GigaScience, 2025, **000**, 1–0
DOI: 10.1093/gigascience/giaf102
Data Note

The Earth Hologenome Initiative: Data Release 1

Nanna Gaun[1], Carlotta Pietroni[1], Garazi Martin-Bideguren[1], Jonas Lauritsen[1], Ostaizka Aizpurua[1], Joana M. Fernandes[2], Eduardo Ferreira[2], Fabien Aubret[3], Tom Sarraude[3], Constant Perry[3], Lucas Wauters[4], Claudia Romeo[1,5], Martina Spada[4], Claudia Tranquillo[4], Alex O. Sutton[6], Michael Griesser[7,8,9,10], Miyako H. Warrington[10,11], Guillem Pérez i de Lanuza[12], Javier Abalos[12,13], Prem Aguilar[14], Ferran de la Cruz[14], Javier Juste[15,16], Pedro Alonso-Alonso[17], Jim Groombridge[18], Rebecca Louch[18], Kevin Ruhomaun[19], Sion Henshaw[20], Carlos Cabido[21], Ion Garin Barrio[21], Emina Šunje[22], Peter Hosner[23,24,25], Ivan Prates[13], Geoffrey M. While[26], Roberto García-Roa[13], Tobias Uller[13], Nathalie Feiner[13,27], Elisa Bonaccorso[28], Pernille Klein-Ipsen[29], Rosalina Rotovnik[29], Antton Alberdi[1,*], and Raphael Eisenhofer[1]

[1]Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark
[2]CESAM & Department of Biology, University of Aveiro, Aveiro, Portugal
[3]Station d'Ecologie Théorique et Expérimentale, CNRS
[4]Università degli Studi dell'Insubria, Varese, Italy
[5]Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Brescia, Italy
[6]School of Environmental and Natural Sciences, Bangor University
[7]Department of Biology, University of Konstanz, Konstanz, Germany
[8]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany
[9]Department of Collective Behaviour, Max Planck Institute of Animal Behaviour, Konstanz, Germany
[10]Luondu Boreal Research Station, Arvidsjaur, Sweden
[11]School of Biological and Medical Sciences, Oxford Brookes University, Headington, OX3 0BP, UK
[12]Ethology Lab, Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, Valencia, Spain
[13]Department of Biology, Lund University, Lund, Sweden
[14]Research Centre in Biodiversity and Genetic Resources, InBIO, CIBIO, Universidade do Porto, Porto, Portugal
[15]Estación Biológica de Doñana (CSIC), Sevilla, Spain
[16]Epidemiology and Public Health, CIBERESP, Madrid, Spain
[17]Department of Animal Ecology and Tropical Biology. University of Würzburg, Würzburg, Germany
[18]Durrell Institute of Conservation and Ecology, School of Natural Sciences, University of Kent, Kent, UK
[19]National Parks and Conservation Service, Ministry of Agro-Industry and Food Security, Government of Mauritius
[20]Mauritian Wildlife Foundation, Vacoas, Mauritius
[21]Aranzadi Science Foundation, Donostia-San Sebastián
[22]University of Sarajevo, Sarajevo, Serbia
[23]Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[24]Center for Global Mountain Biodiversity, University of Copenhagen,Copenhagen, Denmark
[25]Center for Macroecology, Evolution, and Climate, University of Copenhagen, Copenhagen, Denmark
[26]School of Natural Sciences, University of Tasmania, Australia
[27]Max Planck Institute for Evolutionary Biology, Plön, Germany
[28]Instituto Biósfera, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito, Quito, Ecuador
[29]Parasitology and Pathobiology, Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark
*Correspondence address. Antton Alberdi, E-mail: antton.alberdi@sund.ku.dk

Abstract

Background: The Earth Hologenome Initiative (EHI) is a global endeavor dedicated to revisit fundamental ecological and evolutionary questions from the systemic host–microbiota perspective, through the standardized generation and analysis of joint animal genomic and associated microbial metagenomic data.

Results: The first data release of the EHI contains 968 shotgun DNA sequencing read files containing 5.2 TB of raw genomic and metagenomic data derived from 21 vertebrate species sampled across 12 countries, as well as 17,666 metagenome-assembled genomes reconstructed from these data.

Conclusions: The dataset can be used to address fundamental questions about host–microbiota interactions and will be available to the research community under the EHI data usage conditions.

Keywords:

Background

The Earth Hologenome Initiative (EHI) [1] stands as a global scientific undertaking dedicated to revisit fundamental ecological and evolutionary questions from the systemic host–microbiota perspective [2, 3]. This goal is pursued through hologenomics, namely, the joint generation and analysis of host genomic and associated microbial metagenomic data [4]. The EHI unfolds through a 2-level approach with the participation of worldwide researchers

Received: May 27, 2025. **Revised:** August 2, 2025. **Accepted:** August 13, 2025

representing >80 countries. At the initial level, the small- to medium-scale projects are executed, aiming to address taxon- or environment-specific scientific inquiries. While the sampling designs of each project are tailored to particular scientific questions, all projects follow standardized sample collection, metadata acquisition, and data generation procedures [5]. The second level leverages the inherent comparability of previously generated data to explore broad ecological and evolutionary questions requiring extensive taxonomic and geographical representation and larger amounts of data.

