Supplemental information for

Genome enrichment of rare and unknown species from complicated microbiome by nanopore selective sequencing

Yuhong Sun [a, #], Zhanwen Cheng [a, #], Xiang Li [a,b,c], Qing Yang [a],
Bixi Zhao [a], Ziqi Wu [a], Yu Xia [a, b, c *]

[a] School of Environmental Science and Engineering, College of Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[b] State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[c] Guangdong Provincial Key Laboratory of Soil and Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China

# These authors contributed equally to this work
*Corresponding author:
Yu Xia
Address: School of Environmental Science and Engineering, College of Engineering, Southern University of Science and Technology, Shenzhen 518055, China
E-mail: xiay@sustech.edu.cn

Supplemental Items:

- ## Supplemental Text:

  Detailed information on community analysis of the thermophilic anaerobic digester (TAD) community, calculation of the abundance of MAGs, and metabolic capacities of *Verstraetearchaeota* and *Bathyarchaeota* phylum in the TAD community.

  Supplemental_Text_S1. 16S rRNA gene amplicon and community analysis of TAD community.

  Supplemental_Text_S2. Calculation of the abundance and assessment of the quality of MAG

  Supplemental_Text_S3. Versatile metabolic capacities of *Verstraetearchaeota* and *Bathyarchaeota* phylum in TAD community.

  Supplemental_Text_S4. Integration tests for the code.

- ## Supplemental Figures：

  Supplemental_Fig_S1. Selective sequencing report of mock community.

  Supplemental_Fig_S2. Bar plot of reads number of the seven microbial species.

  Supplemental_Fig_S3. community structure of the thermophilic anaerobic digester (TAD) community.

  Supplemental_Fig_S4. Rarefaction analysis of nanopore sequencing data.

  Supplemental_Fig_S5. Selective sequencing report of TAD community.

  Supplemental_Fig_S6. Read-length histograms in the TAD community.

  Supplemental_Fig_S7. The number of sequencing channels over the course of the sequencing run in the TAD community.

  Supplemental_Fig_S8. The quality and quantity of bins obtained for contigs of different lengths.

  Supplemental_Fig_S9. Venn diagram of the number of >1Mbp contigs assembled from canu, unicycler, and metaflye, respectively.

  Supplemental_Fig_S10. A phylogenetic tree was constructed from 57 HQ genomes derived from the TAD community and reference genomes.

  Supplemental_Fig_S11. Taxa detected by normal sequencing and the total reads.

  Supplemental_Fig_S12. Correlation between the number of genes of each genome and the Archael: Bacterial gene ratio.

  Supplemental_Fig_S13. Genomes comparison of MAG56, MAG57, and reference MAGs.

  Supplemental_Fig_S14. Report of human gut microbial community sequenced with metaRUpore.

Supplemental_Fig_S15. Correlation between the ejection rate and the time of normal sequencing of human gut microbiota.

Supplemental_Fig_S16. 3D density plots of t-SNE downscaling results of human gut microbiota.

Supplemental_Fig_S17. Phylogenetic tree of HQ-MAGs assembled from normal sequencing and metaRUpore data

Supplemental_Fig_S18. Human read retention ratio after selective nanopore sequencing.

- ## Supplemental Tables：

Supplemental_Table_S1.xlsx Information on the reference genome of the mock community

Supplemental_Table_S2.xlsx Integrated test of metaRUpore using nanopore reads.

Supplemental_Table_S3.xlsx Flow cells' yield

Supplemental_Table_S4.xlsx Basic statistics on the contigs assembled by Canu, metaFlye, and Unicycler

Supplemental_Table_S5.xlsx Information on the 57 HQ MAGs recovered from TAD community

Supplemental_Table_S6.xlsx Abundance of the 41 HQ MAGs retrived by metaRUpore

Supplemental_Table_S7. xlsx Information of the global genomes collection of the anaerobic reactor (AD) microbiome

Supplemental_Table_S8. xlsx Information of the genomes to build the gene flow figure of Bathyarchaeota

Supplemental_Table_S9.xlsx Previous reports about nanopore sequencing yield

Supplemental results

Supplemental_Text_S1:

16S rRNA gene amplicon and community analysis of TAD community

Primers 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 907R (5'-GGACTACNNGGGTTATCTAAT-3') were used to amplify the V4-V5 region of the 16S rRNA gene. The amplicon product was purified and then subject to shotgun library construction and Illumina high-throughput sequencing on the MiSeq at Novogene Co., Ltd. (Beijing, China) with PE250 strategy. Fastp (Chen et al., 2018) is used to perform quality control of the raw reads obtained from Illumina sequencing. Post-QC reads of 16S rRNA gene amplification were imported into the QIIME 1 (Caporaso et al., 2010) (Quantitative Insights in Microbiology) pipeline to merge pair-end sequences, extract barcodes, split samples, and remove amplification primers. USEARCHV11 was used to obtain OTUs with 97% similarity, then taxonomic assignments were achieved from the Greengenes database (McDonald et al., 2012) with rdp classifier.

Supplemental_Text_S2:

Calculation of the abundance and coverage

For the TAD community, abundance was calculated from both selective sequencing data and normal sequencing data, by mapping these data to the MAGs using minimap2 (Li 2018) (version 2.17) separately using the following flags -ax map-ont -t 40. We used samtools (Li et al. 2009) (version 1.11) to extract SAM file that matched each MAG individually. The abundance of each MAG is calculated by dividing the number of bases in all reads in this SAM file by the total number of bases selectively sequencing or normally sequencing, then normalizing by the size of the MAGs. Analogously, sorted BAM files were used in the calculation of the coverage of the MAGs. For the mock community, coverage was calculated by mapping the sequencing reads to the reference genomes and using the lengths of the reference genomes for normalization using the same method as above. The information on the reference genome is shown in Supplemental_Table_S1. Reference genomes of the seven bacterial strains were obtained by *de novo* assembly of individual nanopore sequencing of these strains using Unicycler. The reference genome sequence of the archaeal strain was downloaded from NCBI (NZ_CP039139.1).

Supplemental_Text_S3:

Versatile metabolic capacities of Verstraetearchaeota and Bathyarchaeota phylum in TAD community.

A complete genome of *Methanosauratus petracarbonis* affiliated with archaeal phylum *Verstraetearchaeota* was recovered as MAG57. The genome size of MAG57 is 1.5M and the GC content is 0.54. The abundance of *Methanosauratus petracarbonis* in TAD community was 0.075 %, which got doubled through selective sequencing, enabling successful retrieval of its entire genome.

MAG57 contains key genes for methane production (*mcrABG* and ancillary genes *mcrCD*) (Ermler et al. 1997) as well as genes for methylamine utilization (*mtaA, mtbA, mtmBC, mtbBC, mttC, mtrH*). The reduction of heterodisulfide (CoM-SS-CoB) to ferredoxin could be accomplished by the coupling of exergonic H2-dependent heterodisulfide reductase (*hdrB*) and F420-non-reducing hydrogenase (*mvhB*).

Meanwhile, the cytosolic complex of F420H2 dehydrogenases (*fpo*) consisted of consecutively located *fpoM, fpoL, fpoN, fpoK, fpoI, fpoH and fpoD*, can reoxidize the reduced ferredoxin while pumping protons across the cytoplasmic membrane to produce a proton gradient that drives the ATP synthesis via an archaeal-type ATP synthase. Additionally, *HdrD*, which is present in three copies in MAG57 and other *Verstraetearchaeota* genomes, may directly interact with the *fpo* complex and act as an energy-converting ferredoxin: heterodisulfide oxido-reductase. Furthermore, genes for hydrogenotrophic and acetoclastic methanogenesis pathways were absent in MAG57, a nearly complete genome of *Methanosauratus petracarbonis* species, consolidating the species' obligate H2-dependent methylotrophic methanogenesis capability (Vanwonterghem et al. 2016; Evans et al. 2019). Notably, while unusual for microorganisms involved in methane metabolism, the exit of adenosine diphosphate (ADP)-forming acetate synthetase (*Acd*) in MAG57 demonstrates that it can convert Acetyl-CoA to acetate, allowing for energy production via substrate-level phosphorylation (Vanwonterghem et al. 2016). Collectively, the coupling of obligate H2-dependent methylotrophic methanogensis and acetate-producing fermentative pathway of *Methanosauratus petracarbonis*'s genomic repertoire found in MAG57, reveals a unique ecological niche for carbon turnover and energy conservation in digestive systems rich of reduced methylated carbon compounds.

