



OPEN

DATA DESCRIPTOR

Representative Metagenomes of Mesophilic Biogas Reactor Across South Korea

Darsha Prabhakaran¹, Young Wook Go¹, Hyunjin Kim², Seongcheol Kang¹
& Byoung-In Sang¹✉

Biogas production through the anaerobic digestion (AD) of organic waste plays a crucial role in promoting sustainability and closing the carbon cycle. Over the past decade, this has driven global research on biogas-producing microbiomes, leading to significant advances in our understanding of microbial diversity and metabolic pathways within AD plants. However, substantial knowledge gaps persist, particularly in understanding the specific microbial communities involved in biogas production in countries such as South Korea. The present dataset addresses one of these gaps by providing comprehensive information on the metagenomes of five full-scale mesophilic biogas reactors in South Korea. From 110 GB of raw DNA sequences, 401 metagenome-assembled genomes (MAGs) were created, which include 42,301 annotated genes. Of these, 187 MAGs (46.7%) were classified as high-quality based on Minimum Information about Metagenome-Assembled Genome (MIMAG) standards. The data presented here contribute to a broader understanding of biogas-specific microbial communities and offers a significant resource for future studies and advancements in sustainable biogas production.

Background & Summary

Anaerobic digestion (AD), which is essential for meeting sustainable energy needs, is a crucial process for generating methane in engineered bioreactors, thereby reducing the reliance on fossil fuels¹. AD converts organic waste, agricultural residues, and renewable primary products into energy and other valuable resources, aligning seamlessly with the circular economy concept². Methane, a prominent end product of AD's methanogenesis step, is vital for meeting societal energy demands and is intricately linked to the composition of the AD microbiome^{3–5}. Microbial metabolism in AD is thermodynamically influenced by the environmental parameters within the reactor⁶. The close relationship between these parameters offers distinct opportunities to improve process efficiency by selective cultivation or modification of microbial communities.

Traditional cultivation-based techniques are limited to studying microorganisms capable of laboratory growth; this limits their ability to fully characterize the diverse microbial communities active in AD. In contrast, metagenome sequencing offers a broader and more comprehensive analysis of all genetic material present in the environment. This approach provides deeper insights into the intricate microbial interactions and processes that drive AD systems⁷. Assembly and binning techniques for metagenome sequencing are crucial to elucidating the “black box” of microbial communities, leading to the discovery of previously unknown microorganisms in AD reactors⁸. This approach enables the generation of metagenome-assembled genomes (MAGs), providing valuable insights into microbial dynamics in AD processes^{8,9}. Geographic variations in AD microbiomes highlight the variability in energy production, waste degradation, and metagenomic composition across locations, making it essential to understand these fluctuations to customize AD procedures for specific regions and to maximize biogas production. However, very few studies have focused on global meta-analyses to systematically address location-dependent differences in AD metagenomes.

In this study, samples were collected from five full-scale biogas reactors located in the Ichon (I_A & I_M), Gunsan (G), Jungrang (J), and Anyang (A) regions to investigate microbial diversity and metabolic potential (Fig. 1). Detailed sample metadata and sequencing strategies are shown in Table S1.

¹Department of Chemical Engineering, Hanyang University, 222 Wangsimniro, Seongdong-gu, Seoul, 04763, Republic of Korea. ²Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94–96, 72076, Tübingen, Germany. ✉e-mail: biosang@hanyang.ac.kr



Fig. 1 Sampling sites of five full-scale biogas plants in South Korea. Detailed sample metadata can be found in Table S1.

