OPEN LETTER

Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project

Mara K.N. Lawniczak[1], Robert P. Davey[2], Jeena Rajan[3], Lyndall L. Pereira-da-Conceicoa[1], Estelle Kilias[4], Peter M. Hollingsworth[5], Ian Barnes[6], Heather Allen[6], Mark Blaxter[1], Josephine Burgin[3], Gavin R. Broad[6], Liam M. Crowley[4], Ester Gaya[7], Nancy Holroyd[1], Owen T. Lewis[4], Seanna McTaggart[2], Nova Mieszkowska[8], Alice Minotto[2], Felix Shaw[2], Thomas A. Richards[4], Laura A.S. Sivess[6], Darwin Tree of Life Consortium

[1]Tree of Life, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
[2]Earlham Institute, Norwich, NR4 7UZ, UK
[3]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, CB10 1SD, UK
[4]University of Oxford, Oxford, OX1 3SZ, UK
[5]Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, UK
[6]Natural History Museum, London, SW7 5BD, UK
[7]Royal Botanic Gardens, Kew, Richmond, TW9 3DS, UK
[8]Marine Biological Association of the UK, Plymouth, PL1 2BP, UK

v1 **First published:** N/A, **N/A**: N/A N/A
Latest published: N/A, **N/A**: N/A N/A

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

Abstract

The vision of the Earth BioGenome Project is to complete reference genomes for all of the planet's ~2M described eukaryotic species in the coming decade. To contribute to this global endeavour, the Darwin Tree of Life Project (DToL, https://darwintreeoflife.org) was launched in 2019 with the aim of generating complete genomes for the ~70k described eukaryotic species that can be found in Britain and Ireland. One of the early tasks of the DToL project was to determine, define, and standardise the important metadata that must accompany every sample contributing to this ambitious project. This ensures high-quality contextual information is available for the associated data, enabling a richer set of information upon which to search and filter datasets as well as enabling interoperability between datasets used for downstream analysis. Here we describe some of the key factors we considered in the process of determining, defining, and documenting the metadata required for DToL project samples. The manifest and Standard Operating Procedure that are referred to throughout this paper are likely to be useful for other projects, and we encourage re-use while maintaining the standards and rules set out here.

Keywords

Species, genomes, biodiversity, long read, conservation, metadata

This article is included in the Tree of Life gateway.

Corresponding author: Mara K.N. Lawniczak (mara@sanger.ac.uk)

Author roles: Lawniczak MKN: Conceptualization, Funding Acquisition, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Davey RP**: Conceptualization, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Rajan J**: Methodology, Writing – Review & Editing; **Pereira-da-Conceicoa LL**: Methodology, Writing – Review & Editing; **Kilias E**: Methodology, Writing – Review & Editing; **Hollingsworth PM**: Methodology, Writing – Review & Editing; **Barnes I**: Funding Acquisition, Methodology, Writing – Review & Editing; **Allen H**: Methodology, Writing – Review & Editing; **Blaxter M**: Funding Acquisition, Methodology, Writing – Review & Editing; **Burgin J**: Methodology, Writing – Review & Editing; **Broad GR**: Methodology, Writing – Review & Editing; **Crowley LM**: Methodology, Writing – Review & Editing; **Gaya E**: Methodology, Writing – Review & Editing; **Holroyd N**: Methodology, Writing – Review & Editing; **Lewis OT**: Methodology, Writing – Review & Editing; **McTaggart S**: Methodology, Writing – Review & Editing; **Mieszkowska N**: Methodology, Writing – Review & Editing; **Minotto A**: Methodology, Writing – Review & Editing; **Shaw F**: Methodology, Writing – Review & Editing; **Richards TA**: Methodology, Writing – Review & Editing; **Sivess LAS**: Methodology, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: The Darwin Tree of Life Project is supported by Wellcome [206194, https://doi.org/10.35802/206194, and 218328, https://doi.org/10.35802/218328]. RPD, AM and FS are funded in part by the UK Biotechnology and Biological Sciences Research Council (BBSRC) Core Strategic Programme grant awarded to the Earlham Institute (BBS/E/T/000PR9817). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: 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 Research , : https://doi.org/

First published: N/A, **N/A**: N/A N/A

Disclaimer

The views expressed in this article are those of the author(s). Publication in Wellcome Open Research does not imply endorsement by Wellcome.

