Data in Brief 57 (2024) 110935

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

16S rRNA metagenomics data on the bacterial communities in Poring Hot Spring, Sabah, Malaysia

Bak Zaibah Fazal[a], Nurshafrina Aida Yahya[a], Clemente Michael Wong Vui Ling[a], Yew Chee Wei[a], Thean Chor Leow[b], Mardani Abdul Halim[a], Krishnan Nair Balakrishnan[a], Cahyo Budiman[a,*], Zarina Amin[a,*]

[a] *Biotechnology Research Institute, University Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*
[b] *Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, UPM, 43400 Serdang, Selangor, Malaysia*

A R T I C L E   I N F O

Article history:
Received 16 August 2024
Revised 5 September 2024
Accepted 6 September 2024
Available online 14 September 2024

Dataset link: Sequence Read Archive (Original data)

Keywords:
Bacterial community
Hot spring
Metagenome
16S rRNA

A B S T R A C T

Hot springs are known to harbor potentially unique microorganisms due to the extreme temperatures in which they thrive and their biotechnologically important enzymes that are active at high temperature, which are beneficial for various industries. Sabah, Malaysia, houses several hot springs, yet knowledge of their microbiological diversity remains limited. Here, the raw sequence data of bacterial communities in a hot spring through metagenomic analysis are revealed. The data were obtained by collecting water and sediment samples from Poring Hot Spring (PHS) in Ranau, Sabah, and their bacterial diversity was analyzed using 16S rRNA amplicon sequencing targeting the V3-V4 regions. The analysis identified bacterial diversity in both water and sediment samples, with 35 phyla, 76 families, and 90 genera in water, and 38 phyla, 114 families, and 128 genera in sediment. Proteobacteria dominated the water samples (87 %), while Cyanobacteria were most abundant in sediment samples (51 %). The most abundant genera in water were Tepidimonas, Hydrogenophilus and Methylothermus, whereas Geitlerinema, Calothrix and Nitrospira dominated the sed-

* Corresponding authors.
E-mail addresses: cahyo@ums.edu.my (C. Budiman), zamin@ums.edu.my (Z. Amin).

https://doi.org/10.1016/j.dib.2024.110935

iment. Sediment samples exhibited higher bacterial richness and diversity compared to water samples, as indicated by α-diversity analysis. Sequences and sample data are deposited in the NCBI Sequence Read Archive under Bioproject ID PRJNA982554 (Accession number: SRX20671661 to SRX20671666) at https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA982554&o=acc_s%3Aa).

Specifications Table

Subject	Environmental Science (Environmental Genomic and Metagenomic)
Specific subject area	*16S rRNA metagenomic sequence of bacterial communities from hot spring*
Type of data	Table, figures, raw 16S rRNA amplicon sequences and analyzed OTUs data.
Data collection	Samples of water and sediment were collected from Poring Hot Spring, Sabah, and subjected to DNA extraction using DNeasy PowerWater and PowerSoil Kits (Qiagen, Germany). The extracted DNA was sequenced using the Illumina Hiseq platform using 16S rRNA primers. The unique or distinctive amplicon sequence variants obtained using the UPARSE pipeline were then grouped into operational taxonomic units (OTUs). Taxonomy assignment was performed using QIIME v.1.7.1 to identify and classify the different microbial taxa present in the samples
Data source location	Institution: Poring Hot Spring City/Town/Region: Ranau/ Sabah Country: Malaysia Latitude and longitude: 6.0458° N, 116.7034° E.
Data accessibility	Repository name: NCBI in Sequence Read Archive (SRA) Data identification number: BioProject PRJNA982554 Direct URL to data: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA982554&o=acc_s%3Aa
Related research article	*None*

1. Value of the Data

- This study provides the first comprehensive metagenomic profiling report on the bacterial diversity of hot springs in Sabah, which has significant implications for bioprospecting and conservation efforts at the site.
- The data can be applied in future comparative studies on bacterial diversity across different thermal water environments.
- The findings can serve as a baseline for monitoring changes in the bacterial community of Sabah's hot springs over time, providing valuable insights for management and conservation purposes.

2. Background

Poring Hot Spring (PHS) in Ranau, Sabah, is not only a sustainable tourism platform [1], but also a promising source for bioprospecting industrial thermophilic bacteria and their thermostable enzymes. The extreme conditions of hot springs are known to support unique microbial communities, making them valuable targets for discovering new, industrially relevant enzymes. 16S rRNA amplicon-based metagenome sequencing is widely utilized for comprehensive assessments of microbial diversity in thermal environments, including hot springs. This method involves sequencing the 16S ribosomal RNA gene, which is highly conserved among bacteria, to identify and quantify bacterial populations within a given sample. By targeting the V3-V4 regions of the 16S rRNA gene, researchers can achieve a detailed understanding of microbial composition and diversity.

