



Data in Brief

Metagenomic analysis of microbial community of an Amazonian geothermal spring in Peru



Sujay Paul, Yolanda Cortez, Nadia Vera, Gretty K. Villena, Marcel Gutiérrez-Correa *

Laboratorio de Micología y Biotecnología, Universidad Nacional Agraria La Molina, Av. La Molina s/n, Lima 12, Peru

ARTICLE INFO

Article history:

Received 5 June 2016

Received in revised form 21 June 2016

Accepted 22 June 2016

Available online 26 June 2016

Keywords:

Peruvian Amazon rainforest

Geothermal/hot spring

Illumina metagenome sequencing

16S rRNA based phylogeny

PICRUST

ABSTRACT

Aguas Calientes (AC) is an isolated geothermal spring located deep into the Amazon rainforest (7°21'12" S, 75°00'54" W) of Peru. This geothermal spring is slightly acidic (pH 5.0–7.0) in nature, with temperatures varying from 45 to 90 °C and continually fed by plant litter, resulting in a relatively high degree of total organic content (TOC). Pooled water sample was analyzed at 16S rRNA V3–V4 hypervariable region by amplicon metagenome sequencing on Illumina HiSeq platform. A total of 2,976,534 paired ends reads were generated which were assigned into 5434 numbers of OTUs. All the resulting 16S rRNA fragments were then classified into 58 bacterial phyla and 2 archaeal phyla. Proteobacteria (88.06%) was found to be the highest represented phyla followed by Thermi (6.43%), Firmicutes (3.41%) and Aquificae (1.10%), respectively. Crenarchaeota and Euryarchaeota were the only 2 archaeal phyla detected in this study with low abundance. Metagenomic sequences were deposited to SRA database which is available at NCBI with accession number SRX1809286. Functional categorization of the assigned OTUs was performed using PICRUST tool. In COG analysis "Amino acid transport and metabolism" (8.5%) was found to be the highest represented category whereas among predicted KEGG pathways "Metabolism" (50.6%) was the most abundant. This is the first report of a high resolution microbial phylogenetic profile of an Amazonian hot spring.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Specifications	
Organism/cell line/tissue	Metagenome of Aguas Calientes (AC) hot spring
Sex	Not applicable
Sequencer or array type	Illumina Genome Analyzer IIX
Data format	Raw data: FASTAQ file
Experimental factors	Environmental sample
Experimental features	16S rRNA genes amplified from the metagenome using Illumina platform followed by bacterial community analysis using QIIME version 1.9.0
Consent	Not applicable
Sample source location	Water Sample, Aguas Calientes (AC) hot spring, Amazon rainforest, Peru

1. Direct link to deposited data

<http://www.ncbi.nlm.nih.gov/sra/?term=SRX1809286>.

* Corresponding author.

E-mail address: mgclmb@lamolina.edu.pe (M. Gutiérrez-Correa).

2. Experimental design, materials and methods

Microbial samples from extreme habitats (extremophiles) represent a huge reservoir of genetic diversity and a potent source for industrially important enzymes [1] but the culturability of microorganisms from extreme habitats is generally very poor. For long time, full length sequences generated by Sanger sequencing of 16S rRNA clone libraries were considered to be the gold standard for phylogenetic analysis but soon it was realized that this procedure is not only expensive but also have many limitations. To overcome the limitations and to gain high taxonomic resolution of bacterial population in extreme habitats, culture-independent 16S rRNAs amplicon based metagenome sequencing became a common practice [2]. Both 454 (Roche) and Illumina platforms are now largely used to study metagenome and microbial diversity [3,4] but the most recent and advanced Illumina MiSeq/HiSeq sequencing platform provides a very distinguish and high quality view of microbial composition than other sequencing technologies [5].

Hot springs are unique sites for extremophilic microorganisms and are of great interest for many years because enzymes obtained from them have been proved to be extremely valuable as biocatalysts for industrial and biotechnological purposes. Moreover, many unknown

microbial species and genes have been revealed in culture-independent microbial diversity assessment of hot springs [6]. The Peruvian Amazon rainforest is one of the most biologically diverse areas on earth and a rich source of several novel microbial species. Peruvian Amazon is endowed with few hot springs but till date none of them was explored in detail to generate high resolution microbial profile. So, the main goal of this study is to generate a high resolution microbial phylogenetic profile of an Amazonian hot spring.

