An integrated metagenome and -proteome analysis of the microbial community residing in a biogas production plant

Vera Ortseifen^{1,2}, Yvonne Stolze¹, Irena Maus¹, Alexander Sczyrba³, Andreas Bremges³, Stefan P. Albaum⁴, Sebastian Jaenicke⁵, Jochen Fracowiak⁶, Alfred Pühler¹ and Andreas Schlüter^{1,*}

¹Genome Research of Industrial Microorganisms, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany

²Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany

³Computational Metagenomics group, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany

⁴Bioinformatics Resource Facility, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany

⁵Bioinformatics and Systems Biology, Justus Liebig University, 35392 Giessen, Germany

⁶Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Winchester Strasse, D-35394, Giessen, Germany

*Corresponding author: Andreas Schlüter

Address: Center for Biotechnology, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany

Phone: +49-521-106-8757

Fax: +49-521-106-89041

Email: aschluet@CeBiTec.Uni-Bielefeld.de

25 **Keywords**

26

27 *biogas microbial community, integrated metagenome-/proteome study, database impact on*

28 *protein identification, contig context information, taxonomic profile,*

29

Abstract

To study the metaproteome of a biogas-producing microbial community, fermentation samples were taken from an agricultural biogas plant for microbial cell and protein extraction and corresponding metagenome analyses. Based on metagenome sequence data, taxonomic community profiling was performed to elucidate the composition of bacterial and archaeal sub-communities. The community's cytosolic metaproteome was represented in a 2D-PAGE approach. Metaproteome databases for protein identification were compiled based on the assembled metagenome sequence dataset for the biogas plant analyzed and non-corresponding biogas metagenomes. Protein identification results revealed that the corresponding biogas protein database facilitated the highest identification rate followed by other biogas-specific databases, whereas common public databases yielded insufficient identification rates. Proteins of the biogas microbiome identified as highly abundant were assigned to the pathways involved in methanogenesis, transport and carbon metabolism. Moreover, the integrated metagenome/-proteome approach enabled the examination of genetic-context information for genes encoding identified proteins by studying neighbored genes on the corresponding contig. Exemplarily, this approach led to the identification of a *Methanoculleus* sp. contig encoding 16 methanogenesis-related gene products, three of which were also detected as abundant proteins within the community's metaproteome. Thus, metagenome contigs provide additional information on the genetic environment of identified abundant proteins.

Introduction

Significant progress has been made regarding the analysis of complex microbial communities. This has been achieved by recent developments in metagenome research such as new high-throughput sequencing techniques and advanced bioinformatics tools for data interpretation (Tanca *et al.*, 2013). Metagenome sequence data provide information about the genetic repertoire of all microorganisms in an environment in a culture-independent approach (Kolmeder and de Vos, 2014). However, the question concerning gene expression leading to distinct metabolic activities cannot be resolved by metagenome analyses (Heyer *et al.*, 2013). Therefore, the goal of metatranscriptome and -proteome analyses is to link genetic information to function. Metatranscriptomics based on next generation sequencing arose as a powerful method that could provide detailed insight into gene transcription reflecting microbial activity in a mixed community (Alberti *et al.*, 2014). There are additional processes interfering at the post-transcriptional level that can be assessed by methods addressing the metaproteome of microbial communities (Hettich *et al.*, 2013). Moreover, the characterization of metaproteomic datasets complements the knowledge of metabolic activities and the linkage to microbial communities and their ecological functions (Wang *et al.*, 2014). The metaproteome approach has been applied for diverse environments, e.g. soil and sediments, the human intestinal tract and oral cavity as well as bioengineered systems. One major challenge already appeared in one of the first metaproteome studies. Schulze *et al.* (2005) analyzed the proteins isolated from dissolved organic matter of four different environments. Only few proteins could be identified due to the lack of corresponding metagenome data (Schulze *et al.*, 2005, Heyer *et al.*, 2015). Even analyzing protein expression in natural and bioengineered systems such as a laboratory-scale activated sludge system resulted in poor protein identification rates because of missing genomic information (Wilmes and Bond, 2004). As a result, different studies focused on integrated approaches using corresponding metagenome datasets (Lauro *et al.*, 2011; Wilmes *et al.*, 2008). An integrative metagenomics/-proteomics

approach has recently been applied to compare healthy and disease conditions of patients suffering from Crohn's disease regarding the intestinal-tract microbiome (Erickson *et al.*, 2012). Likewise, also non-annotated, unassembled metagenome data could be used as basis for the compilation of a specific database for metaproteome analyses (Rooijers *et al.*, 2011). Moreover, a recently published study followed up on the question whether the application of different protein databases for protein identification impacts metaproteome analyses (Tanca *et al.*, 2013). Hereby, not only appropriate protein databases deduced from corresponding metagenome datasets were generated, but also public databases namely NCBI, SwissProt and TrEMBL parsed at different taxonomic levels were used and compared. Tanca *et al.* (2013) could confirm a considerable impact of database selection, also for the trustworthiness of the metaproteome results.

High-throughput omics-technologies help to characterize biogas communities and relationships between community members in detail. Metagenome studies were carried out for biogas-producing microbial communities to uncover their taxonomic profiles and functional potential. First, the community composition was determined by construction of 16S-rRNA gene clone libraries and subsequent sequencing of corresponding amplicons (Huang *et al.*, 2002; Klocke *et al.*, 2007; McHugh *et al.*, 2003; Mladenovska *et al.*, 2003). Moreover, the development of next generation sequencing technologies offers the possibility of whole metagenome shotgun analyses (Stolze *et al.*, 2015; Jaenicke *et al.*, 2011b; Kröber *et al.*, 2009; Schlüter *et al.*, 2008). Based on the metagenome sequence information analyses like functional assignments and taxonomic profiling could be performed. Complementary, high-throughput metatranscriptome sequencing was used for profiling of the metabolically active community from a production-scale biogas plant (Zakrzewski *et al.*, 2012). Obtained results among others indicated a high transcriptional activity of archaeal species. Also, metaproteome

101 analyses of microbial communities involved in biogas-producing fermentations were
102 performed (Hanreich *et al.*, 2013, 2012; Heyer *et al.*, 2013).

103

104 Here, an integrative metagenome/-proteome study was carried out, in which changes in a
105 maize-degrading microbial community in batch fermentation were investigated. Deeply
106 sequenced metagenome datasets of biogas communities from production-scale biogas plants
107 were exploited for the compilation of biogas-specific protein databases to improve the
108 identification of proteins in a metaproteomics approach for one of the biogas plants analyzed.
109 Protein identification results were compared to those obtained by applying public protein
110 databases or heterologous biogas protein databases. Highly expressed proteins within the
111 microbial biogas community were interpreted in the context of pathways important for the
112 biogas production process. Due to the availability of deeply sequenced metagenome data for
113 the biogas plant analyzed, assembled contigs provided context information for genes encoding
114 identified highly expressed proteins. Integrative metagenome/-proteome analysis is discussed
115 as a preferable approach for functional studies addressing complex microbial communities.

Material and Methods

Sampling of a production-scale biogas plant and DNA extraction

At four production-scale biogas plant sites a sample of the fermenter digestate was taken directly from the main fermenter of the plants, respectively. Biogas plant 1, 2 and 3 operate under mesophilic conditions, whereas Biogas plant 4 (BGP4) runs thermophilic at 54°C. Of those four biogas plants BGP 1 and BGP 3 were already described in the literature (Stolze *et al.*, 2015). Moreover, fed substrates differ between those four plants. BGP 1 uses 45% maize silage, 22% sugar beet and 33% poultry manure. In BGP 2 50% maize silage, 10% grass and 40% pig/ cattle manure was used. BGP 3 uses only two fed substrates, namely 67% maize silage and 33% pig manure. BGP 4 operate with 60% maize silage, 30% grass and 10% pig manure as substrates. Metagenome sequence datasets of the four biogas plants are available in the Short Read Archive (SRA).

BGP1: <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357211> and <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357213>,
BGP2: <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357208> and <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357209>,
BGP3: <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357214> and <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357221>,
BGP4: <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357222> and <http://www.ncbi.nlm.nih.gov/sra/?term=SRA357223>.

