Datathons: fostering equitability in data reuse in ecology

Consortium*

*A list of authors and their affiliations appears at the end of the paper

Corresponding author(s): *Stephanie Jurburg (*s.d.jurburg@gmail.com), *Maria J. Álvarez Blanco* (maria.alvbla@gmail.com)

Abstract

In the midst of a looming global biodiversity crisis, approaches to rapidly collect, curate, catalog, and integrate biodiversity data at global scales are more important than ever before[1]. Historically, data collection and reuse have been linked to local access to funding for scientific research and infrastructure, generating blind spots in the distribution of biodiversity data. At the same time, areas with limited access to research funding, where biodiversity data are generated at a slower pace can benefit from data reuse to reinterpret extant data within novel contexts or at different scales in order to further the local development of excellent research. This is especially true for sequence-based biodiversity research, which can be prohibitively expensive. Here, we describe the first Datathon, a three-day event held among microbial ecologists in Argentina and Uruguay to a) improve the openness of local data and develop a rich database of bacterial communities sampled in this region, b) ensure that data providers are credited for data reuse, and 3) encourage and facilitate the reuse of this resource by local researchers through training. The event resulted in the deposition of novel datasets to public databases, the assembly of the largest collection of soil microbiomes in Argentina and Uruguay to date, and the formation of a collaborative consortium that aims to reuse the data in the future. While the event was focused on microbial ecology, this model may serve to further develop equitable data archiving, collection and reuse practices in other areas of ecology.

Main text

Like other branches of life, the global microbiome is under threat [2], and documenting the world's microbial diversity is more urgent than ever before. A global coverage of biodiversity data is essential for developing in-depth ecological knowledge of microbial systems [1] and harnessing them as sources of biotechnological innovation [3]. As DNA sequencing technologies have greatly advanced, cataloging the world's microbiomes is now feasible. Over the past decade, sequencing-based assessments of bacterial diversity (i.e., metabarcoding or amplicon sequencing) have grown exponentially [4], but global blind spots in reusable microbiome data persist [5], often affecting the regions that are predicted to undergo the greatest rates of anthropogenic change, and therefore the greatest biodiversity loss [2].

Sequencing-based biodiversity assessments are necessary for most microbiomes, which cannot be characterized through conventional observation, but require substantial financial investment. Several studies have reported a disproportionately higher availability of microbiome data from wealthier countries [5,6] (Figure 1). For example, a systematic literature review of global soil biodiversity research (much of which relies on sequencing), found that only 8% of the studies surveyed originated from Latin America and Africa [7]. At the same time, as much as 50% of all the sequence data that has already been generated is not properly archived, and therefore is not fully reusable [8].

Improving data archiving practices is a cost-effective first step towards improving the coverage of compiled, global microbiome data, but requires explicit consideration of the associated costs and benefits, especially to the researchers producing the data. In ecology, synthesis research is disproportionately performed by researchers from high income countries (Figure 1), and while data collected from biodiversity blind spots are necessary for global syntheses, synthesis research is seldom performed by scientists from the poorly represented regions, who receive little direct benefit from making their data available.

While data citations allow data creators to receive credit for their work, they do not encourage equitable participation in data reanalyses. Closing the geographic gap between data producers can improve equitability in ecological research and has numerous benefits [9]. First, data reuse in microbiome research allows researchers to produce high-quality research regardless of their access to infrastructure or funding. Second, the prospect of reusing data may serve as an incentive to archive it publicly, increasing the amount and quality of available, reusable microbiome data in countries with limited research funding. Third, as

global participation in synthetic microbiome research increases, so should the diversity of perspectives in the field [10]. Finally, a greater global participation in data reuse may reduce language barriers in synthetic research, which are pervasive [11].

To improve equitability in microbiome synthesis science, it is essential to acknowledge available infrastructures and their limitations, provide educational support and training, build collaborative networks, and credit collaborators [1,10]. We organized a binational data collection and reuse event (*Datathon*) in Argentina and Uruguay, both of which are often poorly represented in global microbiome syntheses [5,6,12]. For example, the Earth Microbiome Project dataset [12], a collection of standardized bacterial metabarcoding data from thousands of samples, contains only nine microbiome samples from Argentina, and none from Uruguay.

