



























































The Earth BioGenome Project 2020: Starting the clock

Harris A. Lewin^{a,b,1} , Stephen Richards^c , Erez Lieberman Aiden^d , Miguel L. Allende^{e,f} , John M. Archibald^g , Miklós Bálint^{h,i} , Katharine B. Barker^j , Bridget Baumgartner^k , Katherine Belov^l , Giorgio Bertorelle^m , Mark L. Blaxterⁿ , Jing Cai^o , Nicolette D. Caperello^c , Keith Carlson^p , Juan Carlos Castilla-Rubio^q , Shu-Miaw Chaw^r , Lei Chen^o , Anna K. Childers^s , Jonathan A. Coddington^t , Dalia A. Conde^{u,v} , Montserrat Corominas^{w,x} , Keith A. Crandall^{y,z} , Andrew J. Crawford^{aa} , Federica DiPalma^{bb} , Richard Durbin^{cc,dd} , ThankGod E. Ebenezer^{ee} , Scott V. Edwards^{ff,gg} , Olivier Fedrigo^{hh} , Paul Flicek^{dd,ii} , Giulio Formenti^{jj} , Richard A. Gibbs^{kk} , M. Thomas P. Gilbert^{ll,mmm} , Melissa M. Goldsteinⁿⁿ , Jennifer Marshall Graves^{oo,pp} , Henry T. Greely^{qq} , Igor V. Grigoriev^{rr,ss} , Kevin J. Hackett^{tt} , Neil Hall^{uu} , David Haussler^{vv,ww} , Kristofer M. Helgen^{xx} , Carolyn J. Hogg^l , Sachiko Isobe^{yy} , Kjetill Sigurd Jakobsen^{zz} , Axel Janke^h , Erich D. Jarvis^{hh,ww} , Warren E. Johnson^{aaa,bbb} , Steven J. M. Jones^{ccc} , Elinor K. Karlsson^{ddd,eee} , Paul J. Kersey^{fff} , Jin-Hyoung Kim^{ggg} , W. John Kress^{hhh} , Shigehiro Kuraku^{iii,jjj} , Mara K. N. Lawniczak^k , James H. Leebens-Mack^{kkk} , Xueyan Li^{lll} , Kerstin Lindblad-Toh^{eee,mmm} , Xin Liuⁿⁿⁿ , Jose V. Lopez^{ooo,ppp} , Tomas Marques-Bonet^{qqq,rrr,sss,ttt} , Sophie Mazard^{uuu} , Jonna A. K. Mazet^{vvv} , Camila J. Mazzoni^{www,xxx} , Eugene W. Myers^{yyy} , Rachel J. O'Neill^{zzz,aaaa} , Sadye Paez^{hh} , Hyun Park^{bbbb} , Gene E. Robinson^{cccc} , Cristina Roquet^{ddd,eee} , Oliver A. Ryder^{fff,ggg} , Jamal S. M. Sabir^{hhhh,iiii} , H. Bradley Shaffer^{jjjj,kkkk} , Timothy M. Shank^{llll} , Jacob S. Sherkow^{cccc,mmmm} , Pamela S. Soltis^{nnnn,oooo} , Boping Tang^{pppp} , Leho Tedersoo^{qqqq,rrrr} , Marcela Uliano-Silvaⁿ , Kun Wang^o , Xiaofeng Weiⁿⁿⁿⁿ , Regina Wetzler^{ssss,tttt} , Julia L. Wilson^{dd} , Xun Xuⁿⁿⁿ , Huanming Yangⁿⁿⁿ , Anne D. Yoder^{vvvv,uuuu} , and Guojie Zhang^{wwww,xxxx,lll,nnn} 

November 2020 marked 2 y since the launch of the Earth BioGenome Project (EBP), which aims to sequence all known eukaryotic species in a 10-y timeframe. Since then, significant progress has been made across all aspects of the EBP roadmap, as outlined in the 2018 article describing the project's goals, strategies, and challenges (1). The launch phase has ended and the clock has started on reaching the EBP's major milestones. This Special Feature explores the many facets of the EBP, including a review of progress, a description of major scientific goals, exemplar projects, ethical legal and social issues, and applications of biodiversity genomics. In this Introduction, we summarize the current status of the EBP, held virtually October 5 to 9, 2020, including recent updates through February 2021. References to the nine Perspective articles included in this Special Feature are cited to guide the reader toward deeper understanding of the goals and challenges facing the EBP.

It is urgent that the EBP move forward. The year 2020 marked a global failure in meeting any of the 20 "Aichi goals" for the preservation of wildlife and

ecosystems (2). The International Union for Conservation of Nature now counts more than 35,000 (28%) of all surveyed species of plants and animals as threatened with extinction (3). The Earth may lose 50% of its biodiversity by the end of this century if nothing is done to mitigate the anthropogenic factors that drive species to extinction and destroy the health of global ecosystems that sustain human existence (2). Degradation of aquatic and terrestrial ecosystems has continued unabated, and we may soon face the possibility of massive ecosystem collapse on a global scale.