The Darwin Tree of Life project (DToL) is a major collaborative undertaking that seeks to generate complete genomes for ~70k eukaryotic species resident in Britain and Ireland. The project involves dozens of institutes and researchers, and consistent data management across all of these people and sites is key to the success of the project. A clear set of standards and rules (e.g.,1) is an important component of any large sequencing project and makes it much easier to adhere to Findable, Accessible, Interoperable, and Reusable (FAIR) data principles[2]. Future studies using DToL reference genomes will require standardised and accurate recording of the environment from which the genomic type specimen was taken along with its many relevant collection properties. Data also need to comply with the requirements for institutional Collection Management Systems, which are based on Darwin Core standards, and the collection management system needs to be capable of exporting data in the Manifest format. For museum and herbarium collections, minimum data standards are necessary for specimens to meet registry standards for acquisition, and then to be locatable in the collections. Accordingly, we have developed a set of instructions to ensure consistency and standardisation of metadata across DToL. This instruction manual, which we call a Standard Operating Procedure (SOP), gives detailed rules and guidance on how to fill in the DToL Sample Manifest, which can be used as a Google Sheet or an Excel file. The manifest and SOP are version controlled live documents and can be found on the DToL project's GitHub (https://github.com/darwintreeoflife/metadata).

DToL samples are collected and processed with the oversight of a Genome Acquisition Laboratory (GAL) where researchers, often taxonomic experts with detailed knowledge of their research organisms, prepare collected specimens into sequencing-ready samples. DToL largely follows the standards set out by the EBP Sample Collection and Processing subcommittee, which can be found on the EBP website[3]. Samples will typically go through different laboratory (e.g., high molecular weight DNA extraction, RNA extraction) and sequencing (e.g., long read, Hi-C) processes to produce a high quality reference genome. To oversee species collections and records for specimens contributing to the project, DToL has a Samples Working Group (SWG) that brings together researchers representing all GALs within the consortium. The SWG also has representatives with expertise in each of six high-level eukaryotic taxonomic areas: plants, lichens and fungi, chordata, arthropods, protists, and other metazoa (mainly comprising non-arthropod invertebrates). The members of the SWG meet twice a month and are tasked with developing a target list of priority species to be sequenced in each phase of the project, standardising metadata collection, and developing, refining, and standardising collection procedures for different taxonomic groups, as well as protocols for DNA barcoding and sample shipment and storage. The group also works alongside DToL's Regulatory Group to develop and refine guidance for ethical and legal compliance when collecting, holding, and transferring material. Activities such as determining the target species list or developing legal documents for collecting are likely to depend on the aims and location of a project, but ensuring good practice for metadata collection is required for projects contributing to the EBP as this enables re-use and interoperability of associated data. Given the potential for wide re-use of the DToL Sample Manifest and accompanying SOP, this Open Letter describes these documents to facilitate their wider use across all EBP projects.

The Sample Manifest contains many core fields that must be provided by a GAL before the sample material is accepted for sequencing. It aims to capture core metadata from a phylogenetically disparate set of species, rather than an exhaustive metadata set for any specific ecosystem. As such, it does not contain fields that are specific to particular taxa or ecosystems, though these data should also be collected in a standardised fashion where relevant. Having a broad range of taxonomic experts on the SWG helped to ensure that the Sample Manifest captures metadata that are common and informative across all taxonomic groups, while also having a wide enough range of defined terms to support all taxonomic groups. For example, ORGANISM_PART, which captures the tissue type that has been sampled, includes terms that are relevant to plants, fungi, and animals. The SWG also determined that some fields must have meaningful data provided in order for a sample to be accepted into the project (e.g., the location from which the sample was collected), while for other fields completion is encouraged but missing data is permitted (e.g., SEX). Some fields accept only terms from a preset controlled vocabulary, e.g. LIFESTAGE, to ensure that consistency of metadata terms is maintained across communities. In other fields we recommend that ontology terms or community-agreed reference formats are used, e.g. we suggest the use of ENVO[4,5] terms within the HABITAT field. It is important to note that some metadata fields are captured at the project level and are therefore not appropriate to link to downstream archived datasets, whereas others are submitted for archival to the European Nucleotide Archive (ENA) and are publicly available.