In this work, metaRUpore has boosted the abundance of *Bathyarchaeota* in TAD community, facilitating its genome recovery as MAG56. MAG56 appeared to be capable of utilizing sugars as a carbon source and generating acetyl-CoA via the

Embden–Meyerhof–Parnas (EMP) pathway (a nearly complete operon of *pfk, tpi, gap, pgk, apg, eno, ppc*) and pyruvate-ferredoxin oxidoreductase (*por*). ADP-forming acetyl-CoA synthase (*acd*) could then produce ATP and acetate, and this fermentative lifestyle was predicted to be the metabolic mode of several *mcr*-devoid *Bathyarchaeota* genomes (Evans et al. 2019; Lazar et al. 2016). Besides that, MAG56 possessed key genes for the autotrophic reductive acetyl-CoA (Wood–Ljungdahl, WL) pathway (*fwd, ftr, mch, cdh*), implying its ability to utilize tetrahydromethanopterin (H4MPT) as the C1-carrier for autotrophic carbon fixation, which is an energy-generating process prevalent in archaea (Feng et al. 2019). Additionally, the critical genes for lipid and benzoate degradation (*lcfB* and *acyP*) found in the MAG56 genome demonstrated its capacity to exploit lipid and benzoate as a source of carbon and energy. These core metabolic potentials of MAG56 are consistent with previous studies, consolidating *Bathyarchaeota*'s organoautotrophic life strategy capable of utilizing a diverse array of carbon sources (Yu et al. 2018; Feng et al. 2019).

Supplemental_Text_S4:

Integration tests for code of metaRUpore.

We have conducted the integration test on metaRUpore workflow using nanopore reads generated in the first one-hour normal sequencing of six different sample types including the TAD and human gut sample used for this study and permafrost top soil sample, a receiving water sample receiving effluent of a domestic wastewater treatment plant and activated sludge sample of another domestic wastewater treatment plant as well as an influent sample of a hospital sewage treatment plant from our previous published studies (Wu et al., 2022). The script has been tested using 10 threads on a local workstation (CPU: Xeon(R) 5220R 2.20 GHz × 24 cores with DDR4 64 Gb × 16 Memory). The results show that for all the samples tested, metaRUpore could finish the analysis within 5 minutes, which will allow for a quick start of the subsequent RU run with the determined reference and target dataset. Relative results are shown in Supplemental_Table_S1.

Reference：

Caporaso J G, Kuczynski J, Stombaugh J, Bittinger K, Bushman F D, Costello E K, Fierer N, Pena A G, Goodrich J K, Gordon J I, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Method*s **2010**: 7 (5), 335-336.

Chen S F, Zhou Y Q, Chen Y R, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018:** 34 (17), 884-890.

Ermler U, Grabarse W, Shima S, Goubeaud M, Thauer R K. 1997. Crystal Structure of Methyl –Coenzyme M Reductase : The Key Enzyme of Biological Methane Formation. *Science* **278**: 1457-1462.

Evans P N, Boyd J A, Leu A O, Woodcroft B J, Parks D H, Hugenholtz P, Tyson G W. 2019. An evolving view of methane metabolism in the Archaea. *Nat Rev Microbiol* **17**: 219–232

Feng X, Wang Y, Zubin R, Wang F. 2019. Core Metabolic Features and Hot Origin of Bathyarchaeota. *Engineering* **5**: 498–504.

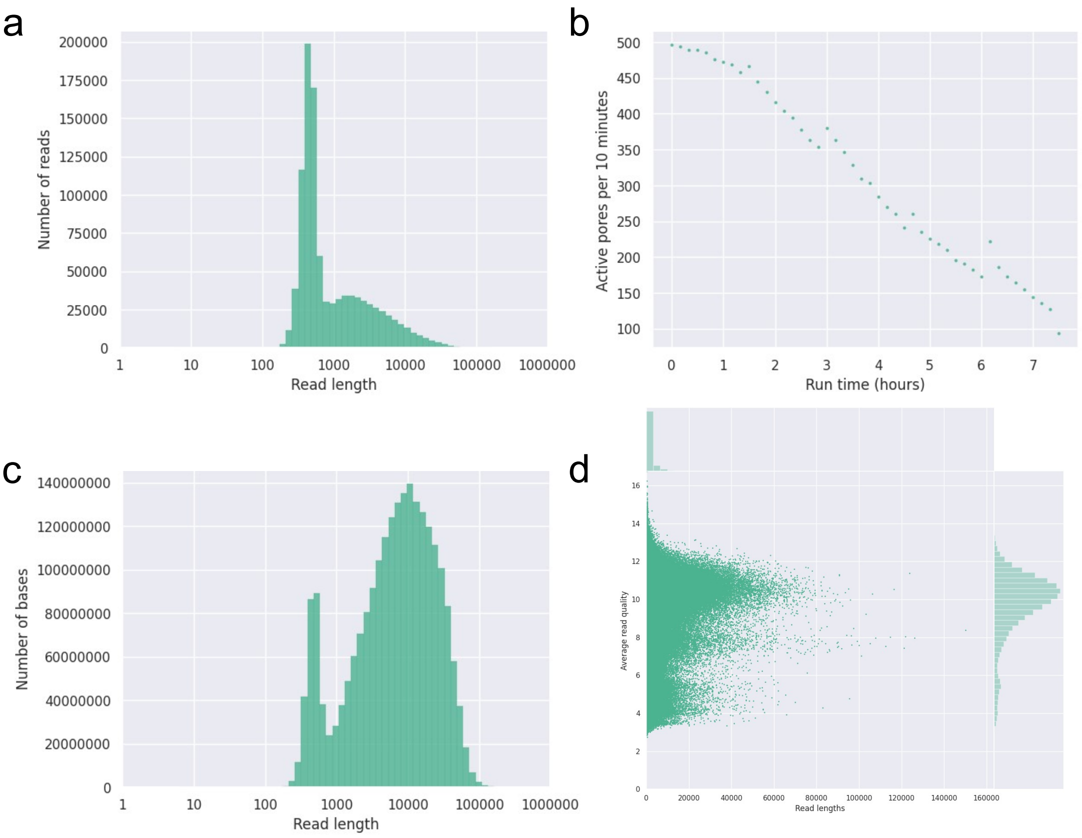
Lazar C S, Baker B J, Seitz K, Hyde A S, Dick G J, Hinrichs K U, Teske A P. 2016. Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environ Microbiol* **18**: 1200–1211.

McDonald D, Price M N, Goodrich J, Nawrocki E P, DeSantis T Z, Probst A, Andersen G L, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **2012**: 6 (3), 610-618.

