Data Article

Datasets generated by shotgun sequencing of metagenomic libraries of the guajataca water reservoir

Berliza M. Soriano a, Laura M. Del Valle-Perez b, Luis Morales-Vale b, Carlosrios-Velazquez b,∗

a Industrial Biotechnology Program University of Puerto Rico, Mayaguez, Puerto Rico
b Biology Department University of Puerto Rico, Mayaguez, Puerto Rico

A R T I C L E  I N F O

Article history:
Received 4 August 2018
Received in revised form
9 November 2018
Accepted 21 November 2018
Available online 27 November 2018

A B S T R A C T

The Guajataca Water Reservoir (GWR) was constructed for irrigation and to bring potable water to the northwestern region of Puerto Rico. The generation of DNA sequencing data from aquatic bodies (AB) using culture-independent approaches allows the investigation of the total microbial diversity as well as the potential anthropogenic impact. Metagenomic libraries were constructed for two GWR sampling sites and genomic information access through shotgun sequencing. After removing the bacterial host cell genome and the library fosmid sequences, the environmental genome was processed through Rapid Annotation using Subsystems Technology for Metagenomes (MG-RAST). The sequences consisted primarily of bacteria (95.70%), followed by viruses (2.94%), other sequences (0.28%) and eukaryote (0.09%). The most abundant species were Enterobacter cloacae (31%), Enterobacter sp. 638 (20%), Enterobacter cancerogenus (10%) and Escherichia coli (11%). Furthermore, the subsystem data showed that 13% of the genes belong to carbohydrates functionality, 12% to clustering-based-subsystems and another 9% related to virulence-disease-and-defense (out of which 8% pertain to genes of antibiotic resistance and toxic compounds). This unique data input will serve...
as a baseline to a better understanding not only the microbial communities present in the AB, but also the microbial activities with potential application in biotechnological and biomedical fields.

© 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         | Figures, FASTQ File|
| How data were acquired | Illumina MiSeq    |
| Data format          | Raw                |
| Experimental factors | Environmental Water Sample, Metagenomic Library Construction, Shotgun Sequencing, and Analysis. |
| Experimental features | The metagenomic library from Guajataca Water reservoir was shotgun sequenced and annotated using MG-RAST. |
| Data source location | Two samples were collected from the Guajataca Water Reservoir located in Quebradillas, Puerto Rico (N 18° 22’ 33.103” W 66° 55’ 19.268” – N 18° 23’ 33.799” W 66° 55’ 20.174”). |
| Data accessibility   | Data are with this article. The metagenomic sequenced Data generated were annotated using MG-RAST, ID mgm4771977.3 and deposited in NCBI database under the BioSample accession SAMN09396994 |

### Value of the data

- This project embodies the first Guajataca Water Reservoir (GWR) metagenomic libraries dataset generation.
- These data provide the scientific community with the first microbial taxonomic profile and functional structure of GWR using metagenomics and shotgun sequencing.
- Data obtained from this data article can be used for a better understanding of the ecology, biology, and anthropogenic impact of unexplored environments such as AB.
- Moreover, these data deliver a comparison point for future metagenomics studies in other water reservoirs around the world.

### 1. Data

#### 1.1. Background

Among aquatic bodies (AB), water reservoirs and dams are essential because of their direct impact on humans. Water reservoirs are artificial man-made lakes to store clean water [1]. A total of 36 public reservoirs are reported to be own by the Commonwealth of Puerto Rico with 21 being considered of significant diversity and volume input [2]. The Guajataca Water Reservoir (GWR), located in Quebradillas, supplies potable water for most the northwestern region. However, its fundamental function is for agricultural irrigation for the flatlands of the area [3]. Metagenomics is a culture-independent approach that overcomes the problem of the uncultivability of most microbes to allow genomic and functional access to AB and many other environments [4]. The data presented within the
data article was generated by shotgun sequencing of combined (1:1) metagenomic libraries made from two GWR samples. GWR's microbial diversity data are presented as a taxonomic profile (Fig. 1) and functional subsystem structure (Fig. 2).

