Data in brief 28 (2020) 104919

Contents lists available at ScienceDirect

Data in brief

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

Data Article

Soil microbiome dataset from Guanica dry forest in Puerto Rico generated by shotgun sequencing

Roberto G. Sotomayor-Mena [a], Carlos Rios-Velazquez [b, *]

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

ARTICLE INFO

Article history:
Received 5 November 2019
Accepted 27 November 2019
Available online 3 December 2019

Keywords:
Metagenomics
Puerto Rico
Subtropical
Dry forest
Microbiome
Functional
Diversity
Library

ABSTRACT

Guanica dry forest (GDF), located in the southwest area or region of Puerto Rico, is among the most preserved subtropical dry forests in the world [1]. To describe the taxonomic diversity and functional profiles of this environment, metagenomic DNA was extracted from a metagenomic library generated from the GDF. The DNA was shotgun-sequenced using Illumina and analyzed using the MG-RAST server. The diversity profile revealed that the most abundant domain was Bacteria (97.8%) followed by Archaea (1.12%), Eukaryota (1.02%) and Viruses (0.03%). Out of the 50 phyla present, the most abundant was Proteobacteria (41.6%) followed by Actinobacteria (18.7%) and Acidobacteria (7.06%). Moreover, a total of 213 orders, 384 families and 791 genus were identified. The functional profile showed abundance of genes related to Carbohydrates (13.16%), Clustering-based subsystems (13.0%), Amino Acids and Derivatives (9.9%) and Protein Metabolism (8.24%). Furthermore, more specific grouping showed that NULL (21.5%) was the most abundant function group, followed by Plant-Prokaryote DOE project (6.05%), Protein biosynthesis (4.82%), Central carbohydrate metabolism (3.98%), DNA repair (2.72%) and Resistance to antibiotics and toxic compounds (2.66%). This dataset is useful in bioprospecting studies with application in biomedical sciences, biotechnology and microbial, population and applied ecology fields.

* Corresponding author.
E-mail address: Carlos.rios5@upr.edu (C. Rios-Velazquez).

https://doi.org/10.1016/j.dib.2019.104919

Specifications Table

Subject	Biology
Specific subject area	Metagenomics
Type of data	FASTQ files, Figures
How data were acquired	Shotgun sequencing MiSeq (Illumina), MG-RAST (Metagenomics Rapid Annotation Using Subsystems Technologies)
Data format	Raw, Processed
Parameters for data collection	A soil sample was collected at 5 cm of depth.
Description of data collection	A metagenomic library of soil from Guanica dry forest in Puerto Rico was generated. Data was obtained by fosmid extraction, shotgun-sequencing and analysis in the MG-RAST pipeline.
Data source location	Guanica dry forest, Puerto Rico (*17°57′56″* N, *66°52′45″* W)
Data accessibility	The Metagenomic data was deposited in the NCBI database under the bioproject accession number PRJNA587232. The processed data files are accessible through the Metagenomics Rapid Annotation using Subsystems Technology server (MG-RAST ID: mgm4863402.3) (https://www.mg-rast.org/linkin.cgi?project=mgp91184).
Related research article:	Jose M. Cruz et al. (2010) Unraveling Activities by Functional-Based Approaches using Metagenomic Libraries from Dry and Rain Forest Soils in Puerto Rico Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology

Value of Data

- This project presents the diversity and functional profiles of soil from the subtropical dry forest in Puerto Rico.
- The profiles generated can be used in comparative studies involving soils from high temperature and low humidity environments.
- The functional profile can be used for bioprospecting studies with applications in biomedical sciences, biotechnology and microbial, population and applied ecology fields.

1. Data

Guanica dry forest (GDF), the second International Biosphere Reserve in Puerto Rico, is among the most well-preserved subtropical forest in the world [1]. The forest spans 11,000 acres of land (including 13 miles of coast), has an annual precipitation of 30 inches and fluctuates between 26.6 °C and 37.7 °C [1]. Here, we describe the diversity (Fig. 1) and functional (Fig. 2) profile of a metagenomic library generated from GDF soil containing 781,199 clones (masterpool). Each individual clone (epi300 − *Escherichia coli*) contains a pCC1FOS vector with a 40kb DNA insert extracted from GDF soil. We performed QIAGEN Plasmid Midi Kit protocol to obtain fosmid DNA containing insert, sequenced it by Illumina shotgun sequencing and plotted the profiles using the MG-RAST server. The raw FASTQ and processed FASTA files were made accessible through the NCBI database and MG-RAST server respectively.

2. Experimental design, materials, and methods

2.1. Sample collection

On June 2006, a 5 cm deep soil sample was aseptically collected from Guanica dry forest in Puerto Rico (*17°57′56″* N *66°52′45″* W) (USDA soil collection permit S-65547). It was transported at 4 °C and stored at −20 °C.

