frontiers | Frontiers in Science

TYPE Frontiers in Science Lead Article
PUBLISHED 04 September 2025
DOI 10.3389/fsci.2025.1514835



OPEN ACCESS

EDITED BY
Mehrdad Hajibabaei,
University of Guelph, Canada

REVIEWED BY
Gane Ka-Shu Wong,
University of Alberta, Canada
Terry Burke,
The University of Sheffield, United Kingdom

*CORRESPONDENCE
Mark Blaxter
mb35@sanger.ac.uk
Harris A. Lewin
Harris.Lewin@asu.edu

RECEIVED 21 October 2024
ACCEPTED 30 June 2025
PUBLISHED 04 September 2025

CITATION
Blaxter M, Lewin HA, DiPalma F, Challis R, da Silva M, Durbin R, Formenti G, Franz N, Guigo R, Harrison PW, Hiller M, Hoff KJ, Howe K, Jarvis ED, Lawniczak MKN, Lindblad-Toh K, Mathews DJH, Martin FJ, Mazzoni CJ, McCartney AM, Mulder N, Paez S, Pruitt KD, Ras V, Ryder OA, Shirley L, Thibaud-Nissen F, Warnow T, Waterhouse RM and the EBP Community of Scientists. The Earth BioGenome Project Phase II: illuminating the eukaryotic tree of life. *Front Sci* (2025) 3:1514835. doi: 10.3389/fsci.2025.1514835



# The Earth BioGenome Project Phase II: illuminating the eukaryotic tree of life

Mark Blaxter[1]*, Harris A. Lewin[2,3]*, Federica DiPalma[4,5], Richard Challis[1], Manuela da Silva[6], Richard Durbin[7,1], Giulio Formenti[8], Nico Franz[9], Roderic Guigo[10,11,12], Peter W. Harrison[13], Michael Hiller[14,15,16], Katharina J. Hoff[17,18], Kerstin Howe[1], Erich D. Jarvis[19,20], Mara K. N. Lawniczak[1], Kerstin Lindblad-Toh[21,22,23], Debra J. H. Mathews[24,25], Fergal J. Martin[13], Camila J. Mazzoni[26,27], Ann M. McCartney[28], Nicola Mulder[29], Sadye Paez[19], Kim D. Pruitt[30], Verena Ras[29,31], Oliver A. Ryder[32,33], Lesley Shirley[1], Françoise Thibaud-Nissen[30], Tandy Warnow[34], Robert M. Waterhouse[35,36] and the EBP Community of Scientists[†]

[1]Tree of Life, Wellcome Sanger Institute, Cambridge, United Kingdom, [2]Global Futures Laboratory, Walton Center for Planetary Health, Arizona State University, Tempe, AZ, United States, [3]Genome Center, Department of Evolution and Ecology, University of California, Davis, Davis, CA, United States, [4]Research and Innovation, Genome British Columbia, Vancouver, BC, Canada, [5]Department of Biological Sciences, University of East Anglia, Norwich, United Kingdom, [6]Fiocruz Biodiversity and Health Biobank, Oswaldo Cruz Foundation–Fiocruz, Rio de Janeiro, Brazil, [7]Department of Genetics, University of Cambridge, Cambridge, United Kingdom, [8]The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, United States, [9]School of Life Sciences, Arizona State University, Tempe, AZ, United States, [10]Computational Biology and Health Genomics, Centre for Genomic Regulation (CRG), Barcelona, Spain, [11]Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain, [12]Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, [13]European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL), Hinxton, United Kingdom, [14]LOEWE Centre for Translational Biodiversity Genomics, Frankfurt, Germany, [15]Senckenberg Research Institute, Frankfurt, Germany, [16]Institute of Cell Biology and Neuroscience, Faculty of Biosciences, Goethe University, Frankfurt, Germany, [17]Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany, [18]Center for Functional Genomics of Microbes, University of Greifswald, Greifswald, Germany, [19]Neurogenetics of Language, The Rockefeller University, New York, NY, United States, [20]Howard Hughes Medical Institute, Chevy Chase, MD, United States, [21]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, [22]SciLifeLab, Uppsala University, Uppsala, Sweden, [23]Broad Institute of MIT and Harvard, Cambridge, MA, United States, [24]Berman Institute of Bioethics, Johns Hopkins University, Baltimore, MD, United States, [25]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Johns Hopkins University, Baltimore, MD, United States, [26]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany, [27]Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany, [28]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, United States, [29]Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [30]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD, United States, [31]Department of Biodiversity and Conservation Biology, University of the Western Cape, Bellville, South Africa, [32]Conservation Science, San Diego Zoo Wildlife Alliance, Escondido, CA, United States, [33]Department of Ecology, Behavior and Evolution, School of Biological Sciences, University of California, San Diego, San Diego, CA, United States, [34]Department of Computer Science, University of Illinois, Urbana-Champaign, Urbana, IL, United States, [35]Environmental Bioinformatics, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland, [36]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[†]The EBP Community of Scientists is composed of the individuals who worked through the Earth BioGenome Project (EBP) International Scientific Committees in the lead-up to and follow-up from a workshop held in Lausanne, Switzerland, in May 2023 that informed the development of this article (see Conflict of interest statement).

10.3389/fsci.2025.1514835

## Abstract

The Earth BioGenome Project (EBP) aims to "sequence life for the future of life" by generating high-quality reference genome sequences for all recognized eukaryotic species, thereby building a rich knowledge base to inform conservation, inspire bioindustry, ensure food security, advance medicine, and establish a deeper understanding of biodiversity. As the EBP works toward completing the original Phase I goal—a reference genome for each of the approximately 10,000 taxonomic families of eukaryotes—milestone publications have demonstrated the transformative potential of the project. The EBP has promoted global collaboration and established core methods and standards. By the end of 2024, EBP-affiliated projects had publicly released 2,000 high-quality genome assemblies, representing more than 500 eukaryotic families. In this article, we present a revised set of goals for Phases I and II of the EBP. For Phase II, we propose generating reference genomes for 150,000 species over 4 years, including representative genomes for at least 50% of all accepted genera and for additional species of biological and economic importance. To deliver Phase II, EBP-affiliated projects will have to release over 3,000 new genomes per month. We review the magnitude of the tasks in sourcing, sequencing, assembling, annotating, and analyzing genomes at this scale, and explore the scientific, technical, social, legal, ethical, and funding challenges associated with them. Success in Phase II will set the stage for sequencing the remaining ~1.5 million named species of Eukaryota and establishing the knowledge platforms necessary for understanding, preserving, and utilizing Earth's biodiversity in an era of rapid environmental change.

KEYWORDS

**biodiversity, conservation, evolution, genomics, DNA sequencing, annotation**

## Key points

- The ongoing success of Phase I of the Earth Biogenome Project (EBP) demonstrates the feasibility of producing reference-quality genomes at scale, enabling the project to achieve its overarching goal: to sequence 1.67 million eukaryotic species in 10 years.
- Using knowledge from Phase I projects, we propose a revised strategy for Phase II: collecting specimens for 300,000 species and sequencing 150,000 species, representing at least half of the eukaryotic genera, in 4 years.
- Technical advances in DNA sequencing, genome assembly, and genome annotation have reduced costs and increased throughput to the point that we envisage globally distributed production of reference-quality genomes for most eukaryotic species for a total cost of about US$3.9 billion—US$800 million less than initially envisioned.
- Key challenges remain, including enhancing global coordination and building communities of users and interested parties; creating an inclusive, global biodiversity genomics workforce; developing effective access and benefit-sharing methodologies; facilitating collection at scale of vouchered specimens; sequencing reference genomes from single-celled and very small organisms; enhancing functional annotation; and building large-scale toolkits for comparative genomics.
- Technological and operational innovations, such as a "sequencing lab in a box," have the potential to radically transform the global capacity for biodiversity genome sequencing, facilitating national benefit-sharing agreements and the realization of societal impacts on Indigenous peoples and local communities.
- We propose the establishment of a US$0.5 billion Foundational Impact Project (FIF) fund to support the immediate use of the genome sequences in conservation, agriculture, biodiversity monitoring, biotechnology, and basic sciences, focused on supporting initiatives in the Global South.

## The Earth BioGenome Project: past, present, and a new future

The Earth BioGenome Project (EBP; see https://www.earthbiogenome.org/) proposed a visionary goal: to sequence all named eukaryotic species in 10 years (1). This science "moonshot" is critical for future planetary and human health as it will transform our biological understanding of all life (2). The EBP has evolved as a network of networks that collectively engages local and global scientific, stakeholder, and public communities to generate a shared genomic resource to advance biodiversity science, underpin essential conservation efforts, and build a more equitable global bioeconomy (1, 3). The EBP originally planned to deliver this revolutionary change in three growing phases over 10 years. A completed 4-year pilot phase has built core methodologies, created standards, and established an ethical framework. In Phase I, which began in 2021, we proposed generating a high-quality reference genome sequence for most of the approximately 10,000 living eukaryotic families (3, 4). Here, we summarize progress in delivering Phase I goals and present a new vision for Phase II.

The EBP has formalized the organizational core of the project, helped to establish and recruit affiliated projects, created open governance principles, empowered committees to advise the project on technical and ethical standards, and planned workflows needed to produce reference-quality genome sequences at scale (5–7). Parallel, rapid advances in single-molecule, long-read, and high-throughput chromatin conformation capture (Hi-C) sequence data generation, as well as in the informatics of genome assembly, have made the production of high-quality, chromosome-scale assemblies much more achievable. In Phase I so far, EBP affiliates have delivered genomes at an inclusive average cost of US$28,000 per species assembled and demonstrated that high-quality genome assemblies can be generated from a wide diversity of taxa. These new genome sequences have been used to shine new light on fundamental and applied biological questions (8–15). Several large project consortia with strong buy-in from biodiversity, genomics, and end-user groups have been funded and started production, including the Vertebrate Genomes Project (VGP) (16), Bat1K (17), the Darwin Tree of Life Project (18), the African BioGenome Project (19), the Aquatic Symbiosis Genomics Project (20), the Norwegian Earth BioGenome Project (https://www.ebpnor.org/), the Catalan Initiative for the Earth BioGenome Project (21), the Canada BioGenome Project (http://earthbiogenome.ca/), the California Conservation Genomics Project (22), and the European Reference Genome Atlas (23) (see https://earthbiogenome.org/affiliated-project-networks). A live summary of EBP progress is maintained on Genomes on a Tree (GoaT; https://goat.genomehubs.org/projects/EBP) (24), an Elastic-search-driven data system organized against a taxonomic tree of all life from the United States National Center for Biotechnology Information (NCBI) Taxonomy database (25).

The original plan for EBP Phase I was to deliver approximately 10,000 genome sequences, one for each eukaryotic family, over a 3-year period (an average of 300 genomes/month). As of September 14, 2024, EBP-affiliated projects had generated 1,667 high-quality genome sequences from fungi, plants, animals, and diverse protists, that met the minimum EBP reference genome metrics (generally 1 Mb contig N50, chromosome-scale scaffolds for all chromosomes with >95% of all sequence in chromosomes, and a base call error rate of less than 1/10,000, summarized as "6.C.Q40"; see https://www.earthbiogenome.org/report-on-assembly-standards). Other researchers deposited 1,798 EBP-quality genomes in International Nucleotide Sequence Database Collaboration (INSDC) databases (Figure 1A). Global production by EBP-affiliated projects was approximately 50 genomes/month in 2023 (see https://tinyurl.com/EBP-by-month-2023-in-GoaT) (Figure 1B). This was double the output in 2022 but is still insufficient to complete the Phase I goal in 3 years.

Our experiences in Phase I have revealed both strengths and limitations in our original strategy. There are still challenges to overcome to complete the goals of Phase I, which will be amplified in Phase II. The more ambitious EBP-affiliated projects have shown that genome sequencing across diversity can be achieved at scale by optimizing all aspects of the sequencing process, from sampling to assembly curation (6, 16, 26, 27). Individual advances are small, but they sum to a significant step-change in genome production. What is clear is that, given funding, EBP-quality reference genomes can be produced at scale, regardless of whether the target is a protist, a fungus, an animal, or a plant (see the Darwin Tree of Life Genome Notes collection for examples of successful chromosomal assembly of specimens of all these taxa; https://wellcomeopenresearch.org/gateways/treeoflife). The main factor limiting reference genome production has been funding, although more than US$200 million has already been raised. Knowledge gained to date, coupled with rapid advances and cost reductions in DNA sequencing, have led us to revise the staging and quality goals of future EBP strategy (Figure 1C). In a revised strategy for Phase I, we will sample from all phyla and from at least 50% of families. In Phase II we propose sequencing, to reference quality, 150,000 additional species, down from 180,000 species. We recognize that collecting strictly to a species list is inefficient, and specimens for many genera, such as deep-sea taxa, may be too cost-prohibitive to acquire. Instead of sequencing one species per genus, as originally proposed, we will sequence representative genomes for at least 50% of genera (80,000 species). We will also prioritize sequencing species of importance to ecosystem health, food security, pandemic control, conservation, and Indigenous peoples and local communities. Importantly, we propose that sequencing to reference quality, rather than draft, should be our goal. In completing Phase II, we will have sequenced about one-tenth of the Earth's known eukaryotic biodiversity.

While many challenges and blockers to completion of Phase I overlap with those of Phase II, the scaling required to sequence 150,000 genomes in 4 years presents unique scientific and social challenges. While Phase II remains challenging, we are optimistic that our goals are achievable and that the data will be transformative.