Before sampling at the sampling tap of the plants, approximately 25 liters of the fermenter digestate were discarded. One liter was then filled into a gastight vessel, respectively, excess air was let out and the bottle hermetically sealed with a screw cap. Within 2 hours, at the most, the bottles were then transferred to the laboratory in a closed styrofoam box to keep the sample material warm. At the laboratory a total amount of approximately 100 g of the fresh

sample were used for DNA extraction in quadruplicates following the protocol described in Schlüter *et al.* (2008), applying minor changes to enable larger sample volumes. Two times 50 ml of the sample were filled into a screw cap tube and frozen at -20°C for subsequent extraction of cytosolic proteins.

Preparation and high-throughput sequencing of metagenome libraries for biogas-producing microbial communities

For preparation and high-throughput sequencing two replicates were used. A total of 100 ng metagenomic DNA was sheared to 270 bp fragments using a focused-ultrasonicator (Covaris Inc.). The sheared DNA fragments were purified and size selected using SPRI beads (Beckman Coulter Inc.). Obtained fragments were blunt-end-repaired, phosphorylated and A-tailed. Subsequently, T-tailed adapters, containing sequences used during cluster formation and Illumina compatible adapters (IDT, Inc.), were ligated to the purified DNA fragments applying the KAPA-Illumina library creation kit (KAPA Biosystems Inc.). The prepared sequencing libraries were quantified by applying the KAPA Biosystem's next-generation sequencing library qPCR kit (KAPA Biosystems Inc.) and run on a Roche LightCycler 480 real-time PCR instrument (Roche Inc.). Sequencing of the quantified sample libraries was performed on the Illumina HiSeq 2000 sequencer using the Illumina TruSeq SBS v3-HS kit, following a 2 x 150 indexed high-output run protocol.

Metagenome assembly and quality control

For each BGP, we first combined all sequencing reads of the two replicates prior to assembly. To further analyze the effect of sequencing depth – and assembly quality –, we randomly subsampled 12.5%, 25%, and 50%, respectively, of BGP3's sequencing reads with seqtk (Heng Li, <https://github.com/lh3/seqtk>). We then used Ray Meta (Boisvert *et al.*, 2012) to assemble the individual metagenomes, with an empirically selected k-mer size of 41 and a

minimal contig size of 1,000 bp. The assemblies were submitted to European Nucleotide Archive (ENA).

Gene prediction and functional annotation

We predicted genes with MetaProdigal (Hyatt *et al.*, 2012), as described previously (Bremges *et al.*, 2015). We used DIAMOND (Buchfink *et al.*, 2015), to perform BLASTP-like searches against NCBI's non-redundant protein sequences (nr), assigning a potential function (best hit) and against KEGG genes (Kanehisa *et al.*, 2016) for pathway analyses. Phylogenetic assignments for each gene were done using MEGAN 5 (Huson *et al.*, 2011).

Estimating metagenomic gene abundance in mesophilic BGP3

For BGP3 only, we aligned the raw sequencing reads to our assembly with bowtie2 (Langmead and Salzberg, 2012) using end-to-end and sensitive mode, and then used samtools (Li *et al.*, 2009) to convert SAM to BAM and sort the alignment file. We estimated each gene's abundance in the metagenome with HTSeq (Anders *et al.*, 2015) using parameters -r pos -s no -t CDS -i ID -m intersection-nonempty -a 0.

Taxonomic profiling of the biogas microbial community

Taxonomic profiling of the biogas community based on metagenome sequence reads was performed by applying the Lowest Common Ancestor (LCA) pipeline as implemented in the metagenome analysis platform MGX. In more detail, a Blast search versus the NCBI nonredundant ('nr') database was conducted with an E-value cut off of $1e^{-5}$ and a sequence identity threshold of at least 80%. The taxonomic assignment for each sequence was derived as the lowest common ancestor of all Blast hits. Moreover, the taxonomy results were visualized as a Krona chart (Ondov *et al.*, 2011), which illustrates the quantification of classified reads in a hierarchical order.

193

194 ***Extraction of cytosolic proteins of a sample from a production-scale biogas plant***

195 Two samples of cytosolic proteins of biogas plant 3 (BGP3) were extracted and identified as
196 described by Wendler *et al.* (2012). In short, lyophilized cell pellets from a biogas
197 fermentation sample were mechanically disrupted with a bead mill. Afterwards cell debris
198 were removed by centrifugation, proteins were extracted using phenol and precipitated with
199 methanol. The cell pellet was washed with 70% (v/v) ethanol as well as acetone.

200

201 ***Analysis of the cytosolic proteome of a biogas fermentation sample by 2D PAGE***

202 In general, cytosolic proteins were separated in two dimension as recently published with
203 some adjustments regarding protein amount and isoelectric focusing (Wendler *et al.*, 2012). In
204 short, purified proteins were rehydrated and quantified with a modified Bradford assay
205 (Bradford, 1976). For isoelectric focusing (IEF) 350 µg of protein was applied on two
206 ImmobilineTM DryStrip (GE Healthcare) with different immobilized non-linear pH gradient.
207 One IPG strip was custom made like described in Wendler *et al.* (2012) with an effective pH
208 range of 4.3-8.3. For the second separation of the same sample a commercially available dry
209 polyacrylamide gel strip with an immobilized non-linear pH gradient 3-10 of 18 cm length
210 was used. For the separation in the second dimension 10.5% Tris-tricine polyacrylamide gels
211 and Protean II IXL Cell (Bio-Rad Laboratories) were used. After fixation, the proteins within
212 the gels were stained using Colloidal Coomassie (Neuhoff *et al.*, 1988).

213

214 ***Identification of proteins by peptide-mass-fingerprint and MS² analyses***

215 Like mentioned above, the gels only differ in slightly different pH gradient. For both gels all
216 distinct abundant protein spots were selected, which resulted in a total of 181 gel slices for the
217 MS measurement (100 gel 1 and 81 gel 2). Hereby, also redundant spots for both gels are

possible. Proteins were in-gel digested with sequencing grade modified trypsin (Promega), dissolved within acetonitrile and trifluoroacetic acid and stored at -20°C before analyses. Protein identifications were performed using the basic protocol of Hansmeier *et al.* (2006), where the digested proteins were measured with ultrafleXtreme (Bruker Daltonics) mass spectrometer.

Set-up of different databases for identification of proteins from the biogas metaproteome

For identification by comparison of peptide mass fingerprinting (PMF) and tandem MS (MS/MS) to the database, the BioTools (Bruker Daltonics) software and the Mascot algorithm were used. Here, various databases were generated. First, metagenome data of all assembled biogas plants were used to construct separate databases. Furthermore, all the data of the four plants were combined into one large database. Also, the metagenome results of the corresponding biogas plant was halved, quartered and eighths to investigate the effect of a lower sequencing depth. Moreover, all protein entries of TrEMBL and NCBI as well as taxonomy restricted version to *Bacteria* and *Archaea* according to Tanca *et al.* (2013) were used for the protein identification.

For the comparison following parameters were set: enzyme: trypsin; missed cleavages: 1; modifications: carbamidomethyl (C) and dynamic modifications: oxidation (M); peptide tolerance: ± 100 ppm; mass values: MH^+ and monoisotopic. For MS^2 mode a mass tolerance of 0.7 Da was accepted.

Results & Discussion

Characterization of an industrial scale biogas plant

To analyze the metaproteome of a biogas-producing microbial community a fermentation sample was directly taken from the main fermenter of the production-scale biogas plant 3 (BGP3) on April, the 17th, 2013. The sampled biogas plant 3 operates under wet fermentation conditions (Stolze *et al.*, 2015, Weiland *et al.*, 2006), characterized by a high liquid and a relatively low total solid percentage. Apart from maize silage, also pig manure was fed as substrates into the system, split into 24 feedings per day with a theoretical retention time of 55 days. The plant consists of a main fermenter (heated to 40°C), a secondary fermenter and a non-heated storage tank. Process parameters collected around the date of sampling showed average physico-chemical parameter values, with expected pH-values and buffering capacity (Table 1). Measured values were stable for weeks before and after the sampling date (data not shown), confirming stability of the biogas process.

