1-18-2022

Standards recommendations for the Earth BioGenome Project

Mara K. Lawniczak
Wellcome Sanger Institute

Richard Durbin
University of Cambridge

Paul Flicek
Wellcome Sanger Institute

Kerstin Lindblad-Toh
Uppsala University

Xiaofeng Wei
China National GeneBank

NSUWorks Citation

Lawniczak, Mara K.; Richard Durbin; Paul Flicek; Kerstin Lindblad-Toh; Xiaofeng Wei; John M. Archibald; William J. Baker; Katherine Belov; Mark L. Blaxter; Tomas Marques-Bonet; Anna K. Childers; Jonathan A. Coddington; Keith A. Crandall; Andrew J. Crawford; Robert P. Davey; Federica Di Palma; Qi Fang; Wilfried Haerty; Neil Hall; Katherine J. Hoff; Kerstin Howe; Erich D. Jarvis; Warren E. Johnson; Rebecca N. Johnson; Paul J. Kersey; Xin Liu; Jose V. Lopez; Eugene W. Myers; Olga Vinnere Pettersson; Adam M. Phillippy; Monica F. Poelchau; Kim D. Pruitt; Arang Rhie; Juan Carlos Castilla-Rubio; Sunil Kumar Sahu; Nicholas A. Salmon; Pamela S. Soltis; David Swarbreck; Françoise Thibaud-Nissen; Sibo Wang; Jill L. Wegrzyn; Guojie Zhang; He Zhang; Harris A. Lewin; and Stephen Richards. 2022. "Standards recommendations for the Earth BioGenome Project." *Proceeding of the National Academy of Sciences of the United States of America* 119, (4). doi:10.1073/pnas.2115639118.

SPECIAL FEATURE: PERSPECTIVE

Standards recommendations for the Earth BioGenome Project

Mara K. N. Lawniczak[a], Richard Durbin[a,b], Paul Flicek[a,c], Kerstin Lindblad-Toh[d,e], Xiaofeng Wei[f], John M. Archibald[g], William J. Baker[h], Katherine Belov[i], Mark L. Blaxter[a], Tomas Marques Bonet[j,k,l,m], Anna K. Childers[n], Jonathan A. Coddington[o], Keith A. Crandall[p], Andrew J. Crawford[q], Robert P. Davey[r], Federica Di Palma[s], Qi Fang[t], Wilfried Haerty[u], Neil Hall[s,u], Katharina J. Hoff[v], Kerstin Howe[a], Erich D. Jarvis[w,x], Warren E. Johnson[y,z], Rebecca N. Johnson[o], Paul J. Kersey[aa], Xin Liu[f], Jose Victor Lopez[bb], Eugene W. Myers[cc], Olga Vinnere Pettersson[dd], Adam M. Phillippy[ee], Monica F. Poelchau[ff], Kim D. Pruitt[gg], Arang Rhie[ee], Juan Carlos Castilla-Rubio[hh], Sunil Kumar Sahu[f,t], Nicholas A. Salmon[a], Pamela S. Soltis[ii], David Swarbreck[u], Françoise Thibaud-Nissen[gg], Sibo Wang[f], Jill L. Wegrzyn[jj,kk], Guojie Zhang[ll,mm,nn,oo], He Zhang[pp], Harris A. Lewin[qq,rr], and Stephen Richards[qq,1]

Edited by Gene Robinson, Entomology, University of Illinois at Urbana–Champaign, Urbana, IL; received September 10, 2021; accepted November 3, 2021

A global international initiative, such as the Earth BioGenome Project (EBP), requires both agreement and coordination on standards to ensure that the collective effort generates rapid progress toward its goals. To this end, the EBP initiated five technical standards committees comprising volunteer members from the global genomics scientific community: Sample Collection and Processing, Sequencing and Assembly, Annotation, Analysis, and IT and Informatics. The current versions of the resulting standards documents are available on the EBP website, with the recognition that opportunities, technologies, and challenges may improve or change in the future, requiring flexibility for the EBP to meet its goals. Here, we describe some highlights from the proposed standards, and areas where additional challenges will need to be met.

Earth BioGenome Project | genomics | ethics | genome assembly

The Earth BioGenome Project (EBP) aims to sequence and characterize reference genomes for each known eukaryotic species (1, 2). The project initiated from early discussions including a working group in 2017 (3), and formally launched at the Wellcome foundation EBP/Vertebrate Genomes Project (VGP) workshop in London on November first, 2018. This ambitious vision builds on previous successes, starting with sequencing the human genome (4) and those of associated model organisms (5–9), through projects completing tens and more recently hundreds of genome sequences (1). A major positive outcome of these efforts—especially those of the VGP—has been a renewed focus on producing high-quality reference genomes, suitable for use as long-term resources faithfully representing the DNA sequence of each species in chromosomal scaffolds with few gaps. One recent outcome of this effort—led by A.R., Shane McCarthy, Olivier Fedrigo, and many members of the VGP—is the publication of the assembly lessons learned from the generation of 16 very high-quality genome assemblies representing major vertebrate lineages (10). Companies such as Pacific Biosciences (PacBio) have worked hand-in-hand with these efforts to improve the quality of long-read sequencing for fast and accurate genome assembly (e.g., ref. 11) and low-input requirements to enable sequencing of species