The Datathon provided support to microbial ecologists in archiving and reusing metabarcoding sequence data and brought researchers together to create a common microbiome dataset for Argentina and Uruguay. By centralizing available raw data and including related bibliographic, technical, and experimental materials in a single online database, we aimed to improve the discoverability and reusability of microbial sequence data in the region, while giving data producers academic credit for their work and creating a valuable resource to foster research in a biodiversity blind spot. Crucially, we aimed to stimulate synthesis research by Datathon participants using the newly deposited sequences.

The Datathon was organized over the course of three days in October 2022 in a hybrid format, and each day focused on a different interconnected aim (Figure 2). On the first day (themed *Inspire*), we held a hybrid symposium focused on the history of synthetic research in ecology, the relevance of 'Open Science' in biodiversity research, and the potential and outcomes of recent global biodiversity data syntheses.

The second day **(**themed *Support*) focused on hands-on sequence data and metadata deposition, and was held remotely to allow all participants access to their work space. Custom, online step-by-step guides were developed to support sequence data preservation and publication process to NCBI's Sequence Read Archives in English and Spanish (freely available, Spanish https://github.com/MariaAlvBla/Dataton-2022/wiki and English versions https://github.com/MariaAlvBla/NCBI-Tutorial/wiki), in line with the FAIR Principles [13]. Guides included a custom metadata sheet, which included specific fields to allow for the rapid integration of all datasets and reanalysis following the Datathon, and to give greater visibility to the original publications of the data creators. In addition to standard NCBI

metadata fields, we included fields for technical information (e.g., DNA extraction method and sample amount), and the DOI for any scientific article accompanying the initial publication of the sequence data. In addition, given the broad range of environments sampled within the context of microbiome research, we developed a three-level ontology, where users could select their sample's realm (e.g., aquatic, mineral, host-associated), broad-scale environment (e.g., soil, freshwater), and complete one description per sample (e.g. "agricultural soil from soy farm"). If data had been made publicly available prior to the Datathon, participants could complete the metadata sheet with the original accession numbers to expedite data integration. All members were invited to join an online, dedicated *Slack* group, which included a helpdesk channel and a networking channel that allowed participants to obtained more personalized help when needed, improving the quality of the deposited datasets.

The third day (themed *Collaborate*) focused on harnessing the deposited data for reuse by participants and on developing collaborative networks. Following a general summary of the collected data, participants created *Slack* channels to brainstorm and develop projects to reuse the data, and secure funding to pursue these ideas. Then, collaborative synthesis projects were voted on, and leaders were collectively selected for each. After the meeting, all participants received a detailed summary of the data resource and synthesis projects, and could opt-in to each. In the 6 months following the Datathon, this collaborative network has already served to exchange knowledge and coordinate future research, procure new funding, and organize upcoming Datathon events for scientists across both Latin America and Sub-Saharan Africa.

In total, 30 scientists participated in the Datathon as data providers, collectively archiving and consolidating 913 samples from 22 projects in NCBI's Sequence Read Archives. Deposition to NCBI ensures that the data remains publicly accessible in the long-term, and facilitates integration with other publicly available datasets (for NCBI accession numbers, see Supplementary 1).Of the contributed projects, 55% (33% of samples) were previously unarchived. Furthermore, the custom data deposition guides remain publicly available as living educational documents for users aiming to deposit microbiome sequence data in the future, and serve as a model for translation into other languages. Notably, the compiled dataset is dominated by soil microbiome samples, likely because the initiative began outreach through the Soil BON [14] network of researchers, illustrating the influence of the networking approach.

Equitable participation from researchers globally through synthesis work can reduce disparities arising from differential access to funding, in turn reducing existing biases in research [6], increasing the scope/breadth of research [15] and bolstering transparency in the receipt of academic credit [16,17], and excellence in science. The rapid, coordinated public archiving of microbiome sequence data from 913 samples within three days demonstrates the tremendous potential of equitable data consolidation approaches to shed light on biodiversity blind spots, in both microbiome research and other areas of ecology. The future reuse of these data by researchers from the region that produced the data will likely advance the collective scientific knowledge of the microbiomes of South America, how they are affected by local anthropogenic change, and how local policies may mitigate microbial diversity loss.