In this study water samples were collected from four random sampling points of AC and mixed in an equal ratio as previously stated by Chan et al. [5] during metagenomic study of a Malaysian hot spring. The temperature and pH were measured on-site. Metagenomic DNA was extracted using PowerWater® DNA Isolation Kits (Mo Bio Laboratories) following manufacturer's protocol. Twenty five nanograms of Nanodrop quantified DNA was used for amplifying the V3–V4 region of 16S rRNA with specific primers which also have a 'tag' sequence that are complementary to Illumina sequence adapter and index primers from the Nextera XT Index kit V2. This round of PCR generates single amplicons of ~530 bp. In the next round of PCR (indexing PCR) Illumina sequencing adapters and dual indexing barcodes are added using limited cycle PCR to give a final product of ~610 bp. The libraries were cleaned using HighPrep PCR (Magbio, Cat # AC-60050) magnetic beads, Qubit quantified and validated for quality by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent). Finally, the cleaned libraries were sequenced in Illumina HiSeq platform at Genotypic Technology Private Limited, Bangalore, India. The Illumina paired end raw reads was quality checked using FastQC tool [7]. QIIME pipelines [8] was used for 16S RNA detection, clustering and OTU picking followed by Biom file generation and statistical analysis. SRA files were deposited to NCBI database under Accession Number SRX1809286. In the present study functional analysis of 16S amplicons was performed using the default settings of PICRUSt version 0.9.1 [9].

After quality filtration and adapter trimming of raw reads, clean sequences were clustered into 5434 operational taxonomic units (OTUs) using a 97% similarity cut off. Rarefaction curve indicated that a reasonable number of individuals were sampled (Fig. 1A). Good's coverage estimator revealed that >99% of the species were estimated, while high values of Chao1 richness estimator (9762) and Shannon diversity index (4.16) indicated that microbial communities in AC are highly rich and diverse. All the resulting fragments were then classified into 58 phyla, 165 classes, 300 orders, 520 families and 954 genera. The top 5 represented phyla were Proteobacteria (88.06%), Thermi (6.43%), Firmicutes (3.41%), Aquificae (1.10%) and Chloroflexi (0.41%) (Fig. 1B); however, an unidentified bacterial phylum (0.10%) was also found among top 10 bacterial phyla. An OTU based phylogenetic tree displayed the genetic diversity among AC microbial community (Fig. 1C). Gammaproteobacteria (86.1%) was found to be the highest represented class in AC microbial community followed by Deinococci (6.43%), Bacilli (3.22%), Betaproteobacteria (1.28%) and Aquificae (1.10%). A Krona chart was constructed to illustrate the distribution pattern of phyla Proteobacteria in AC hot spring (Fig. 2A). Further affiliation revealed that in AC microbial community the most abundant genus was *Acinetobacter* (71.09%) of Moraxellaceae family in which most of the species were unidentified (48.29%); remaining major genera included *Pseudomonas* (8.59%) of Pseudomonadaceae family, *Thermus* (5.99%) of Thermaceae family, *Enhydrobacter* (3.38%) of Moraxellaceae family and *Brevibacillus* (1.52%) of Paenibacillaceae family. The dominance of Proteobacteria, Firmicutes and Chloroflexi phyla in hot spring was reported earlier in few studies [5]. However, in contrast with our data, 16S rRNA based microbial diversity analysis from Little Hot Creek hot springs (temperature 78.7–82.5 °C and pH 6.75–6.97), California, similar to AC hot spring showed the dominance of the phyla Thermodesulfobacteria, *Deinococcus–Thermus*, Thermotogae and Dictyoglomi [10]. Aquificae and Thermotogae are two best known

