Data in Brief 56 (2024) 110829

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Dataset of 313 metagenome-assemble genomes from streamer hot spring water

Jia Hao Tan[a], Kok Jun Liew[a,b], Kian Mau Goh[a,*]

[a] Faculty of Science, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia
[b] Codon Genomics, 42300 Seri Kembangan, Selangor, Malaysia

ARTICLE INFO

*Article history:*
Received 1 July 2024
Revised 17 July 2024
Accepted 6 August 2024
Available online 10 August 2024

Dataset link: Dataset of 313 metagenome-assemble genomes from streamer hot spring water (Original data)

*Keywords:*
Geothermal spring
Hyperthermophiles
MAGs
Microbiota
Thermophiles

## ABSTRACT

This data report presents prokaryotic metagenome-assembled genomes (MAGs) from a hot spring stream with temperatures between 64 and 100°C. The stream water was filtered and the extracted total DNA was sequenced using the Illumina HiSeq 2500 platform. Approximately 80 Gb of raw data were generated, which were subsequently assembled using MEGAHIT v1.2.9. The MAGs were generated using MetaWRAP with binning approaches of MetaBAT2, CONCOCT and MaxBin2. We constructed 25 medium-quality and 24 high-quality archaeal MAGs, and 152 medium-quality and 112 high-quality bacterial MAGs. The fasta files of these MAGs are available in the NCBI database as well as Mendeley Data. Major phyla identified include Bacteroidota, Chloroflexota, Desulfobacterota, Firmicutes, Patescibacteria, Proteobacteria, Spirochaetota, Verrucomicrobiota, Armatimonadota, Nitrospirota, Acidobacteriota, Elusimicrobiota, Planctomycetota, Candidate division WOR-3, Aquificota, Thermoproteota, and Micrarchaeota. This dataset is valuable for studies on thermophilic genomes, reconstruction of biochemical pathways and gene discovery.

[*] Corresponding author.
*E-mail address:* gohkianmau@utm.my (K.M. Goh).

https://doi.org/10.1016/j.dib.2024.110829

Specifications Table

| | |
|---|---|
| Subject | Microbiology- microbiome. |
| Specific subject area | Thermophile metagenome-assembled genomes |
| Data format | Raw and fasta format for MAGs |
| Type of data | Table, Figure and processed Illumina shotgun reads into MAGs |
| Data collection | Environmental genomic DNA was extracted from cells trapped on a piece of filter membrane used to filter a pooled thermal water sample without culturing enrichment in the laboratory. The genomes of 313 prokaryotes were reconstructed from the metagenome dataset. |
| Data source location | City/Country: Sungkai/Malaysia. GPS: 3°59′50.50″N and 101°23′35.51″E |
| Data accessibility | Repository name: Raw reads and sequences of MAGs were submitted to NCBI in the public repository<br>Data identification number: Accession number PRJEB4990<br>Direct URL to data: https://data.mendeley.com/datasets/d7sdbdb3yk/2 and https://www.ncbi.nlm.nih.gov/datasets/genome/?bioproject=PRJEB4990<br>Biosamples: SAMN30304968-SAMN30305280<br>MAGs accession numbers: JANZYR000000000-JAOAKR000000000 |
| Related research article | C.S. Chan, K.G. Chan, Y.L. Tay, Y.H. Chua, K.M. Goh, Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing, Front Microbiol 6 (2015) 177.<br>https://doi.org/10.3389/fmicb.2015.00177. The amplicon 16S rRNA data, the water chemistry and a description of the sampling site can be found in this reference [1]. |

## 1. Value of the Data

- The dataset provides a comprehensive collection of high and medium quality MAGs from thermophilic environments. This collection enriches our understanding of the genomic diversity and metabolic capabilities of thermophilic and hyperthermophilic organisms.
- The dataset contains MAGs of previously uncharacterized or poorly understood microbial phyla. This emphasizes the potential to discover new microbial taxa and understand their ecological roles and evolutionary relationships, thus contributing to the broader field of microbial ecology and taxonomy.
- This dataset serves as a genetic repository for researchers, opening up new avenues for biotechnological exploration, particularly in the development of thermostable enzymes for industrial applications.

