A global initiative for ecological and evolutionary hologenomics

Keywords

Access and benefit-sharing, Biodiversity, Host-microbe, Microbiome, Research network

Author names and affiliations

Aoife Leonard[1], Earth Hologenome Initiative consortium, Antton Alberdi[1]

Correspondence: Antton Alberdi (antton.alberdi@sund.ku.dk)

Earth Hologenome Initiative consortium: Javier Abalos[2], Titus Adhola[3], Windsor Aguirre[4], Ostaizka Aizpurua[1], Shahzad Ali[5], Franco Andreone[6], Fabien Aubret[7], Hefer D. Ávila-Palma[8], Lizbeth Fabiola Bautista Alcantara[9], Juan F. Beltrán[10], Rachel Berg[11], Thomas Bjørneboe Berg[12,13], Sandro Bertolino[14], Daniel T. Blumstein[15], Bazartseren Boldgiv[16], Zbigniew Borowski[17], Jean Philippe Boubli[18], Sven Büchner[19], Carlos Cabido[20], Carlos Camacho[21], Juan C. Chaparro[22], Anne Charmantier[23], Guillermo D'Elía[24], Luis P. da Silva[25,26], Bo Dalsgaard[27], Christophe de Franceschi[23], Ferran de la Cruz[25,26], Noé U. de la Sancha[28,29], Mathieu Denoël[30], Raphael Eisenhofer[1], Nathalie Feiner[2], Joana M. Fernandes[31], Jordi Figuerola[21,32], Leonida Fusani[33], Laura Gangoso[34], Roberto García-Roa[2], Stefania Gasperini[35], Nanna Gaun[1], M. Thomas P. Gilbert[1], Ivan Gomez-Mestre[21], Gary R. Graves[36,37], Jim Groombridge[38], Emilie A. Hardouin[39], Mauricio Hernández[40], L. Gerardo Herrera M.[41], Kathy Hodder[39], Peter A. Hosner[37], Natali Hurtado[42], Javier Juste[21], Sarah C. L. Knowles[43], Kevin D. Kohl[44], Carmi Korine[45], Yurii V. Kornilev[46], Stephanie Kramer-Schadt[47], Xavier Lambin[48], Ella Z. Lattenkamp[1], Jonas Lauritsen[1], Guoliang Li[49], Celeste María López[50], Adrià López Baucells[51], Tali Magory Cohen[52,1], Emiliano Manzo[35], Mélanie Marteau[53], Lynn B. Martin[54], Garazi Martin Bideguren[1], Maria Vittoria Mazzamuto[55], Ara Monadjem[56], Pirmin Nietlisbach[11], Daniel Bilyeli Øksnebjerg[1], Jasmin G. Packer[57], Michael L. Pepke[1], Juan M. Peralta-Sánchez[10], Adrián Perdomo[58], Guillem Pérez i de Lanuza[59], Carlotta Pietroni[1], Mathilde Poyet[60,61,62], Carsten Rahbek[37], Thiruvarangan Ramaraj[28], Yuliaxis Ramayo Caldas[63], Orly Razgour[64], Hugo Rebelo[65], Jiří Reif[66], Rebecca Rimbach[67,68], Ricardo Rocha[43], Rita G. Rocha[25,26], Carlos Rodríguez Fernandes[65], Claudia Romeo[1], Suvi Ruuskanen[69], Scott K. Sakaluk[11], Francesca Santicchia[70], Tom Sarraude[7], Rune Sørås[71], Martina Spada[70], Michael A. Steele[72], Mason R. Stothart[73], Emina Sunje[74,75,76], Alex O. Sutton[77], Marta Szulkin[78], Yu Takahata[79], Charles F. Thompson[11], Kasper Thorup[37], Geizecler Tomazetto[44], Laura Torrent[51], Nia Toshkova[46], Claudia Tranquillo[70], Manfredo Alejandro Turcios-Casco[8], Tobias Uller[2], Isolde van Riemsdijk[1], Guillermo Velo-Antón[80], Goedele Verbeylen[81], Elin Videvall[82,83], Christian C. Voigt[47], Lucas A. Wauters[70], Maren Wellenreuther[84,85], Alexey Yanchukov[86]

[1] Center for Evolutionary Hologenomics (CEH), Globe Institute, University of Copenhagen. Øster Farimagsgade 5, 1353 Copenhagen, Denmark
[2] Department of Biology, Lund University. Sölvegatan 37, 22362 Lund, Sweden
[3] University of Nairobi. P.O. Box 30197-00100, Nairobi, Kenya
[4] Department of Biological Sciences, DePaul University. 2325 North Clifton Avenue, Chicago, IL, 60614, USA
[5] Department of Wildlife and Ecology, Discipline of Zoology, Wildlife Epidemiology and Molecular Microbiology Laboratory (One Health Research Group), University of Veterinary and Animal Sciences. Lahore, Ravi Campus, Pattoki, Pakistan
[6] Museo Regionale di Scienze Naturali. Via Giovanni Giolitti, 36, I-10123 Torino, Italy
[7] Station d'Ecologie Théorique et Expérimentalel, CNRS. 2 route du CNRS, 09200 Moulis France
[8] Asociación para la Sostenibilidad e Investigación Científica en Honduras (ASICH). Tegucigalpa, M.D.C., Francisco Morazán, Honduras
[9] Colección Privada y Centro de Rescate de Fauna Silvestre El Ocotal. Sabanagrande, Francisco Morazán, Honduras
[10] Department of Zoology, University of Seville. Avda. Reina Mercedes, s/n, 41012 Sevilla, Spain
[11] School of Biological Sciences, Illinois State University. Campus Box 4120, Normal, IL 61790-4120, USA
[12] Naturama, Natural History Museum of Southern Denmark. Naturama, Dronningemaen 30, 5700 Svendborg
[13] Department of Biology, University of Southern Denmark. Campusvej 55, 5230 Odense, Denmark
[14] Department of Life Sciences and Systems Biology, University of Turin. Via Accademia Albertina 13, 10123 Torino, Italy
[15] Department of Ecology & Evolutionary Biology, University of California Los Angeles. 621 Young Drive South, Los Angeles, CA 90095-1606, USA
[16] Department of Biology, School of Arts and Sciences, National University of Mongolia. Ikh Surguuliin Gudamj 1, Ulaanbaatar 14201, Mongolia
[17] Department of Forest Ecology, Forest Research Institute. Braci Lesnej 3, Sekocin Stary, 05-090 Raszyn, Poland
[18] School of Science, Engineering and the Environment, University of Salford. G25 Peel Building, Manchester M5 4WT, UK
[19] SENCKENBERG Museum of Natural History. Am Museum 1, D 02826 Görlitz, Germany
[20] Aranzadi Science Society. Zorroagagaina Kalea 11, 20014 Donostia, Gipuzkoa, Spain
[21] Estación Biológica de Doñana, CSIC. Avda. Américo Vespucio 26, 41092 Sevilla, Spain
[22] Museo de Biodiversidad del Perú. Urbanización Mariscal Gamarra A-61, Zona 2, Cusco, Perú
[23] CEFE, Univ Montpellier, CNRS, EPHE, IRD. 1919, route de Mende 34293 Montpellier 5, France
[24] Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile. Campus Isla Teja s/n, Valdivia, Chile
[25] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Universidade do Porto. Campus de Vairão, 4485-661 Vairão, Portugal

[26] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO. Campus de Vairão, 4485-661 Vairão, Portugal
[27] Section for Molecular Ecology & Evolution, Globe Institute, University of Copenhagen. Øster Farimagsgade 5, 1353 Copenhagen, Denmark
[28] DePaul University. CDM 738, 243 South Wabash Avenue, Chicago. IL, 60604, USA
[29] Negaunee Integrative Research Center, The Field Museum of Natural History. 303 E Main St, Negaunee, MI 49866, USA
[30] Laboratory of Ecology and Conservation of Amphibians (LECA), Freshwater and OCeanic Unit of reSEarch (FOCUS), University of Liege. 22 quai van Beneden, 4020 Liège, Belgium
[31] Department of Biology & CESAM, University of Aveiro, Portugal. Campus de Santiago, 3810-193 Aveiro, Portugal
[32] Centro de Investigaciones Biomedicas en Red en Epidemiología y Salud Pública. Madrid, Spain
[33] Konrad Lorenz Institute of Ethology, University of Veterinary Medicine. Savoyenstrasse 1a, 1160 Vienna, Austria
[34] Department of Biodiversity, Ecology and Evolution, Faculty of Biological Sciences, Complutense University of Madrid. C. de José Antonio Novais, 12, Moncloa - Aravaca, 28040
[35] Fondazione Ethoikos. Convento dell'Oseervanza, 53030 Radicondoli, Italy
[36] National Museum of Natural History, Smithsonian Institution. 10th St. & Constitution Ave. NW, Washington, DC 20560, USA
[37] Center for Macroecology, Evolution and Climate, Globe Institute, University of Copenhagen. Universitetsparken 15, 2100 Copenhagen, Denmark
[38] Durrell Institute of Conservation and Ecology, Division of Human and Social Sciences, University of Kent. Canterbury, Kent, CT2 7NR, UK
[39] Life and Environmental Sciences department, Bournemouth University. Talbot campus, BH125BB Poole, United Kingdom
[40] Instituto de Investigaciones en Microbiología, Facultad de Ciencias, National Autonomous University of Honduras. Tegucigalpa 11101, Honduras
[41] Estación de Biología Chamela, Universidad Nacional Autónoma de México. Instituto de Biología, Universidad Nacional Autónoma de México, Apartado Postal 21, 48980, San Patricio, Jalisco, México
[42] Universidad del Bío-Bío. Avda. Andrés Bello 720, Chillán, Ñuble, Chile
[43] Department of Biology, University of Oxford. OX1 3SZ, Oxford, UK
[44] University of Pittsburgh. Pittsburgh, PA, 15224, USA
[45] Mitrani Department of Desert Ecology, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev. Sede Boqer Campus, 8499000 Midreshet Ben-Gurion, Israel
[46] National Museum of Natural History, Bulgarian Academy of Sciences. 1 Tsar Osvoboditel Blvd., 1000 Sofia, Bulgaria
[47] Leibniz Institute for Zoo and Wildlife Research. Alfred-Kowalke-Str. 17, 10315 Berlin, Germany
[48] University of Aberdeen. Zoology building, Tillydrone Avenue, Aberdeen AB24 2TZ, UK
[49] Jiangxi Key Laboratory of Conservation Biology, College of Forestry, Jiangxi Agricultural University. Zhiminda Road 1101, Nanchang 330045, China

