PERSPECTIVE

# The Earth BioGenome Project: Sequencing Life for the Future of Life

Harris A. Lewin[1], Gene E. Robinson[2], W. John Kress[3], William J. Baker[4], Jonathan Coddington[3], Keith A. Crandall[5], Richard Durbin[6], Scott V. Edwards[7], Félix Forest[4], M. Thomas P. Gilbert[9,10], Melissa M. Goldstein[11], Igor V. Grigoriev[12,13], Kevin J. Hackett[14], David Haussler[15], Erich Jarvis[16], Warren E. Johnson[18], Ari Patrinos[18], Stephen Richards[19], Juan Carlos Castilla-Rubio[20], Marie-Anne van Sluys[21], Pamela S. Soltis[22], Xu Xun[23], Huanming Yang[24] and Guojie Zhang[23,25,26]

Submitted to Proceedings of the National Academy of Sciences of the United States of America

**Increasing our understanding of Earth's biodiversity and responsibly stewarding its resources are among the most crucial scientific and social challenges of the new millennium. These challenges require fundamental new knowledge of the organization, evolution, functions and interactions among millions of the planet's organisms. Herein we present a perspective on the Earth BioGenome Project (EBP), a Moon Shot for biology that aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years. The outcomes of the EBP will inform a broad range of major issues facing humanity, such as the impact of climate change on biodiversity, the conservation of endangered species and ecosystems, and the preservation and enhancement of ecosystem services. We describe hurdles the project faces, including data sharing policies that ensure a permanent, freely-available resource for future scientific discovery while respecting Access and Benefit Sharing guidelines of the Nagoya Protocol. We also describe scientific and organizational challenges in executing such an ambitious project, and the structure proposed to achieve the project's goals. The far-reaching potential benefits of creating an open digital repository of genomic information for life on Earth can be realized only by a coordinated international effort.**



biodiversity | genome sequencing | Access and Benefit Sharing | genomics | data science

"Our task now is to resynthesize biology; put the organism back into its environment; connect it again to its evolutionary past; and let us feel that complex flow that is organism, evolution, and environment united."

*New Biology for a New Century,* Carl R. Woese

**Biodiversity: a threatened global resource provides a call to action**

We are only just beginning to understand the full majesty of life on Earth (1). Although 10 to 15 million eukaryotic species and perhaps trillions of bacterial and archaeal species adorn the Tree of Life, approximately 2.3 million are actually known (2) and, of those, fewer than 15,000, mostly microbes, have completed or partially-sequenced genomes (Fig. 1). From this small fraction of Earth's known biome, a significant portion of modern knowledge in biology and the life sciences has emerged. This foundational knowledge has facilitated enormous advances in agriculture, medicine and bio-based industries, and enhanced approaches for conservation of endangered species.

Despite these great advances, the world's biodiversity is largely uncharacterized and increasingly threatened by climate change, habitat destruction, species exploitation, and other human-related activities. The Living Planet Index reported a 58% decline in vertebrate populations during the 42-year period 1970-2012 (3), and the International Union for Conservation of Nature (IUCN) estimated that ~23,000 of ~80,000 species surveyed are approaching extinction (4). We are in the midst of the 6$^{th}$ great extinction event of life on our planet (5), which not only threatens wildlife species but also imperils the global food supply (6). By the year 2050, up to 50% of existing species may become extinct mainly due to natural resource-intensive industries (5). Humanity faces the question of how such massive losses of species diversity will affect the complex ecosystems that sustain life on Earth, including our ability to derive the foods, biomaterials, bioenergy, and medicines necessary to support an expected human population of 9.6 billion by 2050. Ecosystem collapse on a global scale is a real possibility, making the preservation and conservation of terrestrial, marine, freshwater, desert and agricultural ecosystems a global imperative for human survival and prosperity.

Unimaginable biological secrets are held in the genomes of the millions of known and unknown organisms on our planet. This "dark matter" of biology could hold the key to unlocking the potential for sustaining planetary ecosystems upon which we depend and provide life-support systems for a burgeoning world

Harris A. Lewin[1], Gene E. Robinson[2], W. John Kress[3], William J. Baker[4], Jonathan Coddington[3], Keith A. Crandall[5], Richard Durbin[6], Scott V. Edwards[7], Félix Forest[4], M. Thomas P. Gilbert[9,10], Melissa M. Goldstein[11], Igor V. Grigoriev[12,13], Kevin J. Hackett[14], David Haussler[15], Erich Jarvis[16], Warren E. Johnson[18], Ari Patrinos[18], Stephen Richards[19], Juan Carlos Castilla-Rubio[20], Marie-Anne van Sluys[21], Pamela S. Soltis[22], Xu Xun[23], Huanming Yang[24] and Guojie Zhang[23,25,26]

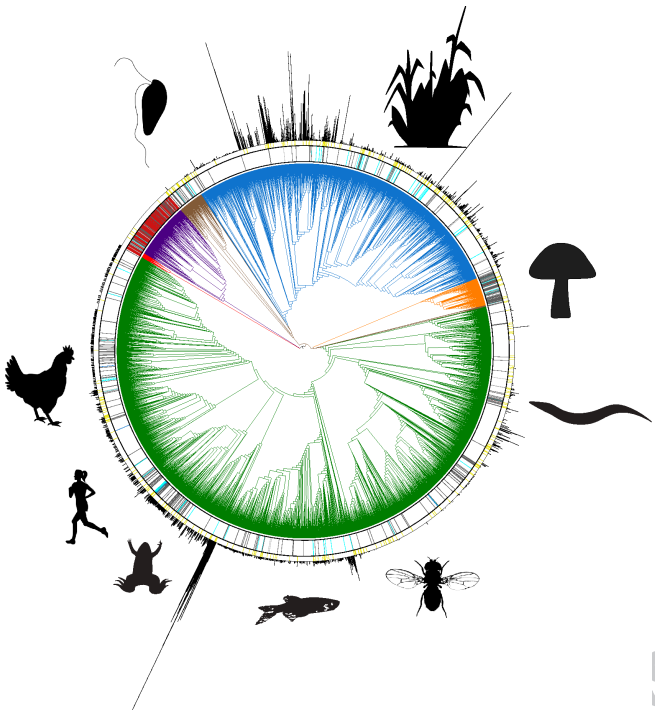

**Fig. 1. Current status of the sequencing of life. Open Tree of Life (opentreeoflife.org) synthesis phylogeny of all of life with resolution to the genus level showing phylogenetic information for Archaea (red), Bacteria (purple), Fungi (orange), Plantae (blue), Protista (brown) and Animalia (green). We show the current state of genomic information available through NCBI's GenBank in the inner circle with complete genomes colored in red, chromosome level in blue, scaffolds in dark grey and contigs in light grey. The second circle shows the transcriptomes available through the NCBI Transcriptome Shotgun Assembly Sequence Database (https://www.ncbi.nlm.nih.gov/genbank/tsa/). Genome size as C-value is displayed in bars around the outer circle, with data for animals extracted from the Animal Genome Size Database (http://www.genomesize.com/), for plants, from the Royal Botanic Gardens, Kew (http://data.kew.org/cvalues/), and for fungi, from the Fungal Genome Size Database (http://www.zbi.ee/fungal-genomesize/).**