## A new EBP Phase II strategy

The EBP is a progressive project, with overlapping rather than stepwise phases. A five-fold increase in reference genome output

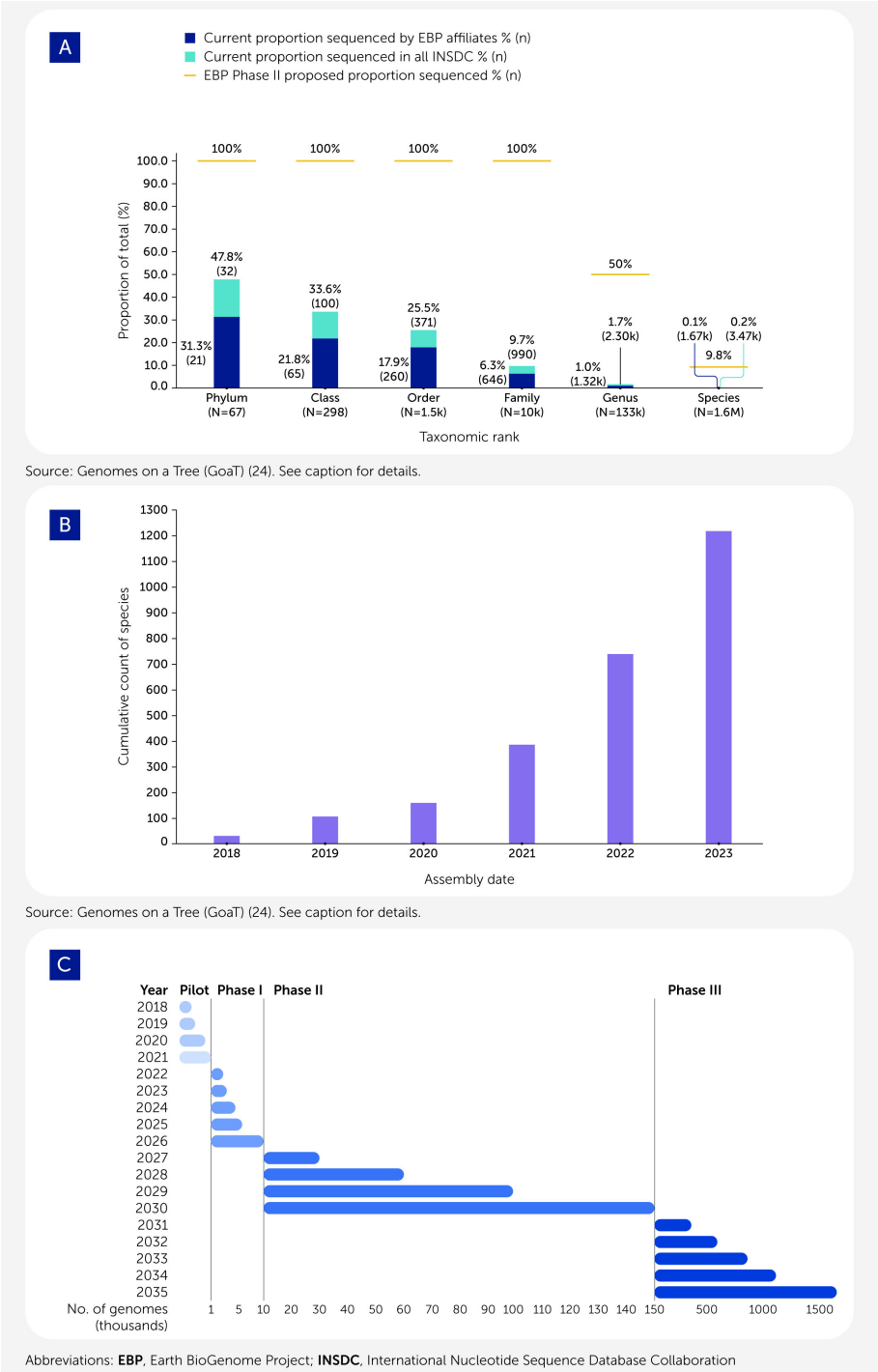



FIGURE 1
Progress toward sequencing all life in the Earth BioGenome Project (EBP). **(A)** The EBP's goal of generating high-quality genomes across eukaryotic life is being realized. By September 2024, over 3,400 genomes with qualities meeting the EBP minimum contiguity standards (contig N50 >1 Mb, scaffold N50 >10 Mb, and >95% of the genome in chromosomal super-scaffolds) had been made available in the open International Nucleotide Sequence Database Collaboration (INSDC) databases, representing nearly 48% of all phyla and nearly 10% of all families (turquoise histogram bars). Of these high-quality genomes, 48% have been generated by EBP-affiliated projects (blue histogram bars). EBP Phase II goals (gold lines) are shown. At the end of Phase II, the EBP aims to complete the sequencing of nearly 10% of all species and the vast majority of all families. Plot based on data presented in Genomes on a Tree (GoaT) using the United States National Center for Biotechnology Information (NCBI) Taxonomy database's taxonomy (see https://goat.genomehubs.org/2024.09.14/) (23). **(B)** The histogram illustrates the accumulation of EBP-standard genomes available in INSDC databases sorted by year of release. Plot based on data presented in GoaT (see https://goat.genomehubs.org/2024.09.14/) (23); assembly-level classification follows the INSDC definitions as outlined at https://www.ncbi.nlm.nih.gov/assembly/help/. **(C)** A timeline for EBP Phases I, II, and III is shown, indicating the approximate timing of each phase in terms of genome sequence delivery. The goals for Phase III are provisional and dependent on the success of Phase II.

rates is needed to achieve the goals of Phase I, and a further 10-fold increase is required to deliver Phase II (over 3,000 genomes per month) (Figure 1C). To deliver this increase, the Phase II proposal is built around three pillars: adaptive sampling, highest genome quality, and equitable global partnerships.

### Pillar 1: Adaptive sampling

Phase II should collect and biobank samples from 300,000 species and sequence 150,000 of these. Sampling will follow the evolving and exacting technical, ethical, and legal standards established during EBP Phase I (https://www.earthbiogenome.org/sample-collection-processing-standards-2024). An adaptive species selection strategy will maximize the number of genera sequenced while also delivering genomes for species that are economically and ecologically important, iconic, of special scientific interest, or of cultural significance to Indigenous peoples and local communities (with their assent).

### Pillar 2: Highest genome quality

Given the radical changes already achieved in genomic technologies, and the promise of further improvement to come, we propose that as many as possible of the 150,000 Phase II genomes be sequenced to EBP reference quality (https://www.earthbiogenome.org/report-on-assembly-standards). Generating genome sequences of high quality will transform their impact both as references for a focal species and collectively across ecosystems, major groups, and the entire field of biology. The technical challenges of generating reference genomes to current EBP standards for microbiota and meiobiota must be met and overcome.

### Pillar 3: Global leadership through equitable partnerships

It is imperative that the EBP has a global base, equitably distributing sample acquisition and data-generation activities and ensuring the equitable realization of the benefits of the work and the resources. To achieve Phase II, much of the species' collection, sample management, sequencing, assembly, annotation, and analysis will have to be based in the Global South and be delivered by EBP partners based in those nations. Genome sequencing will need to be supported at multiple sites in the Global South, including, especially, at laboratories based and sustained in countries with high biodiversity. We propose establishing a Foundational Impact Fund (FIF) to catalyze the realization of these benefits.

These three pillars will also be fundamental to Phase III. By building an equitable global network of cooperating partners promoting best practices in engagement and benefit-sharing, establishing rigorous standards and reproducible methods for the acquisition and sequencing of specimens, and inventing new methods and systems for large-scale annotation and analysis of many thousands of genomes during Phase II, we will be well placed to generate the genomes of all named eukaryotic species on Earth in Phase III.

## What should we sequence?

The EBP ultimately aims to generate reference genome sequences for all 1.67 million named eukaryotic species at the time of this writing—the species formally described by taxonomic communities over the last centuries (1, 4). The precise number of species known on our planet increases as new species are discovered and decreases as species become extinct and taxonomic revision resolves synonymy. Indeed, species are being described at accelerating rates, in part driven by an emerging synergy between molecular, morphological, and machine-learning (ML) approaches to taxonomy (28, 29). The total number of extant species is much greater than those described, with a consensus that we share our planet with at least 10 million other eukaryotes (28, 30–33). While genomics will play a significant role in the discovery and description of new species (34, 35), the EBP will continue to focus on named eukaryotic taxa rather than attempting to sequence and diagnose the vast number of undescribed species.

Not all species are equally accessible for collection and sequencing. We will adaptively prioritize species for sequencing in Phase II using the following four principles in sampling.

i **Phylogenetic diversity:** Phase II will select species representative of previously un- or under-sampled parts of the eukaryotic tree of life. Practically, this can be achieved by aiming to sample a representative for all accessible families (Phase I) and all accessible genera (Phase II).

ii **Conservation**: Phase II will ensure that species subject to conservation efforts, such as the >47,000 species on the International Union for Conservation of Nature (IUCN) Red List of Threatened species, are among the first to be sequenced. A process that allows communities, including representatives of Indigenous peoples and local communities, to nominate species for sequencing will promote engagement and understanding of genome sequencing and the potential value of its outcomes.

iii **Ecological or societal impact**: individual species can play keystone roles in the ecosystems in which they thrive. Species can also be important to human society because they provide ecosystem services, food, or other biomaterials or are pathogens, pests, or predators of valued species. Phase II will contribute to planetary health goals by prioritizing the sequencing of these species.

iv **Exceptional biological interest**: genomic sequencing can be a foundational step in understanding biology. For example, by sequencing species that appear to defy fundamental rules of biology, we can gain a deeper understanding of these rules. By sequencing all species in a well-studied ecosystem, we will better understand the interactions and dependencies that shape and maintain biodiversity.

Simply finding specimens is a key challenge. Species range from widespread to localized, common to rare, and large-bodied to small. The easiest to sample and sequence are those that are widespread, common, and large. The local, rare, and small are challenging to find, identify, *and* sequence. Based on centuries of biodiversity research, accelerated by the recent digitization of species' occurrence records, we have a reasonable overview of global eukaryotic diversity and its distribution, collated, for example, in the Global Biodiversity Information Facility (GBIF; see https://www.gbif.org/). Collecting representatives of every one of the approximately 167,000 living, valid eukaryotic genera (4) in just 4 years is infeasible. About one-third of all genera have only one or two species (Figure 2), and many of these are rare. Many species have been observed once and never recollected, and, sadly, many may be the victims of hidden extinction (36). While we fully expect that many collections made for the EBP will include chance encounters with rare species, it is clear that many other rare species will be practically uncollectable.

For Phase II we propose collecting 300,000 species, twice as many as will be sequenced. The species unsequenced in Phase II will prime Phase III. Campaigns focused on particular taxa (e.g., the VGP or Bat1K) (16), on species of particular concern (e.g., the Australian Threatened Species initiative; see https://threatenedspeciesinitiative.com) (37), on particular modes of life (e.g., the Aquatic Symbiosis Genomes project; see https://www.aquaticsymbiosisgenomics.org/) (20), or based in "genome observatory" sites delivering ecologically linked suites of species for sequencing (22, 38) will be critical in driving synergy between large-scale genomic sequencing and societal, ecological, and community benefits.

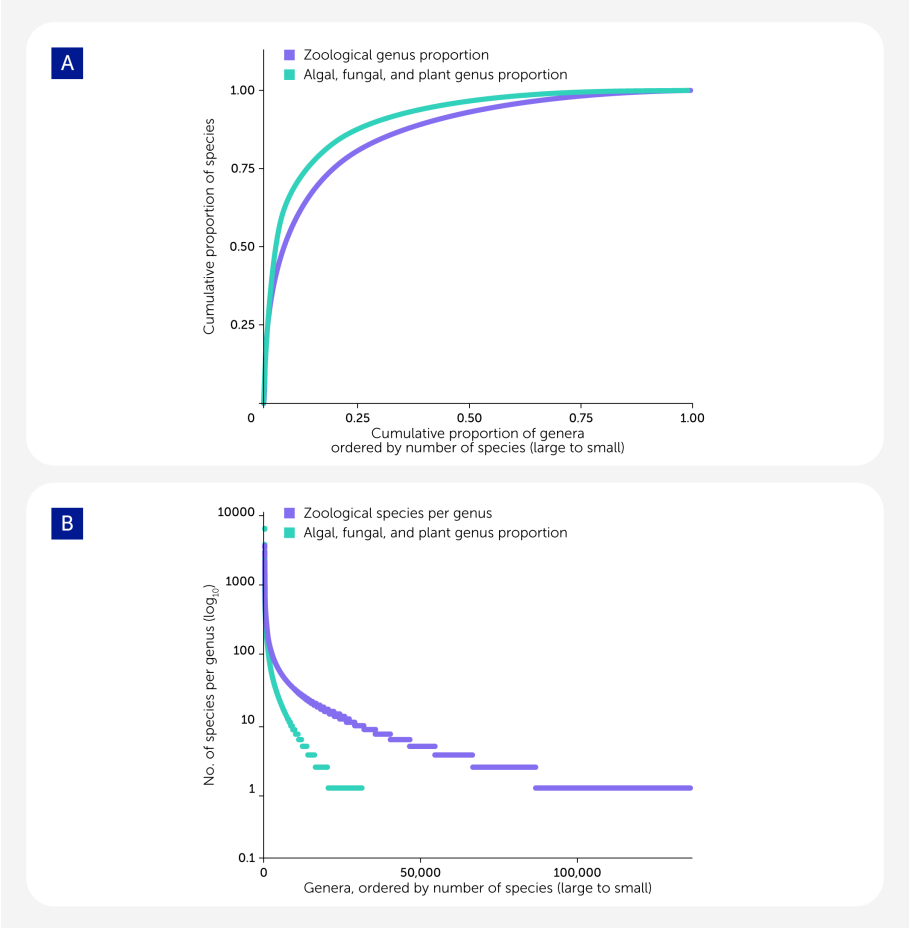



FIGURE 2
The pattern of life's diversity. **(A)** Less than 5% of all genera contain 50% of all eukaryotic species. For species defined under the International Code of Zoological Nomenclature (ICZN; see https://www.iczn.org/the-code/the-code-online/), 4.2% of genera (5,742) contain 50% of described animal species, and for species defined under the International Code for algae, fungi, and plants (ICNafp; see https://www.iapt-taxon.org/nomen/main.php); 2.6% of genera (807) contain 50% of described plant, fungal, algal, and other protist species. **(B)** Most genera contain only one or two species. Plot of the number of species per genus; genera are ordered by the number of species they contain. The most speciose genus defined under the ICNafp is *Hieracium* L. (hawkweeds, 5,524 species), while under the ICZN, the most speciose genus is *Stenus* Latreille (semiaquatic rove beetles, 3,113 species). The analyses presented are based on data available from the Catalog of Life on 31 December 2023 (4). The processed data are available in the Supplementary material.