Such a collapse would have an enormous impact not only on biodiversity, but also on global political stability, and might ultimately affect the survival of our own species. Biological diversity underpins ecosystem services: that is, those services provided by nature that generate food, clean air and water, regulation of critical environmental processes and biogeochemical cycles, and are the basis for deep cultural and esthetic ties between humans and the natural world. Biodiversity is also foundational for the rapidly growing global bioeconomy that exceeds

Author contributions: H.A.L. and S.R. analyzed data; and H.A.L., S.R., E.L.A., M.L.A., J.M.A., M.B., K.B.B., B.B., K.B., G.B., M.L.B., J.C., N.D.C., K.C., J.C.-R., S.-M.C., L.C., A.K.C., J.A.C., D.A.C., M.C., K.A.C., A.J.C., F.D., R.D., T.E.E., S.V.E., O.F., P.F., G.F., R.A.G., M.T.P.G., M.M.G., J.M.G., H.T.G., I.V.G., K.J.H., N.H., D.H., K.M.H., C.J.H., S.I., K.S.J., A.J., E.D.J., W.E.J., S.J.M.J., E.K.K., P.J.K., J.-H.K., W.J.K., S.K., M.K.N.L., J.H.L.-M., X. Li, K.L.-T., X. Liu, J.V.L., T.M.-B., S.M., J.A.K.M., C.J.M., E.W.M., R.J.O., S.P., H.P., G.E.R., C.R., O.A.R., J.S.M.S., H.B.S., T.M.S., J.S.S., P.S.S., B.T., L.T., M.U.-S., K.W., X.W., R.W., J.L.W., X.X., H.Y., A.D.Y., and G.Z. wrote the paper.

The authors declare no competing interest.

This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹To whom correspondence may be addressed. Email: lewin@ucdavis.edu.

This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2115635118/-/DCSupplemental>.

Published January 18, 2022.

\$500 billion each year in just the United States and European Union (4, 5), and it is essential for sustainable food security (6). If biodiversity disappears, so too will the potential for a new inclusive bioeconomy that is possible through a combination of genomics, computational biology, and synthetic biology, identified by the World Economic Forum as key to the fourth Industrial Revolution (7) and estimated to be worth up to US \$3 to 5 trillion per annum (8).

The year 2020 will also be remembered in history as the beginning of the COVID-19 pandemic. The virus that causes COVID-19, SARS-CoV-2, evolved from a bat betacoronavirus (9), possibly finding its way into the human population through an intermediate host that has yet to be identified (10). Spillover of SARS-CoV-2 infection to wildlife, pets, and captive-bred animals demonstrates the interconnectedness of life on Earth, reinforcing the One Health concept that all organisms are interdependent: the health of one impacts the health of all (11). A One Health approach to addressing the biodiversity crisis critically relies on supporting infrastructures, such as the genomic infrastructure that can be provided by the EBP and affiliated projects. The economic disaster and devastating human death toll caused by the pandemic illustrate just how critical it is to have knowledge of potential human pathogens and their hosts before such events arise (12). Clearly, DNA sequence information on the virus and its potential hosts has helped the world to manage and hopefully soon contain COVID-19. Similarly, creating a library of DNA sequences for all known eukaryotic life can contribute critical data necessary to generate effective tools for preventing biodiversity loss and pathogen spread, monitoring and protecting ecosystems, and enhancing ecosystem services [see The Darwin Tree of Life Project Consortium, this issue (13)]. The EBP's proactive stance on understanding the ethical, legal, and social issues surrounding the project will also inform recommendations on access and commercial benefit sharing, equity, and inclusion in the biodiversity genomics community and in indigenous communities within the world's most biodiverse countries [see McCartney et al., this issue (14)].

Organization and Governance

A critical role of the EBP organization is to: develop and promote standards for the scalable production of reference-quality genomes; dissemination of best practices; coordination of sequencing, annotation, data analysis, and training activities; public accessibility of data; and communications about the project's progress. To accomplish these goals, the EBP was established as an international network-of-networks: organizations that specialize in sample acquisition and vouchering; technology centers for sequencing, assembly, and annotation; and affiliated projects with deep expertise with specific taxonomic groups, biomes, and ecosystems (Box 1). In addition, the EBP develops ethical standards for project participation, data sharing, access and benefit sharing of intellectual property derived from whole-genome sequencing [see Sherkow et al., this issue

(15)], and promotes programs for diversity, equity, inclusion, and justice among the project's participants. The EBP Member Institutions and Affiliated Projects are committed to open data access and compliance with the principles of Access and Benefits Sharing under the Convention on Biological Diversity and the Nagoya Protocol (16). The EBP communicates progress and information about the project through its website (<https://www.earthbiogenome.org>), its Twitter handle (@EBPgenome), and other social media accounts, currently with more than 2,000 followers.

The EBP Secretariat is located at the University of California, Davis, and operates under a Memorandum of Understanding between participating institutions available at the EBP website, <https://www.earthbiogenome.org>. The representatives of member institutions have adopted an interim governance structure (*SI Appendix*, Fig. 1).

An interim governance committee is in place, The Earth BioGenome Project Working Group, which as of February 2021 consists of one representative of each of the 43 Memorandum of Understanding-signing institutions (see list on the EBP website, <https://www.earthbiogenome.org>) and 44 affiliated projects (*Dataset S1*; brief summaries of 21 affiliated projects can be found in *SI Appendix*), with membership up 121% and 153%, respectively, since 2018. The Chair of the EBP Working Group coordinates the activities of all the working committees and conducts extensive international outreach for promoting collaboration between member institutions and affiliated projects, implementation of standards, assisting the formation of national and regional projects, and coordination of activities across the EBP network-of-networks. The International Science Committee consists of a chairperson and five subcommittees that are responsible for standards development in the following areas: sample collection and processing, sequencing and assembly, annotation, information technology and informatics, and data analyses. Committee reports are available on the EBP website (<https://www.earthbiogenome.org>) and summarized in this issue. The EBP plans to formally adopt a permanent governance structure in 2021. Those institutions and projects that are interested in joining the EBP should contact the Secretariat using the EBP website for further information.

