Data in Brief 41 (2022) 107989

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

# Metabarcoding data of prokaryotes and eukaryotes inhabiting the phosphogypsum stockpiles on the salt marshes of Huelva (SW Spain)

Patricia Gómez-Villegas[a], José Luis Guerrero[b], Miguel Pérez-Rodríguez[c], Juan Pedro Bolivar[b], Javier Vigara[a], Rosa León[a,*]

[a] *Laboratory of Biochemistry, Center for Natural Resources, Health and Environment (RENSMA), University of Huelva, Avda. de las Fuerzas Armadas s/n, Huelva 21071, Spain*
[b] *Department of Integrated Sciences, Center for Natural Resources, Health and Environment (RENSMA), University of Huelva, Avda. de las Fuerzas Armadas s/n, Huelva 21071, Spain*
[c] *Department of Cell Biology, Physiology and Immunology, University of Córdoba, Campus de Excelencia Internacional Agroalimentario CeiA3, Córdoba, Spain*

ARTICLE INFO

*Article history:*
Received 29 October 2021
Revised 28 January 2022
Accepted 18 February 2022
Available online 22 February 2022

*Keywords:*
Metataxonomy
Metabarcoding
16s rRNA
18s rRNA
Phosphogypsum
Extreme environment

ABSTRACT

Around 100 Mt of phosphogypsum (PG) of extreme acidity and with high concentrations of heavy metals and radionuclides have been deposited on the salt marshes of the Tinto River estuary in Huelva (SW Spain) for more than forty years. The microbial community able to thrive in these adverse conditions remains totally unknown, despite the fact that it can highly influence the biogeochemical cycle of the phosphogypsum components and include new species with biotechnological interest. High throughput sequencing of 16S/18S rRNA encoding genes is a potent tool to uncover the microbial diversity of extreme environments. This data article describes for the first time the prokaryotic and eukaryotic diversity of two water samples collected in the Huelva phosphogypsum stacks. The raw amplicons of the 16S/18S rRNA maker genes for the two phosphogypsum samples and two

DOI of original article: 10.1016/j.aquatox.2022.106103
* Corresponding author.
*E-mail address:* rleon@uhu.es (R. León).

https://doi.org/10.1016/j.dib.2022.107989

reference samples (seawater and the Tinto River water) obtained after sequencing on MiSeq platform are provided. The operational taxonomic units (OTUs) obtained after the treatment and clustering of the obtained reads with the QIIME2 pipeline and their taxonomic assignation performed by comparison with the SILVA database are also presented to complete the information of the article "Exploring the microbial community inhabiting the phosphogypsum stacks of Huelva (SW, Spain) by a high throughput 16S/18S rDNA Sequencing approach".



## Specifications Table

| | |
|---|---|
| Subject | Environmental science |
| Specific subject area | Metataxonomy of environmental samples. |
| Type of data | FASTQ and Excel Tables |
| How the data were acquired | • Raw data were obtained on Illumine MiSeq platform (Illumina, San Diego, California) with Illumina MiSeq Reagent kit V2 (2 × 250 bp)<br>• Processed data were obtained with QIIME 2 (v2020.8) software from the raw data<br>• Operational taxonomic units were taxonomically classified by comparison with the SILVA database |
| Data format | Raw and analysed |
| Description of data collection | Two of the samples were collected from two different water bodies of the phosphogypsum stacks deposited on the Tinto salt marshes of Huelva, Spain<br>PC: Perimeter channel, superficial water<br>UTM coordinates 29 S 684,521, 4,123,390<br>PZ: Piezometer, 3–4 m depth<br>UTM coordinates 29 S 684,536, 4,123,295<br>Other two samples were collected from the seawater and the Tinto River<br>SW: Seawater, near the city of Huelva<br>UTM coordinates 29 S 687,322, 4,113,035<br>TR: Tinto River, near the town of Niebla<br>UTM coordinates 29 S 706,089, 4,138,023<br>Genomic DNA was extracted from each sample and used as template for amplification and sequencing of the corresponding hypervariable regions of the 16S/18S rDNA marker genes |
| Data source location | • Institution: University of Huelva<br>• City/Town/Region: Huelva<br>• Country: Spain |
| Data accessibility | As supplementary material with this article and in the Dryad repository (10.5061/dryad.18931zczx) |
| Related research article | EXPLORING THE MICROBIAL COMMUNITY INHABITING THE PHOSPHOGYPSUM STACKS OF HUELVA (SW SPAIN) BY A HIGH THROUGHPUT 16S/18S rDNA SEQUENCING APPROACH. Submitted to Aquatic toxicology. In press<br>https://doi.org/10.1016/j.aquatox.2022.106103 |