## Challenges, blockers, and proposed solutions to achieve EBP Phase II

We have identified many technical and social challenges that must be overcome to realize the Phase II goal of collecting 300,000 species and producing 150,000 high-quality, annotated genome assemblies from projects worldwide. Below, we present the technical challenges as five interrelated themes:

(1) coordinating the sampling of 300,000 species in 4 years,
(2) progressing from sample to sequence to assembly at an increasing scale,
(3) producing high-quality annotations of 150,000 genomes,
(4) delivering impactful analyses,
(5) integrating innovative, planet-friendly informatics.

We also discuss the enormous challenges in creating a global biodiversity genomics workforce, coordinating such a large project across the planet, and securing funding. Many challenges have a cross-cutting impact and solutions require close collaboration between experts in many domains.

### (1) Coordinated sampling of 300,000 species in 4 years

Sampling 300,000 species presents a set of interlinked organizational, logistic, technical, educational, and social challenges. We must build an adaptive species sampling strategy informed by taxonomic, geographic, and prioritization considerations—and which is legal, ethical, politically sensitive, and culturally aware—and align it with the overlapping constraints and drivers of partner engagement and the availability of local or international funding. Data systems that aggregate biodiversity data, such as GBIF and GoaT, will facilitate the sharing and coordination of EBP Phase II activity.

#### Building a global community rooted in local action

The EBP was envisaged as a hubs-and-spokes organization of regional nodes and taxon-focused projects (1). The human division of the planet into nation-states does not overlap with the ecosystems, biomes, and bioregions that pattern biodiversity. Stewardship of biodiversity is similarly localized, and individuals and groups, including Indigenous peoples and local communities, have a deep local understanding of species diversity (5). The EBP will have the greatest impact if we build on these strong, local foundations. Here, we present a model for EBP regional nodes, based on building autonomous capacity for genomics, from sample acquisition to genome analysis. We emphasize that, in addition to *collecting* locally, we envision regional nodes that will also *sequence*, *assemble*, and *analyze* locally.

The throughput required to meet EBP Phase II goals could be delivered by 25 regional nodes, each collecting an average of 12,000 species and sequencing and assembling at least 6,000 species over a 4-year period. The inception of regional nodes will be driven by local initiative, availability of funding, and assessment of accessible biodiversity. Regional nodes will build on existing local scientific collaborations and knowledge. Sustainable regional nodes will require local skills, capacity, and funding (see the Workforce section below) and will likely take 2 to 3 years to implement.

Sample acquisition relies heavily on human capital and local skills. In contrast to expected savings in sequencing and assembly, as harder-to-source species are targeted, the costs of collecting will be relatively static per species. We envisage only a 40% reduction between Phase I and Phase II, even with the implementation of novel technologies. Much of the required expertise resides in local learned societies, taxon interest groups, national and local biological collections, and Indigenous peoples and local communities. We propose formal recruitment of collector allies to each regional node, who will bring specific taxonomic or habitat expertise and local user-community agendas. To promote sustainable careers, allies could agree to provide specified sets of legally and ethically sourced species, receive guaranteed compensation to recover staff and other costs, and be awarded explicit scientific credit for their work. Species acquisition for EBP sequencing through allies will support the currently underfunded expert taxonomy community, build capacity, and promote engagement with conservationists and other practitioners with the goals and outputs of the EBP.

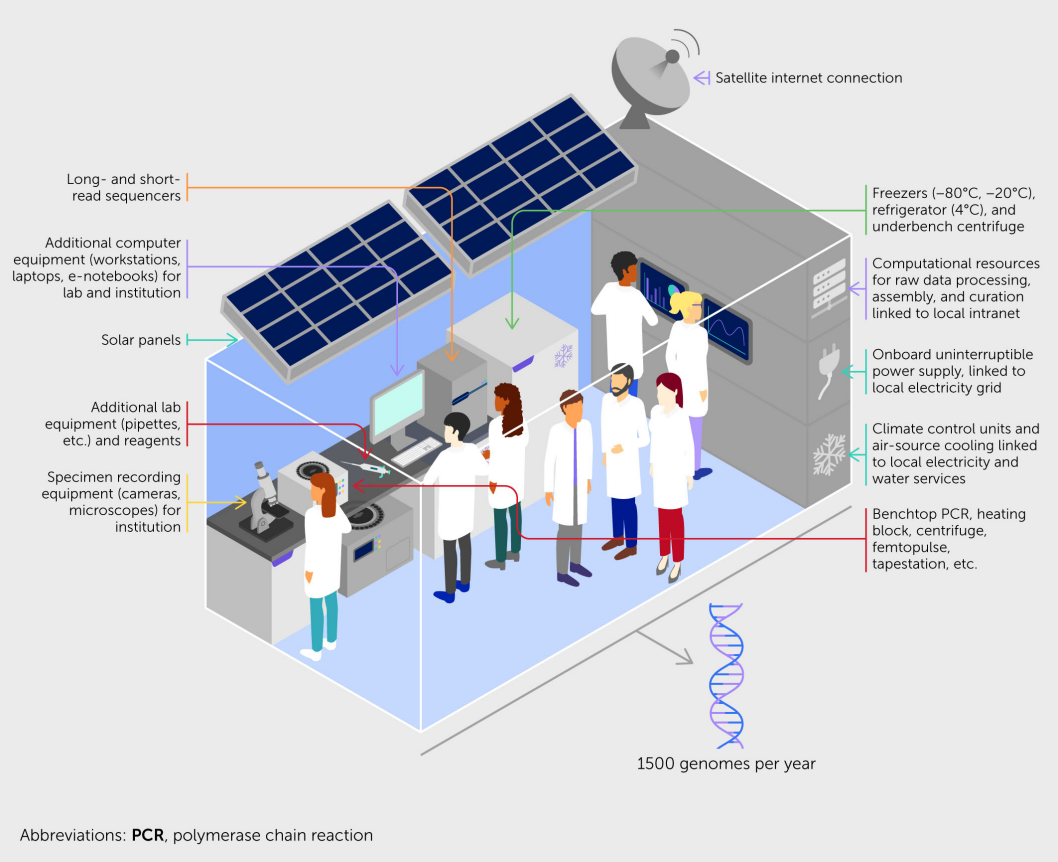
It is essential that regional nodes should be established in biodiverse regions, especially in the low- and middle-income nations of the Global South, which have historically been underrepresented in or excluded from the global scientific commons. New regional nodes should be strongly supported by existing biodiversity genomics centers. Building on existing installed capacity and interest, this will establish a legacy of genomics expertise that can be leveraged for a range of post-genomic work, including FIF projects. One way such capacity could be achieved is through the installation of a complete "genomes from a box" (gBox), specimen-to-sequence laboratory (see Box 1), equipped for EBP data production at scale. A gBox install would be accompanied by support from other established nodes through a system of mutual aid and buy-in from technology companies for reagents and support.

For Phase III, both of these models (regional nodes and biodiversity genomics allies) will have to be expanded to ensure collection from all biomes in an inclusive, just, and ethical manner. Regional nodes will serve as focal points to usher in a post-EBP world of genome-enabled science for conservation, medicine, and bioindustry.

#### Assuring delivery of the highest-quality samples to make the highest-quality genomes

EBP specimens contributing to reference genome assemblies should be accompanied by rich metadata. From sample acquisition to genome publication, the EBP will enforce use of the GBIF Darwin Core standards, which define information that must accompany any globally aggregated biodiversity data record (https://www.gbif.org/standards) (39). This will be coordinated via Darwin Core-compatible metadata management systems such as Symbiota (40). The EBP will redouble efforts to make specimen

**BOX 1** Genomes from a box.




Abbreviations: **PCR**, polymerase chain reaction

Responses to health emergencies such as the West African Ebola virus epidemic and the global COVID-19 pandemic have demonstrated that mobile facilities—built off-site, shipped in transport containers, and ready for use once connected to local electricity and water supplies—can transform crisis response. Our response to the biodiversity crisis needs a similar approach. One of the major hurdles to establishing genomics in a new location is the sourcing of the tools and skills required to run a genomics laboratory. One "leapfrog" solution might be if prospective Earth BioGenome Project (EBP) regional nodes, collaborative funders, and established genome centers worked together to source funding and build and deliver a genomics laboratory, with equipment ready-to-install and with training, reagent supply, and support guaranteed. This model is already in operation for biodiversity work, for example, in the In Situ Laboratory Initiative (https://insitulabs.org/).

A genomics laboratory in a box, or gBox, could be one route to establishing the approximately 25 EBP regional nodes needed to deliver Phase II. We envisage the gBox arriving as a standard shipping container, equipped to act as a high-throughput genomics laboratory. Inside would be all the equipment needed to transform specimens into DNA and RNA, make sequencing libraries, produce long- and short-read sequence data at scale, and turn those data into assembled and annotated genomes.

However, a gBox would be more than just hardware. It would also contain the equipment needed to collect, identify, and store species, ready to be placed in a field laboratory. Laptop computers and digital cameras would be ready to record specimens and their metadata, which would be stored in local databases before being transmitted to global systems, such as Global Biodiversity Information Facility and International Nucleotide Sequence Database Collaboration. An initial tranche of reagents would ship with the gBox, with subsequent shipments available on demand. The computer system would come ready-loaded with best-in-class pipelines developed elsewhere in the EBP and with a commitment to keep these up to date. The collaborating established centers would also offer training in sampling and recording, extraction and other molecular biology operations, instrument operation and use, and transforming raw data into chromosome-scale reference genomes. Training and support would be ongoing, creating a shared virtual laboratory to enable real-time problem solving. All sites with a gBox could participate in a virtual genomics commons to co-learn skills and pass on discoveries.

We intend to build a consortium of reagent suppliers and equipment manufacturers that would support the global collection of gBox labs with guaranteed reagent costing and onsite technical support. This partnership would ensure that the gBoxes run to their full potential. We expect each team, comprising six local laboratory staff members and a similar number of bioinformaticians to be able to deliver at least 1,500 genomes (of ~1 Gbase each) per year.

The gBox would be owned and operated by the receiving institute, which would be free to use it for other sequencing needs, such as urgent viral sequencing for public health or post-genome projects funded by the EBP Foundational Impacts Fund. The EBP will search for visionary funders who seek to partner with us to sponsor the manufacture, shipping, installation, and four-year operational costs of a fleet of gBoxes at EBP regional nodes.

metadata, genomic data, and analyses compatible with the (sometimes conflicting) demands of the FAIR [Findable, Accessible, Interoperable, and Reusable (41)] and CARE (Collective benefit, Authority to control, Responsible, Ethical) (https://www.gida-global.org/care) (42) principles of data governance. Wherever possible, a Traditional Knowledge and Biocultural Label or Notice (https://localcontexts.org/) should be attached.

Methods for high-quality, three-dimensional (3D) imaging of specimens compatible with the use of specimens for genomics are sorely needed. These images would provide an essential digital voucher for specimens subsequently consumed for sequencing. Imaging will be critical as the project progresses, as confidence in species identity may be lower, and sequencing may occur before full taxonomic identification. EBP specimen images could contribute to training resources for artificial intelligence (AI) and other computer-aided species identifications (28, 43, 44). The development of open-data systems and smartphone applications, similar to or allied with the popular iNaturalist platform (https://www.inaturalist.org/), would provide significant benefits.

Biobanking is critical for storing materials for future analyses. Detailed recommendations and protocols for biobanking in the age of genomics are available (45). Regional nodes must put in place secure biobanking of samples, ideally in collaboration with national or regional museums, botanical gardens, and other collections. Expansion of biobanking to explicitly support additional modes of analysis, such as proteomics, metabolomics, and single-cell atlasing, is recommended. Whenever possible, and especially for endangered species, cell lines should be created for future conservation efforts (46, 47).