The EBP's Committee on Ethical, Legal, and Social Issues (ELSI), established in 2020, makes recommendations to the EBP Working Group on legal obligations relating to the Nagoya Protocol on Access and Benefit Sharing; ethical considerations relating to collection of samples, societal concerns, and biosecurity; and collaboration standards (e.g., sample information, digital sequence information, intellectual property, authorship and publication guidelines). The committee's outline of the ELSI issues facing the EBP can be found in this issue (15). A Committee on Diversity, Equity, Inclusion, and Justice (DEIJ) was approved recently by the EBP Working Group. DEIJ recommendations will be based on participatory approaches with fair treatment and meaningful involvement of all people to define processes and practices for creating a welcoming, inclusive, and supportive biodiversity genomics community.

Box 1. The EBP international network-of-networks functions to support the three proposed phases of the EBP

- Phase I: An annotated reference genome for one representative of each taxonomic family of eukaryotes (~9,400 species) in 3 y.
- Phase II: Reference genomes for one representative of each genus (~180,000 species) in years 4 to 7.
- Phase III: Reference genomes for remaining ~1.65 million known eukaryotic species in the final 3 y of the project.

Global Status of Biodiversity Sequencing

Our current ability to investigate the diversity and evolution of Earth's biota is severely constrained by the absence of high-quality genome sequences for most of the species on the eukaryotic tree of life. There are now ~1.84 million taxonomically classified eukaryotic species, but the estimated number of eukaryotic species is 12 to 15 million, including 8.1 million plants and animals (17). The EBP aims to sequence all classified species and to facilitate the discovery and classification of new species. As of March 4, 2021, the International Nucleotide Sequence Database Collaboration (INSDC) contained whole-genome DNA sequence information on 6,480 unique species, representing 81.4% of eukaryotic phyla, 64.7% of classes, 40.1% of orders, 15.5% of families, 2.3% of genera, and just 0.43% of all species (Fig. 1).

However, the assembly quality of these 6,480 species' genomes varies greatly (*SI Appendix*, Fig. 1). A majority (63.1%) of the assemblies falls into the short-read draft category, with contig N50 < 100 kb and scaffold N50 < 10 Mb. A relatively small number of the draft-quality assemblies have achieved greater contiguity using scaffolding methods, such as Hi-C, linked-reads, and optical maps (19). The number of unique eukaryotic species with whole-genome assemblies has more than doubled since 2018 (Fig. 2), most of which are short-read draft quality. The number of reference-quality chromosome-scale assemblies of unique species representing taxonomic families nearly tripled since 2018, from 210 to 583. EBP-affiliated projects produced about half of these new reference-quality assemblies (see below), demonstrating the efficacy of shared goals and standards.

Progress of the EBP toward Phase I Goals

The past 2 y represent the start-up phase of the EBP. The major activities of the international EBP network-of-networks include: the development of standards; the evaluation of strategies for producing reference genomes; organizing regional, national, and transnational projects; and building communities through regular working committee meetings and an annual conference. The "Biodiversity Genomics 2020" conference was held virtually and had 3,000 registrants from 89 countries. The full recording

of the meeting is available (20). The EBP is also developing new initiatives in training, broadening diversity and inclusion in project leadership, and building support for project funding from government agencies and private foundations around the world.

The current line-up of 43 EBP-affiliated projects cover most of the major groups of eukaryotic taxa and represent access to tens of thousands of high-quality samples in museum collections and those from field biologists. The geographic diversity of the institutional members and affiliated projects cover 21 countries across all continents except Antarctica. The first African nodes have recently come on line in 2021 as part of the Africa BioGenome Project. The EBP also aims to expand member institutions and affiliated projects across additional biodiverse regions of the world, including the Indian subcontinent, Southeast Asia, and South America [for example, see Huddart et al. (21), this issue]. With high endemism concentrated in these regions, the ultimate success of the EBP requires building scientific capacity in developing nations and respecting national laws for access and benefit sharing.

EBP-affiliated projects, such as the Darwin Tree of Life Project [see The Darwin Tree of Life Project Consortium, this issue (22)], The Vertebrate Genomes Project, 1000 Fungal Genomes Project, B10K (sequencing 10,000 bird species), and others have led the way in producing publicly accessible high-quality genomes (Table 1 and *SI Appendix*). A Perspective on sequencing of plant genomes is included in this special issue (23). EBP-affiliated sequencing centers around the world are now coming online for the production of reference genomes using a simplified pipeline consisting of long reads and Hi-C (or equivalent), and other scaffolding methods, such as optical mapping, and public domain assembly tools, such as the recently developed hifiasm for generating long-read-based contigs (24) and SALSA for generating Hi-C scaffolds (25). This simplified approach, within the reach of most EBP-affiliated laboratories, yields chromosome-scale assemblies that meet the EBP standard (see above).

The EBP-affiliated projects have sequenced the genomes of 1,719 eukaryotic species, all of which have assemblies deposited in public domain databases (Table 1 and *Dataset S1*). Of these,