The EHI methodologies fully rely on DNA shotgun sequencing, enabling genome-wide analyses of animal hosts [6] and genome-resolved metagenomic analysis of their associated microbial communities [7]. Due to the primary interest in intestinal microbial communities, both data types are primarily sourced from fecal samples, which serve both as a proxy for lower intestinal microbial communities [8, 9] and as a useful data source for population genomic analyses [10]. Alternative sample types, such as blood and tissue samples, are also used when the amount of host DNA in feces is insufficient for host genome analyses. Occasionally, other sample types, such as skin or oral swabs, are also collected in the context of specific projects. Samples are usually obtained from live animals captured in the wild to ensure the collection of unaltered specimens along with relevant metadata about the host. The animals are released immediately after sampling.

This EHI data release includes raw DNA sequencing read files and metagenome-assembled genomes derived from these data [11]. All sequencing data are associated with a rich set of standardized metadata encompassing host phenotype, fieldwork, and laboratory information, which are required for the interpretation of the results.

Data Description

Context

This first EHI data release contains raw sequencing data derived from 21 vertebrate species (Table 1). A total number of 902 samples were collected from animals across 317 sampling events that took place in 12 countries between January 2021 and December 2023 (Fig. 1). The sampling locations spanned 20 biomes, with most samples derived from temperate woodlands, followed by tropical forests, temperate shrublands, lakes or ponds, and polar tundra. All sampled specimens except the Greenland sled dogs (*Canis lupus familiaris*) were wild animals.

Six different types of samples were processed: anal/cloacal swabs ($n = 22$), colon contents ($n = 26$), feces ($n = 891$), oral swabs ($n = 13$), skin swabs ($n = 6$), and skin tissue samples ($n = 5$). For a comparison of the quality of data generated from fecal and anal/cloacal swabs, see Pietroni et al. [5]. From these samples, 963 libraries were sequenced to yield 5,198 GB of data, with an average of 5.39 ± 3.84 GB per sample, representing 33% of the total data generated within the EHI until March 2025. The released data include 6.88% ± 7.14% of low-quality DNA, 16.57% ± 27.52% of DNA mapped to host genomes, and 76.54% ± 28.74% of metagenomic DNA.

The current data release also includes 17,666 metagenome-assembled genomes (MAGs) derived from the binning of individual metagenomic assemblies conducted on the released sequencing data (Fig. 2). These MAGs derive from 15 different vertebrate species (Fig. 3) and have an average completeness value of 83.5% ± 15.3% and contamination value of 1.84% ± 2.07%. The catalog spans 33 phyla, with Bacillota A (7,660 MAGs) and Bacteroidota (5,466 MAGs) encompassing 73.9% of the reconstructed genomes. A total of 15,539 MAGs displayed an average nucleotide identity (ANI) below 95% with respect to any genome available at the R214 GTDB database [12], indicating an average novel species discovery rate of 87.9% [13]. All amphibian and reptile species displayed novel species discovery rates above 90%, with a maximum rate of 97.5%, as observed in the common wall lizard *Podarcis muralis* (Table 1).

Methods

Data were generated using the standardized field, laboratory, and bioinformatic procedures implemented in the EHI, which are explained below.

Sample collection

Sample collection was conducted by the field scientists included in the author list, as specified in the Author Contributions section. Every field researcher received identical sampling guidelines and a standardized EHI sampling kit equipped with barcoded sample collection tubes containing 1 mL DNA/RNA Shield buffer (Zymo Research). In accordance with the manufacturer's guidelines, a 1:10 sample-to-buffer ratio was employed, equating in the case of feces to approximately 100 mg of material. Adhering to EHI sample collection guidelines, samples were systematically accompanied by standardized metadata as outlined by Leonard et al. [1]. Most individual animals contributed at least 2 samples: fecal samples or anal/cloacal swabs were collected to generate gut microbial metagenomic data, while blood or tissue samples were collected to generate host genomic data when the host DNA in feces was insufficient for genome analysis. The samples were frozen within 2 weeks from collection, and details regarding sample preservation procedures prior to freezing were documented in the EHI database.

Laboratory processing

Laboratory sample processing was conducted at the Globe Institute's (University of Copenhagen) molecular laboratory in Copenhagen, Denmark, following the established EHI laboratory protocols [14]. In summary, samples underwent bead-beating before DNA isolation employing silica magnetic beads (G-Biosciences) with solid-phase reversible immobilization. The concentration of DNA extracts was quantified through a Qubit 3 Fluorometer (Thermo Fisher Scientific) using dsDNA HS (High Sensitivity) Assay Kits. Subsequently, DNA was fragmented into approximately 450-bp-long fragments using a Covaris LE220 platform (Covaris). Library preparation followed the ligation-based BEST protocol [15], utilizing a standard input of 200 ng DNA in 24 μL or the closest amount feasible based on the sample DNA concentration. We used 1.5 μL of 20 μM adapters for a 50- to 200-ng DNA input, 1.5 μL of 10 μM for 10–50 ng, 1.5 μL of 5 μM for <10 ng, and 1.5 μL of 2 μM for samples below the quantification range. Libraries underwent quantitative PCR screening to determine the optimal number of library indexing PCR cycles [16], followed by PCR amplification using unique dual index primers with an adjusted number of cycles. The resulting libraries underwent automated capillary electrophoresis using Fragment Analyzer (Agilent) for assessment of fragment-length distribution, adapter dimers, and adapter-to-library molar ratios. Finally, samples were pooled into 21 sequencing batches, and sequencing was performed across multiple lanes of NovaSeq6000 (RRID:SCR_016387) and NovaSeq X (RRID:SCR_024569) platforms (Illumina), generating an average

Table 1: Summary statistics of the animal species represented in the first EHI data release. Detailed metadata tables are available as part of the supporting files.