## 2. Background

Based on our own search, there are approximately 2,000 entries in the NCBI BioProject database related to the metagenomics of geothermal spring, encompassing both amplicon metagenomes and shotgun metagenomes. Thermophilic and hyperthermophilic organisms are often difficult to culture in the laboratory, limiting our understanding of their genomes [2,3]. Metagenomic approaches are therefore invaluable for studying these difficult-to-cultivate lineages. Analyzing thermophilic MAGs can provide new insights into their metabolism, environmental adaptations and biotechnological applications [4,5]. There are over 60 known hot springs with carbonate water in Malaysia [6]. We have previously studied the Sungai Klah (SK) hot spring, the second hottest geothermal spring in the country. This spring is rich in organic matter due to its forested environment and constant supply of plant litter, forming a unique ecosystem with various thermophiles. In 2015, we generated extensive sequence data using an Illumina HiSeq 2500 sequencer, which was then used to predict metabolic diversity in the hot spring, but only at a low resolution as bioinformatics pipelines were still limited at that time [1]. This report therefore aims to reprocess the Illumina raw shotgun data and create MAGs so that the scientific community interested in thermophiles or the hot spring microbiome can benefit from this data.

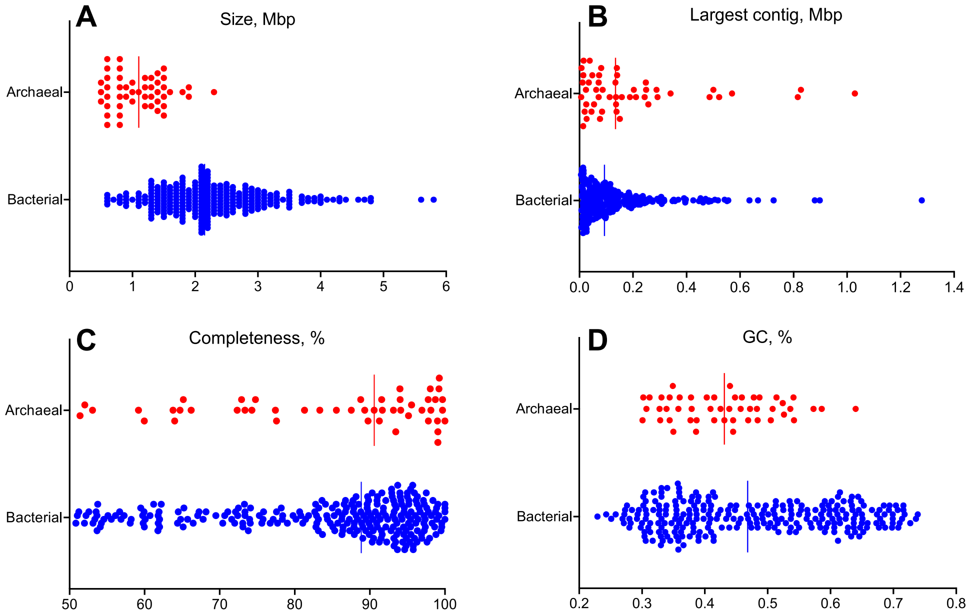

**Fig. 1.** Scatter plots showing the quality metrics for assembled MAGs. Vertical lines indicate the median of each dataset.

## 3. Data Description

'Minimum information about metagenome-assembled genome (MIMAG)' classifies MAGs into high-quality (>90% complete, <5% contamination), medium-quality (≥50% complete, <10% contamination), and low-quality (<50% complete, <10% contamination) drafts, with assemblies below the low-quality threshold considered bins rather than MAGs [7]. Medium and high-quality MAGs from this study were submitted to the NCBI database. Another way to access the fasta sequences of each MAG is to obtain them from the Mendeley Data Respiratory using the link provided above [8]. There are two types of files in the Mendeley Data. A master list in .xlsx format details each MAG, including the following information: name of MAGs, NCBI deposition information (i.e. genome accession number, biosample numbers, coverage), MAG quality and statistics (completeness, contamination, GC, N50, size, number of contigs, scaffold details, N50, N90, L50, L90). Additionally, the Excel file includes the affiliation predicted using the GTDB database and best match (closest_placement_taxonomy), if available [9].

The second file type in the Mendeley Data repository consists of the fasta format files of each MAG. Readers can use the fasta files for further analyzes that are not limited to those mentioned above under 'Value of the data'. It should be noted that this current data was analyzed based on the earlier water sample taken from SK hot spring that was initially published in 2015 [1]. There is another geothermal feature a few meters away from SK, which we named SKY, and MAGs of biofilms from SKY were reported recently [10]. Readers should treat both datasets separately because one consists of MAGs from water (current report), while the other contains MAGs from biofilms from two non-connected hot springs [10].