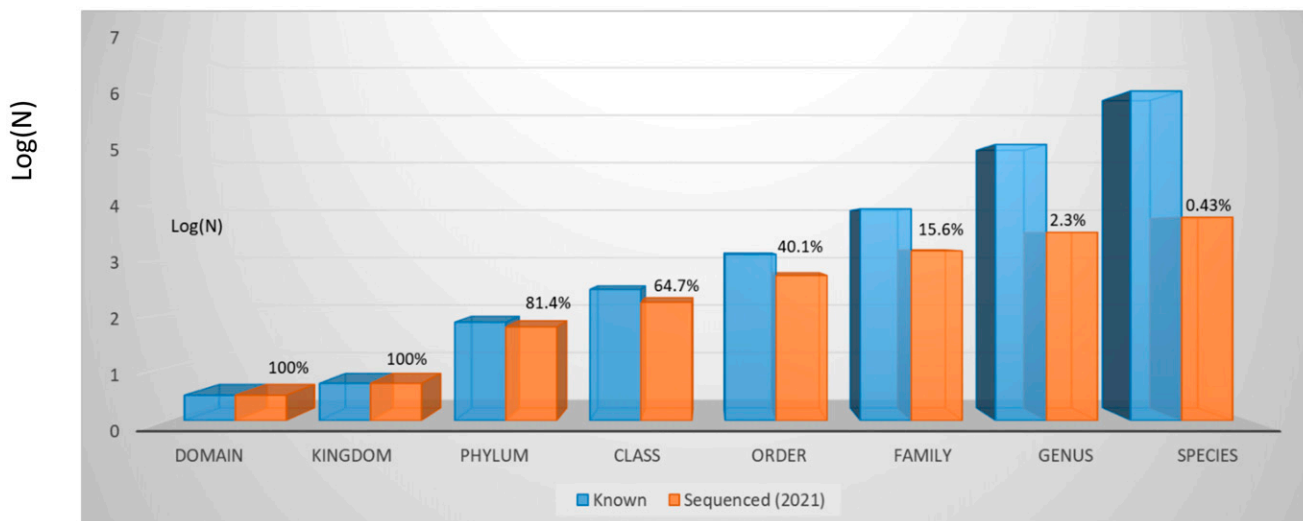


Fig. 1. Global progress in whole-genome sequencing across all eukaryotic taxonomic levels. Data source: National Center for Biotechnology Information, March 4, 2021 (18).

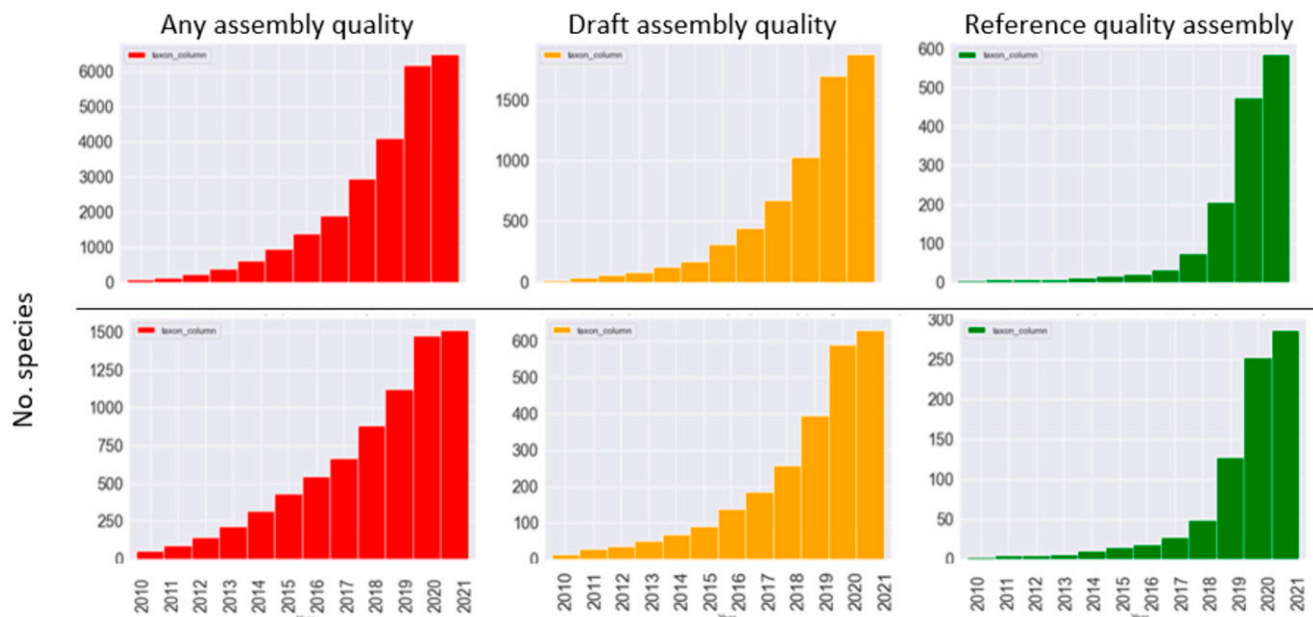


Fig. 2. Year-over-year progress in whole genome sequencing for all eukaryotic taxa (Upper) and family-level (Lower) eukaryotic taxa, 2010 to March 4, 2021. The metrics for draft and reference quality assemblies are given in the text.

316 are reference-quality genomes, constituting ~50% of all the genomes in the INSDC that meet the EBP reference standard. Furthermore, these already represent more than 200 taxonomically distinct nonredundant families. Thus, in the start-up phase, EBP-affiliated projects have sequenced ~2% of extant eukaryotic families to reference-level quality. There are 3,021 family-level reference genomes expected to be completed in 2021. Thus, by the end of 2021, the first full year of the project, we project that ~3,200 taxonomic families will have been sampled with at least one reference genome, corresponding to 34% completion of the EBP Phase I goal.

Other large-scale initiatives with complementary goals have joined EBP as affiliated projects. These include BIOSCAN (26) and the Global Virome Project (27). BIOSCAN aims to DNA barcode every eukaryotic species on Earth, which will be critical to the EBP sample vouchering process and for accessing rare samples for sequencing. Partnership with the Global Virome Project creates an exciting avenue to identify potentially pathogenic

viruses linked with their host species and for codevelopment of biosurveillance strategies (12). Integrated high-level coordination between these projects will have synergistic effects on biodiversity science and societal outcomes. A broad perspective on the scientific challenges and opportunities enabled by large-scale comparative genomics is provided by Stephan et al., this issue (28).