Species	Taxonomy	Sampling events	Individuals	Samples	Data (GB)	Genomes	Percentage new
Calotriton asper	Urodela, Amphibia	5	31	37	230.4	745	95.0
Canis lupus familiaris	Carnivora, Mammalia	14	58	58	333.7	1,252	39.3
Chalcides striatus	Squamata, Reptilia	2	2	2	39.5	0	–
Geospizopsis unicolor	Passeriformes, Aves	1	2	2	18.3	0	–
Lepus europaeus	Lagomorpha, Mammalia	15	25	50	252.6	711	85.4
Lissotriton helveticus	Urodela, Amphibia	16	88	97	444.7	1,590	95.9
Natrix astreptophora	Squamata, Reptilia	2	2	2	32.8	0	–
Perisoreus infaustus	Passeriformes, Aves	2	2	2	32.5	0	–
Plecotus auritus	Chiroptera, Mammalia	1	2	2	42.1	0	–
Podarcis filfolensis	Squamata, Reptilia	9	43	43	174.7	693	91.9
Podarcis gaigeae	Squamata, Reptilia	17	61	61	303.5	1,280	97.3
Podarcis liolepis	Squamata, Reptilia	2	13	13	67.0	232	92.2
Podarcis milensis	Squamata, Reptilia	8	26	26	149.7	590	96.6
Podarcis muralis	Squamata, Reptilia	35	154	165	998.5	2,670	97.5
Podarcis pityusensis	Squamata, Reptilia	12	43	43	220.8	1,046	93.1
Psittacula echo	Psittaciformes, Aves	49	48	50	591.2	123	53.6
Salamandra atra	Urodela, Amphibia	1	2	2	23.8	0	–
Sciurus carolinensis	Rodentia, Mammalia	47	65	120	533.1	1,686	95.8
Sciurus vulgaris	Rodentia, Mammalia	76	74	123	660.5	1,033	72.3
Trichosurus vulpecula	Diprotodontia, Mammalia	2	2	2	20.9	61	88.5
Zonotrichia capensis	Passeriformes, Aves	1	2	2	28.0	0	–

of 5 GB (approximately 16.6 million reads) of 150-bp paired-end sequencing data per sample.

Bioinformatics

The raw sequencing data underwent processing through the automated EHI bioinformatic pipeline [17], which is briefly explained below. The raw, intermediate, and final data were archived in the Electronic Research Data Archive at the University of Copenhagen [18]. Meanwhile, sample locations and pertinent metadata were stored in the EHI Database, built upon the Airtable software. Computation tasks were executed on the local cluster of the Globe Institute (Mjolnir), using custom bioinformatic pipelines based on Snakemake (RRID:SCR_003475) [19] and executed through slurm [20].

In the preprocessing step, fastp [21] was employed for quality filtering, followed by alignment against the reference host genome using Bowtie2 (RRID:SCR_016368) [22]. Mapped reads were retained for genomic analyses, while unmapped reads were isolated using SAMTOOLS (RRID:SCR_002105) [23] for subsequent metagenomic analyses. The unmapped fraction underwent complexity analysis using Nonpareil 3 [24] and microbial fraction estimation using SingleM [25, 26]. Subsequently, metagenomic assemblies were conducted for each individual sample using MEGAHIT v1.2.9 [27], followed by binning using CONCOCT [28], MaxBin2 [29], and MetaBAT2 [30]. Assembly statistics were generated using QUAST v5.2.0 [31]. The bins were subsequently refined using MetaWRAP's refinement module [32] with CheckM [33]. Taxonomic annotation utilized GTDB-tk v2.3.0 [12] against the R214 GTDB database [34], and the phylogenetic tree of MAGs was constructed by pruning the reference genomes using drop.tip function of the ape R package [35].

Data archiving

Raw sequencing data (FASTQ format) were archived at the European Nucleotide Archive (ENA), while draft bacterial genomes (FASTA format) were compiled in a tarball file and archived in Zenodo. We also offer users the option to obtain download links to specific MAGs directly from the EHI database [36]. Metadata specific to this data release, as well as the code used for visualization and summary statistics, are stored in GitHub, with a release frozen in Zenodo. Relevant URLs, DOIs, and accession numbers are mentioned in the Data Availability section.

Data validation and quality control

We implemented numerous measures in the field, laboratory, and bioinformatic procedures to ensure that the generated data were representative of the collected biological samples and comparable across samples obtained by different field researchers across the world [37], as detailed below.

Field quality control

The quality control measures implemented in the field included the usage of standardized sampling kits and guidelines to ensure all samples were collected following identical procedures. All field researchers were informed about the sensitivity of shotgun sequencing procedures regarding environmental contamination and cross-contamination, thus requiring them to employ clean items for storing and manipulating the animals and the samples, using protective synthetic gloves and continuously sterilizing tools. Samples were frozen at or below −18°C, ideally within a day and at maximum within the first 2 weeks after sample collection.

