TCK		1			1
Photobacterium profundum SS9		1	1		
Photorhabdus asymbiotica		1	1		
Photorhabdus luminescens subsp. laumondii TTO1		1	1		
Planctomyces maris DSM 8797	1				1
Plesiocystis pacifica SIR-1		1			1
Polaromonas naphthalenivorans CJ2	2	1	1	2	
Polaromonas sp. JS666	1	1	2		
Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1		1	1		
Proteus mirabilis ATCC 29906		1			1
Proteus mirabilis HI4320		1	1		
Providencia alcalifaciens DSM 30120		1			1
Providencia rettgeri DSM 1131		1			1
Providencia rustigianii DSM 4541		1			1
Providencia stuartii ATCC 25827		1			1
Pseudoalteromonas haloplanktis TAC125		1	1		
Pseudomonas aeruginosa LESB58		1	1		
Pseudomonas aeruginosa PA7		1	1		

Continued on next page

Pseudomonas aeruginosa PACS2		2			2
Pseudomonas aeruginosa PAO1		1	1		
Pseudomonas aeruginosa PAb1		2			2
Pseudomonas aeruginosa UCBPP-PA14		1	1		
Pseudomonas entomophila L48		2	2		
Pseudomonas fluorescens Pf-5		2	2		
Pseudomonas fluorescens Pf0-1		2	2		
Pseudomonas fluorescens SBW25		1	1		
Pseudomonas mendocina ymp	1	2	3		
Pseudomonas putida F1		2	2		
Pseudomonas putida GB-1		2	2		
Pseudomonas putida KT2440		1	1		
Pseudomonas putida W619		3	3		
Pseudomonas stutzeri A1501		3	3		
Pseudomonas syringae pv. aesculi str. 2250		1			1
Pseudomonas syringae pv. aesculi str. NCPPB3681		1			1
Pseudomonas syringae pv. oryzae str. 1_6					
Pseudomonas syringae pv. phaseolicola 1448A		1	1		
Pseudomonas syringae pv. syringae 642		1			1
Pseudomonas syringae pv. syringae B728a		3	3		
Pseudomonas syringae pv. syringae FF5		1			1
Pseudomonas syringae pv. tabaci ATCC 11528		2			2
Pseudomonas syringae pv. tomato K40		3			3
Pseudomonas syringae pv. tomato Max13		3			3
Pseudomonas syringae pv. tomato NCPPB 1108		3			3
Pseudomonas syringae pv. tomato T1		3			3
Pseudomonas syringae pv. tomato str. DC3000		2	2		
Psychrobacter arcticus 273-4		1	1		
Psychrobacter cryohalolentis K5		1	1		
Psychrobacter sp. PRwf-1		1	1		
Pyrobaculum aerophilum str. IM2		1	1		
Ralstonia eutropha H16		1	1		
Ralstonia eutropha JMP134		1	1		
Ralstonia pickettii 12D		2		2	
Ralstonia pickettii 12J	1	4	5		
Ralstonia solanacearum GMI1000		2		2	
Ralstonia solanacearum PSI07		2	2		
Ralstonia solanacearum UW551	1	1			2
Raphidiopsis brookii D9	1				1
Rhizobium etli Brasil 5	1				1
Rhizobium etli CFN 42	1	2	2	1	
Rhizobium etli CIAT 652	1	2	2	1	
Rhizobium leguminosarum bv. trifolii WSM1325		1	1		
Rhizobium leguminosarum bv. trifolii WSM2304	1	1	1	1	
Rhizobium leguminosarum bv. viciae 3841	1	2	3		
Rhizobium sp. NGR234	3	2	1	4	
Rhodobacter capsulatus SB 1003		1	1		
Rhodobacterales bacterium HTCC2150		1			1
Rhodobacterales bacterium HTCC2654	1	3			4
Rhodococcus equi ATCC 33707		3			3
Rhodococcus erythropolis PR4	1	6	6	1	
Rhodococcus erythropolis SK121	1	6			7
Rhodococcus jostii RHA1		4	3	1	
Rhodococcus opacus B4	1	4	3	2	
Rhodopseudomonas palustris BisB5	1		1		
Rhodopseudomonas palustris TIE-1	2		2		
Rhodothermus marinus DSM 4252		2	2		
Rickettsiella grylli		1			1
Robiginitalea biformata HTCC2501		1	1		
Roseobacter denitrificans OCh 114		1	1		