Currently, DNA and RNA extractions maximally compatible with high-quality genomics are achieved from fresh or ultra-low-temperature flash-frozen material. Best practices for sample acquisition and shipping currently rely on live transport or an unbroken cold chain from collection to extraction. These practices are unsustainable on a global scale based on logistical, welfare, cost, and environmental grounds. Approaches that preserve specimens at ambient temperatures will be game-changing in terms of expanding sample collection. These are being explored with some success (48) but remain an urgent development area.

Delivery of EBP goals will require contemporaneous processing of several thousands of species in a single laboratory. Robust laboratory information management systems and electronic lab notebooks are essential. Live aggregation of sample process data from these tools in a workspace such as GoaT (24) would enhance shared learning of best practices. EBP members are already using open platforms to share best practices for collection, storage, and extraction (e.g., through protocols.io; see https://www.protocols.io/workspaces/earth-biogenome-project). EBP Phase II partners can enhance the content of these platforms with protocols modified to work at scale across diverse species.

## (2) From sample to sequence to assembly at increasing scale

To achieve the EBP Phase II goal of 150,000 reference-quality genomes in 4 years, the affiliated projects will have to sequence, on average, 3,125 genomes per month. To deliver this throughput, the EBP must recruit many additional data generation sites in regional nodes and other centers. However, we must also develop improved sequencing and assembly processes, particularly methodologies that enable automated laboratory workflows and improved bioinformatics workflows. Algorithms that better exploit the richness of long-read and long-range data, for example, in building fully haplotype-resolved assemblies, will be essential. The sequencing of single-celled and meiobiotal eukaryotes and the separation of target species from potential cobionts will be more challenging, but successes in these areas are already promising (8, 49–51). These developments will also prime the EBP for Phase III.

### Defining and meeting high-quality genome standards

Following the lead of the VGP (16), the EBP has established exacting but achievable quality metrics for reference assemblies (https://www.earthbiogenome.org/report-on-assembly-standards). We note that it is currently not technically possible to generate assemblies that meet these metrics for some species, largely because of small organism size and, consequently, minimal yield of DNA. Assemblies of such species will be attempted and submitted to the public databases, aiming to meet the EBP *representative* metric (0.1 Mb contig N50, chromosomal level, and 6.C.Q40). On the other end of the quality spectrum, we expect complete and near error-free, i.e., telomere-to-telomere, assemblies for a growing number of species, where contig N50 is the same as chromosomal N50, and all chromosomes are complete (C.C.Q40) (16, 52, 53).

The current recipes for genome sequencing to meet EBP metrics involve a mix of cutting-edge technologies. Three data types are currently used: single-molecule long-read data for contig building, long-range data from Hi-C for scaffolding, and transcriptomic data for accurate annotation. Close collaboration with technology providers will be essential to generate these data types at reduced per-genome costs (Table 1, see section, Costing the new Phase II strategy). These savings must be available worldwide and include equity-based price reductions.

In the future, it may be possible to simplify sequencing so that high-quality genome assemblies can be generated from a single data type, and workflows can be simplified by running long-read, long-range, and transcriptome libraries together on a single platform. Applying this paradigm (one sample, one library, one run, one genome) across biodiversity would put the EBP in a very strong position to deliver Phase III genomes to reference quality. New data types may also prove to be useful. It is already clear that ultralong reads (>100 kb) can be used to deliver much more contiguous, true telomere-to-telomere assemblies (54, 55). The generation of such data for a significant fraction of EBP target species would elevate the quality and value of the genomes produced.

### Overcoming technical challenges to genomics for all biodiversity

Sequencing and assembly procedures for most taxa are robust and ready for Phase II implementation (16, 26, 56), but challenges remain. We estimate that for about half of extant species, less than 1 ng of DNA can be isolated from a single specimen, orders of magnitude less than the input requirements for many current long-read processes. We need to develop robust, transferable protocols that generate genomic sequencing libraries from minuscule inputs without compromising assembly quality. Some success in this area has already been reported (8, 49, 50). More challenging still is the

sequencing of single-celled eukaryotes, including the paraphyletic "protists" and some fungi, several of which have surprisingly large genomes. Some will be sequenced from clonal cultures, but most species are not in culture. EBP-standard genome sequencing of single cells from environmental sources is currently very challenging. Approaches that combine bulk and single-cell data are promising (57–60) but need further development. EBP-affiliated projects are also exploring the problems presented by polyploid genomes, where varying levels of rediploidization make the assembly of distinct sets of homeologous chromosomes difficult (52, 61).

Another issue that impacts biodiversity genomics based on sampling from the wild is that target organisms may be accompanied by mutualist or parasitic symbionts, components of the host microbiome, or by accidentally co-isolated organisms (51, 62). Robust separation of the genomes of these cobionts from that of the target species is essential to avoid misattribution of biological capacity. This work has been facilitated by the recent development of highly sensitive and specific decontamination workflows such as Foreign Contamination Screen (FCS)-GX (63). Nevertheless, close attention to this aspect is essential for the future.

Informatics workflows covering primary assembly, haplotypic duplication removal, scaffolding, decontamination, and pre-curation processing, readying for EBP Phase II, are already openly available in workflow management systems such as Galaxy (64) and Nextflow nf-core (e.g., https://pipelines.tol.sanger.ac.uk/pipelines). The use of AI and ML toolkits to intelligently automate decision-making processes, such as raw data quality control and assembly curation, will make the flow of genomes more efficient and improve output genome quality. Equally important will be the open sharing of process and quality control information, so that issues can be foregrounded and solutions found rapidly for all EBP nodes.

Delivery of these advances will require extensive, focused research and development in the academic and commercial sectors to develop better methods of acquisition and shipping of specimens, extraction of nucleic acids, sequencing, and assembly. We have included an estimate of US$100M for these activities in each of Phases II and III.

## (3) High-quality annotation of 150,000 genomes

Annotation is crucial for understanding the functions encoded in a genome and is the starting point for downstream analyses (65). Over the last decade, major annotation services such as Ensembl (66) and RefSeq (67, 68), and the broader annotation community (69–72) have increased the quality and speed of annotation. Despite these successes, annotation remains a complex and computationally expensive process and a bottleneck to unlocking the value stored in genomes. To meet Phase II goals of annotating 150,000 genomes over 4 years, we need to develop radical new annotation approaches that leverage the diversity of expertise across the global community and optimize the use of available data and computational resources. In particular, new, scalable approaches should be supported, such as cross-genome orthology predictions—as used in TOGA (Tool to infer Orthologs from Genome Alignments) (72)—and AI deployment in gene prediction (73). As currently unexplored branches of the tree of life are illuminated by new reference genome assemblies, we expect to find exceptions to general patterns derived from the current relatively small subsample of annotated genomes (>5,000 eukaryotes, mostly yeast), such as new genetic codes, diversity in splicing mechanisms and patterns, and programmed editing during transcription. These "exceptions that prove the rule" can be identified, defined, and deployed to better understand the functional genomics of all life.

### Annotating genomes to realize their value

The minimal annotation product for every EBP species should be the annotation of protein-coding and conserved non-coding gene types (https://www.earthbiogenome.org/report-on-annotation-standards) (6), accompanied by repeat finding using curated repeat libraries and *de novo* discovery. Many transcription units, perhaps most in lineages such as vertebrates, give rise to more than one mature transcript, and defining the diversity of these isoforms is essential to unpicking the true diversity of genes. Currently, all annotation approaches rely fundamentally on alignment to the genome of transcriptome or protein data, and statistical models of genomic features. For EBP Phase II, the generation of at least 50 million read pairs of short-read transcriptomic data from a single library for every species is the absolute minimum for high-quality annotation. Long-read transcriptomic data are an attractive alternative to standard short-read RNA sequencing as they robustly reveal the diversity of transcript isoforms (74), but are currently expensive compared to short-read data. Concatenation sequencing of full-length complementary DNAs (cDNAs) on the Pacific BioSciences (PacBio) high-fidelity (HiFi) long-read sequencing platform promises to deliver sequence reads for annotation at a reasonable cost (75), and transcript normalization may maximize the utility of these data for annotation. Development and community benchmarking of these and other techniques may usher in an expectation of long-read transcriptomics as standard. High-quality annotation, such as that currently built for "model" species, relies on transcriptome data covering multiple tissues, developmental stages and conditions, and additional functional genomic data. This is unlikely to be achievable for most species, where developmental stages are not collected and dissection into tissue types is impractical. It will be important to build tools that recognize the diversity of alternative splicing, perhaps using information from related species.

### Beyond just transcriptomics

Cells and organisms read their DNA code using a complex mix of sequence-based and epigenetic signals. The development of AI toolkits to predict genes and their likely activity will need significant functional data beyond just deep sampling of mature messenger RNA (mRNA) transcripts. The EBP encourages the generation of additional modalities of functional genomic data for which high-throughput methodologies are already available (such as sequencing non-polyadenylated RNAs and small RNAs, identifying transcription start sites, determining cytosine and adenine DNA

methylation, mapping open chromatin, and defining the patterns of chromatin histone modification), especially for species representative of families. High-quality annotations driven by rich multimodal datasets for a diverse set of carefully chosen species across biodiversity would provide a platform for new comparative approaches that leverage whole-genome alignments between well-annotated and newly sequenced species to simultaneously annotate coding genes and infer orthologous gene loci (72). AI offers an emergent opportunity to achieve high-quality annotation, particularly of protein-coding loci, by "learning" the embedded transcriptional code from well-studied taxa (73). Sharing existing and new high-quality, curated, and transcriptome-validated gene predictions as dense training data for deep learning approaches will be imperative to promote these developments.

Annotation of EBP genomes should aim to meet FAIR and CARE principles and be made publicly available through submission to INSDC databases. FAIR principles demand that annotations should be accompanied by defined metadata, including methods (software tools and parameters), external data used, and agreed-upon quality metrics. Tools such as Benchmarking Using Single Copy Orthologs (BUSCO) (76), compleasm (77), and OMArk (78, 79) that exploit the expectation of the presence of a curated set of single-copy orthologs to assess coding-gene annotation completeness will need to be dynamically updated to maintain precision as Phase II ramps up. Additional quality assessment metrics require development, such as descriptors of ancestral linkage group retention, gene structure congruence, the proportion of genes with transcriptome support, and the number of proteins containing known domains. A standardized tool for multidimensional metric computation from a genome and associated annotation files would contribute to streamlining the entire process of producing reference-quality genomes across the eukaryotic tree of life.

## (4) Delivering impactful analyses

To realize the value in EBP genome assemblies, we need to re-envision how we derive meaning from genomic data at scale. The EBP will, collectively, generate exabytes of raw and analyzed data, and Phase II alone will generate in excess of 200 terabases of assembled genomes. While this wealth of data promises a time of plenty for analytical genomics, it also brings challenges. The drive in the EBP for uniform, high-quality, and open reporting and assessment of assembly metrics will promote the combinability of all the genomes released. In comparative genomics, many computational tasks scale unfavorably with the number of genomes analyzed. Addressing these bottlenecks will require the development, coordination, and integration of research tools, infrastructure, and human resources at an unprecedented scale. The proposed FIF is designed to facilitate these analyses, especially by supporting initiatives, researchers, and organizations in the Global South.

### Comparative and conservation genomics at an unprecedented scale

For comparative genomics analyses, the EBP will have to ensure that products derived from genome sequencing are available for open use. Products will include large-scale whole genome alignments, ancestral linkage group inference (10, 80, 81), repeat and mobile element family data aggregated across species (82), up-to-date and comprehensive gene orthology calls (78), genome-anchored descriptions of conserved functional elements (83–85), and genome-wide description of the 3D structure of each genome (86). To generate these at the new scale of Phase II will require active development of the toolkits used, many of which can currently only scale to tens or hundreds of genomes. The EBP must foster and, where possible, sponsor the exploration of new algorithms and computer architecture for comparative analyses, promoting the inclusion of all data-generating communities and nations.

Globally accessible resources are particularly important for species and ecosystems under threat, where targeted investment in additional data, such as population genomics data, will be crucial in estimating extinction risk, managing wild populations, and understanding the genetic underpinnings of adaptation to local environments. For individual species, expanding population genetic approaches to the whole genome will reveal large-scale structure and illuminate critical details of population interconnectedness (9). At the landscape and ecosystem levels, the availability of complete genomes for many species in an ecosystem will make approaches for investigating species presence, species interaction, or functional capacity using environmentally sourced DNA (eDNA) from water (87), sediments (88), or the atmosphere (89). These approaches will rely on the ability to map eDNA reads to large databases of well-annotated sequences. Metagenomic approaches to exploring biotic diversity diversity, such as the Tara Oceans initiative (90, 91), will be transformed by EBP genome data, again *via* large-scale read mapping to rich, open datasets. The EBP should promote significant pilot projects that explore the use of genomic resources in assessing the diversity, functional capacity, and temporal dynamics of selected ecosystems in "genomic observatories", preferably in the Global South, through the FIF.

### Beyond just the genome

The Encyclopedia of DNA Elements (ENCODE) project has shown the power of deep, coordinated multimodal genomic assays in discovering the function and regulation of genomes in humans and model species (92–94). Similar deep-dive functional genomics analyses of species selected for their phylogenetic disparity or for their potential to illuminate particular evolutionary transitions would be very powerful. We can imagine additional species being selected for a Diversity ENCODE program, developing new approaches to permit multimodal data generation from diverse systems.