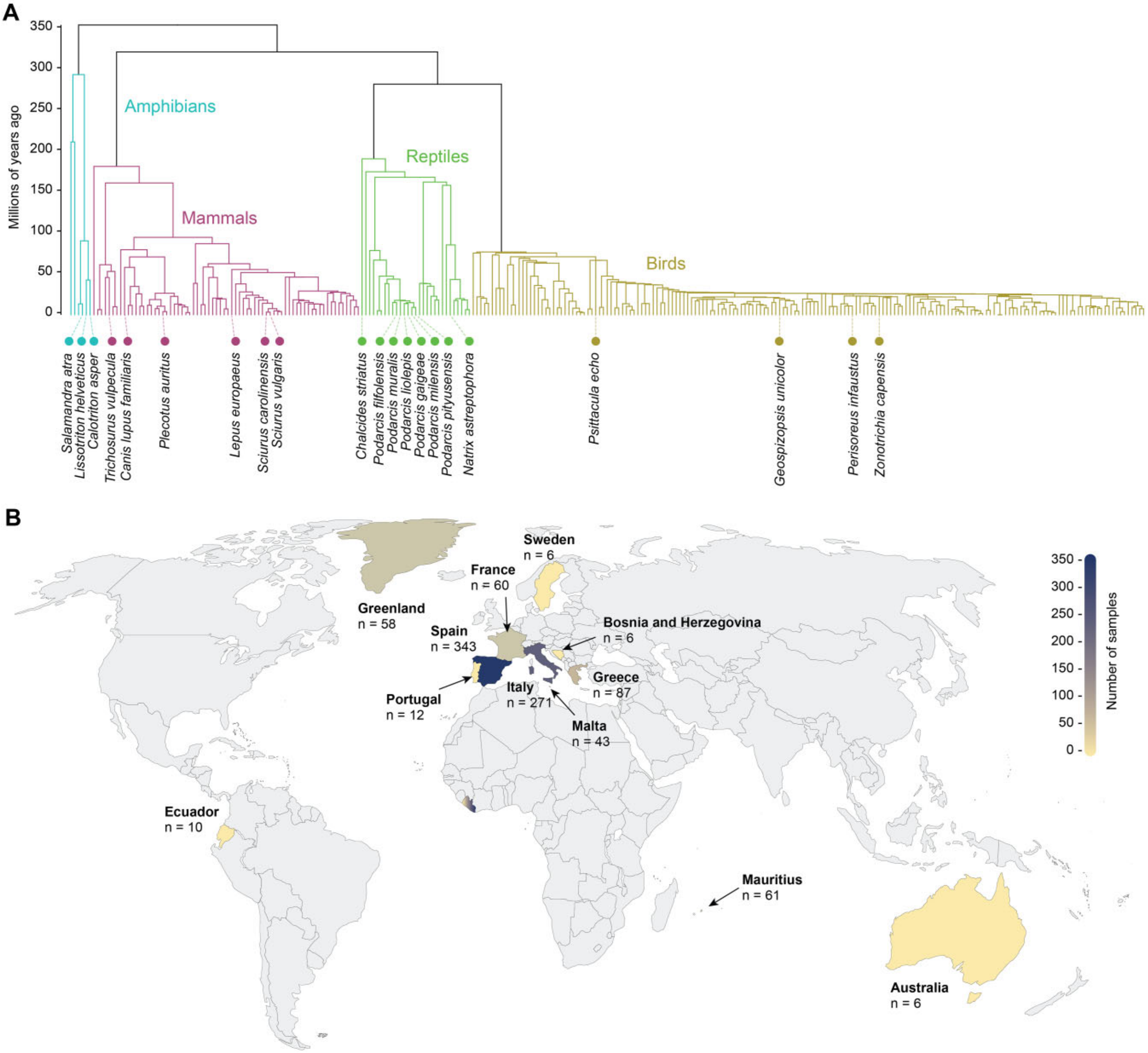

Figure 1: Phylogenetic placement and geographic origin of the samples. (A) Phylogenetic tree of all vertebrate species represented in the EHI collection in 2025 Q1, with the phylogenetic position of the species included in this data release highlighted. (B) World map indicating the number of samples sourced from each of the represented countries.

Time until freezing was recorded as one of the technical metadata variables.

Laboratory quality control

All sampling tubes were prelabeled with identical human-readable (5-digit code with 3 letters and 2 numbers; e.g., ABC99) and machine-readable (QR code) barcodes. Upon arrival at the Globe Institute, samples and metadata sheets were cross-checked and inconsistencies addressed before indexing the samples in the EHI database. This manual quality control also included logging deviations from standard procedures (e.g., overstuffing tubes with sample material) and technical issues such as leaking of sample tubes, which resulted in the disposal of unsuitable samples. All DNA extraction batches included blanks to monitor contamination and were organized according to expected DNA yield to minimize cross-contamination. Due to the variability of sample sources and types, concentrations of all DNA extracts were measured using a Qubit 3 Fluorometer, both to adjust the volumes for library preparation and to account for DNA template amount in statistical analyses. Sequencing adapter molarities were adjusted to the amount of input DNA to minimize the formation of adapter dimers and other artifacts, and all libraries were screened through quantitative PCR (Mx3005p; Agilent) to assess library preparation success and tailor the number of required indexing PCR cycles to each library. All indexed libraries were analyzed through capillary electrophoresis for high-quality measurement of library molarities, to ensure the required amount of sequencing data was generated.