The Challenges Ahead

Although the number of reference-quality genomes at the family level tripled from 2018 to March 4, 2021 (Fig. 2), the EBP will have to produce nearly 3,000 genomes per year to meet the EBP Phase I goal of producing at least one reference genome from all ~9,400 eukaryotic families in 3 y. The main challenges in meeting this target are given in Box 2.

To meet the EBP Phase I goal, the EBP network-of-networks will need to produce nine genomes per day, 365 d/y. Is this feasible? The Wellcome Sanger Institute alone plans to produce 1,500

Table 1. Progress of EBP affiliated projects in whole-genome sequencing and the production of reference genomes

Project name	No. of species	No. of references	No. of families with reference	No. of references 2021	No. of drafts 2021
1000 Fungal Genomes	663	20	10	10	100
B10K (birds)	400	32	29	0	400
Zoonomia (mammals)	130	0	0	0	0
VGP (vertebrates)	128	129	119	200	0
i5K (arthropods)	86	2	2	2	8
Darwin Tree of Life	71	71	14	1,500	0
Tree of Life, Sanger	50	50	50	300	0
LOEWE Center	43	43	43	84	95
Ungulates Genome Project	41	0	0	4	6
CanSeq150	36	21	0	0	0
10KP (plants)	21	2	2	16	42
All other	80	42	30	905	1,505
Total	1,719	412	316	3,021	2,156

All tabulated genomes in the first three columns have been submitted to the INSDC or other public domain databases. Numbers in the last two columns are projected additional species genomes for 2021. A complete table with INSDC project identifiers can be found in [Dataset S1](#). Totals include some species that overlap between projects.

Box 2. Challenges in meeting EBP goals

- Sourcing, vouchering, and permitting thousands of specimens globally
- High molecular weight DNA and RNA isolation at scale
- Sequencing capacity and throughput
- Assembly and curation at scale
- Annotation at scale
- Managing data flows in the context of international current and future data access and sharing regulations
- Whole genome alignments at scale
- Comparative genomic analysis, population genomics, and data visualization at scale

reference-quality genomes in 2021 as part of the Darwin Tree of Life Project, corresponding to four genomes per day. As presented in Table 1, the Institute is already well on its way to achieving this goal in the coming year. The Vertebrate Genomes Project aims to produce six genomes per week to complete its goal of producing high-quality assemblies for species representing 260 vertebrate lineages separated by 50 million y or more from a common ancestor (19), by the end of 2021. With current technology and funded commitments for 2021 by EBP-affiliated sequencing centers, reaching the goal of 9 genomes per day globally, or nearly 3,000 annually, is anticipated (Table 1). The main challenge will be sourcing high-quality taxonomically identified samples for the isolation of high molecular weight DNA and RNA required for long-read DNA sequencing, scaffolding, and annotation. Separate from the current commitments above, about 50% of the taxonomic families could be obtained today from existing collections in the Global Genome Biodiversity Network (*SI Appendix*) (29). Obtaining samples from many countries may require diverse permit processes that can last weeks to years. The EBP is working to develop long-term collaborations to facilitate sample access across the world.

Another critical challenge will be obtaining reference-quality assemblies from small organisms, single-cell eukaryotes, and some green plants. New low-DNA input methods (30) have essentially solved the problem for most metazoans, but not for single-cell eukaryotes that cannot be cultured. Producing reference-quality genomes thus remains a significant challenge for a large part of the eukaryotic tree of life. Setting standards for the generation and storage of the complex set of genomes that characterize green plants will need to accommodate the immense variation in their size, transposable element content, and structure, while enabling research into the molecular and evolutionary processes that have resulted in this enormous genomic variation (23). Recommendations for sample collection and processing are included in this issue. Accelerating the annotation pipeline will also present major challenges as the production of genomes scales up. Planned 2021 annotation throughput is 300, 400, and 500 species for the National Center for Biotechnology Information, Joint Genome Institute, and European Molecular Biology Laboratory–European Bioinformatics Institute, respectively, which remains short of what will be necessary. This issue can be addressed by expanding capacity and creating more efficient genome annotation tools (31). Current recommendations for genome annotation are provided in this issue.

To achieve the outputs required for Phase II and Phase III, dramatic increases in genome sequence production and efficiency will be required. Sequencing one representative for each

of ~165,000 genera in 4 y will require an increase in the throughput of genomes from 9 per day to 123 per day, or 14-fold above the Phase I target. Phase III will require another 10-fold increase above the Phase II target in order to complete the project in 10 y. We are optimistic that within 5 y, sample processing and sequencing technology will improve and costs will be reduced so that reference-quality genomes can be produced for all species for under USD \$1,000 for a 2-Gb genome. We note that the cost, accuracy, and contiguity of assemblies produced today with long reads were not available 2 y ago. High-quality draft assemblies based on long reads can already be produced for ~\$2,000 in reagents and compute per 1-Gb genome average, getting closer to the \$800 originally envisioned for short-read draft-quality genomes (1). Sequencing done for Phases II and III should meet or exceed the minimum standards for short-read-based draft assemblies: contig N50 > 100 Kb, scaffold N50 > 1 Mbp (or chromosome scale for smaller genomes), QV30. Although the EBP aspires to produce chromosome-level assemblies for all species, for uncultured microbial eukaryotes and highly repetitive genomes, the project will sacrifice perfection for progress in the near term.

