

Best-practice guidance for Earth BioGenome Project sample collection and processing: progress and challenges in biodiverse reference genome creation

Mara K. N. Lawnczak^{1,*}, Kevin M. Kocot², Jonas J. Astrin³, Mark Blaxter¹, Cibele G. Sotero-Caio¹, Katharine B. Barker⁴, Anna K. Childers⁵, Jonathan Coddington⁴, Paul Davis¹, Kerstin Howe¹, Warren E. Johnson⁴, Duane D. McKenna^{6,7}, Jeremy G. Wideman⁸, Olga Vinnere Pettersson⁹, Verena Ras^{10,11}, Bernardo F. Santos¹², and the Earth BioGenome Project Samples and Processing Subcommittee¹³

¹Tree of Life, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK

²Department of Biological Sciences & Alabama Museum of Natural History, University of Alabama, Tuscaloosa, AL 35487, USA

³Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig, Adenauerallee 127, 53113 Bonn, Germany

⁴National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, USA

⁵USDA, Agricultural Research Service, Beltsville Agricultural Research Center, Bee Research Laboratory, 10300 Baltimore Avenue, Beltsville, MD 20705, USA

⁶Department of Biological Sciences, University of Memphis, Memphis, TN 38152, USA

⁷Center for Biodiversity Research, University of Memphis, Memphis, TN 38152, USA

⁸Biodesign Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA

⁹SciLifeLab, National Genomics Infrastructure, Department of Immunology, Genetics and Pathology, Uppsala University, SE-751 05 Uppsala, Sweden

¹⁰Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa.

¹¹Department of Biodiversity and Conservation Biology, University of the Western Cape, Cape Town, 7701, South Africa.

¹²Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Center for Integrative Biodiversity Discovery, Invalidenstraße 43, Berlin 10115, Germany

¹³<https://www.earthbiogenome.org/sample-collection-processing>.

*Correspondence address. Mara K. N. Lawnczak, Tree of Life, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK. E-mail: mara@sanger.ac.uk

Abstract

The Earth BioGenome Project has the extremely ambitious goal of generating, at scale, high-quality reference genomes across the entire Tree of Life. Currently in its first phase, the project is targeting family-level representatives and is progressing rapidly. Here we outline recommended standards and considerations in sample acquisition and processing for those involved in biodiverse reference genome creation. These standards and recommendations will evolve with advances in related processes. Additionally, we discuss the challenges raised by the ambitions for later phases of the project, highlighting topics related to sample collection and processing that require further development.

Keywords: Reference genomes, biodiversity, long read sequencing, Hi-C, sample collection

Background

The Earth BioGenome Project (EBP) comprises a network of local, continental, and taxon-focused projects [1], all of which are contributing to the vision that reference genomes across the Tree of Life have and will have huge impacts on our understanding of the world around us and on planetary health [2]. The EBP is currently in its first phase, with many projects already contributing and many more projects that may be initiated in the near future. EBP Phase I proposes the sequencing of a representative species for all of the ~9,500 described eukaryotic families [3]. The EBP is a collective effort, and as such, there is no “EBP Phase 1 target species list.” Instead the EBP member projects are encouraged to create their target species lists to reach this goal collaboratively. For the majority of phyla, a genome is available but many of these do not reach chromosomal level (Fig. 1). With respect to the Phase 1 goal, as of October 2024, there are genomes for 3,000 species [4] from over 1,000 families, of which over 2,000 species [5] from over 750 families [6] have chromosome-level assemblies. Thus, significant work remains to complete Phase 1. This document aims to

provide recommendations for any project that is generating reference genomes.

Recommendations for success in sampling for biodiversity genomics

Here, we list general factors to consider as projects set out to select their target species, and then we discuss at more length some additional considerations. While this list is focused on EBP Phase 1, most of the recommendations are advisable even for projects that already have a clear species target list. This list is focused on larger multicellular organisms or smaller species that are easily cultured, but we discuss below ultra low input approaches for protists and other species where physical size is a limitation. General factors to consider when suggesting a species representative towards the EBP Phase 1 goal include:

- **Community value:** Species should be of broad community use and value including species that are of economic or ecological value, are of conservation concern, have specific

Received: October 18, 2024. Revised: January 30, 2025. Accepted: March 13, 2025

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

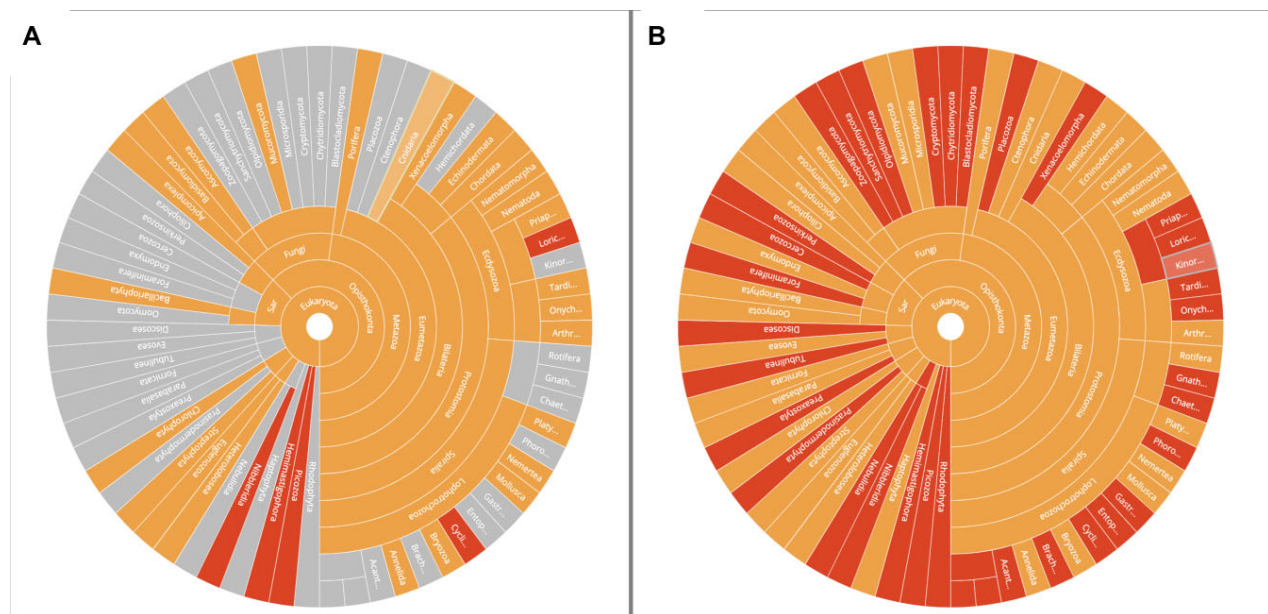


Figure 1: Genome sequencing across eukaryotic phyla from NCBI taxonomy. (A) All assembly levels: red, no available genomes; orange, at least one descendant sequenced by the EBP; grey, at least one descendant has an available genome. GoaT [link](#) for interactive tree. (B) In orange are phyla with at least one EBP-quality assembly, and in red, those with no available EBP-quality genomes. GoaT [link](#) for interactive tree.

scientific interest, or have iconic status. Community value could be assessed through surveys oriented towards target communities (e.g., [7] for the Darwin Tree of Life project).

- **Permissions and availability:** Sampling should be achievable, considering permissions and ethical and legal collection requirements, as well as ease of collection (geography, collecting method).
- **Publicly registered:** The species should be registered with its current name and taxonomy in a publicly available database (we recommend the NCBI Taxonomy Database [8]) and it should be assigned a numeric identifier to assist with tracking taxonomic changes over time. This is covered in more detail below.
- **Taxonomic stability and representation:** The taxon should not be subject to current disagreement and revision. From a taxonomic perspective, it is preferable to sample the type genus of the family, or at least a taxon from the type genus subfamily. The species should be generally considered a “good” biological species (not from a known species complex) and, if possible, sampled from or near the type locality.
- **Voucher availability:** If the organism is large, multiple sub-samples should be taken for tissue and nucleic acid biobanking. For smaller organisms that are likely to be consumed in genome generation, ideally several additional individuals from the same time and place would be collected for morphological and molecular vouchers (e.g., biobanking of DNA, RNA, and/or tissue). Likewise, photographic documentation of living specimens should be undertaken whenever possible.

Additional factors to consider for multicellular organisms include:

- **Genome size, ploidy, sex, and life stage(s):** Where genome sizes and/or ploidy are known (estimates are available via Genomes on a Tree (GoaT); see below), prioritize species with smaller, diploid genomes. This is because data-generation

costs will fall and our ability to assemble high repeat content genomes will improve in future EBP phases. Where sex chromosomes are understood, selecting the heterogametic sex if possible provides a more complete view of the genome. For species with haplodiploid sex determination, the haploid sex should be chosen.

