

Geographic blind spots in an extreme environment microbiome catalog

Matters Arising

Arising from *Jiang et al. Nat Commun* **17**, 4791 (2026). <https://doi.org/10.1038/s41467-026-71145-0>

Francisco P. Chávez¹*

¹ Laboratorio de Microbiología de Sistemas, Departamento de Biología, Facultad de Ciencias, Universidad de Chile, Santiago, Chile.

*Corresponding author: fpchavez@uchile.cl

Jiang *et al.* present the Extreme Environment Microbiome Catalog (EEMC), comprising 78,213 bacterial and archaeal genomes, as a global resource for microbial diversity and antimicrobial discovery¹. Its scale and experimental peptide characterization are substantial contributions. However, a global framing requires evidence that regional data were systematically sought, rather than simply aggregated. A targeted accession audit identifies public datasets whose project identifiers are absent from the published EEMC project list, including datasets from Latin America, Africa and the Global North. These findings identify regional evidence requiring reconciliation with catalog holdings and eligibility criteria; they do not establish whether exclusion rates differ between regions.

We compared EEMC's 238 project-level entries with public ENA records, extending the earlier WGS/metagenomic audit through publication-based accession checks. The expanded register contains 31 candidate projects whose identifiers are absent from the published project list, comprising 212 runs from 204 unique samples and 2.186 trillion raw sequenced bases made public by 29 February 2024 (Figure 1;

Supplementary Table 1). Thirty projects contain WGS-labeled metagenomic libraries; one contains metagenomes labeled WGA despite explicitly amplification-free preparation. These are verified absences from the published project list, not proof that every sample satisfies every inclusion rule or that equivalent genomes could not appear through another accession. Individual-accession reconciliation and documented exclusions are therefore essential.

Nineteen projects concern Chilean environments, three African environments, three Israeli desert projects, two Australian environments, one Spanish saltern project, two Antarctic projects and one multi-region desert project²⁻¹⁶. Under an explicitly stated UNCTAD-based country grouping, 22 concern the Global South and six the Global North; Antarctica and mixed-country sampling remain separate. These counts describe the audited set, not exclusion probabilities. Candidate discovery was targeted rather than geographically exhaustive, and regional counts may reflect differences in search coverage and investigator familiarity with the literature. The concentration of candidates in Chile therefore cannot, by itself, establish preferential exclusion of Global South datasets. Particularly, 14 endolithic BioProjects originate from one study, including 13 newly resolved here⁷. These are culture-derived metagenomes of cyanobacteria and their cohabitants, requiring explicit reconciliation with EEMC's genomic inclusion routes. They are not equivalent to unmanipulated environmental samples. Projects must not be treated as independent publications or given equal ecological weight without considering samples, depth, and environmental diversity. Raw sequence volume is not a measure of additional catalog-quality genomes, taxonomic novelty or antimicrobial-peptide diversity. Those contributions depend on eligibility, sequence quality and overlap with existing EEMC holdings. The present audit identifies datasets requiring assessment but does not quantify their incremental biological contribution.

The added Chilean datasets span El Tatio hot springs, virus-enriched fractions from the high-altitude Huasco wetland, and microbial mats from Salar de Llamara¹³⁻¹⁵. Enrichment for viruses does not invalidate recoverable bacterial or archaeal genomic material. Huasco contains prokaryotic sequences, although no claim is made here that it deposits MAGs meeting EEMC's quality thresholds. Similarly, WGA label from

Llamara should not override its published amplification-free methods. Eligibility should follow the actual genomic material and declared catalog rules, not a thematic emphasis study or a repository label alone.

The audit also identifies six cellular metagenomes from Spanish saltern ponds within PRJNA756245, spanning 6, 14 and 32 percent salinity¹⁶. Other freshwater, ordinary marine and isolated viral fractions in that project are not added to the environmental audit. These northern examples show why a coverage assessment should search both North and South. Further Antarctic resistome studies and genome deposits are recorded separately where raw-read dates, sample reuse, or individual genome accessions remain unresolved^{17–21}. Their bibliographic absence warrants screening, not an inflated quantitative total. EEMC's inclusion of Namib and Mackay Glacier projects likewise prevents a claim of categorical regional exclusion.

Geographic coverage matters because extreme habitats are unevenly distributed. A global assessment estimates that more than 70 percent of dryland area lies in developing countries and includes hyperarid, arid, semiarid and dry-subhumid systems²². This is not a census of all extreme environments, and developing countries are not a universal synonym for the Global South. It nevertheless establishes a strong ecological reason to scrutinize coverage of dryland regions in such countries. Sampling priorities should reflect habitat extent, environmental contrasts and poorly characterized lineages, rather than treating regions as interchangeable map points. Quantifying whether sampling is adequate requires environment-specific denominators; the present audit cannot establish that the South contains most of every extreme habitat.