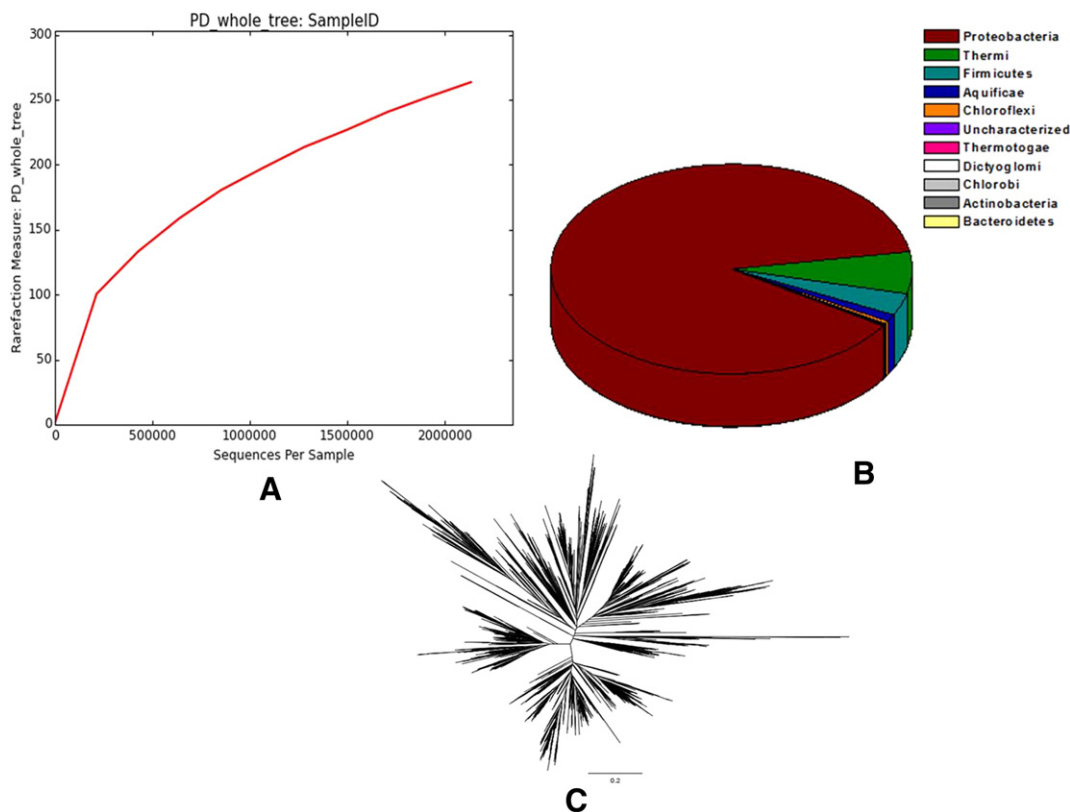


Fig. 1. Metagenomic diversity of Aguas Calientes hot spring (Peru). A: Rarefaction analysis curves of AC bacterial 16S rDNA sequences spanning the V3–V4 region; B: relative abundance of bacterial phyla in AC; and C: OTU based phylogenetic tree of AC microbial community.

