Camuy River Cave Park (CRCP) is an underground cave system located at the subtropical karst carved by the Camuy River in the subtropical moist forest of northern Puerto Rico (Nieves-Rivera, 2003) [1]. This article contains a metagenomic dataset from the microbial and functional diversity of Clara Cave and Empalme Sinkhole water samples. The environmental DNA (eDNA) from the samples was extracted following direct Metagenomic DNA Isolation method, followed by Next-Generation-Sequencing technology (Illumina MiSeq). The sequences were submitted to MG-RAST online server for taxonomic profile generation and functional in silico description of the samples. The data consisted of domain Bacteria (96.69%), followed up by Viruses (2.87%), Eukaryotes (0.37%), and Archaea (0.02%). The data distribution by phyla showed Proteobacteria (92.61%), Bacteroidetes (1.66%), Actinobacteria (1.12%), and Firmicutes (0.48%). The subsystem functional data showed that 12.97% of genes were related to clustering-based subsystems, 11.40% to carbohydrates, and 11.0% to amino acids and derivatives. The metagenome dataset generated will provide an understanding and comparison framework of the microbial composition and functional diversity present in caves.

© 2018 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 table**

| Subject area          | Biology               |
|-----------------------|-----------------------|
| More specific subject area | Metagenomics          |
| Type of data          | Text files and figures |
| How data was acquired | MiSeq (Illumina), MR DNA Laboratories, USA. |
| Data format           | Raw and analyzed      |
| Experimental factors  | Environmental samples |
| Experimental features | The environmental DNA of water samples was extracted, sequenced and annotated. |
| Data source location  | Camuy, Puerto Rico (18°20'85.4"N, 66°49'16.0"W) and (18°20'79.9"N, 66°49'11.2"W) |
| Data accessibility    | Data is with this article. The data of this metagenome is available in the BioSample Submission Portal as Bioproject PRJNA487362 via https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA487362 and Sequence Read Archive (SRA) accession number SRR7741201 via https://www.ncbi.nlm.nih.gov/sra/?term=SRR7741201 |

**Value of the data**

- This project represents the first metagenomic water samples dataset of a cave and a nearby sinkhole generated in Puerto Rico.
- The data obtained provides an insight of microbial composition and functional diversity of cave and sinkhole water samples.
- Also, the profiles will serve for future comparison of taxonomic and functional profile among different types of cave.
- Furthermore, the data can be used to explore potential biomedical, industrial or biotechnological applications.

1. **Data**

Metagenomics has been used for the study of caves, mostly speleothems [2,3] or sediment [4], to unravel the microbial communities and activities in this ecosystem. As a complementary sequence data of speleothem or sediment, data from water samples present in the caves are also important to be generated. The Clara Cave and Empalme Sinkhole are an underground cave system and natural attraction from Camuy River Cave Park at the north of Puerto Rico which is part of a network of natural limestone caves [1]. To our knowledge, these environments have never been studied using metagenomics. The dataset obtained after shotgun sequencing of water samples from the Clara Cave and Empalme Sinkhole are described as taxonomic and functional profiles in Figs. 1 and 2 respectively.

2. **Experimental design, materials and methods**

2.1. **Sample collection**

Three water samples were collected from Clara Cave (18°20’85.4’’ N, 66°49’16.0’’ W) using sterile 100 mL water sampling plastic containers. Two of three water samples of Clara Cave were from water streaming down the cave walls. The other was collected at 0.3-m depth from stagnant water. One water sample was collected from a stream at Empalme Sinkhole, which connects directly to Clara...
The taxonomic analysis revealed the most abundant domain was Bacteria (96.69%), followed up by viruses (2.87%), Eukaryotes (0.37%), other sequences (0.04%), and Archaea (0.02%). The analysis consisted of 52 phyla, from which Proteobacteria (92.61%) showed to be the most abundant, followed up by unclassified for viruses (2.87%), Bacteroidetes (1.66%), Actinobacteria (1.12%), and Firmicutes (0.48%). The metagenome represents the DNA extracted from combined water samples.