Taxonomic classification of the indigenous biogas microbial community in BGP3 based on corresponding metagenome sequence data was accomplished applying the Lowest Common Ancestor (LCA) algorithm implemented in the metagenome analysis platform MGX. The obtained taxonomic profile representing combined results of two independent biological replicates was visualized as a Krona chart (Figure 1). Of 9,986,623 reads 4,138,570 were unassigned. Taxonomic assignments of 5,848,053 sequence reads, about 60% of all reads, are shown in a hierarchical order. Starting from the inner circle, taxonomic classification resulted in two superkingdoms, with 91% (of all metagenome reads) for *Bacteria* and 9% for *Archaea*. The most abundant phylum in BGP3 within the superkingdom *Bacteria* is *Firmicutes* (31%) followed by *Bacterioidetes* (7%). The abundance of further phyla such as *Cloacimonetes* (3%), *Spirochaetes* (3%) and *Proteobacteria* (2%) is in any case below 10%. Going down in the hierarchical order, the most abundant classes within the phylum *Firmicutes* are the *Clostridia* (13%), followed by *Bacilli* (2%). The most abundant orders identified within the

corresponding phylum are the *Clostridiales* (10%), *Lactobacillales* (1%) and *Bacilliales* (0.8%). The predominance of presumptive hydrolytic- and acidogenic bacteria has also been found in other studies, which investigated the microbial composition of biogas-producing communities applying high-throughput 16S rRNA gene amplicon or metagenome sequencing (Stolze *et al.*, 2015, Schlüter *et al.*, 2008, Kröber *et al.*, 2009; Krause *et al.*, 2008). The second most frequent phylum within the domain *Bacteria*, the *Bacterioidetes*, is represented by the classes *Bacteroidia* (3%), *Flavobacteriia* (0.2%) and *Cytophagia* (0.1%). Among the *Bacterioidetes*, the family *Bacteroidales* (3%) occurred most frequently. Members of this family cover a highly diverse spectrum of bacteria that are also assumed to be involved in the degradation of polymers, for example polysaccharides (Jaenicke *et al.*, 2011a).

The superkingdom *Archaea* (9%) is dominated by the phylum *Euryarchaeota* (8%). The class *Methanomicrobia* (8%) dominates the methanogenic community, followed by *Methanobacteria* (0.1%) and *Methanococci* (0.09%). Within the class *Methanomicrobia*, the order *Methanomicrobiales* (7%) featured a preponderance of the methanogens residing in the analyzed reactor, with *Methanoculleus* (6%) being the dominant genus.

In summary, the taxonomic classification of the analyzed microbial community in biogas plant 3 revealed bacterial and archaeal representatives, which can be related to the four metabolic phases of the biogas formation process. All of the identified phyla were previously described to be present in other biogas fermentation reactors, for example in systems that were fed with maize silage supplemented with animal manure (Schlüter *et al.*, 2008; Zakrzewski *et al.*, 2012; Stolze *et al.*, 2015).

Metaproteome analysis of the biogas microbial community by 2D-SDS-PAGE

To obtain insights into the expressed proteins of the microbial community residing in the production-scale biogas plant, a metaproteomics approach was conducted. Corresponding

metagenome sequence data available for the biogas community were exploited to facilitate identification of proteins.

The biogas plant studied in this survey was sampled for protein identification. Since the metaproteome of the biogas community was expected to be very complex, a fractionation was applied. In this approach, the cytosolic fraction of proteins was selected for further analysis. Proteins were extracted from lyophilized cell pellets by mechanical disruption using a bead beating step and subsequent phenol-extraction. Cytosolic proteins were separated in a gel-based approach in two dimensions.

The protein samples were separated using two different immobilized pH gradients (IPG). In detail, the first IPG strip was custom-made as described by Wendler *et al.* (2012) with an effective pH-range of around 4.3 to 8.3, in the following referred to as gel 1. The second strip provides an immobilized non-linear pH gradient of 3 to 10. In the second dimension in which the cytosolic proteins were separated by molecular weight, the procedure with both IPG-strips was identical. After gel-electrophoresis and fixation, the proteins within the gels were stained using Colloidal Coomassie (Neuhoff *et al.*, 1988). Afterwards, all distinct, abundant protein spots were cut out from both gels, which resulted in 181 samples. Since both strips comprise the pH-range of 4.3 up to 8.3, some of the picked spots were represented on both gels.

Gel 1 of the two prepared and stained gels is presented in Figure 2. As expected, most of the cytosolic proteins appear with an isoelectric point in the pH-range of 4 to 6. Surprisingly, not only clouds of stained proteins are visible but also distinct, abundant protein spots. Moreover, also series of spots emerge at the same molecular weight but with slightly different isoelectric points. Comparison of both gels revealed some redundant spots in both profiles. All picked samples from both gels were measured by applying mass spectrometry. For identification, proteins digested with trypsin were subjected to peptide-mass-fingerprint and tandem mass spectrometry (MS²) analyses. One important requirement for a successful protein-identification is the availability of corresponding gene-sequence information to deduce

amino-acid sequences of encoded gene products followed by construction of an appropriate protein-database.

Compilation and definition of protein databases applied for protein identification from peptide mass spectra

A key step in metaproteomics is the identification of proteins, which usually is achieved by comparing peptide mass fingerprints and/or tandem mass spectrometry results to those deduced for reference proteins archived in databases. In first instance, publicly available protein databases such as SWISS-PROT, TrEMBL or UniProt have been used for this task. However, habitat-specific protein databases can be compiled by exploiting appropriate metagenome sequence data of the microbial community analyzed. This approach aims at inferring protein sequences from assembled metagenome sequence data for compilation of a specific protein database. Using deeply sequenced metagenome datasets obtained for four production-scale biogas plants, the latter approach was followed in this study. BGP 1, 2 and 3 operate under mesophilic conditions, whereas BGP4 runs thermophilic at 54°C. Moreover, fed substrates differ between those four plants (detailed process parameters for BGP1, 2 and 4 are given in a Table in Supplementary File 1).

Metagenome sequence data were assembled by the Ray-Meta assembler and protein sequences were deduced for predicted genes to be extracted and stored in biogas-specific protein databases for each individual biogas plant and a combined repository representing all biogas plants. For comparison, all protein entries of TrEMBL and NCBI as well as a taxonomy-restricted NCBI-version (only *Bacteria* and *Archaea*) were used for compilation of additional four protein datasets. Accordingly, a total of twelve different databases were compared including the compiled protein dataset deduced from metagenome sequence data of the biogas plant 3 that was sampled for the metaproteomics approach performed in this study. Table 2 summarizes the characteristics of all constructed databases for identification of

protein spots. BGP1 yielded the largest biogas database comprising 675,231 genes/proteins followed by BGP3. The NCBI database consisted of 51,656,746 protein sequences (October 2014) and the TrEMBL database of 56,010,222 protein sequences (May 2014).

The identification of proteins within a metaproteome applying biogas-specific protein databases

As mentioned above, for coherence of protein identification results in metaproteome studies, availability of a suitable protein-database is crucial. Initially, the question was followed, which protein database is best suited for identification of proteins from the biogas microbial community analyzed. Therefore, all compiled databases were used independently for identification of 181 selected protein-spots originating from both 2D-PAGEs prepared. Obtained results were compared to evaluate suitability of the biogas-plant-specific datasets. Table 3 summarizes the number of identified protein spots as well as identification rates from MS and MS² spectra comparisons. Obtained results clearly show that protein identification rates differ with respect to the protein database applied for identification. Surprisingly, it was possible to identify 67.96% of the protein spots with the database based on metagenome sequences of the sampled biogas plant 3. Remarkably, using one of the other three biogas-specific datasets for the heterologous biogas plants 1, 2 or 4, between 50% and 55% of the 181 protein spots could be determined.

The comparison of public and taxonomy-restricted database led to better identification rates for the later mentioned. This result confirms earlier observations made by Tanca *et al.* (2013). These authors also stated that the use of taxonomy-restricted databases lead to higher identification rates when compared to databases covering broader species diversity. Likewise, the tendency of higher identification rates using NCBI (taxonomy-restricted) compared to TrEMBL (taxonomy-restricted) could be verified. In general, for publicly available protein databases, it was apparent that only a small fraction of protein spots could be identified

indicating that proteins expressed by microorganisms of biogas fermentation samples are not sufficiently represented in these resources.

In addition, the combined database was reduced (halved, quartered and divided into eight pieces) to avoid redundancy of genes and corresponding proteins. The same amount of proteins, respectively 123, could be identified when the appropriate database BGP3 was halved (data not shown). Obviously, the halved database for BGP3 still sufficiently covers the diversity of abundant proteins expressed by biogas community members. On the contrary, identification rates decrease when the BGP3 database was reduced to 25% (100 identified spots) or even 12.5% (89 identified spots) of the original size. In this case, it is very likely that the database does not cover the diversity of abundant proteins.

