DATA NOTE

The old and grumpy biogas metagenome

Andreas Bremges[1,2*], Irena Maus[1], Alfred Pühler[1], Andreas Schlüter[1†] and Alexander Sczyrba[1,2†]

*Correspondence: abremges@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

Biogas important energy source. Clean and awesome. Number of biogas plants in Germany/Europe/worldwide. Key is process optimization, not building more and more plants. Little is known about the microbial community responsible for everything.

Either copy/paste CSP stuff, or Andreas S. writes something nice.

Here, we report the first deeply sequenced metagenome of an agricultural production-scale biogas plant on the Illumina platform [1]. We sequenced 27.3× and 19.3× deeper than previous studies relying on 454 [2] or SOLiD [3] sequencing. These data will enable a deeper exploration of the biogas-producting microbial community, a key step towards process optimization.

Digester management

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 two digesters: a primary digester, where biogas and methane is produced, and a secondary digester, where the digestrate is stored thereafter.

The primary digester is fed hourly with a mixture of maize silage and 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 sampling process, approximately 15 L of fermenter sub-

strate were discarded before aliquots of $1\,L$ were transferred into clean gastight sampling vessels and transported directly to the laboratory.

A $20\,g$ aliquot of the fermentation sample was used for total community DNA preparation as described previously [4].

Metagenomic sequencing

In total, we sequenced four different metagenome shotgun libraries with different insert sizes, resulting in over 23 gigabases scattered across 144 million reads. Table 2 contains the raw numbers.

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.

Sequence quality control

We used Trimmomatic [5], 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 quality 3). Then, we performed an adaptive quality trimming, balancing read length against error rate. We set the target read length to $100\,bp$ and the strictness to 0.2, thus favouring read length over error sensitivity. Finally, reads shorter than 36 bases were discarded.

Table 2 summarizes the impact of quality control on sequencing depth.

Metagenome assembly and quality assessment

2.3.1

We assembled the metagenome with Ray Meta [6], version 2.3.0, 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 217 megabases in 55,563 contigs, with an N50 value of $8,137\,bp$. Table 3 summarizes our results.

2.2.4

We aligned the post-QC sequencing reads to the assembled contigs with bowtie2 [7], version 2.2.1, and used samtools [8], version 1.1, to convert SAM to BAM and thereafter sort the alignment file.

I will run either LAP or ALE to get the probability score. Still struggling with some details.

Gene prediction and annotation

We then used MetaProdigal [9], version 2.6.1, to predict 239,412 protein-coding genes on the assembled contigs. Table 3 also includes these results.

We blasted all predicted genes against the KEGG database [10], release 72.0, using Protein-Protein BLAST [11], version 2.2.29+. Of the 239,412 predicted genes, 230,354 had a match in the KEGG database. We determined the KEGG Orthology (KO) for each gene by mapping the top-scoring BLAST hit to its KO, resulting in 111,380 genes with an assigned KEGG Orthology.

Relating the metagenome and the metatranscriptome

We counted aligned reads in predicted genes with HTSeq [12], version 0.6.1p1. 0.6.1p2
Figure 2 shows where we could be heading. I think it is quite nice, even if not much new insight is provided. But hey, we could buzz *metatranscriptomics* in the title!

Do we want to do this? If so, some additional words are needed.

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 [13]. 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).

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. 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.

AP, ASch

Acknowledgements
AB is supported by a fellowship from the CLIB Graduate Cluster Industrial Biotechnology.

Grants of IM, AP, ASch? Acknowledge Stadtwerke?

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

References

1. Bremges, A., Maus, I., Pühler, A., Schlüter, A., Sczyrba, A.: Name of the GigaScience repository (2014). [DOI:foo/bar]
2. 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]
3. 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]
4. 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]
5. 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]
6. 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]
7. 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]
8. 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]
9. 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]
10. 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]
11. 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]
12. Anders, S., Pyl, P.T., Huber, W.: HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics (2014). [DOI:10.1093/bioinformatics/btu638] [PubMed:25260700]
13. Bremges, A.: GitHub (2014). [https://github.com/abremges/2014-biogas]

Figures

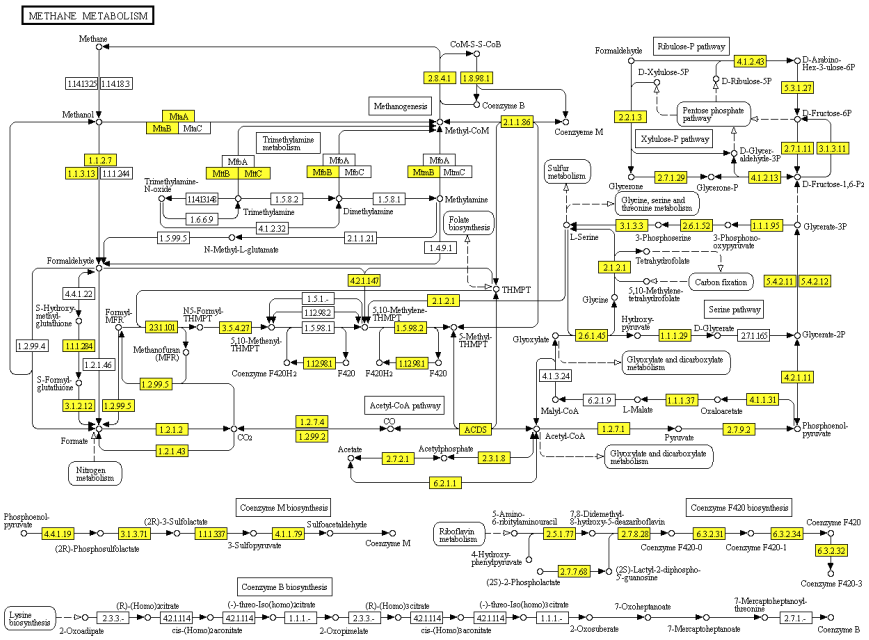

Figure 1 Methane metabolism pathway analysis. Genes reconstructed in our assembly, that are involved in the methane metabolism [PATH:map00680], are highlighted in red. Base pathway image copyrighted by Kanehisa Laboratories.

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.

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	$417.8\,l/kg\,oDM$

TIC only calculated, might be a few digits off

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

Library name	Insert size	Reads, raw	post-QC	Bases, raw	post-QC
GAIIx, Lane 7	183 ± 26	$54,630,090$	$32,383,938$	$8,795,444,490$	$3,899,740,740$
GAIIx, Lane 8	296 ± 49	$74,547,252$	$71,969,779$	$12,002,107,572$	$9,616,193,061$
MiSeq, Run 1.1	214 ± 53	$4,915,698$	$3,632,111$	$761,933,190$	$468,956,057$
MiSeq, Run 1.2	528 ± 117	$1,927,244$	$1,921,175$	$298,722,820$	$277,093,745$
MiSeq, Run 2.1	245 ± 36	$3,840,850$	$3,831,040$	$573,901,713$	$560,751,191$
MiSeq, Run 2.2	531 ± 118	$4,114,304$	$4,103,448$	$614,787,564$	$580,918,665$

Table 3 Metagenome assembly. Some assembly statistics, minimum contig size of $1,000\,bp$.

Assembly metric	Our assembly
Total size	$216,554,757\,bp$
Number of contigs	$55,563$
N50 value	$8,137\,bp$
Largest contig	$319,083\,bp$
Predicted genes	$239,412$
of these, full-length	$160,124\,(66.9\,\%)$
Match in KEGG Genes	$230,354$
of these, assigned KO	$111,380$