The PHS ecosystem includes several microbial habitats, such as thermal fluid and sediment. Consequently, metagenomic data from both sources can provide a more comprehensive and holistic understanding of the microbial diversity in this hot spring. The aim of this research is to explore and characterize the microbial diversity of the Poring hot spring in Sabah, Malaysia. The focus is specifically on identifying thermophilic bacteria and enzymes that could be utilized for bioprospecting purposes. By leveraging 16S rRNA amplicon sequencing, this research seeks to uncover the potential of PHS as a source of novel and industrially important microbial resources.

3. Data Description

The raw data were obtained using Illumina HiSeq sequencing on the V3-V4 regions of the 16S rRNA gene, resulting in an average of 205,330 raw reads for sediment samples (PS) and 194,562 raw reads for water samples (PW) (Table 1). The sequences were processed using UPARSE, generated 819 and 1124 OTUs from sediment and water samples, respectively. The rarefaction curve, shown in Fig. 1 reflects the species richness of the sample.

Table 1
Summary of raw data generated from the 16S rRNA metagenomic sequencing of bacterial communities present in water and sediment samples from Poring hot spring.

Source of sample	Sample name	Raw reads	Cleaned reads	SRA accession number
Water	PW1	217,631	204,375	SRX20671661
	PW2	219,329	207,291	SRX20671662
	PW3	210,402	196,290	SRX20671663
Sediment	PS1	203,935	190,912	SRX20671664
	PS2	202,629	190,163	SRX20671665
	PS3	210,937	197,398	SRX20671666

A total of 1769 unique microbial operational taxonomic units (OTUs) were identified from the bacterial tags of all samples, with 819 OTUs in PHS water and 1124 OTUs in PHS sediment. Among these OTUs, 917 were found in both water and sediment samples, while 210 were unique to the water and 642 were unique to the sediment (Fig. 2).

OTUs from water samples were classified into an average of 35 phyla, 46 classes, 63 orders, 76 families, and 90 genera. In contrast, the OTUs from sediment samples were classified into an average of 38 phyla, 56 classes, 95 orders, 114 families, and 128 genera. The microbial community in PHS water was predominantly composed of *Proteobacteria* (84 %), *Nitrospirae* (8 %), *Chloroflexi (*3 %), *Deinococcus-Thermus* (1.4 %), and *Cyanobacteria* (1 %). Conversely, the PHS sediment was primarily inhabited by *Cyanobacteria* (49 %), *Chloroflexi* (12 %), *Bacteroidetes* (10 %), *Proteobacteria* (8.3 %), *Verrucomicrobia* (3%), and *Firmicutes* (1.6 %) as shown in Fig. 3.

In PHS water, the dominant bacterial classes were *Gammaproteobacteria* (83 %), *Thermodesulfovibrionia* (8 %), *Chloroflexia* (2.2 %), *Deinococci* (1.4 %), and *Oxyphotobacteria* (1 %). Meanwhile, the PHS sediment was mainly dominated by *Oxyphotobacteria* (49 %), *Chloroflexia* (10%), *Bacteroidia* (9 %), *Nitrospira* (9 %), *Gammaproteobacteria* (6 %), *Verrucomicrobiae* (3 %), *Clostridia* (1.5 %), and *Alphaproteobacteria* (1.4 %), as depicted in Fig. 4.

The most abundant bacterial orders in PHS water were *Betaproteobacteriales* (78 %), followed by uncultured bacteria (8 %), *Methylococcales* (5 %), and *Chloroflexales* (2 %). In the PHS sediment, the dominant orders were *Oxyphotobacteria* incertae sedis (32 %), *Nostocales* (14 %), *Chloroflexales* (10 %), *Nitrospirales* (9 %), *Chitinophagales* (7 %), *Betaproteobacteriales* (5 %), *Methylacidiphilales* (3 %), *Clostridiales* (2 %), *Methylococcales* (1 %), and uncultured bacteria (0.8 %) (Fig. 5).

