Recording Sample Metadata for the European Reference Genome Atlas Project[1]

Sample Manifest Standard Operating Procedure

Version: 2.5.1

Published Date: 11.03.2025

Authors: Jennifer A. Leonard, Olga Vinnere Petterson, Seanna McTaggert, Ann McCartney, Luísa Marins, Torsten Struck, Martin Husemann, Carmela Gissi, Isabelle Florent, Katja Reichel, Seanna McTaggart, Felix Shaw, Joana Pauperio, Josephine Burgin, Rita Monteiro and Astrid Böhne

Original DToL manifest authors: Mara K. N. Lawniczak, Robert P. Davey, Jeena Rajan, Lyndall L. Pereira-da-Conceicoa, Estelle Kilias, Peter M. Hollingsworth, Ian Barnes, Heather Allen, Mark Blaxter, Josephine Burgin, Gavin R. Broad, Ester Gaya, Nancy Holroyd, Inez Januszczak, Owen T. Lewis, Liam M. Crowley, Seanna McTaggart, Nova Mieszkowska, Alice Minotto, Joana Pauperio, Radka Platte, Felix Shaw, Laura A. S. Sivess, Thomas A. Richards, and the Darwin Tree of Life Consortium.

Correct, ethical, and comprehensive recording of sample metadata is critical to the long-term utility of the work we do in the ERGA project: these metadata will link the genome sequences to their origins and remaining voucher material and weave our work into the rich fabric of understanding of European eukaryotic biodiversity. Please read this Standard Operating Procedure (SOP) in full before completing the Sample Manifest as it contains detailed guidance on how to record metadata. Guidance on sample submission to sequencing partners should be sought from the sequencing facility. Guidance on submission of vouchering and biobanking material should be

[1]This sample manifest and SOP is a modified version of the Darwin Tree of Life Sample Manifest and SOP. The citation for the DTOL Sample Manifest and SOP is 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 [doi.org]).

sought from the respective collection facility. Suggestions to this manifest should be sent to the ERGA Sampling and Sample Processing committee SSP samples@erga-biodiversity.eu. Please note that submitting the manifest is mandatory for all ERGA-Pilot, BGE-ERGA, and ERGA community genomes..

Table of Contents

Introduction
Completing the Sample Manifest: Overview
- Scope of this document
- The importance of “SPECIMEN_ID”
- Other “_ID”s
- Table 1: Genome Acquisition Labs (GAL) Codes and Contacts
- Manifest Validation Process
- Changes to Uploaded Sample Metadata
- Photographs of Specimen or Sample
- Vouchers of Specimen or Sample

ERGA sample manifest roadmap
Detailed instructions for filling in the Sample Manifest
Column by column instructions for the Metadata Entry tab.
Table 2 Document History

Introduction

Preamble: To be able to register a sample/specimen and its metadata for ERGA, the submitting person must confirm to be an ERGA member and to adhere to the ERGA code of conduct as well as confirm that sampling adhered to ERGA's ethical code of conduct for sampling.

Purpose and responsibility: ERGA aims to generate high-quality genome sequences and to embed these sequences into best practices in scientific research and the landscape of biodiversity science. To do so we must adhere to correct, legal and ethical handling of the specimens, and correct collation of rich metadata describing the specimens. This SOP contains instructions for filling in the metadata manifest. ERGA will not access samples that have incomplete associated metadata, have not been sampled in compliance with legal rules applying to each specimen, and have not been sampled according to ERGA's ethical code of conduct. The legal responsibility for acquiring samples remains with the sample coordinator. By submitting the sampling manifest and providing information on compliance with sampling permits, the sample coordinator guarantees that the sample in question can be legally transferred to the sequencing center indicated, can be vouchered at the indicated collection(s), and has been sampled in compliance with all applicable rules. The responsibility for the oversight of all legal compliance remains with the sample coordinator. Where necessary and applicable, material transfer agreements can be issued and signed by sequencing and collection facilities and the sample coordinator; the oversight of this process lies with the sample coordinator.