Bioinformatic quality control

We employed multiple criteria to assess the quality and representativeness of the generated data. Following standard quality filtering, we removed reads with average phred-scores below q30 (1 sequencing error expected every 1,000 bases) and trimmed reads

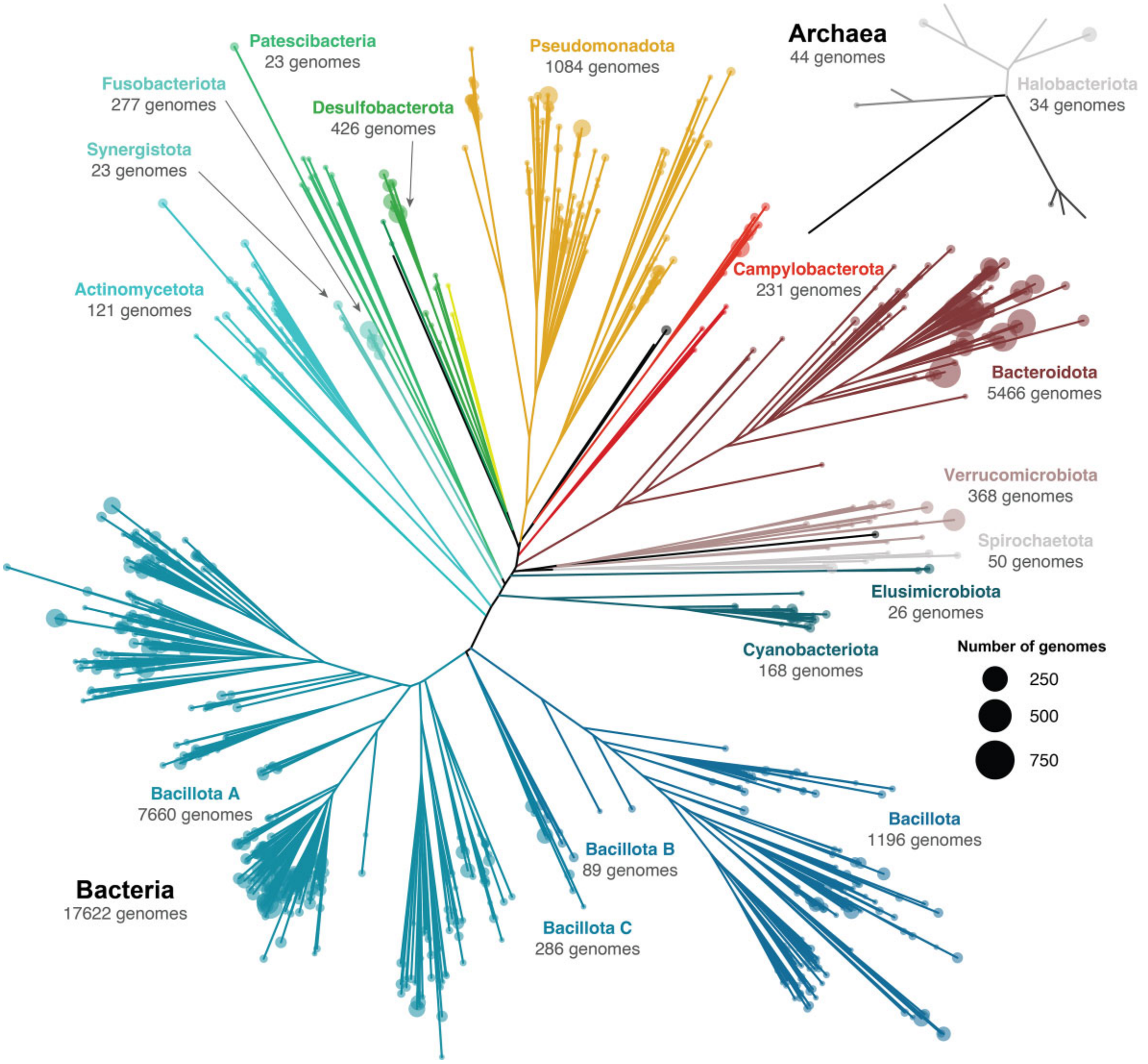

Figure 2: Phylogenetic trees of the EHI-reconstructed bacterial and archaeal genomes. Each tip represents a genus, and the tip size indicates the number of released genomes within the genus.

with low-quality endings and adapter remnants. To further assess library preparation success, we estimated duplication rates using the reads mapped to the host reference genome. Unmapped reads were further screened for complexity using Nonpareil 3, and the microbial read fraction was estimated using SingleM. Through all these measurements, we estimated expected levels of diversity and complexity, which we then used to assess the representativeness of the generated MAGs. Following field standards [38], only bins exceeding 50% completeness and maintaining contamination levels below 10% were considered MAGs to be included in downstream analyses.

Ethics

The EHI is governed by open science principles, adhering to CARE and FAIR data governance frameworks [12, 13], as well as complying with all international, national, and regional regulations stemming from the United Nations' Convention on Biological Diversity [39]. In line with these commitments, the rights and interests of Indigenous peoples are fully considered by actively involving local scientists in research projects. These scientists co-own the samples collected within the EHI framework, as well as the data derived from them. All sample collection, exportation, and data generation strictly adhere to local and international legislation on access and benefit-sharing (ABS) of genetic resources, as outlined in the Nagoya Protocol and implemented through national ABS laws. Accordingly, all sampling, material transfer, and ABS permits are filed in the EHI database. Finally, this data release serves as a testament to our commitment to making the data findable, accessible, interoperable, and reusable (FAIR), ensuring its maximum research and societal impact.