Continued on next page

Roseobacter litoralis Och 149	1	2			3
Roseobacter sp. AzwK-3b		1			1
Roseobacter sp. MED193		2			2
Roseomonas cervicalis ATCC 49957		1			1
Roseovarius nubinhibens ISM	2	1			3
Roseovarius sp. 217	1	2			3
Roseovarius sp. HTCC2601		5			5
Roseovarius sp. TM1035	1				1
Rubrobacter xylanophilus DSM 9941	1	1	2		
Ruegeria pomeroyi DSS-3		5	3	2	
Ruegeria sp. TM1040		1	1		
Saccharophagus degradans 2-40		1	1		
Saccharopolyspora erythraea NRRL 2338		2	1		1
Salinispora arenicola CNS-205		2	2		
Salinispora tropica CNB-440		2	2		
Salmonella enterica subsp. arizonae serovar 62 z4,z23		1	1		
S. enterica subsp. enterica serovar 4,[5],12 - str. CVM23701					
Salmonella enterica subsp. enterica serovar Agona str. SL483		1	1		
S. enterica subsp. enterica serovar Choleraesuis str. SC-B67					
S. enterica subsp. enterica serovar Dublin str. CT_02021853					
S. enterica subsp. enterica serovar Enteritidis str. P125109					
S. enterica subsp. enterica serovar Gallinarum str. 287/91					
S. enterica subsp. enterica serovar Hadar str. RI_05P066					
S. enterica subsp. enterica serovar Heidelberg str. SL476					
S. enterica subsp. enterica serovar Heidelberg str. SL486					
S. enterica subsp. enterica serovar Javiana str. GA_MM04042433					
S. enterica subsp. enterica serovar Kentucky str. CDC 191					
S. enterica subsp. enterica serovar Kentucky str. CVM29188					
S. enterica subsp. enterica serovar Newport str. SL254					
S. enterica subsp. enterica serovar Newport str. SL317					
S. enterica subsp. enterica serovar Paratyphi A str. AKU_12601					
S. enterica subsp. enterica serovar Paratyphi A str. ATCC 9150					
S. enterica subsp. enterica serovar Paratyphi B str. SPB7					
S. enterica subsp. enterica serovar Paratyphi C strain RKS4594					
S. enterica subsp. enterica serovar Saintpaul str. SARA23					
S. enterica subsp. enterica serovar Saintpaul str. SARA29					
S. enterica subsp. enterica serovar Schwarzengrund str. CVM19633					
S. enterica subsp. enterica serovar Schwarzengrund str. SL480					
S. enterica subsp. enterica serovar Tennessee str. CDC07-0191					
Salmonella enterica subsp. enterica serovar Typhi str. CT18		1	1		
S. enterica subsp. enterica serovar Typhi str. E02-1180					
S. enterica subsp. enterica serovar Typhi str. E98-3139					
Salmonella enterica subsp. enterica serovar Typhi str. Ty2		1	1		
S. enterica subsp. enterica serovar Typhimurium str. LT2					
S. enterica subsp. enterica serovar Virchow str. SL491					
S. enterica subsp. enterica serovar Weltevreden str. HI_N05-537					
Serratia odorifera 4Rx13		1			1
Serratia odorifera DSM 4582		1			1
Serratia proteamaculans 568		1	1		
Shewanella benthica KT99		1			1
Shewanella denitrificans OS217		1	1		
Shewanella frigidimarina NCIMB 400		1	1		
Shewanella loihica PV-4		1	1		
Shewanella putrefaciens CN-32		1	1		
Shewanella sp. ANA-3		1		1	
Shewanella woodyi ATCC 51908	1		1		
Shigella boydii CDC 3083-94		1	1		
Shigella boydii Sb227		1	1		
Shigella dysenteriae 1012		1			1
Shigella dysenteriae Sd197		1	1		