Similarly, cell atlasing projects focused on humans and other model species have illuminated the cellular diversity of tissues and systems, revealing the genetic underpinning of complex traits such as immunity and development (95, 96). The Biodiversity Cell Atlas (BCA) initiative proposes expanding the list of species assayed at the single-cell level to explore the dynamics of development and environmental response across eukaryotes (97). The BCA will coordinate with the EBP to sample across diversity to illuminate the evolution and diversification of cell types across life and, in turn, BCA data will enrich genome annotation in targeted species and their relatives, for example, by enhancing understanding of co-expression networks and of the links between the diversity of genes present in a genome and species phenotypes.

## (5) Integrating innovative, planet-friendly informatics

Information technology (IT) strategies impact the whole span of EBP activities from organism identification to dissemination of sequence information through data portals and publications. Implementation of IT solutions on a global scale is challenged by restrictions on data access and sharing under international treaties, (such as the Nagoya Protocol), lack of standardization of metadata, different IT infrastructure capabilities and data across EBP affiliates, and different standards for data analysis, schemas, archiving, and sharing. Computation is energy-intensive, especially when AI is used (98), and it would be self-defeating for the EBP to contribute significantly to climate change-inducing emissions because of the project's hunger for computer power. To address these challenges, EBP affiliates will need to actively promote the use of trusted global commons for laboratory protocols, computational pipelines, raw data, specimen and assembly metadata, genome sequences, and post-genomic data products. Solutions to these challenges will require that the EBP deals openly with issues of data provenance and meets the objectives of both access *and* benefits sharing.

### Data are nothing without linked, trusted metadata

The EBP must promote consistent frameworks for collecting and accessing metadata, the information needed to track provenance, attribution, data processing activity, and public distribution, and to integrate EBP activity with global biodiversity and sequence information data services. The EBP has already taken concrete steps in planning for meeting Phase II requirements (see https://www.earthbiogenome.org/it-and-informatics-standards). Close coordination with INSDC members will be necessary to manage the archiving and sharing of public raw and analyzed data, and the EBP will promote the use of INSDC databases for all outputs. Metadata for public release must be readily combinable and interoperable, based on defined ontologies accepted in the field, and accessible through application programming interfaces. For example, information about the genetic code likely to be used by a species is needed to establish parameters for genome annotation, and sample descriptors including Darwin Core-compatible collection location, time, voucher identifiers, collector, and provenance are needed when submitting the sequence data to public repositories and when using the reference genome in the context of time- and space-resolved population datasets.

In the fast-moving EBP Phase II (over 3,000 genomes released per month, with 30,000 species "in flight" at any one time), real-time, trusted data sharing and integration will be critical. In a global project with overlapping ownership, jurisdictions, and interests, the EBP will strive to ensure the highest standards of explicit sharing of ongoing data generation. For example, GenomeArk (https://www.genomeark.org/) has been built to provide pre-publication access to and sharing of high-quality reference genomes. GoaT (https://goat.genomehubs.org) offers an integrative view of the EBP and affiliated project activities. Data partnerships among research institutions, governments, funding agencies, and the private sector will be needed to ensure the EBP delivers to its full potential. We envision long-term thinking, coordinated action, and committed funding to ensure that data sources created by the EBP will be a lasting legacy.

### Keeping the planet green

The complexity of integrative analysis across thousands of genomes generally scales with the square of the number of species analyzed. The cost of computation (98–100) and consideration of the implied carbon footprint of the EBP favor approaches that generate shared analytic products for wide reuse. The EBP will work toward a "compute once, reuse many" approach, where core analytic products are precomputed for all to reuse. For example, whole-genome, reference-free alignment (101) is costly, with final products best shared rather than regenerated. Similarly, phylogenetic analyses of species and genes requires significant computation, and dynamically updated phylogeny and gene orthology assignments can be generated once and reused many times (102). Refactoring algorithms to support incremental updates when new species' genomes are released—rather than re-running full analyses—can avoid costly whole-dataset recomputations. For example, in phylogenetics, heuristic placement of new taxa updates trees without recalculating from scratch (103).

Workflow management systems are critical to ensure the highest quality of data products, improve automation and scaling, reduce costs and carbon footprint, and meet FAIR and CARE principles. We envisage shared development of open resources widely distributed through workflow hubs. It is clear that AI methods will become widespread in the coming decade and, for the EBP, immediate applications in data tracking and annotation are evident. However, AI is expensive (104). We will need to ensure that EBP data are maximally AI-ready on deposition by providing detailed metadata and extensive quality-controlled training sets.

EBP projects will need to pay close attention to cybersecurity best practices in software, workflows, data storage, and management to protect data integrity and data privacy (e.g., under the Nagoya Protocol). Ensuring equitable access to all EBP data effectively

requires access to global or regional cloud-based storage. Utilizing these resources effectively means that the EBP will have to maximize the compression of raw and analyzed data while ensuring carbon-neutral operation of the chosen data storage providers. Redesigning the EBP informatics workflow to minimize carbon emissions, and partnering with vendors and facilities that demonstrably reduce and offset $CO_2$ from storage and computing, will lower the project's projected carbon footprint (105).

# Building and sustaining a global EBP Phase II community

## Democratizing genomics skills: a global approach to building a skilled workforce

EBP Phase II will require a globally distributed, well-trained, multidisciplinary workforce to address potential challenges. The skill sets required are diverse and include species identification, sample finding, collection and processing, nucleic acid extraction, and genome sequencing, assembly, curation, annotation, and analysis. In addition, the EBP needs to support the development of ethics, data governance, cultural competency, community engagement, benefit-sharing, and leadership.

The EBP will support training and mentorship activities globally, whether through sponsorship of online, open masterclasses and workshops (such as the Biodiversity Genomics Academy; https://thebgacademy.org/) or by cross-project/cross-center internship collaborations. EBP affiliates will need to build capacity by developing a skilled workforce in their geographical areas and establishing mutual aid-based training and mentoring across projects. A functioning regional node producing 1,500 reference-quality genomes per year will require at least six genomics lab technicians and four bioinformaticians, with additional support from collectors, taxonomists, and staff at natural history collections. EBP-affiliated nodes can promote capacity building within the necessary disciplines by supporting biodiversity genomics-focused components in school and university curricula. EBP affiliates should establish mechanisms to integrate Indigenous knowledge, thereby completing the virtuous circle of data sharing for capacity building.

Leadership training and mentorship will be essential, as all nodes will need to coordinate local efforts across the workflow, including community engagement, ethical and legal compliance, sample collection and processing, and generation and release of genomes. Leadership will need to coordinate globally to ensure consistency across the EBP in quality and other metrics, coordinate sampling to minimize overlap in the species being worked on, and ensure effective access and benefit-sharing. EBP representation and input at relevant global gatherings and institutions, such as the Convention on Biological Diversity (CBD) and the United Nations (UN) Climate Change Conferences, should reflect the diversity of EBP projects. The existing EBP governance structure provides initial guidance on many of these issues (https://www.earthbiogenome.org/governance-documents).

Only by establishing this global, diverse, and interconnected EBP workforce will we be able to deliver the 150,000 genomes aimed for in Phase II and build momentum to seed Phase III. Working within an EBP regional node will enable individuals to deliver impactful science and establish future careers in related areas, such as population genomics, comparative genomics, genome function and evolution, phylogenomics, conservation, human genetics, and disease. EBP Phase II can thus be an engine that delivers genomes and builds a workforce skilled in advanced bioeconomy, biotechnology, and medicine.

## Enhancing global coordination

Organizationally, the EBP is a global network of networks. Achieving the goals of Phase II will require open, detailed coordination based on mutual respect, creative compromise, and informed agreement, meeting social, cultural, technical, scientific, and user-value goals. The EBP was established in 2018 under a Memorandum of Understanding and transitioned in 2022 to a permanent governance structure (https://www.earthbiogenome.org/governance-documents). The EBP is composed of affiliated projects that are represented on the Membership Council, a voting body that approves all EBP initiatives and actions. An elected Chair and Executive Council are charged with overall project coordination and facilitating the project's growth. The Executive Council relies on the activities and recommendations of six standing committees (International Scientific Committee; Ethical, Legal, and Social Issues Committee; Justice, Equity, Diversity, and Inclusion Committee; Communications and Public Affairs Committee; Nominations Committee; and Governance Committee) to create policies, guidelines, and white papers, which are discussed, revised, and agreed on by the Membership Council. The International Science Committee delivers to a wide technical remit through five subcommittees: Sample Collection and Processing, Sequencing and Assembly, Genome Annotation, Data Analysis, and Information Technology. An EBP Secretariat was recently established at Arizona State University, United States, to support the completion of Phase I and the initiation of Phase II. The Secretariat coordinates business and meetings, enhances communication between members, ensures integration with affiliated projects, and reaches out to the wider public.

The need for coordination within the EBP network, which already includes 60 affiliated projects with thousands of active participants (https://www.earthbiogenome.org/affiliated-project-networks), can only grow. While all work to the same overarching objectives and standards, different affiliated projects may have distinct goals, driven by their scientific, funding, and cultural environments. EBP coordination roles have been delivered by the voluntary commitment of participants and, more recently, through multi-institutional funding for Secretariat positions. For Phase II scaling, an enhanced Secretariat is essential to link projects at both organizational and technical levels, facilitate cross-training and

other synergies, support the establishment of new projects and regional nodes, integrate EBP efforts with other global biodiversity infrastructures and programs, such as GBIF, and give a voice to the EBP within international and regional policy setting fora (such as the CBD and the Agreement on Marine Biological Diversity of Areas beyond National Jurisdiction under the UN Convention on the Law of the Sea; https://www.un.org/bbnjagreement/en). This enhancement will require further acquisition of dedicated, stable funding.

Rich technical coordination between projects will be the nexus for a shared understanding and collective vision of our effort. The EBP uses GoaT to coordinate the aspirations and progress of each affiliated project (see https://goat.genomehubs.org/projects/ebp). Through GoaT, the EBP deploys effective, real-time systems to resolve species overlap between projects, support the planning of sampling campaigns, underpin the creation of distinctive funding applications, and enhance multilateral collaborations. The EBP should ensure open access to the knowledge being built, from sampling protocols to analysis methods and process management, across affiliated networks. Training programs that are open to qualified applicants everywhere are critical to developing and building global capacity in biodiversity genomics. As an example, the open Biodiversity Genomics Academy offers a self-service menu of courses and modules, dynamically updated by domain experts to reflect best practices and capture the critical details of real-world applications that can be tailored to local needs.

## Costing the new Phase II Strategy

In 2018 we estimated that completion of the EBP would cost US$4.7 billion (1). Based on our experience and developments in laboratory technologies and informatics, we now estimate that Phase I of the EBP (sequencing and annotating the genomes of approximately 10,000 species) can be completed for US$285 million, compared to the US$600 million estimated in 2018 (Table 1). We have also re-estimated projected costs for the subsequent phases of the EBP. Based on achieving additional, reasonable efficiencies of scale and process improvements, we estimate that Phase II can deliver 150,000 high-quality genomes at one-eighth of the current unit cost of genomes in Phase I, even though we now propose sequencing all species to reference rather than short-read draft quality. Excitingly, for Phase III we estimate that all species can be sequenced to EBP reference-quality with a relatively minor (10%) increase in overall cost. With realistic assumptions about future sequencing costs continuing to decrease per species, we now estimate that genome assemblies for the vast majority of the 1.67 million named species can be completed to a uniformly high standard for US$3.9 billion. We note that these costs do not consider the variation associated with genome sequencing in the Global South and other developing areas of the world, where instrumentation and reagent costs are usually higher, but labor and sample collection costs may be lower (106).

We also propose that the EBP should commit to establishing a US$0.5 billion FIF to support research, especially in the Global

TABLE 1 Estimated budget for the Earth BioGenome Project.

| | Phase I | Phase II | Phase III | |
|---|---|---|---|---|
| | Most families | Most genera | All species | Total |
| **No. of species to be sampled and sequenced** | | | | |
| Sampled | 10,000 | 300,000 | 1,360,000 | **1,670,000** |
| Sequenced | 10,000 | 150,000 | 1,510,000 | **1,670,000** |
| **Project costings (US$, millions)** | | | | |
| Sample collection (collection, identification, shipping, and biobanking) | 10 | 250 | 690 | **950** |
| Sequencing (genomic and transcriptomic), assembly, annotation, and databasing | 250 | 500 | 2,000 | **2,750** |
| Research and development for collection, sequencing, and informatics* | | 100 | 100 | **200** |
| Coordination and secretariat funding | 5 | 6 | 10 | **21** |
| **Project core cost (US$, millions)** | **265** | **856** | **2,800** | **3,921** |
| Foundational Impact Fund (FIF) | | 250 | 250 | **500** |
| **Total with FIF (US$, millions)** | **265** | **1,106** | **3,050** | **4,421** |
| Original proposal (US$, millions) | 637 | 1,612 | 2,493 | 4,742 |
| Cost per reference genome (US$) | 26,500 | 6,100 | 1,900 | 2,400** |

*We assume that the technology providers will continue to increase capacity and quality and decrease the per-sample cost of genomic data acquisition, as they have over the last three decades. We do not include these research and development costs here.
**Overall cost per reference genome based on estimated costs.