**Table 1. Genomics-enabled discoveries and applications.**

| Taxon | Application |
|---|---|
| Humans | disease diagnostics; disease risk; human ancestry; drug design; personalized medicine; forensics |
| Livestock & wildlife | disease diagnostics; genomic selection for milk yield, carcass composition; parentage control; conservation of endangered breeds and species; disease models |
| Plants | genomic selection to improve crop yields and other agronomically-important traits; biofuels production; gene editing; conservation of endangered species |
| Insects | gene drives; genome editing; pest control |
| Fungi | synthetic biology; metabolic engineering for drug production and useful chemicals; biofuels production, improved strains for making wine and beer |
| Bacteria | microbiome in health and disease; bioprocessing; detection, surveillance and host response; genomic epidemiology; understanding microbial diversity |
| Viruses | vaccines; gene editing; metagenomics screening |

population. For example, from invertebrates such as sponges, mollusks, tunicates, and cone snails, several FDA-approved drugs

**Table 2. Goals of the Earth BioGenome Project.**

| **Revise and reinvigorate our understanding of biology, ecosystems and evolution** |
|---|
| Better understand evolutionary relationships among all known organisms |
| Fully elucidate the timing, origin, distribution and density of species on Earth |
| Generate new knowledge ecosystem composition and functions |
| Discover new species (80-90% of eukaryote biodiversity) |
| Elucidate genome evolution (gene to chromosome scale) |
| Discover fundamental laws that describe and drive evolutionary processes |
| Enable the conservation, protection, and regeneration of biodiversity |
| Determine the role of climate change on biodiversity |
| Clarify how human activities (pollution, habitat encroachment, etc.) and invasive species affect biodiversity |
| Develop evidence-based conservation plans for rare and endangered species |
| Create genomic resources to restore damaged or depleted ecosystems |
| Maximize returns to society and human welfare (ecosystem services and biological assets) |
| Discover new medicinal resources for human health |
| Enhance control of pandemics |
| Identify new genetic variation for improving agriculture (e.g., yields, disease resistance) |
| Discover novel biomaterials, new energy sources and biochemical |
| Improve environmental quality (soil, air and water) |



**Table 3. Organized communities conducting large-scale genome projects.**

| Community | Lead Center(s) | Sequencing goal |
|---|---|---|
| G10K | BGI, Rockefeller University, Sanger Center, Broad Institute | all vertebrate genomes |
| GIGA | George Washington University, Nova Southeastern University | 7,000 marine invertebrates |
| GAGA | BGI | 300 ant genera |
| I5K | Baylor College of Medicine | 5,000 arthropods |
| 1000 Fungal Genomes Project | DOE Joint Genome Institute | 1,000 fungal species |
| 10KP | BGI | 10,000 plant genomes |