- **Physical sample size:** We recommend a minimal requirement of 10 samples, each weighing more than 10 mg per 1 Gb of genome size, for animals and multicellular/culturable fungi, and culturable protists, and 100 mg per 1 Gb of genome size for plants and multicellular algae. A minimum of three samples is suggested to support the three platforms of long read, chromatin conformation (Hi-C), and transcriptome (RNA-seq) sequencing. Depending on genome size, amplification-based long-read sequencing approaches may still result in a high-quality reference genome for organisms where this is not possible.

The considerations listed above are not intended to be a “must have” list and each project will need to determine what is important for its particular aims. For example, one project might deem completing the genome of a single 9 Gb locust species to be too expensive when those funds could support genomes for 20 species, but another project might deem the locust genome extremely important. In other words, the prioritization of the considerations listed above should be made by each project independently.

In addition to the considerations listed above, there are a number of other areas key to successfully and ethically contributing to EBP goals. Here we set out considerations and recommendations in each of these areas. The guidance focuses on what should be completed for “straightforward” species/specimens—those that are identifiable to species and macroscopic. This is because we are still in the early phase of the project, and when selecting species representatives from families, projects can often prioritize “straightforward” species. However, guidance for the more challenging microscopic or even single-celled organisms is also covered later in the manuscript.

EBP Phase 1 family-representative target species: openly collating community proposals and genomes underway

As the technology needed to generate high-quality reference genomes is improving and becoming increasingly globally accessible, global coordination to avoid duplication of effort (i.e., sequencing the same species) where possible is important. Multiple species for each family should be proposed and collected to provide greater flexibility in achieving Phase 1 goals. As much as possible, this process should be globally visible, with species underway and their associated projects made clear. All projects should use the GoaT system [9, 10] to share their target lists and progress. GoaT has a searchable interface that lists a constantly updated database of progress reported from ongoing reference genome projects and also provides direct and inferred values for genome size, ploidy, and chromosome number. These are extremely useful metrics when selecting target species.

In addition to assisting with global coordination, contributing lists of target species to GoaT allows genomes sequenced through these projects to be tracked and counted both to local and EBP goals. These target lists declare intent, and as such, are likely to undergo revision as projects proceed due to, for example, challenges with acquiring specific target species or extracting adequate DNA and RNA from them. We recommend that larger projects create and maintain a live file listing their target species, identifying taxa selected as EBP family representatives (Fig. 2A), and supply the file URL to the GoaT data curators. The format for the file and a sample template can be found on the website [11]. GoaT archives the submitted file and processes it for display. Larger funded projects are often open to receiving suggestions for target taxa and we encourage genome users to request particular species from the most relevant existing EBP projects (e.g., for an African species, contact AfricaBP).

GoaT can import lists of intent and progress for smaller projects if the project is registered under an umbrella bioproject ID on INSDC and ideally linked to at least one major EBP initiative. Projects should contact one of the major initiatives that best fit their scope to be listed as a contributing lab. The main contact for each project can be found in each project-dedicated page at [12]. Target lists and contact information on GoaT are public and resolution of any overlaps (e.g., Fig. 2B) can be initiated by either project via the contact information provided.

Use cases for exploring available EBP data and also identifying sequencing gaps in taxonomic groups can be found at [13]. Summary information and live progress reports are available as a project-dedicated page for all EBP affiliates at [12]. Each project page contains duplication checkers, where the overlap of target species or progress can be cross-referenced among EBP affiliates. Projects should take advantage of the duplication checkers on GoaT to negotiate and split tasks to maximize the number of species sequenced, and to remove species from target lists by filtering to show those meeting EBP reference genome standards [14].

Ethical collecting

It is paramount that specimens and projects contributing to the EBP are both legally and ethically obtained. Sample collectors should ensure that all local and national permissions for collection are in place, and that there is a record of these permissions that can be referred to if any questions arise as to whether a specimen was legally obtained. This guidance applies to all species, not just those that are of conservation concern. These permissions

will vary widely between countries and jurisdictions within them (e.g., [15, 16] for the European Reference Genome Atlas (ERGA)), and it is beyond the scope of this document to summarize them. Best practice is to ensure that every specimen is collected legally within the applicable frameworks, including national and local rules, rules on endangered species, rules on collecting in protected sites, and rules regarding traditional knowledge. It is also important that permissions are obtained for sequencing, shipping, and publication.

Another complicating factor is that at this phase in the project, some specimens will likely be collected and moved out of their country of origin for sequencing. Customs procedures for importing biological material are often slow and there is a risk of losing precious material due to a loss in maintenance of the cold chain. To avoid this, proper legal documentation for export and import and associated metadata should accompany the specimens when shipped. This can be complicated and factors to consider are described below. When shipping using a courier for the first time, it is advisable to choose a courier (e.g., World Courier or BioCair) that offers dry-ice top-up or a dry-shipping service for a fee. Tissues that do not remain frozen for their entire journey will not yield high-molecular-weight (HMW) DNA or high-quality RNA unless they are in a suitable preservative. Furthermore, unexpected delays due to noncompliance with both export and import regulations are not uncommon when navigating shipments to a new country for the first time.

Regulatory paperwork could include, among others, phytosanitary and veterinary certificates that are usually obtained by country-specific legal entities (ministries/departments of agriculture, national wildlife protection authorities, etc.), as well as other types of import and export permits allowing shipment of biological materials across country borders. Species listed in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES [17]) possess a different set of challenges, as extensive paperwork is required to obtain CITES-specific import and export permits, on top of the usual permits for shipment of biological material. CITES permits must be separately applied for from the local government authority both by the institution in the country of origin, as well as the recipient institution in the country of destination. If both institutions hold CITES Exempt Permits, this greatly simplifies the paperwork exercise. Usually, natural history museums hold these permits, and one might investigate the possibility to import/export the material of interest via those entities. If that option is not available, it is important to file all necessary paperwork with the relevant authority in both countries in advance of the intended shipping date because it can take several months to put in order. When sending CITES-registered species, all the necessary paperwork and customs clearing information will need to be provided to the courier company. BioCair can provide expert services in this regard, whereas many other major shipping companies might need additional guidance. It is also important to point out that some nations only allow export of biological material in special circumstances (e.g., India and China), and Brazil has a very specific set of regulations that should be checked in advance (see [18]).

Beyond the challenges of navigating shipments, the Nagoya Protocol [19] and/or Access and Benefit Sharing (ABS) policies must also be followed when specimens (genetic resources) leave their country of origin. Precise guidance on following the Nagoya Protocol is beyond the scope of this document, especially given that many countries interpret the protocol differently. At a broad level, sample collectors who will be shipping specimens outside their country of origin should contact their local ABS

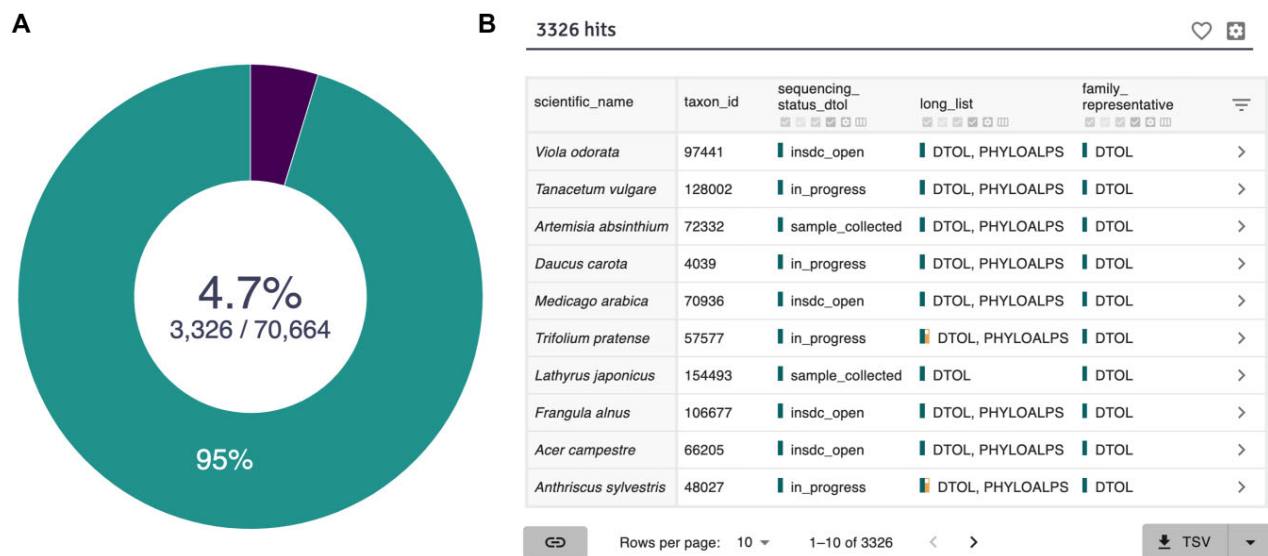


Figure 2: Example of a search for family representatives on GoaT. (A) Report of Darwin Tree of Life (DTOL) species proposed as family representatives out of all >70,000 species on the DTOL wishlist. (B) List of species from the report in (A) with additional columns displaying target (scope) overlaps when present (long_list = DTOL, PHYLOALPS) and current sequencing status within the DTOL genome production pipeline. GoaT search [link](#).

clearing-house [20] to understand the rules, and to obtain a PIC (Prior Informed Consent) document and a MAT (Mutually Agreed Terms on what the benefit is: financial or academic, an acknowledgment on a paper, sharing results, etc). These documents should be written as broadly as possible to support the project's vision. Countries receiving samples should ensure they have further permissions within the MAT to pass the samples on if there is any anticipation that might be required, for example, to support biobanking and morphological vouchering and how data may be shared.