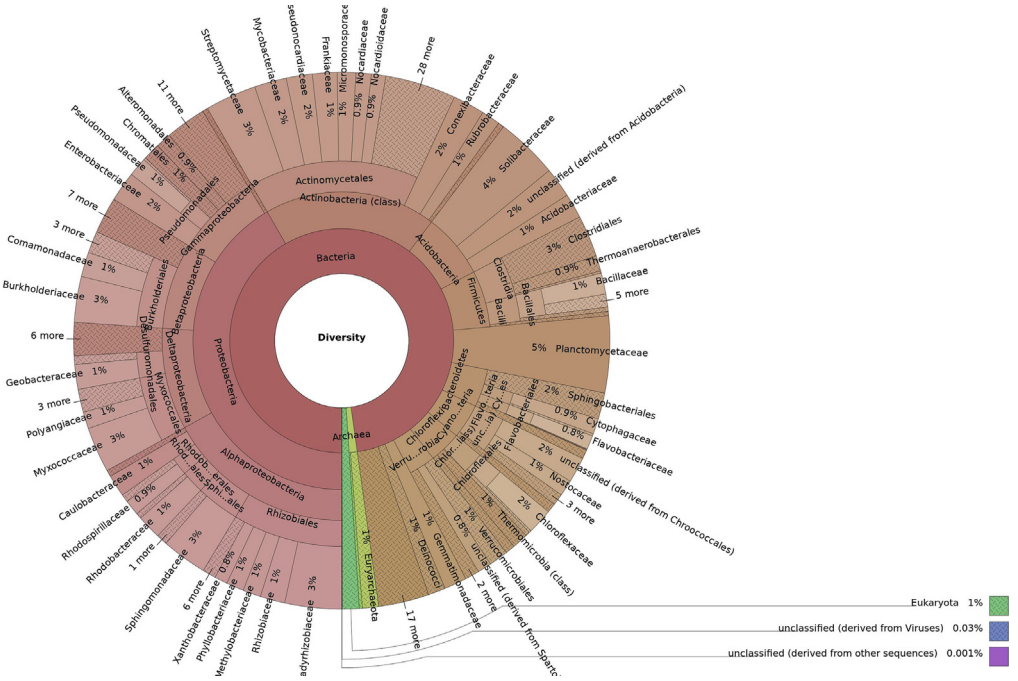

Fig. 1. Taxonomic diversity of the Guanica dry forest metagenome. Using the RefSeq database resource in MG-RAST, the most abundant domain was Bacteria (97.8%) followed by Archaea (1.12%), Eukaryota (1.02%) and Viruses (0.03%). Out of the 50 phyla present, the most abundant were Proteobacteria (41.6%) followed by Actinobacteria (18.7%) and Acidobacteria (7.06%). From the 213 orders detected, the most abundant were Actinomycetales (15.2%), Rhizobiales (8.64%) and Burkholderiales (5.54%). Moreover, a total of 384 families and 791 genus were identified.

2.2. DNA extraction and library preparation

A direct DNA extraction and metagenomic library generation protocols were performed and documented by Cruz et al. [2]. Briefly, the soil sample was sieved and treated with freeze-thaw cycles. A sizing procedure was performed to the extracted DNA and 40 kb fragments were electro-eluted from the agarose gel. The purified DNA was ligated into the fosmid vector pEpiFOS™-5, and packed into lambda phases using the Lucigen MaxPlax lambda packaging extracts. The packed DNA was transduced into EPI300™-T1R and the clones were combined into a masterpool which was stored at −80 °C.

2.3. Metagenome sequencing

Following the manufacturer's protocol, the QIAGEN Plasmid Midi Kit was used to extract fosmid DNA from a masterpool culture incubated for 10 hours at 37 °C. Metagenomic DNA was sent to MR DNA (http://www.mrdnalab.com) for shotgun sequencing. A library was prepared using Nextera DNA Flex library preparation kit (Illumina) following the manufacturer's user guide. The sample underwent the simultaneous fragmentation and addition of adapter sequences. A limited-cycle (6 cycles) PCR was done to add a unique index to the sample. The library was diluted to 4.0 nM and sequenced paired end for 600 cycles using the MiSeq system (Illumina).