[50] Departamento de Ciências Biológicas, Universidade Estadual de Santa Cruz. Ilhéus, Bahia, Brasil
[51] BiBio, Natural Sciences Museum of Granollers. Av/Francesc Macià 51, 08402, Granollers, Spain
[52] Department of Evolution and Ecology, University of California Davis. 1 Shields Ave, Davis, CA 95616, USA
[53] Groupe Mammalogique Normand. 32 route de Pont-Audemer, 27260 Epaignes, France
[54] Centers for Genomics and Global Health and Interdisciplinary Disease Research, University of South Florida. 3720 Spectrum Blvd, Tampa, Florida, 33612, USA
[55] Haub School of Environment and Natural Resources, University of Wyoming. 1000 E University Ave, Laramie, WY, USA, 82071
[56] Department of Biological Sciences, University of Eswatini. University of Eswatini, Private Bag 4, Kwaluseni, Eswatini
[57] The University of Adelaide, Adelaide, Australia. North Terrace, Adelaide SA 5005, Australia
[58] Ornithological Research Institute of Honduras (IOH). Los dolores neighborhood, La fuente Street, Maximo Jerez Av. 549. Tegucigalpa, Francisco Morazán, Honduras
[59] Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València. APT. 22085. 46071 València, Spain
[60] Institute of Experimental Medicine, Kiel University, Germany. UKSH Campus Kiel Haus U27, Schwanenweg 20, 24105 Kiel
[61] Global Microbiome Conservancy. NA
[62] Massachusetts Institute of Technology, United States. NA
[63] Institut de Recerca i Tecnologia Agroalimentaris, IRTA. Torre Marimon, 08140 Caldes de Montbui, Barcelona
[64] Biosciences, University of Exeter. Prince of Wales Road, Exeter, EX4 4PS, UK
[65] cE3c - Centre for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa. 1749-016 Lisboa, Portugal
[66] Institute for Environmental Studies, Faculty of Science, Charles University. Benátská 2, 128 01 Praha 2, Czechia
[67] Department of Behavioural Biology, University of Münster. Badestraße 9, 48149 Münster, Germany
[68] Evolutionary Anthropology, Duke University. 104 Biological Sciences, Durham, NC 27708, USA
[69] Department of Biological and Environmental Science, University of Jyväskylä. Survontie 9C, 40500, Jyväskylä, Finland
[70] Department of Theoretical and Applied Sciences, Università degli Studi dell'Insubria. via J. H. Dunant 3, 21100 Varese, Italy
[71] Department of Biology, Norwegian University of Science and Technology. Trondheim, NO-7491, Norway
[72] Department of Biology and Earth and Env. Science, Wilkes University. Wilkes-Barre, PA 18766, USA
[73] University of Calgary. 3330 Hospital Dr NW, Calgary, AB, T2N 4N1 Canada

[74] Faculty of Science, University of Sarajevo. Zmaja od Bosne 33-35, 71000 Sarajevo, Bosnia and Herzegovina
[75] Laboratory of Functional Morphology, University of Antwerp. Universiteitsplein 1, 2610 Wilrijk, Belgium
[76] Herpetological Association in B&H – ATRA (BH-HU: ATRA), Sarajevo, Bosnia and Herzegovina. Urijan dedina 137, 71000 Sarajevo, Bosnia and Herzegovina
[77] School of Environmental and Natural Sciences, Bangor University. Brambell Building, Deiniol Road, Bangor, Gwynedd, United Kingdom, LL57 2UR
[78] Faculty of Biology, University of Warsaw. University of Warsaw, Biological and Chemical Research Centre, Żwirki i Wigury 101, 02-089 Warsaw, Poland
[79] The Graduate University for Advanced Studies, SOKENDAI. Shonan Village, Hayama, Kanagawa 240-0193 Japan
[80] Facultad de Biología, Universidad de Vigo. Edificio de Ciencias Experimentales, E-36310, Vigo, Spain
[81] Natuurpunt Studie. Michiel Coxiestraat 11, B-2800 Mechelen, Belgium
[82] Department of Ecology and Genetics, Uppsala University. Norbyvägen 18D, 75236 Uppsala, Sweden
[83] Department of Ecology, Evolution and Organismal Biology, Brown University, USA. Providence, RI 02912, USA
[84] The New Zealand Institute for Plant and Food Research Ltd. Nelson, New Zealand
[85] School of Biological Sciences, University of Auckland. Auckland, New Zealand
[86] Zonguldak Bülent Ecevit University. 1 Universitesi Cd., 67100 Zonguldak, Türkiye

Abstract

The Earth Hologenome Initiative is a global collaboration to generate and analyse hologenomic data from wild animals and associated microorganisms using standardised methodologies underpinned by open and inclusive research principles. Initially focused on vertebrates, it aims to reexamine ecological and evolutionary questions by studying host-microbiota interactions from a systemic perspective.

Towards a deeper understanding of hologenomes

Most animals on Earth live in intimate association with microorganisms, which provide their host with an array of functional capabilities that can influence physiological processes and individual fitness [1]. Hologenomics consists of the joint analysis of host genomes, **microbial metagenome**s (see Glossary), and their combined functional attributes, to unravel the underpinnings of such host-microbiota interactions [2]. To date, hologenomic information remains scarce, dominated by **amplicon sequencing** data derived from humans, model organisms, and farm/domestic animals, with limited representation of wild animals (Fig. 1a; Supplemental Information). However, there is an ongoing rise in both the number of animal reference genomes and catalogues of functionally annotated prokaryotic genomes [3,4]. In addition, powerful tools for **shotgun sequencing**-based data generation and analysis have also been recently developed [5–7]. These significant advancements now enable the comprehensive exploration of (meta)genome-wide information in wild host-microbiota systems or holobionts [8] (Fig. 1b). Ecological niche shift, range expansion, invasion, domestication, disease transmission and resistance development, extinction, and adaptation to environmental change are but a few of the eco-evolutionary processes that can be revisited from the integrative perspective of host-microbiota systems [9].

The complexity, spatiotemporal variability, and scale-dependent nature of eco-evolutionary processes require a large and coordinated effort to systematically generate comparable host-microbiota data from a taxonomically, ontogenically, functionally, environmentally, and geographically representative number of animal species worldwide. In response to this need, the Earth Hologenome Initiative (EHI, www.earthhologenome.org) was established to promote, facilitate, coordinate, and standardise hologenomic research on wild organisms. Focused on terrestrial vertebrates in its initial phase, the EHI encompasses projects with diverse study designs and goals following standardised and open-access sample collection and preservation, data generation, and data management criteria. By establishing a robust consortium and implementing an open, transparent database, we strive to actively cultivate collaboration and knowledge transfer, improve data generation quality and efficiency, and yield beyond state-of-the-art scientific outputs.

Goals and anticipated outcomes

The EHI aims to achieve five strategic goals that will benefit the scientific community in terms of scientific outputs, methodological development, infrastructure availability, and networking, all grounded in solid ethical principles (Fig. 1c).

Science

The main scientific objective of the EHI is to investigate eco-evolutionary processes through a host-microbiota lens, by jointly analysing whole-genome (re)sequencing data of wild animal hosts and genome-resolved metagenomic data of associated microbial communities, encompassing large spatio-temporal scales, including host genetic structure, host phenotypic traits, functional microbiome properties, and environmental features. Accounting for these factors increases our capacity to generate higher-order conclusions about animal-microbiota interactions, and their contributions to eco-evolutionary processes. Some examples include microbial contributions to local adaptation, adaptive radiation of host animals, and host contributions to functional diversification and spatial distribution of microorganisms. Collectively, these findings can provide valuable insights for policymakers and managers, empowering the development and implementation of microbiome-aware wildlife conservation and management strategies [10].

Methodology

The primary methodological objective of the EHI is to develop and implement standardised sampling, preservation, and laboratory methods based on open resources and knowledge. This effort aims to address the limited comparability and reproducibility of microbiome studies to date, resulting from the high sensitivity of microbiome analyses to cross-contamination and variations in sample collection, preservation, and data generation methods [11]. All procedures are openly shared with the research community through the EHI website and developed so that they can be reproduced, automated, and deployed in different laboratories, with maximum cost-effectiveness.

Infrastructure

The main infrastructure goal of the EHI is to pioneer a novel strategy for transparent and reproducible research. This entails providing other researchers with direct access to real-time information on project progress, sample processing updates and data generation status. In addition, we aim to maximise outcomes from fieldwork efforts by making biological samples and DNA extracts available for other researchers, whenever allowed by legislation. This will aid multidisciplinary research investigations and help to decrease costly field sampling and disturbance of wild animal populations while fostering collaboration among researchers.

Network

Disentangling the complexity of host-microbiota interactions in wild animals requires joint contributions of researchers encompassing multiple disciplines, geographic regions, and career

development stages. This network is essential to make the most informed decisions regarding the scientific questions to be addressed, collect samples worldwide, and analyse and interpret the generated results. The EHI network provides an opportunity to build international bridges to foster multi-directional knowledge transfer among researchers interested in ecological and evolutionary hologenomics.