Upon metagenomic sequencing and binning, 401 MAGs were reconstructed with completeness of $\geq 50\%$ and contamination of $\leq 10\%$. Among these, 74, 79, 80, 93, and 65 representative MAGs originated from the metagenomes I_A, I_M, G, J, and A, respectively. Furthermore, 187 MAGs (46.7%) were classified as high-quality MAGs, with completeness $\geq 90\%$ and contamination $\leq 5\%$, based on the Minimum Information about Metagenome-Assembled Genome (MIMAG) standards¹⁰. These MAGs were taxonomically assigned to two archaeal and 36 bacterial phyla based on the Genome Taxonomy Database (GTDB)^{11–15}, with a total of five archaeal and 182 bacterial MAGs. Archaeal MAGs were affiliated with Halobacteriota ($n = 3$) and Methanobacteriota ($n = 2$) phyla. Bacterial MAGs were primarily from Bacteroidota ($n = 85$), Bacillota_A ($n = 65$), Bacillota_G ($n = 34$), Cloacimonadota ($n = 11$), Bacillota_D ($n = 11$), Desulfobacterota ($n = 10$), Verrucomicrobiota ($n = 10$), Actinomycetota ($n = 8$), Chloroflexota ($n = 8$), Pseudomonadota ($n = 7$), Synergistota ($n = 7$), Acidobacteriota ($n = 5$), DTU030 ($n = 5$), Armatimonadota ($n = 4$), Atribacterota ($n = 4$), Hydrogenedentota ($n = 3$), Myxococcota ($n = 3$), Omnitrophota ($n = 2$), Fibrobacterota ($n = 2$), Riflebacteria ($n = 2$), Sumerlaeota ($n = 2$), UBP6 ($n = 1$), UBA1439 ($n = 1$), Latescibacterota ($n = 1$), Planctomycetota ($n = 1$), Bacillota_F ($n = 1$), Bacillota_E ($n = 1$), and Bacillota_H ($n = 1$). The detailed taxonomic assignments of all MAGs are provided in Table S2 and Fig. 2.

Genes were identified at the MAGs level to construct a community-level gene catalog. After gene calling and deduplication, 42,301 genes were identified. Their functions were annotated using the Hidden Markov Model for Emission and Recognition (HMMER)^{16,17}. Genome sets containing high-quality MAGs were further annotated and distilled using Distilled and Refined Annotation of Metabolism (DRAM)¹⁸. The metabolic profiles of these high-quality MAGs annotated by DRAM with advanced parameters such as a bit score threshold of 60 and a reverse search bit score threshold of 350 are shown in Fig. 3.

Methods

Sample collection and preparation. In this study, samples were collected from five mesophilic biogas reactors operating within a temperature range of 35–40 °C, located across four regions in South Korea, identified by GPS coordinates (I_A = 37.335048, 127.413917; I_M = 37.335048, 127.413917; G = 35.97861, 126.71139; J = 37.557499, 127.057625; and A = 37.402233, 126.895660). The sampling period spanned from August 2022 to August 2023. Additional details regarding the sampling sites and substrates are presented in Fig. 1 and Table S1,