We constructed 152 medium-quality and 112 high-quality bacterial MAGs, and 25 medium-quality and 24 high-quality archaeal MAGs. Fig. 1 shows an overview of the quality of the MAGs compiled. In brief, we have constructed 49 archaeal MAGs, primarily consisting of Thermoproteota (27 MAGs) and Micrarchaeota (10 MAGs). Additionally, a total of 264 medium- and high-quality bacterial MAGs were obtained (Fig. 2). The following phyla have at least 10 MAGs each: Bacteroidota (36 MAGs), Chloroflexota (20 MAGs), Desulfobacterota (10 MAGs), Firmicutes (16 MAGs), Patescibacteria (10 MAGs), Proteobacteria (29 MAGs), Spirochaetota (15 MAGs), and Ver-

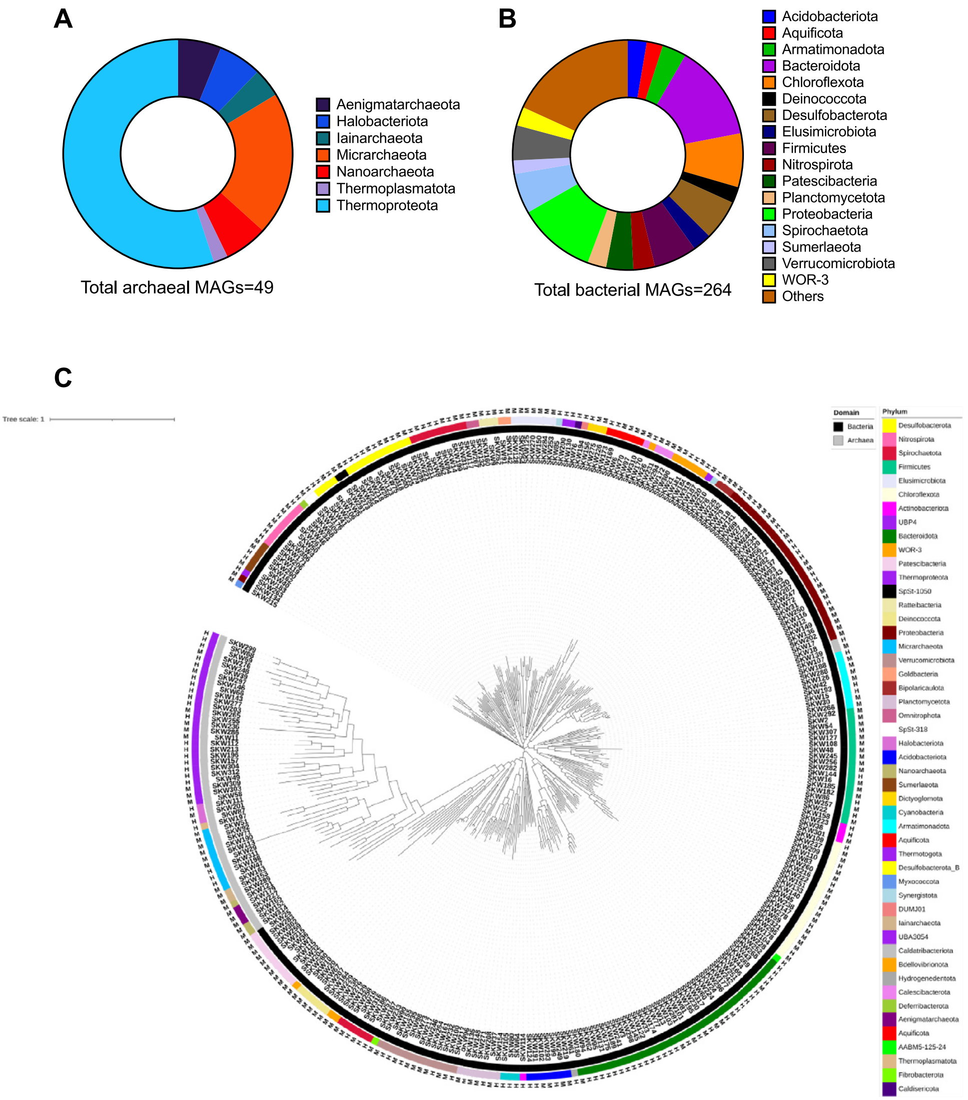

**Fig. 2.** Major archaeal (A) and bacterial (B) MAG phyla constructed from hot spring samples. (C) Phylogenomic tree (H: high-quality MAG, M: medium-quality MAG).