Ethics

Characterising the diversity, distribution, and structure of animal-microbiota interactions across Earth is only achievable through the inclusive collaboration of diverse researchers and communities around the world, and responsible sharing of local genetic resources and knowledge. As such, the EHI is governed by open science, embracing **CARE** and **FAIR** data governance principles [12,13], and complying with all international, national and regional regulations stemming from the United Nations' Convention on Biological Diversity (www.cbd.int).

Progress status and future steps

Conceived in 2020, the EHI is progressing through three phases (Fig. 2a). The initial developmental phase (2020–2023) focused on building the administrative, logistic and technical framework, and generating preliminary data (Supplemental information). This resulted in the distribution of 170 standardised EHI sampling kits among 86 participants, which yielded a cumulative submission of over 8,500 samples from 244 distinct vertebrate species across the world (Fig. 2b). The generation of hologenomic information from these samples has already produced >10 **terabases** of sequencing data, with an even split between host genomic and metagenomic data. **Reference genome**s with variable assembly levels were only available for 19% of the analysed species (Supplemental information), which highlights the ongoing need to coordinate information exchange with reference genome-generating consortia [14]. The other half of the generated data, namely the metagenomic fraction, yielded nearly 30,000 draft bacterial genomes (Fig. 2c). Fewer than 15% of these genomes were annotated at the species level, underscoring the fact that wild vertebrates harbour an enormous quantity of bacteria heretofore undescribed by genomic science. The knowledge generated in this preliminary phase has been crucial in shaping the design and guiding future steps. During the next consolidation phase (2024–2026) we are initiating targeted studies concentrating on one or a few focal taxa, following study-specific designs. The subsequent integration step (2027–2029) will entail the utilisation or reutilisation of hologenomic data derived from diverse individuals, populations, and species, with the goal of unravelling global patterns and addressing overarching biological questions at the forefront of science.

The Earth Hologenome Initiative boldly steps into the realm of a largely unexplored scientific territory, with significant work ahead to achieve a comprehensive overview of host-microbiota systems or holobionts across taxa and biomes. With over 15 active projects and more than 60 participating agreements in place, a rich scientific output is anticipated. Furthermore, the initiative's transparent procedures and collaborative ethos are likely to attract a wave of innovative researchers equipped with fresh perspectives and advanced analytical skills. This

collective drive is poised to propel pioneering hologenomic research to deepen our insights into the intricate interconnectedness of Earth's life forms, while tackling challenges at the frontiers of the global research, biodiversity and sustainability agendas [15].

Acknowledgements

The Earth Hologenome Initiative could not be conceived without the trust and economic support provided by the Danish National Research Foundation through the grant DNRF143, and the Carlsberg Foundation through the grant CF20-0460. The authors also acknowledge the NERC (grant NE/M018660/1), and University of Exeter Europe Network Fund for the economic support provided to ongoing projects embedded within the EHI. M. Denoël is a Research Director at F.R.S.-FNRS (Fonds de la Recherche Scientifique). C. Rodríguez Fernandes thanks the support of cE3c through an assistant researcher contract (FCiência.ID contract #366) and FCT (Fundação para a Ciência e a Tecnologia) for Portuguese National Funds attributed to cE3c within the projects UIDB/00329/2020, UIDP/00329/2020, and LA/P/0121/2020. We also thank all the researchers who participated in the fieldwork, as well as the support people who facilitated the administrative paperwork, with a special mention for the Head of the Globe Institute Anders J. Hansen, who has signed dozens of participation agreements. Finally, we acknowledge that some of the elements included in the figures have been obtained and modified from the Integration and Application Network, University of Maryland—Center for Environmental Science (https://ian.umces.edu/media-library/symbols).

Glossary

Microbial metagenome: collective genetic material derived from a microbial community in a particular sample.
Amplicon sequencing data: genetic information obtained through sequencing of a selectively amplified DNA region.
Shotgun sequencing data: genetic information obtained through sequencing total DNA.
CARE principles: research principles that consider the rights and interests of Indigenous peoples to ensure collective benefit, authority to control, responsibility and ethics in working with research data.
FAIR principles: research principles that aim to facilitate data sharing by making research data findable, accessible, interoperable and reusable.
Terabase: genomic data of a magnitude of one trillion (10^12) nucleotide bases.
Reference genome: a representative version of a genome that serves as a template for comparing and analysing the genetic material of individuals within a species.

References

1. McFall-Ngai, M. *et al.* (2013) Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U. S. A.* 110, 3229–3236
2. Alberdi, A. *et al.* (2022) Disentangling host-microbiota complexity through hologenomics. *Nat. Rev. Genet.* 23, 281–297

3. Parks, D.H. *et al.* (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 50, D785–D794
4. Rhie, A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746
5. Shaffer, M. *et al.* (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* 48, 8883–8900
6. Olm, M.R. *et al.* (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868
7. Uritskiy, G.V. *et al.* (2018) MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6, 1–13
8. Bordenstein, S.R. and Theis, K.R. (2015) Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLoS Biol.* 13, e1002226
9. Hird, S.M. (2017) Evolutionary biology needs wild microbiomes. *Front. Microbiol.* 8, 725
10. West, A.G. *et al.* (2019) The microbiome in threatened species conservation. *Biol. Conserv.* 229, 85–98
11. Aizpurua, O. *et al.* (2023) Field and laboratory guidelines for reliable bioinformatic and statistical analysis of bacterial shotgun metagenomic data. *Crit. Rev. Biotechnol.* DOI: 10.1080/07388551.2023.2254933
12. Wilkinson, M.D. *et al.* (2019) Addendum: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 6, 6
13. Jennings, L. *et al.* (2023) Applying the “CARE Principles for Indigenous Data Governance” to ecology and biodiversity research. *Nature Ecology & Evolution* 7, 1547–1551
14. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
15. Peixoto, R.S. *et al.* (2022) Harnessing the microbiome to prevent global biodiversity loss. *Nat Microbiol* 7, 1726–1735

Figures

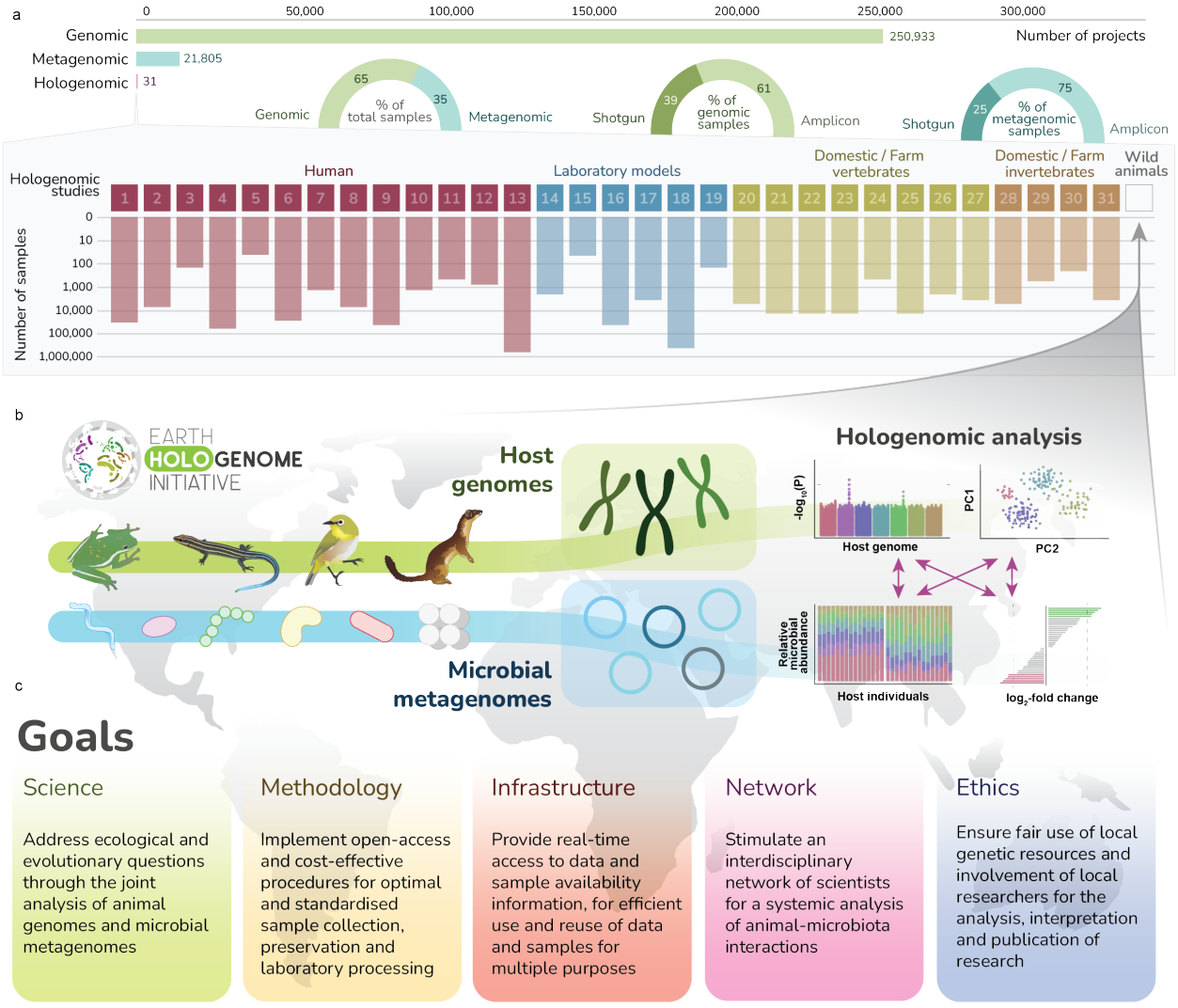

Figure 1. **Overview of hologenomic datasets available at public INSDC databases and the goals of the Earth Hologenome Initiative**. a) A programmatic search (details in supplemental information) for linked animal genomic and microbial metagenomic information at international research data archives yielded 31 research projects, none of them based on wild animals. b) The Earth Hologenome Initiative is conceived to fill this knowledge gap, by generating paired animal genomic and microbial metagenomic data in wild animals worldwide within an open and collaborative framework. c) Overview of the five strategic goals of the EHI. INSDC stands for International Nucleotide Sequence Database Collaboration, and interconnects DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive) and NCBI (National Center for Biotechnology Information, USA) databases.

