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], Felix Eikmeyer[1], Anika Winkler[1], Andreas Albersmeier[1], Alfred Pühler[1], Andreas Schlüter[1†] and Alexander Sczyrba[1,2*†]

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

Abstract

Background: a presentation of the interest or relevance of these data for the broader community

Findings: a very brief preview of the data type(s) produced, the methods used, and information relevant to data validation

Conclusions: a short summary of the potential uses of these data and implications for the field

Keywords: Biogas; Metagenome; Sequencing; Assembly; Annotation

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 is currently the only way to obtain unbiased insights into community composition and the genetic potential of key community members.

Here, we report the first deeply sequenced metagenome of an agricultural production-scale biogas plant on the Illumina platform [2]. We sequenced 27.3× and 19.3× deeper, respectively, than previous studies relying on 454 [3] or SOLiD [4] sequencing. Metatranscriptomic sequencing of total community RNA complements the metagenome. Combined, these data will enable a deeper exploration of the biogas-producting 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 and 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 DNA isolation

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.

Aliquots of 20 g of the fermentation sample were used for total community DNA preparation as described previously [5].

Metagenomic sequencing

In total, we sequenced four different metagenome shotgun libraries with different insert sizes, resulting in 144 million reads yielding more than 23 gigabases sequence information. Table 2 summarizes the statistics of the sequencing approach.

On Illumina's Genome Analyzer IIx system, we sequenced two libraries with an average insert size of 250 nt and 450 nt, respectively, applying the Paired-End DNA Sample Preparation Kit (Illumina Inc.) as described by the manufacturer and generating $2 \times 161\,bp$ paired-end reads.

On Illumina's MiSeq system, we sequenced two further libraries with an average insert size of 190 nt and 690 nt, respectively, applying the Nextera DNA Sample Preparation Kit (Illumina Inc.) as described by the manufacturer and generating $2 \times 155\,bp$ paired-end reads.

Metatranscriptomic sequencing

On Illumina's Genome Analyzer IIx system, we sequenced further metatranscriptome library with an average insert size of 130nt and 380nt, respectively, applying the Paired-End DNA Sample Preparation Kit (Illumina Inc.) as described by the manufacturer and generating $2 \times 160\,bp$ paired-end reads.

search the damn transcript to double-check Irena's info - seems wrong to me

Sequence quality control

We used Trimmomatic [6], 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 2 summarizes the impact of quality control on sequencing depth.

Metagenome assembly and quality assessment

We assembled the metagenome with Ray Meta [7], version 2.3.1, using a k-mer size of 31 and a minimum contig length of $1,000\,bp$. This resulted in a total assembly size of approximately 228 megabases in $54,489$ contigs, with an N50 value of $9,796\,bp$. Table 3 summarizes our results.

We aligned the post-QC sequencing reads to the assembled contigs with bowtie2 [8], version 2.2.4, and used samtools [9], version 1.1, to convert SAM to BAM and thereafter sort the alignment file.

Insert mapping statistics (roughly 80% could be mapped per sample), mention insert sizes as reported by Picardtools (agreement with library prep?), and try to understand the ALE score of -5390031450.700606

Gene prediction and annotation

We then used MetaProdigal [10], version 2.6.1, to predict $250,596$ protein-coding genes on the assembled contigs. Table 3 also includes these results.

We blasted all predicted genes against the KEGG database [11], release 72.0, using Protein-Protein BLAST [12], version 2.2.29+. Of the $250,596$ predicted genes, $191,766$ 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 xxx genes with an assigned KEGG Orthology.

Kegg-Gene to KO mapping contained a bug, will be fixed asap!

Relating the metagenome and the metatranscriptome

We counted aligned reads in predicted genes with BEDTools, version 2.22.0, [13]. Figure 2 shows metagenomic vs. metatranscriptomic coverage in RPKM units.

I'm not 100% happy with this.

Availability

Data accession

The datasets supporting the results of this article are available in the [repository name] repository, [unique persistent identifier and hyperlink to datasets in http:// format].

Data needs to be submitted to SRA (raw reads) and GigaDB (everything).

Reproducibility

The complete workflow is organized in a single GNU Makefile and available on GitHub [14]. Starting from the raw read files, available from SRA and/or GigaDB, all data and results can be reproduced by a simple invocation of *make*. Excluding the KEGG analysis, which relies on a commercial license of the KEGG database, all steps are performed using free and open-source software.

