Data in Brief 51 (2023) 109726

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Sediment microbiome diversity and functional profiles of unprotected arid-tropical natural wetlands in South Africa revealed by shotgun metagenomics data

Henry Joseph Oduor Ogola [a], Grace N. Ijoma [a,*], Joshua Nosa Edokpayi [b]

[a] Department of environmental Sciences, College of Agriculture and Environmental Sciences, University of South Africa (UNISA), Florida Campus, Roodepoort, 1709, South Africa
[b] Water and Environmental Management Research Group, Faculty of Science, Engineering and Agriculture, University of Venda, Thohoyandou 0950, South Africa

ARTICLE INFO

Article history:
Received 24 May 2023
Revised 22 October 2023
Accepted 23 October 2023
Available online 27 October 2023

Dataset link: Shotgun metagenomics data of sediment microbiome of unprotected arid-tropical natural wetlands in South Africa, Mendeley Data, V1 (Original data)
Dataset link: Shotgun sequencing of sediment microbiome unprotected arid-tropical natural wetland in northeastern South Africa - BioProject (Original data)

Keywords:
Shotgun metagenomics
Microbial diversity
Functional annotation
bioBakery 3

ABSTRACT

The Limpopo province, located in the arid-tropical region in northeastern South Africa, is renowned for its diverse natural wetlands, some of which are currently unprotected. These wetlands play a crucial role in preserving biodiversity, purifying water, controlling floods, and supporting agricultural production for rural communities. Unfortunately, human activities such as agricultural effluents, run-offs, domestic wastewater, and plastics pollution, along with the impacts of climate change, are mounting pressures on these ecosystems. However, there is limited information on the microbial ecology of natural wetlands in this region, considering the changing anthropogenic activities. The data presented represents the first report on the microbial and functional diversity of sediment microbiomes associated with unprotected arid-tropical natural wetlands in South Africa. Metagenomic shotgun sequencing was performed on sediment samples from ten different wetlands using the Illumina NextSeq 2000 platform. Taxonomic profiling of 328,625,930 high-quality sequencing reads using the MetaPhlAn v3.0 pipeline revealed that Bacteria were the most abundant kingdom (54.5 %), followed by

* Corresponding author.
E-mail address: ijomagn@unisa.ac.za (G.N. Ijoma).

https://doi.org/10.1016/j.dib.2023.109726

Viruses (0.40 %), Archaea (0.01 %), and Eukaryota (0.36 %). Among bacteria, the most prevalent taxa belonged to the phylum Proteobacteria, particularly the classes Gammaproteobacteria and Betaproteobacteria, which accounted for 83 % of bacterial sequences. The Terrabacteria group, consisting of the phyla Firmicutes and Actinobacteria, made up 3 % of the bacterial population. The abundance of these top bacterial taxa varied across different wetland samples, both at the genus and species levels. In addition, hierarchical clustering based on Bray-Curtis dissimilarity distances of fungal, protist, archaea, and virus species showed distinct clustering of sediment samples from different wetlands. Functional annotation of the metagenomes identified 1224–1702 enzyme classes, 84,833–198,397 gene families, and 280–400 pathways across the various wetland sediments. The data provide crucial baseline information on the microbial and functional diversity of sediment communities in arid tropical wetlands. This knowledge will contribute to a better understanding of these unique environments and can aid in their management and conservation efforts in rural South Africa.