## Value of the Data

- This dataset provides information on the prokaryotic and eukaryotic microbial population present in the phosphogypsum stacks of Huelva, revealing an unexpected biodiversity
- Raw data obtained from Illumina MiSeq sequencing platform can be processed with different bioinformatics pipelines to analyze the microbial population of the sampled locations.
- This dataset demonstrates that high throughput sequencing of the 16S/18S rRNA genes of environmental samples is a potent tool for the metataxonomic analysis of microbial communities.
- The data provide useful information that can serve to compare the microbial population of this highly polluted and acidic environment with other similar locations in the world.
- This dataset can reveal the existence of new extremophilic species with interesting biotechnological applications.

## 1. Data Description

The data presented in this paper are the results of sequencing the hypervariable regions of the 16S rRNA and 18S rRNA encoding genes for the genomic DNA samples obtained from four environmental samples. Two of these samples were obtained at the south of the zone 2 of the phosphogypsum stacks located on the Tinto salt marshes in Huelva (Fig. 1). The other two were obtained from the seawater and Tinto River and included in this study as reference samples.

The Genomic DNA from the two phosphogypsum locations and the two reference samples was isolated and its quality was assessed (Table 1).

The data corresponding to the forward and reverse raw pair-end sequences (without barcode and primer sequences) obtained after sequencing the V3-V4 hypervariable regions of the 16S

**Fig. 1.** Aspect of the phosphogypsum stacks of Huelva.

**Table 1**
Purity of the Genomic DNA isolated from the two phosphogypsum and the two reference samples.

| Sample description | Sample Name | Concentration (ng/µl) | 260/280 ratio | Volumen (µl) | % G + C |
|---|---|---|---|---|---|
| Tinto River (TR) | D3_16S | 33.74 | 1.81 | 25 | 55.27 |
| Seawater (SW) | D4_16S | 5.84 | 1.58 | 25 | 52.55 |
| Perimeter channel (PC) | D7_16S | 6.27 | 2.04 | 20 | 55.24 |
| Piezometer (PZ) | D8_16S | 4.09 | 1.59 | 20 | 55.35 |
| Tinto River (TR) | D3_18S | 33.74 | 1.81 | 25 | 54.08 |
| Seawater (SW) | D4_18S | 5.84 | 1.58 | 25 | 50.86 |
| Perimeter channel (PC) | D7_18S | 6.27 | 2.04 | 20 | 50.72 |
| Piezometer (PZ) | D8_18S | 4.09 | 1.59 | 20 | 53.80 |

**Table 2**
List of FASTQ files included as supplementary material which contain the raw data (forward and reverse) obtained from the sequencing the 16sRNA and the 18sRNA libraries: PC, perimeter channel; PZ, piezometer; TR, Tinto River; SW, seawater.