Author contributions: M.K.N.L.., R.D., P.F., K.L.-T., X.W., J.M.A., W.J.B., K.B., M.L.B., T.M.B., A.K.C., J.A,C., K.A.C., A.J.C., R.P.D., F.D.P., Q.F., W.H., N.H., K.J.H., K.H., E.D.J., W.E.J., R.N.J., P.J.K., X.L., J.V.L., E.W.M., O.V.P., A.M.P., M.F.P., K.D.P., A.R., J.C.C.-R., S.K.S., N.A.S., P.S.S., D.S., F.T.-N., S.W., J.L.W., G.Z., H.Z., H.A.L., and S.R. wrote the paper; M.K.N.L. chaired the Sample Collection and Processing Committee, and led the development of those standards; R.D. chaired the Assembly Standards Committee and led the development of assembly standards; P.F. chaired the Annotation Standards Committee and led the development of annotation standards; K.L.-T. chaired the Analysis Committee and led the development of analysis standards; X.W. chaired the IT Standards Committee and led the development and design of IT standards; J.M.A., W.J.B., K.B., M.L.B., T.M.B., A.K.C., J.A.C., K.A.C., A.J.C., R.P.D., F.D.P., Q.F., W.H., N.H., K.J.H., K.H., E.D.J., W.E.J., R.N.J., P.J.K., X.L., J.V.L., E.W.M., O.V.P., A.M.P., M.F.P., K.D.P., A.R., J.C.C.-R., S.K.S., N.A.S., P.S.S., D.S., F.T.-N., S.W., J.L.W., G.Z., H.Z., and H.A.L. contributed to Earth BioGenome Project standards development; S.R. led standards development for all subgroups and was Chair of the Earth BioGenome Project International Scientific Committee for the development of these standards and contributed to Earth BioGenome Project standards development.

Competing interest statement: P.F. is a member of the scientific advisory boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: srichards@ucdavis.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2115639118/-/DCSupplemental.

Published January 18, 2022.

with minute individual size (12). The major lessons from these efforts are the necessity of long reads, the need to sequence single individuals to reduce haplotype complexity, the necessity of long-range linked read technology, such as Hi-C to place assembled sequences in a chromosomal context, high-quality QC pipelines, and iterative improvement of assembly pipelines for continued increases in quality and decreases in cost.

The five EBP technical standards committees (Sample Collection and Processing, Sequencing and Assembly, Annotation, Analysis, and IT and Informatics) have built on these and other advances to recommend quality standards and organizational principles at all steps of the process from ethical sample collection to sharing analyses, to guide forward progress (13). In most cases the committees have refrained from highly specific recommendations, as the rapid pace of research in all areas often changes the best approach within months. We have produced recommendations rather than a definitive "how to" set of commandments, and we will update these recommendations annually as EBP projects progress and technologies improve. Here we describe these recommendations, also found on the EBP website (13).

Results

Sample Collection and Processing Recommendations. ***Ethical collecting.*** It is of utmost importance that specimens and data contributing to the EBP are legally obtained and projects are ethically conducted. All collection activities should carefully follow institutional and national protocols, including but not limited to prior informed consent, compliance with the Nagoya Protocol of the Convention for Biological Diversity (14), and endangered species legislation. Sample collectors should ensure that all local and national permissions for collection are in place, and that a record of these permissions is maintained for referral in the case of questions on a specimen's legality. These permissions can vary widely among countries, so it is beyond the scope of this document to summarize them. A complicating factor is that some specimens will be collected and moved out of their country of origin for sequencing. Where the country of origin has implemented the Nagoya Protocol with local laws, they must be followed. Again, precise guidance on how to follow the Nagoya Protocol is beyond the scope of this document, as countries interpret the protocol differently. We recommend that organizations receiving samples ensure they have permissions within a Material Transfer Agreement to pass the samples on if there is any anticipation that such transfer might be required. The reader is also directed to additional guidance presented by the EBP's Ethical, Legal, and Social Issues committee (15). Beyond conforming to rules and regulations, collecting methods must be ethical. For example, projects should consider sampling strategies that aim to avoid overcollection, such as lineage focused bio-blitzes supervised by a group of taxonomic experts.

Identifying EBP phase 1 family-representative target species: Open collation of community proposals and genomes underway. There are over 9,000 taxonomic families within Eukaryota, including 6,470 Metazoa, 1,052 Chromista, 757 Fungi, 964 Plantae, and 221 Protozoa in the Catalogue of Life checklist (www.catalogueoflife.org/annual-checklist/2019). An important collective step toward EBP ambitions is to gather the list of target species proposed to be family representatives to fulfill the phase 1 goal of a high-quality reference genome for every eukaryotic family (2). To provide greater flexibility in achieving phase 1 goals, we suggest that multiple species for each family should be proposed and collected simultaneously.

As much as possible, this process should be globally transparent and open to input from the wider community that may not be actively generating reference genomes but who will benefit from their availability. To assist with this transparency and the need for community input, we have created an open EBP Family_Reference_Proposals spreadsheet to document proposals for target taxa. Please note there are two tabs available for community entries: "Family Reference Suggestions" collates recommendations from the community on ideal family-level target species, and "Family Reference Projects" records reference genome sequencing projects already underway (more below) that are targeting a relatively small and clearly defined set of taxa. For larger-scale projects that aim to sequence hundreds or thousands of species, target lists are likely to undergo revision as projects proceed. Therefore, we have created two distinct approaches (for small defined projects or larger projects with funding) to facilitate global transparency of target species lists depending on the size of the project. In both cases, species and their associated projects will be searchable and displayed on the "Genomes on a Tree" service (https://goat.genomehubs.org/).