The Sample Manifest is processed by COPO, a data brokering system that collects, aggregates, and validates metadata for life science communities[6] such as DToL and also performs data submission alongside these metadata into public repositories such as the ENA. COPO's use for biodiversity genomics will be covered in a separate manuscript. Sample collectors upload the Sample Manifest to COPO which initiates tracking of a sample from collection to sequencing, and performs first pass validation checks against accepted manifest values as determined by the SWG. The metadata are then submitted to the ENA with secondary validation via the Tree of Life sample checklist to ensure the samples comply against the minimum metadata and standard format for BioSamples registered as part of the DToL project. All DToL project samples and associated data are published in the DToL data portal (https://portal.darwintreeoflife.org/), and subsequently linked to the publication of the data on the DToL data portal.

In the next sections, we describe factors considered by the SWG that led to the evolution of the Sample Manifest and SOP and its management.

Version control

The Sample Manifest and SOP are living documents. Version control is important to track the inevitable changes that will occur as we develop our understanding of the use to which metadata will be made (and adjust the defined terms lists), or include new modalities (as new fields). These changes need to be managed across both sample collection and technical systems development, and therefore require rigorous planning to ensure compatibility. The SWG is responsible for recording the changes made in each revision (which currently happen twice per year), retaining all archived versions of the SOP, and naming updated documents with a new version number. We hope that changes will become fewer and fewer as the project progresses and encourage projects adopting the Sample Manifest and SOP to find their latest versions at the DToL project GitHub.

Standardisation of terms -- drop-down menus and data validation

Whenever a field has a limited set of possible entries (i.e. is not free text, numeric, or unique) we have added drop-down menus that provide the exact terms permitted as entries for that field. In the Excel and Google versions of the Sample Manifest, we have set these fields to flag any entries that do not comply with the rules for that field to help alert the person carrying out data entry that they have entered an invalid term. One note of caution: if data are copied and pasted into the Excel or Google manifest, the underlying data validation is overwritten and error flags no longer appear for invalid terms. This level of detail towards keeping metadata entries standardised across different contributors helps enormously when performing wider analyses on submissions because people will tend to have their own preferred terms (for example, the sex of a specimen could be F, f, fem, female, FEMALE, Female, etc). Wherever we have been able to add in pre-defined terms, we have done so (all terms that exist in the drop-down menus can be read on the "Data Validation" tab of the Sample Manifest). We also declare mandatory formats for fields that are often problematic such as dates (all use the ISO 8601 standard) and GPS coordinates (in decimals rather than degrees). We also only permit three distinct and defined 'missing data' terms. These missing data terms are defined in the SOP and differentiate between 1) NOT_APPLICABLE, meaning data are missing because the field is not meaningful for that sample, 2) NOT_COLLECTED, meaning the sample did not have this particular piece of metadata collected though it theoretically could have been collected, and 3) NOT_PROVIDED, meaning the sample might have this missing entry updated at later point, such as for VOUCHER_ID, where a physical voucher of a specimen is later accessioned and curated into a collection or repository.

Most metadata fields are self-explanatory (e.g., taxonomic information, date of collection), but several require further explanation, which can be found in the SOP and are elaborated on below.

"SERIES": The scale of the DToL project is large enough that collecting partners are asked to accumulate at least 50 samples prior to submitting their manifest for validation or shipping for sequencing. These steps can be time consuming (validation) and expensive (shipping) and this is not linear (e.g., one sample can cost just as much to ship as 100 samples). For this reason, the first field in the Sample Manifest is "SERIES" and this simply serves to keep track of how many individual samples (i.e. tubes containing tissue) have been gathered in any one 'batch'.

"RACK_OR_PLATE_ID", "TUBE_OR_WELL_ID": unique barcoded tubes, racks, and plates are expected as part of submission for sequencing in DToL. Sample providers are urged to scan rather than type in barcodes. There are many benefits to using pre-barcoded tubes, racks, and plates even though they add costs. Sample mix-ups due to labeling issues rapidly become more costly (in time and budget) than pre-barcoded consumables. DToL uses only FluidX tubes and racks.