Obtained identification rates for heterologous biogas plants are in agreement with the biogas-process-type, namely mesophilic or thermophilic and other process parameters such as fed substrates of the production-scale biogas plants corresponding to the biogas-specific protein databases. As described above, biogas plants BGP1, -2 and -3 represent mesophilic systems, whereas BGP4 operates under thermophilic conditions.

Subsequently, the question was addressed, whether the results using heterologous metagenome datasets are suitable for application on non-corresponding metaproteome analyses. Therefore, the identifications were evaluated by comparing the results for the same spot between different protein databases assuming that the corresponding protein database represents the correct result. All biogas plant related databases were compared regarding their concordance of the best hit based on the Protein Basic Local Alignment Search Tool (BLASTp) (Altschul *et al.*, 1990). The results indicated, that the largest proportion of concordant identification results corresponds to the group showing 100% sequence identity, respectively 52 up to 70% (Supplement Table S1). Around 6 to 13% of the identified proteins show identity values of more than 98% and 10 to 27% have a similarity of less than 98% but

more than 80%. Therefore, it can be concluded that in 64% to 77% of all cases, obtained identification results are reliable since at least a highly similar protein was detected based on application of different biogas-specific databases.

It can be concluded, that integrated metagenome/-proteome analyses for a production-scale biogas plant resulted in protein identification rates of over 67%. Moreover, obtained results showed that the exploitation of metagenome sequence data derived from heterologous biogas plants for compilation of specific protein databases is preferable in comparison to public databases. The comparison of the compliance between the metagenome of different biogas plants shows, that the application of heterologous protein databases representing other biogas communities is suitable for protein identification. Hence, compilation of specific protein databases based on corresponding metagenome datasets is very favorable regarding protein identification in metaproteome analyses for biogas communities.

Functional classification of identified proteins in the metaproteome from the biogas microbial community

A detailed view on the protein identification results using the corresponding protein database (BGP3) uncovered functions of importance in the context of biogas production.

For functional classification, identified protein sequences of all analyzed spots were grouped based on comparisons against the KEGG database by means of the Protein Basic Local Alignment Search Tool (BLASTp) (Altschul *et al.*, 1990) and curated manually. Obtained results are summarized in Table 4 sorted by the assigned pathway. In almost all cases more than one KEGG pathway association was given, all of which are individually listed in the last column.

Table 4 shows the best BLAST hits of 123 identified spots representing 105 non-redundant gene IDs. Corresponding gene products could be classified into seven groups, of which three

are larger functional groups, namely ABC transport (24), carbon metabolism (11) and methane metabolism (19). Surprisingly, proteins belonging to ABC transport form the largest group with 24 members. As expected, all 19 identified methanogenesis enzymes originate from *Archaea*. Moreover, for only 19 proteins no function could be predicted. Proteins, which could not be classified to a KEGG pathways are produced by both, *Bacteria* as well as *Archaea*.

FPKM (Fragments Per Kilobase of a gene per Million reads) values were calculated for the target genes encoding identified proteins from the biogas community. Seven out of ten most abundant genes with FPKM values between 15.02 and 25.30 belong to the group assigned to ABC transport. For genes affiliated to methane metabolism FPKM values from 0.3 up to 5.89 were calculated, in which all four genes featuring FPKM values of > 5 originate from *Methanoculleus bourgensis*. Each of these four gene products formed distinct spots within the gels indicating that they represent unique proteins. *Methanoculleus bourgensis* MS2^T is described as a dominant methanogenic archaeon in many biogas plants and is involved in the methanogenesis (Maus *et al.*, 2014). Concerning the abundance of *Methanoculleus bourgensis* MS2^T, obtained results are in accordance to literature findings. Moreover, it has been described that although *Archaea* represent only 4 to 5% of the microbial community in biogas fermentation, methanogenesis enzymes appeared to be highly expressed (Hanreich *et al.*, 2013). Identified proteins of this study provide an almost complete set of enzymes required for methanogenesis, which have also been identified in the work of Hanreich *et al.* (2012, 2013) and Heyer *et al.* (2013). However, not only the methanogenesis enzymes were identified in this study, but also corresponding genes and their context as deduced from contig sequence information was documented (see below).

In general, most of the identifications are in good agreement with metabolic enzymes described in the literature for biogas producing microbial communities. Moreover, the list of

abundant gene products that are of importance for the biogas process could be complemented. For example, different sugar ABC transporter components, an iron-uptake system, a glyceraldehyde-3-phosphate dehydrogenase, enolases, an alcohol dehydrogenase and a succinyl-glutamate desuccinylase/aspartoacylase were not detected previously.

Localization of genes for identified proteins on metagenome contigs and context information

The metagenome of the sampled biogas community can provide additional information since it was assembled to longer continuous sequences, so called contigs, on which in most cases several genes are encoded. Therefore, an integrated metagenome/metaproteome study offers the advantage to locate genes encoding identified proteins on corresponding contigs providing context information. This means that target genes can be analyzed regarding their genetic context, which offers the opportunity to gain further functional insights. Commonly, genes contributing to the same functional context are often clustered or organized in operons. Accordingly, the assembled metagenome is a source of valuable information on the genetic content of organisms encoding identified metaproteome target proteins. Here, three contigs, each representing different pathway types (see Table 4), were chosen exemplarily to illustrate the valuable additional information derived from context analyses. The first contig (Figure 5 A) encodes enzymes involved in the methanogenesis pathway. Three protein spots, namely 1, 77 and 99 on gel 1 (see Figure 2), were identified as methyl-coenzyme M reductase subunits A, B and G, respectively. Corresponding genes are all located on one contig, which was taxonomically assigned to the superkingdom *Archaea* (see Figure 5 A, contig 59072000012, yellow arrows). Additionally, 16 out of 22 genes encoded on contig 59072000012 were annotated to belong to the same functional context, namely methanogenesis. This pathway is not completely encoded on the contig, but in total the genes of five methyl-coenzyme M

reductase subunits (A, B, C, D, G), five formylmethanofuran dehydrogenase subunits (A, B, C, D, F), as well as one other methanogenesis related protein and subunits of CoB/CoM heterodisulfide reductase and methyl-viologen-reducing hydrogenase are present (Figure 5 A, orange arrows). All encoded gene products are key enzymes in the formation of methane (Thauer *et al.*, 2008). Abundance of proteins identified as methanogenesis enzymes in the biogas metaproteome, lead to the conclusion that the corresponding pathway plays a crucial role in the habitat analyzed. Similar results were obtained for different biogas-producing fermenters (i.e. Ziganshin *et al.*, 2012; Liu and Whitman, 2008; Garcia *et al.*, 2000).

Secondly, a contig was analyzed which encodes basic fermentation pathways. Protein spot 59 on gel 1 was identified as an enolase assigned to *Streptococcus equinus* and encoded on contig 58257000014 (see Figure 5 B, yellow arrow). Commonly, enolases are involved in carbon metabolism. For example, an enolase catalyzes the reversible reaction from 2-phosphoglycerate to phosphoenolpyruvate (Raghunathan *et al.*, 2014). In total, contig 58257000014 harbors 22 genes. Five of these genes encode proteins involved in amino acid and/or glycerolipid metabolism (gene 2, phosphoribosyl-ATP pyrophosphatase; gene 4, glycerate kinase; gene 8, glutamine amidotransferase; gene 9, UDP-N-acetyl-muramyl peptide synthase; gene 12, phosphoglucosamine mutase), but are not organized in a single cluster on this contig (Smith and Ames, 1965; Doughty *et al.*, 1966; Weng and Zalkin, 1987; van Heijenoort, 2001). In addition, there is a cluster consisting of three genes encoding subunits of a CRISPR-Cas-system (genes 18, 19, 20, Figure 5 B, blue arrow cluster). This system is responsible for conferring defense against exogenous DNA from e.g. phages or conjugative plasmids (Barrangou, 2015). Gene 3 on the contig encodes a pullulanase, an enzyme involved in degradation of starch, amylopectin and pullulan found in plants (Figure 5 B, blue arrow) (Nisha and Satyanarayana, 2013; Domań-Pytka and Bardowski, 2004). Bacterial species producing pullulanases may play an important role in the degradation of plant-derived

carbohydrates. They contribute to the hydrolysis step of biomass conversion (Ali Shah *et al.*, 2014). In summary, the contig encoding an enolase gene provides further functional information on the corresponding microorganism.