Specifications Table

Subject	Microbial Ecology, Genomics and Molecular Biology
Specific subject area	Metagenomics
Type of data	Figures and fastq files
How the data were acquired	Shotgun DNA sequencing using Illumina NextSeq 2000 platform.
Data format	Raw (fastq.gz files) and analyzed data in Krona charts, Stacked plots, heatmaps and tables
Description of data collection	Sediment samples were collected 10 different small unprotected wetlands (bogs and marshes) representative of the typical arid-tropical natural wetlands in Limpopo province of South Africa. Environmental DNA was purified from the composite sediment samples of each wetland using standardized methods, and DNA libraries sequenced on an Illumina NextSeq 2000 platform with 2 × 150 bp sequencing. Raw data was analyzed using bioBakery 3 platform consisting of sequence-level quality control and contaminant depletion guidelines (KneadData), MetaPhlAn v3.0 for taxonomic profiling, and HUMAnN v3.0 for functional profiling.
Data source location	Institution: University of Venda and University of South Africa City/Town/Region: Thohoyandou, Limpopo Province Country: South Africa Latitude and longitude (and GPS coordinates) for collected samples/data: S1 (24° 0′ 35″, S 29° 30′ 36″ E); S2 (23° 29′ 24″ S 29° 24′ 39.6″ E); S3 (24° 37′ 51.6″ S, 29° 44′ 2.4″ E); S4 (24° 43′ 33.6″ S, 29° 46′ 8.4″ E); S5 (23° 38′ 60″ S, 27° 45′ 50.4″ E); S6 (23° 34′ 48″ S, 28° 6′ 28.8″ E); S7 (23° 53′ 27.6″ S, 30° 16′ 33.6″ E); S8 (23° 53′ 42″ S, 30° 15′ 25.2″ E); S9 (22° 59′ 2.4″ S, 30° 26′ 34.8″ E); and S10 (22° 58′ 26.4″, 30° 29′ 9.6″ E)
Data accessibility	The SRA data are publicly available under the following repository: Repository name: National Center for Biotechnology Information Data identification number: BioProject ID PRJNA972844 and SRA accession numbers SRX20358958 to SRX20358949 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/PRJNA972844 The Supplementary Data related to these datasets are available under the following repository: Repository name: Mendeley Data

(*continued on next page*)

Data identification number: DOI: 10.17632/mm25y745hz.1
Direct URL to the data: https://doi.org/10.17632/mm25y745hz.1

1. Value of the Data

- The data generated can be used to offer valuable insights into sediment microbial diversity, ecological dynamics and potential function in Limpopo Province's arid-tropical wetlands.
- This dataset establishes a baseline for understanding the stability and health of the fragile yet unprotected wetland ecosystems, and monitoring the impact of pollution and climate change in such ecosystems.
- The publicly available raw datasets can support further studies, such as the analysis antibiotic resistance genes, which can help enhance our understanding of the wetland ecosystem's potential implications for human health.

2. Objective

Despite the diverse unprotected natural wetlands in the arid-tropical Limpopo province, South Africa, which provide ecosystem services supporting rural livelihoods through agricultural production [1,2], they are facing mounting pressures from various anthropic activities [1–3]. These pressures are aggravated by the effects of rapid climate change, resulting in prolonged seasonality and overexploitation of some wetlands [4]. Herein, our objective was to investigate the microbial community structure, diversity, species richness, and functional metabolic potential in the sediment samples collected from ten different natural wetlands in the Limpopo province, South Africa. The ultimate goal is to delineate the current ecological dynamics and health of these ecosystems and understand their implications for future economic well-being and public health of the dependent rural population in the catchment area. To achieve this, we employed the high-throughput Illumina NextSeq 2000 platform for shotgun sequencing of the sediment microbiome.

3. Data Description

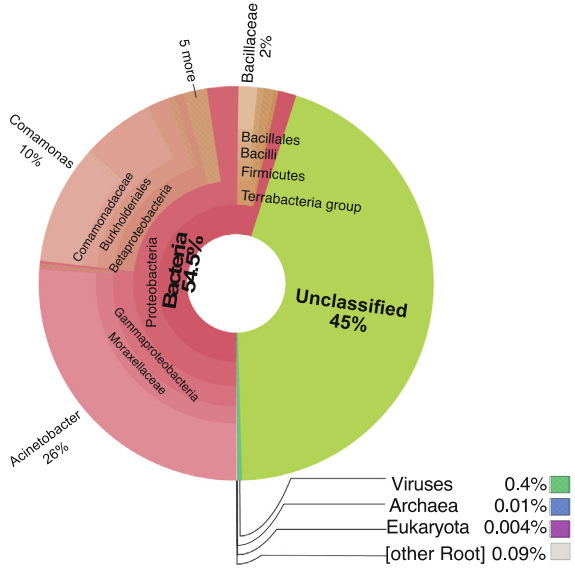
3.1. Taxonomic profiling of microbial community