In 2018, we estimated a total EBP cost of USD \$4.7 billion. This is significantly less than the original USD \$2.7 billion (1991 dollars) cost of sequencing the human genome, comparable with USD \$5.2 billion today. We note that producing complete telomere-to-telomere assemblies for all human chromosomes is a mission that is now being realized (32), and that the true cost of sequencing the human genome is significantly higher than the original USD \$2.7 billion price tag. Reference-quality genomes currently being produced by the EBP's sequencing nodes are of far greater quality (i.e., continuity, completeness, phasing) than the original "complete" human genome sequence [e.g., Rhie et al. (19)], and can now be produced for about USD \$10,000 per 2-Gb genome, including transcriptome data for annotation. This amount is 20% of the cost of a similar quality assembly only 3 y ago when the original estimates were made. The project will save about USD \$186 million in Phase I due to these improvements, bringing the total cost of Phase I down to \$414 million from \$600 million.

The EBP has embraced the strategy of supporting funding efforts by states and nations: for example, the California Conservation Genomics Project and 1000 Chilean Genomes (*SI Appendix*), and EBP-Colombia (21). This effort has proven highly successful as it allows for local and regional concerns to be addressed in the funding drive. For example, in Australia there is great interest in conserving endangered marsupial species [see Hogg, this issue (33)]. This has led to a funded project that will produce five new marsupial reference genomes in 2021 (Table 1). Other examples include the Catalan Initiative for the Earth BioGenome Project, which aims to prioritize sequencing of endemic species with the goal of eventually sequencing all species in the Catalan territories (*SI Appendix*). National funding also provides an inherent mechanism for compliance with national laws on access and benefit sharing, which may prove essential for building trust, and ultimately obtaining all taxonomically classified species for sequencing. Capacity building in developing countries will be a direct benefit of participation.

Conclusions

The past year has been one of great progress for the EBP, marking the start of the clock for completing Phase I of the project.

There are many challenges ahead in meeting Phase II and Phase III goals. Clearly, the ultimate aim of sequencing 1.84 million eukaryotes cannot be achieved by a single country or private entity. The coordinated efforts of thousands of scientists and institutions around the world are needed to produce ~9,400 family reference genomes in 3 y. The project needs significant amounts of new funding, but the investments required on a global scale should be obtainable given the importance of the project to conserving and enhancing ecosystem services in the context of climate change and promoting a new bioeconomy. Despite limited financial resources for coordination, the EBP international network-of-networks has matured as the world's most technically advanced organization to tackle the grand challenge of sequencing all known eukaryotes, identifying their genes and functions, advancing our understanding of the evolution of life on Earth, and developing a complete genomic characterization of Earth's critical ecosystems. Based on a survey of institutional members and affiliates, the EBP now includes more than 5,000 scientists and technical staff around the world who are dedicated to EBP's mission. The EBP has unleashed

tremendous passion and energy among the project's participants, particularly its younger generation of scientists and the general public.

Given the precarious condition of Earth's biodiversity, it is essential that the EBP and its affiliated projects achieve their ambitious goals. In the words of David Attenborough, "Extinction is forever—so our action must be immediate." Every eukaryotic species is the product of millions of years of evolution. Recorded in their genomes are secrets that can fundamentally change our understanding of the evolution of life on Earth—its very existence and essence—and may lead to radical new approaches for mitigating the effects of climate change on biodiversity, improving agriculture, growing a sustainable global bioeconomy, saving species and repairing ecosystems, and preventing future pandemics. Let us go forth and sequence!

Acknowledgments

We thank Prof. Beth Shapiro and Fritz J. Sedlazeck for their editorial comments on the manuscript.