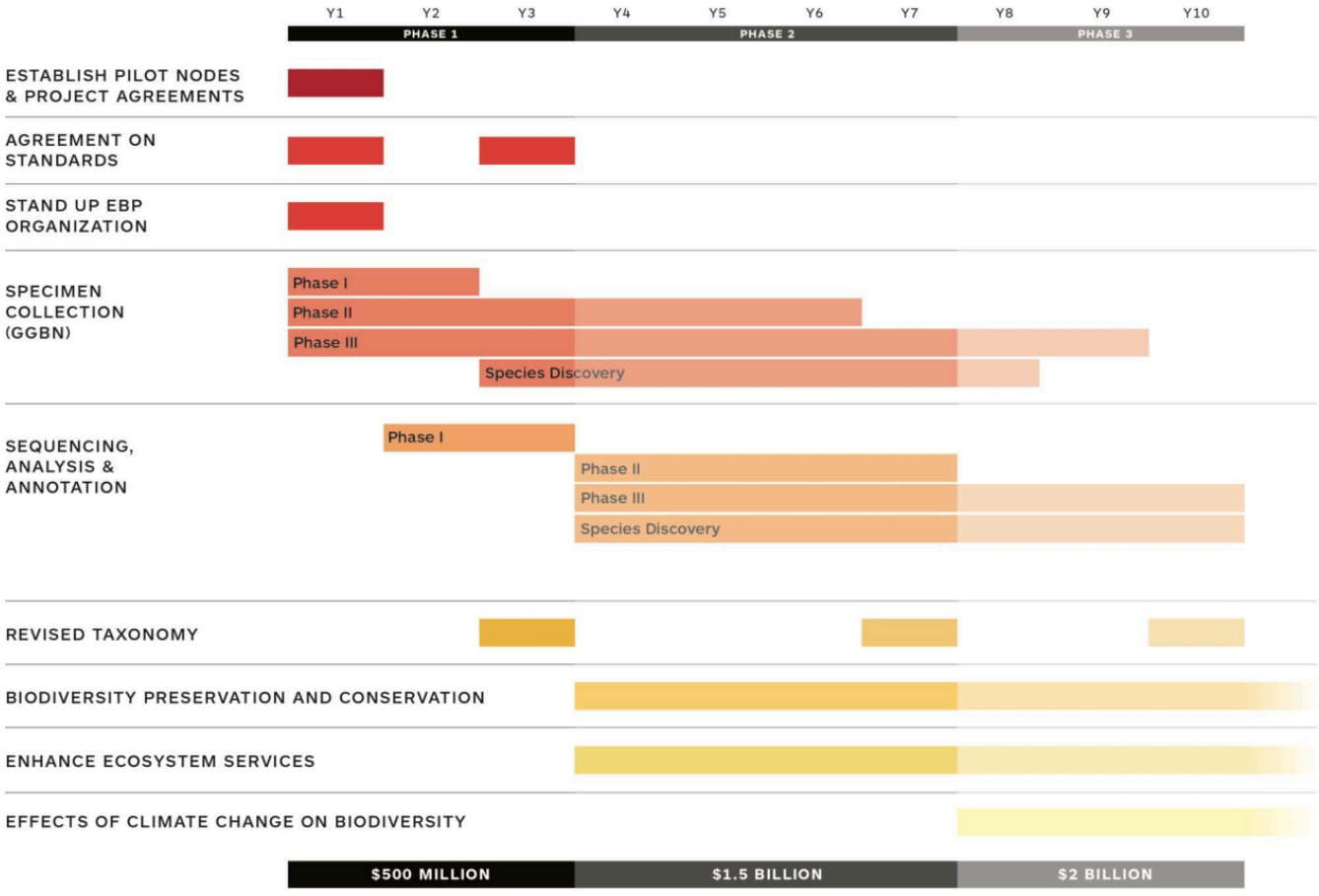


**Fig. 2. Proposed roadmap for the Earth BioGenome Project.**

have been developed for treating cancer (e.g., cytarabine, initially isolated from a sponge), virus infections, and pain (7). More than 25 marine-derived drugs are in preclinical or clinical trials. Fungi are the basis for fermentation to produce wine, beer and bread, and knowledge gained from yeast genomics has led to improved production strains for brewers, vintners and bioenergy production from waste streams. From the more than 391,000 known species of plants (https://stateoftheworldsplants.com/2-016/), hundreds of drugs for treating pain (e.g., opiates) and chronic diseases, including cancer (e.g., taxol), have been produced and commercialized. Plants are also the basis of large industries, such as food, rubber and second-generation bio-ethanol production. Thus, sequencing and annotating the vast number of previously uncharacterized genomes will continue to result in the discovery of many new useful genes, proteins, and novel metabolic pathways. These organisms and their genomes will provide the raw materials for genome engineering and synthetic biology approaches to produce valuable bio-products at industrial scale, as has been accomplished for artemisinin, which is used to treat malaria and nematode infections (8). Sustainable production of goods and bio-inspired materials is the foundation for a healthy planet, especially as humanity transitions from petrochemical inputs to reduce carbon emissions and other greenhouse gases.

Many pioneering discoveries have been made using species across the phylogenetic spectrum as a direct result of the genomics revolution. With initial efforts focused primarily on our own species, the United States government's initial investment of $3 billion to sequence and annotate the human genome, plus significant contributions from the Wellcome Trust and other international funding bodies, resulted in an entirely new field of medicine and more than $1 trillion of direct economic benefit (9). Driven by the technological advances made during the Human Genome Project, sequencing of other genomes across the Tree of Life have contributed to expanding scientific knowledge and the global economy (Table 1).

**Sequencing All Eukaryotic Life: Why Now?**

Powerful advances in genome sequencing technology, informatics, automation, and artificial intelligence, have propelled humankind to the threshold of a new beginning in understanding, utilizing, and conserving biodiversity. For the first time in history, it is possible to efficiently sequence the genomes of all known species, and to use genomics to help discover the remaining 80 to 90 percent of species that are currently hidden from science. While organized efforts are underway to sequence Bacteria and Archaea (e.g. (10)), there is currently no parallel effort for the Eukarya.

A conceptual argument for sequencing eukaryotic life was made by Stephen Richards in 2015 (11). Richards argued that current technology would permit sequencing a vast number of species, and a phylogenetic approach to stratifying samples for sequencing would accelerate scientific discovery. Independently, in November 2015, an exploratory meeting that included representatives from research universities and major international and U.S. federal funding agencies (SI Table S1) was held at the Smithsonian Institution to discuss the rationale, strategies

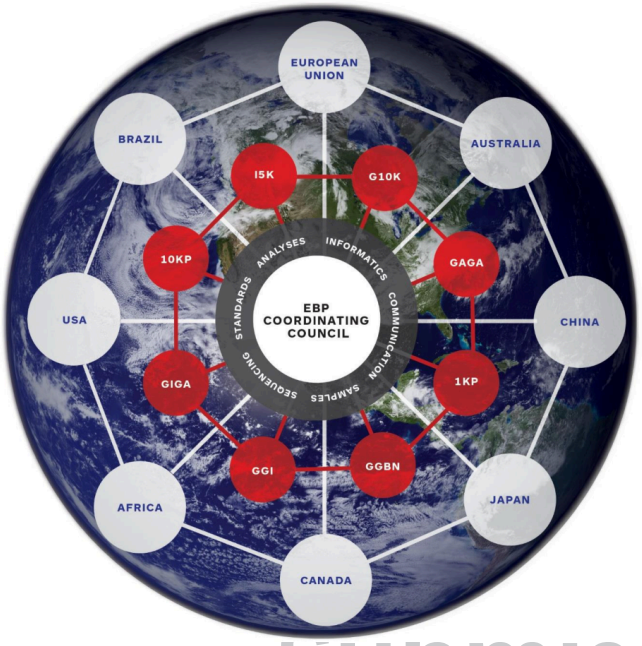


**Fig. 3. Earth BioGenome Project organizational model. The schematic shows the main features of the proposed organizational model for the EBP as a global network of communities. The outer ring of shows a global network of interacting nodes involved in DNA sequencing and informatics. The geographical locations are representative and not intended to be completely inclusive of all participating countries or political entities. The communities in the inner ring are also representative and not intended to be inclusive of all taxon-related communities and organizations that are supporting the goals of the EBP. The EBP Coordinating Council will have representation from all participating nodes, communities and organizations, as well as representation from the public and private sectors. GGBN, Global Genome Biodiversity Network; GGI, Global Genome Initiative; GIGA, Global Invertebrate Genomics Alliance; i5K, GAGA, Global Ant Genome Alliance; Initiative to Sequence 5000 Arthropod Genomes; G10K; Initiative to Sequence 10,000 Vertebrate Genomes; 1KP; Initiative to Sequence 1000 Plant Genomes.**

and feasibility of sequencing all life on Earth, a venture termed *The Earth BioGenome Project* (EBP). The consensus view that emerged from the meeting was that the time was right to consider a global initiative to sequence eukaryotic life on Earth. A subsequent EBP workshop and a major conference on genomics and biodiversity organized by the Smithsonian Institution and BGI (China) were held in Washington, D.C., in February, 2017. There, the EBP Working Group endorsed a project roadmap and organizational structure for completion of the sequencing aspect of the project in 10 years (12) as described below.