Requirements

I will log runtime and memory in the final run. My goal is to list hardware requirements and CPU time needed to reproduce all results.

Discussion

Potential use cases.

Metagenomic and metatranscriptomic profiling of the biogas-producing microbial community. Highlight, that methane metabolism pathway is widely covered, but still room for improvement, i.e. sequence deeper. Possibly mention new data generated within the CSP? Tricky to phrase it without trashing this data set.

Identification of metaproteomic data out there (cite Vera, in preparation, and Magdeburg).

Ultimate goal: process optimization by biological insights.

Can be written once we agreed upon the rest.

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. FE sampled stuff. AW and AA sequenced stuff. AP provided funding. ASch revised the paper. AScz conveived of many of the analyses and revised the paper. ASch and AScz jointly directed the project. All authors read and approved the final manuscript.

FIXME: Phrasing of middle authors' contributions

Acknowledgements

AB, IM, and FE are supported by a fellowship from the CLIB Graduate Cluster Industrial Biotechnology.

ASch: Biogas Marker, Biogas Core

Acknowledge Stadtwerke?

Author details

[1]Center for Biotechnology, Bielefeld University, Germany. [2]Faculty of Technology, Bielefeld University, 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] [PubMed:19777226]
2. Bremges, A., Maus, I., Pühler, A., Schlüter, A., Sczyrba, A.: Name of the GigaScience repository (2015). [DOI: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). [PubMed Central:PMC3027613] [DOI:10.1371/journal.pone.0014519] [PubMed:21297863]
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). [PubMed Central:PMC3395570] [DOI:10.1186/1754-6834-5-41] [PubMed:22673110]
5. 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] [PubMed:18597880]
6. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014). [PubMed Central:PMC4103590] [DOI:10.1093/bioinformatics/btu170] [PubMed:24695404]
7. 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). [PubMed Central:PMC4056372] [DOI:10.1186/gb-2012-13-12-r122] [PubMed:23259615]
8. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012). [PubMed Central:PMC3322381] [DOI:10.1038/nmeth.1923] [PubMed:22388286]
9. 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). [PubMed Central:PMC2723002] [DOI:10.1093/bioinformatics/btp352] [PubMed:19505943]
10. 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] [PubMed:22796954]
11. 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). [PubMed Central:PMC3965122] [DOI:10.1093/nar/gkt1076] [PubMed:24214961]

12. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: architecture and applications. BMC Bioinformatics **10**, 421 (2009). [PubMed Central:PMC2803857] [DOI:10.1186/1471-2105-10-421] [PubMed:20003500]
13. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010). [PubMed Central:PMC2832824] [DOI:10.1093/bioinformatics/btq033] [PubMed:20110278]
14. Bremges, A.: GitHub (2015). [https://github.com/abremges/2015-biogas-cebitec]

Figures

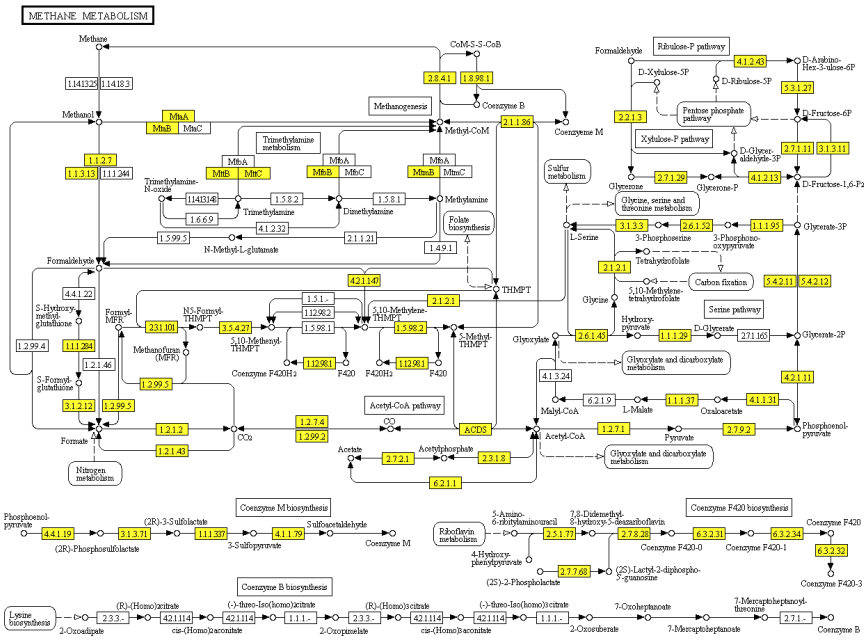

Figure 1 Methane metabolism pathway analysis. Genes reconstructed in our assembly, that are involved in the methane metabolism [PATH:map00680], are highlighted in yellow. Base pathway image copyrighted by Kanehisa Laboratories.

