DATA NOTE

Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant

Andreas Bremges[1,2*], Irena Maus[1], Peter Belmann[1,2], Felix Eikmeyer[1], Anika Winkler[1], Andreas Albersmeier[1], Alfred Pühler[1], Andreas Schlüter[1†] and Alexander Sczyrba[1,2†]

[*]Correspondence:
abremges@cebitec.uni-bielefeld.de
[1]Center for Biotechnology, Bielefeld University, 33615 Bielefeld, Germany
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: The production of biogas takes place under anaerobic conditions and involves microbial decomposition of organic matter. Most of the participating microbes still have to be considered as unknown and non-cultivable. Accordingly, shotgun metagenome sequencing currently is the method of choice to obtain insights into community composition and the genetic repertoire.

Findings: Here, we report the deeply sequenced metagenome and metatranscriptome of a complex biogas-producing microbial community from an agricultural production-scale biogas plant. We assembled the metagenome and e.g. reconstructed most genes involved in the methane metabolism, a key pathway involving methanogenesis performed by methanogenic *Archaea*. This exemplary result indicates sufficient sequencing coverage for most downstream analyses.

Conclusions: Sequenced at least one order of magnitude deeper than previous studies, our metagenome data will enable novel insights into community composition and the genetic potential of important community members. Moreover, mapping of transcripts to reconstructed genome sequences will enable the identification of active metabolic pathways in target organisms.

Keywords: Biogas; Anaerobic digestion; Wet fermentation; Methanogenesis; Metagenomics; Metatranscriptomics; Sequencing; Assembly

Data description

Background

Production of biogas by means of anaerobic digestion of biomass is becoming increasingly important as biogas is regarded a clean, renewable and environmentally compatible energy source [1]. Moreover, generation of energy from biogas relies on a balanced carbon dioxide cycle.

The process of biogas production takes place under anaerobic conditions and involves microbial decomposition of organic matter, yielding methane as the main final product of the fermentation process. Complex consortia of microorganisms are responsible for biomass decomposition and biogas production. The majority of the participating microbes are still unknown, as is their influence on reactor performance. Since most of the organisms within biogas communities are non-cultivable

by today's conventional microbiological techniques, sequencing of metagenomic total community DNA currently is the best way to obtain unbiased insights into community composition and the metabolic potential of key community members.

Here, we describe the deeply sequenced metagenome and metatranscriptome of an agricultural production-scale biogas plant on the Illumina platform [2]. We sequenced the metagenome 27× and 19× deeper, respectively, than previous studies applying 454 or SOLiD sequencing [3, 4], primarily focusing on community composition. Metatranscriptomic sequencing of total community RNA, 230× deeper than previously reported [5], complements the metagenome. Combined, these data will enable a deeper exploration of the biogas-producing microbial community, with the objective to develop rational strategies for process optimization.

Digester management and process characterization

The biogas plant, located in North Rhine Westphalia, Germany, features a mesophilic continuous wet fermentation technology characterized recently [6]. It was designed for a capacity of $537\,kW_{el}$ combined heat and power (CHP) generation. The process comprises three digesters: a primary and secondary digester, where the main proportion of biogas is produced, and a storage tank, where the digestate is fermented thereafter.

The primary digester is fed hourly with a mixture of 72 % maize silage and 28 % liquid pig manure. The biogas and methane yields at the time of sampling were at 810.5 and 417.8 liters per kg organic dry matter ($l/kg\,oDM$), respectively. After a theoretical retention time of 55 days, the digestate is stored in the closed, non-heated final storage tank. Further metadata are summarized in Table 1.

Sampling and library construction

Samples from the primary digester of the aforementioned biogas plant were taken in November 2010. Prior to the sampling process, approximately $15\,l$ of the fermenter substrate were discarded before aliquots of $1\,l$ were transferred into clean gastight sampling vessels and transported directly to the laboratory.

For the metagenome, aliquots of $20\,g$ of the fermentation sample were used for total community DNA preparation as described previously [7].

For the metatranscriptome, a random-primed cDNA library was prepared by an external vendor (Vertis Biotechnologie AG). Briefly, total RNA was first treated with 5′-P dependent Terminator exonuclease (Epicentre) to enrich for full-length mRNA carrying 5′ CAP or triphosphate structures. Then, first-strand cDNA was synthesized using a N6 random primer and M-MLV-RNase H reverse transcriptase, and second-strand cDNA synthesis was performed according to the Gubler-Hoffman protocol [8].