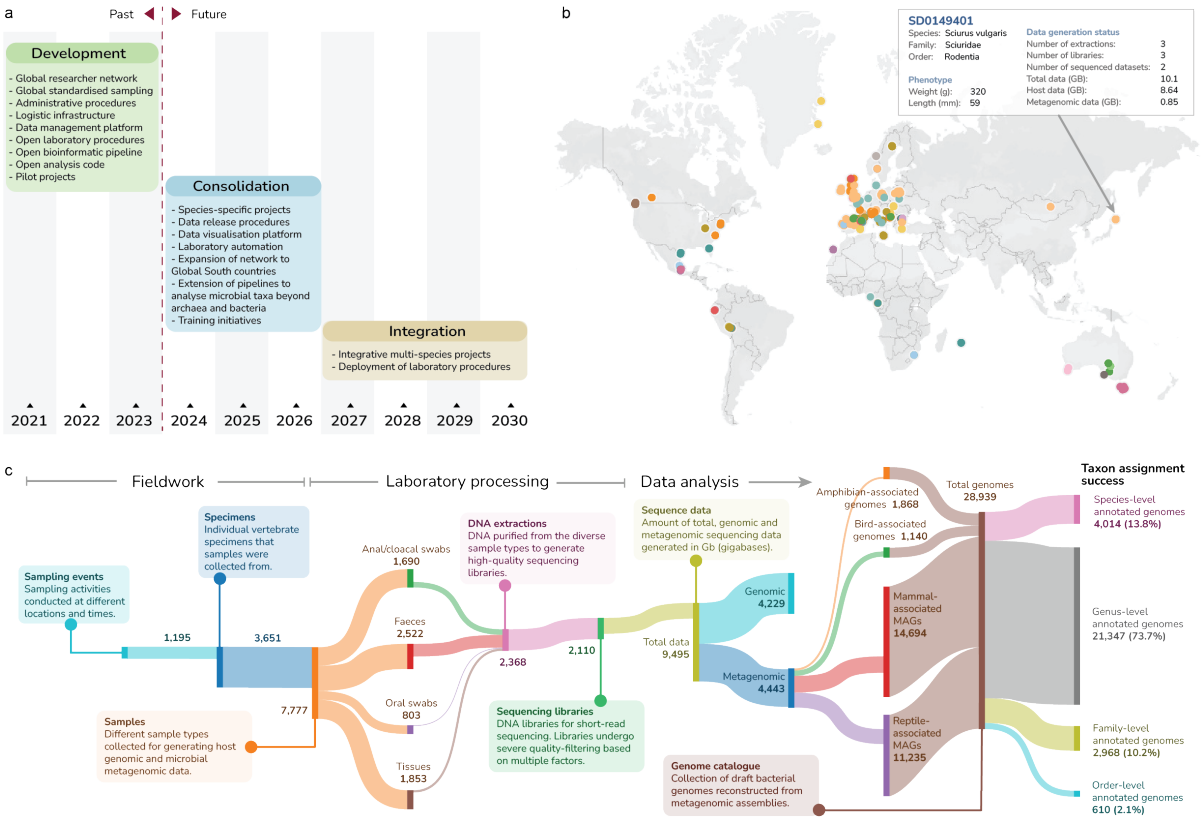

Figure 2. **Overview of sample collection and data processing of the EHI in 2023**. a) A 10-year roadmap indicating the main tasks associated with the three phases of the Earth Hologenome Initiative. b) World map with the geographic locations of sampled animals as shown on the www.earthhologenome.org website. Only samples stored in the EHI biobank and indexed in the database are displayed. The map reflects the geographical bias of the developmental phase of the EHI, which was primarily limited to Europe while administrative and logistic aspects of worldwide sampling were being addressed. c) Sankey diagram displaying an overview of sample and data amounts throughout the EHI workflow.

Supplemental information

A global initiative for ecological and evolutionary hologenomics

Leonard A, Earth Hologenome Initiative consortium, Alberdi A.
Correspondence: Alberdi, A. (antton.alberdi@sund.ku.dk)

Table of contents

A. Hologenomic data search

We conducted a programmatic search for paired vertebrate genomic and associated microbial metagenomic datasets available at the databases of the International Nucleotide Sequence Database Collaboration (INSDC), namely DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive) and NCBI (National Center for Biotechnology Information, USA). The analyses were conducted in R [1], using the packages tidyverse [2], taxize [3], httr [4] and readr [5]. An overview of the employed methodology is provided here, while the detailed bioinformatic code can be found in the following Github repository: github.com/earthhologenome/hologenomic_data_search

Data project search

Study or Bioproject searches were carried out on September 1st 2023 through ENA's API [6], using the following structure:

```
https://www.ebi.ac.uk/ena/portal/api/search?result=study&query=[CONDITIO
NS]&fields=study_accession,parent_study_accession,tax_id
```

Metagenomic data search

For metagenomic data, we searched for the below listed keywords combinations in the 'description', 'study_title', 'study_name', 'study_alias', 'project_name' and 'keywords' fields of all studies (Bioprojects) indexed at ENA, yielding 21,805 studies.

- "*metagenome*" + "*gut*"
- "*microbiome*" + "*gut*"
- "*metagenome*" + "*gut*"
- "*metagenome*" + "*intestin*"
- "*microbiome*" + "*intestin*"
- "*metagenome*" + "*intestin*"
- "*metagenome*" + "*fecal*"
- "*microbiome*" + "*fecal*"
- "*metagenome*" + "*fecal*"
- "*metagenome*" + "*faecal*"
- "*microbiome*" + "*faecal*"
- "*metagenome*" + "*faecal*"

Genomic data search and filtering

For genomic data, we searched for the keywords '* genom*' and 'Genome*' in the 'description', 'study_title', 'study_name', 'study_alias', 'project_name' and 'keywords' fields of all studies (Bioprojects) indexed at ENA. The search yielded 250,933 studies, which were subsequently filtered by tax_id (taxonomic identification code) to only retain data derived from amphibians, reptiles, birds and mammals.

Experiment information of paired studies

Connections between studies retrieved in the metagenomic and genomic searches were identified by pairing ‘study_accession’ and ‘parent_study_accession’ fields across all studies. This initial pairing yielded 95 projects, whose experiment accessions were fetched through an iterative search using the following structure:

```
for (STUDY_ACCESSION in STUDY_ACCESSIONS){
      https://www.ebi.ac.uk/ena/portal/api/search?result=read_experiment
      &query=study_accession=[STUDY_ACCESSION]&fields=study_accession,pa
      rent_study_accession,tax_id
}
```

The search retrieved 125,704 experiments associated with the 95 paired projects. Subsequently, only studies containing both experiments with ‘GENOMIC’ and ‘METAGENOMIC’ values in the ‘library_source’ field were retained. This filtering yielded 53 studies, which were reduced to 31 after a manual quality-check (Supplementary Table S1). Manually discarded studies primarily included metagenomic (from faeces and gut contents) and genomic (from isolates) data originally sampled from animal hosts, yet without host genomic data. Finally, summary statistics were calculated to obtain the numbers of amplicon- and shotgun-based experiments for generating genomic and metagenomic data in the target studies, which were visualised in Figure 1a.

We must note that the results might be missing projects lacking the search words in the metadata, while the number of samples might be overestimated, as the available metadata did not enable us to verify whether genomic and metagenomic datasets were derived from the same individual animals.

Table S1. **Overview of the 31 studies containing animal genomic and associated microbial metagenomic data**. Note the presence of 32 rows, because one of the studies (PRJEB43192) contains data on two species.

Accession	Taxid	Phylum	Order	Species	Group
PRJEB40770	256318	Chordata	Primates	*Homo sapiens*	Human
PRJEB40771	9606	Chordata	Primates	*Homo sapiens*	Human
PRJNA358636	9606	Chordata	Primates	*Homo sapiens*	Human
PRJNA46305	646099	Chordata	Primates	*Homo sapiens*	Human
PRJNA46941	408170	Chordata	Primates	*Homo sapiens*	Human
PRJNA48489	646099	Chordata	Primates	*Homo sapiens*	Human
PRJNA51441	646099	Chordata	Primates	*Homo sapiens*	Human
PRJNA661099	9606	Chordata	Primates	*Homo sapiens*	Human
PRJNA74933	646099	Chordata	Primates	*Homo sapiens*	Human
PRJNA74937	646099	Chordata	Primates	*Homo sapiens*	Human