Considerations for selection of phase 1 family representatives. A variety of factors should be considered when proposing a species representative of a family, and these are listed here in order of importance: 1) permissions and availability: sampling is achievable taking into consideration permissions and legal requirements; 2) community value: of broad community use and value (this could be assessed through surveys oriented toward target communities, e.g., https://docs.google.com/forms/d/e/1FAIpQLSfOA4GRA3mkLLnN3fng9ZM8RdMIHusqPqfMDPePy8a0sE72Ug/viewform for the Darwin Tree of Life [DToL] project); 3) publicly registered: the species is registered with its current name and taxonomy in a publicly available database (we recommend the National Center for Biotechnology Information [NCBI] Taxonomy Database https://www.ncbi.nlm.nih.gov/taxonomy) and assigned a Tree of Life ID (TOLID; see below) to assist with tracking name and taxonomy changes over time; 4) physical size: considering today's technology limits, we propose a requirement for 10 samples, each weighing more than 10 mg per 1 Gb of genome size, for animals, fungi, and protists, and 100 mg per 1 Gb of genome size for plants, including a minimum of three samples to support the three platforms of long read, Hi-C, and transcriptome sequencing; 5) species representative: generally considered to be a "good" biological species (not from a known species complex) and if possible, sampled from or near the type locality, which will help make the sample robust to taxonomic ambiguities or changes; 6) genome size: where genome sizes and ploidy are known, prioritizing species with smaller, diploid genomes (because costs of data generation are projected to fall and our ability to assemble large high-repeat content genomes will improve in future EBP phases); 7) taxonomic stability: not currently subject to disagreement or revision.

Metadata to place each genome in its biological and ecological context. The assemblies for EBP species must be accompanied by robust and comprehensive metadata, including the collection event and collector identification in a common set of agreed metadata fields. Each contributed specimen should be identified to species level by a taxonomic expert and, whenever possible, material from the same specimen should be

independently DNA barcoded using appropriate markers and the data deposited into BOLD (www.boldsystems.org) and the International Nucleotide Sequence Database Collaboration (16) (INSDC, https://www.insdc.org) databases (GenBank, European Nucleotide Archive [ENA], DNA Databank of Japan [DDBJ]). These DNA barcodes will serve both to ensure that species with reference genomes have assembly-derived barcodes matching independently generated DNA barcode data and that the reference genome corresponds to the named taxon (i.e., that no sample swaps have occurred during processing). Given the importance of ensuring the specimen is truly representative of the species to which it is assigned, where these standards are not achievable it is advisable to substitute a different representative for the family.

Consensus data practices should be established early in the project planning to ensure that metadata fields and terms are standardized, and ideally that dedicated tools are provided to assist with capture and brokering of metadata into INSDC and project-level repositories [e.g., COPO, in use by the United Kingdom Darwin Tree of Life project (https://copo-project.org (17)]. Coordination of best practice in metadata collection is delivered by the EBP IT/Informatics Best Practices committee, which draws on the efforts to standardize metadata collected by the DToL project (https://darwintreeoflife.org). We recommend when possible that high-quality, informative images should accompany each contributed specimen, and these should be made publicly available. Ideally, these images will be deposited in the BOLD database to accompany the DNA barcode (and in a yet-to-be designed portal for EBP that supports access to all metadata for each sequenced species).

Specimens used to generate reference genomes should be vouchered. These vouchers should take as many complementary forms as is practical, including museum voucher specimens (unless the whole specimen is required for sequencing or legal/ethical issues prevent sacrificing the source individual), tissue vouchers (discussed further below), viably frozen cell lines, image vouchers, and molecular vouchers of extracted RNA and DNA. Vouchers should, wherever possible be deposited in publicly accessible collection facilities located in the country of origin of the specimen, or where there is excess material, possibly spread across multiple repositories. It is recommended that subsamples and sample derivatives be stored in a Global Genome Biodiversity Network member institution and linked to the voucher using a unique identifier. Fig. 1 shows an example of a collected species during physical processing on dry ice, which is recommended to maintain high molecular weight DNA and RNA integrity.

EBP sample collection in the future. Current best-practice assembly guidelines are to generate a combination of data types, including long-read (PacBio HiFi, and Oxford Nanopore Technologies long-reads [ONT]), long-range (Hi-C paired reads and perhaps linked reads or Bionano optical maps), and transcriptomic (RNA-seq Illumina short read, PacBio Iso-seq, or ONT cDNA-PCR) data from the same specimen wherever possible, aiming for the heterogametic sex where relevant. EBP partners are developing protocols that support minimal extraction of material sufficient for any of these types of data generation. Furthermore, current standard operating procedures for nucleic acid extraction for genomics effectively discard material that may be useful in the future, such as proteins and metabolites, and do not explicitly preserve the microbiome or live cells. While data generation from these materials is currently out of scope, this is unlikely to be true in years to come and the retention of relevant material to add additional layers of data to the high-quality reference genomes would be prudent. Thus, for specimens where material is available in excess, considerations should be given to appropriate storage of additional samples to future-proof these specimens as much as possible. For specimens where all material is used in the data generation, perhaps typically discarded supernatants should be retained for future investigations.

We also encourage activities that simplify and streamline standard procedures, such that it becomes easier to "containerize" extraction, sequencing, and assembly activities. Containerization means a world in which a shipping container or crate could house everything needed to go from sample to sequence and would build capacity in low- and middle-income countries, often harboring the greatest biodiversity. This will allow all nations to deliver to EBP goals while building capacity and expertise and enhancing global science.

Sequencing and Assembly Recommendations. The primary goal is to obtain for each species a sufficiently complete, accurate, and contiguous representative genome sequence that can provide a reference for further genomic analysis. This represents a single haplotype at each locus, and so can be provided by a single haploid assembly from an inbred or outbred individual, whatever the ploidy. Representation of alternate haplotypes from the sampled individual, if outbred, is desirable to illustrate genetic variation, but secondary.