Beyond the rules and regulations, collecting methods must be ethical. The EBP maintains a dedicated Ethical, Legal, and Social Issues (ELSI) committee, which has published detailed recommendations for responsible sample collection, data sharing, and benefit-sharing [5]. These recommendations closely align with the Kunming–Montreal Global Biodiversity Framework (KMGBF). Adopted at the Fifteenth Conference of the Parties (COP15) to the Convention on Biological Diversity (CBD), the KMGBF represents an ambitious international commitment to halt and reverse biodiversity loss by 2030 and achieve “living in harmony with nature” by 2050. Building on earlier biodiversity targets, the KMGBF includes specific goals and targets that emphasize the need for sustainable use of biodiversity, fair and equitable benefit-sharing from genetic resources, and meaningful engagement with Indigenous Peoples and local communities [21]. These priorities resonate with the core principles of the EBP, reinforcing the project's commitment to inclusive and responsible biodiversity research that respects national legislation, traditional knowledge, and ABS requirements.

In practice, the KMGBF emphasizes the importance of informed consent from rights holders, equitable sharing of benefits (financial or otherwise), and capacity-building measures that support local research and conservation efforts. These provisions support the same ethical considerations championed by the EBP's ELSI committee. Consequently, EBP researchers are encouraged to familiarize themselves with the KMGBF's goals and targets to better align genomic research with current international standards for biodiversity conservation and social equity. Further considerations from the EBP's ELSI committee can be found in Sherkow et al. [5].

EBP initiatives are further encouraged to adhere to two complementary sets of data governance principles: CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) and FAIR (Findable, Accessible, Interoperable, and Reusable). The CARE principles emphasize respectful data stewardship that prioritizes the rights and interests of Indigenous Peoples and local communities [22]. The FAIR principles, meanwhile, guide the scientific community to publish genomic and biodiversity data in ways that maximize discoverability and reusability [23]. By following both CARE and FAIR, EBP researchers can ensure that ethical considerations and best practices in data management go hand in hand.

Effort should also be made to build sustainable partnerships with Indigenous Peoples and local communities [15]. Overcollection of any species should be avoided. Projects should consider what the best sampling strategies might be to avoid overcollection, e.g., lineage-focused bioblitzes with a group of taxonomic experts.

Confident species identification

Each contributed specimen should be identified to species level by a taxonomic expert and, ideally, material from the same specimen should be independently DNA barcoded using appropriate markers and the data deposited publicly on the Barcode of Life Data System (BOLD [24]) or in an INSDC database. These DNA barcodes will ensure that species with reference genomes have independently generated DNA barcode data, that the DNA barcodes match the resulting reference genome, and that no sample swaps occur along the way. Although barcoding is advised, it may not always be feasible and work is now underway by the EBP Sample Collection and Processing Subcommittee to develop guidance on how to proceed with EBP goals given the challenges in confident species identification of a large proportion of Earth's biodiversity.

Robust and complete metadata

Robust and complete metadata of all types must accompany the family-level representatives for EBP. Metadata fields and terms should be standardized and we recommend drawing on the extensive efforts to standardize collection metadata already completed

by the Biodiversity Information Standards community (originally the Taxonomic Databases Working Group (TDWG) [25]). Both the Darwin Core (DwC) standard [26] and the Access to Biological Collection Data (ABCD) Schema [27] offer standardized ways of sharing biodiversity occurrences and collections data. While overall similar, DwC is somewhat less complex in structure, and ABCD is able to convey information at a more granular level. The GGBN Data Standard is an extension of ABCD and DwC particularly for molecular samples [28]. The Darwin Tree of Life (DTOL) project has adapted standards for collections for biodiversity genomics [29], and the latest DTOL guidance can be found at [30]. Other projects have already built on the foundation provided by DTOL (e.g., [31]) and it is important to maintain the ontology of terms used in these different projects to prevent drift whereby terms become project-specific. Therefore it is advisable to use these metadata schemas as they are or to ensure that modifications are made in consultation. If project-specific metadata differ substantially from these recommendations, other INSDC sample checklists (e.g., [32] or [33]) are perfectly acceptable, provided that as many metadata fields are completed as is possible and reasonable.

The EBP recommends that every specimen should have a ToLID (Tree of Life identification) assigned. A ToLID is a unique, easy-to-communicate identifier that provides species recognition, numerically and uniquely differentiates between specimens of the same species, and adds some taxonomic context. ToLIDs facilitate internal and external communication about the samples and help the EBP track all sequencing projects. It is also worth remembering that a specimen can contain multiple organisms, so ToLID can disambiguate between target and off-target organisms within a given sample/specimen allowing other specimen products to be published unambiguously. Further, when a genome assembly is submitted, the ToLID of the specimen that was used for long read data generation should be used to name the assembly. ToLIDs are not a replacement for INSDC BioSample records, which hold all of the metadata associated with the sample. Every sample should have both. ToLIDs can be requested at [34] (instructions on the website).

To facilitate tracking of sequencing status at the species level, and retrieving biodiversity genomics data via stable taxon identifiers, the EBP recommends assigning taxon identifiers (TaxIds) in the NCBI taxonomy as early as possible in the creation of a draft target list. TaxIds are stable numerical identifiers assigned by the NCBI taxonomy expert team to all taxa and taxon ranks. A species-level TaxId is necessary not only for registering a ToLID for a specimen but also for declaration of intent on GoAT and for generation of BioSample and assembly submissions. Specific guidelines have been created [35] to identify the need to-, what type of- and how to- request taxids for planned sequencing targets. This starts at checking for the availability of a TaxId for the target species using the ENA or NCBI taxonomy query services. If a TaxId at species or infraspecies rank for the target specimen does not exist, it needs to be requested from the ENA or NCBI taxonomy curators as described in the guide.

Taxonomy is an active process, and revisions to species definitions, such as synonymization of existing names and splitting of existing concepts, and to the higher organisation of species into genera, families, etc., are constant. The NCBI taxonomy aims to represent the best-supported taxonomy for all of life and strives to be up-to-date. To ensure that the NCBI taxonomy remains current, researchers are encouraged to discuss directly with the NCBI taxonomy curators (or the curators at ENA) whenever there is an opportunity to improve the NCBI taxonomy, including taxonomic revisions and reconciliations.

It is important to develop a robust mechanism for updating the names and identifiers associated with a specimen if misidentification or taxonomic reclassification occurs while a reference genome is being produced. This mechanism should also have a policy to determine what gets changed at different points in the process. If a misidentification or a taxonomic name change occurs early in the process of data generation, then fully updating to a new TaxId, scientific name, and ToLID would be appropriate. Once the assembly has been published, we recommend correcting the taxonomic information and ToLID in any cases of misidentifications, even after release. However, we suggest that taxonomic changes such as synonym preferences or merges that do not hamper the identification of the species should only be cause for amendments within six months after assembly release and the ToLID should remain as is to relieve the projects and databases from the burden of frequent minor changes.

Vouchering

Vouchers from each specimen contributing to a reference genome should be preserved when possible and guidance on various biobanking topics is available [36, 37]. Ideally, there will be both morphological vouchers preserving the diagnostic characters of the taxon (or image vouchers if no material sample can be preserved), and molecular vouchers, such as tissue vouchers (discussed below), viably frozen cell lines, and/or extracted RNA and DNA. Vouchers should be prepared in a manner that ensures long-term physical preservation of the specimen and straightforward association with its metadata. Vouchers should be deposited in publicly accessible collection facilities, such as natural history museums or herbaria, located in the country of origin for each sequenced species, or where there is excess material, possibly spread across multiple repositories. It is advisable to arrange in advance of the work commencing where remaining material (including specimen and molecular vouchers) should reside after the work is completed, and this should be reflected in any relevant legal transfer agreements. It can otherwise require secondary legal agreements to transfer materials to a third party. If remaining material is accessible, tracking additional uses of the material (e.g., by using the ToLID) is important to provide awareness to future users of the material that additional data may have already been generated on the specimen.