Subsystem functional analysis presented that 11.0% of genes are related to clustering-based subsystems, 11.0% to amino acids and derivatives, 10.0% to carbohydrates, 6.0% to miscellaneous, and 5.0% to membrane transport. In addition, genes related to metabolism of proteins (6.0%), DNA (4.0%), RNA (4.0%), iron (3.0%), aromatic compounds (3.0%), sulfur (2.0%), nitrogen (1.0%), phosphorous (1.0%), potassium (0.8%), and secondary metabolism (0.2%). The metagenome represents the DNA extracted from combined water samples.
Cave exit. Temperature measures of samples ranged from 18 °C to 20 °C. The samples were transported to the laboratory where pH measures were taken; pH values ranged from 8.38 to 9.93.

2.2. DNA extraction

For DNA extraction, approximately 50 mL of each water sample were taken to combine it into one sample. The protocol of Metagenomic DNA Isolation Kit from Water (Epicentre, USA, 11) was followed for 200 mL of sample. Extracted high molecular weight (40 kbp) DNA was sent to MR DNA (http://www.mrdnalab.com), where a genomic library was prepared using Qubit® dsDNA HS Assay Kit (Life Technologies) and Nextera DNA Sample Preparation Kit (Illumina) following the manufacturer’s instructions. After using 50 ng of DNA to prepare the library, it underwent fragmentation and addition of adapter sequences. Following the library preparation, the final concentration of DNA was 7.22 ng/μL and the average library size was 2355 bp, this one determined by using Agilent 2100 Bioanalyzer (Agilent Technologies). The library was diluted to 10.0pM and sequenced using the 600 cycle v3 Reagent Kit (Illumina) on the MiSeq (Illumina).

2.3. Taxonomic and functional insight

To unravel the microbial diversity and functional in silico of the samples, the metagenomic sequence was examined using the Metagenomic Rapid Annotation using Subsystems Technology (MG-RAST) online server [5]

Acknowledgements

We would like to thank the Maximizing Access to Research Careers (MARC) U*STAR Program (NIH-T34GM008419) for supporting this project. In addition, we thank the administration of Camuy River Cave Park, especially Maribel Acaba for welcoming and allowing us to collect the samples. Finally, special thanks to Ricky Padilla-Del Valle and Luis R. Morales for their guidance during this project.

Transparency document. Supplementary material

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.11.028.

References

[1] Á.M. Nieves-Rivera, Mycological survey of Río Camuy Caves Park, Puerto Rico, J. Cave Karst Stud. 65 (2003) 22–28.
[2] M. Ortiz, A. Legatzki, J.W. Neilson, B. Fryslie, W.M. Nelson, R.A. Wing, R.M. Maier, Making a living while starving in the dark: metagenomic insights into the energy dynamics of a carbonate cave, ISME J. 8 (2014) 478–491. https://doi.org/10.1038/ismej.2013.159.
[3] M.L.Z. Mendoza, J. Lundberg, M. Ivarsson, P. Campos, J.A.A. Nylander, T. Sallstedt, L. Dalen, Metagenomic analysis from the interior of a speleothem in Tjuv-Ante’s cave, Northern Sweden, PLoS One 11 (3) (2016), https://doi.org/10.1371/journal.pone.0151577.
[4] S. De Mandal, A.K. Panda, E. Lalmunmawii, S.S. Bisht, N.S. Kumar, Illumina-based analysis of bacterial community in Khuangcherapuk cave of Mizoram, Northeast India, Genom. Data 5 (2015) 13–14. https://doi.org/10.1016/j.gdata.2015.04.023.
[5] F. Meyer, D. Paarmann, M. D’Souza, R.A. Edwards, The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes, BMC Bioinform. 9 (2008) 386. https://doi.org/10.1186/1471-2105-9-386.