"SPECIMEN_ID": this is a key field that uniquely numbers each single genetic entity or individual, wherever biology allows this. Multiple samples may be taken from the same specimen (e.g., different tissues from one individual animal are put into different tubes) and the original genetic entity information must be carefully tracked across these derivative tubes. Therefore, each sample tube is assigned a SPECIMEN_ID, which is a unique identifier generated by the GAL (sample provider) that reflects the genetic identity of the organism from which the sample originated. If two samples taken from the same specimen are different (e.g., blood in one tube, liver in another), this is captured in further fields (ORGANISM_PART, SIZE_OF_TISSUE_IN_TUBE). Some organisms are too small to be collected in tubes (e.g. cell-sorted environmental protists are prepared in 96 or 384 well plates). These entities are identified by the combination of a unique plate barcode and the well ID, e.g. plate001_A1, plate001_A2, etc. The concept of SPECIMEN_ID is of major importance in building high quality reference genomes because multiple rounds of sequencing or combining different data types are sometimes required to generate sufficient coverage for an assembly. The standard approach of DToL is to sequence single genetic individuals and not to use composite templates containing different genetic individuals. Therefore, when a sequencing library is depleted, top-up sequencing requires another extraction from a sample from the same SPECIMEN_ID whenever possible. It is always recommended to return to material from the same specimen where circumstances allow (i.e., if the quality of the original data was good enough to continue to sequence from that specimen and more tissue from that specimen exists). Building assemblies from long read data arising from multiple genetically distinct organisms of the same species adds significant challenges. Many sample providers will already have unique ways of labeling their specimens, so we support these schemes, but ask that a prefix is added to the unique identifier for each specimen that makes it clear which project partner contributed the specimen (e.g., NHM for Natural History Museum London).

"DIFFICULT_OR_HIGH_PRIORITY_SAMPLE": this is a field that allows sample contributors to declare that this sample is difficult to collect (e.g., rare, or from a place that is difficult to access) or is a high priority and should move to the front of the sequencing queue, for example due to a conservation need. This field can have additional terms added as other projects may have other reasons to flag samples for special attention. For example, in a later version of the Sample Manifest, we added "FULL_CURATION" to the drop-down menu for this field. This allows partners to indicate that the resulting genome is designated as a family-level representative and thus should receive comprehensive (or full, as opposed to rapid) curation to fix any issues that have arisen along the automated assembly pipeline.

"PURPOSE_OF_SPECIMEN": The vast majority of the samples that are submitted for the DToL project are for "REFERENCE_GENOME_SEQUENCING", but sometimes samples are submitted for other reasons, and we capture these potential reasons here. For example, the DToL project aims to generate a DNA barcode for every specimen, which helps confirm species identification but also serves as an independently generated tag for the sample, aiding in identifying sample mixups. For some species (particularly very small taxa from which there is not enough tissue to undertake separate DNA barcoding and genome sequencing), additional individuals are collected as barcoding proxy specimens and used to confirm species identity. Thus, one of the terms in this field is "DNA_BARCODING_ONLY". Finally, our project is also carrying out population genetic studies on some organisms, so tissues may be submitted for "SHORT_READ_SEQUENCING".

Use of universal identifiers and collating information globally

Species do not respect national boundaries, and many projects are international in conception and delivery. The global biodiversity genomics effort being marshalled by the Earth BioGenome Project thus needs coordination and interoperability. Several information systems are being built to promote synergy and communication between these projects, and feed directly into the metadata collection and reporting process.

The Genomes on a Tree (GoaT) (https://goat.genomehubs.org/) platform provides a one-stop resource for genome size, karyotype, and genome sequencing information on species across the eukaryotic tree of life. GoaT has collated information from direct measurements reported in the literature and from submitted genome assemblies, and uses these data to estimate the expected genome sizes of unstudied taxa. GoaT also collates reports from ongoing major genome sequencing projects, allowing the wider community to see what species are in progress and what their statuses are.