South, to improve technologies for biodiversity genomics and deploy the wealth of the reference genome sequences into conservation, biodiversity enhancement, and biotechnological and biopharmaceutical applications. Thus, the full cost of the integrated EBP vision is estimated at US$4.42 billion, spread over a ten-year timeline (Table 1), with US$1.11 billion required for Phase II and US$3.1 billion for Phase III.

## Building a global EBP funding strategy

Obtaining the US$3.9 billion required to collect, sequence, and annotate 1.67 million eukaryotic species is a considerable challenge. Phase I—sequencing 10,000 species—is decentralized to individual affiliated projects, and the larger projects have raised upwards of US$200 million. Securing funding for the completion of Phase I and initiating Phase II is a high priority and has almost been realized. Successes to date have leveraged the vision of the global project to generate considerable enthusiasm from public and private funding sources, and the EBP is pursuing multiple strategies to achieve funding goals. Ideally, the EBP requires pooled funding from multiple geographic regions to deliver improved coordination and outreach, thereby maximizing scientific and societal benefits. We recognize that funding may come with important stipulations, such as open and free access to data, considerations relating to intellectual property, benefit-sharing, capacity development and building, and partnership with Indigenous peoples and local communities, that may complicate achieving the project's goals.

Attractive possibilities for funding include pre-competitive consortia as well as pooling resources from public agencies, major research universities and institutions, and private companies. Non-governmental organizations, not-for-profits, and the general public have shown interest in funding EBP activities. Crowdfunding among scientists has been effective in raising funds (e.g., for the VGP) and could be expanded to fill important taxonomic gaps. We anticipate that individual philanthropy will play an important role in achieving the project's end goals, and much effort is already underway to work with visionary philanthropists who appreciate the planet-critical nature of the EBP (e.g., the Minderoo Foundation's "OceanOmics" initiative; https://www.minderoo.org/oceanomics).

As noted above, the EBP's impact can be maximized through rich rewards delivered by other modalities of analysis, such as deep functional genomics, proteomics, metabolomics, or single-cell atlasing, applied to a wide diversity of species. The biobanks established or enhanced for the EBP could provide essential support for these additional programs of work, which would need significant additional funding if attempted on a large scale.

Including the FIF, the total of US$4.42 billion required to fulfill the EBP goal of sequencing and analyzing 1.67 million species in 10 years is very reasonable for a global effort with such a lasting impact. The EBP offers extraordinary value for money if one compares the project cost to the US$3 billion Human Genome Project (nearly US$6 billion in inflation-adjusted dollars; see 107), the US$10 billion cost of the Webb Telescope (108), or the US$13 billion cost of discovering the Higgs boson at the Large Hadron Collider (109).

## Justice, equality, diversity, and inclusion in access and benefit-sharing across the EBP

In delivering its mission, the EBP will need to creatively solve a series of issues arising from its commitments to justice, equality, diversity, and inclusion. These issues range from unequal infrastructure and prohibitive costs to ineffective or one-sided communication spanning all the technical and logistical challenges discussed (Table 2). The EBP acknowledges concerns surrounding access to the unprecedented volume of digital sequence information (DSI) that the project will generate, as well as the benefits that can be derived from EBP data. The EBP also recognizes the rights of countries, Indigenous peoples, and local communities that contribute to the collection of genetic resources (5), works to ensure that these rights and interests are respected and advanced throughout, and strives to cultivate a culture of working together to harness the power of DSI for the betterment of humanity. It has been estimated that Indigenous peoples and local communities steward 80% of the Earth's remaining biodiversity (110) and thus proactive engagement with Indigenous peoples and local communities is important because the wealth of intergenerational, place-based knowledge can provide an enhanced understanding of the Earth's biodiversity and how to protect, use, and conserve it. Recognizing the rights and interests of all these communities is crucial for the EBP to achieve its Phase II goals and lay an inclusive and equitable foundation for Phase III.

The EBP remains fully committed to sustainable development and the fair and equitable sharing of benefits arising from the use of genetic resources, operationalized through the 2014 CBD Nagoya Protocol. At the 15th Conference of the Parties (COP15) in 2022, a multilateral mechanism for benefit-sharing was included in the Kunming-Montreal Global Biodiversity Framework (111). This mechanism highlights not only monetary benefit-sharing but also the reinforcement of value creation and sharing, emphasizing inclusive and open access to DSI and the need to develop and build capacity, including technology transfer to bridge the gap between developed and developing countries. Additional international agreements, such as the UN Declaration on the Rights of Indigenous Peoples (112), as well as national laws also shape the work and ethics of the EBP.

Anticipating and addressing the ethical, legal, and social justice issues that the EBP will face during Phase II will accelerate the realization of the project's mission. The EBP will directly contribute to the goals of the Kunming-Montreal Global Biodiversity Framework and other treaties, including the UN Convention on the Law of the Sea (113), the International Treaty on Plant Genetic Resources for Food and Agriculture (114), and the emerging Pandemic Prevention, Preparedness, and Response Agreement

10.3389/fsci.2025.1514835

TABLE 2 Technical and social challenges the Earth BioGenome Project (EBP) must face.

| Technical and social challenges | Issues | Progress required |
|---|---|---|
| Sample collection and processing | • Unequal distribution of biobanking/vouchering infrastructure<br>• Diverse permissions regulations leading to inequitable species collection and inadequate metadata | • Connect providers to existing infrastructure for sample deposition<br>• Create and enforce policies ensuring fair attribution<br>• Generate sustainable sample shipment mechanisms<br>• Enforce transparency and compliance, and reduce duplicated efforts |
| Sequencing | • Unequal infrastructure distribution and costs<br>• Differing technological equipment and reagent storage costs<br>• Prohibitive infrastructure upkeep costs for sustained participation in the Global South | • Lobby for local infrastructure<br>• Seek discounted service rates<br>• Prepare hands-on sequencing training |
| Assembly and annotation | • Unequal access to resources<br>• Limited accessibility and/or restricted access due to paywalls<br>• Deprioritization of curation in laboratories in the Global South | • Design all code, workflows, and standards as open access<br>• Deploy software utilizable in resource-limited settings (e.g., CPU/GPU use)<br>• Deliver hands-on assembly, curation, and annotation training using open-access content<br>• Develop tools to handle data in compressed formats |
| Downstream analysis | • Unequal infrastructure distribution and costs<br>• Biased datasets for training models<br>• Computationally intensive<br>• Unequal capacity to translate genomics into applications | • Develop and use downstream analysis tools responsibly and sustainably<br>• Reduce species biases and sequencing duplication<br>• Promote international collaboration |
| Workforce and training | • Funding for capacity building and knowledge transfer<br>• Unidirectional knowledge sharing<br>• Lack of diversity in STEM | • Scale equitable and inclusive training models<br>• Support career paths for underrepresented groups<br>• Provide reciprocal bidirectional training with local partners<br>• Invest in project coordination and communication |
| Engagement | • Unfair distribution of benefits/burdens<br>• Distrust from legacy extraction and exploitation<br>• Inadequate engagement with trans-sectoral interested parties and inclusion of their worldviews throughout the project | • Prioritize cultural awareness and inclusion of other worldviews and value systems<br>• Engage internationally with communities throughout the research process<br>• Strengthen existing partnerships and co-build new ones<br>• Obtain appropriate consent and mutually agreed-upon terms before project onset |
| Communication and coordination | • Disparities in digital technologies and high-speed internet access<br>• Priorization of, and balance between, FAIR and CARE principles<br>• Ineffectiveness of global communicational coordination | • Engage with policymakers in Global South nations<br>• Reconcile FAIR and CARE principles<br>• Invest in centralized outreach and coordination<br>• Dismantle power imbalances where possible |

Note: The suggested solutions are non-exhaustive and many apply to multiple stages of each work category.
Abbreviations: **CARE**, collective benefit, authority to control, responsible, ethical; **CPU/GPU**, central processing unit/graphics processing unit; **FAIR**, findable, accessible, interoperable, and reusable; **STEM**, science, technology, engineering, and medicine.

(115). The EBP will serve as an essential partner in enabling all nations and peoples to progress and share the benefits of global biodiversity genomics fairly and equitably. As the EBP advances into its next phases, it aims to establish mechanisms that ensure that the use of DSI leads to tangible benefits for countries, communities, and peoples. These mechanisms will include exploring models for non-monetary benefit-sharing, such as capacity development and building initiatives, technology transfer, and the development of partnerships that promote sustainable development in regions of origin, the retention of young researchers, and ensuring that the FIF is well subscribed to and equitably disbursed.

The EBP will contribute to several UN Sustainable Development Goals (SDGs) that relate to biodiversity and genetic diversity. Specifically, reference genomes will contribute to SDG 2 (Zero Hunger), SDG 14 (Life Below Water), and SDG 15 (Life on Land). EBP genomes can be directly used as primary indicators of genetic diversity (e.g., runs of homozygosity) or as a basis for cost-efficient technologies (e.g., eDNA monitoring) to monitor genetic diversity in populations over time (supporting the achievement of SDGs 14.4.1, 15.5.1, and 15.9.1). The latter will be extremely useful for conservation programs and management of marine and terrestrial resources, including animal and plant breeding (SDGs 2.5.1 and 2.5.2).

## The way forward

Based on our experience in Phase I of the EBP, in Phase II we propose sequencing 150,000 species—representing at least half of all known genera—to reference quality over 4 years. To build momentum for the remaining 90% of named eukaryotic species, we propose that Phase II includes collecting an additional 10% of the planet's fungi, protists, plants, and animals, biobanking them for sequencing in the early years of Phase III and safeguarding them for future research.

As the EBP proceeds, it will strive to be inclusive and equitable, recognizing the biodiversity richness of the Global South by establishing and sustaining genomics capacity, particularly in countries with high biodiversity. It is vitally important to build a genomic commons where data are accessible and can be shared frictionlessly. It is also critical to show and enhance the utility of the data by driving high-impact demonstration projects in conservation, biodiversity assessment, biopharmaceutical discovery, and bioproduct identification. By sequencing phylogenetic breadth as a driver, we will span the full eukaryotic tree of life, and by diving deep into specific taxonomic groups, or complete local ecosystems, we will demonstrate the riches that could come with Phase III sequencing of all eukaryotic biota.

The revised estimated cost of Phase II of the EBP is US$1.1 billion, including US$0.25 billion from the new FIF; this is down from the US$1.6 billion estimate made in 2018. Securing funding is one of the most pressing tasks faced by the EBP. By attracting funds that facilitate the sustainable establishment of biodiversity genomics within institutions in developing economies, we will be able to not only robustly deliver Phase II but also generate proof of concept for Phase III, "sequencing all life for the future of life". Understanding the origins and evolution of life on Earth is a human pursuit equivalent to understanding the origins and evolution of the universe. Beyond this, the wealth of practical applications that will emerge from sequencing eukaryotic life, ranging from conservation to climate adaptation and ecosystem preservation, likely makes the EBP the most ambitious and beneficial project in the history of science.

## Supplementary material

The data used to plot Figure 2 are available as a workbook in GoogleDocs [https://docs.google.com/spreadsheets/d/1zgnRMJ1L8Iujs2ItYrtFNviQeBDZWGRS9R0-IzE4YKQ/edit?usp=sharing] and as a zipped file of the three pages of the workbook in tab-separated values-format deposited in Zenodo: https://doi.org/10.5281/zenodo.12709327.

## Statements

### Author contributions

MB: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Project administration, Validation, Writing – original draft, Writing – review & editing.

HL: Conceptualization, Funding acquisition, Project administration, Validation, Writing – original draft, Writing – review & editing.

FD: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

RC: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

MD: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

RD: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

GF: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

NF: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

RG: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

PH: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

MH: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

KJH: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

KH: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

EJ: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

ML: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

KL-T: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

DM: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

FM: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

CM: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

AM: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

NM: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

SP: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

KD: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

VR: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

OR: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

LS: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

FT-N: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

TW: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

RW: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

EBP Community of Scientists: Writing – original draft, Writing – review & editing, Conceptualization, Funding acquisition.