Reuse potential

The Earth Hologenome Initiative was established to promote high-quality, open hologenomic research on wild animals and their associated microorganisms. This data release, like those to follow, reflects our commitment to fostering collective efforts to understand and conserve biodiversity on our planet. Following the norms set in the Bermuda Principles, Fort Lauderdale agreement, and Toronto International Data Release Workshop [40], the authors kindly request users to respect the rights of the many researchers who invested significant effort in collecting samples and generating data for primary research. For 1 year following this article's publication, anyone wishing to use these data to investigate animal or microbial ecological and evolutionary ques-

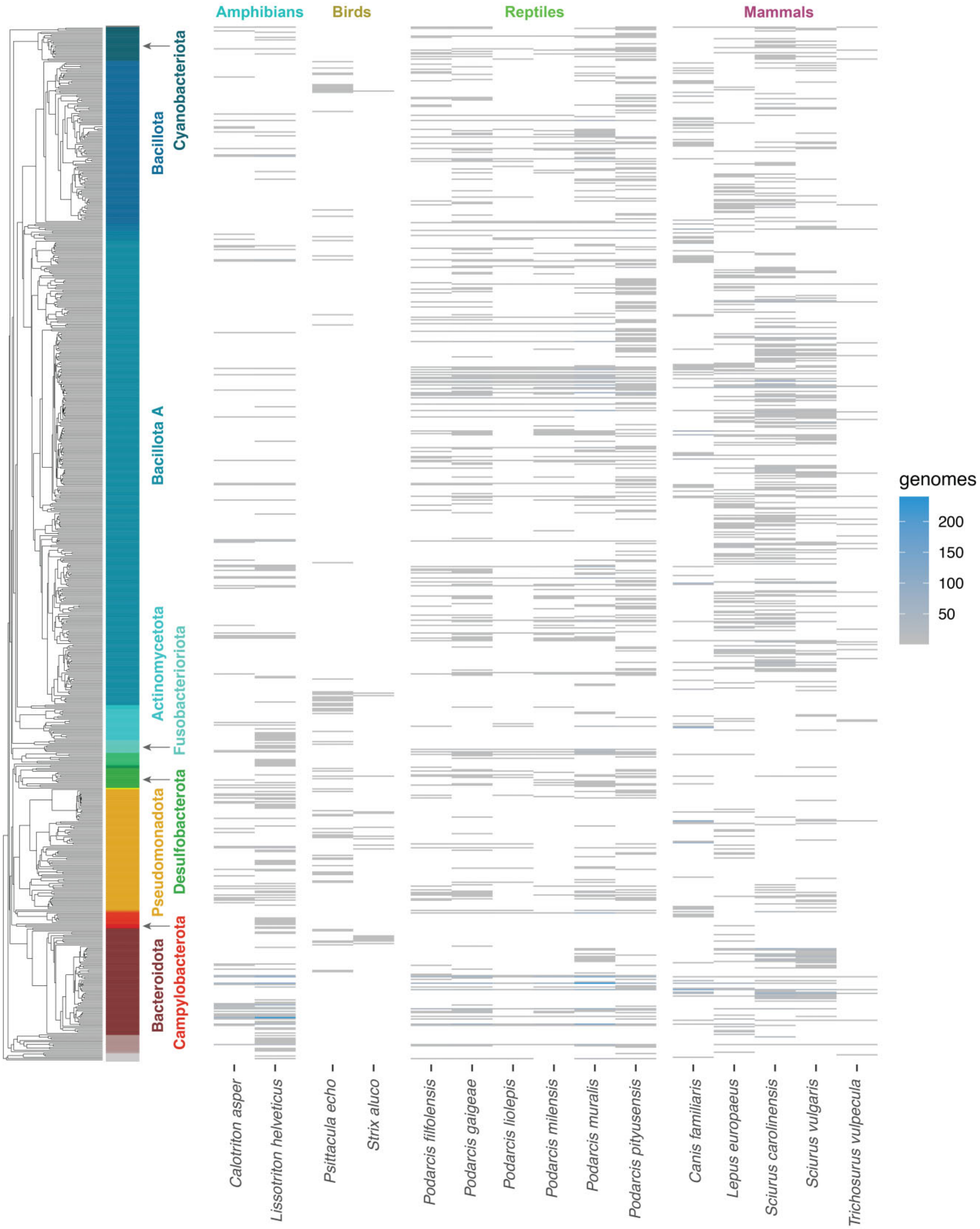

Figure 3: Host breadth of the reconstructed bacterial taxa. Only genomes reconstructed from individual assemblies are displayed in this figure. *Chalcides striatus*, *Geospizopsis unicolor*, *Natrix astreptophora*, *Plecotus auritus*, *Salamandra atra*, and *Zonotrichia capensis* did not yield any metagenome-assembled genomes from individual assemblies. Note that only the most abundant bacterial phylum names are displayed for the sake of visualization. Exact data can be found in the supplementary materials.

tions should first contact the corresponding author. Following this communication, the EHI Management will facilitate discussions between interested users and the original researchers to ensure efforts are coordinated with the people that are already working with these data.

Availability of Source Code and Requirements

Project name: Earth Hologenome Initiative Data Release 1
Project homepage: https://github.com/earthhologenome/EHI_data_release_1
Operating system(s): Platform independent
Programming language: R
License: CC0

Abbreviations

ABS: access and benefit-sharing; ANI: average nucleotide identity; EHI: Earth Hologenome Initiative; ENA: European Nucleotide Archive; FAIR: findable, accessible, interoperable, and reusable; GB: gigabases; MAG: metagenome-assembled genome.