Tree of Life IDentifiers (ToLIDs) are a universal system of species and specimen identifiers that aid communication and reporting of genomics efforts. This system uses a controlled-vocabulary text string to simply record the taxonomy of a collected specimen, and place it in the context of other specimens collected for the same taxon. ToLID prefixes capture – in seven to nine characters – the major taxon, subgroup, and species name of a specimen. A numeric suffix records whether the specimen is the first, second, etc. individual (i.e. genetic identity) from this taxon to be collected and processed. Thus, aRanTem1 identifies the first sampled individual of the frog *Rana temporaria*, xgPerPere3 the third sampled individual of the mollusc *Peregriana peregra*. TOLIDs are centrally assigned on request at id.tol.sanger.ac.uk.

Wider use of the Sample Manifest and SOP

In this manuscript, we have summarised our recommendations for collecting in a standardised fashion the minimal information that should accompany any genomic type specimen for biodiversity genomics projects. We welcome the re-use and modification of the DToL Sample Manifest and SOP that accompany this document for other projects. However, we have one important request. We ask that any project adopting the DToL Sample Manifest continue to use the fields (i.e., column headers) we have named and defined in this SOP identically to the definition that accompanies them here. If new projects find the fields are not adequate, they should define new fields, preferably ones which are already used within the biodiversity community and can be aligned with other initiatives. ENA checklists have already been developed using many of these fields which allows cross comparison between samples registered using different sample standards and the Tree of Life checklist in particular was developed with the intention of reuse across different biodiversity reference and monitoring projects. To facilitate coherent development of a standard community approach, we encourage communication (via dtol_swg@sanger.ac.uk) from other developing EBP initiatives to discuss and resolve any issues with the manifest/SOP.

Given the effort involved in each genome that the DToL project releases, including species prioritisation, collection, identification, preservation, extraction, sequencing, assembly, curation, and annotation, together with the immeasurable value of each species' genome in the years to come, having meticulous and complete metadata records is a core principle in the DToL project. Therefore, the evolution of the Sample Manifest and SOP that are described here is the result of many discussions and real-world attempts to gather metadata on real taxonomically diverse samples. We hope it provides a useful template for other biodiversity genomics projects.

Data availability

No data are associated with this article.

Author contributions

The need for a standardised metadata recording scheme was identified by Darwin Tree of Life Project members at the beginning of the project. MKNL created the original DToL Sample Manifest and wrote the accompanying SOP with major developmental input from RPD. Additions from members of the Samples

Working Group including LLP, IB, GRB, EG, PH, HA, NH, LS, MB greatly improved the clarity and applicability for the breadth of taxa for which the project aims to generate reference genomes. FS and AM contributed to manifest development and built the COPO platform. JR and JB developed the ENA ToL checklist. MKNL wrote the paper with input from RPD, MB, PH, IB, JB, OTL and GRB.

Acknowledgements

Patrick Adkins, Rob Mrowicki, Kes Scott-Somme, Joanna Harley, Suzie Warham, and Helen Jenkins (MBA); Chris Fletcher, Lauren Hughes, Inez Januszczak and Olga Sivell (NHM) all road-tested the manifest and provided feedback for improvements.

References

1. Stevens I, Mukarram AK, Hörtenhuber M, *et al.*: **Ten simple rules for annotating sequencing experiments.** *PLoS Comput Biol.* 2020; **16**(10): e1008260.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | Publisher Full Text | Free Full Text
3. **Report on Sample Collection and Processing Standards.** [cited 18 Jun 2021].
Reference Source
4. Buttigieg PL, Morrison N, Smith B, *et al.*: **The environment ontology: contextualising biological and biomedical entities.** *J Biomed Semantics.* 2013; **4**(1): 43.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Buttigieg PL, Pafilis E, Lewis SE, *et al.*: **The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation.** *J Biomed Semantics.* 2016; **7**(1): 57.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Shaw F, Etuk A, Minotto A, *et al.*: **COPO: a metadata platform for brokering FAIR data in the life sciences [version 1; peer review: 1 approved, 1 approved with reservations].** *F1000Res.* 2020; **9**: 495.
Publisher Full Text