Bibliography

1. Nuñez, M. A., Chiuffo, M. C., Pauchard, A. & Zenni, R. D. Making ecology really global. *Trends Ecol. Evol.* **36**, 766–769 (2021).
2. Averill, C. *et al.* Defending Earth's terrestrial microbiome. *Nat. Microbiol.* (2022)
3. Vuong, P., Chong, S. & Kaur, P. The little things that matter: how bioprospecting microbial biodiversity can build towards the realization of United Nations Sustainable Development Goals. *npj Biodivers.* **1**, 4 (2022).
4. Ahmed, S. A. J. A. *et al.* Large scale text mining for deriving useful insights: A case study focused on microbiome. *Front. Physiol.* **13**, 933069 (2022).
5. Guerra, C. A. *et al.* Blind spots in global soil biodiversity and ecosystem function research. *Nat. Commun.* **11**, 3870 (2020).
6. Abdill, R. J., Adamowicz, E. M. & Blekhman, R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* **20**, e3001536 (2022).
7. El Mujtar, V., Muñoz, N., Prack Mc Cormick, B., Pulleman, M. & Tittonell, P. Role and management of soil biodiversity for food security and nutrition; where do we stand? *Glob. Food Sec.* **20**, 132–144 (2019).
8. Jurburg, S. D., Konzack, M., Eisenhauer, N. & Heintz-Buschart, A. The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Commun. Biol.* **3**, 474 (2020).

9. Aubin, I. *et al.* Managing data locally to answer questions globally: The role of collaborative science in ecology. *J. Veg. Sci.* **31**, 509–517 (2020).

10. Oduaran, O. H. & Bhatt, A. S. Equitable partnerships and the path to inclusive, innovative and impactful human microbiome research. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 683–684 (2022).

11. Konno, K. *et al.* Ignoring non-English-language studies may bias ecological meta-analyses. *Ecol. Evol.* **10**, 6373–6384 (2020).

12. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

13. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

14. Guerra, C. A. *et al.* Tracking, targeting, and conserving soil biodiversity. *Science* **371**, 239–241 (2021).

15. Melles, S. J. *et al.* Diversity of practitioners publishing in five leading international journals of applied ecology and conservation biology, 1987–2015 relative to global biodiversity hotspots. *Écoscience* **26**, 323–340 (2019).

16. Eichhorn, M. P., Baker, K. & Griffiths, M. Steps towards decolonising biogeography. *Front. Biogeogr.* **12**, (2020).

17. Hazlett, M. A., Henderson, K. M., Zeitzer, I. F. & Drew, J. A. The geography of publishing in the Anthropocene. *Conservation Science and Practice* **2**, (2020).

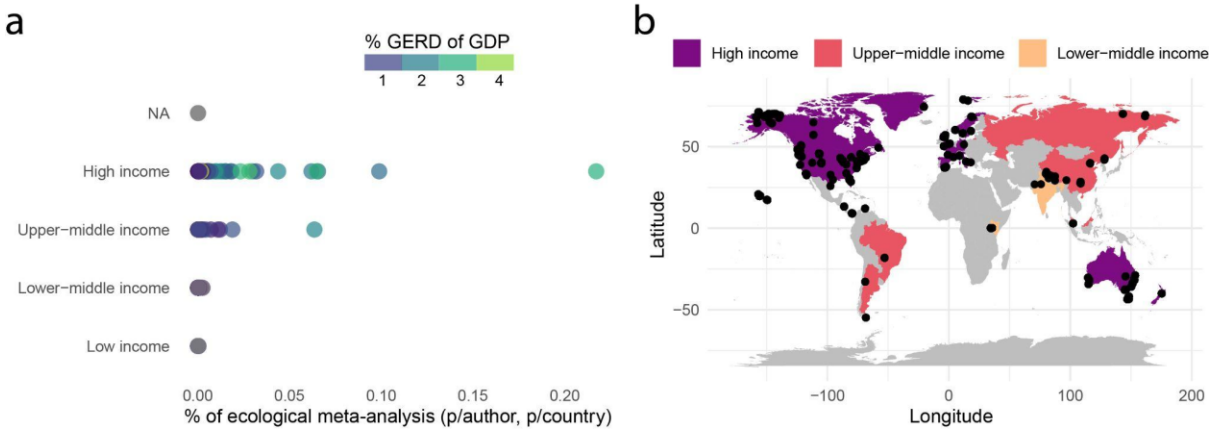

Figure 1. Data reuse and archiving in ecology are disproportionately performed by researchers from wealthy countries. a. Most published ecological syntheses are authored by researchers residing in high income countries. We systematically searched ecology journals for articles which mentioned meta analyses using Web of Science (*Ecology* category, and "meta-analy*" OR "metaanaly*" OR "meta analy*" in all fields), and identified 2446 articles. The affiliations of each author in each manuscript were counted as an authorship in that country, and countries were classified according to the 2021 World Bank income groups. Countries are coloured by Gross domestic expenditure on R&D (GERD) as a percentage of GDP, as reported by the World Bank (2021). b. Most data from the Earth Microbiome Project was sampled in high income countries.