Metadata are collected manually by the sample coordinator using a defined spreadsheet, referred to as the ERGA SAMPLE MANIFEST. This document allows integration into the data management and brokering platform system COPO (http://copo-project.org). COPO allows for dry runs of metadata upload to validate compliance to format requirements. COPO will link to a database that tracks all samples and their associated metadata as they progress from collection to genome assembly. The sequencing data will be archived in the ENA (https://www.ebi.ac.uk/ena/browser) for all sequenced samples with the information provided in the metadata. The update of mandatory information initially set to "NOT_PROVIDED" after initial manifest validation is currently in place

Raising issues: This SOP will be reviewed on a biannual basis by the SSP committee to incorporate feedback from the community. Elements of this SOP are subject to discussion, development and change. We expect that there will be questions to answer and lessons learned to share. Please raise specific issues by emailing the SSP at samples@erga-biodiversity.eu. Please also raise issues with the current manifest and SOP on the corresponding GitHub issue tracker at https://github.com/ERGA-consortium/COPO-manifest/issues. For questions concerning the brokering of the manifest over COPO please reach out to EI.COPO@earlham.ac.uk.

Completing the Sample Manifest: Overview

Scope of this document

This SOP guides you through recording of your samples' metadata into the ERGA sample manifest. **Specific guidance on preparing samples** is not covered by this SOP. Please contact samples@erga-biodiversity.eu to be referred to your corresponding taxonomic subcommittee. **Submission of samples** is also not covered by this SOP. Please contact the involved sequencing partner to obtain necessary information.

The importance of "SPECIMEN_ID"

The ERGA SPECIMEN_ID must reflect the genetic identity of the individual, serving to link the various samples, images, vouchers, DNA barcodes, etc. that derive from one individual organism together. The ERGA SPECIMEN_ID allows resampling the same individual specimen (i.e. the same haplotypes) if needed, e.g., in the case of requiring more DNA to create a library. For example, ten different individual specimens each in their own tube would have ten distinct ERGA SPECIMEN_IDs, even if they are all from the same species, clone or culture. However, samples of a single specimen split across ten tubes would result in each of those ten tubes having the same ERGA SPECIMEN_ID.

Other "_ID"s

Genome Acquisition Labs, "GALs" (the entity performing the actual genome sequencing), can attribute identifiers to each SPECIMEN (GAL_SAMPLE_IDs). Examples of the formats of the ERGA partners and companies for their internal sample tracking are provided in Table 1. Such identifiers need to be provided from the GAL to the sample coordinator, please contact your GAL if those identifiers are available to you upon manifest upload.

The Genome Acquisition Labs GAL_SAMPLE_IDs and COLLECTOR_SAMPLE_IDs can refer to an individual organism or something else (e.g., a soil sample could be represented by the COLLECTOR_SAMPLE_ID and a specimen taken from within that collection of soil be assigned a SPECIMEN_ID). The COLLECTOR_SAMPLE_ID is the identifier assigned by the collector to the specimen or the sample, hence the use of the term SAMPLE rather than SPECIMEN in this metadata field. If a collector collects a sample that could have mixed genotypes or species, this will have a single COLLECTOR_SAMPLE_ID, and will need to be split further into specimens, each of which is assigned a unique SPECIMEN_ID.

It is permitted to have identical names for any or all of three categories (COLLECTOR_SAMPLE_ID, GAL_SAMPLE_ID and SPECIMEN_ID). The SPECIMEN_ID is the only ID that is required for a sample to enter the ERGA workflow and metadata upload to commence. We ask sample providers to complete metadata collection and upload before commencing sequencing to guarantee a sample adheres to ERGA's standards.

Management of COLLECTOR_SAMPLE_ID, GAL_SAMPLE_ID and their relationship to SPECIMEN_ID is the responsibility of the sample coordinator uploading the manifest.

Table 1: Genome Acquisition Labs (GAL) Codes and Contacts

Please reach out to samples@erga-biodiversity.eu if your GAL is not in the list.