1.2. Taxonomic and functional unraveling

See Figs. 1 and 2.
2. Experimental design, materials and methods [5,6]

2.1. Metagenomic library generation

For the generation of high molecular weight metagenomic libraries, GWR water samples were collected from shallow waters (0–1 m depth). Direct DNA extraction was performed using a Metagenomic DNA Isolation Kit for Water. DNA of approximate 40 kbp was processed according to the specifications of the CopyControl™ Fosmid Library Production Kit from Epicentre. In brief, DNA was end-repaired, precipitated and ligated into a CopyControl pCC1FOS vector. MaxPlax Lambda phages were used to package the recombinant DNA and to transduce it to EPI300-T1R host cells. Libraries were plated, and the number of clones determined as well as the insertion percentage by enzymatic digestion with *Bam*H1 (New England Biolabs).

2.2. Metagenome sequencing and pCCFOS1 vector removal

A midiprep kit (QIAGEN) was used as described by the manufacturer to extract and purify the MetaDNA from the GWR metagenomic library. The sample was then sent to the MR DNA for sequencing. Following the Qubit® dsDNA HS Assay Kit (Life Technologies) and Nextera DNA Sample Preparation Kit (Illumina) in verbatim, a genomic library was generated. DNA samples (50 ng) were diluted (2.5 ng/µL), fragmented, and finally the addition of adapter sequences. The library was diluted to 12 pM to be sequenced using a 600 cycle v3 Reagent Kit (Illumina) on the MiSeq (Illumina). To isolate the metagenome sequence, the pCCFOS1 vector was removed using the Gene Indices Sequence Cleaning and Validation script. Finally, a Blastn was performed to corroborate that no sequences from E. coli DH10B or the plasmid pCC1FOS where present in the metagenome.
2.3. Taxonomic and functional insights

The clean metagenomic sequences were submitted to the Rapid Annotation and subsequently examined using Subsystems Technology for Metagenomes (MG-RAST) online server to develop a functional \textit{in silico} description of the GWR.

Acknowledgments

We want to thank the Puerto Rico Louis Stokes Alliance for Minority Participation [NSF-1400868] and the Howard Hughes Medical Institute [HHMI-52007566]. We recognize the collaboration of Dr. Carlos Santos-Flores, Dr. Gustavo Martinez and the personnel of the UPR-Mayaguez Agriculture Experimental Station for bringing the transportation for the sample collection.

Transparency document. Supporting information

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.11.114.

References

[1] National Geographic Society, Reservoirs, Retrieved (09 October 2012). from (https://www.nationalgeographic.org/encyclopedia/reservoir/).
[2] (n.d). Retrieved 11 April 2018, from (https://enciclopediapr.org/en/encyclopedia/reservoirs-of-puerto-rico/).
[3] (n.d.). Retrieved 04 September 2017, from (https://pr.water.usgs.gov/public/rt/pr_lakes/lake_50010800.html).
[4] Committee on Metagenomics: Challenges and Functional Applications, Board on Life Sciences, N. Division on Earth and Life Studies, Why metagenomics?, 1st ed.,The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet, 12–13, National Academic Press, Washington, 2007.
[5] R. Padilla-Del Valle, L.R. Morales-Vale, C. Ríos-Velázquez, Unraveling the microbial and functional diversity of Coamo thermal spring in Puerto Rico using metagenomic library generation and shotgun sequencing, Genom. Data 11 (2017) 98–101. https://doi.org/10.1016/j.gdata.2016.12.010.
[6] R. Rabelo-Fernandez, K. Santiago-Morales, L. Morales-Vale, C. Rios-Velazquez, The metagenome of \textit{Caracolus marginella} gut microbiome using culture-independent approaches and shotgun sequencing, Data Brief 16 (2018) 501–505. https://doi.org/10.1016/j.dib.2017.11.043.