| 16S rRNA | | | 18S rRNA | | |
|---|---|---|---|---|---|
| D7_16S_1.fq | PC, forward | Suppl. Mat S1 | D7_18S_1.fq | PC, forward | Suppl. Mat S2 |
| D7_16S_2.fq | PC, reverse | Suppl. Mat S1 | D7_18S_2.fq | PC, reverse | Suppl. Mat S2 |
| D8_16S_1.fq | PZ, forward | Suppl. Mat S1 | D8_18S_1.fq | PZ, forward | Suppl. Mat S2 |
| D8_16S_2.fq | PZ, reverse | Suppl. Mat S1 | D8_18S_2.fq | PZ, reverse | Suppl. Mat S2 |
| D3_16S_1.fq | TR, forward | Suppl. Mat S1 | D3_18S_1.fq | TR, forward | Suppl. Mat S2 |
| D3_16S_2.fq | TR, reverse | Suppl. Mat S1 | D3_18S_2.fq | TR, reverse | Suppl. Mat S2 |
| D4_16S_1.fq | SW, forward | Suppl. Mat S1 | D4_18S_1.fq | SW, forward | Suppl. Mat S2 |
| D4_16S_2.fq | SW, reverse | Suppl. Mat S1 | D4_18S_2.fq | SW, reverse | Suppl. Mat S2 |

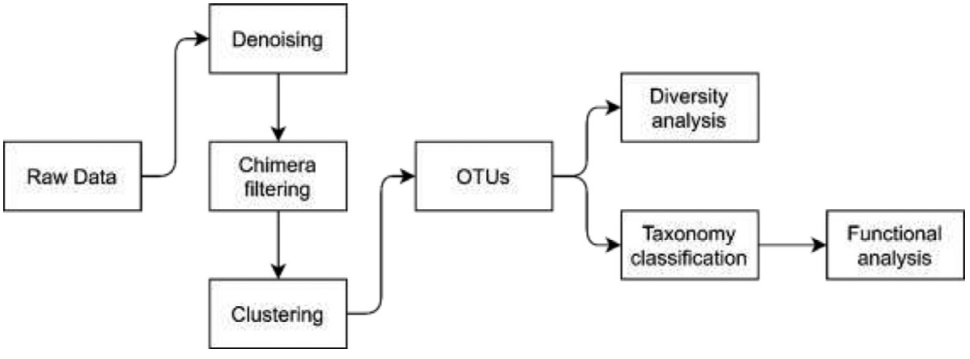


**Fig. 2.** Workflow of data analysis from obtaining of the raw reads to the species annotation.

**Table 3**
Number of raw, filtered and merged reads, and the number of clustered Operational Taxonomic Unit, for each sample. Adapted from Gómez-Villegas et al. 2022 [1].

| Code | Sample | Raw Reads | After denoising | Merged inputs | Non chimeric reads | Mean quality | | Observed OTUs |
|---|---|---|---|---|---|---|---|---|
| | | | | | | $Q_{30}$ (%) | Q Score | |
| D7_16S | PC_16S | 19,129 | 14,995 | 11,338 | 10,555 | 94.60 | ≥ 36 | 680 |
| D7_18S | PC_18S | 10,067 | 8864 | 8716 | 6076 | 93.36 | ≥ 36 | 38 |
| D8_16S | PZ_16S | 17,084 | 13,236 | 10,254 | 9409 | 94.54 | ≥ 36 | 596 |
| D8_18S | PZ_18S | 30,362 | 21,890 | 19,607 | 15,080 | 89.74 | ≥ 36 | 186 |
| D3_16S | TR_16S | 189,164 | 175,352 | 145,395 | 145,395 | 95.12 | ≥ 36 | 348 |
| D3_18S | TR_18S | 160,136 | 126,280 | 123,047 | 103,365 | 90.99 | ≥ 36 | 133 |
| D4_16S | SW_16S | 206,161 | 190,181 | 179,924 | 171,677 | 95.17 | ≥ 36 | 838 |
| D4_18S | SW_18S | 203,677 | 172,627 | 162,688 | 111,906 | 92.24 | ≥ 36 | 399 |

*Sample names:* PC, perimeter channel; PZ, piezometer; TR, Tinto River; SW, Seawater.

rRNA gene (supplementary material, S1) and the V9 hypervariable regions of the 18S rRNA gene (supplementary material, S2) are available in the Supplementary Material of this publication in compressed FASTQ format files Table 2. summarizes the naming for the included files.