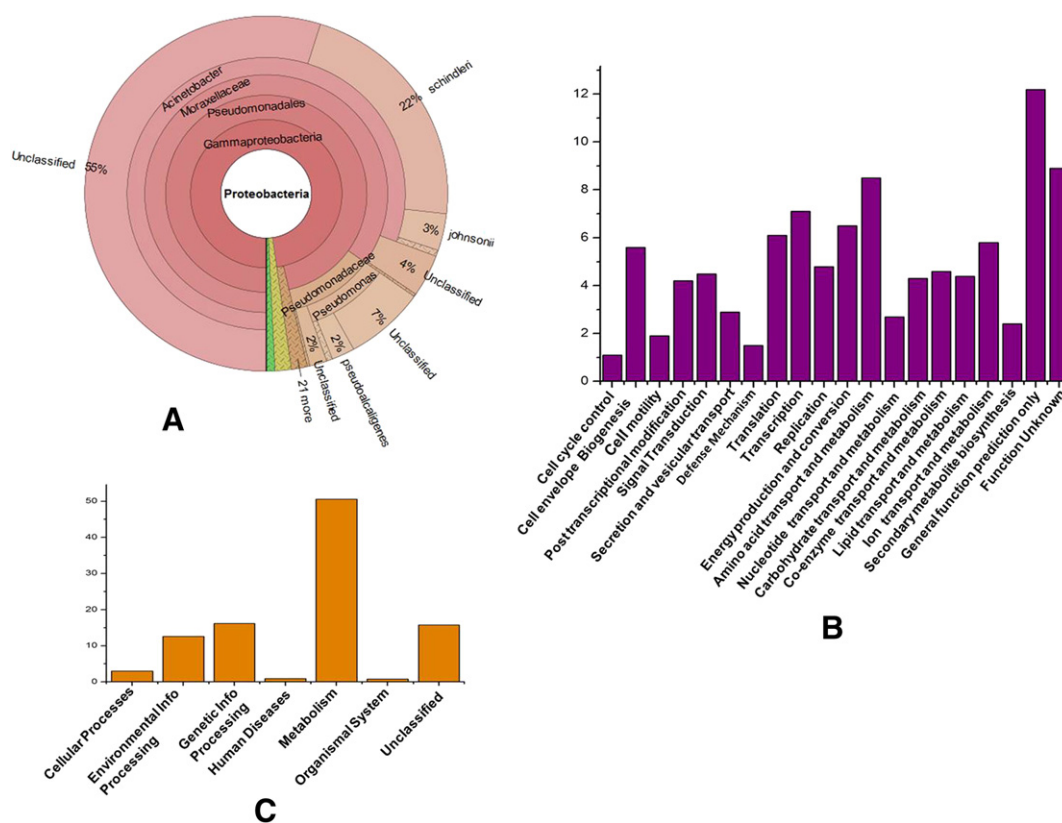


Fig. 2. Abundance and functional predictions of bacterial diversity of Aguas Calientes hot spring (Peru). A: Summary of the phylum Proteobacteria detected in AC in a multilevel chart (Krona); B: COGs metagenome functional predictions of OTUs; and C: KEGG metagenome functional predictions of OTUs using PICRUST tool.

hyperthermophilic bacterial phyla that show high heat tolerance. In this study an unknown genus (1.09%) of Aquificae dominates the sample, the other two low abundant genera detected in this study were *Hydrogenobacter* (0.01%) and *Hydrogenivirga* (0.00%); while in Thermotogae the genus *Fervidobacterium* (0.09%) was found to be the most represented one.

Undoubtedly, microbial diversity in hot springs is greatly affected by the pH of the water. Water samples from an acidic hot spring (pH 3.5–4) and a circumneutral hot spring (pH 7.2–7.4) were analyzed previously [11] and the results revealed that Thermotogae and Gammaproteobacteria dominated the acidic hot spring, while Thermodesulfobacteria, Gammaproteobacteria and Betaproteobacteria monopolized the circumneutral hot spring. In alkaline hot springs, *Thermus* [12], *Hydrogenobacter* [13], *Caldicellulosiruptor*, *Dictyoglomus* and *Fervidobacterium* [14] were reported to be the most represented genera. AC hot spring is slightly acidic to circumneutral in nature (pH 5.0–7.0) and in agreement with previous reports monopolization of Gammaproteobacteria was also found in this study. Regarding archaeal community structure Euryarchaeota and Crenarchaeota were the two major represented Archaeal phyla detected in AC but with low read count. In a recent study Thaumarchaeota and Crenarchaeota were found in both acidic and alkaline thermal spring environments, while Euryarchaeota was only found in the acidic environment [11]. Further affiliation of archaeal community in this study revealed that different unidentified families of the order pGrfC26 under Crenarchaeota and few methanogenic members of the family Methanosaetaceae, Methanobacteriaceae and Methanosarcinaceae under Euryarchaeota were also present in AC hot spring. Among which the complete genome sequence of a strain of *Methanobacterium thermoautotrophicum*, one of the best methane producing thermophilic archaeon are available in the database [15].

In this study, PICRUST, a modern tool designed to infer metagenomic information from 16S amplicon sequencing data, was used [9]. PICRUST

uses the OTU table of assigned taxa and their relative distribution to generate the relative abundance of functional categories based on sequenced genomes. Predicted abundance of gene categories (COGs) (Fig. 2B) and metabolic pathways (KEGG) (Fig. 2C) revealed that in COG analysis, despite of “General” (12.2%) and “Unknown” (8.9%) function categories, the highest represented category at second tier was “Amino acid transport and metabolism” (8.5%), followed by “Transcription” (7.1%) and “Energy production and conversion” (6.5%) (Fig. 2B). Among predicted KEGG pathways “Metabolism” (50.6%) was the most abundant category at first tier followed by “Genetic information processing” (16.2%), “Unclassified” (15.8%) and “Environmental information processing” (12.6%) (Fig. 2C). In agreement with our KEGG functional analysis data, “Metabolism” category followed by “Genetic information processing” was also reported previously in a Malaysian hot spring microbial community [5].