One of the key strategic issues for the EBP is the goal of sequencing every species as opposed to one representative member of each family or genus. This important objective requires a strong rationale to justify the cost. Evolution, co-evolution and conservation ultimately occur at the species level, and ecology is defined by the interactions among species. Therefore, understanding evolutionary and other biological processes, such as adaptation, speciation, the fate of endangered species, the reasons for extinctions, the importance of species to the functioning of ecosystems, and the possibility of restoring ecosystems critical to human survival, all require knowledge at the species level. In modern biology, the most powerful way to gain insights into the origins, evolution and biological functions of a species is through genomics. Moreover, taxonomy is a human construct, even when based on phylogeny, and the number of species in a recognized taxon is in part historical artifact and convenience as well as the product of evolution. Consider, for example, the genus *Candida*, which includes about 25% of all known yeast species. Some species are important human pathogens, while others are associated with wine spoilage. The metabolic and phenotypic diversity in this genus is enormous, and no single species can possibly represent the unique biology contained within it. Similar taxonomic diversity exists in plants, such as the genus *Astragalus* with more than 3,200 species. In insects, one order alone, the Coleoptera (beetles), has nearly 400,000 identified species in 30,000 genera across 176 families, which represents about 25% of all classified eukaryotic life, with a predicted 1.5 million beetle species inhabiting the planet (13). Sampling just one species per genus or family would not give a realistic assessment of the evolutionary complexity of these, or many other, groups. While recognizing that it may not be feasible to obtain samples for every species, pragmatism does not negate the primary scientific and societal need for trying to do so.

**Project Goals and Anticipated Outcomes**

Revise and reinvigorate our understanding of biology and evolution. The EBP has identified a broad set of scientific goals and projected economic, social, and environmental returns to society and human welfare (Table 2). These goals require assembled whole genome sequences collected in a robust Tree of Life framework (2) to derive the fundamental evolutionary principles that drive eukaryotic genomic and phenotypic evolution. As of October 2017, there were only 2,534 unique eukaryotic species with sequenced genomes in the National Center for Biotechnology Information (NCBI) database, which is less than 0.2% of the known eukaryotes. Moreover, of these, only 25 species meet the standard for contig N50, scaffold N50, and other metrics proposed for reference genomes by the G10K organization (SI Fig. S1; SI Fig. S2). At all levels of assembly quality, only 25 species, mostly fungi, are assembled to "complete genome" status as defined by NCBI. The phylogenetic distribution of existing reference quality assemblies is also highly skewed, with only seven eukaryotic phyla represented. Thus, there is a critical need to obtain annotated genomes from across the eukaryotic Tree of Life to answer important scientific questions and to provide a solid foundation for future biological discoveries and innovations. In fact, the eukaryotic Tree of Life itself is poorly understood, and large-scale phylogenetic syntheses have resolved or validated only a small fraction of polytomies (2). Thus, the EBP will not only take advantage of a phylogenetic framework, but will contribute significantly to a better understanding of how all biological diversity is related, as well as the relative and absolute timing of diversification events (14).

Many questions can only be answered if all genomes of a single group of organisms are available (15). One of these scientific challenges is to resolve conflicting phylogenetic relationships in the eukaryotic Tree of Life, especially among the deepest branches (16). Whole genome sequences may be particularly powerful in this regard, because problems arising in phylogenetic tree topologies from the use of small numbers of genes can be resolved (16, 17). Whole genome sequencing allows for selection of the most informative gene set for accurate phylogenetic inference. A complete set of sequenced and annotated eukaryotic genomes will also greatly expand our knowledge and understanding of the effects of incomplete lineage sorting and

horizontal gene transfer on eukaryote phylogenomic analyses, and their functional role in eukaryote evolution. A well-supported eukaryotic Tree of Life is essential for properly classifying the millions of presently undiscovered, unnamed, and unclassified organisms as their genome sequences become available. We anticipate that new methods for whole genome sequencing of unicellular organisms (18) will contribute to the discovery of new eukaryotic species and their correct taxonomic classification, insights into divergence times, and ultimately resolving some of the most contentious phylogenetic relationships in the Tree of Life.

Another major scientific challenge that will be addressed is the understanding of how genomes evolve, from the base-pair to the chromosome level. Sequencing genomes of extant species will enable reconstruction of the evolutionary history of eukaryotic genomes from a computed ancestral state (19, 20). The evolutionary history of point mutations, duplications, deletions, insertions, translocations, inversions, fusions, and fissions is crucial to our understanding of the relationships between genotype and phenotype (21), and the changes in genomic architecture that led to multicellularity and organismal complexity. The computational reconstruction of the karyotype of the ancestral eukaryote and key ancestral genomes on the evolutionary paths to higher taxa will enhance understanding of the evolution of genes in the context of their surrounding regulatory DNA and enable resolution of long-standing controversies on the role of chromosome rearrangement in adaptive evolution and speciation (19, 22).

Facilitate the conservation, protection, and regeneration of biodiversity. There is a clear and urgent need to understand the impact of natural and human factors on biodiversity. Climate change and habitat destruction are having tremendous impacts on both marine and terrestrial ecosystems (23). We know that species numbers and diversity are rapidly declining, but other than storing germplasm in frozen collections in the event of a natural or human-induced disaster, limited means are available to systematically preserve, protect, and restore endangered species in the wild. Therefore, resources are needed to uncover and better document genomic diversity, especially of endangered species. Sampling genomic diversity in populations reveals the frequency and distribution of genetic polymorphisms as baseline data on population fitness necessary for the design of conservation programs (reviewed in (24)). In endangered species with severely reduced populations, breeding programs based on genomic data from a few individuals can be implemented that avoid inbreeding, eliminate recessive lethal alleles, and increase disease resistance (25). Although much can be learned by sequencing a single diploid or polyploid individual, characterization of genomic diversity of the more than 23,000 species currently listed as endangered by IUCN is a high priority and a goal of the EBP. The Earth BioGenome Project will spur development of urgently needed new conservation management approaches based on genomic diversity of endangered species.

In addition to a taxonomically driven format for sequencing genomes, the EBP also will work to establish bio-observatories that use genomics to obtain a baseline understanding of how climate change affects global biodiversity. These observatories are especially critical in areas that have large numbers of endangered species. A network of instrumented bio-observatories in biodiversity hotspots can provide real-time information on species numbers, distribution and fluxes. We anticipate that some needs can be met by the National Ecological Observatory Network (NEON) created by the U.S. National Science Foundation, its counterpart in China (the Chinese Ecological Research Network; http://english.iae.cas.cn/rh/as/201311/t20131121_113005.html), the plot-based ForestGEO (http://www.forestgeo.si.edu/) and MarineGEO (https://marinegeo.si.edu/) programs at the Smithsonian Institution, and other national and international networks. Local resources, such as those developed by the UC California Conservation Genomics Consortium (https://ucconservationgenomics.eeb.ucla.edu/), will also be essential. Substantial opportunities will be afforded to students and citizens of all ages to participate in the monitoring of biodiversity, discovery of new species, and collection of environmental DNA (eDNA) samples, thus increasing public interest and participation in the EBP.