For prokaryotes, the raw reads obtained were 19,129 in the Perimeter channel and 13,236 in the Piezometer. For eukaryotes, the raw reads were 10,067 in the Perimeter channel and 21,890 in the Piezometer. These raw data was treated as indicated in the schematic workflow (Fig. 2) to yield 680 prokaryotic and 38 eukaryotic Operational Taxonomic Units in the Perimeter channel, and 596 prokaryotic and 186 eukaryotic Operational Taxonomic Units in the Piezometer.

The number of reads and of Operational Taxonomic Units generated from these raw sequences after denoising, merging, chimera filtering and clustering are summarized in Table 3.

The obtained Operational Taxonomic Units were classified at different taxonomic levels by comparison with the SILVA database as described in material and methods. The results for the prokaryotic and eukaryotic microorganisms are available in the supplementary material (files S3 and S4).

All effective tags grouped by 97% DNA sequence similarity into Operational Taxonomic Units are compiled in supplementary, as well as their classification at different taxonomic levels by comparison with SILVA database as described in material and methods. The results for the prokaryotic (Suppl. Mat. File S3) and eukaryotic (Suppl. Mat. File S4) microorganisms are available in the supplementary material of this article and in Dryad repository (10.5061/dryad.18931zczx). A detailed comparative analysis of the most abundant genera at each location is shown elsewhere [1].

## 2. Experimental Design, Materials and Methods

### *2.1. Collection of samples*

Two of the samples were collected from two different water bodies of the phosphogypsum stacks deposited on the Tinto salt marshes in Huelva, Spain. The first sample was collected from the perimeter drainage channel that surrounds the zone 2, collecting leachates from the stored phosphogypsum (UTM coordinates 29 S 684,521, 4,123,390). The second one was taken from a piezometer located in the border of the same zone that receives underground leachates with a depth of 3–4 m (UTM coordinates 29 S 684,536, 4,123,295). Samples from the seawater and the lower course of the Tinto River, collected near the cities of Huelva (UTM coordinates 29 S 687,322, 4,113,035) and Niebla (UTM coordinates 29 S 706,089, 4,138,023), respectively, have also been included for comparison.

### *2.2. Genomic DNA extraction*

Genomic DNA isolation was performed using the GeneJet Genomic Purification Kit (Thermo Fisher Scientific, Waltham, MA, USA) and the biomass obtained from 10 L of water from each of the described locations. The biomass was obtained by filtering through 0.7 µm glass fiber filters (Whatman, GF/G) and centrifugation at 12 000 x g. The genomic DNA was quantified using Nanodrop Spectrophotometer ND-1000 (Thermo Fisher Scientific). The quality of the obtained DNA was verified by the A260/A280 ratio and by electrophoretic analysis in 2% agarose gel as previously described [2].

### *2.3. Library construction and amplicon sequencing*

The hypervariable V3-V4 16S rDNA region was amplified with primers 1380F/1510R, and the corresponding eukaryotic hypervariable V9 18S rDNA region was amplified with the primers 341F/806R. In both cases, amplifications were done with the Phusion® High-Fidelity PCR Master Mix (New England Biolabs, MA, USA) and using the genomic DNA isolated at the indicated points as a template. For each sample, the amplicons were purified with the Qiagen Gel Extraction Kit (Qiagen, Germany), pooled and used to generate two libraries, one for the 18S rRNA and one for the 16S rRNA, using NEBNext® UltraTM DNA Library Prep Kit for Illumina. The quality of the eight libraries generated was assessed and quantified using Qubit and agarose electrophoresis as QC procedures. The pooled libraries were sequenced with the Illumina MiSeq Reagent kit V2, using a 2 $\times$ 250 bp paired-end strategy, in the Illumine MiSeq platform (Illumina, San Diego, California) following the manufactureś protocol.