Metagenomic and metatranscriptomic sequencing

We sequenced one metatranscriptome and two metagenome shotgun libraries on Illumina's Genome Analyzer IIx system, applying the Paired-End DNA Sample Preparation Kit (Illumina Inc.) as described by the manufacturer and generating $2\times161\,bp$ paired-end reads. On Illumina's MiSeq system, we sequenced three further metagenome shotgun libraries, applying the Nextera DNA Sample Preparation Kit

(Illumina Inc.) as described by the manufacturer and generating $2 \times 155\,bp$ paired-end reads. Our sequencing efforts, yielding 35 gigabases in total, are summarized in Table 2.

Metagenome assembly

Prior to assembly, we used Trimmomatic [9], version 0.32, for adapter removal and moderate quality trimming. After adapter clipping, using Trimmomatic's *Truseq2-PE* and *Nextera-PE* templates, we removed leading and trailing ambiguous or low quality bases (below Phred quality scores of 3). Table 3 summarizes the effect on sequencing depth, more than 25 gigabases of sequence data passed quality control.

We assembled the metagenome with Ray Meta [10], version 2.3.1, trying a range of k-mer sizes from 21 to 61 in steps of 10. To estimate the inclusivity of the set of assemblies, we aligned the post-QC sequencing reads to the assembled contigs with bowtie2 [11], version 2.2.4. We then used samtools [12], version 1.1, to convert SAM to BAM, sort the alignment file, and calculate the mapping statistics. Based on total assembly size, contiguity, and the percentage of mapped back metagenomic reads, we selected the assembly produced with a k-mer size of 31. Here, we assembled approximately 228 megabases in $54,489$ contigs greater than $1,000\,bp$, with an N50 value of $9,796\,bp$. 77 % (79 %) of metagenomic (metatranscriptomic) reads mapped back to the assembly.

Gene prediction and annotation

We used MetaProdigal [13], version 2.6.1, to predict $250,596$ protein-coding genes on the assembled contigs. We blasted the protein sequences of all predicted genes against the KEGG database [14], release 72.0, using Protein-Protein BLAST [15], version 2.2.29+. Of the $250,596$ predicted genes, $191,766$ (76.5 %) had a match in the KEGG database, using an Evalue cutoff of 10^{-6}. We determined the KEGG Orthology (KO) for each gene by mapping the top-scoring BLAST hit to its orthologous gene in KEGG, resulting in $109,501$ genes with an assigned KEGG Orthology. Table 4 summarizes our results.

Relating the metagenome and the metatranscriptome

To illustrate potential use cases, we first counted the number of reads within genes using BEDTools [16], version 2.22.0, and highlighted metagenomic and metatranscriptomic coverage of the methane metabolism pathway in Figure 1. We therefore assembled the majority of genes involved in the methane metabolism from our metagenomic data, with accompanying metatranscriptomic data suggesting active gene expression for many.

For a second example, we calculated the reads per kilobase per million mapped reads (RPKM) for each gene as a crude measure for abundance (metagenome) or expression (metatranscriptome). Figure 2 relates the two. Here, we accentuated all genes assigned to either of the three known types of methanogenic pathways. Hydrogenotrophic methanogenesis, i.e. the reduction of CO_2 with hydrogen, appears to be highly expressed in the reactor analyzed, which is in agreement with results obtained via 454 amplicon and metatranscriptome sequencing [5].

Discussion

We report extensive metagenomic and metatranscriptomic profiling of the microbial community from a production-scale biogas plant. Given the unprecedented sequencing depth and established bioinformatics, our data are of great interest to the biogas research community in general and microbiologists working on biogas-producing microbial communities in particular. In a first applied study, our metagenome assembly was used to improve the characterization of a metaproteome generated from biogas plant fermentation samples and to investigate the metabolic activity of the microbial community [17].

By sharing our data, we want to actively encourage its reuse. This will hopefully result in novel biological and biotechnological insights, eventually enabling a more efficient biogas production.

Availability

Data accession

The datasets supporting the results of this article are available in the European Nucleotide Archive (ENA) under study accession PRJEB8813.