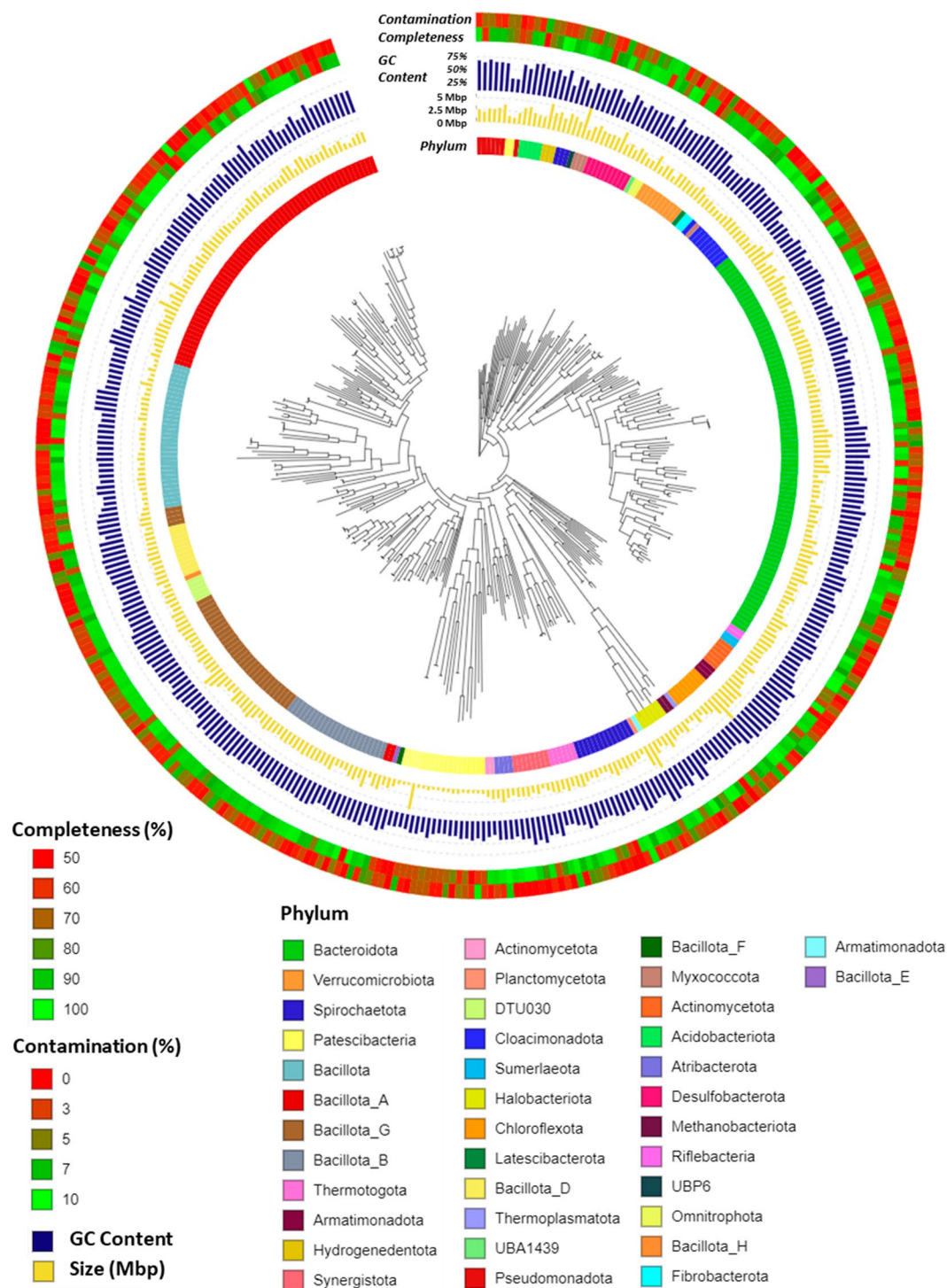


Fig. 2 Phylogenomics tree of MAGs reconstructed using full-scale biogas reactors in South Korea. 49 COG domains were used to build a maximum-likelihood phylogenomic tree. Detailed MAG taxonomy assignments, associated with completeness and contamination information, can be found in Table S2, and the list of 49 COD domain information can be seen in Table S3.

respectively. All the samples were collected during periods of stable operation. To maintain anaerobic conditions, the collected, unfiltered samples were stored in airtight bags and frozen at -80°C until DNA extraction.

DNA extraction, sequencing, and sequence quality assessment. The genomic DNA of the five biogas reactor samples was extracted using a Fast DNATM Spin Kit for soil (MP Biomedicals, LLC, Solon, OH, USA) according to the manufacturer's instructions.