PRJNA74951	646099	Chordata	Primates	*Homo sapiens*	Human
PRJNA74955	646099	Chordata	Primates	*Homo sapiens*	Human
PRJNA9558	9606	Chordata	Primates	*Homo sapiens*	Human
PRJNA9557	7955	Chordata	Cypriniformes	*Danio rerio*	Laboratory
PRJNA10621	10116	Chordata	Rodentia	*Rattus norvegicus*	Laboratory
PRJNA17401	410661	Chordata	Rodentia	*Mus musculus*	Laboratory
PRJNA608517	410661	Chordata	Rodentia	*Mus musculus*	Laboratory
PRJNA9559	10090	Chordata	Rodentia	*Mus musculus*	Laboratory
PRJNA98095	10117	Chordata	Rodentia	*Rattus rattus*	Laboratory
PRJNA43379	7959	Chordata	Cypriniformes	*Ctenopharyngodon idella*	Domestic
PRJNA73299	27706	Chordata	Centrarchiformes	*Micropterus salmoides*	Domestic
PRJEB43192	8030	Chordata	Salmoniformes	*Salmo salar*	Domestic
PRJEB43192	9031	Chordata	Galliformes	*Gallus gallus*	Domestic
PRJNA10804	9031	Chordata	Galliformes	*Gallus gallus*	Domestic
PRJNA10709	9940	Chordata	Artiodactyla	*Ovis aries*	Domestic
PRJNA10718	9823	Chordata	Artiodactyla	*Sus scrofa*	Domestic
PRJNA49537	9925	Chordata	Artiodactyla	*Capra hircus*	Domestic
PRJNA10726	9615	Chordata	Carnivora	*Canis lupus*	Domestic
PRJNA78065	307972	Echinodermata	Aspidochirotida	*Apostichopus japonicus*	Invertebrate
PRJNA10637	7091	Arthropoda	Lepidoptera	*Bombyx mori*	Invertebrate
PRJNA45885	7213	Arthropoda	Diptera	*Ceratitis capitata*	Invertebrate
PRJNA9555	7460	Arthropoda	Hymenoptera	*Apis mellifera*	Invertebrate

B. Availability of reference genomes

The standard procedures of the EHI include mapping the reads of all types of analysed samples to a reference host genome. This genome should ideally belong to the studied host species, yet availability of reference genomes is still limited. The following table (Table S2) shows the list of analysed species and employed reference genomes, which in 24 cases (21%) belonged to the same species, in 49 cases (43%) to another species from the same genus, in 39 cases (34%) to a member of the same family, and in 2 cases (2%) to species from the same order. Among the employed reference genomes, 41 had chromosome-level quality according to NCBI Assembly level definitions (highest quality, chromosomes are resolved), 65 had scaffold-level quality (medium-quality, long DNA sequences yet without chromosome assignment) and 8 had contig-level quality (low-quality, short DNA sequences without spatial structure).

Table S2. **Reference genomes employed for each analysed animal**. Relatedness refers to the lowest common taxonomic level between the analysed species and the species used as reference.

Analysed species	**Employed reference genome**	**Relatedness**	**Quality**	**Accession**
Acanthodactylus erythrurus	*Podarcis muralis*	Family	Chromosome	GCF_004329235.1
Acrocephalus scirpaceus	*Muscardinus avellanarius*	Genus	Scaffold	GCA_004027005.1
Apodemus flavicollis	*Apodemus sylvaticus*	Genus	Chromosome	GCA_947179515.1
Apodemus sylvaticus	*Apodemus sylvaticus*	Species	Chromosome	GCA_947179515.1
Barbastella barbastellus	*Corynorhinus townsendii*	Family	Scaffold	GCA_026230045.1
Canis familiaris	*Canis familiaris*	Species	Chromosome	GCA_011100685.1
Cathartes aura	*Cathartes aura*	Species	Scaffold	GCA_000699945.1
Chaerephon pumila	*Molossus molossus*	Family	Scaffold	GCF_014108415.1
Chalcides striatus	*Cryptoblepharus egeriae*	Family	Contig	GCA_030015325.1
Chloroceryle aenea	*Chloroceryle aenea*	Species	Scaffold	GCA_013399075.1
Chloroceryle amazona	*Chloroceryle amazona*	Species	Scaffold	GCA_027560015.1
Chloroceryle americana	*Chloroceryle inda*	Genus	Scaffold	GCA_027560095.1
Chloroceryle inda	*Chloroceryle inda*	Species	Scaffold	GCA_027560095.1
Coragyps atratus	*Cathartes aura*	Family	Scaffold	GCA_000699945.1
Cyanoloxia rothschildii	*Cardinalis cardinalis*	Family	Scaffold	GCA_014549065.1
Dasyurus geoffroii	*Sarcophilus harrisii*	Family	Chromosome	GCF_902635505.1
Dendrocincla merula	*Campylorhamphus procurvoides*	Family	Scaffold	GCA_013396655.1
Donacobius atricapilla	*Donacobius atricapilla*	Species	Scaffold	GCA_013397315.1
Emberiza cirlus	*Emberiza elegans*	Genus	Scaffold	GCA_024865655.1
Eptesicus nilssonii	*Eptesicus fuscus*	Genus	Scaffold	GCF_000308155.1
Eptesicus serotinus	*Eptesicus fuscus*	Genus	Scaffold	GCF_000308155.1
Eutamias sibiricus	*Eutamias sibiricus*	Species	Chromosome	GCA_025594165.1

Falco eleonorae	*Falco peregrinus*	Genus	Chromosome	GCA_023634155.1
Glauconycteris sp.	*Eptesicus fuscus*	Family	Scaffold	GCF_000308155.1
Habia rubica	*Cardinalis cardinalis*	Family	Scaffold	GCA_014549065.1
Hemitriccus griseipectus	*Tyrannus tyrannus*	Family	Scaffold	GCA_026770345.1
Hipposideros ruber	*Hipposideros armiger*	Genus	Scaffold	GCF_001890085.2
Hypocnemoides maculicauda	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Hypsignathus monstrosus	*Rousettus aegyptiacus*	Family	Chromosome	GCA_014176215.1
Iberolacerta aranica	*Lacerta viridis*	Family	Contig	GCA_900245905.1
Iberolacerta bonnali	*Lacerta viridis*	Family	Contig	GCA_900245905.1
Isleria hauxwelli	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Lacerta viridis	*Lacerta viridis*	Species	Contig	GCA_900245905.1
Lasiorhinus latifrons	*Vombatus ursinus*	Family	Scaffold	GCA_900497805.2
Lissotriton helveticus	*Pleurodeles waltl*	Order	Chromosome	GCA_002915635.3
Lissotriton vulgaris	*Pleurodeles waltl*	Order	Chromosome	GCA_002915635.3
Marmota sibirica	*Marmota himalayana*	Genus	Scaffold	GCA_005280165.1
Microtus agrestis	*Microtus agrestis*	Species	Contig	GCA_902806775.1
Miniopterus schreibersii	*Miniopterus schreibersii*	Species	Scaffold	GCA_004026525.1
Momotus coeruliceps	*Momotus momota*	Genus	Scaffold	GCA_028565915.1
Momotus momota	*Momotus momota*	Genus	Scaffold	GCA_028565915.1
Mops thersites	*Molossus molossus*	Family	Scaffold	GCF_014108415.1
Muscardinus avellanarius	*Muscardinus avellanarius*	Species	Scaffold	GCA_004027005.1
Myodes glareolus	*Myodes glareolus*	Species	Scaffold	GCF_902806735.1
Myodes rufocanus	*Myodes glareolus*	Genus	Scaffold	GCF_902806735.1
Myotis bechsteinii	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis blythii	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis bocagii	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis brandtii	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis capaccinii	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis crypticus	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis daubentonii	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis emarginatus	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis escalerai	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myotis mystacinus	*Myotis myotis*	Genus	Scaffold	GCF_014108235.1
Myrmelastes hyperythrus	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Myrmoborus myotherinus	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Nyctalus noctula	*Pipistrellus pipistrellus*	Family	Chromosome	GCA_903992545.1
Oneillornis salvini	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1

Paroaria gularis	*Diglossa brunneiventris*	Family	Chromosome	GCA_019023105.1
Parus major	*Parus major*	Species	Chromosome	GCA_001522545.3
Passer montanus	*Passer montanus*	Species	Scaffold	GCA_014805655.1
Percnostola fortis	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Perisoreus infaustus	*Aphelocoma coerulescens*	Family	Scaffold	GCA_013398375.1
Peromyscus maniculatus	*Peromyscus maniculatus*	Species	Chromosome	GCA_003704035.3
Petrogale xanthopus	*Macropus giganteus*	Family	Chromosome	GCA_028627215.1
Phlegopsis nigromaculata	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Phoenicurus ochruros	*Oenanthe oenanthe*	Family	Scaffold	GCA_013399995.1
Pipra fasciicauda	*Pipra filicauda*	Genus	Scaffold	GCA_003945595.2
Platyrinchus coronatus	*Tyrannus tyrannus*	Family	Scaffold	GCA_026770345.1
Plecotus auritus	*Plecotus auritus*	Species	Contig	Unpublished
Plecotus austriacus	*Plecotus auritus*	Genus	Contig	Unpublished
Podarcis filfolensis	*Podarcis muralis*	Genus	Chromosome	GCF_004329235.1
Podarcis gaigeae	*Podarcis muralis*	Genus	Chromosome	GCF_004329235.1
Podarcis liolepis	*Podarcis muralis*	Genus	Chromosome	GCF_004329235.1
Podarcis muralis	*Podarcis muralis*	Species	Chromosome	GCF_004329235.1
Podarcis pityusensis	*Podarcis muralis*	Genus	Chromosome	GCF_004329235.1
Progne tapera	*Progne subis*	Genus	Scaffold	GCA_022316685.1
Psammodromus manuelae	*Zootoca vivipara*	Family	Chromosome	GCF_011800845.1
Pseudoromicia sp.	*Pipistrellus pipistrellus*	Family	Chromosome	GCA_903992545.1
Psittacula echo	*Psittacula krameri*	Genus	Scaffold	GCA_002870145.1
Ramphocelus carbo	*Diglossa brunneiventris*	Family	Chromosome	GCA_019023105.1
Rhinolophus alcyone	*Rhinolophus ferrumequinum*	Genus	Chromosome	GCA_004115265.3
Rhinolophus euryale	*Rhinolophus ferrumequinum*	Genus	Chromosome	GCA_004115265.3
Rhinolophus ferrumequinum	*Rhinolophus ferrumequinum*	Species	Chromosome	GCA_004115265.3
Rhinolophus hipposideros	*Rhinolophus ferrumequinum*	Genus	Chromosome	GCA_004115265.3
Rhinolophus landeri	*Rhinolophus ferrumequinum*	Genus	Chromosome	GCA_004115265.3
Salamandra atra	*Pleurodeles waltl*	Family	Chromosome	GCA_026652325.1
Sceloporus aeneus	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus bicanthalis	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus grammicus	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus horridus	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus mucronatus	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus spinosus	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus subniger	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sceloporus torquatus	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1