Monitoring organismal diversity in bio-observatories, particularly in remote locations, will spur development and use of new technologies, such as portable DNA sequencers, advanced sensor technologies, and secure bio-data transmission. Efficient technologies have already been created for identifying and categorizing species in their native habitats based on short DNA segments (26). The EBP will coordinate its activities with such organizations as the International Barcode of Life (iBOL; http://www.ibolproject.org/), the Global Genome Biodiversity Network (GGBN) (27), and major biodiversity collections around the world to contribute large numbers of new vouchered species for the barcoding effort and eventual whole genome sequencing. A digital repository of annotated eukaryotic genome sequences will facilitate new methods and approaches for studying genomic ecology at different spatial and temporal scales, which is necessary for obtaining a multidimensional and dynamic view of life on Earth.

Maximize returns to society and human welfare. The myriad ways by which ecosystems and the biodiversity comprising them contribute to the benefit of society and human welfare are termed ecosystem services. These services employ the full range of nature's natural products and materials and also serve as a template for imitating nature's biological functions and processes. The human population explosion and the rapid spread of medical and agricultural pests and diseases as a result of global interconnectivity are compelling examples of the need for new resources that will contribute to feeding, protecting, and improving Earth's ecosystems. An urgent demand exists for new sources of food proteins that can be produced cheaply and at scale, new medicines for treating the increasing frequency of chronic diseases plaguing human populations, new strategies for controlling outbreaks of zoonotic diseases, and new resources for maintaining and improving the quality of soil, air and water. With less than 0.2% of known eukaryotic genomes sequenced, most at draft level, and only a small fraction of nature's 285,000 known natural compounds replicated in the laboratory, we have barely scratched the surface in identifying new genetic resources for delivery of ecosystem services. Thus, the full value of nature, in particular our tropical forests and other biodiversity-rich hot spots, is likely to be grossly underestimated. Annual revenues in the U.S. alone from genetically engineered plants and microbes are estimated at more than $300 billion, or about 2% of gross domestic product (28). Obtaining the genetic blueprints for all eukaryotic life, and eventually the vast numbers of Bacteria and Archaea, will create a powerful source of discovery for improving and increasing ecosystem services.

**A 10-year Road Map**

The successful, systematic sequencing of all life on Earth will not be possible without an organized and sustained effort. Although geographic and taxon-based communities are working on related initiatives, few of them are coordinated with each other, and most are not focused on whole genome sequencing. A global "network of communities," proposed as an EBP organizational structure (see below), is a realistic strategy for achieving the grand challenge goal of sequencing all life on Earth. Among the examples of organized taxon-based communities, those working on cultured and uncultured Bacteria and Archaea are benefitting from well-defined objectives and global funding sources. The National Microbiome Initiative (NMI) was launched in 2016 by the U.S. Office of Science and Technology Policy and is supported by $121 million from multiple stakeholder federal agencies (29). The NMI will receive an additional $400 million in support from private companies and philanthropies. The Earth Microbiome Project is an ambitious parallel global effort to characterize microbial diversity (10).

Because the scientific communities that work on Bacteria and Archaea are relatively well-organized by taxon, the EBP Working Group decided to similarly focus its planning effort on sequencing all eukaryotes using a taxonomically driven format. Although several taxon-based initiatives for sequencing the genomes of eukaryotic species exist (Table 3), a common strategy is lacking among these groups, and large swaths of the eukaryotic Tree of Life are not represented. The EBP Working Group has focused specifically on a strategy and justification to sequence and annotate all eukaryotic genomes, including vertebrates, invertebrates, plants, fungi, and microbial eukaryotes, among others. In this paper, we have explored the rationale, feasibility and challenges associated with sequencing eukaryotes almost exclusively, but the ultimate goal of the EBP is to work with other organizations to support sequencing of the full diversity of life on Earth.

The EBP roadmap calls for sequencing and annotating ~1.5 million known eukaryotic species in three phases over a 10-year period using a phylogenomic approach (Fig. 2). During the three years of phase I, one of the most important goals is to create annotated chromosome-scale reference assemblies for at least one representative species of each of the ~9,000 eukaryotic taxonomic families. Nucleotide divergence and divergence time will be additional factors in the selection of species so that balance across eukaryotic taxa is achieved. High-quality reference assemblies (minimum standard of 2.3.2.Qv40; see SI Fig. S1) at the family level will ensure robustness of comparative genomic analyses by providing complete gene sets as well as ordered and oriented syntenic blocks created by genome scaffolding methods (15, 19). In addition, these genomes will be useful for classification of extant and new species, identifying genetic changes associated with specialized traits in specific lineages, *in silico* reference-assisted scaffolding of assemblies produced in phase II and phase III of the project (30), *in silico* reconstruction of ancestral genomes, and rescue of species from extinction (15, 31). A full description of the roadmap, overall strategy, and estimated costs can be found in the Appendix, Supplementary Text.

**Challenges and Opportunities**

Sample acquisition. The EBP will fully catalog the genome content of eukaryotic biodiversity and make the data openly available as a permanent foundation for future scientific discovery whilst ensuring compliance with the Nagoya Protocol, as discussed below. The main challenge is the development of a global strategy for the collection of voucher specimens that are preserved adequately to enable production of high-quality genome assemblies. The distributed nature of Earth's biodiversity and the location of biodiversity hotspots in remote parts of the world, such as the Amazon Basin or Borneo, makes collection of many organisms a distinct challenge. For the EBP to be successful it is crucial to involve institutions whose mission it is to procure and preserve the world's biodiversity, such as natural history museums, botanical gardens, zoos, and aquaria . For example, the collections of the botanical gardens of the world comprise about one third of all species of plants and more than 40% of all endangered plant species (32), which will be an invaluable resource for the EBP. The EBP also has GGBN as a committed partner, which is the world's major resource of tissues and DNA from voucher specimens (see SI Supplementary Text). It will be essential to involve scientists in countries where a significant fraction of the world's biodiversity resides, such as Brazil, Colombia, Peru, Madagascar, Malaysia, and Indonesia. A goal of the EBP is to globalize its activities through novel partnerships that build scientific capacity in developing countries, including the capacity to utilize, not just create, our shared genomics legacy.

To accelerate the acquisition of voucher specimens, the EBP also plans to capitalize on the burgeoning citizen scientist movement (fueled by the internet and social media) and new autonomous robotic technologies. There is an exciting opportunity for citizen science to contribute to collecting and identifying sequence-ready specimens and performing data analysis. For example, the University of California Conservation Genomics Consortium's CALeDNA program (http://www.ucedna.com) will involve 1,000 citizen scientists who will collect 18,000 environmental samples by the end of 2018. Radically new technologies may be developed and deployed for sample collection, such as the use of aerial, terrestrial and aquatic autonomous drones equipped with high-resolution cameras that can enable species collection and identification, and telecommunications with taxonomic experts (33). Conceivably, such robotic devices can also be equipped with automated DNA extraction devices and portable DNA sequencers for rapid species identification. The development of such robots is feasible given the current state of relevant technologies, and offers an excellent opportunity for interdisciplinary collaboration and breakthrough innovation.