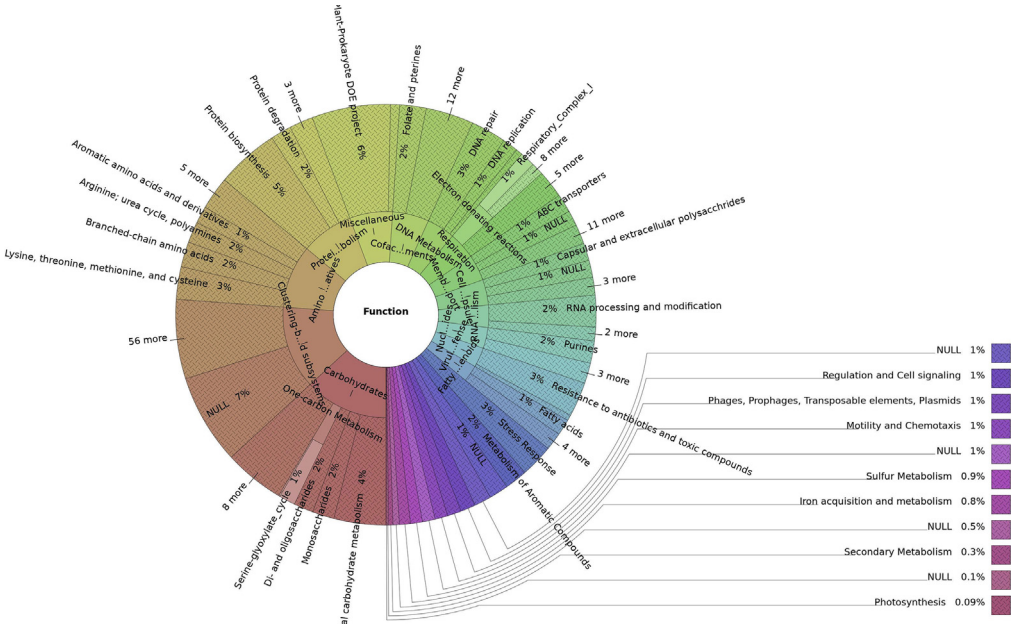

Fig. 2. Functional profile of Guanica dry forest metagenome. Using the subsystems database resource in MG-RAST, the most abundant functions under a broad grouping classification (level 1) were Carbohydrates (13.16%), Clustering-based subsystems (13.0%), Amino Acids and Derivatives (9.9%) and Protein Metabolism (8.24%). Under more specific grouping (level 2), the more abundant were NULL (21.5%), Plant-Prokaryote DOE project (6.05%), Protein biosynthesis (4.82%), Central carbohydrate metabolism (3.98%), DNA repair (2.72%) and Resistance to antibiotics and toxic compounds (2.66%).

2.4. Sequence processing

FastQC was used to determine sequence quality [3]. FastX toolkit was used to trim sequences to Phred scores higher than 30, remove the Illumina sequencing adapters and convert the FastQ files to FASTA [4]. The Gene Indices Sequence Cleaning and Validation script (SeqClean) was used to remove pEpiFOS™-5 and EPI300™-T1R DNA sequences using a minimum identity of 99% and 97% respectively [5]. The base genome of *E. coli* DH10B was used to remove the EPI300™-T1R DNA sequences, and the sequence of pCC1FOS was used to remove the pEpiFOS™ DNA sequences.

2.5. Taxonomic and functional insight

The processed sequences were uploaded to the Metagenomics Rapid Annotation using Subsystems Technology server (MG-RAST, www.mg-rast.org) [6]. This platform was used to perform *in-silico* analyses of the sequences. The taxonomic (Fig. 1.) and functional (Fig. 2.) profiles were generated using the RefSeq and Subsystem database, respectively.

Acknowledgements

The work was supported by the Research training Initiative for Student Enhancement: Enhancing Biomedical Achievements in Science and Engineering (RISE-E-BASE) program [NIH-5R25GM127191-02] and by the United Stated Department of Agriculture [USDA-CSREES: 2007- 02386]. Special thanks to Luis Morales-Vale for his help during the sequence processing and to Jo Handelsman Laboratory for their training on metagenomics library generation.

Conflict of 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] Bosque Estatal de Guanica, Bosques de Puerto Rico, Departamento de Recursos Naturales y Ambientales de Puerto Rico, Hojas de Nuestro Ambiente, November 2008, p. 034. http://drna.pr.gov/historico/biblioteca/publicaciones/hojas-de-nuestro-ambiente/34-Guanica.pdf. (Accessed 21 October 2008).
[2] J.M. Cruz, M.A. Ortega, J.C. Cruz, P. Ondina, R. Santiago, C. Ríos-Velázquez, Unraveling activities by functional-based approaches using metagenomic libraries from dry and rain forest soils in Puerto Rico. Current Research Technology and Education Topics, Appl Microbiol Microbial Biotechnol 2 (2) (2010) 1471–1478.
[3] FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[4] FASTX. http://hannonlab.cshl.edu/fastx_toolkit/.
[5] SeqClean. https://sourceforge.net/projects/seqclean/.
[6] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, J. Wilkening, The metagenomics RAST serverea public resource for the automatic phylogenetic and functional analysis of metagenomes, BMC Bioinf. 9 (1) (2008) 386.