Sceloporus variabilis	*Sceloporus tristichus*	Genus	Chromosome	GCA_016801065.1
Sciurus carolinensis	*Sciurus carolinensis*	Species	Chromosome	GCA_902686445.2
Sciurus vulgaris	*Sciurus vulgaris*	Species	Chromosome	GCA_902686455.2
Sclerurus caudacutus	*Sclerurus mexicanus*	Genus	Scaffold	GCA_013396755.1
Stelgidopteryx ruficollis	*Hirundo rustica*	Family	Chromosome	GCA_015227805.2
Strix aluco	*Strix occidentalis*	Genus	Scaffold	GCA_002372975.2
Sylvia atricapilla	*Sylvia atricapilla*	Species	Chromosome	GCA_009819655.1
Tachycineta albiventer	*Tachycineta bicolor*	Genus	Scaffold	GCA_025960845.1
Taraba major	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Tarentola mauritanica	*Gekko gecko*	Family	Scaffold	GCA_029375565.1
Thamnomanes schistogynus	*Sakesphorus luctuosus*	Family	Scaffold	GCA_013396695.1
Timon lepidus	*Lacerta viridis*	Family	Contig	GCA_900245905.1
Trichosurus vulpecula	*Trichosurus vulpecula*	Species	Chromosome	GCA_011100635.1
Turdus albicollis	*Turdus rufiventris*	Genus	Scaffold	GCA_013186435.1
Tyrannus melancholicus	*Tyrannus savana*	Genus	Scaffold	GCA_013399735.1
Xenops minutus	*Furnarius figulus*	Family	Scaffold	GCA_013397465.1
Xiphorhynchus elegans	*Xiphorhynchus elegans*	Species	Scaffold	GCA_013401175.1
Zosterops virens	*Zosterops hypoxanthus*	Genus	Scaffold	GCA_013399795.1

C. Research and management strategy of the EHI

The EHI is conceived as a methodological, logistic, and social infrastructure to harbour diverse projects based on unified data generation and knowledge sharing criteria. The initiative implements a flexible research strategy within which projects with different study designs, scientific goals, and spatio-temporal scales are accommodated (Fig. S1). The ultimate operational aim is to provide the means to scale up the data generated for a given study (e.g., small-scale research based on a single animal species) to larger studies (e.g., large-scale study spanning multiple animal species) addressing broader scientific questions, thus fostering collaboration while overcoming issues related to technical biases and key metadata availability.

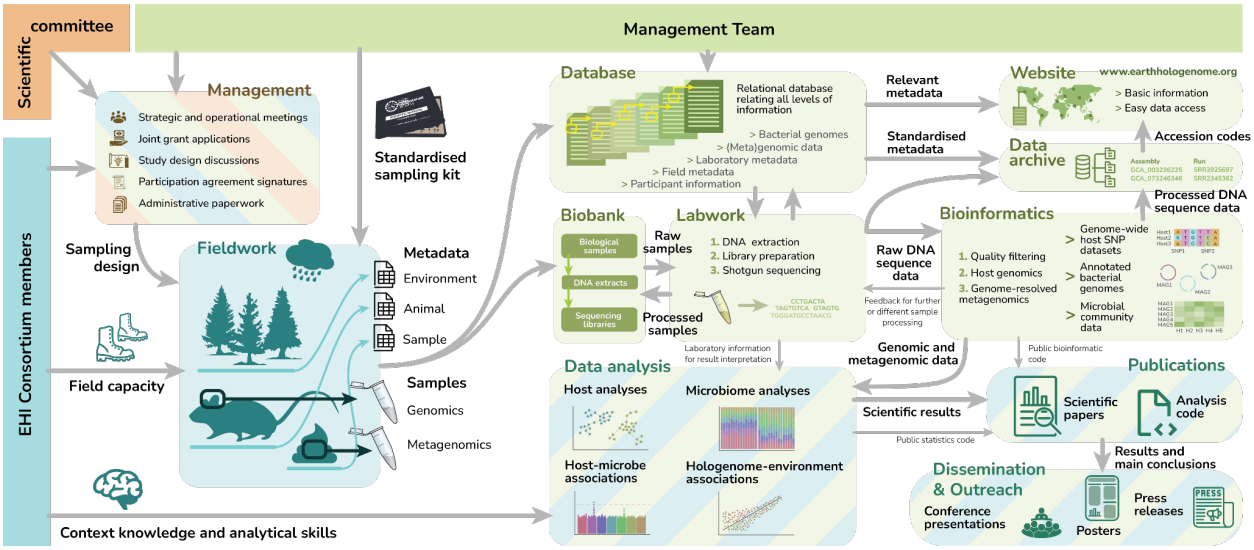

Figure S1. **Overview of the research and management strategy of the Earth Hologenome Initiative**. The activities are conducted by one or multiple participating bodies, including the Management Team, the EHI Consortium Members, and the Scientific Committee. The three bodies are involved in different actions regarding core EHI management. The field work is the responsibility of the consortium members or participants, while data generation and management are responsibilities of the management team. Data analysis, publication, dissemination and outreach are conducted as a collaboration between the management team and the consortium members. Detailed explanations of each activity are provided in Supporting Information C.

Governance and management

The governance model of the Earth Hologenome Initiative is built upon three bodies. The EHI Consortium is the main governance body composed of all researchers with an active EHI participation agreement. The EHI consortium meets twice a year to discuss the development of the initiative and key decisions. The EHI management team, consisting of researchers affiliated to the coordinating institution (Globe Institute, University of Copenhagen), is responsible for the overall management of the initiative, providing the basis for administrative, research and logistic capacities. The Scientific Committee is the advisory group that meets once a year to discuss the

strategic goals of the initiative, which consists of 10 international researchers with diverse backgrounds, expertise and career development stages.

Participation, administration and funding

Participation in the EHI is open to all researchers interested in generating and analysing hologenomic data in wild animals, and more than 230 researchers from around the world have already joined the initiative (Fig. S2). Participation can happen through sample collection, data analysis, or both. The initial expression of interest, which must be sent through the sign-up form available on the EHI website (www.earthhologenome.org/signup), is followed by the signature of the EHI participation agreement, which establishes the legal framework between the participating individual or institution and the coordinating institution. This agreement outlines the samples expected to be collected, establishes the legal responsibilities of each party for sample collection, export and import, and sets the terms of expense coverage, which can be shared or fully covered by the EHI management, depending on the available resources and strategic goals.

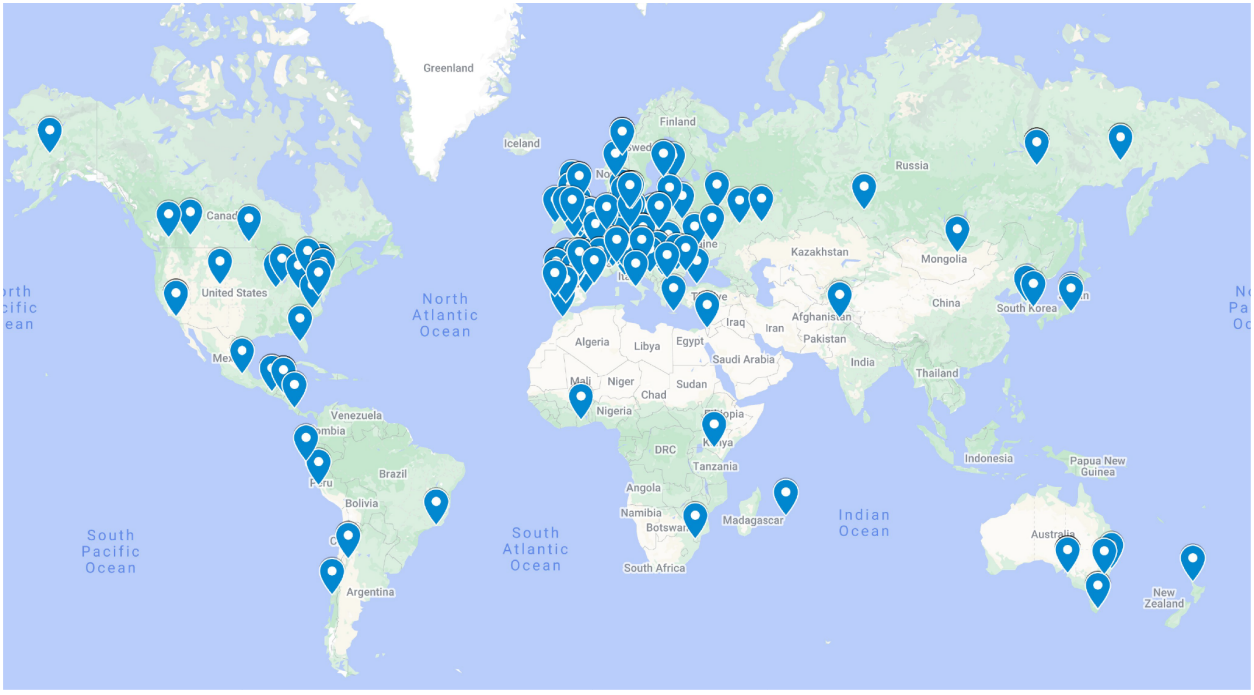

Figure S2. **Geographic locations of EHI participants**.