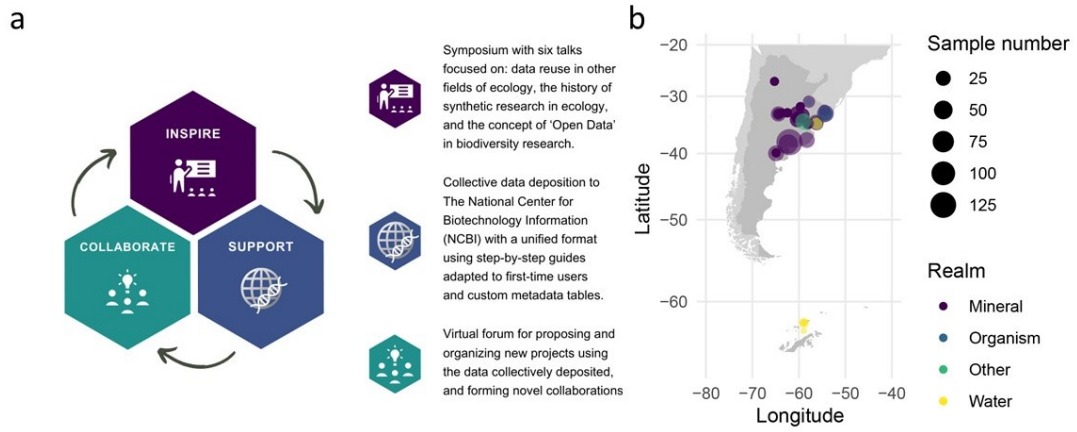

Figure 2. Equitable data archiving and reuse practices must consider the needs of participants. a. In designing the Datathon, we considered the need for background information on the potential of data synthesis, providing support during data deposition, and the establishment of collaborative networks, to be of primary importance in furthering microbiome synthesis research. b. The project resulted in the deposition and consolidation of 913 samples, and a much improved coverage of the Argentina-Uruguay region.

***Consortium members**

Stephanie Jurburg [1]*

María J. Álvarez Blanco [1, 2]*

Antonis Chatzinotas [1, 3, 4]

Anahita Kazem [2, 5]

Birgitta König-Ries [2, 5]

Doreen Babin [6]

Kornelia Smalla [6]

Victoria Cerecetto [6, 7]

Gabriela Fernandez-Gnecco [6, 8]

Fernanda Covacevich [8, 9]

Emilce Viruel [10]

Yesica Bernaschina [11]

Carolina Leoni [7, 11]

Silvia Garaycochea [7, 12]

Jose. A Terra [13]

Pablo Fresia Coronel [14]

Eva Lucía Margarita Figuerola [15, 16]

Luis Gabriel Wall [15, 17]

Julieta Mariana Covelli [17]

Ana Carolina Agnello [18]

Esteban Emanuel Nieto [18]

Sabrina Festa [18]

Lina Edith Dominici [19]

Marco Allegrini [20]

María Celina Zabaloy [20]

Marianela Estefanía Morales [20, 21]

Leonardo Erijman [23, 23]

Anahi Coniglio [24]

Fabricio Dario Cassán [24]

Sofia Nievas [24]

Diego M. Roldán [25, 26]

Rodolfo Menes [26, 27]

Patricia Vaz Jauri [28, 29]

Carla Silva Marrero [28]

Adriana Montañez Massa [28]

María Adelina Morel Revetria [28, 30]

Ana Fernández-Scavino [31]

Luciana Pereira Mora [31]

Soledad Martínez [32]

Juan Pablo Frene [33, 34]

Affiliations of consortium members

1. Department of Environmental Microbiology, Helmholtz Centre for Environmental Research (UFZ), 04318 Leipzig, Germany.

2. Data and Code Unit (iBID), German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Saxony, Germany.

3. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany.

4. Institute of Biology, Leipzig University, 04103 Leipzig, Germany.

5. Department of Mathematics and Computer Science, Friedrich Schiller University Jena, 07743 Jena, Thüringen, Germany.

6. Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Epidemiology and Pathogen Diagnostics, 38104 Braunschweig, Germany.