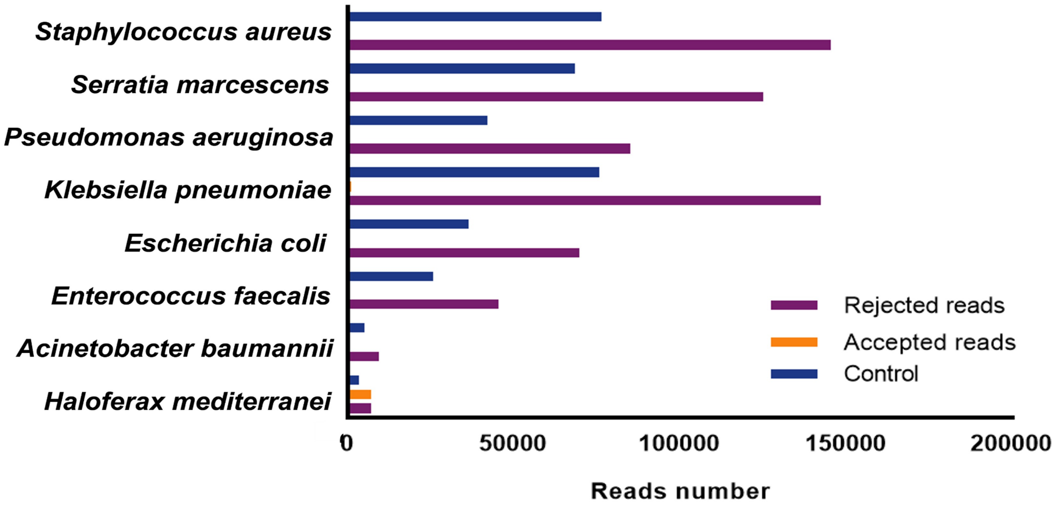
Vanwonterghem I, Evans P N, Parks D H, Jensen P D, Woodcroft B J, Hugenholtz P, Tyson G W. 2016. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat Microbiol* **1**: 1–9.

Yu T, Wu W, Liang W, Alexander M, Hinrichs K. 2018. Growth of sedimentary Bathyarchaeota on lignin as an energy source. *Proc Nat Acad Sci* **115**: 6022-6027.

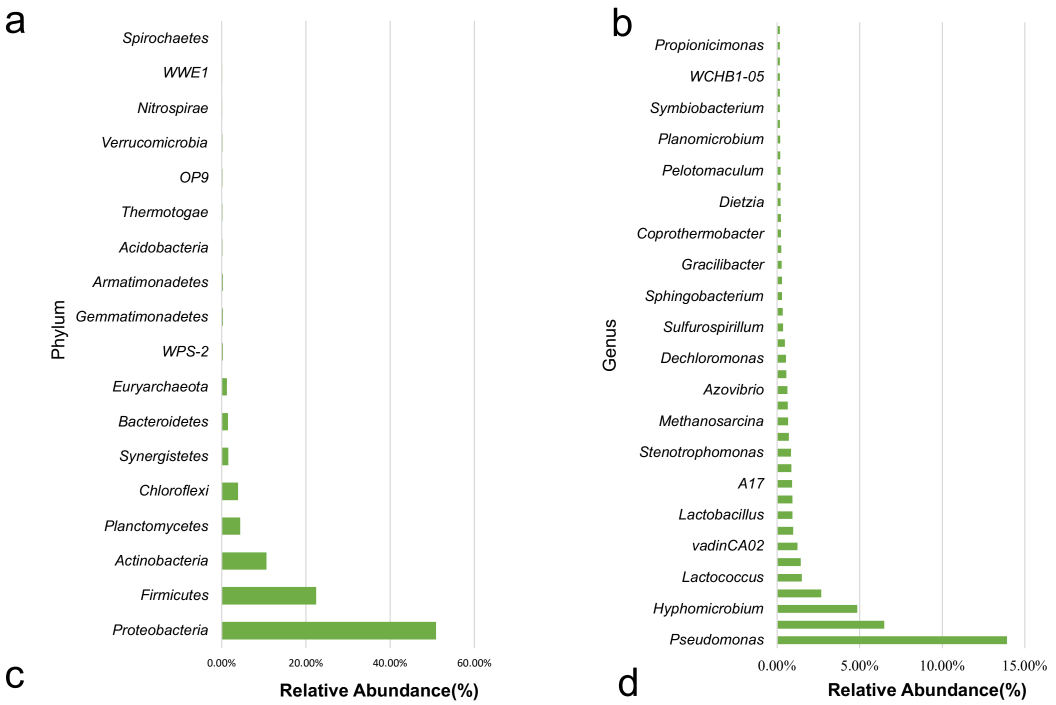
Supplemental Figures



Supplemental_Fig_S1. Selective sequencing report of mock community. a) Histogram of lengths after log transformation. b) Number of total active pores over time. c) Weighted histogram of read lengths after log transformation. d) Plot of read lengths versus average read quality.



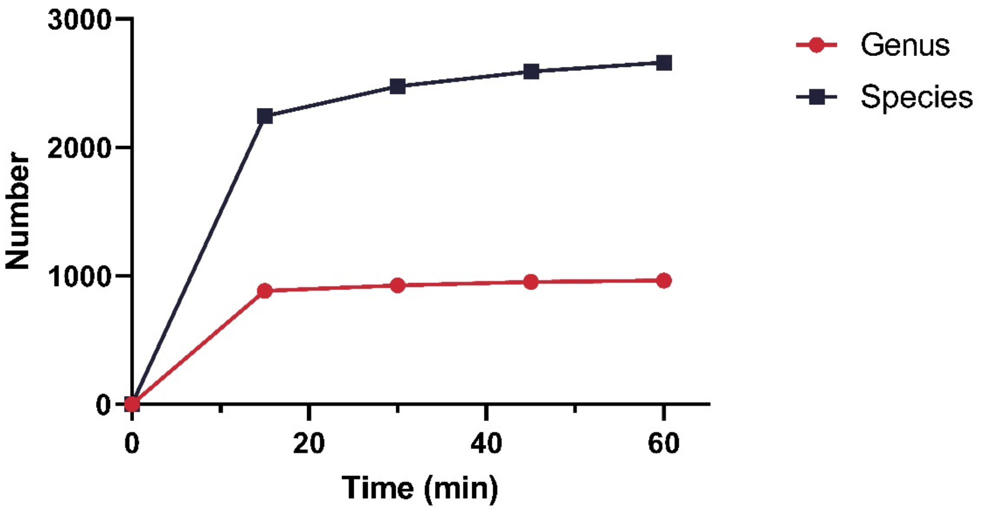
Supplemental_Fig_S2. Bar plot of reads number of mock community sequencing. The number of rejected and accepted reads of the RU channels and reads number in the control channels are respectively shown.



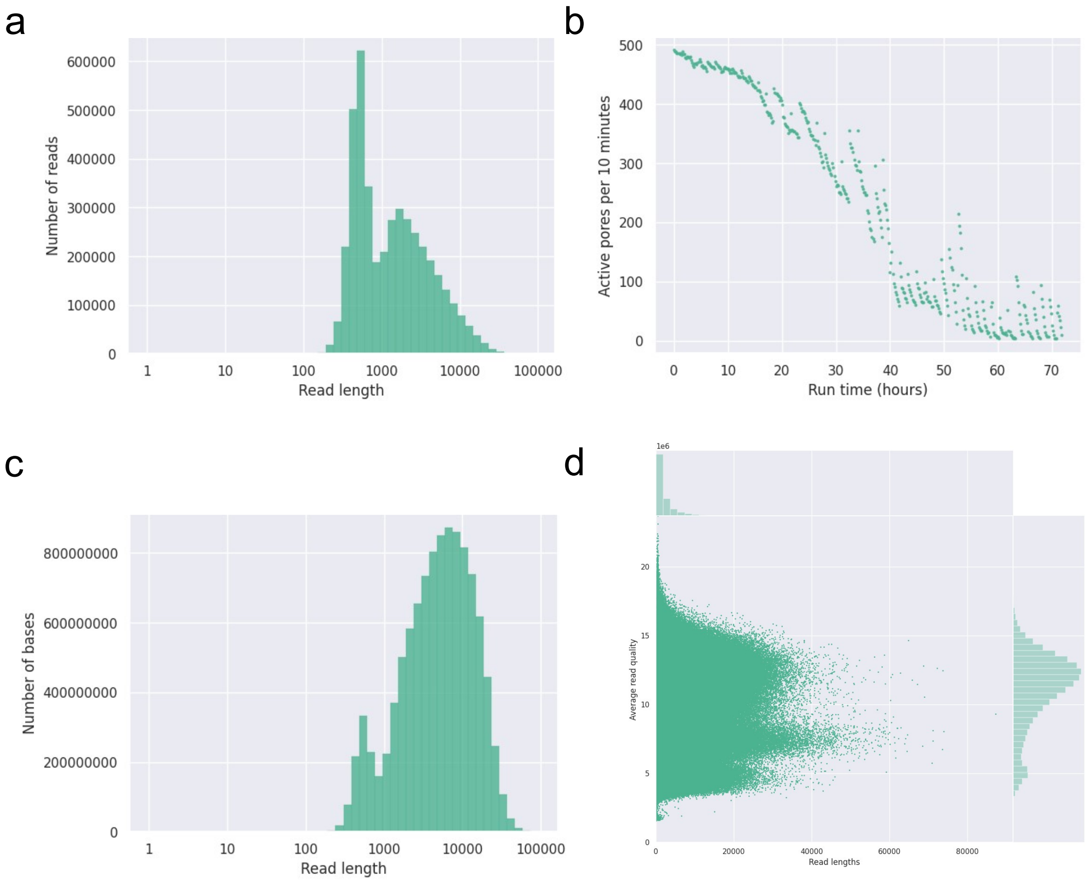
	Phylum	Class	Order	Family	Genus
Classified Ratio	99.73%	99.03%	94.34%	78.96%	46.80%