Intermediate results for the review process are deposited in the project's GitHub repository for now [18], and will be uploaded to GigaDB [2] upon acceptance.

Reproducibility

The complete workflow is organized in a single GNU Makefile and available on GitHub [18]. All data and results can be reproduced by a simple invocation of *make*. To further support reproducibility, we bundled all tools and dependencies into one Docker container available on DockerHub [19]. *docker run* executes the aforementioned Makefile inside the container. Reproduction requires roughly 74 GiB memory and 200 GiB storage.

Excluding the KEGG analysis, which relies on a commercial license of the KEGG database, all steps are performed using free and open-source software.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

AB conveived and performed all bioinformatic analyses and wrote the paper. IM investigated all metadata and drafted part of the data description. PB implemented the accompanying Docker container. FE collected the study material. AW and AA provided the sequencing service. AP acquired funding and revised the paper. ASch and AScz jointly directed the project and extensively revised the paper. All authors read and approved the final manuscript.

Acknowledgements

AB, IM, and FE are supported by a fellowship from the CLIB Graduate Cluster Industrial Biotechnology. We gratefully acknowledge funding by the German Federal Ministry of Food and Agriculture (BMEL), grant number 22006712 (joint research project Biogas-Core) and the German Federal Ministry of Education and Research (BMBF), grant number 03SF0440C (joint research project Biogas-Marker). We acknowledge support of the publication fee by Deutsche Forschungsgemeinschaft and the Open Access Publication Funds of Bielefeld University. Lastly, we wish to thank C. Titus Brown, Kornél L. Kovács, and Sebastien Boisvert for detailed and constructive open peer reviews.

Author details

[1]Center for Biotechnology, Bielefeld University, 33615 Bielefeld, Germany. [2]Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany.

References

1. Weiland, P.: Biogas production: current state and perspectives. Appl. Microbiol. Biotechnol. **85**(4), 849–860 (2010). doi:10.1007/s00253-009-2246-7
2. Bremges, A., Maus, I., Belmann, P., Eikmeyer, F., Winkler, A., Albersmeier, A., Pühler, A., Schlüter, A., Sczyrba, A.: Supporting data and materials for "Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant". GigaScience Database (2015). doi:http://dx.doi.org/foo/bar
3. 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., Goesmann, A.: Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. PLoS ONE **6**(1), 14519 (2011). doi:10.1371/journal.pone.0014519
4. Wirth, R., Kovács, E., Maróti, G., Bagi, Z., Rákhely, G., Kovács, K.L.: Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. Biotechnol Biofuels **5**, 41 (2012). doi:10.1186/1754-6834-5-41
5. Zakrzewski, M., Goesmann, A., Jaenicke, S., Jünemann, S., Eikmeyer, F., Szczepanowski, R., Al-Soud, W.A., Sørensen, S., Pühler, A., Schlüter, A.: 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 (2012). doi:10.1016/j.jbiotec.2012.01.020
6. Stolze, Y., Zakrzewski, M., Maus, I., Eikmeyer, F., Jaenicke, S., Rottmann, N., Siebner, C., Puhler, A., Schluter, A.: Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation conditions. Biotechnol Biofuels **8**, 14 (2015). doi:10.1186/s13068-014-0193-8
7. Schlüter, A., Bekel, T., Diaz, N.N., Dondrup, M., Eichenlaub, R., Gartemann, K.H., Krahn, I., Krause, L., Krömeke, H., Kruse, O., Mussgnug, J.H., Neuweger, H., Niehaus, K., Pühler, A., Runte, K.J., Szczepanowski, R., Tauch, A., Tilker, A., Viehöver, P., Goesmann, A.: 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 (2008). doi:10.1016/j.jbiotec.2008.05.008
8. Gubler, U., Hoffman, B.J.: A simple and very efficient method for generating cdna libraries. Gene **25**(2–3), 263–269 (1983). doi:10.1016/0378-1119(83)90230-5
9. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014). doi:10.1093/bioinformatics/btu170
10. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J.: Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. **13**(12), 122 (2012). doi:10.1186/gb-2012-13-12-r122
11. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012). doi:10.1038/nmeth.1923
12. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009). doi:10.1093/bioinformatics/btp352
13. Hyatt, D., LoCascio, P.F., Hauser, L.J., Uberbacher, E.C.: Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics **28**(17), 2223–2230 (2012). doi:10.1093/bioinformatics/bts429
14. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. **42**(Database issue), 199–205 (2014). doi:10.1093/nar/gkt1076
15. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: architecture and applications. BMC Bioinformatics **10**, 421 (2009). doi:10.1186/1471-2105-10-421
16. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010). doi:10.1093/bioinformatics/btq033
17. Kohrs, F., Wolter, S., Benndorf, D., Heyer, R., Hoffmann, M., Rapp, E., Bremges, A., Sczyrba, A., Schlüter, A., Reichl, U.: Fractionation of biogas plant sludge material improves metaproteomic characterization to investigate metabolic activity of microbial communities. Submitted (2015)
18. GitHub Repository. `https://github.com/metagenomics/2015-biogas-cebitec`
19. DockerHub Registry. `https://registry.hub.docker.com/u/metagenomics/2015-biogas-cebitec`