### 2.4. Bioinformatics and data analysis

The analysis of the raw data was carried out using QIIME 2 (v2020.8) [3]. First of all, raw data were demultiplexed, using the q2-demux plugin, and filtered, to get clean data, by trimming and truncating low-quality regions, dereplicating reads, and filtering chimeras, using DADA2 [4] (via q2-dada2). Then, the reads were organized in operational taxonomic units using *de novo* clustering method (via q2-vsearch) from VSEARCH [5]. The clustering was performed grouping at 97% identity to create 97% operational taxonomic units . The operational taxonomic units were classified at each taxonomic rank using the *q2-feature-classifier* plugin (via *classify-sklearn* method) and the SILVA database [6]. The SILVA database was applied as two different pre-trained classifiers, specially curated, for 16SV3V4 and 18SV9 regions sequenced. Annotation was performed with a 0.7 threshold.

## Ethics Statements

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

## 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.

## CRediT Author Statement

**Patricia Gómez-Villegas:** Investigation, Writing – review & editing; **José Luis Guerrero:** Investigation, Writing – review & editing; **Miguel Pérez-Rodríguez:** Software, Data curation, Writing – review & editing; **Juan Pedro Bolivar:** Writing – review & editing; **Javier Vigara:** Supervision, Writing – review & editing; **Rosa León:** Conceptualization, Supervision, Writing – review & editing.

## Acknowledgments

The authors thank Fertiberia S.A. for its support in obtaining the water samples used in this study. P. Gómez-Villegas acknowledges the financial support of the University of Huelva (EPIT 2016–17). This research was funded by University of Huelva and the Operative FEDER Program-Andalucía 2014-2020 (UHU-1257518 and UHU-1255876); The SUBV. COOP.ALENTEJO-ALGARVE-ANDALUCIA 2021; The European Regional Development Fund through the Agencia Estatal de Investigación (research grant PID 2019–110438RB-C22) and the Andalucian government ($I + D + i$-JA-PAIDI-Retos project 2020- PY20_00728).

## Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2022.107989.

## References

[1] P. Gómez-Villegas, J.L. Guerrero, M. Pérez-Rodriguez, J.P. Bolívar, A. Morillo, J. Vigara, R. León, Exploring the microbial community inhabiting the phosphogypsum stacks of Huelva (SW, Spain) by a high throughput 16S/18S rDNA Sequencing approach, Aquat. Toxicol. (2022) 106103, doi:10.1016/j.aquatox.2022.106103.
[2] P. Gómez-Villegas, J. Vigara, R. León, Characterization of the microbial population inhabiting a solar saltern pond of the Odiel Marshlands (SW Spain), Mar. Drugs 16 (9) (2018) 332, doi:10.3390/md16090332.
[3] E. Bolyen, J.R. Rideout, M.R. Dillon, N.A. Bokulich, C.C. Abnet, G.A. Al-Ghalith, H. Alexander, E.J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J.E. Bisanz, K. Bittinger, A. Brejnrod, C.J. Brislawn, C.T. Brown, B.J. Callahan, A.M. Caraballo-Rodríguez, J. Chase, … J.G. Caporaso, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, Nat. Biotechnol. 37 (8) (2019) 852–857, doi:10.1038/s41587-019-0209-9.
[4] B.J. Callahan, P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, S.P. Holmes, DADA2: high-resolution sample inference from Illumina amplicon data, Nat. Methods 13 (7) (2016) 581–583, doi:10.1038/nmeth.3869.
[5] T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics, PeerJ 4 (2016) e2584, doi:10.7717/peerj.2584.
[6] 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, doi:10.1093/nar/gks1219.