Sample	Chao1	Reads	Shannon
TAD	7545.83	237778	8.74

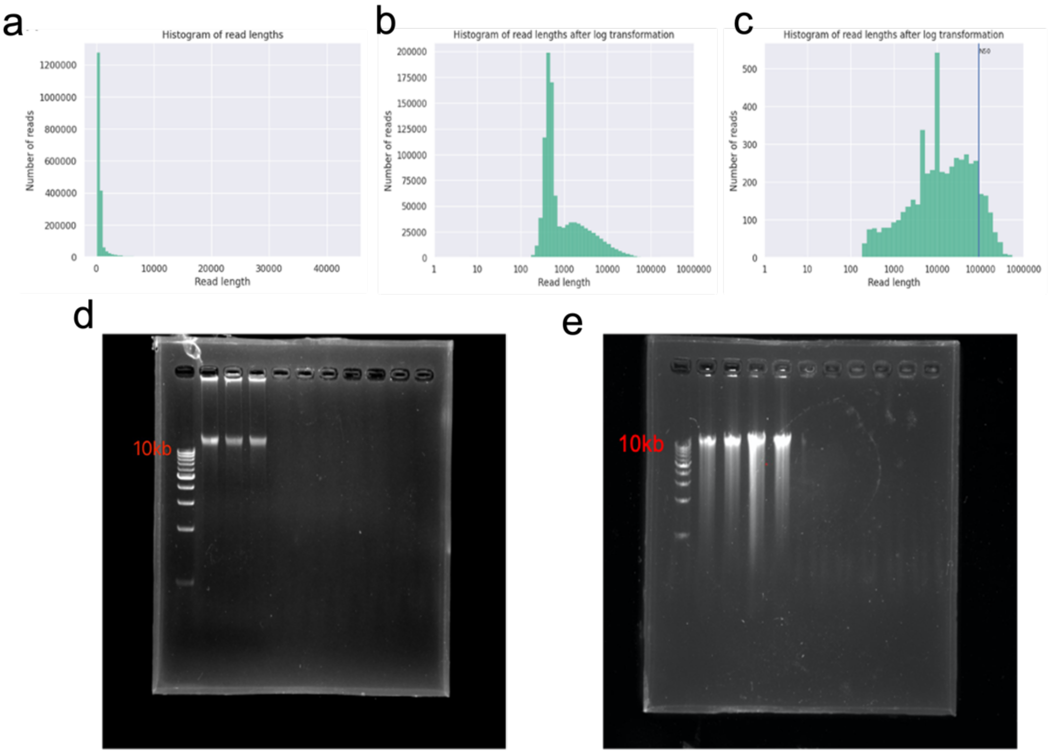
Supplemental_Fig_S3. community structure of the thermophilic anaerobic digester (TAD) community. a) Phylum, b) Genus level community structure of the TAD community. c) Classified ratio of each taxonomy level. d) Alpha diversity index based on metagenome extracted 16S rRNA.



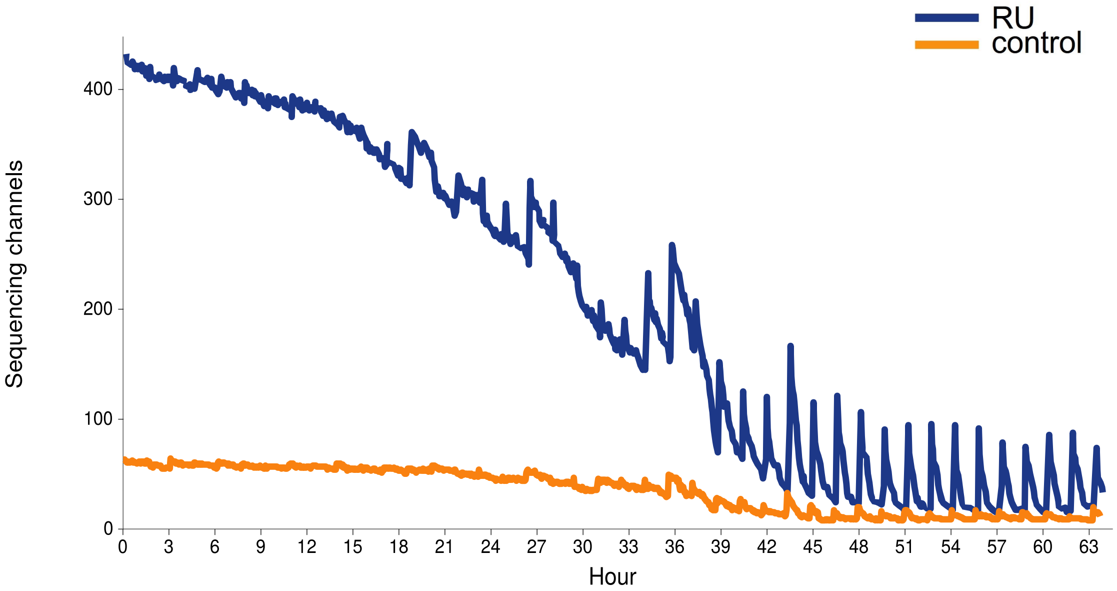
Supplemental_Fig_S4. Rarefaction analysis of nanopore sequencing data. The Y-axis is the number of species or genus annotated by Centrifuge. The curve is close to saturation at 60min.