Continued on next page

Shigella flexneri 2a str. 2457T		1	1		
Shigella flexneri 2a str. 301		1	1		
Shigella flexneri 5 str. 8401		1	1		
Shigella sonnei Ss046		1	1		
Shigella sp. D9		1			1
Sinorhizobium medicae WSM419	1	3	1	3	
Sinorhizobium meliloti 1021	2	1	2	1	
Solibacter usitatus Ellin6076		2	2		
Sorangium cellulosum 'So ce 56'	1	7	8		
Sphaerobacter thermophilus DSM 20745	2	1	3		
Sphingobacterium spiritivorum ATCC 33300		2			2
Sphingobacterium spiritivorum ATCC 33861		2			2
Sphingobium japonicum UT26S		3	3		
Sphingomonas sp. SKA58		5			5
Sphingomonas wittichii RW1	1	1	2		
Sphingopyxis alaskensis RB2256		3	3		
Spirosoma linguale DSM 74		2	1	1	
Stackebrandtia nassauensis DSM 44728		3	3		
Stappia aggregata IAM 12614		1			1
Starkeya novella DSM 506		2	2		
Stenotrophomonas maltophilia K279a		2	2		
Stenotrophomonas maltophilia R551-3		1	1		
Stigmatella aurantiaca DW4/3-1	1	1			2
Streptomyces avermitilis MA-4680		1	1		
Streptomyces flavogriseus ATCC 33331		1			1
Streptomyces griseus subsp. griseus NBRC 13350		2	2		
Streptomyces roseosporus NRRL 11379		2			2
Streptomyces scabiei 87.22		1	1		
Streptomyces sp. ACT-1		2			2
Streptomyces sp. ACTE		3			3
Streptosporangium roseum DSM 43021		5	5		
Sulfitobacter sp. EE-36	1				1
Sulfitobacter sp. NAS-14.1	1	7			8
Sulfurihydrogenibium sp. YO3AOP1		1	1		
Sulfurovum sp. NBC37-1		3	3		
Synechococcus sp. CC9311		1	1		
Synechococcus sp. PCC 7002	1		1		
Synechococcus sp. RCC307		1	1		
Synechococcus sp. RS9917		1			1
Synechococcus sp. WH 5701	1				1
Synechococcus sp. WH 7803		1	1		
Syntrophobacter fumaroxidans MPOB		1	1		
Thauera sp. MZ1T	1		1		
Thermincola sp. JR		1	1		
Thermobaculum terrenum ATCC BAA-798	1	1	2		
Thermobispora bispora DSM 43833		2	2		
Thermocrinis albus DSM 14484		1	1		
Thermomicrobium roseum DSM 5159		1		1	
Thermosediminibacter oceani DSM 16646		1	1		
Thermus thermophilus HB27		1	1		
Thioalkalivibrio sp. HL-EbGR7	1		1		
Thioalkalivibrio sp. K90mix	1	1	2		
Thiobacillus denitrificans ATCC 25259		2	2		
Thiomicrospira crunogena XCL-2		4	4		
Thiomonas intermedia K12	1	1	2		
Tolumonas auensis DSM 9187		1	1		
Trichodesmium erythraeum IMS101	1		1		
Tsukamurella paurometabola DSM 20162		1	1		
Turicibacter sp. PC909		1			1
Variovorax paradoxus S110	1		1		