Fieldwork, sample collection and preservation

Participants provide sampling expectations to the management team, who prepare fieldwork kits that include sample collection tubes prefilled with the preservation buffer and pre-labeled with EHI codes and barcodes. Kits are shipped to the collaborator, who performs the field work following standardised guidelines. EHI sample collection guidelines aim at minimising environmental contamination that can jeopardise shotgun sequencing-based analyses. The standard sampling effort includes collection of biological material that enables characterisation

of the lower intestinal microbiome (e.g., faeces, anal/cloacal swabs, gut contents), as well as host genomic analysis (e.g., blood, muscle, oral swab). The specific selection of the sample type depends on the biological features of the animals (e.g., whether animals defecate when captured), the administrative/ethical limitations of sample collection (e.g., whether animals can be manipulated or euthanised), and the aims of the study. Complementary samples such as skin or gland swabs can also be collected as required by the aims of the projects. Samples are then shipped to the Globe Institute, where they are biobanked and processed for data generation.

Metadata collection and archiving

Sample collection is paired with the gathering of standardised metadata. Field metadata is collected at three levels, covering sampling events (e.g., temporal, geographical and environmental attributes), animal (e.g., species, age, sex, biometrics) and sample (e.g., sample type, preservation conditions) information. Such multi-layered information enables host genomic and microbiota variation to be associated with host and environmental attributes, while accounting for biases introduced by sample types and preservation strategies. Standardised metadata is stored in the central EHI database, and partially displayed (e.g., resolution of geographic coordinates is reduced for conservation reasons) in real time on the EHI website (www.earthhologenome.org).

Laboratory procedures

Samples are processed following the standardised open-access EHI laboratory workflow available at www.earthhologenome/laboratory. The standard protocol, which enables liquid handler automation, consists of i) chemical digestion (lysis buffers) and mechanical cell lysis (bead-beating) to release as much DNA as possible from eukaryotic and different kinds of prokaryotic cells, ii) DNA isolation using silica magnetic beads combined with solid-phase reversible immobilisation to remove as many inhibitors as possible, iii) DNA shearing using either sonication of enzymatic fragmentation to adjust DNA fragment-sizes to efficient short-read sequencing, iv) adapter ligation-based sequencing library preparation to deal with total DNA, v) library amplification with dual index identifiers with adjusted number of PCR cycles for each library to ensure highest quality libraries and vi) library pooling for multiplexed short-read sequencing. All libraries are initially screened at a sequencing depth of ca. 5GB (around 16.6 million reads), to later resequence them, and/or complement them with long-read sequencing, based on the data generation needs informed by bioinformatic analyses.

Bioinformatic procedures

Bioinformatic data processing is conducted using an open-access pipeline available at www.earthhologenome/bioinformatics. The pipeline is based on *snakemake* (for workflow management), *conda* (for environment management) and *slurm* (for computational job management), and directly managed from the EHI database. The pipeline first quality-filters raw data before splitting the metagenomic from the host genomic fraction through mapping reads against a reference host genome. The genomic fraction is outputted as a compressed mapping

file, while the metagenomic fraction undergoes the genome-resolved metagenomic pipeline consisting of (co)assembly, binning, annotation and quantitation. All relevant raw, intermediate and final data sets are locally stored in the Electronic Research Data Archive (ERDA) of the University of Copenhagen, and uploaded to the European Nucleotide Archive (ENA) for long-term storage and data sharing.

Data analysis

Data analysis is performed in collaboration between the participants and the EHI management team, with the possibility to incorporate external collaborators that bring complementary expertise. Typical analyses include population structure and genomic diversity analyses using genome-wide SNP datasets, taxonomic, phylogenetic and functional diversity of microbiomes based on community-level annotated bacterial genome data, spatial and/or temporal modelling of microbiome attributes, correlative modelling of host genome and microbiome features with environmental variables, and association modelling between host genomic and microbiome features, to cite but a few.

Data access, ownership rights and publications

Raw data, host genomic mapping files, pre-processed metagenomic files without host data, metagenomic assemblies and metagenome-assembled genomes (MAGs) are uploaded to ENA along with all relevant standardised metadata [6], complying with international data standards [7,8] and FAIR principles [9]. The data is shared under temporary usage limitations that aim at properly crediting the effort conducted for its generation. Once a dedicated article is published, the data are released from restrictions. Availability and status of biological samples, which belong to the EHI participants involved in collecting and processing them, is displayed in the EHI website to enable alternative uses after agreement with sample owners. Publications derived from data analysis acknowledge all researchers who made a significant contribution with co-authorships, and free access to all EHI-derived publications is ensured at least through archiving of an institutional-hosted post-peer-review version of every publication.

D. Detailed overview of the 10-year roadmap

The Earth Hologenome Initiative is an ongoing endeavour delivered over three phases with distinct priorities, goals, and outcomes (Fig. S3).

Development

Phase 1: development (2020 – 2023)

The Earth Hologenome Initiative (EHI) was conceptualised in 2020 as a pivotal endeavour within the framework of the Center for Evolutionary Hologenomics (CEH), Copenhagen, Denmark. Over the initial three-year span, the EHI focused on establishing the foundational management, logistical, and scientific infrastructure required to orchestrate a global-scale initiative for hologenomic data generation and analysis. In order to avoid excessive complexity, in this initial phase, the taxonomic breadth of host species has been limited to terrestrial vertebrates. Presently, the EHI boasts an internal relational database housing more than 150,000 data entries, a publicly accessible website (www.earthhologenome.org) offering fundamental information alongside a subset of the stored data (www.earthhologenome.org/database), optimally streamlined pipelines for managing laboratory samples (www.earthhologenome.org/laboratory), and bioinformatics processes (www.earthhologenome.org/bioinformatics), in addition to an ensemble of R scripts designed for data visualisation, analysis, and modelling (www.earthhologenome.org/analysis).

As of Q4 of 2023, more than 230 research groups from over 70 countries have conveyed an interest in participating, resulting in the signing of more than 50 participation agreements. Distribution of 170 standardised EHI sampling kits among participants has yielded a cumulative submission of over 7,700 samples representing 175 distinct vertebrate species across the world. Notably, more than 1,500 of these samples, derived from 115 different host species, have already undergone sequencing, culminating in the generation of 9.5 terabases (TB) – equivalent to over 31.6 billion reads or > 3,000 human genomes – of hologenomic data. Roughly half of this data were mapped to the reference genomes of the animal hosts or, when unavailable, the genomes of their closest relatives. Reference genomes were only available for 19% of the analysed species, which highlights the need to coordinate information exchange with reference genome-generating consortia [10], to ensure availability of high-quality host genomic data required for population genomic and association analyses. The other half of the generated data, namely the metagenomic fraction, yielded nearly 30,000 draft bacterial genomes. Less than 15% of these genomes were annotated at the species level, underscoring the fact that wild vertebrates harbour an enormous quantity of bacteria heretofore undiscovered by science.

Phase 2: consolidation (2024 – 2025)

Through the first half of 2023, the EHI achieved an average production of 130 datasets per month. Consequently, by 2024, the EHI is projected to have amassed a volume of data sufficient for addressing an initial set of inquiries regarding technical aspects of hologenomic

data generation in wild animals, and diverse biological questions specific to studied taxa. This phase will witness the dissemination of expertise in hologenomic data generation and analysis for various sample types and vertebrate taxa, with the aim of providing invaluable technical insights to the wider community. We will also expand the network to data scientists, who will contribute to make the most of the generated data, and we will explore the feasibility to expand the taxonomic breadth of the initiative beyond terrestrial vertebrates. During this period, biological studies will focus on one or a few focal taxa, adhering to study-specific designs. Notably, more than ten such studies are already active at various stages of data generation and analysis, with expected publication within the aforementioned timeframe (Box 1). Among the array of topics anticipated for exploration during this phase are, for instance, hologenomic diversity loss across island sizes in rodents and reptiles, responses to climate shifts in bats, and hologenomic responses to active conservation actions in newts (www.earthhologenome.org/projects). Throughout the consolidation phase, the EHI will remain committed to refining and updating procedures in line with the latest technological advancements. Additionally, the establishment of new collaborations and the promotion of joint research initiatives within the EHI's overarching framework are also anticipated outcomes.

Box S1. **Overview of some of the ongoing Phase 2 projects within the Earth Hologenome Initiative**. The four showcased studies encompass mammals, reptiles, amphibians and birds, focus on host species in North America, Europe, and Japan, and involve 19 institutions from 13 countries. More detailed descriptions of these and other projects can be found at www.earthhologenome.org/projects.

Invasion hologenomics
Through characterising the host genomic and microbiome variation across their native and allochthonous ranges, this study aims to understand the ecological interactions between the Eurasian red squirrel (*Sciurus vulgaris*) and the Eastern grey squirrel (*S. carolinensis*).
> *Involved institutions*: *University of Galway, Bangor University, Bournemouth University, University of Insubria, CIBIO-InBIO, IZW Berlin, The Graduate University for Advanced Studies - SOKENDAI, Scottish Wildlife Trust, University of Calgary, Duke University, Wilkes University, University of Copenhagen.*

Elevational adaptation hologenomics
This study aims at identifying joint animal genomic and microbial metagenomic signatures of adaptation in wall lizards (*Podarcis muralis*) across 4 altitudinal transects conducted in the Pyrenees (Spain and France).
> *Involved institutions*: Lund University, University of Copenhagen, SETE-CNRS, University of Valencia.