rucomicrobiota (13 MAGs). Other phyla are shown in the Fig. 2 and the .xlsx master list (Mendeley Data). Multiple MAGs belong to poorly understood candidate divisions or environmental samples, including the HRBIN17, WOR-3, DUMJ01, ABY1, SZUA-567, UBA4820, UBA3054, UBA5377, UBA9087, BMS3BBIN04, UBA8468, UBA4092, EX4484-205, PGYV01, UBA2361, UBA10030, SM23-61, UBA11346, UBA5829, JAAYUW01, BSN033, and others.

## 4. Experimental Design, Materials and Methods

The GPS coordinates for the Sungai Klah (SK) hot spring is 3°59′50.50′′N and 101°23′35.51′′E. In 2015, data on metagenomic shotgun sequencing of water samples collected from the SK stream were published. A pooled sample mixture of water and sediment was collected from multiple points along the hot spring stream, where temperatures ranged from 64 to 100°C. Microbial cells trapped on a 0.45-µm pore size filter (Sartorius, Göttingen, Germany) were subjected to total DNA extraction using the Metagenomic DNA Isolation Kit (Epicentre, Wisconsin, USA). The metagenome library was prepared using Illumina Nextera DNA Sample Preparation Kit, following the manufacturer's protocols. Shotgun sequencing was performed using the Illumina HiSeq 2500 sequencer (San Diego, CA, USA), utilizing a 150 bp paired-end approach with dual indexing [1]. Earlier reports presented data on microbial diversity and gene functional analyses but did not include MAGs due to the limitations of bioinformatic pipelines at the time. Here in this current report, we used the raw data, amounting to 80 Gb (SRA accession: ERX345285), underwent trimming and filtering with Cutadapt v3.3 (parameters: -a IlluminaAdapters.fa -A IlluminaAdapters.fa -e 0.1 -O 13 -q 30 –trim-n -m 50), followed by *de novo* assemblies using MEGAHIT v1.2.9 [11] (parameters: –min-count 2 –k-list 21,29,39,59,79,99,119,141). The metaWRAP v1.3 pipeline [12], incorporating MetaBAT2 v2.12.1, CONCOCT v1.0.0, and MaxBin2 v2.2.6 algorithms, was used for the binning process. To generate medium- to high-quality MAGs, the Bin_refinement (parameters: -c 50 -x 10; MAGs were filtered for >50% completeness and <10% contamination), Blobology, Quant_bins, and Reassembled_bin modules within the metaWRAP pipeline were utilized. The quality of the MAGs was assessed using the CheckM v1.0.12 program [13]. Barrnap v0.9 was used to predict the rRNA sequences, while ARAGORN v1.2.40 was used to identify the tRNA sequences from each MAG [14]. The metaWRAP's Classify_bins module (utilizing NCBI nt and NCBI taxonomy database) and GTDB-Tk v1.7.0 (utilizing GTDB reference data version r202) were used for MAGs taxonomy assignment [9,14].

## Limitations

The sequencing of the metagenome was performed using one of the early versions of Illumina flow cells and kits, resulting in an average raw read length of 2×150bp on the Illumina HiSeq 2500 sequencer. This read length is shorter than the latest standards, such as the NovaSeq 6000, which can generate 2×250bp reads. However, the total generated raw data was about 80 Gb, which should compensate for the limitation described earlier.

## Ethics Statement

This research did not involve human subjects, animals, or any species requiring ethical approval.

## CRediT Author Statement

**Jia Hao Tan:** Methodology. **Kok Jun Liew:** Methodology, Software, Visualization, Writing- Reviewing and Editing. **KM Goh**: Conceptualization, Investigation, Data curation, Validation, Writing, Original draft preparation.

## Data Availability

Dataset of 313 metagenome-assemble genomes from streamer hot spring water (Original data) (Mendeley Data).

## Acknowledgements

This research is funded by the Malaysia Ministry of Higher Education fund Fundamental Research Grant Scheme (FRGS) [grant number FRGS/1/2023/STG02/UTM/02/1] and Universiti Teknologi Malaysia Fundamental Research Grant [grant number 22H63].