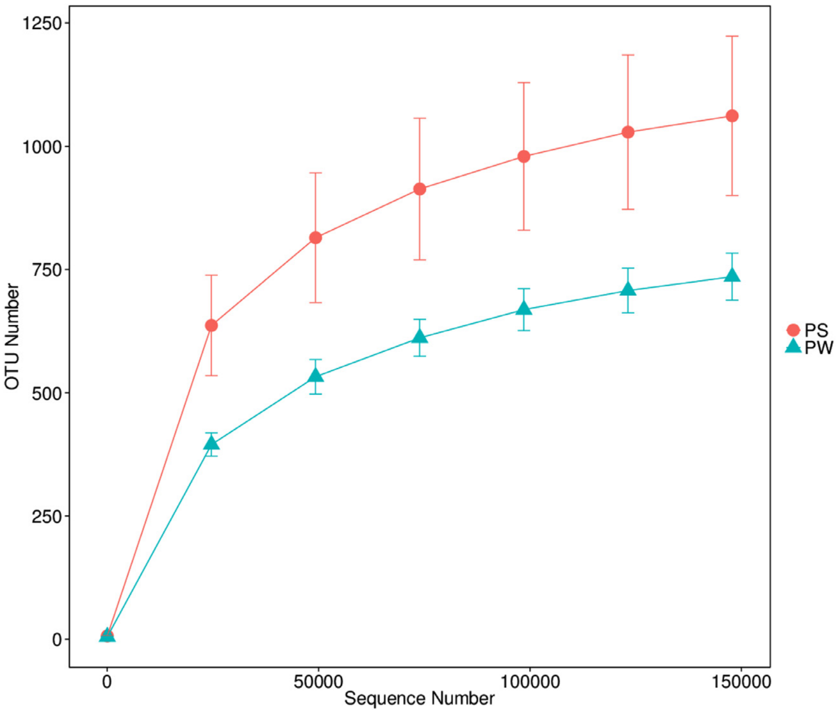

Fig. 1. The rarefaction curve shows the alpha-diversity analysis, reflecting the variation in OTUs abundance across each sample. PW and PS represent the water and sediment sample from PHS, respectively.

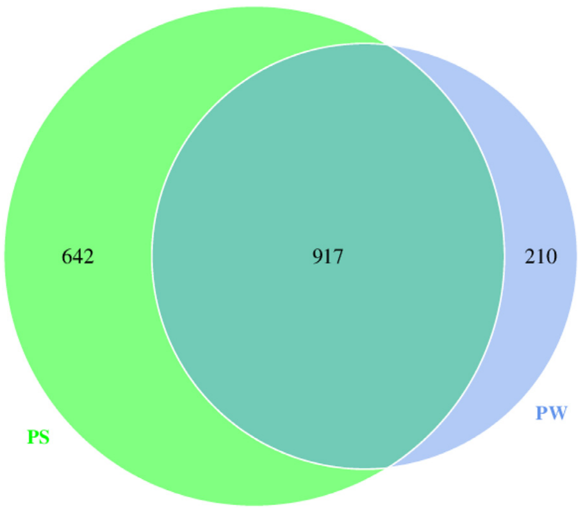

Fig. 2. The Venn diagram shows the number of unique and shared OTUs in water and sediment samples. PW and PS represent water and sediment sample from PHS, respectively.

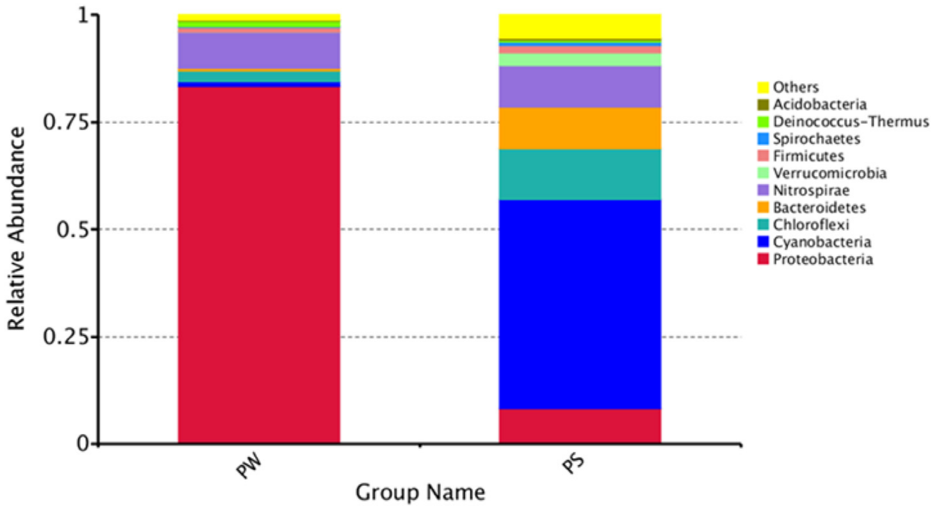

Fig. 3. The relative abundance of the ten dominant bacterial phyla identified from Poring Hot Spring. PW and PS indicate water and sediment sample from PHS, respectively.