Unequal access to sequencing and bioinformatics infrastructure adds a further reason for careful ascertainment: these resources are often concentrated in high-income countries²³. Where research budgets and technical capacity are constrained, producing costly shotgun metagenomic datasets can require a substantial share of available resources. Overlooking eligible, publicly available datasets from such settings can therefore compound an existing inequity: barriers first limit data generation, and incomplete synthesis then limits the contribution of data that were successfully produced. This is an argument for systematic discovery and transparent eligibility assessment, not preferential inclusion or relaxed genomic quality criteria.

Temporal coverage also needs explicit treatment. Five unlisted projects released 42 WGS runs from 41 samples, totaling 0.597 Tbp, after the cutoff but before submission on 23 June 2025: Union Glacier, Laguna Lejía, Pampa Lirima, Barrancas Blancas and north Danakil brines²⁴⁻²⁶. The five Danakil shotgun samples became public in October 2024; registration of their shared BioProject in 2019 does not make them pre-cutoff data. A further Antarctic release occurred during review²⁶. These are update opportunities rather than cutoff violations. A versioned corpus and an explicit refresh policy would make the temporal meaning of ‘global’ assessable.

Participation is a separate dimension. EEMC’s 26 authors list institutions only in China and Serbia: 21 and five, respectively; none lists a Latin American or African institution. China’s classification within the Global South precludes describing this as absent Global South authorship. Conversely, the six authors of Llamara 2018 and the six of Danakil 2024 all list French institutions, with no Chilean or Ethiopian institutional affiliation, respectively^{15,27}. Both have a France-affiliated corresponding author. These are selected examples, not a prevalence estimate or evidence of nationality, contribution quality, or absent local engagement. They show that southern sample provenance and locally anchored scientific participation are different properties.

Broader research documents unequal local authorship in international conservation, while microbiological genomics has explicitly called for inclusive contribution recognition and stronger local participation^{23,28}. Those findings cannot be transferred quantitatively to extreme-environment microbiology. They support seeking local researchers and relevant community expertise from study design through analysis, interpretation, and leadership, including corresponding authorship when contributions warrant it. Regional specialists can also help discover datasets and resolve ecosystem terminology in global syntheses. Substantive collaboration should be encouraged without imposing honorary authorship or inferring misconduct from institutional addresses.

Editors evaluating global catalogs should require reproducible searches, region-by-environment coverage summaries, accession-level exclusions, and justified update policies. Review should draw on relevant regional expertise, and authors should explain how local knowledge and contributions were incorporated

where applicable. The published record does not establish that editorial decisions or author affiliations caused these gaps, nor does this audit quantify their effects on taxonomic novelty or peptide discovery. These findings make transparent geographic validation an editorial responsibility. Global resources should demonstrate that eligible evidence from the Global South was sought and assessed with the same methodological care as evidence from the Global North. The present audit identifies concrete datasets requiring reconciliation; it does not estimate regional exclusion rates or the effects of their inclusion on the catalog's biological findings.

Data availability

Supplementary Table 1 lists sampling region, operational North/South grouping, environmental stresses, release dates, counts, provenance and eligibility qualifications. Supplementary Data 1 provides the machine-readable table, ENA snapshots, run-level selections, screening register and reproducible count and figure code. The targeted audit has no regional eligibility denominator.

Code availability

The supplied count scripts reproduce the quantitative audit from archived metadata and project classifications. Figure 1 maps reported ENA sample coordinates; coincident sampling locations may overlap. Background EEMC locations are shown only where repository coordinates are available.

Author disclosure

F.P.C. is a co-author of Marcoleta et al. (2022), reference 17. This study is retained in the screening register and is not given special weight or added to unresolved read counts.