Quantitative assembly standards. Where sufficient DNA and material is available from a single individual (or clonal colony), currently more than around 100 ng DNA per gigabase genome, we propose a minimum reference standard of 6.C.Q40 (i.e., >1 Mb NG50 contig continuity and chromosomal scale NG50 scaffolding, with less than 1/10,000 nucleotide error rate (see *SI Appendix*, Table S1, taken from ref. 10, for notation and further information). When the chromosome NG50 is smaller than a megabase this will be C.C.Q40, with "C" denoting having achieved full chromosome lengths. This standard is now

Fig. 1. An example of the documentation that should occur as a sample is being processed. The SPECIMEN_ID (the Natural History Museum, London [NHMUK] barcode under the fly) is photographed alongside the specimen and the barcoded tubes into which different samples of that specimen will be placed. The metadata tracking sheet would have three entries for this fly, where collection-related information would be identical, but tissue type and tissue size would vary (e.g., head, thorax, and abdomen, each in a separate tube). Image credit: M.K.N.L.

achievable with reasonable effort using multiple alternative long-read–based strategies, and is comparable to the standard of the human GRCh37 (hg19) and mouse GRCm38 (mm10) references that have proven to be of high utility for many years. Alongside the contiguity and error-rate goals, we propose the following additional criteria (*SI Appendix*, Table S1): fewer than 5% false duplications, more than 90% kmer completeness, more than 90% of the sequence assigned to candidate chromosomal sequences, more than 90% single copy conserved genes [e.g., as inferred using BUSCO (18)] complete and single copy, and more than 90% of transcripts from the same species mappable. Suggestions of and links to standard tools for measuring these metrics are provided on the EBP website (13).

We believe that these are achievable goals for most and perhaps ultimately all species for which large enough high-quality samples can be obtained. However, we recognize that for many reasons (e.g., sample quality, very large genomes, polyploidy, cost expediency) they may not be met in the first instance. Interim references that do not meet the standard can be very useful and should be valued. Nevertheless, there will be a continuing EBP goal to revisit these and bring them up to the target standard as that becomes practical.

For species for which there is more limited DNA available per individual, additional technology development is required both in generation of sequence from limiting input, and in the assembly of genomes from multiple individuals. We expect that improved approaches will bring more and more "small" species into group above. In the meantime, we propose the interim standard outlined in the 4.5.Q40 column of *SI Appendix*, Table S1. For as yet uncultured single-celled eukaryotes, we expect that a metagenomics-like standard will be determined, based on experience in the prokaryotic and emerging microeukaryote metagenome fields. No specific recommendations are made at this time.

Additional requirements. In addition to the quantitative requirements above, our experience has shown that currently all (combinations of) automated processes generate assemblies with a variety of remaining errors, some of which are relatively easy to address and should be corrected before submission to public databases (19). The full recommendations also propose that projects meet a wider set of quality control criteria, including separation of the organelle genomes of target species from their nuclear genomes, separation of the genomes of any symbionts or cobionts, explicit identification of primary and alternate haplotype assemblies, that the sequence includes only A, C, G, T, and N base calls, and that sequences should not begin or end with Ns. The reader is referred to the full recommendations on the EBP website for full details (13).

Identification and naming of chromosomal-scale scaffolds can be achieved with Hi-C two-dimensional maps and comparison to existing karyotyping and linkage maps where available. If chromosome naming already exists for a given species, it should be reflected in the new assembly so long as the chromosomes match to those of the prior naming. If no previous names exist, we recommend naming chromosomes by size, taking into account scaffolds that can be assigned to belong to a certain chromosome, but could not be unambiguously placed (unlocalized scaffolds). An alternative that is applicable in some cases is to name chromosomes after those in a closely related species with an established nomenclature; we only recommend this if there are no major interchromosomal rearrangements identified between the species (i.e., all chromosomes are in one-to-one correspondence, but may have within chromosome rearrangements).

INSDC project structure and nomenclature. For a reference genome to count toward the EBP goals, it must be submitted to the INSDC Genomes Division for open access use by the scientific community. When this submission is made, the assembly is associated with a "data" BioProject object, which can be part of a hierarchy by being assigned to an "umbrella" BioProject. We suggest the structure shown in Fig. 2, with a data project per assembly, an umbrella project for the sample (note that under this there may be assemblies of separate symbiont or cobiont species), and then above an umbrella BioProject corresponding to the overall project. When creating an umbrella project in INSDC, please request that it be linked to the EBP BioProject object (PRJNA533106, https://www.ncbi.nlm.nih.gov/bioproject/533106). If an assembly contributes to other larger-scale efforts (e.g., Global Invertebrate Genomics Alliance or VGP), you can request it also be linked to the respective umbrella BioProjects. These requests go through an approval process.

To be submitted to the INSDC databases to a BioProject, an assembly needs to be assigned to a TaxonID entry in the NCBI Taxonomy database. Although TaxonID identifiers can be created for informal taxa, such as *Maylandia* sp. "pearly", we would like EBP genomes to be associated with taxonomically valid species names, and we urge EBP-affiliated projects to work with appropriate taxonomists to identify samples to a species and, where necessary, to establish the species name in the standard manner in the literature.

Furthermore, in addition to the numerical identifiers generated for assemblies by the public databases, we request projects to adopt the TOLID standard short nomenclature for samples and assemblies, as used by the VGP and DToL. This takes the form .. For example, ilAlcRepa1.1 identifies the first assembly of an insect lepidopteran (two characters "il"), Alcis (three characters "Alc"), repandata (four characters "Repa") individual one ("1"), assembly version one (0.1). Unique TOLID designations have been generated to cover all ~485,000 species that already have data in INSDC or are found in Britain and Ireland. Other species can be added on request. Details of the two-letter prefix assignment, which partitions the tree of life, and of the currently assigned identifiers, are at https://gitlab.com/wtsi-grit/darwin-tree-of-life-sample-naming. A server to view and assign unique individual identifiers for specimens is available at https://id.tol.sanger.ac.uk/, with instructions for registering new users and projects.