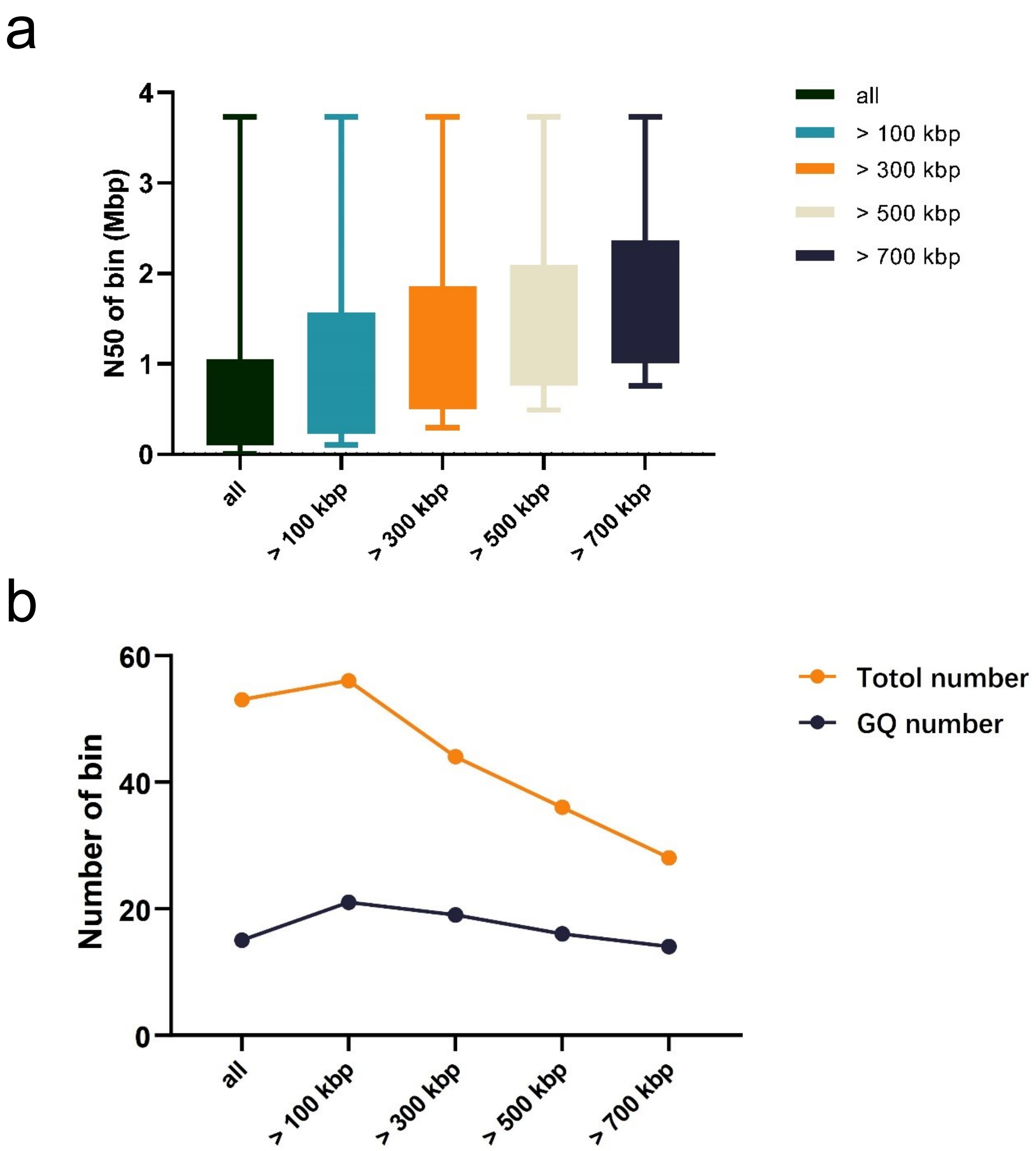
Supplemental_Fig_S5. Report of TAD community sequenced with metaRUpore. a) Histogram of lengths after log transformation. b) Number of total active pores over time. c) Weighted histogram of read lengths after log transformation. d) Plot of read lengths versus average read quality.



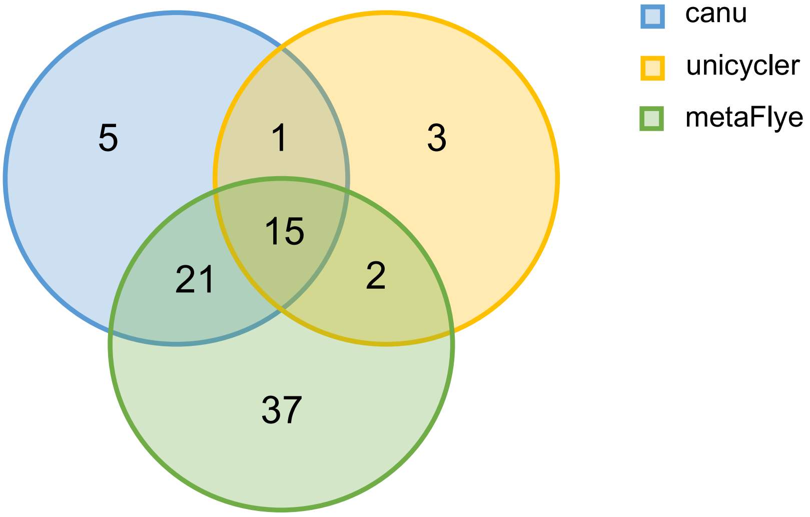
Supplemental_Fig_S6. Read-length histograms of a) rejected reads and b) total reads in RU runs as well as c) control runs in the TAD community. Gel image of d) 4 samples of the TAD community and e) the 3 samples of the human gut.



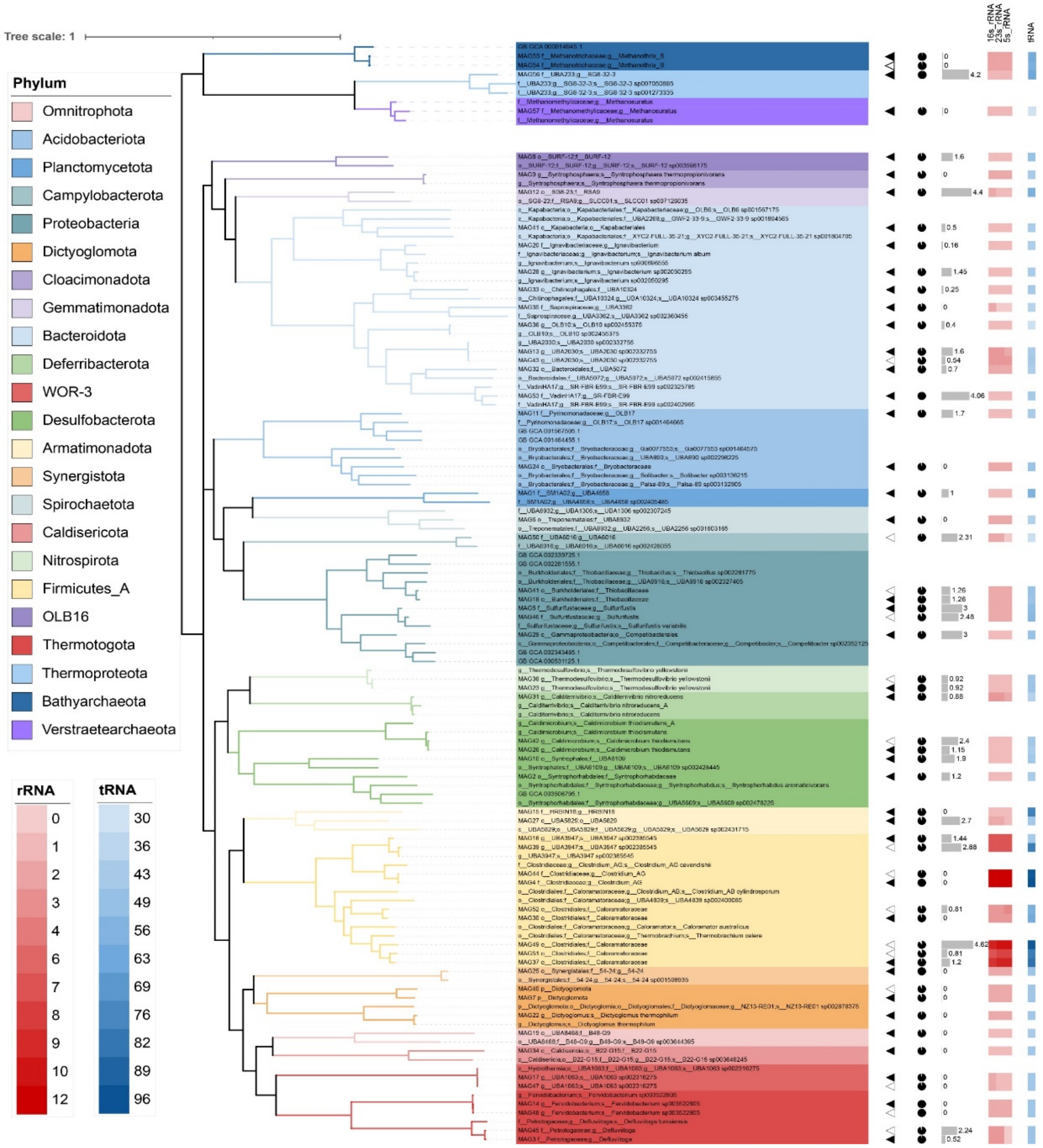
Supplemental_Fig_S7. The number of sequencing channels over the course of the sequencing run in TAD community. It shows that active pore loss speed of RU-channels was faster than that of the control channels by the slop of the line.