It is recommended that molecular vouchers be stored in a GGBN member institution (of which there are many, see [38]) for sample accessibility, and linked to the morphological voucher using a unique ID, ideally a globally unique ID (GUID [39]), or at least following the Darwin Core triplet structure ([unique institute acronym][collection acronym, unique in the institute][voucher number, unique in the collection]). Any facility that owns and/or manages collections of nonhuman genomic/molecular samples can apply to become a GGBN member institution, by following the instructions at [40]. When two institutions share material via a material transfer agreement (MTA), it may be worth considering a specific provision that allows the remaining material to be shared with a third party if that is in the interest of the institution providing the material.

Although we strongly advocate for vouchers of specimens in publicly accessible collection facilities, we recognize that these are not uniformly available in all countries, raising potential issues with respect to export/import, access and benefit sharing, and other issues addressed in the KMGBF. Unfortunately, there is no clear solution and best practices with respect to vouchersing may simply not be possible for all contributors. In such cases, we

note that retention of vouchers in small institutional collections would still be preferable to not retaining vouchers at all.

High-quality images should accompany each contributed specimen, and these should be made publicly available. Although specimen digitization can involve complex photography systems, which can yield exquisitely detailed images, any photographic voucher—even one taken with a smartphone—is better than none at all. Photographic documentation of specimens should include all relevant morphological axes. If the specimens were collected while actively engaging in identifiable behaviors (mating, feeding), these should be photographed and noted as well as the time of collection. These images can be deposited in the BioImage Archive [41], GBIF [42], or iDigBio [43]. For genomes accompanied by a DNA barcode, the BOLD database may also be an option for image archiving, as well as GGBN if samples are biobanked in a member repository.

Sample to sequence

The Phase 1 ambition for EBP is to sequence species that are good family-level representatives and the guidelines above indicate the primary considerations for selecting appropriate species. Ultimately, what is selected and sequenced is an individual or a set of individuals and should be recognized as such. Furthermore, additional considerations when selecting the precise specimen(s) that will be sequenced to represent the species and the family are discussed below.

Specimen and tissue selection

The sex of the specimen and the particular life-history stages and tissues that are best for different data types should be considered carefully. Where relevant and possible, it is preferable to sample from the heterogametic sex to provide data for both sex chromosomes. Recommendations on the best life stages and best tissues to target to achieve the highest qualities and quantities of DNA, RNA, and nuclei will vary depending on the taxon. When possible, if a species has been described from a given life stage, that life stage should be selected for sequencing to decrease the likelihood of misidentification. Specific tissues might be prioritized or avoided based on additional species (cobionts) that might be present in or associated with those tissue types (e.g., sequencing gut tissue may yield off-target sequences that may or may not be desirable). The exact tissue types recommended for HMW DNA and RNA for the wide range of target taxa are beyond the scope of this document. They undoubtedly will change as we experience successes and failures in extracting nucleic acids, preparing libraries, sequencing, assembly, and annotation. For annotation using RNA-seq, see the EBP guidelines [44]. Briefly, it is recommended to collect a diversity of tissue types whenever possible, factoring in previous understanding of tissues for the focal taxa that have representative or higher-than-average transcript diversity. We are beginning to gain a better understanding of the life stages and tissues for a wide range of taxa that provide the best quantities and qualities of material, along with the most reliable protocols [45].

The target individual should be collected from the wild rather than from a laboratory colony, zoo, or culture collection (when possible), to capture natural genetic diversity, avoid inbreeding effects, and accurately reflect ecological and evolutionary contexts [46]. The size of the samples needed from a specimen will depend on the taxon and tissue sampled and the genome size, and the precise guidance around required input material is likely to be a rapidly moving target as required quantities decrease and our ability to achieve high-quality extracts across a wide range

of taxa increases. Our current recommendations for animals and multicellular fungi are at least 10 mg and preferably closer to 100 mg of tissue per 1 Gb of genome size for each sample as this tends to be sufficient for long-read data generation without amplification. For plants and multicellular algae we suggest at least 100 mg and preferably 1,000 mg of tissue per 1 Gb of genome for each sample. Multiple samples from the same specimen should be prioritized over single samples from different specimens, and given the current need for samples to be directed down three different processes (Hi-C, HMW DNA extraction, RNA extraction), we recommend at least 10 samples meeting these standards per specimen where possible, and per species where not. For taxa that are known to be difficult (e.g., many marine invertebrates) and might require many extraction attempts, at least 20 samples should be taken given the significant delays and costs incurred by the need for an additional collection trip. If taking a sufficient number of samples from a single specimen is not possible, then additional specimens, ideally from the same locality and collected at the same time, should be preserved to reach similar quantities of tissue. This level of replication gives slack in the system for repeat extractions where sufficient quantities or qualities of data have not been achieved and also provides material for biobanking, enabling future expansion of results with new approaches (e.g., protein, metabolite analysis or new or improved genome/transcriptome sequencing technologies). We strongly recommend getting in touch with the facility that is likely to complete the sequencing prior to specimen collection in the field to check their sample requirements (i.e., tube type, number of specimens required, if they accept samples preserved in ethanol or not, etc.) because these vary depending on the facility and their experience in processing certain taxonomic groups.

Specimen processing

As an organism is processed, it should be photographed alongside a tracking identifier (e.g., a SPECIMEN_ID, which could be the ToLID) and alongside the barcodes of the tubes into which it is processed (Fig. 3). These photographs are in addition to any that might be taken to document the living specimen and are useful for resolving sample-tracking problems that can arise. We strongly encourage the use of barcoded tubes and scanning of these tubes rather than handwriting identifiers on tubes and manual entry of identifiers into tracking systems as this is prone to error. For samples in the dozens or hundreds, this can be done with a simple single-tube scanner or even with a phone and an application such as EpiCollect [47]. For larger projects processing many hundreds or thousands of specimens, rack scanners can be used to scan whole racks of barcoded tubes before sample processing.

Living specimens should be processed into tubes on dry ice and, from that point forward, held at -70°C or below (e.g., in liquid nitrogen). Specimens that have died before processing tend to have damaged and degraded DNA and RNA, or can become highly contaminated with bacteria. Precautions should be taken if such material is to be used for reference genome generation: even if high-quality DNA can be obtained from this material, it may be largely bacterial. Specimens should be taken to a site where dry ice or liquid nitrogen are available, humanely killed, and rapidly processed into small, lentil-sized pieces (Fig. 4) while freezing, for example using a Petri dish on dry-ice (as pictured in Fig. 3) and a scalpel.

Small pieces of specimens sitting on dry-ice usually yield high-quality DNA as long as the freezing process is rapid. These small tissue pieces can then be placed into prechilled barcoded tubes—e.g., in Fig. 3, the fly could be cut into head, thorax, and abdomen, and each piece put in a separate tube. Currently, for animals and



Figure 3: An example of the documentation that should occur as a sample is being processed. Here the SPECIMEN_ID (the NHM barcode under the fly) is photographed alongside the specimen and the barcoded tubes to which different samples of that specimen are destined. The metadata tracking sheet would thus 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). Photograph by M.K.N.L.