Nicer (?) alternative: 2 colors, metagenome and metatranscriptome!

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	$40\,^{\circ}C$
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	$417.8\,l/kg\,oDM$

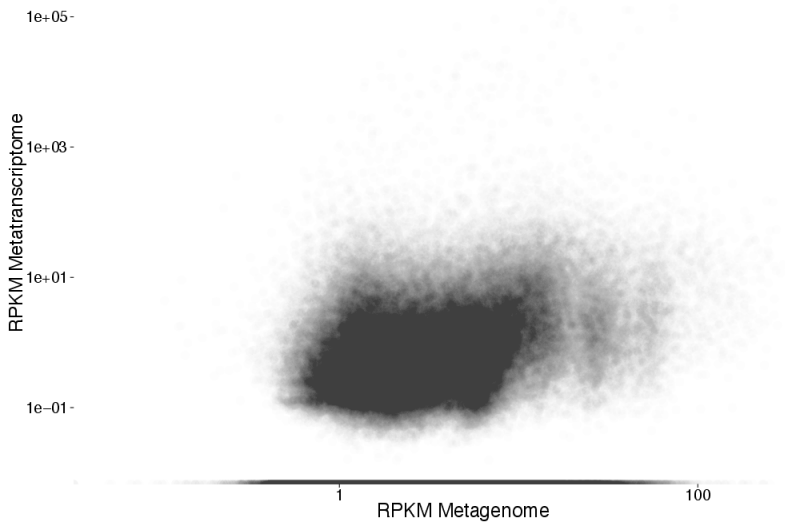

Figure 2 Relating the metagenome and metatranscriptome. Highlighted are genes involved in methanogenesis, color coded by pathway type: CO2 to methane [MD:M00567] green, methanol to methane [MD:M00356] red, and acetate to methane [MD:M00357] blue.

Spotted a mistake in KEGG analysis, figure colors will be re-added asap!

Table 2 Metagenomic and metatranscriptomic sequencing. Initial sequencing statistics and the impact of quality control.

Library name	Type	Insert size	Reads, raw	post-QC	Bases, raw	post-QC
GAIIx, Lane 6	RNA, TruSeq	202 ± 49	$78,752,308$	$73,165,986$	$12,679,121,588$	$8,455,809,264$
GAIIx, Lane 7	DNA, TruSeq	157 ± 19	$54,630,090$	$48,925,129$	$8,795,444,490$	$5,426,329,184$
GAIIx, Lane 8	DNA, TruSeq	298 ± 32	$74,547,252$	$73,642,681$	$12,002,107,572$	$9,746,408,945$
MiSeq, Run 1.1	DNA, Nextera	173 ± 53	$4,915,698$	$4,915,449$	$761,933,190$	$610,841,270$
MiSeq, Run 1.2	DNA, Nextera	522 ± 88	$1,927,244$	$1,927,126$	$298,722,820$	$295,503,323$
MiSeq, Run 2.1	DNA, Nextera	249 ± 30	$3,840,850$	$3,840,582$	$573,901,713$	$573,663,794$
MiSeq, Run 2.2	DNA, Nextera	525 ± 90	$4,114,304$	$4,114,086$	$614,787,564$	$614,573,705$
Total	N/A		xxx	xxx	xxx	xxx

Move insert size metrics from Picardtools to another table, add mapping statistics

Table 3 Metagenome assembly. Some 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	$333,979\,bp$
Predicted genes	$250,596$
of these, full-length	$172,372\,(69\,\%)$
Match in KEGG Genes (10)	$241,153$
Match in KEGG Genes (1e-3)	$200,214$
Match in KEGG Genes (1e-6)	$191,766$
Match in KEGG Genes (1e-9)	$184,251$
of these, assigned KO	xxx, xxx

KOs to be added asap, see above.