7. Instituto Nacional de Investigación Agropecuaria (INIA), Área de Recursos Naturales, Producción y Ambiente, Estación Experimental INIA Las Brujas, Ruta 48 km 10, Canelones, Uruguay.

8. Instituto de Investigaciones en Biodiversidad y Biotecnología-Consejo Nacional de Investigaciones Científicas y Técnicas (INBIOTEC-CONICET), Mar del Plata, Buenos Aires, Argentina.

9. Instituto Nacional de Tecnología Agropecuaria, Estación Experimental Agropecuaria Balcarce (INTA, EEA Balcarce), Balcarce, Buenos Aires, Argentina.

10. Instituto de Investigación Animal del Chaco Semiárido (IIACS), Centro de Investigaciones Agropecuarias (CIAP), Instituto Nacional de Tecnología Agropecuaria (INTA), Tucumán, Argentina.

11. Instituto Nacional de Investigación Agropecuaria (INIA), Sistema Vegetal Intensivo, Estación Experimental INIA Las Brujas, Ruta 48 km 10, Canelones, Uruguay.

12. Instituto Nacional de Investigación Agropecuaria (INIA), Área Mejoramiento Genético y Biotecnología Vegetal, Estación Experimental INIA Las Brujas, Ruta 48 km 10, Canelones, Uruguay.

13. Instituto Nacional de Investigación Agropecuaria (INIA), Sistema Arroz-Ganadería, Estación Experimental INIA Treinta y Tres, Ruta 8 km 282, Treinta y Tres, Uruguay.

14. Unidad Mixta Pasteur + INIA (UMPI), Institut Pasteur de Montevideo, Montevideo, Uruguay.

15. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina.

16. Instituto de Biociencias, Biotecnología y Biología Traslacional, Departamento de Fisiología y Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina.

17. Laboratorio de Bioquímica y Biología de Suelos, Centro de Bioquímica y Microbiología de Suelos, Universidad Nacional de Quilmes (UNQ), Bernal, Buenos Aires, Argentina.

18. Centro de Investigación y Desarrollo en Fermentaciones Industriales (CINDEFI, CONICET-UNLP), La Plata, Argentina.

19. Centro de Investigación y Desarrollo en Tecnología de Pinturas y Recubrimientos (CIDEPINT, CICPBA-CONICET-UNLP), La Plata, Argentina.

20. Centro de Recursos Naturales Renovables de la Zona Semiárida (CERZOS, CONICET-UNS), Bahía Blanca, Buenos Aires, Argentina.

21. Departamento de Agronomía, Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina.

22. Instituto de Investigaciones en Ingeniería Genética y Biología Molecular "Dr Héctor N Torres" (INGEBI, CONICET-UBA), Buenos Aires, Argentina.

23. Departamento de Fisiología, Biología Molecular y Celular "Dr Héctor Maldonado", Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA).

24. Laboratorio de Fisiología Vegetal y de la Interacción Planta Microorganismo (LFVIPM), Instituto de Investigaciones Agrobiotecnológicas (INIAB-CONICET), Facultad de Ciencias Exactas Físico-Químicas y Naturales, Universidad Nacional de Río Cuarto (UNRC), Río Cuarto, Córdoba, Argentina.

25. Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), Ministerio de Educación y Cultura, Montevideo, Montevideo, Uruguay.

26. Laboratorio de Ecología Microbiana Medioambiental, Microbiología, Facultad de Química y Unidad Asociada del Instituto de Química Biológica, Facultad de Ciencias, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

27. Laboratorio de Microbiología, Unidad Asociada del Instituto de Química Biológica, Facultad de Ciencias, Universidad de la República (UdelaR), Montevideo, Montevideo, Uruguay.

28. Laboratorio de Microbiología de Suelos, Instituto de Ecología y Ciencias Ambientales, Facultad de Ciencias, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

29. Laboratorio de Interacción Planta-Microorganismo, Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), Montevideo, Montevideo, Uruguay.

30. Laboratorio de Microbiología Molecular, Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), Montevideo, Montevideo, Uruguay.

31. Área Microbiología, Departamento de Biociencias, Facultad de Química, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

32. Laboratorio de Biotecnología, Departamento de Biociencias, Unidad de Análisis de Agua, Facultad de Química, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

33. Future Food Beacon of Excellence, University of Nottingham, LE12 5RD, United Kingdom.

34. School of Biosciences, University of Nottingham, LE12 5RD, United Kingdom.