Sediment samples from 10 different small unprotected wetlands, mainly bogs and marshes 1 acre in size, representative of the typical arid-tropical natural wetlands of Limpopo Province (Supplementary Data Fig S1) [5], South Africa, were collected in August 2021. The raw fastq files has been deposited in NCBI SRA database as BioProject ID PRJNA972844 and Biosample accession numbers SRX20358958 to SRX20358949 (https://www.ncbi.nlm.nih.gov/sra/PRJNA972844) [6–15]. Overall, Illumina NextSeq 2000 platform sequencing generated 328,625,930 quality sequencing reads, with 54.5 % were assigned to Bacteria, 0.4 % to Viruses, 0.01 % to Archaea, and 0.004 % to Eukaryota, while other reads were not classified or unassigned (Fig. 1a). For bacterial taxa, phylum *Proteobacteria* (with >80% of proteobacterial sequences being members of class *Gammaproteobacteria* and *Betaproteobacteria*), and Terrabacteria group (phylum *Firmicutes* and *Actinobacteria*) were the most abundant (Fig. 1a) across the sampling sites. The most abundant (>0.1 % relative abundance bacterial, archaeal, fungi and viruses' phyla detected in each sampling sites is provided in Fig 1b.

3.2. Functional annotation of metabolism pathways and gene families

HUMAnN v3.0 tool [5] was used to investigate pathway inference and gene families with Pfam domains. A total of 1224–1702, 84,833-198,397 and 280–400 features identified as enzyme

a)



b)

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Bacteria; Proteobacteria	73.7	93.2	87	17.2	82.6	74.9	81.5	87.4	82.9	67.8
Bacteria; Firmicutes	19.2	1.9	3.8	57.6	11.9	21.3	10.9	5.9	11.4	22.8
Bacteria; Actinobacteria	2.4	2.2	4.6	11.8	0.7	1.1	2.9	1.9	3.4	1.8
Bacteria; Bacteroidetes	2.7	1.1	1.3	7.2	3.4	1.4	2.6	2.7	1.3	5.3
Bacteria; k__Bacteria_Otu8074	1	0.9	1.3	2.6	0.6	0.7	1.2	1	0.5	1
Viruses; Uroviricota	0	0	0.6	0.5	0.2	0.1	0	0.5	0	0.3
Bacteria; Planctomycetes	0.2	0.1	0.4	0.4	0	0	0.1	0.1	0	0.1
Archaea; Euryarchaeota	0.1	0.1	0.1	0.7	0	0	0.1	0	0	0.1
Bacteria; Cyanobacteria	0.2	0.1	0.2	0.4	0.1	0.1	0.1	0.1	0.1	0.1
Bacteria; Fusobacteria	0.1	0	0	0.2	0.1	0	0.1	0	0.1	0.1
Bacteria; Acidobacteria	0.1	0	0.1	0.2	0	0	0	0	0	0
Eukaryota; Ascomycota	0.1	0	0	0.1	0	0	0.1	0	0	0
Bacteria; Spirochaetes	0.1	0	0	0.1	0	0	0	0	0	0.1
Bacteria; Tenericutes	0.1	0	0	0.1	0	0	0	0	0	0.1
Bacteria; Deinococcus-Thermus	0	0	0.1	0.1	0	0	0	0	0	0
Bacteria; Chloroflexi	0	0	0.1	0.1	0	0	0	0	0	0
Bacteria; Verrucomicrobia	0	0	0.1	0.1	0	0	0	0.1	0	0
Bacteria; Gemmatimonadetes	0	0	0.1	0.1	0	0	0	0	0	0
Eukaryota; Chordata	0	0	0	0.1	0	0	0	0	0	0
Bacteria; Thermotogae	0	0	0	0.1	0	0	0	0	0	0

Fig. 1. Taxon abundance identified by MetaPhlAn pipeline of metagenomes of the arid tropical wetland sediments of Limpopo Province, South Africa. a) Krona chart of the overall relative abundance of taxa across the sampling sites. The most dominant bacteria classes were *Gammaproteobacteria* (27 %), *Betaproteobacteria* (19 %), *Bacilli* (3 %), *Clostridia* and *Actinomycetia* at 1 % relative abundance each. The other five betaproteobacterial taxa detected included *Acidovorax, Delftia, Variovorax, Diaphorobacter* and *Hydrogenophaga*. b) Heatmap showing the relative abundance of the major phyla of the metagenomes in each wetland sediment samples. Non-bacterial taxa have been colored differently to improve clarity.

Table 1
Summary of major metabolic pathways by relative abundance (RPK) across the sediment samples metagenomes.