References

1. Jiang, P. *et al.* The Extreme Environment Microbiome Catalog (EEMC): a global resource for microbial diversity and antimicrobial discovery. *Nature Communications* **17**, (2026).
2. Veloso, M. *et al.* Diversity, Taxonomic Novelty, and Encoded Functions of Salar de Ascotán Microbiota, as Revealed by Metagenome-Assembled Genomes. *Microorganisms* **11**, (2023).
3. Finstad, K. M. *et al.* Microbial community structure and the persistence of cyanobacterial populations in salt crusts of the hyperarid atacama desert from genome-resolved Metagenomics. *Front. Microbiol.* **8**, (2017).
4. Schulze-Makuch, D. *et al.* Transitory microbial habitat in the hyperarid Atacama Desert. *Proc. Natl. Acad. Sci. U. S. A.* **115**, (2018).
5. Castro-Severyn, J. *et al.* Living to the High Extreme: Unraveling the Composition, Structure, and Functional Insights of Bacterial Communities Thriving in the Arsenic-Rich Salar de Huasco Altiplanic Ecosystem. *Microbiol. Spectr.* **9**, (2021).
6. Guajardo-Leiva, S., Pedrós-Alió, C., Salgado, O., Pinto, F. & Díez, B. Active crossfire between cyanobacteria and cyanophages in phototrophic mat communities within hot springs. *Front. Microbiol.* **9**, (2018).
7. Murray, B., Dailey, M., Ertekin, E. & DiRuggiero, J. Draft Metagenomes of Endolithic Cyanobacteria and Cohabitants from Hyper-Arid Deserts. *Microbiol. Resour. Announc.* **10**, (2021).
8. Martínez-Alvarez, L. *et al.* With a pinch of salt: metagenomic insights into Namib Desert salt pan microbial mats and halites reveal functionally adapted and competitive communities. *Appl. Environ. Microbiol.* **89**, (2023).
9. Zablocki, O., van Zyl, L. J., Kirby, B. & Trindade, M. Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. *Viruses* **9**, (2017).
10. Johnson, S. S., Chevrette, M. G., Ehlmann, B. L. & Benison, K. C. Insights from the metagenome of an acid salt lake: The role of biology in an extreme depositional environment. *PLoS One* **10**, (2015).
11. Podell, S. *et al.* Assembly-Driven Community Genomics of a Hypersaline Microbial Ecosystem. *PLoS One* **8**, (2013).
12. Jordaan, K. *et al.* Hydrogen-Oxidizing Bacteria Are Abundant in Desert Soils and Strongly Stimulated by Hydration. *mSystems* **5**, (2020).
13. Salgado, O. *et al.* Global phylogenomic novelty of the CasI gene from hot spring microbial communities. *Front. Microbiol.* **13**, (2022).
14. Eissler, Y., Dorador, C., Kieft, B., Molina, V. & Hengst, M. Virus and potential host microbes from viral-enriched metagenomic characterization in the high-altitude wetland, Salar de Huasco, Chile. *Microorganisms* vol. 8
15. Gutiérrez-Preciado, A. *et al.* Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nat. Ecol. Evol.* **2**, (2018).
16. López-García, P. *et al.* Metagenome-derived virus-microbe ratios across ecosystems. *ISME Journal* **17**, (2023).
17. Marcoleta, A. E. *et al.* The highly diverse Antarctic Peninsula soil microbiota as a source of novel resistance genes. *Science of The Total Environment* **810**, 152003 (2022).
18. Centurion, V. B., Silva, J. B., Duarte, A. W. F., Rosa, L. H. & Oliveira, V. M. Comparing resistome profiles from anthropogenically impacted and non-impacted areas of two South Shetland Islands – Maritime Antarctica. *Environmental Pollution* **304**, (2022).
19. Zhang, T., Ji, Z., Li, J. & Yu, L. Metagenomic insights into the antibiotic resistome in freshwater and seawater from an Antarctic ice-free area. *Environmental Pollution* **309**, (2022).

20. Santos, A., Burgos, F., Martínez-Urtaza, J. & Barrientos, L. Metagenomic Characterization of Resistance Genes in Deception Island and Their Association with Mobile Genetic Elements. *Microorganisms* **10**, (2022).
21. Andreani-Gerard, C. M., Cambiazo, V. & González, M. Biosynthetic gene clusters from uncultivated soil bacteria of the Atacama Desert. *mSphere* **9**, (2024).
22. Plaza, C. *et al.* Soil resources and element stocks in drylands to face global issues. *Sci. Rep.* **8**, (2018).
23. Heinz, E., Holt, K. E., Meehan, C. J. & Sheppard, S. K. Addressing parachute research and removing barriers for lmic researchers in microbial genomics. *Microbial Genomics* vol. 7 Preprint at <https://doi.org/10.1099/MGEN.0.000722> (2021).
24. Arros, P. *et al.* Life on the edge: Microbial diversity, resistome, and virulome in soils from the union glacier cold desert. *Science of The Total Environment* **957**, 177594 (2024).
25. Lezcano, M. Á. *et al.* Hyperexpansion of genetic diversity and metabolic capacity of extremophilic bacteria and archaea in ancient Andean lake sediments. *Microbiome* **12**, (2024).
26. Medeiros, W. B. *et al.* Antarctic soil prokaryotic diversity: a dataset of 319 metagenome-assembled genomes from Deception and Livingston Islands. *Microbiol. Resour. Announc.* <https://doi.org/10.1128/mra.01346-25> (2026) doi:10.1128/mra.01346-25.
27. Gutiérrez-Preciado, A. *et al.* Extremely acidic proteomes and metabolic flexibility in bacteria and highly diversified archaea thriving in geothermal chaotropic brines. *Nat. Ecol. Evol.* **8**, (2024).
28. Miller, J., White, T. B. & Christie, A. P. Parachute conservation: Investigating trends in international research. *Conserv. Lett.* **16**, (2023).