Transporter systems for the import of substrates into the cell are of importance for the biogas digestion process. Comparison of the protein sequence represented by spot 66 with database entries led to the annotation of a mannoside ABC transporter system protein (see Figure 5 C, contig 333000015, yellow arrow). The corresponding contig was assigned to the superkindom *Bacteria* and specifically to the species *Halanaerobium saccharolyticum*. This contig is of interest, since it harbors six additional genes forming a cluster of genes representing the same functional context as the target gene (see Figure 5 C, contig 333000015, orange arrows). Six genes are part of the ABC transporter gene cluster transcribed in the same direction. The seventh gene of the contig encodes a histidine kinase.

Currently, it is not possible to predict the substrate for the ABC transport system encoded on contig 333000015 since corresponding gene products only show limited similarity to known systems. Although the identified histidine kinase gene is highly conserved (Mascher *et al.* 2006, Mascher 2014), its functional context also remains to be determined.

509 **Conclusions**

510 In summary, this study is the first integrated metagenome/metaproteome study of a
511 production-scale biogas plant. It could be stated, that availability of a corresponding
512 metagenome dataset is crucial for metaproteome analysis. If available, metagenome datasets
513 related to the analyzed system should be chosen before publicly available databases are used.
514 Selection of a suitable metagenome should consider comparable metadata, for example
515 temperature or substrate profile. Furthermore, the effect of sequencing depth suggests
516 investing in next-generation sequencing techniques.

517 The information regarding biogas formation based on metaproteome data shown in this
518 analysis is in good accordance to recently published studies. Based on the deeply sequenced
519 metagenome dataset, it was possible to assign the majority of proteins to functions and to
520 reduce the proportion of unknown proteins in comparison to previous studies. Application of
521 state-of-the-art metaproteomics methods predictably will lead to highly resolved
522 metaproteome data.

523 An integrated metagenome/-proteome approach using assembled genomes offers the
524 opportunity to examine contig context information for genes encoding identified proteins of
525 the metaproteome. Allocation of identified proteins to contigs and exploitation of context
526 information for target genes illustrates that the outlined approach further specifies pathways
527 operating in the biogas plant analyzed. Prospective improvements in metagenome assembly
528 algorithms and binning tools may enable genome-centered metaproteomics. With this
529 approach, expression of dominant proteins can be analyzed against the background of genome
530 information thus allowing prediction of functions for individual organisms. In the context of
531 biogas production, genome-enabled metaproteomics is believed to provide deeper insights
532 into functional roles of dominant community members, thus facilitating implementation of
533 new optimization strategies for the biogas process.

534

535 **Acknowledgements**

536 VO, IM & AB acknowledge support from the CLIB Graduate Cluster “Industrial Biotechnology” co-financed by
537 the Ministry of Innovation of North Rhine-Westphalia. The authors would like to thank R. Hoffrogge for advice.
538 The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User
539 Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-
540 05CH11231. The bioinformatics support of the ‘Bielefeld-Gießen Center for Microbial Bioinformatics—BiGi’
541 consortium within the German Network for Bioinformatics Infrastructure (de.NBI) is gratefully acknowledged.
542 AS also acknowledges funding by the BMBF (FKZ 03SF0440C) and the BMEL (FKZ 22404015).

References

- Alberti, A., Belser, C., Engelen, S., Bertrand, L., Orvain, C., Brinas, L., Cruaud, C., Giraut, L., Da Silva, C., Firmo, C. and Aury, J.M., 2014. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC genomics*, 15(1).
- Ali Shah, F., Mahmood, Q., Maroof Shah, M., Pervez, A. and Ahmad Asad, S., 2014. Microbial ecology of anaerobic digesters: the key players of anaerobiosis. *SCI. WORLD. J.*, 2014.
- 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.
- Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31, 2:166-9.
- Barrangou R., 2015. The roles of CRISPR-Cas systems in adaptive immunity and beyond. *Curr Opin Immunol.* 2015 Jan 6;32C:36-41.
- Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J., 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.*, 13, 12:R122.
- Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.*, 72(1-2): 248-254.
- Bremges, A., Maus, I., Belmann, P., Eikmeyer, F., Winkler, A., Albersmeier, A., Pühler, A., Schlüter, A., Sczyrba, A., 2015. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. *Gigascience*, 4:33.
- Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 12(1): 59-60.
- Domań-Pytka, M., Bardowski, J., 2004. Pullulan degrading enzymes of bacterial origin. *Crit Rev Microbiol.* : 30(2):107-21.
- Doughty, C.C., Hayashi, J.A., Guenther, H.L., 1966. Purification and properties of D-glycerate 3-kinase from *Escherichia coli*. *J Biol Chem.* ;241(3): 568-72.
- Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N.C., Fraser, C.M., Hettich, R.L., Jansson, J.K., 2012. Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. *PLoS One* 7.
- Garcia, J.L., Patel, B.K., Ollivier, B., 2000. Taxonomic, phylogenetic, and ecological diversity of methanogenic *Archaea*. *Anaerobe*: 6(4):205-26.

574 Hanreich, A., Heyer, R., Benndorf, D., Rapp, E., Pioch, M., Reichl, U., Klocke, M., 2012. Metaproteome
575 analysis to determine the metabolically active part of a thermophilic microbial community producing
576 biogas from agricultural biomass. *Can. J. Microbiol.* , 58(7), 917-922.

577 Hanreich, A., Schimpf, U., Zakrzewski, M., Schlüter, A., Benndorf, D., Heyer, R., Rapp, E., Pühler, A., Reichl,
578 U., Klocke, M., 2013. Metagenome and metaproteome analyses of microbial communities in mesophilic
579 biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. *Syst.*
580 *Appl. Microbiol.* 36, 330–338.

581 Hansmeier, N., Chao, T.C., Puhler, A., Tauch, A. and Kalinowski, J., 2006. The cytosolic, cell surface and
582 extracellular proteomes of the biotechnologically important soil bacterium *Corynebacterium efficiens* YS-
583 314 in comparison to those of *Corynebacterium glutamicum* ATCC 13032. *Proteomics*, 6(1): 233-250.

584 Hettich, R.L., Pan, C., Chourey, K., Giannone, R.J., 2013. Metaproteomics: harnessing the power of high
585 performance mass spectrometry to identify the suite of proteins that control metabolic activities in
586 microbial communities. *Anal. Chem.* 85, 4203–14.

587 Heyer, R., Kohrs, F., Benndorf, D., Rapp, E., Kausmann, R., Heiermann, M., Klocke, M., Reichl, U., 2013.
588 Metaproteome analysis of the microbial communities in agricultural biogas plants. *N. Biotechnol.* 30, 614–
589 622.

590 Heyer, R., Kohrs, F., Reichl, U. and Benndorf, D., 2015. Metaproteomics of complex microbial communities in
591 biogas plants. *Microbial biotechnology*, 8(5), pp.749-763.

592 Huang, L.N., Zhou, H., Chen, Y.Q., Luo, S., Lan, C.Y., Qu, L.H., 2002. Diversity and structure of the archaeal
593 community in the leachate of a full-scale recirculating landfill as examined by direct 16S rRNA gene
594 sequence retrieval. *FEMS Microbiol. Lett.* 214, 235–240.

595 Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N. and Schuster, S.C., 2011. Integrative analysis of
596 environmental sequences using MEGAN4. *Genome research*, 21(9): 1552-1560.

597 Hyatt, D., LoCascio, P.F., Hauser, L.J., Uberbacher, E.C., 2012. Gene and translation initiation site prediction in
598 metagenomic sequences. *Bioinformatics*, 28, 17:2223-30.

599 Jaenicke, S., Ander, C., Bekel, T., Bisdorf, R., Dröge, M., Gartemann, K.H., Jünemann, S., Kaiser, O., Krause,
600 L., Tille, F., Zakrzewski, M., Pühler, A., Schlüter, A., Goesmann, A., 2011a. Comparative and joint
601 analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. *PLoS*
602 *One* 6.

603 Jaenicke, S., Zakrzewski, M., Jünemann, S., Pühler, A., Goesmann, A., Schlüter, A., 2011b. Analysis of the
604 metagenome from a biogas-producing microbial community by means of bioinformatics methods, in:
605 Handbook of Molecular Microbial Ecology II. pp. 403–414.

606 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for
607 gene and protein annotation. *Nucleic Acids Res.* 44(D1):D457-62.

608 Klocke, M., Mähnert, P., Mundt, K., Souidi, K., Linke, B., 2007. Microbial community analysis of a biogas-
609 producing completely stirred tank reactor fed continuously with fodder beet silage as mono-substrate. *Syst.*
610 *Appl. Microbiol.* 30, 139–151.