^aDepartment of Evolution and Ecology, College of Biological Sciences, University of California, Davis, CA 95616; ^bDepartment of Population Health and Reproduction, University of California, Davis, CA 95616; ^cUniversity of California Davis Genome Center, University of California, Davis, CA 95616; ^dDNA Zoo and The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030; ^eCenter for Genome Regulation, Universidad de Chile 3425 Santiago, Chile; ^fFacultad de Ciencias, Universidad de Chile 3425 Santiago, Chile; ^gDepartment of Biochemistry & Molecular Biology, Dalhousie University, Halifax, NS B3H 4H7, Canada; ^hLOEWE Centre of Translational Biodiversity Genomics, Senckenberg Leibniz Institution for Biodiversity and Earth System Research 60325 Frankfurt am Main, Germany; ⁱInstitute for Insect Biotechnology, Justus-Liebig University 35392 Giessen, Germany; ^jGlobal Genome Biodiversity Network Secretariat, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560; ^kRevive & Restore, Sausalito, CA 94965; ^lSchool of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia; ^mDepartment of Life Sciences and Biotechnology, University of Ferrara 44121 Ferrara, Italy; ⁿTree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, United Kingdom; ^oSchool of Ecology and Environment, Northwestern Polytechnical University 710072 Xi'an, China; ^pThe Novim Group, University of California, Santa Barbara, CA 93106; ^qSpacetime Ventures 05449-050 São Paulo, Brazil; ^rBiodiversity Research Center, Academia Sinica 11529 Taipei, Taiwan; ^sBee Research Laboratory, Beltsville Agricultural Research Center, US Department of Agriculture, Agriculture Research Service, Beltsville, MD 20705; ^tGlobal Genome Initiative, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560; ^uConservation Science, Species360 Conservation Science Alliance, Bloomington, MN 55425; ^vDepartment of Biology, University of Southern Denmark 5230 Odense M, Denmark; ^wDepartment of Genetics, Microbiology, and Statistics, Universitat de Barcelona 08028 Barcelona, Spain; ^xCatalan Society for Biology, Institute for Catalan Studies 08001 Barcelona, Spain; ^yDepartment of Biostatistics & Bioinformatics, Computational Biology Institute, George Washington University, Washington, DC 20052; ^zDepartment of Biostatistics & Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC 20052; ^{aa}Department of Biological Sciences, Universidad de los Andes 111711 Bogotá, Colombia; ^{ab}Genome British Columbia, Vancouver, BC V5Z 0C4, Canada; ^{ac}Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom; ^{ad}Wellcome Sanger Institute, Cambridge CB10 1SA, United Kingdom; ^{ae}UniProt, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, United Kingdom; ^{af}Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; ^{ag}Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138; ^{ah}Laboratory of the Neurogenetics of Language, The Rockefeller University, New York, NY 10065; ^{ai}European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; ^{aj}Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10065; ^{ak}Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030; ^{al}GLOBE Institute, University of Copenhagen 1350 Copenhagen, Denmark; ^{am}University Museum, Norwegian University of Science and Technology 7491 Trondheim, Norway; ^{an}Department of Health Policy and Management, George Washington University, Washington, DC 20052; ^{ao}School of Life Sciences, La Trobe University, Bundoora, VIC 3086, Australia; ^{ap}Institute for Applied Ecology, University of Canberra, Bruce, ACT 2617, Australia; ^{aq}Stanford Law School, Stanford University, Stanford, CA 94305; ^{ar}US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; ^{as}Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720; ^{at}Office of National Programs, US Department of Agriculture, Agricultural Research Service, Beltsville, MD 20705; ^{au}Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, United Kingdom; ^{av}Genome Institute, University of California, Santa Cruz, CA 95060; ^{aw}HMMI, Chevy Chase, MD 20815; ^{ax}Australian Museum Research Institute, Australian Museum, Sydney, NSW 2000, Australia; ^{ay}Department of Frontier Research and Development, Kazusa DNA Research Institute, Chiba 292-0818, Japan; ^{az}Department of Biosciences, University of Oslo, Oslo 0316, Norway; ^{baa}Walter Reed Biosystematics Unit, Smithsonian Institution, Suitland, MD 20746; ^{bab}Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Front Royal, VA 22630; ^{bac}Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada; ^{bad}Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605; ^{bae}Broad Institute of MIT and Harvard, Cambridge, MA 02142; ^{baf}Royal Botanic Gardens, Kew, Richmond TW9 3AE, United Kingdom; ^{bag}Division of Life Sciences, Korea Polar Research Institute 21990 Incheon, South Korea; ^{bah}Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012; ^{bai}Department of Genomics and Evolutionary Biology, National Institute of Genetics 411-8540 Shizuoka, Japan; ^{baa}Laboratory for Phylinformatics, RIKEN Center for Biosystems Dynamics Research 650-0047 Hyogo, Japan; ^{baa}Department of Plant Biology, University of Georgia, Athens, GA 30602; ^{baa}State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences 650223 Yunnan, China; ^{baa}Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University 752 36 Uppsala, Sweden; ^{baa}BGI-Research, Beijing Genomics Institute-Shenzhen 518083 Shenzhen, China; ^{baa}Department of Biological Sciences, Halmos College of Arts and Sciences, Nova Southeastern University, Dania Beach, FL 33004; ^{baa}PPP Guy Harvey Oceanographic Center, Dania Beach, FL 33004; ^{baa}Institute of Evolutionary Biology, Pompeu Fabra University, Consejo Superior de Investigaciones Científicas, Parc de Recerca Biomedica de Barcelona 08003 Barcelona, Spain; ^{baa}Catalan Institute of Research and Advanced Studies 08010 Barcelona, Spain; ^{baa}Centre Nacional d'Anàlisi Genòmica, Centre for Genomic Regulation, Barcelona Institute of Science and Technology 08028 Barcelona, Spain; ^{baa}Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona 08193 Barcelona, Spain; ^{baa}Bioplatforms Australia, Macquarie University, Sydney, NSW 2109, Australia; ^{baa}One Health Institute, University of California Davis, CA 95616; ^{baa}Berlin Center for Genomics in Biodiversity Research 14195 Berlin, Germany; ^{baa}Evolutionary Genetics Department, Leibniz Institute for Zoo and Wildlife Research 10315 Berlin, Germany; ^{baa}Max Planck Institute for Molecular Cell Biology and Genetics 01307 Dresden, Germany; ^{baa}Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269; ^{baa}Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269; ^{baa}Division of Biotechnology, Korea University 02841 Seoul, Korea; ^{baa}Department of Entomology, Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801; ^{baa}Systematics and Evolution of Vascular Plants Associated Unit to Consejo Superior de Investigaciones Científicas, Departament de Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona 08193 Bellaterra, Spain; ^{baa}Laboratoire d'Ecologie Alpine, University Grenoble Alpes, University Savoie Mont Blanc, CNRS 38000 Grenoble, France; ^{baa}Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027; ^{baa}Division of Biology, Department of Evolution, Behavior, and Ecology, University of California, San Diego, La Jolla, CA 92039; ^{baa}Department of Biological Sciences, Faculty of Science, King Abdulaziz University 21589 Jeddah, Saudi Arabia; ^{baa}Centre of Excellence in Bionanoscience Research, King Abdulaziz University 21589 Jeddah, Saudi Arabia; ^{baa}La Kretz Center for California Conservation Science, Institute of Environment and Sustainability, University of California, Los Angeles, CA 90024; ^{baa}Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095; ^{baa}Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543; ^{baa}College of Law, University of Illinois at Urbana-Champaign, Champaign, IL 61820; ^{baa}Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; ^{baa}Biodiversity Institute, University of Florida, Gainesville, FL 32611; ^{baa}Jiangsu Key Laboratory for Bioresources of Saline Soils, Jiangsu Provincial Key Laboratory of Coastal Wetland Bioresources and Environmental Protection, Jiangsu Synthetic Innovation Center for Coastal Bio-agriculture, School of Wetlands, Yancheng Teachers University 224002 Yancheng, China; ^{baa}Center of Mycology and Microbiology, University of Tartu 50411 Tartu, Estonia; ^{baa}College of Science, King Saud University 11451 Riyadh, Saudi Arabia; ^{baa}Research and Collections, Natural History Museum of Los Angeles County, Los Angeles, CA 90007; ^{baa}Biological Sciences, University of Southern California, Los Angeles, CA 90089; ^{baa}Department of Biology, Duke University, Durham, NC 27708; ^{baa}Duke Center for Genomic and Computational Biology, Duke University, Durham, NC 27708; ^{baa}Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen 2100 Copenhagen, Denmark; and ^{baa}China National Genebank, Beijing Genomics Institute 51803 Shenzhen, China