Acknowledgments

The EHI could not be conceived without the trust and economic support provided by the Danish National Research Foundation through grant DNRF143 and the Carlsberg Foundation through grant CF20-0460. Additionally, C.R. was funded by the European Union through an MSCA Postdoctoral Fellowship (HORIZON-MSCA-2021-PF-01; Grant ID: 101066225). Special thanks go to Anders J. Hansen, head of the Globe Institute, for signing numerous participation agreements and to project managers Aoife Leonard and Ella Lattenkamp, whose management ensured the smooth handling of the practical aspects of the EHI. The Estación Biológica Cantábrica and Asociación Amigos de Doñana made possible the collection of bat samples. J.F. was supported by a PhD grant from FCT—Fundação para Ciência e Tecnologia (PD/BD/150645/2020). We also acknowledge financial support from the UID Centre for Environmental and Marine Studies (CESAM; LA/P/0094/2020) through national funds. P.A.H. was supported by research grant 25925 from VILLUM FONDEN. J.A. was supported by Margarita Salas (contract no. MS21-053) from the University of Valencia. P.K.I. and R.R. were supported by 5 smaller national funds, as well as the Sirius Dog Sled Patrol of Denmark, and veterinarian Lone Lykke Hansen.

Authors Contributions

N.G., R.E., and A.A. wrote the manuscript. N.G., C.P., G.M.B., and J.L. contributed to the data generation. R.E., O.A., and A.A. conducted the data analysis. J.F. and E.F. collected the *Chalcides striatus* and *Natrix astreptophora* samples. F.A., T.S., and C.P. collected *Podarcis muralis* samples. G.M.B. collected samples of *Podarcis muralis*, *Podarcis liolepis*, and *Calotriton asper*. L.W., C.R, M.S., and C.T. collected the *Sciurus vulgaris* and *Sciurus carolinensis* samples. A.O.S., M.G., and M.H.W. collected the *Perisoreus infaustus* samples. G.P.L., J.A., P.A., and F.C. collected *Podarcis muralis* and *Podarcis pityusensis* samples. F.C., R.G-R., and T.U. collected *Podarcis pityusensis* samples. N.F. and J.A. collected *Podarcis filfolensis* samples. N.F., J.A., G.M.W., and I.P. collected *Podarcis gaigeae* samples. T.U., N.F., G.M.W., and I.P. contributed with *Podarcis milensis* samples. R.E. collected the *Trichosurus vulpecula* samples. J.J. and P.A. collected the *Plecotus auritus* samples. P.H. and E.B. collected *Zoonotrichia capensis* and *Geospizopsis unicolor* samples. P.K.I. and R.R. collected the *Canis lupus familiaris* samples.

Funding

Danish National Research Foundation, DNRF (ROR: 00znyv691): DNRF143

Carlsberg Foundation (ROR: 01kpjmx04): CF20-0460

European Commission (ROR: 00k4n6c32) (HORIZON-MSCA-2021-PF-01): 101066225

Fundação para Ciência e Tecnologia (ROR: 00snfqn58): PD/BD/150645/2020

CESAM (ROR: 03q36qm19): LA/P/0094/2020

Villum Fonden (ROR: 05nqkay65): 25925

Fundación Margarita Salas: MS21-053

Data Availability

Raw sequencing data belonging to the first EHI data release are available at the European Nucleotide Archive, under Bioproject accession number PRJEB76898, which is nested within the Earth Hologenome Initiative's umbrella Bioproject PRJEB51837. A tarball containing fasta files of all MAGs was deposited in Zenodo [41]. Details of the specific sample and data accession numbers, their associated metadata, and the code used for visualization and summary statistics can be found in GitHub [42], with a snapshot in Zenodo [43]. In addition, the GitHub repository is also archived in Software Heritage [44]. The overview of all EHI data is available at the EHI database [36].

Competing Interests

The authors declare that they have no competing interests.

References

1 Leonard A, Alberdi A, Earth Hologenome Initiative Consortium, . A global initiative for ecological and evolutionary hologenomics. Trends Ecol Evol. 2024;39(7):616–20. https://doi.org/10.1016/j.tree.2024.03.005.

2 McFall-Ngai M, Hadfield MG, Bosch TCG, et al. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci USA. 2013;110(9):3229–36. https://doi.org/10.1073/pnas.1218525110.

3 Bordenstein SR, The Holobiont Biology Network, Holobiont Biology Network. The disciplinary matrix of holobiont biology. Science. 2024;386(6723):731–32. https://doi.org/10.1126/science.ado2152.

4 Alberdi A, Andersen SB, Limborg MT, et al. Disentangling host–microbiota complexity through hologenomics. Nat Rev Genet. 2022;23:281–97. https://doi.org/10.1038/s41576-021-00421-0.

5 Pietroni C, Gaun N, Leonard A, et al. Hologenomic data generation and analysis in wild vertebrates. Methods Ecol Evol. 2025;16(1):97–107. https://doi.org/10.1111/2041-210X.14456.

6 Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29(1):51–63. https://doi.org/10.1016/j.tree.2013.09.008.

7 Taş N, Jong AE, Li Y, et al. Metagenomic tools in microbial ecology research. Curr Opin Biotechnol 2021;67:184–91.

8 Hernández M, Ancona S, Hereira-Pacheco S, et al. Comparative analysis of two nonlethal methods for the study of the gut bacte-

rial communities in wild lizards. Integr Zool. 2023;18(6):1056–71. https://doi.org/10.1111/1749-4877.12711.