611 Kolmeder, C.A., de Vos, W.M., 2014. Metaproteomics of our microbiome - developing insight in function and
612 activity in man and model systems. *J Proteomics*, 97: 3-16.

613 Krause, L., Diaz, N.N., Edwards, R. a, Gartemann, K.-H., Krömeke, H., Neuweyer, H., Pühler, A., Runte, K.J.,
614 Schlüter, A., Stoye, J., Szczepanowski, R., Tauch, A., Goesmann, A., 2008. Taxonomic composition and
615 gene content of a methane-producing microbial community isolated from a biogas reactor. *J. Biotechnol.*
616 136, 91–101.

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

621 Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9, 4:357-9.

622 Lauro, F.M., DeMaere, M.Z., Yau, S., Brown, M. V, Ng, C., Wilkins, D., Raftery, M.J., Gibson, J.A.E.,
623 Andrews-Pfannkoch, C., Lewis, M., Hoffman, J.M., Thomas, T., Cavicchioli, R., 2011. An integrative
624 study of a meromictic lake ecosystem in Antarctica. *ISME J.* 5, 879–895. doi:10.1038/ismej.2010.185

625 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009.
626 The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 16:2078-9.

627 Liu, Y., Whitman, W.B., 2008. Metabolic, phylogenetic, and ecological diversity of the methanogenic *Archaea*.
628 *Ann N Y Acad Sci.*;1125:171-89.

629 Mascher, T., Helmann, J.D., Unden, G., 2006. Stimulus perception in bacterial signal-transducing histidine
630 kinases. *Microbiol Mol Biol Rev.*: 70(4):910-38.

631 Mascher., T., 2014. Bacterial (intramembrane-sensing) histidine kinases: signal transfer rather than stimulus
632 perception. *Trends Microbiol.*: 22(10):559-65.

633 Maus, I., Wibberg, D., Stantscheff, R., Stolze, Y., Blom, J., Eikmeyer, F.G., Fracowiak, J., König, H., Pühler, A.,
634 Schlüter, A., 2014. Insights into the annotated genome sequence of *Methanoculleus bourgensis* MS2T,
635 related to dominant methanogens in biogas-producing plants. J Biotechnol. 201:43-53.

636 McHugh, S., Carton, M., Mahony, T., O'Flaherty, V., 2003. Methanogenic population structure in a variety of
637 anaerobic bioreactors. FEMS Microbiol. Lett. 219, 297–304.

638 Mladenovska, Z., Dabrowski, S., Ahring, B.K., 2003. Anaerobic digestion of manure and mixture of manure
639 with lipids: biogas reactor performance and microbial community analysis. Water Sci. Technol. 48, 271–
640 278.

641 Neuhoff, V., Arold, N., Taube, D. and Ehrhardt, W., 1988. Improved staining of proteins in polyacrylamide gels
642 including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie
643 Brilliant Blue G-250 and R-250. Electrophoresis, 9(6): 255-262.

644 Nisha, M., Satyanarayana, T., 2013. Recombinant bacterial amylopullulanases: developments and perspectives.
645 Bioengineered.: 4(6):388-400.

646 Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser.
647 BMC Bioinformatics, vol. 12, no. 1, p. 385.

648 Raghunathan, K., Harris, P.T., Spurbeck, R.R., Arvidson, C.G., Arvidson, D.N., 2014. Crystal structure of an
649 efficacious gonococcal adherence inhibitor: an enolase from *Lactobacillus gasseri*. FEBS Lett.;
650 588(14):2212-6.

651 Rooijers, K., Kolmeder, C., Juste, C., Doré, J., de Been, M., Boeren, S., Galan, P., Beauvallet, C., de Vos, W.M.,
652 Schaap, P.J., 2011. An iterative workflow for mining the human intestinal metaproteome. BMC Genomics
653 12, 6.

654 Schlüter, A., Bekel, T., Diaz, N.N., Dondrup, M., Eichenlaub, R., Gartemann, K.H., Krahn, I., Krause, L.,
655 Kroemeke, H., Kruse, O., Mussnug, J.H., Neuweiger, H., Niehaus, K., Pühler, A., Runte, K.J.,
656 Szczepanowski, R., Tauch, A., Tilker, A., Viehöver, P., Goesmann, A., 2008. The metagenome of a
657 biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-
658 pyrosequencing technology. J. Biotechnol. 136, 77–90.

659 Schulze, W.X., Gleixner, G., Kaiser, K., Guggenberger, G., Mann, M., Schulze, E.-D., 2005. A proteomic
660 fingerprint of dissolved organic carbon and of soil particles. Oecologia 142, 335–343.

661 Smith, D.W., Ames, B.N., 1965. Phosphoribosyladenosine monophosphate, an intermediate in histidine
662 biosynthesis. J Biol Chem. 240:3056-63.

663 Stolze, Y., Zakrzewski, M., Maus, I., Eikmeyer, F., Jaenicke, S., Rottmann, N., Siebner, C., Pühler, A., Schlüter,
 664 A., 2015. Comparative metagenomics of biogas-producing microbial communities from production-scale
 665 biogas plants operating under wet or dry fermentation conditions. *Biotechnology for Biofuels*.
 666 Tanca, A., Palomba, A., Massimo, D., Cubeddu, T., Fraumene, C., Biossa, G., Pagnozzi, D., Addis, M.F., Uzzau,
 667 S., 2013. Evaluating the impact of different sequence databases on metaproteome analysis: Insights from a
 668 lab-assembled microbial mixture. *PLoS One* 8.
 669 Thauer, R.K., Kaster, A.-K., Seedorf, H., Buckel, W., Hedderich, R., 2008. Methanogenic *Archaea*: ecologically
 670 relevant differences in energy conservation. *Nat. Rev. Microbiol.* 6, 579–91.
 671 van Heijenoort, J., 2001. Recent advances in the formation of the bacterial peptidoglycan monomer unit. *Nat.*
 672 *Prod. Rep.* 18: 503-19.
 673 Wang, D.Z., Xie, Z.X. and Zhang, S.F., 2014. Marine metaproteomics: current status and future directions. *J.*
 674 *Proteomics*, 97: 27-35.
 675 Weiland, P., 2006. Biomass digestion in agriculture: A successful pathway for the energy production and waste
 676 treatment in germany. *Engineering in Life Sciences*; 6(3): 302–309.
 677 Wendler, S., Hürtgen, D., Kalinowski, J., Klein, A., Niehaus, K., Schulte, F., Schwientek, P., Wehlmann, H.,
 678 Wehmeier, U.F. and Pühler, A., 2013. The cytosolic and extracellular proteomes of *Actinoplanes* sp.
 679 SE50/110 led to the identification of gene products involved in acarbose metabolism. *J. Biotechnol.*,
 680 167(2): 178-189.
 681 Weng, M.L., Zalkin, H., 1987. Structural role for a conserved region in the CTP synthetase glutamine amide
 682 transfer domain. *J. Bacteriol.*;169(7):3023-8.
 683 Wilmes, P., Bond, P.L., 2004. The application of two-dimensional polyacrylamide gel electrophoresis and
 684 downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* 6, 911–
 685 920.
 686 Wilmes, P., Wexler, M., Bond, P.L., 2008. Metaproteomics provides functional insight into activated sludge
 687 wastewater treatment. *PLoS One* 3.
 688 Zakrzewski, M., Goesmann, A., Jaenicke, S., Jünemann, S., Eikmeyer, F., Szczepanowski, R., Al-Soud, W.A.,
 689 Sørensen, S., Pühler, A., Schlüter, A., 2012. Profiling of the metabolically active community from a
 690 production-scale biogas plant by means of high-throughput metatranscriptome sequencing. *J. Biotechnol.*
 691 158, 248–58.
 692 Ziganshin, A.M., Ziganshina, E.E., Kleinstuber, S., Pröter, J., Ilinskaya, O.N., 2012. Methanogenic community
 693 dynamics during anaerobic utilization of agricultural, *Acta Naturae*, 4(4), 91-97.