Ongoing EBP sequencing and assembly challenges. Given rapidly evolving technologies and bioinformatic tools, we have deliberately avoided defining hard and fast rules about how genomes should be sequenced and assembled. Continued efforts are still required to utilize future sequencing technology improvements by combining them with new assembly algorithms to continue to improve both assembly quality and reduce costs for future phases of the EBP. Despite major achievements (e.g., ref. 19), improved technologies are still required for the large number of eukaryotic species with very small individuals, and genomics of unculturable single-celled eukaryotes remains a major challenge requiring innovation before large scale production is advisable.

Annotation Recommendations. High-quality annotation is required to transform reference genome sequences into actionable knowledge. In the most expansive definition, genome annotation is an

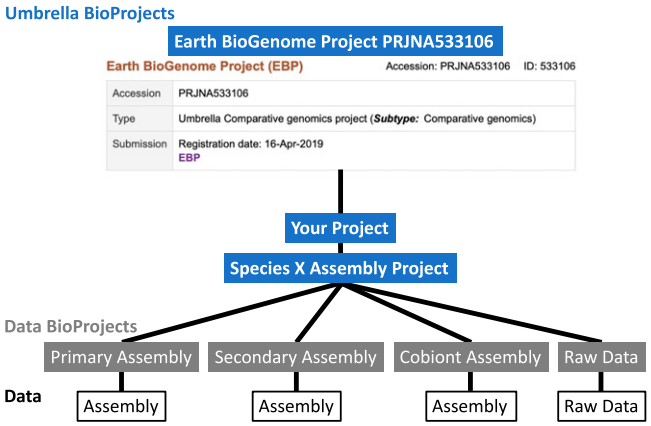

Fig. 2. Recommended EBP INSDC assembly BioProject submission structure. Alternatives on this theme could include maternal assembly and paternal assembly at the DataBioProject levels, when complete genome assemblies are generated for both haplotypes of a diploid sexually reproducing species. "Your Project" represents a project sequencing multiple genomes, such as the VGP or DToL.

accounting of the role and history of each base pair in the genome. Unlike genome assembly for which completeness and accuracy can be exactly defined, for most eukaryotic species it is likely impossible that an annotation can be constructed that would account for all possible conditions and sequence variation. However, the completeness and quality of gene annotation are substantially higher when transcriptomic data from the same species (ideally the same individual that has had its genome sequenced) is incorporated. For this reason, we recommend collection of transcriptomic data for all sequenced species. Indeed, some annotation pipelines require such data.

Genome features to be annotated. We propose that required annotation for all genomes include the feature classes listed in the left-hand column of *SI Appendix*, Table S2. Additionally, the feature classes listed in the right-hand column may be annotated in some genomes.

For genes, the supporting evidence shall be provided at the level of gene or transcript annotation (see *SI Appendix*, Table S3 for evidence types) if possible or at the level of feature set if not. Note that hybrid annotation approaches are common using multiple evidence categories across a single transcript or gene. In general, all the evidence categories used for annotation should be recorded.

As described in *SI Appendix*, Table S3, evidence data from other species (i.e., homology information) is commonly used. The value of this information is a function of the type of information and the sequence divergence between the target genome and the source of the evidence. The use of transcriptomic data from other species for annotation is possible but should be carefully evaluated because alignment errors across will negatively impact annotation accuracy.

As much as possible, coding and noncoding genes should be assigned a molecular function and, when assigned, the method for assigning such function must be specified. Common methods for assigning function include homology, orthology to a reference genome, matching to hidden Markov models, and gene ontology terms.

Annotation description. For a given genome, similar features will be collected into feature sets that will be further described by the software and parameters used to create the annotation. The full annotation set for each feature class may consist of one or more feature sets. For example, a full protein-coding annotation set may have a generalized feature set annotated with a particular methodology and a specific feature set, such as immune genes, annotated with a different methodology. Appropriate descriptors thus must exist at the level of the annotation set, the feature set and, where possible, each annotated feature. Descriptors for annotation sets and feature sets are largely interchangeable and are reported together.

Annotation set and feature set descriptors are:

- Accession number(s) of all transcriptomic data used in the annotation. The use of each data within the annotation process must also be specified. These data may be specifically generated for the annotation or previously available.
- Defined protein sequence data sets ideally specified by, for example, UniProt release number or the annotation set translated specific by, for example, Ensembl release number.
- Accession numbers of genome assemblies used for alignments and information about any annotation on those assemblies used for projection.
- Release numbers of supporting information used for annotation such as repeat libraries or hidden Markov model profiles.
- Software versions and parameter settings used.

Annotation feature descriptors are:

- Where annotated features are based on accessioned evidence sequences such as protein sequences, these accession numbers should be associated with the annotated object.
- Where annotated features, such as genes, are named the process for naming must be described and the supporting information used for the naming (e.g., Hugo Gene Nomenclature Committee access date, reference genome annotation, and so forth) must be specified.

Annotation sharing and coordination. All annotation must be freely available without any restrictions on further use (equivalent to public domain or CC0 licensing).

Annotation must be mappable to sequence records contained in the INSDC, although the provider may use different nomenclature for a given sequence, such as "chromosome 1," as long as the corresponding INSDC sequence identifier is incorporated into the annotation release. Annotation sets will be named based on the formal assembly name. Annotation shall be provided in standard file formats such as GFF3 with standard descriptive attribute tags.