Figures

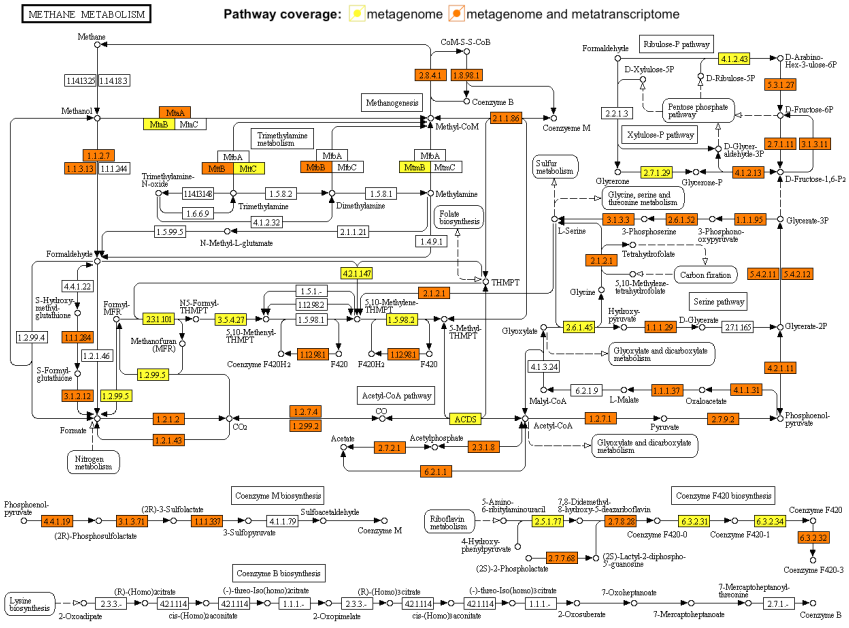

Figure 1 Methane metabolism pathway analysis. Genes reconstructed in our assembly, that are involved in the methane metabolism [PATH:ko00680], are highlighted: Genes with only metagenomic support in yellow, genes with also metatranscriptomic support, suggesting active gene expression, in orange. Methane is metabolized from CO_2, methanol, or acetate. KEGG pathway map courtesy of Kanehisa Laboratories.