Pathway	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Metabolism	69 %	68 %	70 %	64 %	73 %	69 %	64 %	65 %	68 %	78 %
Human diseases	8 %	8 %	8 %	9 %	8 %	8 %	9 %	11 %	10 %	5 %
Cellular processes	6 %	8 %	6 %	9 %	5 %	7 %	8 %	8 %	5 %	5 %
Genetic information processing	9 %	7 %	5 %	8 %	7 %	7 %	7 %	8 %	6 %	4 %
Organismal systems	4 %	5 %	5 %	5 %	4 %	5 %	7 %	4 %	5 %	4 %
Environmental information processing	4 %	4 %	4 %	4 %	4 %	4 %	5 %	4 %	5 %	3 %

classes, gene families and pathways were inferred from the sediment's metagenomes, indicating the diversity and complexity of metabolic functions within these microbial ecosystems. Summary of the top pathways identified based on average abundance is presented in Table 1. Additionally, the relative abundance of the annotated enzyme classes, and pathways are provided as Supplementary Data Table S1, and S2, respectively. Overall, there is prominence of relative pathway abundance (RPK) in "Metabolism" pathways (ranging from 64–78 %), and the variation in the relative abundance of other pathways categories such as "Human diseases (5–11 %)," "Cellular processes (5–8 %)," "Genetic information processing (4–9 %)," "Organismal systems (4–7 %)," and "Environmental information processing (3–5 %)" among the various sediment samples..

4. Experimental Design, Materials and Methods

4.1. Sampling sites and samples collection

Sediment samples were collected from 10 unprotected natural wetlands spread across the arid-tropical region of Limpopo Province, in the northeastern South Africa. Sediments soil samples were collected at three different sites from the surface of the bed substrate (0–10 cm deep) using dredge sampler (Kajak, KC-Denmark) for each wetland. At each site, a multi-point mixed sampling method was used, where five soil subsamples were collected randomly with an area of 2 $\times$ m and then mixed into one sample (the sediment) for DNA extraction (20 g) and soil properties data (300–400 g. For DNA extraction, approximately 20 g sediments were collected in a centrifuge tube and immediately frozen in liquid N_2 and stored at -80 °C.

4.2. DNA extraction, library preparation and sequencing

Environmental DNA was purified from the composite sediment samples using the DNeasy® PowerSoil Pro Kit (Qiagen, Germany), following the manufacturer's instructions. The quality and quantity of the purified DNA were determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA and the Qubit™ dsDNA BR Assay Kit using a Qubit™ 4 Fluorometer (Thermo Fisher Scientific, USA), respectively. For downstream library preparation and sequencing, DNA extracts with A260:A280 ratio between 1.8 and 2.0 and DNA concentrations of 20–150 ng/μl were used.

DNA libraries were prepared using the Nextera XT® DNA Library Preparation Kit (Illumina Inc., San Diego, CA, United States) and IDT Unique Dual Indexes® Tagmentation Kit (Illumina Inc., San Diego, CA, United States) with total DNA input of 1 ng. Briefly, environmental genomic DNA was initially fragmented using a proportional amount of Illumina Nextera XT fragmentation enzyme. Barcoding of each sampling sites was undertaken by addition of unique dual indexes to each sample followed by 12 cycles of PCR to construct libraries. The resultant DNA libraries were purified using AMpure XP® magnetic beads (Beckman Coulter, Massachusetts, United States) and eluted according to manufacturer's instructions and eluted in QIAGEN EB® buffer (Qiagen, Germany). DNA libraries were quantified using Qubit 4 fluorometer and Qubit™ dsDNA HS Assay

Kit. The sequencing of the constructed libraries was performed on an Illumina NextSeq 2000 platform 2 × 150 bp at CosmosID Inc., (Germantown, MD, USA). The raw high throughput sequencing data has been deposited into the NCBI Sequence Read Archive database as BioProject ID PRJNA972844 and SRA accession numbers SRX20358958 to SRX20358949.

4.3. Sequence pretreatment, assembly and taxonomic identification

Illumina NextSeq2000 platform sequencing generated between 23,618,694 and 42,062,226 paired-end reads per sample. Shotgun sequences was analyzed using bioBakery 3 platform (http://segatalab.cibio.unitn.it/tools/biobakery). bioBakery 3 includes updated sequence-level quality control and contaminant depletion guidelines (KneadData), MetaPhlAn v3.0 for taxonomic profiling, and HUMAnN v3.0 for functional profiling [16]. The primary steps included initial removal of reads mapping to the human reference database and basic quality control using KneadData (https://github.com/biobakery/kneaddata) using default settings. Then, the quality filtered sequences were used for profiling the composition of microbial communities (Bacteria, Archaea and Eukaryotes) using MetaPHlAn v3.0 using default setting [16]. Functional genes and pathways were annotated using HUMAnN v3.0 and its UniRef 50, Pfam, and MetaCyc pathway databases using read sequences that were trimmed and quality filtered using KneadData [16].