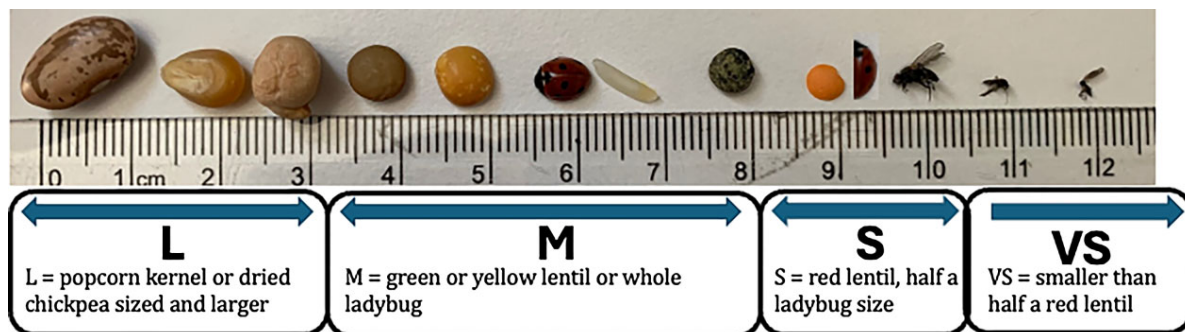


Figure 4: A scale showing the sizes of tissues that should typically be aimed for in a single tube. In some cases (e.g., the flies in the Very Small (VS) category), there is no option to provide more tissue, so multiple separate specimens will be needed. The exact amount of tissue to supply depends on the taxon and the genome size, so it is best to discuss this in detail with the sequencing facility before beginning species collections. Facilities will not want to receive pieces of tissue much larger than the Large (L) size shown here as the tissue will be frozen solid and it is extremely difficult to break suitable sized pieces off without compromising the DNA or RNA integrity (e.g., by thawing). Plants will generally need substantially more tissue than is shown here. As advised in the main text, for each species, we recommend aiming for 10 tubes, each containing a piece of tissue in the S-L range (typically 20–50 mg). Photograph by M.K.N.L.

multicellular fungi, we recommend one piece of >10 mg tissue per tube to support different workstreams without compromising the temperature of the remainder of the material through freeze-thaw cycles. Plants and multicellular algae should also be processed into small pieces to support rapid freezing, but larger volumes of tissue might be placed in tubes because up to 10 times more tissue may be required to achieve adequate quantities of DNA for these groups. For many taxa where the majority of the specimen will be consumed in the process of data generation (e.g., most insects), it may be advisable to grind the whole organism to a fine powder in liquid nitrogen prior to further laboratory work to

avoid different data types being generated from different tissues (e.g., Hi-C data coming from the head, and long-read data coming from the abdomen, each tissue with distinct associated cobiont taxa).

Sequencing the very small

For organisms where achieving milligrams of tissue is not possible, including meiofauna (i.e., animals <1 mm in size) and unicellular eukaryotes, ultra-low-input (ULI) protocols can be adopted to generate long-read data from single microscopic organisms.

Protocols leveraging long-range PCR [48, 49] or multiple displacement amplification [50, 51] for microscopic organisms are available with caveats that DNA integrity and genome size are important considerations and these protocols do not perform well on large, highly repetitive genomes [49, 51]. We advise that single specimens be used to generate long-read data, even for the smallest organisms. A second individual may be needed to produce a transcriptome (but see [49]). Commercially available cDNA library preparation kits make it possible to obtain high-quality transcriptome data from a single minute animal or even single cells (e.g., [52]). Coassembly of individual single amplified genomes (SAGs) has been moderately successful for some protist species, though not yet to the standards desired by the EBP [53, 54]. However, pooling several individuals (ideally related individuals of the same sex/mating type) may be necessary for Hi-C. The ULI approaches might not be sufficient for achieving EBP assembly-quality standards, but at this point of technology development, it is the best that can be done for many microscopic taxa. When multiple individuals must be pooled for small organisms, species and collection sites should be chosen to ideally yield several specimens; these lots need to be collected at the same time and location, and carefully determined taxonomically to ascertain whether they belong to the same species. DNA barcodes on individuals should be attempted where possible. If this is not possible for the specimens that will be used for reference genome data generation, then it is advisable to collect additional specimens from the same time and location and generate barcode data for these as a proxy identification approach. In the case of unicellular eukaryotes that are difficult-to-determine species, refraining from pooling cells will ensure that only a single species is sequenced. Ideally, images of the cell/organism sequenced should be taken for vouchers purposes.

Cold chain challenges

Situations in which preserving samples from living organisms without access to dry-ice or liquid nitrogen are likely to become common as the EBP progresses. We are still learning which preservatives offer the best chance at successful long-read and long-range sequencing, and we recommend sharing successful protocols openly and early using the EBP protocols.io workspace (see below). As of now, we suggest that if there is no possibility of rapid processing and preservation of a specimen from living to -70°C or below, then samples should be processed into small lentil-sized pieces in an excess (high preservative volume to tissue volume ratio) of 100% ethanol for HMW DNA and Hi-C, and in RNAlater for RNA, and that these are then stored at the coldest temperature possible. Lower percentages of ethanol are not advised because they seem to result in more degraded DNA, but this is still a grey area and different taxa may have different requirements. High-quality genomes have been generated from specimens stored in lower percentages of ethanol for long periods, but here we offer general guidance on best-known practices rather than what might work.

For small insects that may not require further processing and so be preserved whole, compromising the cuticle to permit ethanol penetration is critical for preserving HMW DNA [55] and has resulted in high-quality reference genomes even when insects were shipped for over 1 week at room temperature [56]. The ratio of tissue to preservative should also be considered because water in tissues may dilute the preservative. It is advised to change ethanol twice within 24 hours following preservation for tissues with high water content. We recommend a $>20:1$ preservative-to-tissue ratio. As soon as access to a -70°C freezer is available, the

samples should be frozen and records should be kept on how long samples were held at room temperature. For dissected vertebrate tissue, if liquid nitrogen is not readily available, storage in a preservation liquid for up to 1 week at 4°C before flash-freezing in the laboratory has resulted in high quality HMW-DNA and Hi-C [57].

Experience has shown that preservation in media designed for nucleic acid protection (such as DNAGuard or RNAlater) is not optimal for subsequent Hi-C sequencing because these media induce excessive trans interactions, likely because of disruption of nuclei [57]. For preservation of RNA, ethanol is not suitable. RNAlater-ICE does not produce a precipitate upon cooling and may be better than standard RNAlater for storage of RNA-preserved material at below-freezing temperatures. Intensive further testing of preservatives and their ability to protect HMW DNA, RNA, and material suitable for Hi-C across the Tree of Life is likely to be an active area of development over the coming years, and sharing successful preservation protocols will be valuable for the EBP's ambitions (see below). While this summarizes our current understanding of preservation when cold-chain access is limited, best-practice guidance is to snap-freeze tissue immediately from ethically killed specimens and to maintain that cold chain until the point of extraction. This is superior in terms of HMW DNA yield and fragment size and is possible even in remote field settings with the use of charged dry-ice shippers that can maintain ultracold temperatures for a couple weeks.

An EBP protocols.io community

Collection, extraction, and library generation protocols are as important to retain and share, according to the FAIR principles [23], as the sample-associated metadata. We recommend sharing collection, preservation, and extraction protocols in open-source repositories where a unique Digital Object Identifier (DOI) is assigned to every document. We advise including these DOIs in publications arising from the genomes produced under the EBP umbrella. As our knowledge about the biochemical and genetic makeup of previously understudied taxa increases, so will the knowledge base behind the appropriate handling of samples of these taxa and their genetic material. It will increase the reproducibility of research and provide an invaluable resource for the global community.

Biologists worldwide are already actively developing and releasing protocols for ethically collecting specimens, collecting comprehensive metadata, vouchering samples, preserving and processing samples, extracting RNA and HMW DNA, sequencing RNA and DNA, performing Hi-C, etc. These protocols are abundant but still primarily focused on a small number of taxonomic groups and are also typically hidden in research manuscripts rather than released as step-by-step protocols. As we learn what modifications or entirely different approaches work best for different taxonomic groups, we encourage open sharing of this information as early as possible in protocol form. To assist with this, we have created an EBP protocols.io workspace [58]. We recommend that the biodiversity genomics community publish their genome-relevant protocols at protocols.io (this is free and results in a citable DOI) and then link their protocols to the EBP workspace. To do this, users must join the EBP workspace and, once their protocol is published, simply link it as shown in Fig. 5.

In addition to contributing new protocols to the EBP workspace, we also recommend extensive commenting and forking of existing protocols. Commenting on existing protocols can help rapidly share information on successes and failures, as well as minor tips and recommendations that improve the chances of success.

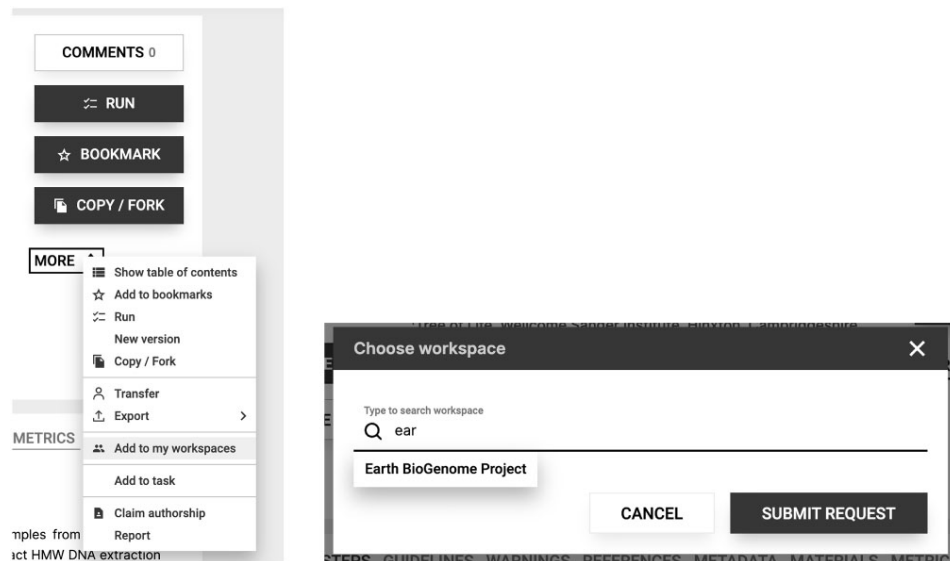


Figure 5: How to link protocols to the EBP workspace. Step 1. Join the EBP workspace at protocols.io/workspaces/earth-biogenome-project (access may require a free login). Step 2. Click on the “MORE” menu as shown in the left panel above and select “Add to my workspaces.” Step 3. Select “Earth BioGenome Project” as the workspace and add the protocol.