Amphibian evolutionary hologenomics
In this study, we are analysing hologenomic features of fire-bellied and yellow-bellied toads (*Bombina bombina* and *B. variegata*), their hybrids, and how their evolutionary pathways are affected by chytrid fungus, a highly deadly disease for amphibians.
> ***Involved institutions***: *Croatian Institute for Biodiversity, Hungarian National History Museum, University of Copenhagen*.

Scavenger hologenomics

Aimed at understanding the contribution of intestinal microorganisms to scavenging, in this study we are studying microbial communities in the stomach and colon of turkey vultures (*Cathartes aura*) and black vultures (*Coragyps atratus*) in southern North America.
> *Involved institutions*: *Smithsonian Museum of Natural History, University of Copenhagen, The Pennsylvania State University*.

Phase 3: integration (2026 – 2029)

Subsequent to the consolidation phase, a continued focus on addressing host species-specific inquiries is anticipated. However, the advanced phase will entail the utilisation or reutilisation of hologenomic data derived from diverse individuals, populations, and species, with the goal of unravelling global patterns and addressing overarching biological questions at the forefront of science. Notably, the comprehensive characterisation of thousands of microbial strains and communities across animals inhabiting various continents, all executed through standardised protocols and enriched with uniform metadata, will facilitate investigations into matters such as microbial diversification in light of host evolution, and diffusion of antimicrobial resistance shaped by host ecology. In this pursuit, ongoing collaborations with reference genome-generating consortia, including B10K [11] and Bat1K [12], will also contribute to hologenomic analyses aimed at unveiling whether and how microorganisms have shaped animal evolutionary processes, such as dietary diversification or adaptive radiation.

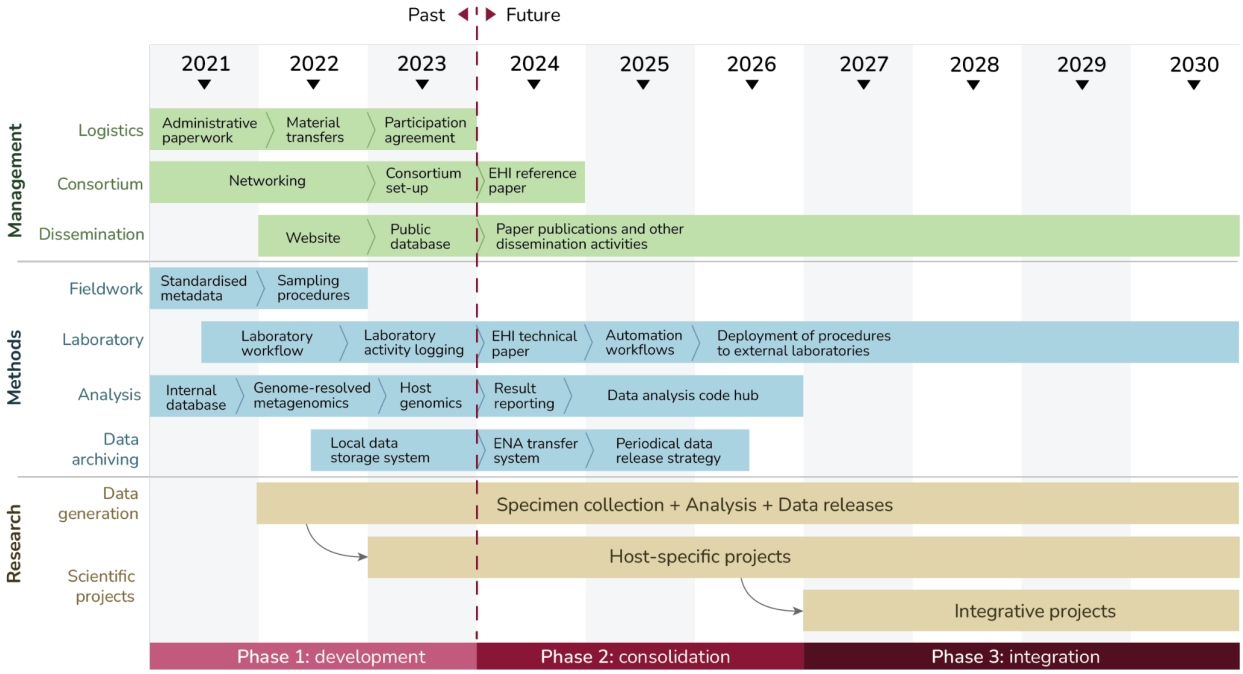

Figure S3. **Gantt-chart of the 10-year roadmap.** The approximate temporal assignment of multiple management, methods development, and research tasks is divided into three phases with distinct priorities, goals and outcomes.

Management

Logistics

Administrative paperwork: preparation of the procedures regarding obtaining sampling licences, export and import permits, and other paperwork related to the collection, transfer and archiving of biological samples.
Material transfers: preparation of logistics behind the transfer of biological materials, including courier options, strategies to maintain cold chain, and other aspects related with sample transfer without quality loss.
Participation agreement: preparation of the legal document to be signed by the legal representatives of the EHI Management Team (Head of the Globe Institute at the University of Copenhagen, Denmark) and the participating researcher or institution, which specifies the terms of the research collaboration within the EHI.

Consortium

Networking: global communication activities to raise awareness of the aims and procedures of the EHI to build up a diverse and multidisciplinary consortium of researchers from all around the world.
Consortium set-up: collaborative effort to define the structure and procedures of the EHI.
EHI reference paper: peer-reviewed manuscript defining the goals of the EHI, its research and management strategy, and the roadmap for the coming years.

Dissemination

Website: design of the public database www.earthhologenome.org, containing the essential information about the EHI.
Public database: design and development of the public database displayed on the website (www.earthhologenome.org/database) containing detailed information about data generation.
Paper publications: scientific articles containing the main outputs of the hologenomic research conducted on EHI data.
Other dissemination activities: participation in conferences, public outreach activities and other actions aimed at disseminating the knowledge generated within the EHI.

Methods

Fieldwork

Standardised metadata: define standard metadata for sampling events, animal individuals and samples based on existing standard checklists and ontologies, and design field and electronic metadata sheets.
Sampling procedures: describe overall sampling procedures to ensure samples are collected and preserved in the best way possible.

Laboratory

Laboratory workflow: standardised pipeline for DNA extraction, DNA shearing, and library preparation using cost-effective procedures.
Laboratory activity logging: procedures to ensure every step conducted in the laboratory and the location of each intermediate sample (DNA extracts, libraries) is logged in the internal database.
EHI technical paper: manuscript based on the initial 1,000 samples processed in the EHI, delving into the challenges of sample processing and high-quality data generation from diverse sample types.
Automation workflows: procedures to scale-up sample processing by automatising the laboratory procedures using custom programs for Opentrons OT2 liquid-handling workstations.
Deployment of procedures to external laboratories: adjustment of laboratory procedures and benchmarking to ensure EHI data can be generated in other laboratories without compromising comparability of results.

Analysis

Internal database: relational database interlinking all management information of the EHI, including collaborators, sampling events, samples, laboratory data, and bioinformatic outputs.
Genome-resolved metagenomics: bioinformatic pipeline to reconstruct and annotate draft bacterial genomes from metagenomic data.
Host genomics: bioinformatic pipeline to generate whole-genome SNP data.
Result reporting: automatised procedure to generate a summary report with the most relevant information about sample origin, quantity and quality, and representativeness of the results.
Data analysis code hub: Github organisational repository containing a suite of useful exploratory, modelling and visualisation R scripts for processing EHI data.

Data archiving

Local data storage system: design of the workgroup space structure at the University of Copenhagen's Electronic Research Data Archive, to store raw, intermediate and final data generated and processed within the EHI in an interoperable fashion.
ENA transfer system: development of procedures to transfer raw data, preprocessed data, metagenomic assemblies and metagenome-assembled genomes (MAGs) to the European Nucleotide Archive (ENA).
Periodical data release strategy: design of the strategy and procedures to archive EHI data in public repositories and release for third-party usage.

Research

Data generation

Sample collection, preservation, biobanking and laboratory processing using standardised procedures, followed by bioinformatic data analyses to generate research outputs that will be stored in public databases in periodical data releases.

Scientific projects

Development of scientific projects adhering to EHI standards, beginning with small-to-medium scale species-specific studies in Phase 2, and transitioning to larger integrative projects relying on data generated in multiple projects in Phase 3.

E. Supplementary references

1. R Core Team, R. and Others (2013) R: A language and environment for statistical computing. at <https://apps.dtic.mil/sti/citations/AD1039033>
2. Wickham, H. *et al.* (2019) Welcome to the tidyverse. *J. Open Source Softw.* 4, 1686
3. Chamberlain, S.A. and Szöcs, E. (2013) taxize: taxonomic search and retrieval in R. *F1000Res.* 2, 191
4. Wickham, H. and Wickham, M.H. (2023) Package “httr”[Online]. Available: http://cran.ma.imperial.ac.uk/web/packages/httr/httr.pdf. [Accessed: 10-Sep-2023]
5. Wickham, H. *et al.* (2023) Package “readr.” */packages/readr …* at <http://cran.uni-muenster.de/web/packages/readr/readr.pdf>
6. Burgin, J. *et al.* (2023) The European Nucleotide Archive in 2022. *Nucleic Acids Res.* 51, D121–D125
7. Ten Hoopen, P. *et al.* (2017) The metagenomic data life-cycle: standards and best practices. *Gigascience* 6, 1–11
8. Bowers, R.M. *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731
9. Wilkinson, M.D. *et al.* (2019) Addendum: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 6, 6
10. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
11. Jarvis, E.D. *et al.* (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346, 1320–1331
12. Teeling, E.C. *et al.* (2018) Bat biology, genomes, and the Bat1K project: to generate chromosome-level genomes for all living bat species. *Annu. Rev. Anim. Biosci.* 6, 23–46