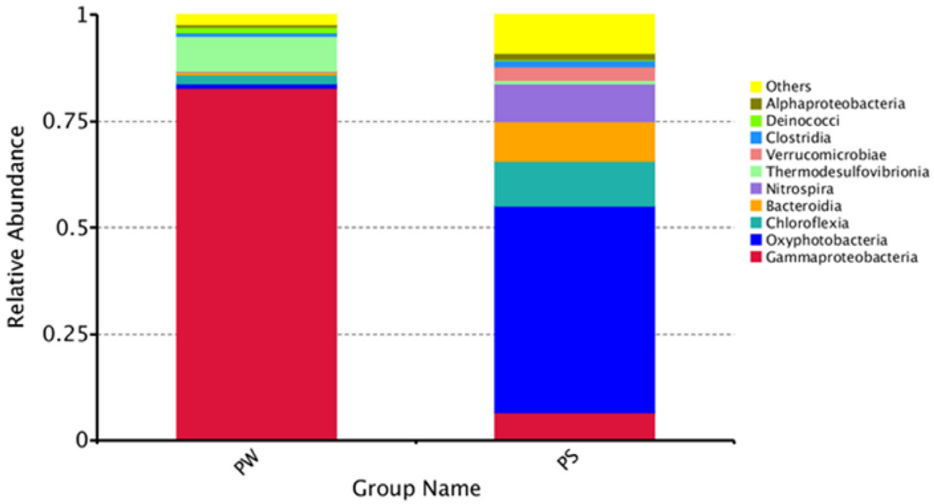

Fig. 4. The relative abundance of the ten dominant bacterial classes identified from Poring Hot Spring. PW and PS indicate water and sediment sample from PHS, respectively.

The most prevalent bacterial families in PHS water were *Burkholderiaceae* (58 %), *Hydrogenophilaceae* (19 %), *Methylohalobiaceae* (5 %), and *Chloroflexaceae* (2 %). In PHS sediment, the dominant bacterial family was unidentified, accounting for 32 % of the total bacterial population. Other significant families in the sediment included *Nostocaceae* (14 %), *Nitrospiraceae* (9 %), *Chloroflexaceae* (7 %), *Saprospiraceae* (7 %), *Roseiflexaceae* (3 %), *Methylacidiphilaceae* (2 %), *Burkholderiaceae* (2 %), and *Hydrogenophilaceae* (1 %) (Fig. 6).

The dominant bacterial genera in PHS water were *Tepidimonas* (58 %), *Hydrogenophilus* (18 %), *Methylothermus* (5 %), and *Chloroflexus* (2 %). In PHS sediment, the major bacterial genera were *Geitlerinema* PCC-8501 (32 %), *Calothrix* PCC-6303 (12 %), *Nitrospira* (9 %), *Chloroflexus* (7 %), un-

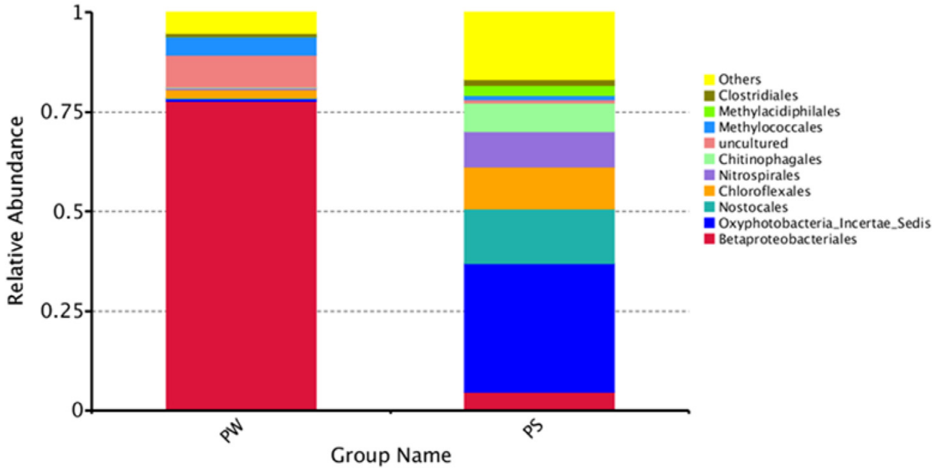

Fig. 5. The relative abundance of the ten dominant bacterial orders identified from Poring Hot Spring. PW and PS indicate water and sediment sample from PHS, respectively.