Commenting and forking are easily achieved on any protocol through the “COMMENTS” and “COPY/FORK” clickable buttons that are present on every protocol. We suggest using comments to indicate if the protocol worked for specific species and to include higher taxonomic information, e.g., “This protocol results in high-quality HMW DNA for *Hapalochlaena lunulata*, an octopus. The DNA was successfully sequenced using PacBio HiFi.” Forking can be used where an existing protocol has been modified more extensively to improve results, e.g., for a particular tissue or taxon.

In the process of trying to go from specimen to high-quality reference genome, we learn through unexpected successes and failures the tips and tricks that are not protocol-worthy but still should be shared. If we had a way to share these anecdotes, we might save people treading down the same dead-end roads or give them a tip that opens up huge possibilities. Therefore, we have created a community-owned Google sheet called “EBP community collection of anecdotes” at [59]. This is set up to simply collect experiences of anyone working in the general area of reference genomes for biodiversity. We ask that the EBP community continue to populate this resource with their tips and tricks and we have provided some guidelines at the top of the document.

Fostering inclusive collaboration for enhanced impact

For the EBP to succeed, we must all work towards a more inclusive and equitable global biodiversity genomics community. Strengthening collaboration and engagement across the biodiversity genomics stakeholder community will accelerate progress and maximize scientific impact, benefit sharing, equity, diversity, inclusion, and just practices. The biodiversity genomics stakeholder community includes individuals from all relevant subdisciplines of biology: those traditionally engaged, such as geneticists, genomicists, and bioinformaticians; experts in taxonomy, natural history and ecological and evolutionary theory, including ecologists, evolutionary biologists, taxonomists, and systematists; and those often overlooked, such as individuals working in natural history col-

lections who are involved in vouchers and specimen curation, staff at sequencing facilities who assist with project design and data generation, and those providing support in other aspects of project design, specimen acquisition, data generation, and curation.

The success of EBP is intrinsically linked to the collaborative efforts of scientists across a wide range of disciplines. Researchers working in natural history collections or at biobanks, taxonomists, systematists, genomicists, bioinformaticians, and support staff at sequencing facilities all play crucial roles. These professionals collect, identify, preserve, and maintain specimens, infer phylogenetic relationships, and study the evolution of traits and geographic distributions of organisms. Their work enables the genomic research community to access well-curated and accurately identified specimens, ensuring the correct identification and systematic placement of species-representative reference genomes. It is essential for all involved to acknowledge and strengthen these relationships, advocating for the vital work done by museum scientists, systematists, taxonomists, and technical support staff. Additionally, relevant taxonomic and systematic work, such as species descriptions and evolutionary context studies, and IDs of collection material examined, should be cited in genomics publications. Ensuring that individuals who collected and identified specimens are included as authors on resulting papers fosters a collaborative approach and gives credit where it is due [60].

Enhanced collaborations, e.g., among genomicists, systematists, taxonomists, museum scientists, bioinformaticians, and sequencing facility staff, improve the scientific rigor of genomic research and recognize the critical contributions of experts in taxonomy, evolution, and organismal biology [61, 62]. Systematics and taxonomy provide an indispensable framework for genomics research. Experts in these fields often lead advances in imaging, informatics, ecology, evolution, genetics, and genomics (e.g., [63–65]), and often yield important new discoveries at the interface between traditional subdisciplines of biology (e.g., [66, 67]). Presently, many groups of organisms are studied by only a few

living taxonomic experts, making reliable identification challenging, at best. Alarming, limited taxonomic training in undergraduate curricula, fewer living expert taxonomists mentoring students and postdocs, and limited funding for taxonomic research stand to exacerbate this issue [62]. Collaborative efforts can provide valuable training for the next generation of integrative biodiversity scientists.

Close collaborations among genomicists, museum scientists, and taxonomists can also prevent duplicative sampling efforts, minimize the ecological impact of specimen collection, conserve resources, and ensure efficient sampling. With genomic sequencing now being more cost-effective, it is crucial to maximize the value of each collected specimen [68]. The genomics community, with its expertise in bioinformatics and data management, can significantly contribute to biodiversity informatics and digitization initiatives [69]. Researchers collaborating with museums can advocate for best practices in specimen preservation that will facilitate future genomic work and ensure the deposition of voucher specimens in public collections [70]. Of course, many scientists in the field of biodiversity genomics also engage in systematics and taxonomy and/or work in museum settings, demonstrating the interdisciplinary nature of modern genomics research (e.g., [71]).

All members of the EBP stakeholder community, including genomicists, systematists, taxonomists, museum scientists, bioinformaticians, sequencing facility staff, and others, play important interdependent roles in this endeavor. Strengthening collaboration among these communities will enhance the quality and impact of research, ensuring it is ethical, sustainable, and inclusive. Through such partnerships we can more fully and effectively explore and preserve the diversity of life on Earth for generations to come.

Looking to the future

Current best-practice assembly guidelines are to generate a combination of data types including long-read (PacBio HiFi and/or ultra-long ONT), long-range (Hi-C), and RNAseq (Illumina short read, PacBio Kinnex, or ONT cDNA-PCR) data from the same specimen wherever possible, aiming for the heterogametic sex when this is relevant. Typically, separate samples are used for these different applications, but we should be developing protocols that support minimal extraction of material sufficient for any of these types of data generation (e.g., nuclei coextracted with RNA). Furthermore, in all current extraction efforts, we discard material that we might one day look back on and regret, such as proteins and metabolites. While data generation from these materials is currently out of scope, this is unlikely to be true in years to come. Retaining relevant material to add data layers to the high-quality reference genomes would be prudent. Thus, for specimens where samples are available in excess, considerations should be given to preserving replicate samples and appropriate storage to future-proof these samples as much as possible. For specimens where all material is used in data generation, perhaps typically discarded supernatants should be retained for future investigations.

As sequencing proceeds into phases comprising more fine-scale taxonomic coverage, we foresee that getting specimens identified *in vivo* before killing and freezing them will be challenging for many species-rich but poorly known taxa, including most invertebrate groups. Reliable identifications often depend on careful examination under the microscope by experts, of which there are typically only a few for entire taxonomic families; this would require holding the specimens out of the cold chain for a length

of time that could compromise the DNA and RNA. This highlights the importance of appropriately preserved morphological voucher specimens and imaging. Further, advances in DNA barcoding reference libraries can be of enormous importance, so that specimens can be collected first and identified later. One possibility would be to have a fourth tissue sampling used initially only for DNA barcoding to attempt identification by comparing the sequence to available public databases. Such identifications would then inform the decisions about whether a sample should be further processed for genomic sequencing.

We also encourage activities that simplify and streamline standard operating procedures and protocols to make it easier to “containerize” extraction, sequencing, and assembling activities. Containerization means a world in which a simple portable lab could house everything needed to go from sample to sequence and would build capacity in the Global South, in the nations that often harbor the greatest biodiversity. This relieves pressure on the Nagoya requirements and the budget spent on expensive shipping costs to maintain the cold-chain, but more importantly, is better for global science.

Numerous unexplored opportunities also lie within the realm of artificial intelligence (AI) and automation, waiting to be unlocked and harnessed. AI-driven robotic systems could increase efficiency by automating sample collection and optimizing workflows, significantly reducing the time and resources required (e.g., [72]). This integration could enhance precision and accuracy through AI-based quality control and precise data annotation, ensuring only high-quality samples contribute to genome sequencing. Costs could thus be reduced through labor savings and resource optimization, while the speed of genome sequencing could be accelerated by automated processing and parallelization of samples. Furthermore, AI could be used to facilitate intelligent sampling strategies, promoting diversity in collected samples, and enabling real-time monitoring of environmental conditions. Integrating biodiversity data into centralized databases and advanced analytics enhances data accessibility and analysis. Overall, the synergy of AI and automation may improve the quality and speed of reference genome generation. Reference genomes from across the Tree of Life enabled by a global community and multiple technologies will contribute to a deeper understanding of biodiversity for applications in conservation, ecology, and evolutionary biology.