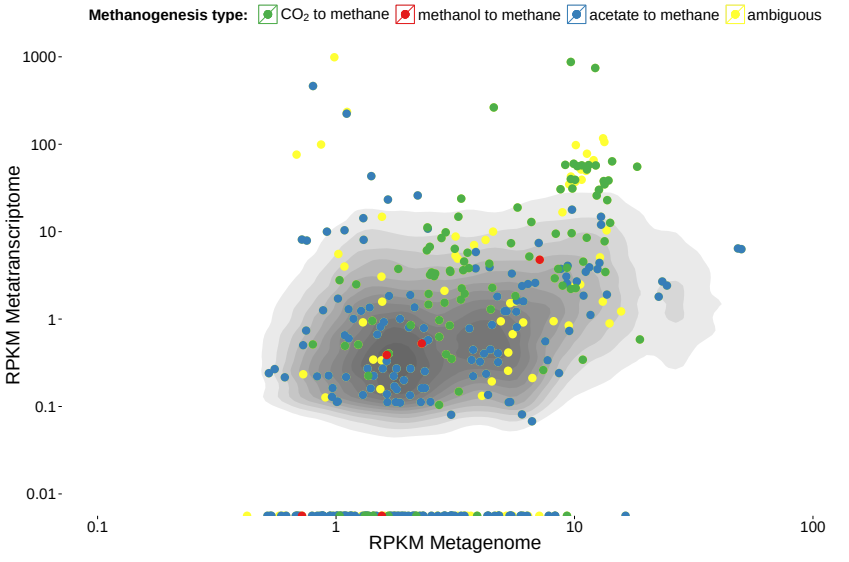

Figure 2 Relating the metagenome and metatranscriptome. Highlighted are genes involved in methanogenesis, color coded by pathway type: CO_2 to methane [MD:M00567] green (96 genes), methanol to methane [MD:M00356] red (5 genes), and acetate to methane [MD:M00357] blue (209 genes). Common genes, shared between pathway types, are yellow (80 genes). In the background a two-dimensional density estimation for all $250,596$ genes.

Tables

Table 1 Characteristics of the studied biogas plant. Primary digester, sampling date: Nov 15, 2010.

Process parameter	Sample
Net volume	$2041\,m^3$
Dimensions	$6.4\,m$ high, diameter of $21\,m$
Electrical capacity	$537\,kW_{el}$
pH	7.83
Temperature	$40\,^{\circ}\mathrm{C}$
Conductivity	$22.10\,mS/cm$
Volative organic acids (VOA)	$5327\,mg/l$
Total inorganic carbon (TIC)	$14397\,mg/l$
VOA/TIC	0.37
Ammoniacal nitrogen	$2.93\,g/l$
Acetic acid	$863\,mg/l$
Propionic acid	$76\,mg/l$
Fed substrates	$72\,\%$ maize silage, $28\,\%$ pig manure
Organic load	$4.0\,kg\,oDM\,m^{-3}\,d^{-1}$
Retention time	$55\,d$
Biogas yield	$810.5\,l/kg\,oDM$
Methane yield	$417.8\,l/kg\,oDM$

Table 2 Overview of the different sequencing libraries.

Accession	Library name	Library type	Insert size[1]	Cycles	Reads	Bases
ERS697694	GAIIx, Lane 6	RNA, TruSeq	202 ± 49	2×161	$78,752,308$	$12,679,121,588$
ERS697688	GAIIx, Lane 7	DNA, TruSeq	157 ± 19	2×161	$54,630,090$	$8,795,444,490$
ERS697689	GAIIx, Lane 8	DNA, TruSeq	298 ± 32	2×161	$74,547,252$	$12,002,107,572$
ERS697690	MiSeq, Run A1[2]	DNA, Nextera	173 ± 53	2×155	$4,915,698$	$761,933,190$
ERS697691	MiSeq, Run A2[2]	DNA, Nextera[3]	522 ± 88	2×155	$1,927,244$	$298,722,820$
ERS697692	MiSeq, Run B1[2]	DNA, Nextera	249 ± 30	2×155	$3,840,850$	$573,901,713$
ERS697693	MiSeq, Run B2[2]	DNA, Nextera[3]	525 ± 90	2×155	$4,114,304$	$614,787,564$

[1]Insert sizes determined with Picard tools. [2]Partial runs. [3]This Nextera library was sequenced twice.

Table 3 Metagenomic and metatranscriptomic sequencing and quality control.

Library type	Reads, raw	post-QC	Bases, raw	post-QC
Metagenome (total)	$143,975,438$	$137,365,053$	$23,046,897,349$	$17,267,320,221$
Metatranscriptome	$78,752,308$	$73,165,986$	$12,679,121,588$	$8,455,809,264$

Table 4 Metagenome assembly statistics, minimum contig size of $1,000\,bp$.

Assembly metric	Our assembly
Total size	$228,382,457\,bp$
Number of contigs	$54,489$
N50 value	$9,796\,bp$
Largest contig	$333,979\,bp$
Mapped DNA reads	$105,461,596\,(77\,\%)$
Mapped RNA reads	$57,436,058\,(79\,\%)$
Predicted genes	$250,596$
of these, full-length	$172,372\,(69\,\%)$
Match in KEGG Genes	$191,766$
of these, assigned KO	$109,501$
of these, in KEGG pathways	$61,100$