Requirements for annotation providers. Annotation meeting EBP standards can be produced either by centralized annotation services, such as those provided by Ensembl (20) at European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) and RefSEq. (21) at the NIH's NCBI or by other groups, either as part of the generation of the genome sequence or by groups provided distributed annotation services that are targeted by phylogenetic scope or other reason.

For the centralized annotation services, genomes will only be annotated after they have been submitted to an INSDC database (GenBank, ENA, DDBJ), have been accessioned with an GCA_XX identifier, and have been publicly released. Supporting transcriptome data for annotation (including transcriptome assemblies where they have been used) must also be submitted, accessioned, and released via a recognized archival database, such as ArrayExpress, Gene Expression Omnibus, or INSDC. Other groups proposing themselves as central

annotation services in the context of EBP are expected to follow these requirements of annotation and use of openly available genome sequences and evidence data.

Estimating annotation quality. Annotation quality is an estimate of completeness and accuracy. It should be based on multiple factors, including: 1) deep evolutionary conservation; 2) comparative data from closely related species in the same clade, ideally with at least one gold standard reference for the clade to compare to; and 3) expressed sequence evidence from the species or individual being annotated. Moreover, annotation quality assessment encompasses both the need to capture the loci (i.e., the completeness of the gene set) as well as the details of the transcript set.

For the foreseeable future, annotation quality will be measured in a relative scale as compared to other genomes within the same clade. Finished genome annotation is unlikely to be possible for nearly all species. When the quality of annotation for given species has fallen substantially below the relative level for similar species, it should be prioritized for reannotation.

Basic statistics for protein-coding and nonprotein-coding genes should be produced, including the number of genes and, for protein coding genes, the number of exons per gene.

Protein-coding gene annotation should also be compared to gene-based assessments of the assembly quality. For example, BUSCO results from the final gene predictions should be compared to the estimated BUSCO score on the genome assembly to ensure that these metrics are comparable. Repeat annotation should be compared with a measure of repeat completeness in the genome assembly, such as the long terminal repeat index value (22).

Ongoing annotation challenges. A number of basic and applied genome annotation research challenges in genome annotation remain to be addressed, including: 1) the creation of robust, community-defined measures of annotation quality; 2) increasing portability and automation of current genome annotation pipelines; and 3) developing training and providing funding to access cloud-based or deployable annotation tools to diversify the community of scientists who can use and contribute to genome annotation and the development of annotation tools.

Analysis Recommendations. One of the challenges of capturing analysis standards for the diverse needs of a genomics community is that there are many scientific goals and practical uses for genome data. Substantial variation in genome size and complexity, life history, and other features means that analysis tools may not work on all species and groups of species. Here we highlight some key analysis areas, across a broad set of questions related to genome evolution, population genomics, and biodiversity, and list some currently available tools (*SI Appendix*, Table S4). Many of these tools will need to be altered to handle larger datasets and to ask novel questions only answerable at the scale of the EBP.

Key types of analysis. Genome analysis targets a broad set of questions related to genome evolution, population genomics, and biodiversity. We outline some of the more common research areas:

1. Alignments and synteny analysis of related species: An alignment forms the basis for comparative analysis across species, allowing comparison of the same genome features in many species. Multiple tools for sequence alignment exist, but most need to be scaled up to handle the expected thousands of genomes.
2. Repeat content and evolution: Catalogs of simple sequence repeats and mobile elements will be generated during genome annotation. Repeat elements may be exapted into novel regulatory elements or transcripts that may regulate gene function. Based on their similarity and distribution within and between genomes, it is possible to form and test hypotheses of the evolution of repeat content and their contributions to adaptation.
3. Partial or whole-genome duplication: Genome size is a function of gain and loss of sequence, including genes, repeats, segmental, and genome duplications. Many evolutionary lineages have undergone whole or partial genome duplication. Such duplications allow the divergent evolution of duplicated genes. Using synteny between species and internal similarity of genome sections allows for analysis of gene gain and loss.
4. Updating the evolutionary tree: Species trees provide a comparative framework for multiple analyses, such as inferring evolutionary constraint, detecting positive selection, delineating species boundaries/hybridization events, and estimating evolutionary relationships. The tree may be built using different types of genomic sequence data, including ancestral repeats, genes, and fourfold degenerate codons. While our understanding of family-level relationships across the tree of life is fairly stable, more closely related species can vary significantly across genomic regions through processes of hybridization, incomplete lineage sorting, and horizontal gene transfer.
5. Evolutionary constraint: To identify functional elements in the genome of each species, evolutionarily constrained regions need to be identified. These will include not only transcribed protein-coding and RNA genes, but also regulatory elements. Given sufficient data (such as a large number of related species) constraint can be detected at the single-base level.
6. Analysis of gene content, gene family expansion/contraction, and selection on protein-coding genes: Both gene family expansions and contractions as well as positive selection on specific regions of a protein are important for species evolution. Typically, a majority of genes are 1:1 orthologs between larger species groups (~70% of mammalian genes). To detect gene family expansions, whole-genome alignments and synteny can be used. Such studies can also detect horizontal (lateral) gene transfer.
7. Analysis of noncoding transcripts: Noncoding transcripts have a key role in genome regulation and function, and should be identified as part of the genome annotation process. Noncoding transcripts typically evolve more rapidly than protein-coding genes, so performing analyses to study the evolution of noncoding transcripts in different species is important. Analysis can be performed either by direct comparison of transcripts between related species or by comparison using the whole-genome alignment/synteny.
8. Intraspecific variation, conservation, biodiversity, and adaptation: As more and more species become endangered or critically endangered, the need for genomic information to guide conservation efforts increases. To generate data for these analyses, whole genome resequencing data from multiple populations with 10 to 20 individuals each are ideal; however, important information can still be gleaned from the two haplotypes present in a single diploid individual (i.e., the genome reference). With multiple populations, not only can markers for managing populations be generated, but signals of adaptation can also be identified.