A special challenge for the EBP will be obtaining whole genome sequence information from cultured and uncultured single cell eukaryotes. For example, there are more than 34,000 known and perhaps 107,000 unidentified species in the Chromista and Protista Kingdoms (34). Sequencing of microbial eukaryotes will be paramount for resolving the phylogenies within the Protista and Chromista, and for understanding how early eukaryotic life evolved. As discussed above, recent advances in single cell sequencing technology (18) have opened new horizons for understanding microbial evolution. The methods involve separating single cells using flow cytometry, in situ lysis in micro-well plates, and whole genome amplification followed by library production and sequencing (18). At present, genome coverage varies significantly, but technological improvements are now yielding genomes with up to 80% coverage. The usefulness of single cell genomics in elucidating undiscovered microbial life has been demonstrated for Bacteria and Archaea (35) and should also prove invaluable for identifying and characterizing microbial eukaryotes.

Computation and Data Science. The EBP will generate opportunities and challenges for new tools to visualize, compare and understand the connection of genome sequence to the evolution of phenotype, organism and ecosystems. It is noteworthy that for storing sequence reads, assembling reads into genome sequences, aligning the genomes of related species, and annotating gene models, the computational challenge has already been surmounted by the information industry. For example, storage and distribution of reference genomes, annotations and analyses will likely require less than 10 Gigabytes per species or ~20 Petabytes in total (see SI Supplementary Text), well within current capabilities of the International Nucleotide Sequence Database Collaboration of NCBI, EMBL-EBI and DDBJ. Storage of the underlying sequence read data for the completed EBP is more challenging at ~200 Petabytes (see SI Supplementary Text for calculations). Commercial vendors and at least one academic project (CERN) have already surpassed this storage capacity (https://home.cern/about/updates/2017/07/cern-data-centre-passes-200-petabyte-milestone). Furthermore, we expect costs and capabilities to improve before the highest data generation years of the project. There are new technologies on the horizon that will help support genome storage needs, including 3-D memory, integrated computing technologies that overcome the I/O bottleneck, and faster networks enhanced by optical switching (36).

Similarly, computing requirements are very large but tractable. Mammalian sized long read genome assemblies currently require ~100 processor-weeks. The later phases of the EBP will require ~10,000 simultaneous assemblies running in parallel – a scale already approached by academic supercomputer centers such as those at universities in Texas, Pittsburg, Illinois and San Diego, and exceeded by commercial cloud providers such as AWS, Microsoft, Google, Alibaba, and others around the world. Although current tools are already capable of completing the project, there is no doubt that assembly, alignment, and annotation algorithms implemented in both hardware and software will, in the future, all need to be improved for efficiency, accuracy and application to difficult genomes such as very large, very repetitive or very polymorphic genomes.

The EBP promises the opportunity to envision and develop new computational tools and analysis methods to maximize our understanding and utilization of large amount of data generated by the project. This challenge will require new architectures, algorithms and software for improved quality, efficiency, and cost effectiveness as well as data analysis, big data visualization and sharing. For example, graph representations of diploid references are required – especially for more polymorphic genomes (37). New visualization tools enabling comparative inspection of gene loci and synteny across the Tree of Life will be required. We anticipate that annotation tools focused on a gene family across many species using gene-specific knowledge (e.g., transmembrane domain or three-dimensional structure conservation) will perform better than the current tools designed for the annotation of all genes. Comparative analyses could identify the degree of selection on each base-pair in every species, enabling the study of the evolution of gene regulation as well as gene and protein structure. Building phylogenetic trees with genomic data at this scale will require a new set of bioinformatics requirements and a modular framework for integrating phylogenetic data and computing on such trees. Better methods to visualize genome evolution via structural changes, including segmental duplications, inversions, translocations, insertions and deletions, must be created. In addition, improved methods to associate phenotype with change in comparative genome and transcriptome sequence are needed. We also envision new tools for assessing how protein sequence variation changes the efficiencies of enzymes, the binding constants of molecular interactions and the comparative systems biology of different eukaryotic cells and tissues. Finally, the advances in computer science and the internet have enabled new possibilities for large scale data sharing, such as open access to the EBP's computational "lab book." With a sufficiently rich platform, raw sequence data, assemblies, alignments, phylogenetic trees and automated annotations can be instantly disseminated. The EBP will promote these tools for equitable worldwide sharing of data, analysis tools and data mining resources.

Access and Benefit Sharing. In 2010, the Nagoya Protocol provided guidelines on access to genetic resources as well as fair and equitable sharing of benefits arising from their utilization under the Convention on Biological Diversity (CBD). Nagoya, an international convention, requires its member countries to create laws and policies within their own legal systems to address points outlined in the Convention. Thus, Nagoya is international but implemented at the national level. The Convention sets "minimum standards" and countries can go beyond those standards if they wish. A number of important publications can be found on the Nagoya Protocol and its associated impact on genetic research and collections (e.g., (34, 38)). Most countries are still working on their national implementations of Nagoya; however, some uncertainty remains as to what the final legal requirements will be.

Users of biological resources are now responsible for complying with regulations on biodiversity use at the national level, including those regulations associated with Access and Benefit Sharing. The EBP will adhere to the principles of the Nagoya Protocol by: 1) requiring participants to comply with regulations on biodiversity use at the national level, and 2) using the established tools and resources on Access and Benefit Sharing. Specifically, the EBP aims to provide fair, equitable, open and rapid access to, and sharing of the benefits of, the eukaryotic genomes of planet Earth.

In order to ensure proper documentation of genetic resources and Access and Benefit Sharing compliance, the EBP will promote downstream monitoring and tracking of utilized genetic and genomic resources. Systems for Access and Benefit Sharing compliance have been developed to meet this need, both at the national and international levels. For example, at the international level, the GGBN Data Standard (27, 39) was developed for the exchange of information on genetic samples housed in biological repositories globally. The Standard requires that genetic samples provided for research by GGBN member institutions (i.e., non-human biological repositories) be associated with permitting and other legal information associated with Access and Benefit Sharing. At the national level, a new Brazilian law (Law 13.123/15) expands the interpretation of "access to genetic resources" to include research related to molecular taxonomy, phylogeny, molecular ecology, and molecular epidemiology, as well as the use of information from genetic sequences published in databases. A national-level registration system, SisGen, allows Brazilian biological material to be legally accessed and shipped abroad for research and provides an interface for registration, notification, and accreditation. Such standards and registration

systems will be essential for the success of the EBP.