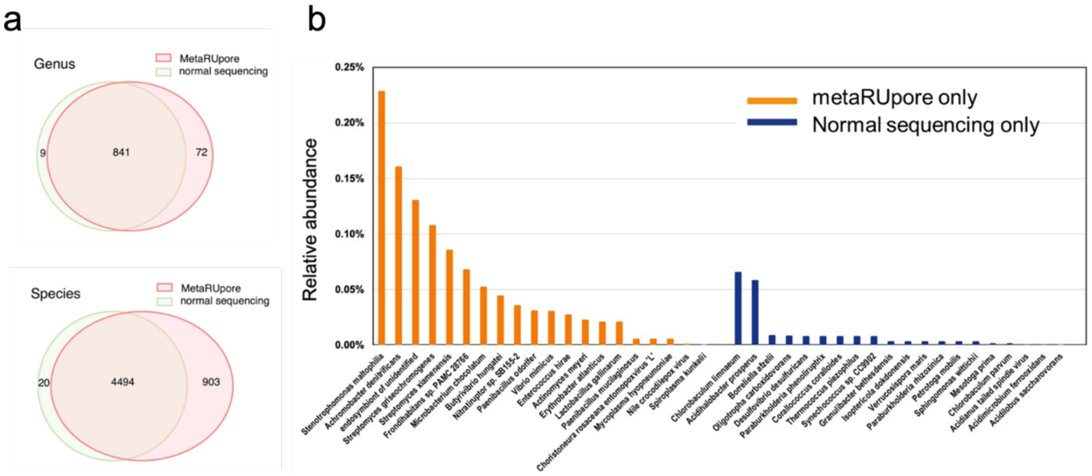
Supplemental_Fig_S8. The quality and quantity of bins obtained for contigs of different lengths. We grouped the contigs <1M into five categories: >700 kbp, >500 kbp, >300 kbp, >100 kbp, and all contigs and binned them separately. As a result, binning with >100kb contigs could achieve the greatest balance between quantity and quality of MAGs, so we finally chose 100kbp as a tradeoff for binning. a) N50 of the bin obtained from contigs of different length groups. b) Number or good-quality number of the bin obtained from contigs of different length groups. Good quality bins mean they have > 80% SCG-completenessaand < 5% contamination, with the potential to be corrected to high-quality bins.



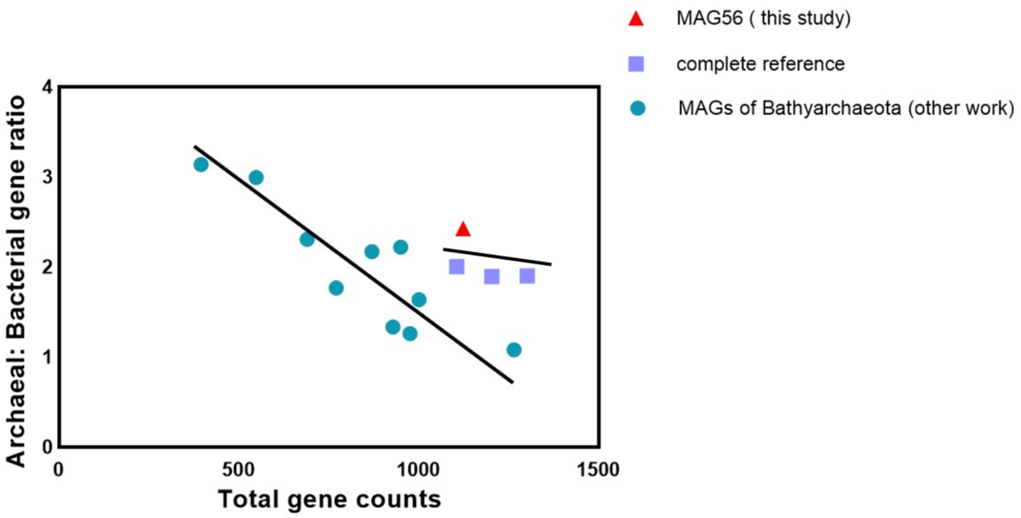
Supplemental_Fig_S9. Venn diagram of the number of >1Mbp contigs assembled from Canu, Unicycler, and metaFlye, respectively. We assembled the nanopore data with Canu, Unicycler, and metaFlye, respectively, and de-duplicated them by dRep with a relatedness threshold of ANI > 0.95. We found that the three tools produced duplicate >1 Mbp contigs, but each tool was able to assemble additional contigs.



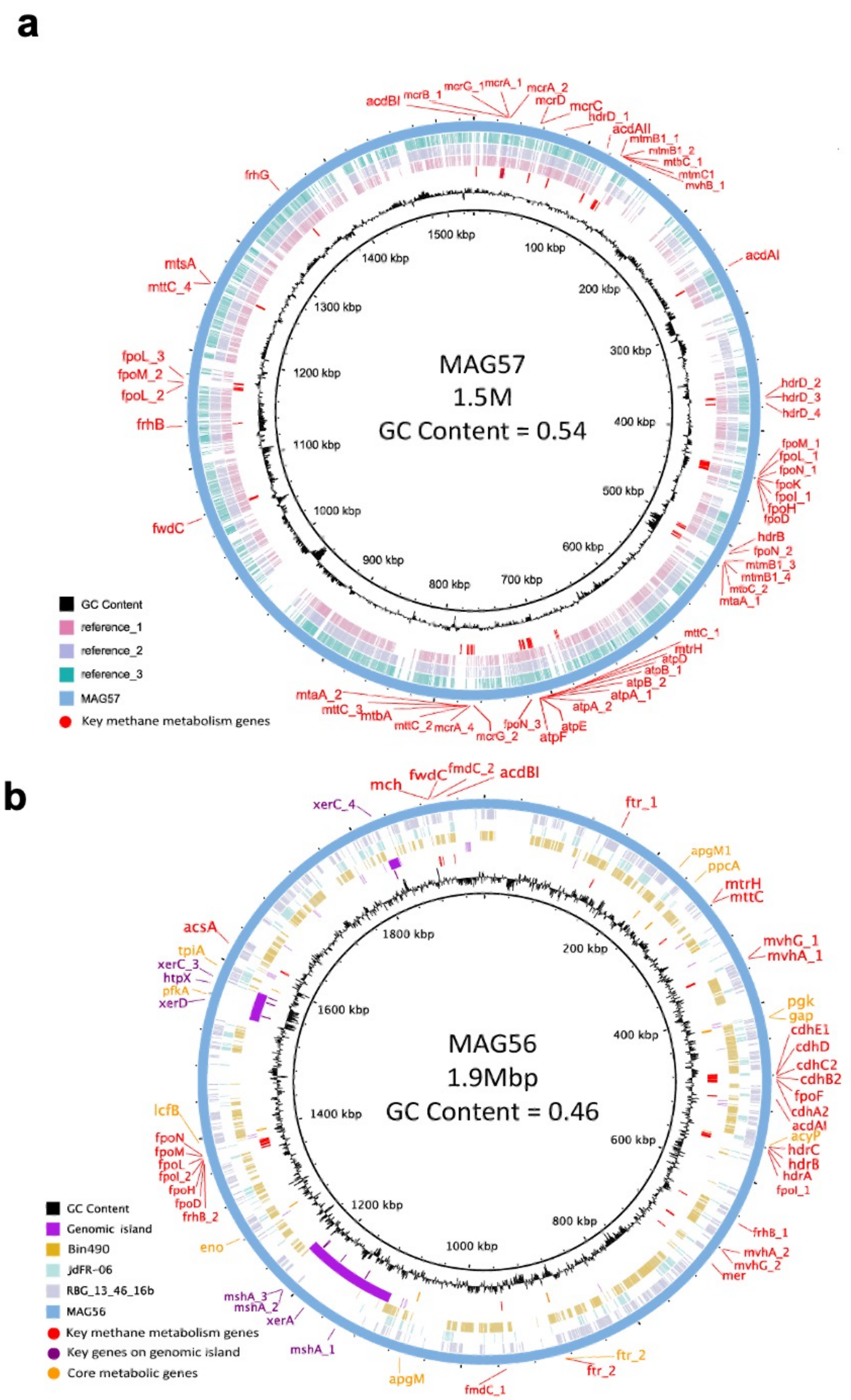
Supplemental_Fig_S10. A phylogenetic tree was constructed from 57 HQ genomes derived from the TAD community and reference genomes. The solid triangles represent the 41 MAGs assembled from the metaRUpore dataset and the hollow triangles represent the 16 MAGs assembled from the normal sequencing dataset. The different colored branches of the tree represent phyla, the pie chart represents genomic SCG-completeness and the bar chart represents genomic contamination. The copy number of 16S rRNA, 23S rRNA, and 5S rRNA is represented by the red heat map from left to right, while the copy number of tRNA is represented by the blue heat map.