## Members of the EBP Community of Scientists

Alexandre Aleixo, Miguel Allende, Jonas Astrin, Miklós Bálint, Katherine Barker, Ian Barnes, Kathy Belov, Giorgio Bertorelle, Iliana Bista, Mark Blaxter, Tomas Marques Bonet, Irus Braverman, Titus Brown, Jing Cai, Nicolette Caperello, Juan Carlos Castilla Rubio, Shu-Miaw Chaw, Haidan Chen, Lei Chen, Anna K. Childers, Robert Cook-Deegan, Montserrat Corominas, Shannon Corrigan, Keith A. Crandall, Andrew J. Crawford, Manuela da Silva, Robert Davey, Alice Dennis, Federica Di Palma, Richard Durbin, Jay Evans, Samuel Eziuzor, Olivier Fedrigo, Marc Palmada Flores, Giulio Formenti, Nico M. Franz, Arthur Georges, Anita Ghansah, M Thomas P Gilbert, Melissa Goldstein, Henry T. Greely, Roderic Guigo, Kevin Hackett, Neil Hall, Peter Harrison, Uljana Hesse, Katharina J. Hoff, Carolyn Hogg, Kerstin Howe, Maui Hudson, Ozede Nicholas Igiehon, Sachiko Isobe, Kjetill Sigurd Jakobsen, Erich Jarvis, Rebecca N Johnson, Steven Jones, Nathaniel K. Jue, Elinor K. Karlsson, Sally Katee, Paul Kersey, Jin-Hyoung Kim, Kevin Kocot, Tiffany Kosch, W. John Kress, Josiah Kuja, Shigehiro Kuraku, Malathi Lakshmikumaran, Mara Lawniczak, James Leebens-Mack, Harris Lewin, Qiye Li, Xueyan Li, Kerstin Lindblad-Toh, Xin Liu, Jose V. Lopez, Jianguo Lu, Jian Ma, Meike Mai, Roksana Majewska, Ntanganedzeni Mapholi, Luisa S. Marins, Fergal J. Martin, Debra JH Mathews, Camila J. Mazzoni, Catherine McCarthy, Ann M McCartney, Duane D. McKenna, Phillip Morin, Anne WT Muigai, Gene Myers, Ellis C. O'Neill, Rachel J. O'Neill, Sadye Paez, Adam Phillippy, Monica Poelchau, Kim D. Pruitt, Verena Ras, Arang Rhie, Emillio Righi, Gene Robinson, Lily Rodriguez, Hugues Roest Crollius, Cristina Roquet, Oliver A. Ryder, Sunil Kumar Sahu, Cynthia Saloma, Bernardo Santos, H. Bradley Shaffer, Timothy M. Shank, Taukondjo Shikongo, Heitor Shimizu, He Shunping, Pamela Soltis, Cibele Sotero-Caio, Ciara Stauton, David Swarbreck, Boping Tang, Francoise Thibaud-Nissen, Bashir Bolaji Tiamiyu, Andrew Torrance, Krystal S. Tsosie, Marcela Uliano-Silva, Andrew Veale, Sonja Vernes, Olga Vinnere Pettersson, Kun Wang, Robert Waterhouse, Claudia C. Weber, Jill Wegrzyn, Xiaofeng Wei, Regina Wetzer, Jeremy Wideman, Jason Williams, Linda Wong, Charlotte J. Wright, Joseph M. Yracheta, Guojie Zhang, and He Zhang.

Further information regarding the EBP Community of Scientists authors, including their EBP roles, affiliations, funding sources, and ORCID numbers, is available on Zenodo: https://zenodo.org/records/13762754 (doi: 10.5281/zenodo.13762753).

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

## Funding

The authors declared financial support was received for this work and/or its publication. The authors acknowledge funding from the following sources:

MB: Wellcome Trust grants 206194 and 218328, Gordon and Betty Moore Foundation grant CBMF8897, Biodiversity Genomics Europe Project funded by Horizon Europe under the Biodiversity, Circular Economy and Environment sectors (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 24.00054; and by the United Kingdom Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

HL: Arizona State University, United States, University of California, Davis, and the Monell Foundation.

FD: Genome BC, Canada.

RC: Wellcome Trust grants 206194 and 218328.

RD: Wellcome Trust grants 226458 and 207492.

PH: Wellcome Trust grants 222155 and 226458, Biodiversity Genomics Europe Project funded by Horizon Europe under the REA.B.3 sectors; co-funded by SERI under contract number 24.00054; and by UKRI under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme, Gordon and Betty Moore Foundation grant CBMF8897, the European Molecular Biology Laboratory (EMBL) Planetary Biology Transversal Theme and core funds.

KH: Wellcome Trust grants 206194 and 218328, Gordon and Betty Moore Foundation grant CBMF8897, Biodiversity Genomics Europe Project funded by Horizon Europe under the REA.B.3 sectors; co-funded by SERI under contract number 24.00054; and by UKRI under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

MH: LOEWE-Centre for Translational Biodiversity Genomics (LOEWE/1/10/519/03/03.001(0014)/52).

ML: Wellcome Trust grants 206194 and 218328, Biodiversity Genomics Europe Project funded by Horizon Europe under the REA.B.3 sectors; co-funded by SERI under contract number 24.00054; and by UKRI under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

CM: the Biodiversity Genomics Europe Project funded by Horizon Europe under the REA.B.3 sectors; co-funded by SERI under contract number 24.00054; and by UKRI under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

FT-N: the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH).

KP: NCBI, NLM, NIH.

SP: Monell Foundation.

RW: the Swiss National Science Foundation (SNSF) grant 202669 and the Biodiversity Genomics Europe Project funded by Horizon Europe under the REA.B.3 sectors; co-funded by SERI under contract number 24.00054; and by UKRI under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

The funders of the EBP Community of Scientists are listed in full in a document available on Zenodo: https://zenodo.org/records/13762754 (doi: 10.5281/zenodo.13762753). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

## Conflict of interest

The authors declared that this article was commissioned and the structure subsequently discussed and designed during a workshop organized and funded by Frontiers Media SA, held in Lausanne, Switzerland, on May 7–10, 2023. All the authors listed, except for the EBP Community of Scientists members, were participants in the workshop. The workshop organizer and funder was not involved in the study design, collection, analysis, interpretation of data, or the final decision to submit it for publication.

The authors declared that this work was conducted in the absence of financial relationships that could be construed as a potential conflict of interest.

The authors RG, DM, CM and RW declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The authors declared that no generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note



# References

1. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci* (2018) 115(17):4325–33. doi: 10.1073/pnas.1720115115

2. Blaxter M, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, et al. Why sequence all eukaryotes? *Proc Natl Acad Sci* (2022) 119(4):e2115636118. doi: 10.1073/pnas.2115636118

3. Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, et al. The Earth BioGenome Project 2020: starting the clock. *Proc Natl Acad Sci* (2022) 119(4):e2115635118. doi: 10.1073/pnas.2115635118

4. Bánki O, Roskov Y, Döring M, Ower G, Hernández Robles DR, Plata Corredor CA, et al. Catalogue of life checklist. Catalogue of Life Foundation (2023). doi: 10.48580/dgr6n

5. Sherkow JS, Barker KB, Braverman I, Cook-Deegan R, Durbin R, Easter CL, et al. Ethical, legal, and social issues in the Earth BioGenome Project. *Proc Natl Acad Sci* (2022) 119(4):e2115859119. doi: 10.1073/pnas.2115859119

6. Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, et al. Standards recommendations for the earth bioGenome project. *Proc Natl Acad Sci* (2022) 119(4):e2115639118. doi: 10.1073/pnas.2115639118

7. Mc Cartney AM, Anderson J, Liggins L, Hudson ML, Anderson MZ, TeAika B, et al. Balancing openness with Indigenous data sovereignty: an opportunity to leave no one behind in the journey to sequence all of life. *Proc Natl Acad Sci* (2022) 119(4): e2115860119. doi: 10.1073/pnas.2115860119

8. Stevens L, Martínez-Ugalde I, King E, Wagah M, Absolon D, Bancroft R, et al. Ancient diversity in host-parasite interaction genes in a model parasitic nematode. *Nat Commun* (2023) 14(1):7776. doi: 10.1038/s41467-023-43556-w

9. du Plessis SJ, Blaxter M, Koepfli K-P, Chadwick EA, Hailer F. Genomics reveals complex population history and unexpected diversity of Eurasian otters (Lutra lutra) in Britain relative to genetic methods. *Mol Biol Evol* (2023) 40(11):msad207. doi: 10.1093/molbev/msad207

10. Wright CJ, Stevens L, Mackintosh A, Lawniczak M, Blaxter M. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nat Ecol Evol* (2024) 8(4):777–90. doi: 10.1038/s41559-024-02329-4

11. Mulhair PO, Crowley L, Boyes DH, Lewis OT, Holland PWH. Opsin gene duplication in Lepidoptera: retrotransposition, sex linkage, and gene expression. *Mol Biol Evol* (2023) 40(11):msad241. doi: 10.1093/molbev/msad241

12. Mulhair PO, Holland PWH. Evolution of the insect Hox gene cluster: comparative analysis across 243 species. *Semin Cell Dev Biol* (2024) 152–153:4–15. doi: 10.1016/j.semcdb.2022.11.010

13. Mulhair PO, Crowley L, Boyes DH, Harper A, Lewis OT, Holland PW, et al. Diversity, duplication, and genomic organization of homeobox genes in Lepidoptera. *Genome Res* (2023) 33(1):32–44. doi: 10.1101/gr.277118.122

14. Zimmermann B, Montenegro JD, Robb SMC, Fropf WJ, Weilguny L, He S, et al. Topological structures and syntenic conservation in sea anemone genomes. *Nat Commun* (2023) 14(1):8270. doi: 10.1038/s41467-023-44080-7

15. Dussex N, Tørresen OK, van der Valk T, Le Moullec M, Veiberg V, Tooming-Klunderud A, et al. Adaptation to the High-Arctic island environment despite long-term reduced genetic variation in Svalbard reindeer. *iScience* (2023) 26(10):107811. doi: 10.1016/j.isci.2023.107811

16. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* (2021) 592(7856):737–46. doi: 10.1038/s41586-021-03451-0

17. Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, et al. Six reference-quality genomes reveal evolution of bat adaptations. *Nature* (2020) 583(7817):578–84. doi: 10.1038/s41586-020-2486-3

18. Blaxter M, Mieszkowska N, Palma FD, Holland P, Durbin R, Richards T. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc Natl Acad Sci* (2022) 119(4):e2115642118. doi: 10.1073/pnas.2115642118

19. Ebenezer TE, Muigai AWT, Nouala S, Badaoui B, Blaxter M, Buddie AG, et al. Africa: sequence 100,000 species to safeguard biodiversity. *Nature* (2022) 603(7901):388–92. doi: 10.1038/d41586-022-00712-4

20. McKenna V, Archibald JM, Beinart R, Dawson MN, Hentschel U, Keeling PJ, et al. The Aquatic Symbiosis Genomics Project: probing the evolution of symbiosis across the tree of life. *Wellcome Open Res* (2021) 6:254. doi: 10.12688/wellcomeopenres.17222.2

21. Corominas M, Marquès-Bonet T, Arnedo MA, Bayés M, Belmonte J, Escrivà H, et al. The Catalan initiative for the Earth BioGenome Project: contributing local data to global biodiversity genomics. *NAR Genom Bioinform* (2024) 6(3):lqae075. doi: 10.1093/nargab/lqae075

22. Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, et al. Landscape genomics to enable conservation actions: the California conservation genomics project. *J Hered* (2022) 113(6):577–88. doi: 10.1093/jhered/esac020

23. Mazzoni CJ, Ciofi C, Waterhouse RM. Biodiversity: an atlas of European reference genomes. *Nature* (2023) 619(7969):252. doi: 10.1038/d41586-023-02229-w

24. Challis R, Kumar S, Sotero-Caio C, Brown M, Blaxter M. Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Res* (2023) 8:24. doi: 10.12688/wellcomeopenres.18658.1

25. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* (2020) 2020:baaa062. doi: 10.1093/database/baaa062

26. Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. *GigaScience* (2021) 10(1):giaa153. doi: 10.1093/gigascience/giaa153

27. Howard C, Denton A, Jackson B, Bates A, Jay J, Yatsenko H, et al. On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species. bioRxiv [preprint] (2025). doi: 10.1101/2025.04.11.648334

28. Hartop E, Srivathsan A, Ronquist F, Meier R. Towards large-scale integrative taxonomy (LIT): resolving the data conundrum for dark taxa. *Syst Biol* (2022) 71(6):1404–22. doi: 10.1093/sysbio/syac033

29. Borba VH, Martin C, MaChado-Silva JR, Xavier SCC, de Mello FL, Iñiguez AM. Machine learning approach to support taxonomic species discrimination based on helminth collections data. *Parasit Vectors* (2021) 14(1):230. doi: 10.1186/s13071-021-04721-6

30. Wiens JJ. How many species are there on Earth? Progress and problems. *PloS Biol* (2023) 21(11):e3002388. doi: 10.1371/journal.pbio.3002388

31. May RM. Tropical arthropod species, more or less? *Science* (2010) 329(5987):41–2. doi: 10.1126/science.1191058

32. Blaxter M. Counting angels with DNA. *Nature* (2003) 421(6919):122–4. doi: 10.1038/421122a

33. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the ocean? *PloS Biol* (2011) 9(8):e1001127. doi: 10.1371/journal.pbio.1001127

34. Stevens L, Félix M-A, Beltran T, Braendle C, Caurcel C, Fausett S, et al. Comparative genomics of 10 new Caenorhabditis species. *Evol Lett* (2019) 3(2):217–36. doi: 10.1002/evl3.110

35. Slos D, Sudhaus W, Stevens L, Bert W, Blaxter M. Caenorhabditis monodelphis sp. n.: defining the stem morphology and genomics of the genus Caenorhabditis. *BMC Zool* (2017) 2:1–5. doi: 10.1186/s40850-017-0013-2

36. Boehm MMA, Cronk QCB. Dark extinction: the problem of unknown historical extinctions. *Biol Lett* (2021) 17(3):20210007. doi: 10.1098/rsbl.2021.0007

37. Hogg CJ, Ottewell K, Latch P, Rossetto M, Biggs J, Gilbert A, et al. Threatened Species Initiative: empowering conservation action using genomic resources. *Proc Natl Acad Sci* (2022) 119(4):e2115643118. doi: 10.1073/pnas.2115643118

38. Fiedler PL, Erickson B, Esgro M, Gold M, Hull JM, Norris JM, et al. Seizing the moment: the opportunity and relevance of the California Conservation Genomics Project to state and federal conservation policy. *J Hered* (2022) 113(6):589–96. doi: 10.1093/jhered/esac046

39. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: an evolving community-developed biodiversity data standard. *PloS One* (2012) 7(1):e29715. doi: 10.1371/journal.pone.0029715