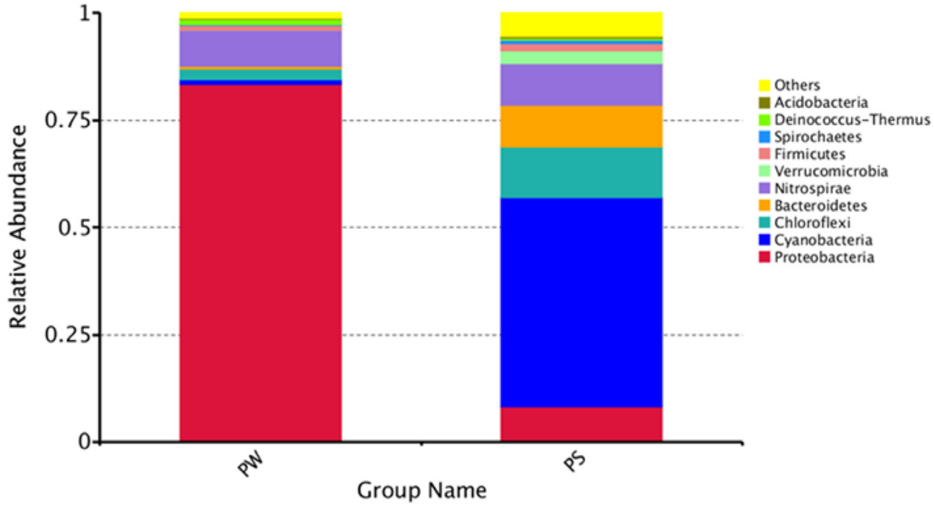

Fig. 6. The relative abundance of the ten dominant bacterial families identified from Poring Hot Spring. PW and PS indicate water and sediment sample from PHS, respectively.

cultured *Saprospiraceae* (6 %), *Roseiflexus* (3 %), uncultured *Methylacidiphilaceae* (3 %), and *Tepidimonas* (2 %) (Fig. 7).

4. Experimental Design, Materials and Methods

4.1. Sample collection and DNA extraction

Water and sediment samples were collected from Poring Hot Spring (PHS), a popular recreational site in Ranau, Sabah, Malaysia (6.0458° N, 116.7034° E). The water sample was designated as PW and the sediment sample as PS. DNA was extracted from these samples using the DNeasy

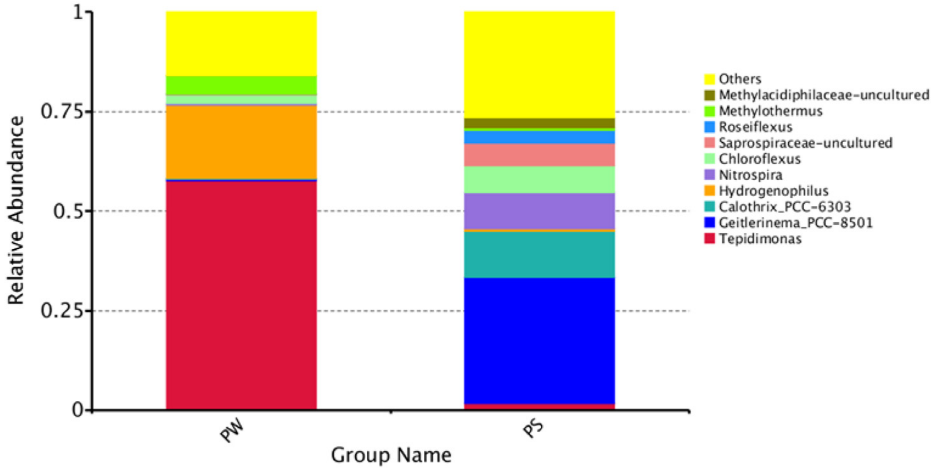

Fig. 7. The relative abundance of the ten dominant bacterial genera identified from Poring Hot Spring. PW and PS indicate water and sediment sample from PHS, respectively.

PowerWater and PowerSoil kits (Qiagen, Germany), following the manufacturer's protocols. The quality of the extracted DNA was evaluated using a Qubit 2.0 Fluorometer (Invitrogen, USA).