Figures

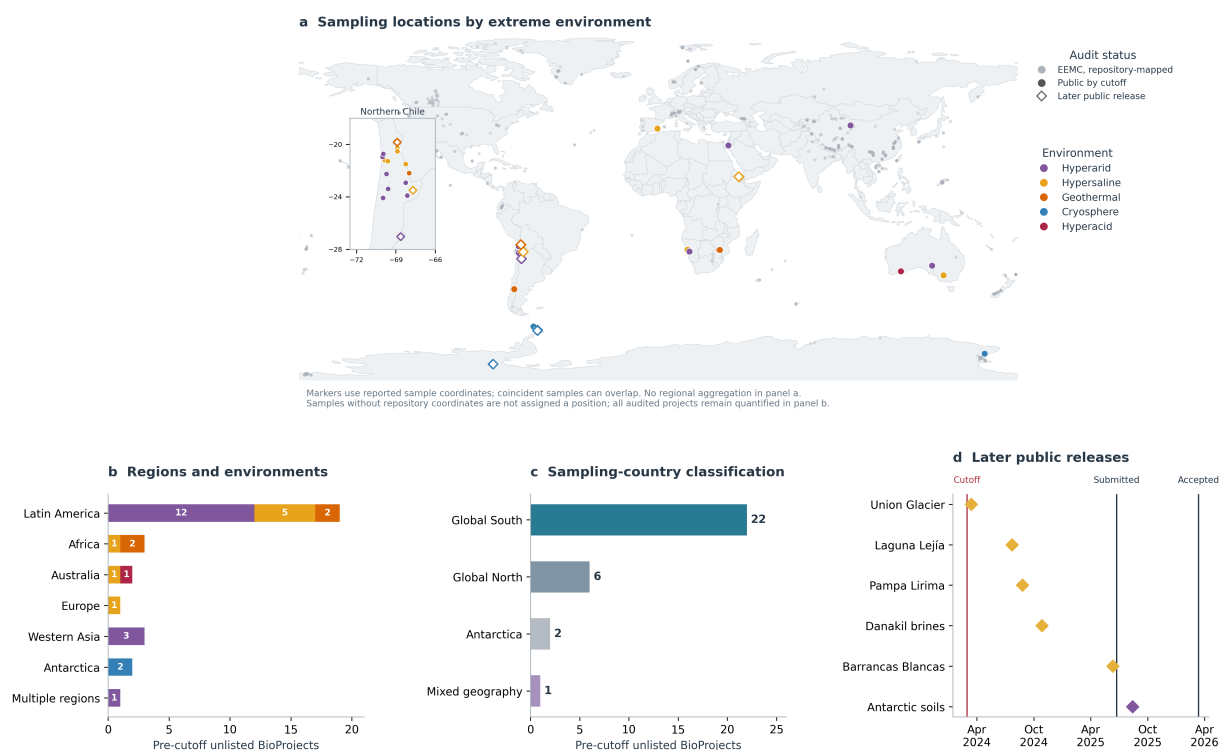


Figure 1. Expanded geographic and temporal audit of EEMC. a, reported sampling locations, colored by primary extreme environment. Filled circles indicate data published by the cutoff; open diamonds indicate later releases. Gray background markers show repository-mapped EEMC locations. Samples sharing coordinates may overlap; the inset enlarges northern Chile. b, pre-cutoff unlisted projects by sampling region and primary extreme-environment class; secondary stresses are retained in Supplementary Table 1. c, Sampling-country grouping based on UNCTAD developed/developing economy categories as an operational North/South convention, with Antarctica and mixed-country projects unassigned. d, Five releases after the cutoff and before submission, plus one during review. The 31-project pre-cutoff register includes Huasco viral-enriched metagenomes, and Llamara amplification-free libraries labeled WGA. This is an ascertainment audit, not an estimate of regional selection bias. BioProjects are not independent studies: fourteen culture-derived endolithic projects originate from a single study.[7] Their genomic eligibility is separately annotated.