**Coordination and Governance**

A Network of Scientific Expertise. Achieving the goals of the EBP will require coordination on a global scale, as well as scientific excellence and experienced leadership (Fig. 3). The EBP Working Group, currently comprising this paper's authors, has well-known experts from the genomics, informatics, systematics, evolutionary biology, biorepositories and conservation communities, an ethicist and an expert on innovation. Some of them have already led or currently lead large-scale genome projects, such as those at the Sanger Center (UK) and BGI (China). The community of interested scientists is rapidly growing, and there is strong international support for this effort (12). In the near future, representatives from government, private industry, civil society, international organizations, and private foundations will be integrated into the governance structure of the EBP. Broad representation in governance is desirable for the EBP's global public good mandate to ensure inclusive societal benefits worldwide, and to establish stable and sustainable funding for accessibility to state-of-the-art technology in genomics, computing, data science, and biotechnology.

Several large sequencing centers are supporting the goals of the EBP, including BGI (China), Baylor College of Medicine (USA), the Sanger Institute (UK), and Rockefeller University (USA). The São Paulo Research Foundation (FAPESP), one of the major research funding organizations in Brazil, will establish an EBP node in São Paulo to be the initial location serving Latin America, adding to the global hub-and-spokes model envisioned for the EBP (Fig. 3). Other institutions and service providers with significant sequencing capacity will be encouraged to participate.

If the genomes of Earth's biome are to be compared and fully decoded there must be standards for creating, comparing and analyzing genome assemblies so that the genome information will be useful to the broadest possible scientific community. The issue of standards will be particularly challenging to coordinate across the sequencing and informatics nodes and among the different taxon-based communities. However, if this is not accomplished, the final anticipated outcomes may fall short of project expectations. To address this issue, authors of this paper (R.D. and H.L.) have recently developed a set of standards for quality assessment of whole genome assemblies (see SI Fig. S1). This standard can be applied to characterizing any whole genome assembly and can be readily adopted by other communities. In addition, the G10K community is working on a push-button assembly pipeline that uses data from a variety of sources to produce highly contiguous chromosome-scale assemblies. These approaches can be extended to all the communities comprising the EBP. An EBP central Coordinating Council, comprised of the leaders of each network community, and the sequencing and informatics nodes, will be responsible for developing and promulgating standards for sequencing, annotation and downstream analysis (Fig. 3).

A Network of Communities. A growing fraction of biodiversity genome sequencing is being performed by taxon-based communities of experts, such as the Genome 10K organization (G10K)(40), i5K (41), B10K (https://b10k.genomics.cn/), GAGA (http://antgenomics.dk/), and GIGA (42) (see Table 2 for ongoing genome projects affiliated with the EBP). Groups such as G10K and i5K have made progress in planning and implementation of their projects, and have served as a model for the EBP and other existing and developing groups working across the phylogenetic spectrum. It is anticipated that the leadership of phylum and class-level taxonomic groups will become part of the Governing Council of the EBP, and that the EBP will endeavor to provide support to all of these projects. The EBP will be a global effort – a network of communities and individuals – all focused on this grand challenge.

**Total Project Cost & Economic Benefit**

With the current cost of $1,000 USD (and plummeting rapidly) for sequencing an average vertebrate-sized genome to draft level, genomes of all ~1.5 million known eukaryotes, up to 100,000 new eukaryotic species, and a defined number of eDNA samples from biodiversity hot zone collection sites, can be sequenced to a high-level of completeness and accuracy for approximately $4.7 billion USD (SI Table S3). This includes costs for sequencing instruments, sample collection, ~9,000 reference-quality genomes, data storage, analysis, visualization and dissemination, and project management. Incredibly, this is less than the cost of creating the first draft human genome sequence ($2.7 billion USD) in today's dollars ($4.8 billion USD)! New funds raised for the EBP will be leveraged by the hundreds of millions of dollars already committed to genome projects around the world.

The economic impact of the EBP is likely to be very large and globally distributed. Using the Human Genome Project as an example, the return on U.S. federal investment was estimated at 141:1 in the U.S. alone as of 2012 (9). An entire industry was created, with a workforce size of more than 47,000 people generating nearly $1 trillion dollars in economic activity. The technologies arising from investments in genomics are having a profound effect on human medicine, veterinary medicine, renewable energy development, food and agriculture, environmental protection, industrial biotechnology, the justice system and national security. For example, a recent report of the U.S. National Academy of Sciences states that annual revenues in the U.S. from genetically engineered plants and microbes to be least $300 billion (43). China, the United Kingdom, Canada, France, Japan and other countries have also made sizable investments in genomics research and now have mature industries that are contributing to their national economies.

These economic returns from the Human Genome Project to date have resulted just from sequencing the human genome and those of a relatively small number of model organisms. With less than <0.2% eukaryotic species sequenced to date, there is significant potential for discoveries that will impact human, animal and environmental health and the food and agriculture system, and multiple manufacturing industries. While it is not possible to predict the economic impact, it is quite reasonable to assume that sequencing the remaining 99.8% of eukaryotic species will yield returns similar to or exceeding those of the Human Genome Project. Importantly, the distribution of much of the world's biodiversity in developing regions could bring tremendous economic benefits to those countries under the Nagoya Protocol as discussed above. The EBP will comply with Access and Benefit Sharing laws through partnerships with organizations such as the Amazon Third Way Initiative and the Amazon Bank of Codes (44).

## Conclusions

The Earth BioGenome Project is arguably the most ambitious proposal in the history of biology. If successful, the EBP will completely transform our scientific understanding of life on earth and provide new resources to cope with the rapid loss of biodiversity and habitat changes that are primarily due to human

activities and climate change. Fundamental knowledge of Earth's biodiversity may also lead to new food sources, revolutionary bio-inspired materials, and innovations to treat human, animal, and plant diseases. Significant challenges remain in executing the EBP, the most substantial of which are sample acquisition, the related issue of Access and Benefit Sharing, and funding. The greatest legacy of the EBP will be the gift of knowledge – a complete Digital Library of Life – that contains the collective biological intelligence of 3.5 billion years of evolutionary history. This knowledge will guide future discoveries for generations and may ultimately determine the survival of life on our planet.


## Acknowledgements

The authors thank Mike Trizna for compiling data from NCBI for the summary of genome quality; Manuela da Silva and Scott Miller for their contributions to the section on Access and Benefit Sharing; Katie Barker for providing data on GGBN and Brazilian biodiversity; David Stern for help in quantifying genome data across the tree of life and the development of Figure 1; Ilia Leitch for providing genome size data for plants, and Claudia Lutz for helpful editorial suggestions.