4.4. Statistics

Unless stated otherwise, R statistical software v3.5.2 [17] was used for statistics of the data. Heatmaps were generated with the pheatmap package [18] or heatmap.2 function within the gplots package [19].

Ethics Statements

The authors declare that there are no ethical issues with the data presented. The study did not conduct human or animal experiments, and that you did not collect social media data.

Data Availability

Shotgun metagenomics data of sediment microbiome of unprotected arid-tropical natural wetlands in South Africa, Mendeley Data, V1 (Original data) (Mendeley)

Shotgun sequencing of sediment microbiome unprotected arid-tropical natural wetland in northeastern South Africa - BioProject (Original data) (NCBI)

CRediT Author Statement

Henry Joseph Oduor Ogola: Data curation, Formal analysis, Investigation, Writing – original draft, Visualization, Writing – review & editing; **Grace N. Ijoma:** Conceptualization, Investigation, Writing – original draft, Visualization, Funding acquisition, Resources, Writing – review & editing; **Joshua Nosa Edokpayi:** Conceptualization, Funding acquisition, Resources, Writing – review & editing.

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.

Acknowledgments

Authors also would like to extend the acknowledgement to CHPC, Pretoria for providing high performance computing facilities for metagenomics. The authors are thankful to Stanley Skhuna Ndlovu for assisting in sample collection. This research was supported by a grant received from the Department of Science and Innovation (DSI) of South Africa, DST/CON 0197/2017, administered via the Technology Innovation Agency (TIA).

References

[1] E. Mandishona, J. Knight, Inland wetlands in Africa: a review of their typologies and ecosystem services, Prog. Phys. Geogr. Earth Environ. 46 (2022) 547–565, doi:10.1177/03091333221075328.
[2] W. Jogo, R. Hassan, Balancing the use of wetlands for economic well-being and ecological security: the case of the Limpopo wetland in southern Africa, Ecol. Econ. 69 (2010) 1569–1579, doi:10.1016/j.ecolecon.2010.02.021.
[3] J.M. Letsoalo, M.J. Potgieter, Domestic waste disposal in a small urban wetland area by Ga-Makanye Community, Limpopo Province, South Africa, South African Geogr. J. 103 (2021) 374–380, doi:10.1080/03736245.2020.1824804.
[4] A.O. Adeeyo, S.S. Ndlovu, L.M. Ngwagwe, M. Mudau, M.A. Alabi, J.N. Edokpayi, Wetland resources in South Africa: threats and metadata study, Resources 11 (2022), doi:10.3390/resources11060054.
[5] H.J.O. Ogola, G. Ijoma, J. Edokpayi, Shotgun Metagenomics Data of Sediment Microbiome of Unprotected Arid-Tropical Natural Wetlands in South Africa, Mendeley Data, 2023 V1, doi:10.17632/mm25y745hz.1.
[6] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358958 (2023).
[7] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358957 (2023).
[8] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358956 (2023).
[9] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358955 (2023).
[10] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358954 (2023).
[11] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358953 (2023).
[12] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358952 (2023).
[13] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358951 (2023).
[14] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358950 (2023).
[15] *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRX20358949 (2023).
[16] F. Beghini, L.J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A.M. Thomas, M. Valles-Colomer, G. Weingart, Y. Zhang, M. Zolfo, C. Huttenhower, E.A. Franzosa, N. Segata, Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3, Elife 10 (2021) e65088, doi:10.7554/eLife.65088.
[17] R Core Team, R: A Language and Environment for Statistical Computing (Version 3.5. 2), 2018, R Foundation for Statistical Computing, Vienna, Austria, 2019 https://www.R-project.org.
[18] R. Kolde, Pheatmap: Pretty Heatmaps, R Package, 2012 Version. 1 726 https://CRAN.R-project.org/package=pheatmap.
[19] G.R. Warnes, B. Bolker, L. Bonebakker, R.W. Gentleman, H.A. Liaw, T. Lumley, gplots: Various R programming Tools For Plotting Data, R package version 3.0. 1.1. 2019, 2019 https://CRAN.R-project.org/package=gplots.