9 Ingala MR, Simmons NB, Wultsch C, et al. Comparing microbiome sampling methods in a wild mammal: fecal and intestinal samples record different signals of host ecology, evolution. Front Microbiol 2018;9:803. https://doi.org/10.3389/fmicb.2018.00803.

10 Kohn MH, York EC, Kamradt DA, et al. Estimating population size by genotyping faeces. Proc Biol Sci. 1999;266(1420);657–63. https://doi.org/10.1098/rspb.1999.0686.

11 Quince C, Walker AW, Simpson JT, et al. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44. https://doi.org/10.1038/nbt.3935.

12 Chaumeil P-A, Mussig AJ, Hugenholtz P, et al. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38(23):5315–16. https://doi.org/10.1093/bioinformatics/btac672.

13 Jain C, LM Rodriguez-R, Phillippy AM, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114. https://doi.org/10.1038/s41467-018-07641-9.

14 Pietroni C, Alberdi A. The Earth Hologenome Initiative Laboratory Workflow. 2023. https://www.earthhologenome.org/laboratory. Accessed 2023 January 12.

15 Carøe C, Gopalakrishnan S, Vinner L, et al. Single-tube library preparation for degraded DNA. Methods Ecol Evol. 2018;9(2):410–19.

16 Murray DC, Coghlan ML, Bunce M. From benchtop to desktop: important considerations when designing amplicon sequencing workflows. PLoS One. 2015;10(4):e0124671. https://doi.org/10.1371/journal.pone.0124671.

17 Eisenhofer R, Alberdi A. The Earth Hologenome Initiative Bioinformatics Workflow. 2023. https://www.earthhologenome.org/bioinformatics. Accessed 2023 January 12.

18 UCPH. ERDA: Electronic Data Archive at the University of Copenhagen. 2025. https://erda.dk. Accessed 11 August 2025.

19 Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–22.

20 Yoo AB, Jette MA, Grondona M. SLURM: simple Linux utility for resource management. In: Job scheduling strategies for parallel processing. Berlin, Germany: Springer; 2003:44–60.

21 Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.

22 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–59. https://doi.org/10.1038/nmeth.1923.

23 Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–79. https://doi.org/10.1093/bioinformatics/btp352.

24 Rodriguez-R LM, Gunturu S, Tiedje JM, et al. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems. 2018;3(3). https://doi.org/10.1128/mSystems.00039-18.

25 Woodcroft BJ, Aroney STN, Zhao R, et al. SingleM and Sandpiper: robust microbial taxonomic profiles from metagenomic data. Biorxiv. https://www.biorxiv.org/content/biorxiv/early/2024/01/31/2024.01.30.578060. A ccessed 26 Mar 2024.

26 Eisenhofer R, Alberdi A, Woodcroft BJ. Quantifying microbial DNA in metagenomes improves microbial trait estimation. ISME Commun. 2024;4(1):ycae111. https://doi.org/10.1093/ismeco/ycae111.

27 Li D, Liu C-M, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–76. https://doi.org/10.1093/bioinformatics/btv033.

28 Alneberg J, Bjarnason BS, de Bruijn I, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11(11):1144–46. https://doi.org/10.1038/nmeth.3103.

29 Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32(4):605–7. https://doi.org/10.1093/bioinformatics/btv638.

30 Kang DD, Li F, Kirton E, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359. https://doi.org/10.7717/peerj.7359.

31 Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–75. https://doi.org/10.1093/bioinformatics/btt086.

32 Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6(1):1–13. https://doi.org/10.1186/s40168-018-0541-1.

33 Parks DH, Imelfort M, Skennerton CT, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043–55. https://doi.org/10.1101/gr.186072.114.

34 Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996–1004. https://doi.org/10.1038/nbt.4229.

35 Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–90. https://doi.org/10.1093/bioinformatics/btg412.

36 Alberdi A. The Earth Hologenome Initiative Database. 2025. www.earthhologenome.org/database. Accessed 11 Aug 2025.

37 Aizpurua O, Dunn RR, Hansen LH, et al. Field and laboratory guidelines for reliable bioinformatic and statistical analysis of bacterial shotgun metagenomic data. Crit Rev Biotechnol. 2023;1–19.

38 Bowers RM, Kyrpides NC, Stepanauskas R, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35(8):725–31. https://doi.org/10.1038/nbt.3893.

39 www.cbd.int.

40 Birney E, Hudson T, Green E, et al. Prepublication data sharing. Nature. 2009;461:168–70.

41 Alberdi A. Earth Hologenome Initiative (EHI) Data Release 1: sequence files of metagenome assembled genomes (MAGs) [Data set]. Zenodo. 2025. https://doi.org/10.5281/zenodo.16689666.

42 Alberdi A. Earth Hologenome Initiative Data Release 1. Github. 2025. https://github.com/earthhologenome/EHI_data_release_1.

43 Alberdi A. Earth Hologenome Initiative (EHI) Data Release 1: metadata files and analysis code (1.0.3). Zenodo. 2025. https://doi.org/10.5281/zenodo.15347437.

44 Gaun N, Pietroni C, Martin-Bideguren G, et al. The Earth Hologenome Initiative: data Release 1 (Version 1) [Computer software]. Software Heritage. 2025. https://archive.softwareheritage.org/swh:1:snp:c4861440bd494ef8ab8c9d4390d1492a934501f6.

Received: May 27, 2025. **Revised:** August 2, 2025. **Accepted:** August 13, 2025