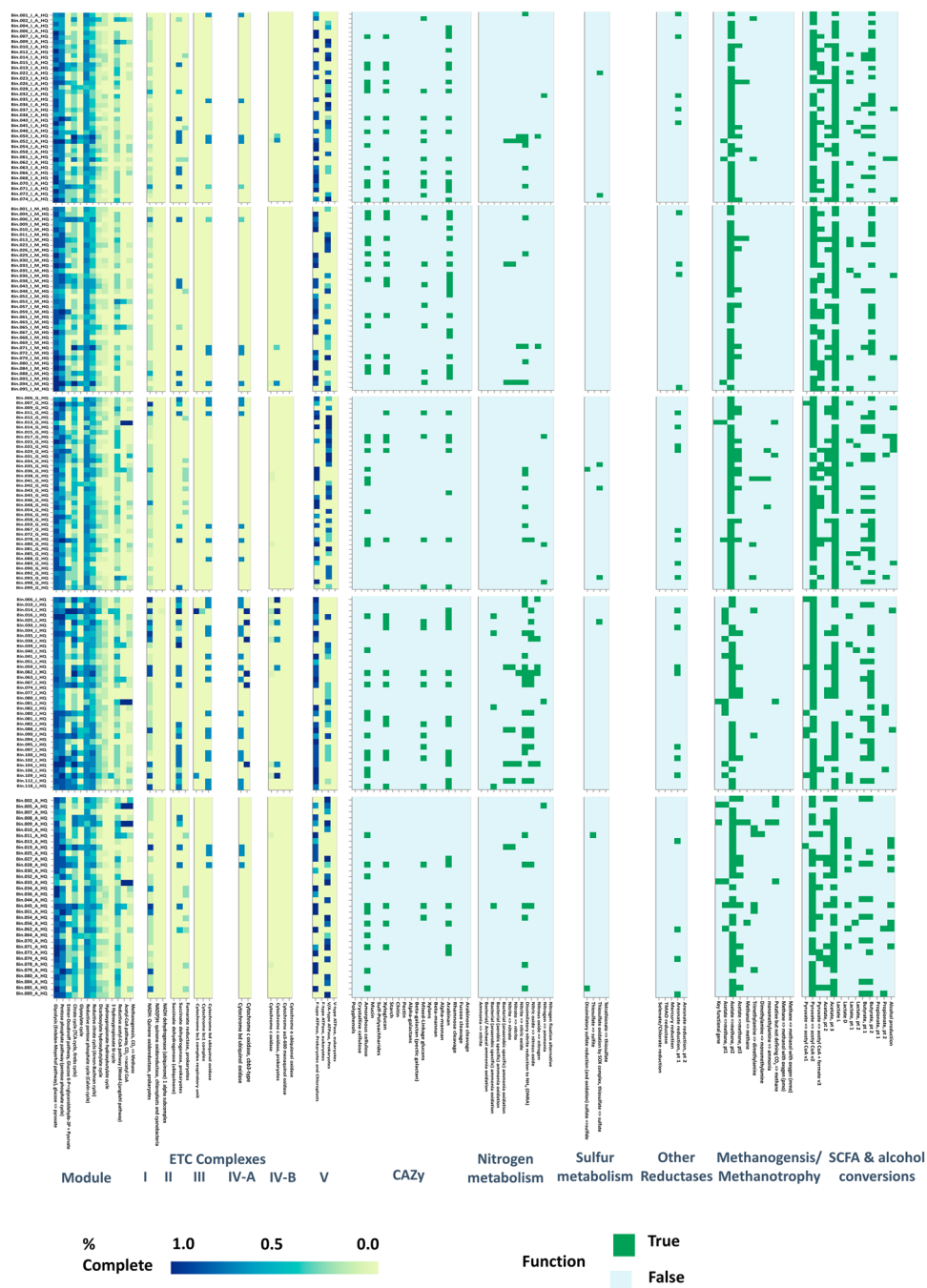


Fig. 3 DRAM annotation of metabolic profiles for high-quality MAGs across all genome sets. Detailed annotations, gene sequences, metabolite products, and summary are submitted in figshare⁵⁶ repository.

The concentration of the extracted double-stranded DNA was determined using an Infinite M200 PRO microplate reader (Tecan Austria GmbH, Grödig, Austria). The extracted DNA was then sent to Macrogen Inc., Seoul, South Korea for DNA sequencing using a NovaSeq 6000 platform with a read length of 151 base pairs (bp) for paired-end sequencing (X 2) in high-output mode (H). Raw reads were trimmed using the Joint Genome Institute (JGI) RQCFilter pipeline (BBTools v38.22)¹⁹ with specific parameters (library option = frag and trim the low-quality read ends to remove low-quality bases; minq = Trim both ends). The quality of the trimmed reads from the JGI RQCFilter pipeline (BBTools v38.22)¹⁹ was checked using FastQC v0.12.1²⁰, and additional contaminating reads were removed using Trimmomatic v0.36²¹.

Metagenomic assembly. Clean reads of each sample were assembled using different assemblers: metaSPAdes - v3.15.3²², MEGAHIT v1.2.9 (x2 using different parameters; Meta-large and Meta-sensitive)²³, and IDBA-UD - v1.1.3²⁴ with a minimum contig length of 300 bp. The assemblies were evaluated using

QUAST-v4.4^{25,26} and Compare Assembled Contig Distributions – v1.1.2²⁷. Based on these results, MEGAHIT v1.2.9 - Meta-sensitive was selected as the best assembler for all samples, except for sample A, which was better suited for the MEGAHIT v1.2.9 - Meta-large assembler.