694 Table 1: Operating conditions of the production-scale biogas plant studied in this survey.

Parameters	Range	BGP3 ^{c)}
<i>pH</i>	7.4 – 8.0	7.53
<i>Conductivity (mS/cm)</i>	22.95 – 32.50	21.3
<i>VOA^{a)} (mg/l)</i>	2,050 – 6,500	3,391
<i>TIC^{b)} (mgCaCO₃/l)</i>	8,500 – 15,000	14.714
<i>VOA^{a)}/TIC^{b)}</i>	0.11 – 0.55	0.23
<i>NH₄-N (g/kg)</i>	1.4 – 2.85	(2.9 (g/l)) 3.15
<i>Nickel (mg/kg)</i>	--	7.7
<i>Cobalt (mg/kg)</i>	--	2.1
<i>Selenium (mg/kg)</i>	--	1.5
<i>Molybdenum (mg/kg)</i>	--	5
<i>Temperature (°C)</i>	38.5 – 41	40 (mesophilic)
<i>Fed substrates (%)</i>	n.d.	Maize silage (67), pig manure (33)
<i>Retention time (days)</i>	n.d.	81
<i>Biogas yield (l/kg oDMe)</i>	n.d.	528.5
<i>% Methane</i>	n.d.	52.4

a) VOA: Volatile organic acids
b) TIC: Total inorganic carbon
c) Fermentation sample was taken on 17th of April 2013 and Metadata based on 25th of April 2013.

695

696

697

698 Table 2: Characteristics of databases constructed based on metagenome data of BGP1 to 4 for
699 metaproteome analysis of the microbial community residing in BGP3.

Database	Read count	Assembly size	N50 ^{a)}	Number of contigs	Longest contig	Number of deduced protein sequences
BGP 1	532,062,042	613,272,056	9,719	145,055	455,167	675,231
BGP 2	283,504,064	429,748,082	9,125	101,277	491,925	460,667
BGP 3	534,648,332	567,186,175	11,275	124,866	469,394	603,237
BGP 4	559,535,284	356,575,948	12,391	73,607	484,162	384,893

a) Median of contig lengths

700

701 Table 3: Identification rates of analyzed protein spots using different proteome databases.

No.	Database	No. of identified spots	identification rate (%) ^{a)}
1	BGP 3	123	67.96
2	BGP 1	99	54.70
3	BGP 2	96	53.04
4	BGP 4	91	50.28
5	NCBI	41	22.65
6	NCBI taxonomy-restricted ^{b) c)}	71	39.23
7	TrEMBL	56	30.94
12	TrEMBL taxonomy-restricted ^{b) d)}	67	37.02

a) Based on 181 spots analysed

b) Taxonomy-restricted implies a restriction on the kingdoms *Bacteria* and *Archaea*

c) Database update October 2014

d) Database update May 2014

703 Table 4: Classification of identified proteins for BGP3 based on BLASTp analyses against the
704 metabolic pathway database Kyoto Encyclopedia of Genes and Genomes (KEGG).

Pathway ^{a)}	Gene ID	Predicted Function	Taxonomic Classification ^{b)}	E-value	FPKM ^{c)}
ABC transporter (24)	contig-102000013_3	maltose-binding periplasmic protein	<i>Succinatimonas sp. CAG:777</i>	0	1.29
	contig-13466000027_10	peptide ABC transporter substrate-binding protein	<i>Thermotoga naphthophila Subdoligranulum sp. 4_3_54A2FAA</i>	0	0.35
	contig-155000023_19	sugar ABC transporter ATP-binding protein		0	0.40
	contig-204000000_4	Extracellular ligand-binding receptor	<i>Clostridium ultunense</i>	0	20.69
	contig-220000016_7	ABC transporter substrate-binding protein	<i>Persephonella marina</i>	0	15.74
	contig-241000006_1	ABC transporter substrate-binding protein	<i>Anoxybacillus flavithermus</i>	0	0.88
	contig-2650000011_3	Ferric iron ABC transporter, iron-binding protein	<i>Halanaerobium saccharolyticum</i>	0	20.96
	contig-28060000000_1	ABC transporter substrate-binding protein	<i>Butyrivibrio sp. FCS014</i>	0	15.02
	contig-28506000014_3	sugar ABC transporter substrate-binding protein	<i>bacterium UASB14</i>	0	0.16
	contig-292000007_2	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)	<i>Halanaerobium saccharolyticum subsp. saccharolyticum DSM 6643</i>	0	0.16
	contig-29979000009_1	hypothetical protein	<i>Caldicellulosiruptor acetigenus</i>	0	18.68
	contig-333000015_6	Mannoside ABC transport system, sugar-binding protein	<i>Halanaerobium saccharolyticum</i>	0	19.44
	contig-33357000008_11	sugar ABC transporter substrate-binding protein	<i>bacterium UASB14</i>	0	0.67
	contig-34566000014_1	ABC transporter substrate-binding protein	<i>Anoxybacillus flavithermus</i>	0	0.79
	contig-34566000014_8	peptide ABC transporter substrate-binding protein	<i>Aerophobetes bacterium SCGC AAA255-F10</i>	0	0.20
	contig-3485000002_1	sugar ABC transporter substrate-binding protein	<i>Saccharibacillus sacchari Candidatus Caldatriobacterium saccharofermentans</i>	0	17.20
	contig-4032000006_2	hypothetical protein		0	0.76
	contig-58039000013_4	family 5 extracellular solute-binding protein	<i>Truepera radiovictrix</i>	0	13.83
	contig-58436000019_4	sugar ABC transporter substrate-binding protein	<i>bacterium UASB14</i>	0	11.11
	contig-58895000005_83	peptide ABC transporter substrate-binding protein	<i>Halothermothrix orenii</i>	0	0.40
	contig-634000010_4	ABC transporter substrate-binding protein	<i>Treponema brennaborensense</i>	0	10.36
	contig-6432000013_2	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)	<i>Halanaerobium saccharolyticum subsp. saccharolyticum DSM 6643</i>	0	0.20
	contig-7306000009_3	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)	<i>Halanaerobium saccharolyticum subsp. saccharolyticum DSM 6643</i>	1.5E-27	16.61
	contig-8051000014_27	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)	<i>Halanaerobium saccharolyticum subsp. saccharolyticum DSM 6643</i>	0	1.71
Carbon metabolism (11)	contig-10000019_68	Phosphate acetyltransferase	<i>Methanosarcina vacuolata Z-761</i>	0	0.26
	contig-16000016_66	enolase	<i>Enterococcus cecorum</i>	0	0.35
	contig-2292000014_3	glyceraldehyde-3-phosphate dehydrogenase	<i>bacterium UASB14</i>	0	0.17
	contig-33000019_6	fructose-bisphosphate aldolase	<i>[Clostridium] cellulolyticum</i>	0	4.13
	contig-3670000015_10	mannitol-1-phosphate 5-dehydrogenase	<i>Youngiibacter fragilis 232.1</i>	0	3.48