3. Nucleotide sequence accession number

Metagenome sequence data is available in NCBI SRA under accession number <http://www.ncbi.nlm.nih.gov/sra/?term=SRX1809286>.

Competing interest

Authors declare that there are no competing interests.

Acknowledgements

This work was supported by grants from Consejo Nacional de Ciencia y Tecnología de Peru [No. 002-2013-CONCYTEC-FONDECYT] and Programa Nacional de Innovación para la Competitividad y Productividad, Ministry of Production of Peru [No. 109-FINCYT-FIDECOM-PIPEA-2012 and No. 143-FINCYT-IB-2013].

References

- [1] T. Satyanarayana, C. Raghukumar, S. Shivaji, Extremophilic microbes: diversity and perspectives. *Curr. Sci.* 89 (2005) 78–90.
- [2] P. Dudhagara, A. Ghelani, R. Patel, R. Chaudhari, S. Bhatt, Bacterial tag encoded FLX titanium amplicon pyrosequencing (bTEFAP) based assessment of prokaryotic diversity in metagenome of Lonar soda lake, India. *Genomics Data* 4 (2015) 8–11.
- [3] G.B. Gloor, R. Hummelen, J.M. Macklaim, R.J. Dickson, A.D. Fernandes, R. Macphee, G. Reid, Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS One* 5 (2010) e15406, <http://dx.doi.org/10.1371/journal.pone.0015406>.
- [4] S. Sekar, A.A.E.A. Zintchem, J. Keshri, I. Kamika, M.N.B. Momba, Bacterial profiling in brine samples of the Emalaheni water reclamation plant, South Africa, using 454-pyrosequencing method. *FEMS Microbiol. Lett.* 359 (2014) 55–63.
- [5] 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, <http://dx.doi.org/10.3389/fmicb.2015.00177>.
- [6] A. Ghelani, R. Patel, A. Mangrola, P. Dudhagara, Cultivation-independent comprehensive survey of bacterial diversity in Tulsi Shyam Hot Springs, India. *Genomics Data* 4 (2015) 54–56, <http://dx.doi.org/10.1016/j.gdata.2015.03.003>.
- [7] S. Andrews, FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/2010>.
- [8] 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, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C.A. Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W.A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7 (2010) 335–336.
- [9] M.G.I. Langille, J. Zaneveld, J.G. Caporaso, D. McDonald, D. Knights, J.A. Reyes, J.C. Clemente, D.E. Burkpile, R.L. Vega Thurber, R. Knight, R.G. Beiko, C. Huttenhower, Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31 (2013) 814–821.
- [10] T.J. Vick, J.A. Dodsworth, K.C. Costa, E.L. Shock, B.P. Hedlund, Microbiology and geochemistry of Little Hot Creek, a hot spring environment in the Long Valley Caldera. *Geobiology* 8 (2010) 140–154.
- [11] B. Wemheuer, R. Taube, P. Akyol, F. Wemheuer, R. Daniel, Microbial diversity and biochemical potential encoded by thermal spring metagenomes derived from the Kamchatka Peninsula. *Archaea* 2013 (2013) <http://dx.doi.org/10.1155/2013/136714> (Article ID 136714).
- [12] K.B. De León, R. Gerlach, B.M. Peyton, M.W. Fields, Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geysir Basin, Yellowstone National Park. *Front. Microbiol.* 4 (2013) 330, <http://dx.doi.org/10.3389/fmicb.2013.00330>.
- [13] W. Hou, S. Wang, H. Dong, H. Jiang, B.R. Briggs, J.P. Peacock, Q. Huang, L. Huang, G. Wu, X. Zhi, W. Li, J.A. Dodsworth, B.P. Hedlund, C. Zhang, H.E. Hartnett, P. Dijkstra, B.A. Hungate, A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing. *PLoS One* 8 (2013) e53350, <http://dx.doi.org/10.1371/journal.pone.0053350>.
- [14] K. Sahn, P. John, H. Nacke, B. Wemheuer, R. Grote, R. Daniel, G. Antranikian, High abundance of heterotrophic prokaryotes in hydrothermal springs of the Azores as revealed by a network of 16S rRNA gene-based methods. *Extremophiles* 17 (2013) 649–662, <http://dx.doi.org/10.1007/s00792-013-0548-2>.
- [15] D.R. Smith, L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, J.N. Reeve, Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.* 179 (1997) 7135–7155.