MAGs generation, quality assessment, and annotation. The contigs from the MEGAHIT v1.2.9 assembler were grouped into bins with the following binners: MaxBin2 - v2.2.4^{28,29} and CONCOCT - v1.1³⁰. To enhance the quality of the resulting binned contigs, optimization was performed using DAS Tool - v1.1.2^{31–34}, resulting in bins with $\geq 50\%$ completeness and $\leq 10\%$ contamination. Subsequently, all bins were subjected to quality assessment and filtering using CheckM - v1.0.18³⁵ under the MIMAG threshold¹⁰, categorizing them as high- and medium-quality bins. Furthermore, filtered bins were assigned to unique MAGs using Extract Bins as Assemblies from BinnedContigs - v1.0.2²⁷ and turned into annotated genomes using the Rapid Annotation Subsystem Technology (RAST) Pipeline, RASTtk - v1.073^{36–44}. The taxonomy of each MAG-containing genome set was annotated using GTDB-Tk - v2.3.2^{11–15,45,46} with version r214. Additionally, the metabolic roles of MAGs were functionally annotated using HMMER^{16,17} and DRAM¹⁸ by searching for functional marker genes.

Phylogenomics. A total of 401 bacterial and archaeal MAGs were obtained using GTDB-Tk - v2.3.2^{11–15,45,46}. The phylogenomics tree (Fig. 2) was constructed using the Insert Genome into SpeciesTree - v2.2.0⁴⁷, utilizing 49 core clusters of orthologous group (COG) gene families (see Table S3). The genomes were inserted into curated multiple sequence alignments (MSAs) for each COG family using Muscle⁴⁸. The curated alignments were trimmed using GBLOCKS⁴⁹ to remove the poorly aligned sections. Subsequently, the MSAs were concatenated and used for the tree construction. The final tree was visualized and annotated using the iTOL web tool.

Data Records

All sequencing products associated with this project can be found under National Center for Biotechnology Information (NCBI) BioProject (ID: PRJNA1104272). Raw reads from metagenomic sequencing were deposited in the NCBI with specific Sequence Read Archive (SRA) numbers⁵⁰. The assemblies were deposited in GenBank^{51–55} under the same Bioproject submission. The NCBI SRA and GenBank accession numbers for each sample are listed in Table S1. MAGs, the functional annotation of non-redundant genes identified using HMMER, and all metabolic profiles of high-quality MAGs annotated using DRAM are deposited in Figshare repository⁵⁶.

Technical Validation

All raw data processing steps, software, and parameters used in this study are described in the “Methods” section. The quality scores for the raw and clean reads of the five metagenomes were assessed using FastQC v0.12.1²⁰. The best assembler for each sample was selected using QUAST-v4.4^{25,26} and Compare Assembled Contig Distributions – v1.1.2²⁷. The quality of the MAGs was assessed using CheckM- v1.0.18³⁵.

Code availability

All versions of the third-party software and scripts used in this study are described and referenced in the Methods section.

Received: 11 September 2024; Accepted: 17 December 2024;

Published online: 03 February 2025

References

- Venkateshwaran, K., Bocher, B., Maki, J. & Zitomer, D. Relating anaerobic digestion microbial community and process function: supplementary issue: water microbiology. *Microbiol. insights* **8**, MBI. S33593 (2015).
- Campanaro, S. *et al.* New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. *Biotechnol. for biofuels* **13**, 1–18 (2020).
- Campanaro, S., Treu, L., Kougias, P. G., Luo, G. & Angelidaki, I. Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water res.* **140**, 123–134 (2018).
- Theuerl, S., Klang, J., Heiermann, M. & De Vrieze, J. Marker microbiome clusters are determined by operational parameters and specific key taxa combinations in anaerobic digestion. *Bioresour. technol.* **263**, 128–135 (2018).
- Ziels, R. M. *et al.* Microbial rRNA gene expression and co-occurrence profiles associate with biokinetics and elemental composition in full-scale anaerobic digesters. *Microbial biotechnol.* **11**, 694–709 (2018).
- Oh, S. T. & Martin, A. D. Thermodynamic equilibrium model in anaerobic digestion process. *Biochem. J. Eng.* **34**, 256–266 (2007).
- Lagkouvardos, I., Overmann, J. & Clavel, T. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut microbes* **8**, 493–503 (2017).
- Alivisatos, A. P. *et al.* A unified initiative to harness Earth's microbiomes. *Science* **350**, 507–508 (2015).
- Vanwonterghem, I., Jensen, P. D., Ho, D. P., Batstone, D. J. & Tyson, G. W. Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Current opinion in biotechnol.* **27**, 55–64 (2014).
- Bowers, R. M. *et al.* Corrigendum: Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnol.* **36** (2017).
- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
- Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids res.* **50**, D785–D794 (2022).
- Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnol.* **36**, 996–1004 (2018).
- Rinke, C. *et al.* A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946–959 (2021).
- Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic acids res.* **46**, W95–W101 (2018).
- Eddy, S. R. Accelerated profile HMM searches. *PLoS computational biol.* **7**, e1002195 (2011).

18. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic acids res.* **48**, 8883–8900 (2020).
19. Bushnell, B., Rood, J. & Singer, E. BBMerge—accurate paired shotgun read merging via overlap. *PloS one* **12**, e0185056 (2017).
20. Bioinformatics group at the Babraham Institute. FastQC source, version 0.12.1. *Babraham Bioinformatics* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2023).
21. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
22. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome res.* **27**, 824–834 (2017).
23. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
24. Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
25. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
26. Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V. & Gurevich, A. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics* **32**, 3321–3323 (2016).
27. Arkin, A. P. *et al.* KBase: the United States department of energy systems biology knowledgebase. *Nature biotechnol.* **36**, 566–569 (2018).
28. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
29. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 1–18 (2014).
30. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature methods* **11**, 1144–1146 (2014).
31. Sieber, C. M. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature microbiol.* **3**, 836–843 (2018).
32. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**, 1–11 (2010).
33. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1–9 (2009).
34. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).
35. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome res.* **25**, 1043–1055 (2015).
36. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC genomics* **9**, 1–15 (2008).
37. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids res.* **42**, D206–D214 (2014).
38. Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports* **5**, 1–6 (2015).
39. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656–664 (2002).
40. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids res.* **25**, 3389–3402 (1997).
41. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids res.* **25**, 955–964 (1997).
42. Cobucci-Ponzano, B., Rossi, M. & Moracci, M. Translational recoding in archaea. *Extremophiles* **16**, 793–803 (2012).
43. Meyer, F., Overbeek, R. & Rodriguez, A. FIGfams: yet another set of protein families. *Nucleic acids res.* **37**, 6643–6654 (2009).
44. Croucher, N. J., Vernikos, G. S., Parkhill, J. & Bentley, S. D. Identification, variation and transcription of pneumococcal repeat sequences. *BMC genomics* **12**, 1–13 (2011).
45. Chivian, D. *et al.* Metagenome-assembled genome extraction and analysis from microbiomes using KBase. *Nature Protocols* **18**, 208–238 (2023).
46. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* **11**, 1–16 (2010).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).
48. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids res.* **32**, 1792–1797 (2004).
49. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biol.* **56**, 564–577 (2007).
50. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP505515> (2024).
51. Sang, B. I. & Prabhakaran, D. Metagenome sequencing assembly of Ichon_A, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBEFFB0000000000> (2024).
52. Sang, B. I. & Prabhakaran, D. Metagenome sequencing assembly of Ichon_M, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBEFFC0000000000> (2024).
53. Sang, B. I. & Prabhakaran, D. Metagenome sequencing assembly of Gunsan, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBEFFD0000000000> (2024).
54. Sang, B. I. & Prabhakaran, D. Metagenome sequencing assembly of Jungrang, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBEFFE0000000000> (2024).
55. Sang, B. I. & Prabhakaran, D. Metagenome sequencing assembly of Anyang, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBEFFF0000000000> (2024).
56. Sang, B. I. & Prabhakaran, D. Representative Metagenomes of Mesophilic Biogas Reactor Across South Korea. *figshare* <https://doi.org/10.6084/m9.figshare.26962630> (2024).

Acknowledgements

This work was supported by Korea Environment Industry and Technology Institute (KEITI) through Development of Demonstration technology for energy conversion using waste resources Program, funded by Korea Ministry of Environment (MOE; 2022003480001). This research was supported by Plastic Free Specialized Graduate Program through the Korea Environmental Industry & Technology Institute(KEITI) funded by the Ministry of Environment (MOE).

Author contributions

G.W. conducted field sampling, D.P. conducted DNA extraction and metagenomic data analysis, produced all figures, and wrote the draft under the supervision of B.I. H.K. initiated the study, S.I. revised the draft, B.I. supervised the study and revised the draft, All authors reviewed and contributed to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-024-04315-8>.

Correspondence and requests for materials should be addressed to B.-I.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025