40. Gries C, Gilbert EE, Franz NM. Symbiota – a virtual platform for creating voucher-based biodiversity information communities. *Biodivers Data J* (2014) 2):e1114. doi: 10.3897/BDJ.2.e1114

41. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* (2016) 3:160018. doi: 10.1038/sdata.2016.18

42. Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, et al. The CARE principles for indigenous data governance. *Data Sci J* (2020) 19(1):43. doi: 10.5334/dsj-2020-043

43. Emerson BC, Borges PAV, Cardoso P, Convey P, deWaard JR, Economo EP, et al. Collective and harmonized high throughput barcoding of insular arthropod biodiversity: toward a Genomic Observatories Network for islands. *Mol Ecol* (2023) 32(23):6161–76. doi: 10.1111/mec.16683

44. Chua PYS, Bourlat SJ, Ferguson C, Korlevic P, Zhao L, Ekrem T, et al. Future of DNA-based insect monitoring. *Trends Genet* (2023) 39(7):531–44. doi: 10.1016/j.tig.2023.02.012

45. Corrales C. *Biodiversity biobanking – a handbook on protocols and practices*. Sofia: Pensoft (2023). doi: 10.3897/ab.e101876

46. Paez S, Kraus RHS, Shapiro B, Gilbert MT, Jarvis ED, Vertebrate Genomes Project Conservation Group, et al. Reference genomes for conservation. *Science* (2022) 377(6604):364–6. doi: 10.1126/science.abm8127

47. Ryder OA, Onuma M. Viable cell culture banking for biodiversity characterization and conservation. *Annu Rev Anim Biosci* (2018) 6(1):83–98. doi: 10.1146/annurev-animal-030117-014556

48. Dahn HA, Mountcastle J, Balacco J, Winkler S, Bista I, Schmitt AD, et al. Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *GigaScience* (2022) 11:giac068. doi: 10.1093/gigascience/giac068

49. Roberts NG, Gilmore MJ, Struck TH, Kocot KM. Multiple displacement amplification facilitates SMRT sequencing of microscopic animals and the genome of the gastrotrich Lepidodermella squamata (Dujardin 1841). *Genome Biol Evol* (2024) 16(12):evae254. doi: 10.1093/gbe/evae254

50. Lee Y-C, Ke H-M, Liu Y-C, Lee H-H, Wang M-C, Tseng Y-C, et al. Single-worm long-read sequencing reveals genome diversity in free-living nematodes. *Nucleic Acids Res* (2023) 51(15):8035–47. doi: 10.1093/nar/gkad647

51. Vancaester E, Blaxter M. Phylogenomic analysis of Wolbachia genomes from the Darwin Tree of Life biodiversity genomics project. *PloS Biol* (2023) 21(1):e3001972. doi: 10.1371/journal.pbio.3001972

52. Cheng H, Asri M, Lucas J, Koren S, Li H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat Methods* (2024) 21(6):967–70. doi: 10.1038/s41592-024-02269-8

53. Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV, Formenti G, Fungtammasan A, et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods* (2022) 19(6):687–95. doi: 10.1038/s41592-022-01440-3

54. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* (2023) 41(10):1474–82. doi: 10.1038/s41587-023-01662-6

55. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science* (2022) 376(6588):44–53. doi: 10.1126/science.abj6987

56. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* (2022) 40(9):1332–5. doi: 10.1038/s41587-022-01261-x

57. Massana R, López-Escardó D. Metagenome assembled genomes are for eukaryotes too. *Cell Genom* (2022) 2(5):100130. doi: 10.1016/j.xgen.2022.100130

58. Mangot J-F, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* (2017) 7(1):41498. doi: 10.1038/srep41498

59. Sieracki ME, Poulton NJ, Jaillon O, Wincker P, de Vargas C, Rubinat-Ripoll L, et al. Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci Rep* (2019) 9(1):6025. doi: 10.1038/s41598-019-42487-1

60. Labarre A, López-Escardó D, Latorre F, Leonard G, Bucchini F, Obiol A, et al. Comparative genomics reveals new functional insights in uncultured MAST species. *ISME J* (2021) 15(6):1767–81. doi: 10.1038/s41396-020-00885-8

61. Wang Y, Yu J, Jiang M, Lei W, Zhang X, Tang H. Sequencing and assembly of polyploid genomes. *Polyploidy Methods Protoc* (2023) 2545:429–58. doi: 10.1007/978-1-0716-2561-3_23

62. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit - interactive quality assessment of genome assemblies. *G3 (Bethesda)* (2020) 10(4):1361–74. doi: 10.1534/g3.119.400908

63. Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* (2024) 25(1):60. doi: 10.1186/s13059-024-03198-7

64. Larivière D, Abueg L, Brajuka N, Gallardo-Alba C, Grüning B, Ko BJ, et al. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nat Biotechnol* (2024) 42(3):367–70. doi: 10.1038/s41587-023-02100-3

65. Guigó R. Genome annotation: from human genetics to biodiversity genomics. *Cell Genom* (2023) 3(8):100375. doi: 10.1016/j.xgen.2023.100375

66. Rigden DJ, Fernández XM. The 2024 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res* (2024) 52(D1):D1–9. doi: 10.1093/nar/gkad1173

67. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res* (2024) 52(D1):D33–43. doi: 10.1093/nar/gkad1044

68. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* (2016) 44(D1):D733–45. doi: 10.1093/nar/gkv1189

69. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* (2024) 21(7):1349–63. doi: 10.1038/s41592-024-02298-3

70. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* (2021) 3(1):lqaa108. doi: 10.1093/nargab/lqaa108

71. Bruna T, Lomsadze A, Borodovsky M. GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data (2023). Available at: https://www.genome.org/cgi/doi/10.1101/gr.278373.123

72. Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, et al. Integrating gene annotation with orthology inference at scale. *Science* (2023) 380(6643): eabn3107. doi: 10.1126/science.abn3107

73. Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, et al. Helixer–*de novo*prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. bioRxiv [preprint] (2023). doi: 10.1101/2023.02.06.527280

74. Haj Abdullah Alieh L, Cardoso de Toledo B, Hadarovich A, Toth-Petroczy A, Calegari F. Characterization of alternative splicing during mammalian brain development reveals the extent of isoform diversity and potential effects on protein structural changes. *Biol Open* (2024) 13(10):bio061721. doi: 10.1242/bio.061721

75. Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat Biotechnol* (2024) 42(4):582–6. doi: 10.1038/s41587-023-01815-7

76. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol* (2019) 1962:227–45. doi: 10.1007/978-1-4939-9173-0_14

77. Huang N, Li H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* (2023) 39(10):btad595. doi: 10.1093/bioinformatics/btad595

78. Altenhoff AM, Warwick Vesztrocy A, Bernard C, Train C-M, Nicheperovich A, Prieto Baños S, et al. OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Res* (2024) 52(D1):D513–21. doi: 10.1093/nar/gkad1020

79. Nevers Y, Rossier V, Train CM, Altenhoff A, Dessimoz C, Glover N. Quality assessment of gene repertoire annotation with OMArk. *Nat Biotechnol* (2025) 43(1):124–33. doi: 10.1101/2022.11.25.517970

80. Simakov O, Marlétaz F, Yue J-X, O'Connell B, Jenkins J, Brandt A, et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* (2020) 4(6):820–30. doi: 10.1038/s41559-020-1156-z

81. Damas J, Corbo M, Kim J, Turner-Maier J, Farré M, Larkin DM, et al. Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc Natl Acad Sci* (2022) 119(40):e2209139119. doi: 10.1073/pnas.2209139119

82. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res* (2016) 44(D1):D81–9. doi: 10.1093/nar/gkv1272

83. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* (2003) 299(5611):1391–4. doi: 10.1126/science.1081331

84. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* (2020) 587(7833):240–5. doi: 10.1038/s41586-020-2876-6

85. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, et al. Evolutionary constraint and innovation across hundreds of placental mammals. *Science* (2023) 380(6643):eabn3943. doi: 10.1126/science.abn3943

86. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) 596(7873):583–9. doi: 10.1038/s41586-021-03819-2

87. Jaquier M, Albouy C, Bach W, Waldock C, Marques V, Maire E, et al. Environmental DNA recovers fish composition turnover of the coral reefs of West Indian Ocean islands. *Ecol Evol* (2024) 14(5):e11337. doi: 10.1002/ece3.11337

88. Kjær KH, Winther Pedersen M, De Sanctis B, De Cahsan B, Korneliussen TS, Michelsen CS, et al. A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* (2022) 612(7939):283–91. doi: 10.1038/s41586-022-05453-y

89. Clare EL, Economou CK, Bennett FJ, Dyer CE, Adams K, McRobie B, et al. Measuring biodiversity from DNA in the air. *Curr Biol* (2022) 32(3):693–700.e5. doi: 10.1016/j.cub.2021.11.064

90. Alexander H, Hu SK, Krinos AI, Pachiadaki M, Tully BJ, Neely CJ, et al. Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *MBio* (2023) 14(6):e0167623. doi: 10.1128/mbio.01676-23

91. Sunagawa S, Acinas SG, Bork P, Bowler C, Eveillard D, Gorsky G, et al. Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* (2020) 18(8):428–45. doi: 10.1038/s41579-020-0364-5

92. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012) 489(7414):57–74. doi: 10.1038/nature11247

93. Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al. Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Res* (2014) 24(7):1209–23. doi: 10.1101/gr.159384.113

94. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* (2010) 330(6012):1775–87. doi: 10.1126/science.1196914

95. Park J-E, Botting RA, Domínguez Conde C, Popescu D-M, Lavaert M, Kunz DJ, et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* (2020) 367(6480):eaay3224. doi: 10.1126/science.aay3224

96. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. *Science* (2020) 370(6518):eaba7721. doi: 10.1126/science.aba7721

97. Anonymous. Evolution at the cellular level. *Nat Ecol Evol* (2023) 7(8):1155–6. doi: 10.1038/s41559-023-02133-6

98. Lannelongue L, Aronson HG, Bateman A, Birney E, Caplan T, Juckes M, et al. GREENER principles for environmentally sustainable computational science. *Nat Comput Sci* (2023) 3(6):514–21. doi: 10.1038/s43588-023-00461-y

99. Lannelongue L, Grealey J, Bateman A, Inouye M. Ten simple rules to make your computing more environmentally sustainable. *PloS Comput Biol* (2021) 17(9): e1009324. doi: 10.1371/journal.pcbi.1009324

100. Grealey J, Lannelongue L, Saw W-Y, Marten J, Méric G, Ruiz-Carmona S, et al. The carbon footprint of bioinformatics. *Mol Biol Evol* (2022) 39(3):msac034. doi: 10.1093/molbev/msac034

101. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* (2020) 587(7833):246–51. doi: 10.1038/s41586-020-2871-y

102. Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* (2006) 34(suppl_1):D572–80. doi: 10.1093/nar/gkj118

103. Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, et al. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst Biol* (2019) 68(2):365–9. doi: 10.1093/sysbio/syy054

104. Crawford K. Generative AI's environmental costs are soaring - and mostly secret. *Nature* (2024) 626(8000):693. doi: 10.1038/d41586-024-00478-x

105. Lannelongue L, Grealey J, Inouye M. Green algorithms: quantifying the carbon footprint of computation. *Adv Sci (Weinh)* (2021) 8(12):2100707. doi: 10.1002/advs.202100707

106. Helmy M, Awad M, Mosa KA. Limited resources of genome sequencing in developing countries: challenges and solutions. *Appl Transl Genom* (2016) 9:15–9. doi: 10.1016/j.atg.2016.03.003

107. United States National Human Genome Research Institute. Human Genome Project fact sheet [online] (2024). Available at: https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

108. Dreier C. How much does the James Webb Space Telescope cost? *The Planetary Society* (2021). Available at: https://www.planetary.org/articles/cost-of-the-jwst

109. Knapp A. How much does it cost to find a Higgs boson? *Forbes* (2012). Available at: https://www.forbes.com/sites/alexknapp/2012/07/05/how-much-does-it-cost-to-find-a-higgs-boson/

110. Cresswell I, Janke T, Johnston EL, et al. *Australia State of the Environment 2021*. Canberra, ACT: Australian Government Minister for the Environment (2021). doi: 10.26194/f1rh-7r05

111. United Nations Environment Programme. Convention on Biological Diversity. Decision adopted by the Conference of the Parties to the Convention on Biological Diversity. 15/4. Kunming-Montreal Global Biodiversity Framework [decision CBD/COP/DEC/15/5]. Conference of the Parties to the Convention on Biological Diversity Fifteenth meeting – Part II (2022). Available at: https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-04-en.pdf

112. United Nations General Assembly. United Nations declaration on the rights of indigenous peoples. United Nations General Assembly 107th plenary meeting (2007). Available at: https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf.

113. United Nations. United Nations Convention on the Law of the Sea. Montego Bay: UN (1982). Available at: https://www.un.org/depts/los/convention_agreements/convention_overview_convention.htm.

114. Food and Agriculture Organization of the United Nations. United States International Treaty on Plant Genetic Resources for Food and Agriculture. Thirty-First Session of the Conference of the Food and Agriculture Organization of the United Nations. Madrid: UN (2001). Available at: https://www.fao.org/plant-treaty/en/.

115. World Health Organization. Pandemic prevention, preparedness and response agreement [online]. (2025). Available at: https://www.who.int/news-room/questions-and-answers/item/pandemic-prevention–preparedness-and-response-accord.