Continued on next page

Veillonella parvula ATCC 17745		1			1
Verrucomicrobium spinosum DSM 4136		1			1
Vibrio furnissii CIP 102972		1			1
Vibrio splendidus LGP32		1	1		
Weissella paramesenteroides ATCC 33313		1			1
Xanthobacter autotrophicus Py2	3	7	7	3	
Xanthomonas axonopodis pv. citri str. 306		1	1		
Xanthomonas campestris pv. campestris str. 8004		1	1		
Xanthomonas campestris pv. campestris str. ATCC 33913		1	1		
Xanthomonas campestris pv. campestris str. B100		1	1		
Xanthomonas campestris pv. musacearum NCPPB4381		1			1
Xanthomonas campestris pv. vasculorum NCPPB702		1			1
Xanthomonas campestris pv. vesicatoria str. 85-10		1	1		
Xanthomonas fuscans subsp. aurantifolii str. ICPB 10535		1			1
Xanthomonas fuscans subsp. aurantifolii str. ICPB 11122		1			1
Xanthomonas oryzae pv. oryzae KACC10331		1	1		
Xanthomonas oryzae pv. oryzae MAFF 311018		1	1		
Xanthomonas oryzae pv. oryzae PXO99A		1	1		
Xylella fastidiosa 9a5c		1	1		
Xylella fastidiosa Ann-1		2			2
Xylella fastidiosa Dixon		2			2
Xylella fastidiosa M12		1	1		
Xylella fastidiosa M23		1	1		
Xylella fastidiosa Temecula1		1	1		
Yersinia aldovae ATCC 35236		1			1
Yersinia bercovieri ATCC 43970		1			1
Yersinia enterocolitica subsp. enterocolitica 8081		1	1		
Yersinia frederiksenii ATCC 33641		1			1
Yersinia intermedia ATCC 29909		1			1
Yersinia kristensenii ATCC 33638		1			1
Yersinia mollaretii ATCC 43969		2			2
Yersinia pestis Angola		1	1		
Yersinia pestis Antiqua		1	1		
Yersinia pestis CA88-4125		1			1
Yersinia pestis CO92		1	1		
Yersinia pestis FV-1		1			1
Yersinia pestis KIM 10		1	1		
Yersinia pestis KIM D27		1			1
Yersinia pestis Nepal516		2	1		1
Yersinia pestis Pestoides A		1			1
Yersinia pestis Pestoides F		1	1		
Yersinia pestis Z176003		1	1		
Yersinia pestis biovar Antiqua str. B42003004		1			1
Yersinia pestis biovar Mediaevalis str. K1973002		1			1
Yersinia pestis biovar Microtus str. 91001		1	1		
Yersinia pestis biovar Orientalis str. F1991016		1			1
Yersinia pestis biovar Orientalis str. IP275		1			1
Yersinia pestis biovar Orientalis str. India 195		1			1
Yersinia pestis biovar Orientalis str. MG05-1020		1			1
Yersinia pestis biovar Orientalis str. PEXU2		1			1
Yersinia pseudotuberculosis IP 31758		1	1		
Yersinia pseudotuberculosis IP 32953		1	1		
Yersinia pseudotuberculosis PB1/+		1	1		
Yersinia pseudotuberculosis YPIII		1	1		
Yersinia rohdei ATCC 43380		1			1
Yersinia ruckeri ATCC 29473		1			1
Zunongwangia profunda SM-A87		3	3		
alpha proteobacterium BAL199	1	4			5
bacterium Ellin514	1				1
marine actinobacterium		1			1