	contig-391000003_8	enolase	<i>Firmicutes bacterium CAG:631</i>	0	3.26
	contig-45750000007_3	enolase	<i>Holdemania massiliensis</i>	0	2.63
	contig-58257000014_6	enolase	<i>Streptococcus equinus</i>	0	1.67
	contig-59109000012_22	lactate dehydrogenase	<i>Anaerosalibacter sp. ND1</i>	0	0.60
	contig-812000008_30	enolase	<i>Caldicoprobacter oshimai</i>	0	0.91
	contig-8372000004_13	3-hydroxybutyryl-CoA dehydratase	<i>Clostridium sp. M62/1</i>	0	1.91
Glycerolipid metabolism (2)	contig-13208000005_1	glycerol kinase	<i>Clostridiales bacterium VE202-26</i>	0	1.35
	contig-641000029_1	MULTISPECIES: glycerol kinase	<i>Coproccoccus</i>	0	1.78
	contig-10000010_47	N5-methyltetrahydromethanopterin:coenzyme M methyltransferase subunit H	<i>Methanosarcina thermophila TM-1</i>	0	0.30
	contig-11215000006_5	methylenetetrahydromethanopterin reductase	<i>Methanoculleus bourgensis</i>	0	5.66
	contig-14000027_1	Methanol methyltransferase corrinoid protein	<i>Methanosarcina thermophila TM-1</i>	0	0.25
	contig-14000027_2	Methanol:corrinoid methyltransferase	<i>Methanosarcina thermophila TM-1</i>	0	0.26
	contig-149000013_14	methyl-coenzyme M reductase	<i>Methanobacterium formicicum</i>	0	0.19
	contig-30282000010_10	methyl-coenzyme M reductase	<i>Methanoculleus bourgensis</i>	0	3.38
	contig-30282000010_6	methyl-coenzyme M reductase	<i>Methanoculleus bourgensis</i>	0	5.89
	contig-30282000010_7	methyl-coenzyme M reductase	<i>Methanoculleus bourgensis</i>	0	2.99
Methane metabolism (19)	contig-31000029_162	Methyl coenzyme M reductase alpha subunit	<i>Methanosarcina thermophila CHTI-55</i>	0	0.30
	contig-31000029_163	Methyl coenzyme M reductase gamma subunit	<i>Methanosarcina thermophila CHTI-55</i>	0	0.25
	contig-31000029_166	Methyl coenzyme M reductase beta subunit	<i>Methanosarcina thermophila TM-1</i>	0	0.32
	contig-41285000028_3	methylenetetrahydromethanopterin dehydrogenase	<i>Methanoculleus bourgensis</i>	0	5.48
	contig-58738000017_27	methylenetetrahydromethanopterin reductase	<i>Methanoculleus bourgensis</i>	0	1.83
	contig-59072000012_1	methyl-coenzyme M reductase subunit A	<i>uncultured archaeon</i>	0	4.43
	contig-59072000012_2	methyl-coenzyme M reductase subunit G	<i>uncultured archaeon</i>	0	4.91
	contig-59072000012_5	methyl-coenzyme M reductase subunit B	<i>uncultured archaeon</i>	0	4.56
	contig-729000016_2	methyl-coenzyme M reductase	<i>Methanoculleus bourgensis</i>	0	5.18
	contig-799000025_17	methyl-coenzyme M reductase	<i>Methanoculleus bourgensis</i>	0	1.85
	contig-799000025_20	methyl-coenzyme M reductase subunit G	<i>uncultured archaeon</i>	0	2.15
RNA degradation (3)	contig-58866000005_4	chaperone protein DnaK	<i>Clostridium paraputrificum</i>	0	0.26
	contig-93000013_1	chaperone protein DnaK	<i>Mageeibacillus indolicus</i>	0	3.06
	contig-93000013_29	molecular chaperone GroEL	<i>Clostridium josui</i>	0	3.27
	contig-1182000015_2	s-adenosylmethionine:tRNA ribosyltransferase-isomerase	<i>Firmicutes bacterium CAG:114</i>	0	0.11
	contig-13000011_26	Small heat shock protein HSP16.5	<i>Methanosarcina sp. Kolksee</i>	0	0.41
	contig-14996000023_2	flagellin	<i>Treponema brennaborens</i>	0	4.57
Other (27)	contig-178000011_10	isopropanol dehydrogenase	<i>Mageeibacillus indolicus</i>	0	3.46
	contig-18069000025_2	putative biotin ligase	<i>Candidatus Cloacimonas acidaminovorans str. Evry</i>	2.1E-35	0.38
	contig-22000003_7	Superoxide dismutase (Mn/Fe)	<i>Methanosarcina thermophila TM-1</i>	0	0.34
	contig-2228000018_1	integrase	<i>Thermaerobacter marianensis</i>	0	0.11
	contig-	isopropanol dehydrogenase	<i>Clostridium ultunense</i>	0	1.55

	2373000017_21				
	contig-2555000002_1	aspartyl/glutamyl-tRNA amidotransferase subunit B	<i>Thermincola potens</i>	0	0.14
	contig-26417000014_6	antigen	<i>Methanosarcina thermophila TM-1</i>	0	0.26
	contig-3050000018_2	methylmalonyl-CoA epimerase	<i>Marteella endophytica</i>	0	0.17
	contig-31920000024_3	luciferase	<i>Methanoculleus bourgensis</i>	0	0.44
	contig-3413000026_3	lipoprotein	<i>Defluviitoga tunisiensis</i>	0	0.17
	contig-38000023_61	FAD-dependent oxidoreductase	<i>Alkaliphilus oremlandii</i>	0	0.71
	contig-4152000003_2	methylated-DNA--protein-cysteine methyltransferase	<i>Alkaliphilus oremlandii</i>	0	0.21
	contig-423000028_3	2'-5' RNA ligase	<i>Clostridium botulinum</i>	1.4E-26	0.20
	contig-49000002_16	lipoprotein	<i>Defluviitoga tunisiensis</i>	0	1.94
	contig-50050000015_1	histidine kinase	<i>Methanoculleus bourgensis</i>	0	0.60
	contig-52360000008_2	recombinase	<i>Firmicutes bacterium CAG:41</i>	0	0.40
	contig-5482000013_4	DNA primase	<i>[Clostridium] cellulolyticum</i>	0	0.42
	contig-5668000002_8	surface antigen BspA-like	<i>Trichomonas vaginalis G3</i>	0	0.48
	contig-58425000028_28	quinol dehydrogenase	<i>Campylobacter ureolyticus</i>	0	0.19
	contig-58526000019_8	flagellar protein	<i>Treponema bryantii</i>	0	9.44
	contig-59174000000_34	cell wall hydrolase/autolysin	<i>Caldicellulosiruptor hydrothermalis 108</i>	1.1000 2E-42	0.81
	contig-82000010_35	imidazole glycerol phosphate synthase subunit hisF	<i>Ruminococcus champanellensis</i>	0	1.07
	contig-8372000004_5	superoxide dismutase	<i>Solirubrobacterales bacterium URHD0059</i>	0	1.89
	contig-903000026_1	lon protease	<i>Firmicutes bacterium CAG:114</i>	0	0.07
unknown (19)	contig-13289000009_24				2.23
	contig-1663000024_2				0.48
	contig-21261000019_1				25.30
	contig-23000022_6	hypothetical protein	<i>Clostridiales bacterium 9403326</i>	0	1.73
	contig-24000025_13				1.21
	contig-265000021_7	hypothetical protein	<i>Pseudobacteroides cellulosolvens</i>	0	1.09
	contig-33988000006_1	hypothetical protein	<i>Acaryochloris sp. CCME 5410</i>	7E-09	0.12
	contig-3422000003_42				1.31
	contig-3605000001_4				3.87
	contig-361000024_2	hypothetical protein QT03_C0001G1291	<i>archaeon GW2011_AR10</i>	8.7E-23	0.56
	contig-3886000005_2	hypothetical protein	<i>Proteiniphilum acetatigenes</i>	0	0.26
	contig-505000025_18	MULTISPECIES: hypothetical protein	<i>Bacteroides</i>	0	0.33
	contig-510000014_19	hypothetical protein	<i>Clostridium lundense</i>	0	0.21
	contig-5770000002_2	hypothetical protein	<i>bacterium UASB270</i>	0	20.81
	contig-58074000015_56	hypothetical protein, partial	<i>Clostridiales bacterium DRI-13</i>	2.2E-07	0.30
	contig-61000003_2				0.85
	contig-8123000003_1	hypothetical protein	<i>Caldicoprobacter oshimai</i>	0	0.19
	contig-8372000004_10	hypothetical protein	<i>Clostridiales bacterium 9403326</i>	0	1.44
	contig-8372000004_9	hypothetical protein	<i>Clostridiales bacterium 9403326</i>	0	1.58

- a) Pathway hits with less than two proteins or with predicted function but without pathway association were combined in the group "others"
- b) Classification based on Prodigal_v2.6.2
- c) Fragments Per Kilobase of a gene per Million reads (FPKM) values calculated for the target genes based on sequencing reads

Figure 1: Taxonomic composition of assigned reads of the microbial community in the analyzed biogas plant 3 (BGP3) determined by applying the Lowest-Common-Ancestor (LCA) algorithm implemented in the MGX metagenomics analysis platform and visualized by Krona tools (Ondov *et al.*, 2011). In total 4,138,570 reads (41%) were unassigned.

Figure 2: Identification and labeling of protein spots on a 2D-SDS-PAGE (Gel 1) for separation of the cytosolic metaproteome extracted from microorganisms residing in a fermentation sample obtained from BGP3.

Figure 3: Annotation of three selected metagenomic contigs encoding proteins identified by metaproteome analyses for BGP3. Three exemplarily chosen contigs from the metagenome assembly for BGP3 provide context information regarding genes encoding identified metaproteome proteins. Arrows specify genes and their transcriptional orientation. Yellow: Gene encoding the protein identified in the metaproteome analysis performed. Orange: Genes functionally associated to the target gene. Blue: Genes not functionally associated to the target gene, but interesting for the biogas community. Grey: Genes without any functional predictions or putative role for the biogas community.

Figure 1

Figure 2

Figure 3