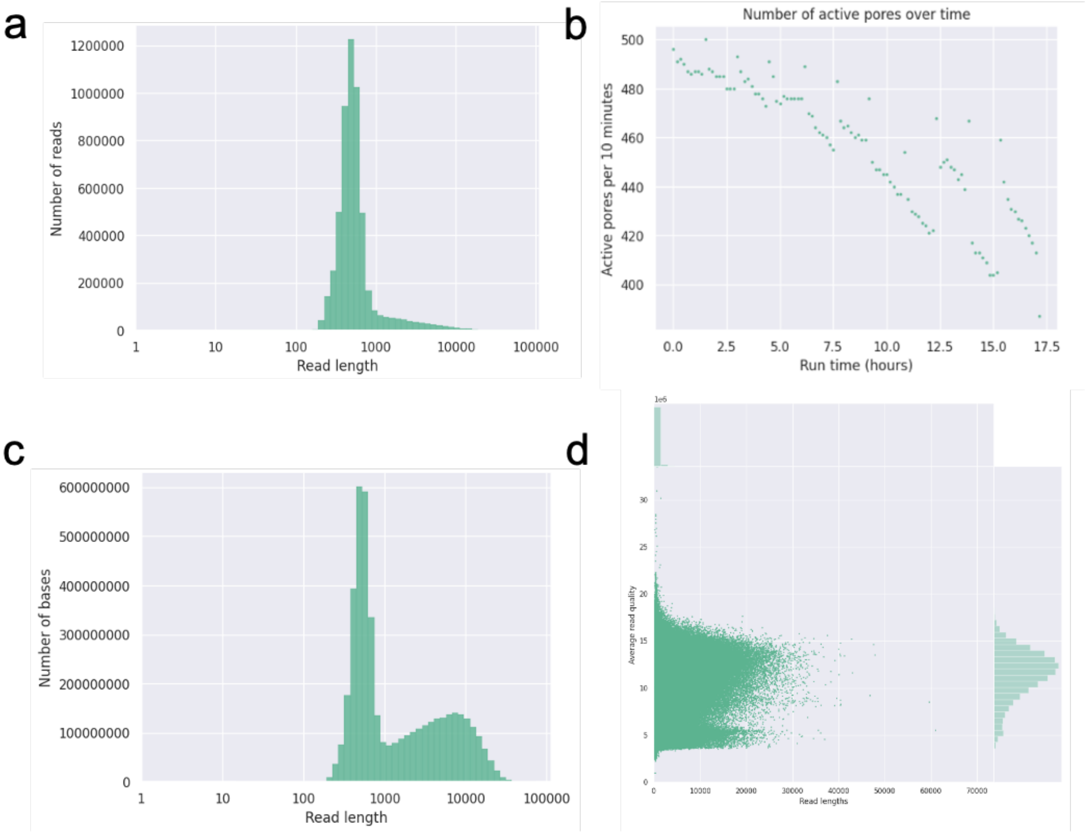
Supplemental_Fig_S11. a) Venn plot showing the number of taxa detected by normal sequencing dataset and the total reads (both ejected and received reads) of RU-channels. b) Bar plot showing the relative abundance of the 20 species detected only in normal sequencing dataset (blue bar) and the top 20 species detected only in the total reads dataset (Orange bar).



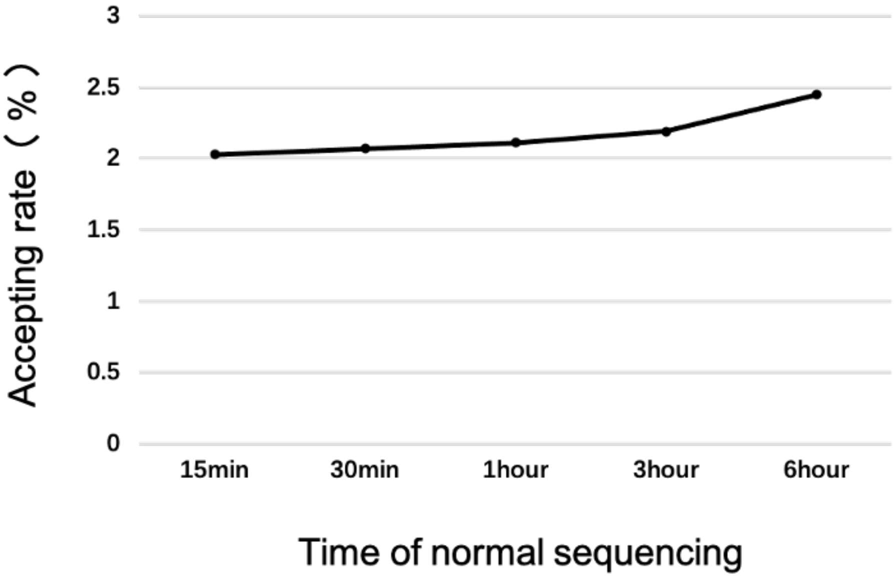
Supplemental_Fig_S12. Correlation between the number of genes of each genome and the Archaeal: Bacterial gene ratio.