**1.** Wilson EO (1999) *The Diversity of Life* (W. W. Norton).
**2.** Hinchliff CE, *et al.* (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* 112(41):12764-12769.
**3.** WWF (2016) Living Planet Report 2016: Risk and resilience in a new era. (Gland, Switzerland).
**4.** IUCN (2017) IUCN 2016 : International Union for Conservation of Nature annual report 2016. (Gland, Switzerland), p 48.
**5.** Ceballos G, Ehrlich PR, & Dirzo R (2017) Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proc Natl Acad Sci U S A* 114(30):E6089-E6096.
**6.** International B (2017) *Mainstreaming Agrobiodiversity in Sustainable Food Systems: Scientific Foundations for an Agrobiodiversity Index* (Biodiversity International, Fiumicino, Italy).
**7.** Sharma V & Sarkar IN (2013) Leveraging biodiversity knowledge for potential phyto-therapeutic applications. *Journal of the American Medical Informatics Association* 20(4):668-679.
**8.** Ro D-K, *et al.* (2008) Induction of multiple pleiotropic drug resistance genes in yeast engineered to produce an increased level of anti-malarial drug precursor, artemisinic acid. *BMC Biotechnology* 8(1):83.
**9.** Wadman M (2013) Economic return from Human Genome Project grows. *Nature*.
**10.** Gilbert JA, Jansson JK, & Knight R (2014) The Earth Microbiome project: successes and aspirations. *BMC Biology* 12(1):69.
**11.** Richards S (2015) It's more than stamp collecting: how genome sequencing can unify biological research. *Trends Genet* 31(7):411-421.
**12.** Pennisi E (2017) Sequencing all life captivates biologists. *Science* 355(6328):894-895.
**13.** Stork NE, McBroom J, Gely C, & Hamilton AJ (2015) New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc Natl Acad Sci U S A* 112(24):7519-7523.
**14.** Hedges SB, Marin J, Suleski M, Paymer M, & Kumar S (2015) Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution* 32(4):835-845.
**15.** Jarvis ED (2016) Perspectives from the Avian Phylogenomics Project: Questions that Can Be Answered with Sequencing All Genomes of a Vertebrate Class. *Annual Review of Animal Biosciences* 4(1):45-59.
**16.** Shen X-X, Hittinger CT, & Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature ecology & evolution* 1(5):126-126.
**17.** Burki F (2014) The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harbor Perspectives in Biology* 6(5):a016147.
**18.** Rinke C, *et al.* (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* 9(5):1038-1048.
**19.** Kim J, *et al.* (2017) Reconstruction and evolutionary history of eutherian chromosomes. *Proc Natl Acad Sci U S A* 114(27):E5379-E5388.
**20.** Paten B, Zerbino DR, Hickey G, & Haussler D (2014) A unifying model of genome evolution under parsimony. *BMC Bioinformatics* 15:206-206.
**21.** Kumar S, Dudley JT, Filipski A, & Liu L (Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends in Genetics* 27(9):377-386.
**22.** Infante JJ, Dombek KM, Rebordinos L, Cantoral JM, & Young ET (2003) Genome-wide amplifications caused by chromosomal rearrangements play a major role in the adaptive evolution of natural yeast. *Genetics* 165(4):1745-1759.
**23.** Pecl GT, *et al.* (2017) Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science* 355(6332).
**24.** Casillas S & Barbadilla A (2017) Molecular Population Genetics. *Genetics* 205(3):1003-1035.
**25.** Steiner CC, Putnam AS, Hoeck PEA, & Ryder OA (2013) Conservation Genomics of Threatened Animal Species. *Annual Review of Animal Biosciences* 1(1):261-281.
**26.** Kress WJ & Erickson DL (2012) DNA Barcodes: Methods and Protocols. *DNA Barcodes: Methods and Protocols*, eds Kress WJ & Erickson DL (Humana Press, Totowa, NJ), pp 3-8.
**27.** Droege G, *et al.* (2016) The Global Genome Biodiversity Network (GGBN) Data Standard specification. *Database* 2016:baw125-baw125.
**28.** Carlson R (2016) Estimating the biotech sector's contribution to the US economy. *Nature Biotechnology* 34:247.
**29.** Bouchie A (2016) White House unveils National Microbiome Initiative. *Nat Biotech* 34(6):580-580.
**30.** Kim J, *et al.* (2013) Reference-assisted chromosome assembly. *Proc Natl Acad Sci U S A* 110(5):1785-1790.
**31.** Lewin HA, Larkin DM, Pontius J, & O'Brien SJ (2009) Every genome sequence needs a good map. *Genome Res* 19(11):1925-1928.
**32.** Mounce R, Smith P, & Brockington S (2017) Ex situ conservation of plant diversity in the world's botanic gardens. *Nature Plants* 3(10):795-802.
**33.** Marlow J, *et al.* (2017) Opinion: Telepresence is a potentially transformative tool for field science. *Proceedings of the National Academy of Sciences* 114(19):4841-4844.
**34.** McCluskey K, *et al.* (2017) The U.S. Culture Collection Network Responding to the Requirements of the Nagoya Protocol on Access and Benefit Sharing. *mBio* 8(4):e00982-00917.
**35.** Rinke C, *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431-437.
**36.** Stephens ZD, *et al.* (2015) Big Data: Astronomical or Genomical? *PLoS Biology* 13(7):e1002195.
**37.** Novak AM, *et al.* (2017) Genome Graphs. *bioRxiv*.
**38.** Seberg O, *et al.* (2016) Global Genome Biodiversity Network: saving a blueprint of the Tree of Life – a botanical perspective. *Annals of Botany* 118(3):393-399.
**39.** Droege G, *et al.* (2014) The Global Genome Biodiversity Network (GGBN) Data Portal. *Nucleic Acids Research* 42(D1):D607-D612.
**40.** Koepfli K-P, Benedict Paten, Scientists tGKCo, & O'Brien SJ (2015) The Genome 10K Project: A Way Forward. *Annual Review of Animal Biosciences* 3(1):57-111.
**41.** Consortium iK (2013) The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *Journal of Heredity* 104(5):595-600.
**42.** Scientists GCo (2014) The Global Invertebrate Genomics Alliance (GIGA): Developing Community Resources to Study Diverse Invertebrate Genomes. *Journal*

*of Heredity* 105(1):1-18.
**43.** National Academies of Sciences E & Medicine (2017) *A Proposed Framework for Identifying Potential Biodefense Vulnerabilities Posed by Synthetic Biology: Interim Report* (The National Academies Press, Washington, DC) p 51.
**44.** Nobre CA, *et al.* (2016) Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. *Proceedings of the National Academy of Sciences* 113(39):10759-10768.