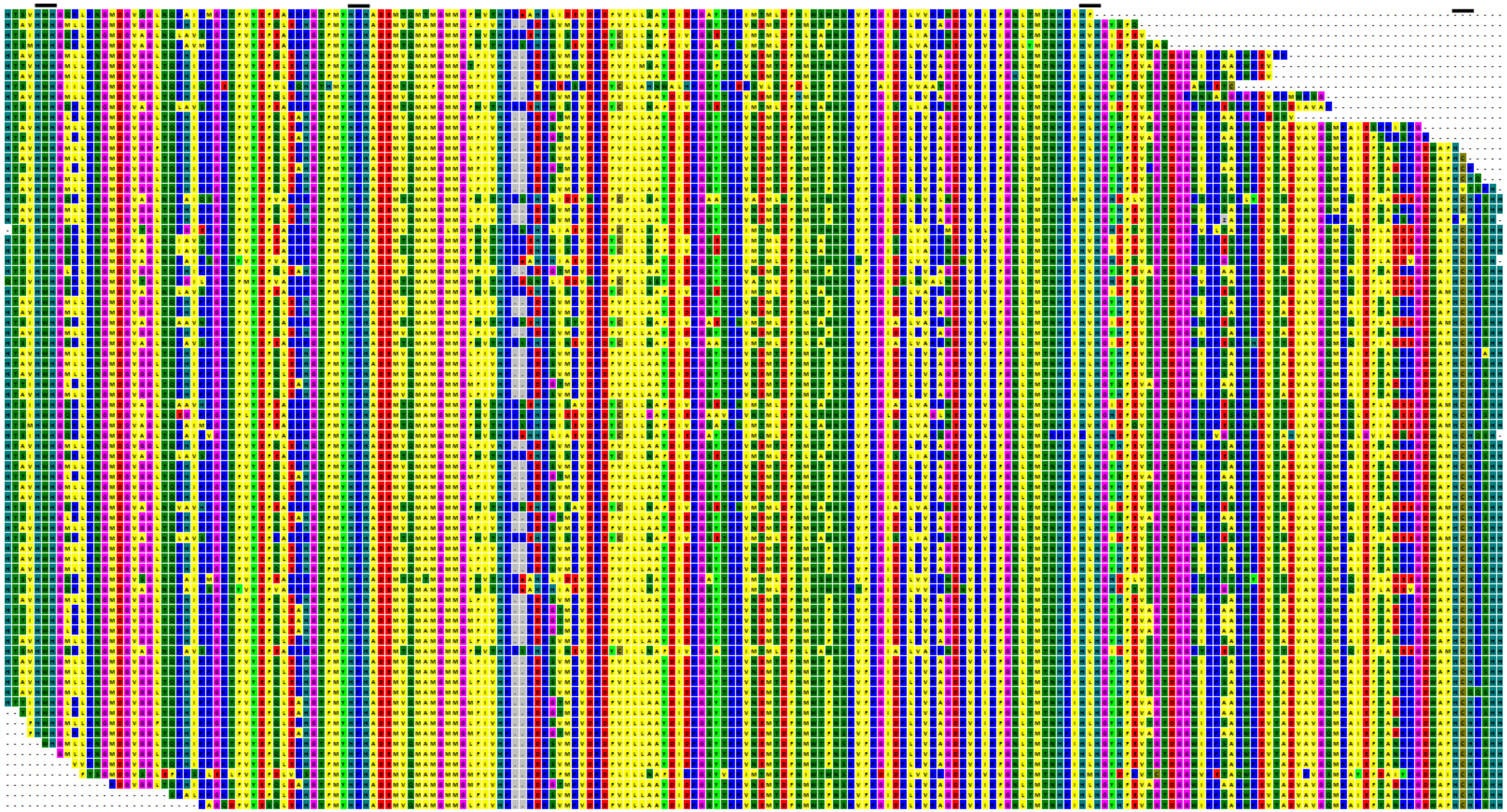


Figure A.1: Alignment of translated reads encoding type-B two-domain laccases: Translated reads obtained from a marine metagenome were compared to the profile HMM representing the type-B two-domain laccase enzyme using the HMMER3 [Eddy, 2011] software. The identified reads carrying genes for putative laccases were aligned to the same profile HMM. The alignment was manually adjusted. Only a subset of the identified translated reads are shown in the alignments. The black bars above the alignment indicate the four cbr. The alignment presents clearly that the complete region between cbr1 and cbr4 was covered by the translated metagenome reads.

Nomenclature

AJAX	asynchronous JavaScript and XML
API	application programming interface
BLAST	basic local alignment search tool
bp	base pair
cbr	copper binding region
CD	coding sequence
cDNA	copy DNA
CDS	coding sequence
COG	Clusters of Orthologous Groups
DNA	deoxyribose nucleic acid
EC	Enzyme Commission
EGT	environmental gene tag
EST	expressed sequence tag
GS	Genome Sequencer
HMM	hidden Markov model

KEGG	Enzyme Commission
LSU	large subunit
mcr	methyl-coenzyme M reductase
mRNA	messenger RNA
NBC	Naïve Bayesian Classifier
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NGS	next generation sequencing
nr	non-redundant NCBI protein sequence database
OLE	ornate large extremophilic RNA
OTU	operational taxonomic unit
PCR	polymerase chain reaction
PFAM	protein family
RDP	Ribosomal Database Project
RNA	ribonucleic acid
RNase P	Ribonuclease P
rRNA	ribosomal RNA
SAMS	Sequence Analysis and Management System
SLP	single-linkage preclustering
SSU	small subunit
tmRNA	transfer-messenger RNA
tRNA	transfer RNA

Acknowledgements

First, I would like to thank Prof. Dr. Jens Stoye and Prof. Dr. Alfred Pühler for their support and helpful advices throughout my PhD project. I am also very grateful to Andreas Schlüter and Alexander Goesmann for their truly interest for my PhD project and their invaluable comments.

I want to acknowledge the funding obtained from the "CLIB Graduate Cluster", which supported me to develop my scientific career by attending retreats and soft skill courses.

This thesis would not have been possible without the help from my colleagues and friends from the Computational Genomics Group at Bielefeld University, in particular the support from Thomas Bekel, Oliver Rupp, Sebastian Jaenicke and Sebastian Jünemann. Moreover, I wish to thank my colleagues for corrections of this work, and especially I wish to thank Dr. Erin Kennerly Parker for proof-reading the manuscripts and her friendship in spite of the distance. Besides, I would like to thank Antje Rademacher, Felix Eikmeyer, Luka Ausec and Regina Bisdorf for the pleasant and fruitful collaborations.

I am very grateful to my parents for their encourage throughout my studies and my sister for getting me through the good and the bad times. Finally, I thank all my friend who supported me during the last three years that are not mentioned here.

Erklärung

Ich, Martha Zakrzewski, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel benutzt habe.

Bielefeld, 20. Juni 2012

. . . .
Martha Zakrzewski