9. Environmental DNA and ecological sampling: eDNA analysis facilitates the identification of threatened and nonthreatened species within ecosystems. Currently, analyses typically use PCR amplicon metabarcoding, but the availability of whole genomes would allow unbiased sampling approaches. By generating a high-quality digital library, based on whole-genome sequences to underpin the identification of eDNA samples, the EBP can accelerate this work. Thus, it is important to deposit not just genome sequences into INSDC databases, but also derived barcodes to appropriate barcode public repositories (e.g., BOLD: www.boldsystems.org) to accelerate their growth and enable identification of species from nonbarcode sequences.

Supporting the EBP analysis mission. To support the EBP analysis mission, we note four key needs: 1) enhancing analysis tools to handle large datasets; 2) having sufficient computer resources and ways to share the data; 3) fostering collaboration between nongenomicists and genomicists; and 4) ensuring that the whole EBP community has access to tutorials for different types of genome analysis.

IT and Informatics Recommendations. ***Access and sharing policies.*** EBP supports the FAIR principles (findability, accessibility, interoperability, and reusability) (23) and we recommend that no restrictions are placed on data access by submitting consensus reference genomes and corresponding sample metadata to INSDC-managed data repositories. There are possible caveats that may arise from the Convention on Biological Diversity Access and Benefit Sharing protocols and national laws, but the EBP recommends open deposition in INSDC databases unless local laws make this impossible.

IT infrastructure. The EBP infrastructure requirements are large, not only for the genome assembly data that need to be archived, but also the storage and computing resources needed for intermediate analyses. These infrastructures need to be interoperable with repositories and data-management platforms that house EBP data and metadata. This will require consensus across platforms to facilitate clear and transparent information exchange. We recommend the development of a mechanism to share the infrastructure capabilities for each affiliated project (Fig. 3) with the aim of ensuring global participation.

IT and informatics areas requiring future research. The IT and informatics challenges are not conceptual but practical. While the computational resources required are large, they are certainly not intractable on the global scale. Thus, the practicalities of making pipelines and computational resources available to all is a significant organizational problem to be solved. A key factor for success will be the identification of distinct "modules" within the data processing and outputs of EBP to enable parallel research on key computational tasks without affecting overall progress. Some of these are very clear: for example, assembly software can be improved without negatively affecting downstream genomic analyses.

Discussion

This set of recommendations is the first from the EBP community, and specifically addresses phase 1 of the project. As might be expected, there is more clarity for the early steps in the process (sample collection, sequencing, and assembly) than for the later steps (annotation, analysis, global sharing at scale, and so forth). The comparative genomics analysis of tens of thousands

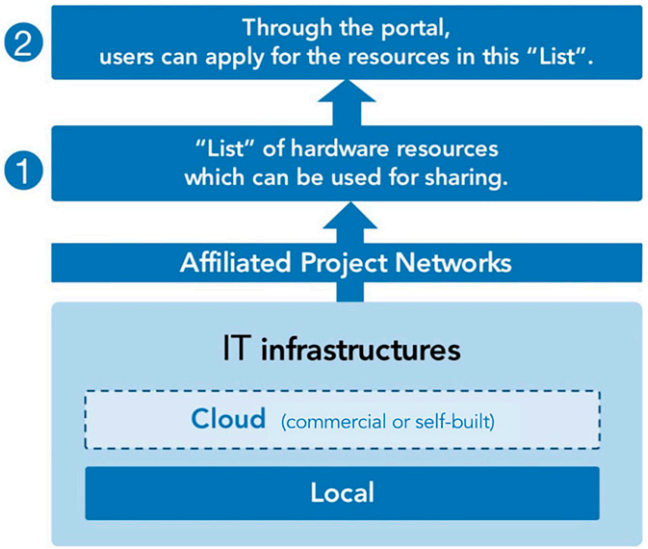

Fig. 3. Recommended future EBP data service portal. Researchers would easily apply for the hardware resources that are shared by the EBP affiliated project networks.

of reference genomes will require active exploration and collective learning once the phase 1 references are available. However, coordination in the analysis space is critical to accelerate returns to society and EBP members are committed to continued efforts in this area. We will revisit and update recommendations as the EBP progresses. For phases 2 and 3 we will have gained experience and improved technologies in sampling, sequencing, and analysis, to fill gaps where no recommendations can be made at the current time.

For individual groups or large consortia wishing to coordinate with the EBP, there are two general ways to contribute. For an individual laboratory, where the goal is to generate a reference genome worthy of being a long-term resource, we believe the recommendations here are critical to assuring success. Coordination of target species via the mechanisms, such as the EBP Family_Reference_Proposals spreadsheet described above, and rapid public release of references to the INSDC with genome note manuscripts will accelerate all fields. For large consortia, we recommend formally joining the EBP as described on the EBP website (13) and contributing groups of species coordinated with other EBP members.

Finally, we are excited that we are technically able to complete phase 1 reference genomes to an extremely high quality. We hope that these guidelines will influence and improve the quality of individual reference genome sequences, whomever generates them, to create annotated, analyzed, world-wide accessible references that will serve humanity for the foreseeable future.

Data Availability. There are no data underlying this work.