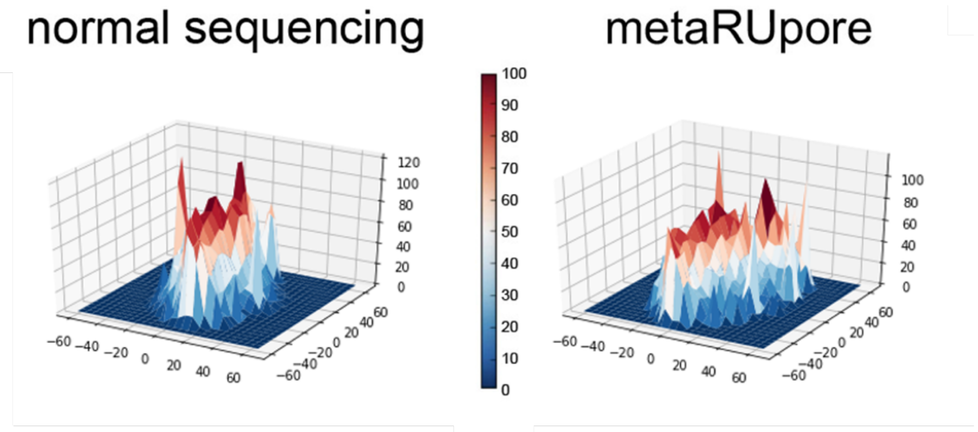
Supplemental_Fig_S13. a) Genomes comparison of MAG57 of the *Verstraetearchaeota* phylum and reference MAGs. The outermost ring stands for the circular genome of MAG57 reconstructed by metaRUpore. The second to fourth circles from the outside represent the MAGs of phylum *Verstraetearchaeota* reconstructed by short reads-only assembly method, which was mapped to MAG57. The two innermost circles from the outside to the inside indicated the key methane metabolism predicted genes by Prokka and GC content, respectively. b) Genomes comparison of MAG56 of the *Bathyarchaeota* phylum and reference MAGs. The outermost ring stands for the circular genome of MAG56 of the reconstructed by metaRUpore. The second to fourth circles from the outside represent the MAG, which was mapped to MAG56. The fifth purple circle represents the genomic island. The sixth circle from the outside indicated the key methane metabolism genes (red), key genes on a genomic island (purple) and core metabolic genes (orange) predicted by Prokka and the innermost circles represent GC content.



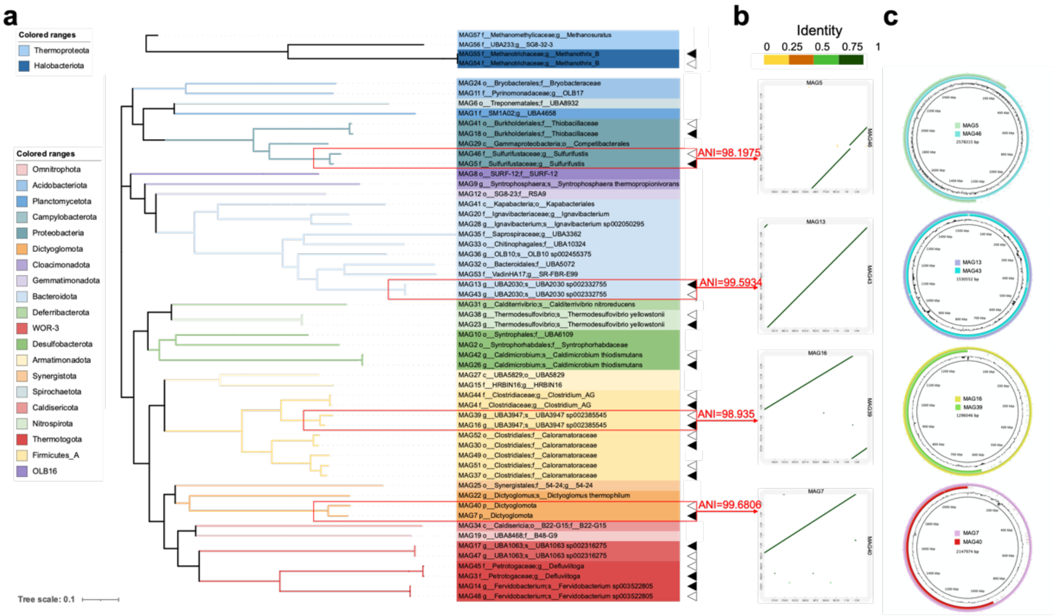
Supplemental_Fig_S14. Report of human gut microbial community sequenced with metaRUpore. a) Histogram of lengths after log transformation. b) Number of total active pores over time. c) Weighted histogram of read lengths after log transformation. d) Plot of read lengths versus average read quality.



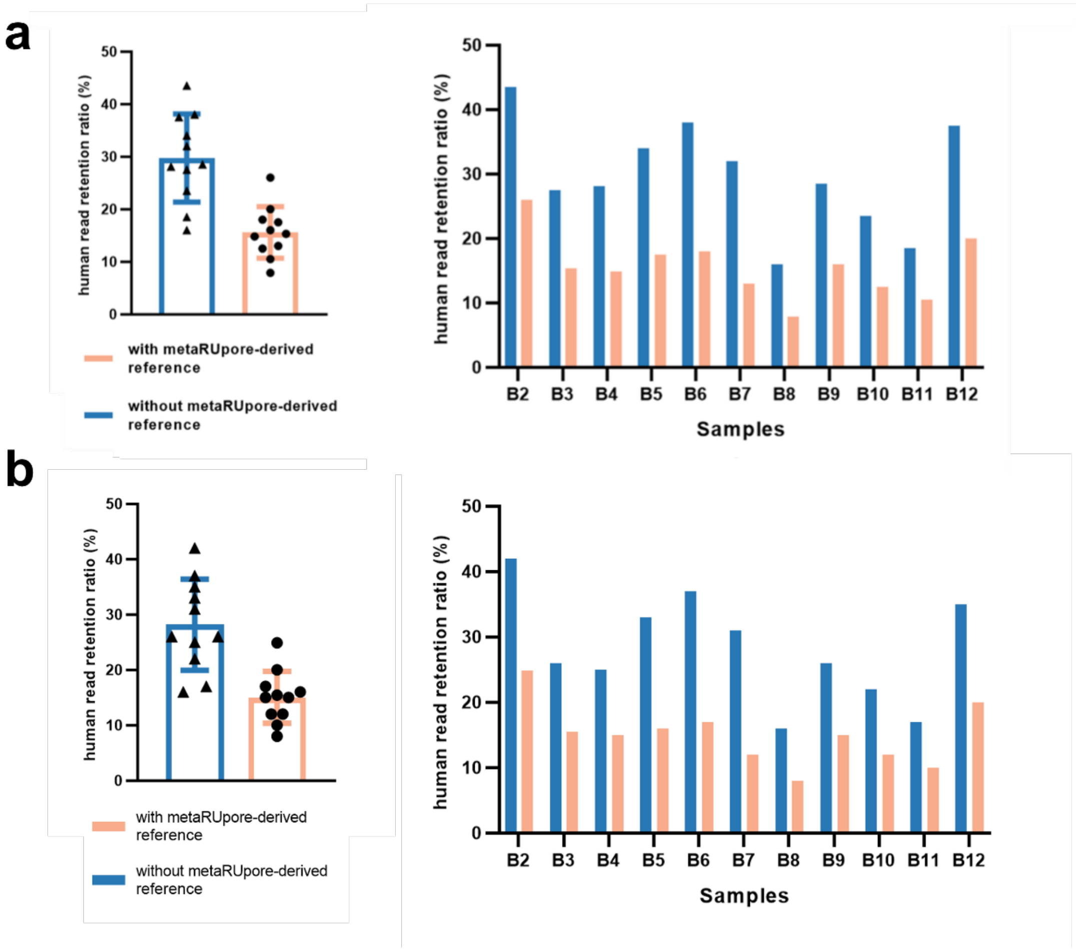
Supplemental_Fig_S15. Correlation between the accepting rate and the time of normal sequencing.



Supplemental_Fig_S16. 3D density plots of t-SNE downscaling results for normal sequencing datasets and selective sequencing datasets by metaRUpore at four base frequencies, showing that metaRUpore renders the human hut community structure homogenous.



Supplemental_Fig_S17. a) Evolutionary tree of high-quality MAGs assembled from normal sequencing and metaRUpore. The hollow triangle represents MAGs assembled from normal sequencing data, while the solid triangle represents MAGs assembled from metaRUpore data. b) The dot plot and c) ring plot demonstrate that while the metaRUpore-recovered genome was evidently larger in size, the regions that can be aligned between the two genomes are highly consistent.



Supplemental_Fig_S18. Human reads retention ratio after selective nanopore sequencing with or without metaRUpore-derived reference set. a) and b): ReadUntil results using GRCh38 and hg19 assembly of the human genome as targets for ejection, respectively.