Abbreviations

ABS: access and benefit sharing; BOLD: Barcode of Life Database; EBP: Earth BioGenome Project; Gb: gigabase; GoAT: Genomes on a Tree; HMW: High molecular weight; INSDC: International Nucleotide Sequence Database Collection; MTA: Material Transfer Agreement; ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences; TaxId: taxonomic identifier; ToLID: Tree of Life identifier.

Acknowledgements

The authors thank Harris Lewin and Federica di Palma for comments.

Authors' Contributions

The EBP Sample Collection and Processing Subcommittee members conceptualized the manuscript. M.K.N.L. drafted the original manuscript. All authors contributed to review and editing. EBP

Sample Collection and Processing Subcommittee as of 2025: Mara K. N. Lawniczak (Chair); and Jonas Astrin, Anna Childers, Kevin Kocot, Duane McKenna, Olga Pettersson, Verena Ras, Bernardo Santos, and Jeremy Wideman (all members, listed alphabetically).

Funding

M.K.N.L., M.B., P.D., K. H., and C.G.S.C. are funded by the Wellcome Trust quinquennial award to the Wellcome Sanger Institute (grant number 220540/Z/20/A). O.V.P. is funded by the SciLifeLab (Science for Life Laboratory, Sweden), and the Swedish Research Council (RFI/VR). K.M.K. is funded by National Science Foundation grants 1846174, 2138994, 2321308, and 2001303. V.R. is funded by the NIH Common Fund Award/NHGRI grant number U24HG006941 and National Institutes of Health (OD), National Institution of Biomedical Imaging and Bioengineering, and NIH award number 1 U2C EB 032224-01. D.D.M. is funded by National Science Foundation grants 2110053 and 1937815. J.G.W. is funded by the National Science Foundation grant DBI-2119963, BII: Mechanisms of Cellular Evolution. A.K.C. is supported by the US Department of Agriculture, Agricultural Research Service (USDA-ARS), Bee Research Laboratory in-house appropriated research project 8042-21000-291-000-D.

Data Availability

Not applicable

Competing Interests

The authors declare that they have no competing interests.

References

- Affiliated Project Networks. Earth BioGenome Project. <https://www.earthbiogenome.org/affiliated-project-networks>. Accessed 10 March 2025.
- Blaxter M, Archibald JM, Childers AK, et al. Why sequence all eukaryotes? *Proc Natl Acad Sci USA*. 2022;119(4):e2115636118. <https://doi.org/10.1073/pnas.2115636118>.
- Catalogue of Life. COL. <https://www.catalogueoflife.org/>. Accessed 10 March 2025.
- Genomes on a Tree (GoaT). EBP Umbrella - EBP taxa with assemblies out of all Eukaryotic taxa in INSDC. https://goat.genomehubs.org/2024.09.14/report?report=arc&x=assembly_span%20AND%20bioproject%3DPRJNA533106&rank=phylum%2Cclass%2Corder%2Cfamily%2Cgenus%2Cspecies&includeEstimates=true&excludeAncestral%5B0%5D=bioproject%2CexcludeMissing%5B0%5D=bioproject&caption=%2A%2AEBP%20Umbrella%2A%2A%20-%20EBP%20taxa%20with%20assemblies%20out%20of%20all%20Eukaryotic%20taxa%20in%20INSDC&taxonomy=ncbi&result=taxon. Accessed 10 March 2025.
- Genomes on a Tree (GoaT). complete genome, chromosome for bioproject=prjna533106. [https://goat.genomehubs.org/2024.09.14/search?query=tax_rank%28species%29%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome&result=taxon&includeEstimates=true&summaryValues=count&taxonomy=ncbi&size=100&offset=0&fields=ebp_standard_criteria%2Cebp_standard_date%2Cassembly_level%2Cbioproject&names=&ranks=#tax_rank\(family\)%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome](https://goat.genomehubs.org/2024.09.14/search?query=tax_rank%28species%29%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome&result=taxon&includeEstimates=true&summaryValues=count&taxonomy=ncbi&size=100&offset=0&fields=ebp_standard_criteria%2Cebp_standard_date%2Cassembly_level%2Cbioproject&names=&ranks=#tax_rank(family)%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome). Accessed 10 March 2025.
- Genomes on a Tree (GoaT). complete genome, chromosome for bioproject=prjna533106. [https://goat.genomehubs.org/2024.09.14/search?query=tax_rank%28family%29%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome&result=taxon&includeEstimates=true&summaryValues=count&taxonomy=ncbi&size=100&offset=0&fields=ebp_standard_criteria%2Cebp_standard_date%2Cassembly_level%2Cbioproject&names=&ranks=#tax_rank\(family\)%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome](https://goat.genomehubs.org/2024.09.14/search?query=tax_rank%28family%29%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome&result=taxon&includeEstimates=true&summaryValues=count&taxonomy=ncbi&size=100&offset=0&fields=ebp_standard_criteria%2Cebp_standard_date%2Cassembly_level%2Cbioproject&names=&ranks=#tax_rank(family)%20AND%20bioproject%3Dprjna533106%20AND%20assembly_level%3Dcomplete%20genome%2Cchromosome). Accessed 10 March 2025.
- Proposing priority species for genome sequencing in, and working with the Darwin Tree of Life project. Google Docs. <https://docs.google.com/forms/d/e/1FAIpQLSfOA4GRA3mkLLnN3fng9ZM8RdMIHusqPqfMDPePy8a0sE72Ug/viewform>. Accessed 10 March 2025.
- NCBI. Home—Taxonomy. <https://www.ncbi.nlm.nih.gov/taxonomy>. Accessed 10 March 2025.
- Challis R, Kumar S, Sotero-Caio C, et al. Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Res*. 2023;8:24. <https://doi.org/10.12688/wellcomeopenres.18658.1>.
- Genomes on a Tree (GoaT). Website. <https://goat.genomehubs.org/>. Accessed 10 March 2025.
- Genomes on a Tree (GoaT). Submissions. <https://goat.genomehubs.org/submissions>. Accessed 10 March 2025.
- Genomes on a Tree (GoaT). Projects. <https://goat.genomehubs.org/projects>. Accessed 10 March 2025.
- Genomes on a Tree (GoaT). Use Cases. https://goat.genomehubs.org/help/use_cases. Accessed 10 March 2025.
- EBP. Report on Assembly Standards. Earth BioGenome Project. <https://www.earthbiogenome.org/report-on-assembly-standards>. Accessed 10 March 2025.
- Mc Cartney AM, Head MA, Tsosie KS, et al. Indigenous peoples and local communities as partners in the sequencing of global eukaryotic biodiversity. *NPJ Biodivers*. 2023;2:8. <https://doi.org/10.1038/s44185-023-00013-7>.
- Böhne A, Fernández R, Leonard JA, et al. Contextualising samples: supporting reference genomes of European biodiversity through sample and associated metadata collection. *NPJ Biodivers*. 2024;3:26. <https://doi.org/10.1038/s44185-024-00053-7>.
- CITES. <https://cites.org/eng>. Accessed 10 March 2025.
- Frequent Asked Questions proposed by the Academic Sector Chamber from the Genetic Heritage Management Council. https://www.gov.br/mma/pt-br/assuntos/bioeconomia/patrimonio-genetico/conselho-de-gestao-do-patrimonio-genetico-cg-en-1/camaras-tematicas-e-setoriais/camaras-setoriais-1/camara-setorial-da-academia/csa-faq-collections_08_01_2021.pdf. Accessed 10 March 2025.
- The Nagoya Protocol, 2015. <https://www.cbd.int/abs/default.shtml>.
- The ABS Clearing House, 2025. <https://www.cbd.int/abs/theabsch.shtml>.
- COP15: Final text of Kunming–Montreal Global Biodiversity Framework. Convention on Biological Diversity. <https://www.cbd.int/article/cop15-final-text-kunming-montreal-gbf-221222>. Accessed 28 January 2025.
- Carroll SR, Garba I, Figueroa-Rodríguez OL, et al. The CARE principles for indigenous data governance. *Data Sci J*. 2020;19:43. <https://doi.org/10.5334/dsj-2020-043>.
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and steward-