Acknowledgments

The work of K.D.P. and F.T.-N. was supported by the Intramural Research Program of the National Library of Medicine, NIH. K.L.-T. is a distinguished professor funded by the Swedish Research Council. J.A.C. is supported by the Global Genome Initiative of the National Museum of Natural History, Smithsonian Institution. E.D.J. is supported by the Howard Hughes Medical Institute (HHMI). P.F. is supported by Wellcome Grant WT108749/Z/15/Z and the European Molecular Biology Laboratory. J.L.W. is supported by National Science Foundation Grant

DBI:IIBR:CAREER #1943371. This work was supported in part by the US Department of Agriculture, Agricultural Research Service. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The US Department of Agriculture is an equal opportunity provider and employer.

[a]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom; [b]Department of Genetics, University of Cambridge, Cambridge CB3 0DH, United Kingdom; [c]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, United Kingdom; [d]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [e]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University 751 23 Uppsala, Sweden; [f]China National GeneBank, Shenzhen 518120, China; [g]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS B3H 4R2, Canada; [h]Department of Accelerated Taxonomy, Royal Botanic Gardens, Kew, Surrey TW9 3AE, United Kingdom; [i]School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, NSW 2006, Australia; [j]Institute of Evolutionary Biology, Consejo Superior de Investigaciones Científicas-Universitat Pompeau Fabra, Parc de Rechercha Biomédica Barcelona 08003 Barcelona, Spain; [k]Catalan Institution of Research and Advanced Studies 08010 Barcelona, Spain; [l]Centre Nacional d'Anàlisi Geonòmica - Centre for Genomic Regulation, Barcelona Institute of Science and Technology 08028 Barcelona, Spain; [m]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona 08193 Barcelona, Spain; [n]Bee Research Laboratory, Beltsville Agricultural Research Center, US Department of Agriculture, Agricultural Research Service, Beltsville, MD 20705; [o]Smithsonian Institution, National Museum of Natural History, Washington, DC 20560-0105; [p]Computational Biology Institute and Department of Biostatistics & Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052; [q]Department of Biological Sciences, Universidad de los Andes 111711 Bogotá, Colombia; [r]Engineering Biology, Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, United Kingdom; [s]Genome British Columbia, Vancouver, BC V5Z 0C4, Canada; [t]BGI-Shenzhen, Beishan Industrial Zone, Shenzhen 518083, China; [u]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, United Kingdom; [v]Institute of Mathematics and Computer Science, Center for Functional Genomics of Microbes, University of Greifswald 17489 Greifswald, Germany; [w]Vertebrate Genomes Lab, The Rockefeller University, New York, NY 10065; [x]HHMI, Chevy Chase, MD 20815; [y]Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Front Royal, VA 22630; [z]The Walter Reed Biosystematics Unit, Museum Support Center MRC-534, Smithsonian Institution, Suitland, MD 20746-2863; [aa]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; [bb]Halmos College of Arts and Sciences, Guy Harvey Oceanographic Center, Nova Southeastern University, Dania Beach, FL 33004; [cc]Department of Systems Biology, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany; [dd]Science for Life Laboratory in Uppsala, Uppsala University 75123 Uppsala, Sweden; [ee]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, MD 20894; [ff]National Agricultural Library, USDA Agricultural Research Service, Beltsville, MD 20705; [gg]National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894; [hh]Spacetime Ventures, São Paulo 05449-050, Brazil; [ii]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; [jj]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269; [kk]Institute for Systems Genomics, Computational Biology Core, University of Connecticut, Storrs, CT 06269; [ll]Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen 1165 Copenhagen, Denmark; [mm]China National Genebank, BGI-Shenzhen 518083 Shenzhen, China; [nn]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences 650223 Kunming, China; [oo]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences 650223 Kunming, China; [pp]BGI-Qingdao, BGI-Shenzhen 266555 Qingdao, China; [qq]University of California Davis Genome Center, University of California, Davis, CA 95616; and [rr]Department of Evolution and Ecology, University of California, Davis, CA 95616

1 H. A. Lewin *et al.*, The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2115635118 (2022).
2 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).
3 E. Pennisi, Sequencing all life captivates biologists. *Science* **355**, 894–895 (2017).
4 E. S. Lander *et al.*; International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5 R. H. Waterston *et al.*; Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
6 Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
7 M. D. Adams *et al.*, The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
8 *C. elegans* Sequencing Consortium, Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018 (1998).
9 A. Goffeau *et al.*, The yeast genome directory. *Nature* **387**, 5–6 (1997).
10 A. Rhie *et al.*, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
11 M. R. Vollger *et al.*, Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* **84**, 125–140 (2020).
12 S. B. Kingan *et al.*, A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes (Basel)* **10**, 62 (2019).
13 The Earth BioGenome Project Working Group, The Earth BioGenome Project Website, https://www.earthbiogenome.org (2021).
14 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. 1st ed. [ebook] (Secretariat of the Convention on Biological Diversity, Montreal, Montreal: United Nations, 2011). https://www.cbd.int/abs/text/ (Accessed 21 February 2015).
15 J. S. Sherkow *et al.*, Ethical, legal, and social issues in the Earth BioGenome Project. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115859119 (2022).
16 M. Arita, I. Karsch-Mizrachi, G. Cochrane, The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
17 F. Shaw *et al.*, COPO-linked open infrastructure for plant data, in *Semantic Web Applications and Tools for Life Science* (*SWAT4LS*, 2015), pp 181–182.
18 R. M. Waterhouse *et al.*, BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
19 K. Howe *et al.*, Significantly improving the quality of genome assemblies through curation. *Gigascience* **10**, giaa153 (2021).
20 B. L. Aken *et al.*, The Ensembl gene annotation system. *Database (Oxford)* **2016**, baw093 (2016).
21 K. D. Pruitt *et al.*, RefSeq: An update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
22 S. Ou, J. Chen, N. Jiang, Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
23 M. D. Wilkinson *et al.*, The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).