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] C.S. Chan, K.G. Chan, Y.L. Tay, Y.H. Chua, K.M. Goh, Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing, Front. Microbiol. 6 (2015) 177, doi:10.3389/fmicb.2015.00177.

[2] K.M. Goh, S. Shahar, K.-G. Chan, C.S. Chong, S.I. Amran, M.H. Sani, I.I. Zakaria, U.M. Kahar, Current status and potential applications of underexplored prokaryotes, Microorganisms. 7 (2019). https://doi.org/10.3390/microorganisms7100468.

[3] M.S. Urbieta, E.R. Donati, K.-G. Chan, S. Shahar, L.L. Sin, K.M. Goh, Thermophiles in the genomic era: biodiversity, science, and applications, Biotechnol. Adv. 33 (2015) 633–647, doi:10.1016/j.biotechadv.2015.04.007.

[4] C. Yang, D. Chowdhury, Z. Zhang, W.K. Cheung, A. Lu, Z. Bian, L. Zhang, A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data, Comput. Struct. Biotechnol. J. 19 (2021) 6301–6314, doi:10.1016/j.csbj.2021.11.028.

[5] Z. Zhou, P.Q. Tran, A.M. Breister, Y. Liu, K. Kieft, E.S. Cowley, U. Karaoz, K. Anantharaman, METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks, Microbiome 10 (2022) 33, doi:10.1186/s40168-021-01213-8.

[6] C.S. Chan, K.-G. Chan, R. Ee, K.-W. Hong, M.S. Urbieta, E.R. Donati, M.S. Shamsir, K.M. Goh, Effects of physiochemical factors on prokaryotic biodiversity in Malaysian circumneutral hot springs, Front. Microbiol. 8 (2017) 1252, doi:10.3389/fmicb.2017.01252.

[7] R.M. Bowers, N.C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T.B.K. Reddy, F. Schulz, J. Jarett, A.R. Rivers, E.A. Eloe-Fadrosh, S.G. Tringe, N.N. Ivanova, A. Copeland, A. Clum, E.D. Becraft, R.R. Malmstrom, B. Birren, M. Podar, P. Bork, G.M. Weinstock, G.M. Garrity, J.A. Dodsworth, S. Yooseph, G. Sutton, F.O. Glöckner, J.A. Gilbert, W.C. Nelson, S.J. Hallam, S.P. Jungbluth, T.J.G. Ettema, S. Tighe, K.T. Konstantinidis, W.T. Liu, B.J. Baker, T. Rattei, J.A. Eisen, B. Hedlund, K.D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G.W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D.H. Parks, A.M. Eren, L. Schriml, J.F. Banfield, P. Hugenholtz, T. Woyke, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, Nat. Biotechnol. 35 (2017) 725–731, doi:10.1038/nbt.3893.

[8] K.J.G.K.M. Liew, Dataset of 313 metagenome-assemble genomes from streamer hot spring water", Mendeley Data, V2, Mendeley Data (2024).

[9] P.A. Chaumeil, A.J. Mussig, P. Hugenholtz, D.H. Parks, GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database, Bioinformatics 36 (2020) 1925–1927, doi:10.1093/bioinformatics/btz848.

[10] K.J. Liew, S. Shahar, M.S. Shamsir, N.B. Shaharuddin, C.H. Liang, K.-G. Chan, S.B. Pointing, R.K. Sani, K.M. Goh, Integrating multi-platform assembly to recover MAGs from hot spring biofilms: insights into microbial diversity, biofilm formation, and carbohydrate degradation, Environ. Microbiome 19 (2024) 29, doi:10.1186/s40793-024-00572-7.

[11] D. Li, R. Luo, C.M. Liu, C.M. Leung, H.F. Ting, K. Sadakane, H. Yamashita, T.W. Lam, MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices, Methods 102 (2016) 3–11, doi:10.1016/j.ymeth.2016.02.020.

[12] G.V. Uritskiy, J. DiRuggiero, J. Taylor, MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis, Microbiome 6 (2018) 158, doi:10.1186/s40168-018-0541-1.

[13] D.H. Parks, M. Imelfort, C.T. Skennerton, P. Hugenholtz, G.W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, Genome Res. 25 (2015) 1043–1055, doi:10.1101/gr.186072.114.

[14] D. Laslett, B. Canback, ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences, Nucleic Acids Res. 32 (2004) 11–16, doi:10.1093/nar/gkh152.