- ship. Sci Data. 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>
24. BOLD—The Barcode of Life Data Systems. <http://www.boldsystems.org>. Accessed 10 March 2025.
 25. Biodiversity Information Standards (TDWG). <https://www.tdwg.org/>. Accessed 10 March 2025.
 26. Darwin Core. <https://www.tdwg.org/standards/dwc/>. Accessed 10 March 2025.
 27. Access to Biological Collection Data (ABCD) Schema. tdwg.org/standards/abcd/. Accessed 10 March 2025.
 28. GGBN. ggbn.org/ggbn_portal/site/wf?p=GGBN_Data_Standard. Accessed 10 March 2025.
 29. Lawniczak MKN, Davey RP, Rajan J, et al. Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project. Wellcome Open Res. 2022;7:187. <https://doi.org/10.12688/wellcomeopenres.17605.1>.
 30. DTOL metadata guidance. github.com/darwintreeoflife/metadata.
 31. ERGA-sample-manifest. github.com/ERGA-consortium/ERGA-sample-manifest.
 32. EMBL-EBI ENA Browser <https://www.ebi.ac.uk/ena/browser/checklists>. Accessed 10 March 2025.
 33. NCBI Biosample Packages—BioSample. ncbi.nlm.nih.gov/biosample/docs/packages/. Accessed 10 March 2025.
 34. ToLID—Tree of Life Identifiers. <https://id.tol.sanger.ac.uk/>. Accessed 10 March 2025.
 35. Blaxter M, Pauperio J, Schoch C, et al. Taxonomy Identifiers (TaxId) for Biodiversity Genomics: a guide to getting TaxId for submission of data to public databases. Wellcome Open Res. 2024;9:591. <https://doi.org/10.12688/wellcomeopenres.22949.1>.
 36. Corrales C, Astrin JJ. Biodiversity Biobanking—A Handbook on Protocols and Practices, vol. 1. Sofia: Pensoft, 2023. <https://doi.org/10.3897/ab.e101876>.
 37. ISBER. isber.org/page/BPR. Accessed 21 May 2024.
 38. GGBN. https://www.ggbn.org/ggbn_portal/members. Accessed 10 March 2025.
 39. Leach PJ, Salz R, Mealling MH. RFC 4122: a universally unique Identifier (UUID) URN namespace. IETF Datatracker. <https://datatracker.ietf.org/doc/html/rfc4122>. Accessed 10 March 2025.
 40. Membership—GGBN Wiki. <https://wiki.ggbn.org/ggbn/Membership>. Accessed 20 March 2025.
 41. EMBL-EBI BioImage Archive <https://www.ebi.ac.uk/bioimage-archive/>. Accessed 10 March 2025.
 42. GBIF. <https://www.gbif.org/>. Accessed 10 March 2025.
 43. Integrated Digitized Biocollections. <https://www.idigbio.org/>. Accessed 10 March 2025.
 44. EBP Report on Annotation Standards. Earth BioGenome Project. <https://www.earthbiogenome.org/report-on-annotation-standards>. Accessed 10 March 2025.
 45. Howard C, Denton A, Jackson B, et al. On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species. bioRxiv. 2025. <https://doi.org/10.1101/2025.04.11.648334>.
 46. Frankham R. Genetic adaptation to captivity in species conservation programs. Mol Ecol. 2008;17:325–33. <https://doi.org/10.1111/j.1365-294X.2007.03399.x>.
 47. Free and easy-to-use mobile data-gathering platform. <https://fiv.e.epicollect.net/>. Accessed 10 March 2025.
 48. Schneider C, Woehle C, Greve C, et al. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny meta-zoans (Collembola). Gigascience. 2021. <https://doi.org/10.1093/gigascience/giab035>.
 49. Laumer C. Picogram input multimodal sequencing (PiMmS). protocols.io. 2023. <https://dx.doi.org/10.17504/protocols.io.rm7vzywy5lx1/v1>.
 50. Lee Y-C, Ke H-M, Liu Y-C, et al. Single-worm long-read sequencing reveals genome diversity in free-living nematodes. Nucleic Acids Res. 2023;51:8035–47. <https://doi.org/10.1093/nar/gkad647>.
 51. Roberts NG, Gilmore MJ, Struck TH, et al. Multiple displacement amplification facilitates SMRT sequencing of microscopic animals and the genome of the gastrotrich *Lepidodermella squamata* (Dujardin, 1841). Genome Biol Evol. 2024;16(12):evae254. <https://doi.org/10.1093/gbe/evae254>.
 52. Varney RM, Funch P, Kocot KM, et al. A new species of Echinoderes (Cyclorhagida: Echinoderidae) from the San Juan Islands, Washington State, USA, and insights into the kinorhynch transcriptome. Zool Anz. 2019;282:52–63. <https://doi.org/10.1016/j.jcz.2019.06.003>.
 53. Mangot J-F, Logares R, Sánchez P, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci Rep. 2017;7:41498. <https://doi.org/10.1038/srep41498>.
 54. Sørensen MES, Zlatogursky VV, Onuț-Brännström I, et al. A novel kleptoplastidic symbiosis revealed in the marine centrophilid *Meringosphaera* with evidence of genetic integration. Curr Biol. 2023;33:3571–84.e6. <https://doi.org/10.1016/j.cub.2023.07.017>.
 55. Teltscher F, Lawniczak M. Squishing insects for preservation of HMW DNA in the field. protocols.io. 2023. <https://dx.doi.org/10.17504/protocols.io.4r3l2224jl1y/v1>.
 56. Nsango SN, Ayala D, Agbor J-P, et al. A chromosomal reference genome sequence for the malaria mosquito, *Anopheles nili*, Theobald 1904. Wellcome Open Res. 2024;9:599. <https://doi.org/10.12688/wellcomeopenres.23198.1>.
 57. Dahn HA, Mountcastle J, Balacco J, et al. Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. Gigascience. 2022;11:giac068. <https://doi.org/10.1093/gigascience/giac068>.
 58. Earth BioGenome Project—research workspace on. protocols.io. protocols.io/workspaces/earth-biogenome-project. Accessed 10 March 2025.
 59. EBP community collection of anecdotes. Google Docs. <https://tinyurl.com/EBP-Anecdotes>. Accessed 10 March 2025.
 60. Meier R. Citation of taxonomic publications: the why, when, what and what not: species citations. Syst Entomol. 2016;42(6):301–4. <https://doi.org/10.1111/syen.12215>.
 61. Drew LW. Are we losing the science of taxonomy? As need grows, numbers and training are failing to keep up. Bioscience. 2011;61:942–46. <https://doi.org/10.1525/bio.2011.61.12.4>.
 62. Löbl I, Klausnitzer B, Hartmann M, et al. The silent extinction of species and taxonomists—an appeal to science policymakers and legislators. Diversity. 2023;15:1053. <https://doi.org/10.3390/d15101053>.
 63. Faulwetter S, Vasileiadou A, Kouratoras M, et al. Thanos Dailianis, Arvanitidis C. Micro-computed tomography: introducing new dimensions to taxonomy. Zookeys. 2013;263:1–45. <https://doi.org/10.3897/zookeys.263.4261>.
 64. Meyer C, Duffy E, Collins A, et al. The US Ocean biocode. Mar Technol Soc J. 2021;55:140–41. <https://doi.org/10.4031/MTSJ.55.3.33>.

65. Pante E, Schoelinck C, Puillandre N. From integrative taxonomy to species description: one step beyond. *Syst Biol.* 2015;64:152–60. <https://doi.org/10.1093/sysbio/syu083>.
66. McKenna DD, Shin S, Ahrens D, et al. The evolution and genomic basis of beetle diversity. *Proc Natl Acad Sci USA.* 2019;116:24729–737. <https://doi.org/10.1073/pnas.1909655116>.
67. Stiller J, Feng S, Chowdhury A-A, et al. Complexity of avian evolution revealed by family-level genomes. *Nature.* 2024;629:851–60. <https://doi.org/10.1038/s41586-024-07323-1>.
68. Funk VA. Collections-based science in the 21st century. *J Syst Evol.* 2018;56:175–93. <https://doi.org/10.1111/jse.12315>.
69. Guralnick R, Hill A. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics.* 2009;25:421–28. <https://doi.org/10.1093/bioinformatics/btn659>.
70. Colella JP, Stephens RB, Campbell ML, et al. The open-specimen movement. *Bioscience.* 2021;71:405–14. <https://doi.org/10.1093/biosci/biaa146>.
71. Droege G, Barker K, Astrin JJ, et al. The Global Genome Biodiversity Network (GGBN) Data Portal. *Nucleic Acids Res.* 2014;42:D607–12. <https://doi.org/10.1093/nar/gkt928>.
72. Wühl L, Pylatiuk C, Giersch M, et al. DiversityScanner: robotic handling of small invertebrates with machine learning methods. *Mol Ecol Resour.* 2022;22:1626–38. <https://doi.org/10.1111/1755-0998.13567>.