- 1 H. A. Lewin *et al.*, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4325–4333 (2018).
- 2 S. Diaz *et al.*, *Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, 2019).
- 3 International Union for Conservation of Nature, The IUCN Red List of Threatened Species. Version 2020-3. <https://www.iucnredlist.org>. Accessed 1 April 2021.
- 4 Bioeconomy Capital, Bioeconomy dashboard: Economic metrics. <http://www.bioeconomycapital.com/bioeconomy-dashboard>. Accessed 1 April 2021.
- 5 Natural Resources Institute Finland, Finnish bioeconomy in numbers. <https://www.luke.fi/en/natural-resources/finnish-bioeconomy-in-numbers/>. Accessed 1 April 2021.
- 6 National Academy of Sciences, *The Challenge of Feeding the World Sustainably: Summary of the US-UK Scientific Forum on Sustainable Agriculture* (National Academies Press, Washington, DC, 2021).
- 7 World Economic Forum, The Global Risks Report 2021, ed. 16. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2021.pdf. Accessed 1 April 2021.
- 8 McKinsey Global Institute, The Bio Revolution—Innovations transforming economies, societies, and our lives. https://www.mckinsey.com/~media/McKinsey/Industries/Pharmaceuticals%20and%20Medical%20Products/Our%20Insights/The%20Bio%20Revolution%20Innovations%20transforming%20economies%20societies%20and%20our%20lives/May_2020_MGI_Bio_Revolution_Report.pdf. Accessed 1 April 2021.
- 9 P. Zhou *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
- 10 J. Damas *et al.*, Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22311–22322 (2020).
- 11 G. Hansen, J. Mazet, J. Rushton, C. Stroud, What happens after disease X: Using One Health to prevent the next pandemic. *NAM Perspectives*, 10.31478/202011c (2020).
- 12 W. J. Kress, J. A. K. Mazet, P. D. N. Hebert, Opinion: Intercepting pandemics through genomics. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13852–13855 (2020).
- 13 M. L. Blaxter, *et al.*, Why sequence all eukaryotes? *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115636118 (2021).
- 14 S. A. McCartney *et al.*, Balancing openness with indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all life. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115860118 (2021).
- 15 J. S. Sherkow *et al.*, Ethical, legal, and social issues in the Earth BioGenome Project. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115859118 (2021).
- 16 Secretariat of the Convention on Biological Diversity, *Text and Annex of the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization to the Convention on Biological Diversity* (United Nations, Montreal, ed. 1, 2011). <https://www.cbd.int/abs/text/default.shtml>. Accessed 21 February 2015.
- 17 Y. Roskov *et al.*, Eds., *Species 2000 & ITIS Catalogue of Life, 2020-12-01* (Species 2000: Naturalis, Leiden, the Netherlands, 2020). <https://www.catalogueoflife.org>.
- 18 National Center for Biotechnology Information, Data from “Eukaryota.” NCBI. <https://www.ncbi.nlm.nih.gov/assembly?term=eukaryotes>. Accessed 4 March 2021.
- 19 A. Rhie *et al.*, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
- 20 C. Leech, “Biodiversity Genomics 2020” (video recording, 2020). <https://www.youtube.com/playlist?list=PLpziXi8KUFQXTq6YelpmBCicIRA1KB6eU>. Accessed 3 December 2021.
- 21 J. Huddart, A. J. Crawford, A. L. Luna Tapia, S. Restrepo, F. DiPalma, EBP-Colombia and the bioeconomy: Genomics in the service of biodiversity conservation and sustainable development. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115641118 (2021).
- 22 The Darwin Tree of Life Project Consortium, Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115642118 (2021).
- 23 W. J. Kress *et al.*, Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115640118 (2021).
- 24 H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- 25 J. Ghurye *et al.*, Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
- 26 D. Hobern, BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. *Genome* **64**, 161–164 (2021).
- 27 D. Carroll *et al.*, The Global Virome Project. *Science* **359**, 872–874 (2018).
- 28 T. Stephan *et al.*, Darwinian genomics and diversity in the Tree of Life. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115644118 (2021).
- 29 G. Droege *et al.*, The Global Genome Biodiversity Network (GGBN) data portal. *Nucleic Acids Res.* **42**, D607–D612 (2014).
- 30 S. B. Kingan *et al.*, A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes (Basel)* **10**, 62 (2019).
- 31 K. L. Howe *et al.*, Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
- 32 K. H. Miga *et al.*, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
- 33 C. J. Hogg *et al.*, Threatened Species Initiative: Empowering conservation action using genomic resources. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115643118 (2021).