4.2. PCR and Shotgun sequencing

Bacterial 16S ribosomal RNA gene was amplified using the Phusion High-Fidelity PCR Master Mix (New England Biolabs, USA) with bacterial 16S primers 341F/785R [2] targeting the V3-V4 region of the 16S rRNA gene. The resulting rRNA amplicon libraries were barcoded and sequenced on the Illumina HiSeq platform (Illumina, USA).

4.3. Bioinformatics analysis

Paired-end reads were sorted and matched to their respective samples using unique barcodes. Barcode and primer sequences were trimmed, and the paired-end reads were merged using FLASH (V1.2.7) [3], a precise and rapid tool that merges paired-end reads where the reads partially overlap with the opposite end of the same DNA fragment. The merged sequences, termed raw tags, were filtered under specific conditions to generate high-quality clean tags using the Qiime (V1.7.0) quality control process [4]. These tags were compared against the Gold database reference using the UCHIME algorithm to identify and eliminate chimera sequences, leaving only effective tags for further analysis [5]. The UPARSE pipeline (V7.0.1001) [6]) was used to analyze all effective tags. Sequences with a similarity of ≥97 % were assigned to the same operational taxonomic units (OTUs). Abundance information for OTUs were normalized using the sequence number standard corresponding to the sample with the least sequences. Subsequent analysis of alpha diversity was performed based on this normalized output data. Phylogenetic relationships of representative sequences were determined using MUSCLE (Version 3.8.31) [7]. A representative sequence for each OTU was selected to further annotate the species at each taxonomic rank (threshold: 0.8–1) using QIIME v.1.7.1 [8] by comparing representative sequences against the SSUrRNA database of SILVA Database [9].

Limitations

Not applicable.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRediT Author Statement

Zarina Amin: Funding acquisition, Conceptualization, Supervision, Methodology, Data Curation, Writing – Review & Editing; **Cahyo Budiman, Clemente Michael Wong Vui Ling, Yew Chee Wei, Thean Chor Leow:** Supervision, Methodology, Writing – Review & Editing; **Bak Zaibah Fazal:** Investigation, Resources, Formal Analysis, Writing-Original Draft; **Nurshafrina Aida Yahya:** Investigation, Resources, Writing-Original Draft; **Mardani Abdul Halim, Krishnan Nair Balakrishnan:** Formal Analysis, Writing- Review & Editing.

Data Availability

Sequence Read Archive (Original data) (NCBI Database).

Acknowledgments

This research was supported by the Ministry of Higher Education of Malaysia through the Fundamental Research Grant Scheme (KPT Grant Code: FRGS/1/2018/STG05/UMS/02/4; UMS Grant Code: FRG0489-2018). We would also like to express our gratitude to Universiti Malaysia Sabah for supporting this project under UMS GUG0421-2/2019. We thank Sabah Biodiversity Centre (SaBC) for providing access licence (JKM/MBS.1000-2/2 JLD. 12 (100)) and Sabah Park for research permit (TTS/IP/100-6/2 JLD. 10 (26)) that enable the research in Poring Hot Spring.

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] A.K. Hua, Sustainable tourism in Poring Hot Spring Sabah: an experience, Int. J. Acad. Res. Environ. Geogr. 3 (1) (2016) 24–28.
[2] A. Klindworth, E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, F.O. Glöckner, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies, Nucleic Acids Res. 41 (1) (2013) e1-e1.
[3] T. Magoč, S.L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies, Bioinformatics 27 (21) (2011) 2957–2963.
[4] J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Peña, J.K. Goodrich, J.I. Gordon, G.A. Huttley, QIIME allows analysis of high-throughput community sequencing data, Nat. Methods 7 (5) (2010) 335–336.

[5] R.C. Edgar, B.J. Haas, J.C. Clemente, C. Quince, R. Knight, UCHIME improves sensitivity and speed of chimera detection, Bioinformatics 27 (16) (2011) 2194–2200.
[6] R.C. Edgar, UPARSE: highly accurate OTU sequences from microbial amplicon reads, Nat. Methods 10 (10) (2013) 996–998.
[7] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinform. 5 (2004) 1–19.
[8] J. Kuczynski, J. Stombaugh, W.A. Walters, A. González, J.G. Caporaso, R. Knight, Using QIIME to analyze 16S rRNA gene sequences from microbial communities, Curr. Protocols Microbiol. 27 (1) (2012) 1E–15.
[9] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F.O. Glöckner, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Res. 41 (D1) (2012) D590–D596.