GAL	Code Model	Number of digits	Contact Person Email address (as of June 2024)
SANGER INSTITUTE (SAN)	SAN0000000	7	Nancy Holroyd,Ian Still, Radka Platte treeoflifesamples@sanger.ac.uk
EARLHAM INSTITUTE (EI)	EI_00000	5	Seanna McTaggart, seanna.mctaggart@earlham.ac.uk
CENTRO NACIONAL DE ANÁLISIS GENÓMICO (CNAG)			Marta Gut, marta.gut@cnag.eu
SCILIFELAB (SCI)			Olga Vinnere Pettersson, olga.pettersson@igp.uu.se
WEST GERMAN GENOME CENTRE (WGGC)			Peter Nürnberg, nuernberg@uni-koeln.de
NGS COMPETENCE CENTER TÜBINGEN (NCCT)			Nicolas Casadei, Nicolas.Casadei@med.uni-tuebingen.de
FUNCTIONAL GENOMIC CENTER ZURICH (FGCZ)			Simon Oliver Grueter simon.oliver.grueter@fgcz.ethz.ch
GENOSCOPE (GEN)			Pedro Oliveira, pcoutool@genoscope.cns.fr
LAUSANNE GENOMIC TECHNOLOGIES FACILITY (LGTF)			Julien Marquis, contactGTF@unil.ch
DNA SEQUENCING AND GENOMICS LABORATORY, HELSINKI GENOMICS CORE FACILITY (HGCF)			Petri Auvinen, petri.auvinen@helsinki.fi
BERN NGS (NBE)			Pamela Nicholson, pamela.nicholson@vetsuisse.unibe.ch
NORWEGIAN SEQUENCING CENTRE (NSC)			Ave Tooming, ave.tooming-klunderud@ibv.uio.no
UNIVERSITY OF BARI (UBA)	UBA0000000	7	Carmela Gissi, carmela.gissi@uniba.it
UNIVERSITY OF FLORENCE (FL)			Claudio Ciofi, claudio.ciofi@unifi.it
NEUROMICS SUPPORT FACILITY, UANTWERP, VIB (NSF)			Mojca Strazisar, Mojca.Strazisar@uantwerpen.vib.be
SVARDAL LAB, ANTWERP (SVL)			Hannes Svardal, hannes.svardal@uantwerpen.be
LEIBNIZ INSTITUTE FOR THE ANALYSIS OF BIODIVERSITY CHANGE (LIB)	NA	NA	Lars Podsiadlowski, l.podsiadlowski@leibniz-lib.de
INDUSTRY PARTNER (IP)	NA	NA	as provided by sequencing company

Manifest Validation Process

Choose whether you prefer to use the Sample Manifest from the google spreadsheet or another option (e.g., epicollect, ARCGIS).

Google spreadsheet: The Google spreadsheet can be used by ***making a copy*** and using it as an online spreadsheet, or by downloading it and entering data locally. If you choose to do the latter, please download as an XLS/XLSX (Microsoft Excel format) file to ensure that the data validation fields are retained.

Please carefully read the guidance in this SOP for each field, and attempt to get your submitted manifests as close to the guidance as possible (you will be provided with a mock example to facilitate input). You can test your manifest using the COPO demo website to spot e.g. formatting errors before the submission. If your sample requires metadata fields or terms that are not present in the manifest, please contact samples@erga-biodiversity.eu to discuss and define new fields or terms.

We recommend that you retain a copy in Excel (XLS/XLSX) or Google spreadsheet format so as not to lose the data validation given the likelihood that further edits will be required, even after upload to COPO.

Once you have completed entering all metadata, the initial check **upon submission to COPO** will confirm that each TAXON_ID maps to the correct species name. Please read the guidance on TAXON_ID carefully as you should be able to ensure that each TAXON_ID matches a species name in advance of submitting your manifest. The most common issues are a misspelling in the SCIENTIFIC_NAME or the TAXON_ID fields, a species for which no TaxonID is available in the NCBI TaxonomyDB, or a change in the taxonomy not reflected in NCBI TaxonomyDB. These will need to be addressed before the manifest can be validated.

If any issues with the information provided within the sample manifest are identified (e.g., missing mandatory entries, duplicate rows, incorrect date formats), the sample manifest will be returned to you to resolve these issues; within COPO, this will be an iterative process pointing you towards malformatted or missing information. Once this process is complete the manifest is considered to be “validated”.

When a manifest is validated, each sample will be allocated a “ToLID” that reflects both the species and the SPECIMEN_ID (i.e., the genetic identity of the sample) by COPO. The ToLID helps to track the submitted samples through the sequencing process and acts as assembly names when the data are submitted to the INSDC databases. It is constructed from two letters indicating the general area of the taxonomy the species derives from (il indicating Insecta, Lepidoptera) and then seven letters derived from the species binomen (AriAgre for *Aricia agrestis*) and then a number that increments for each specimen added. For more information please visit ToLID.

Once you have ensured that your manifest is ready for validation, follow the guidance of the GAL and collections (if applicable at this stage) involved for sample submission.

When data are submitted to ENA for release (as part of BioSample, raw data and assembly submissions), the submissions will include all of the fields below indicated by **ENA_submission**. If the field name is in **turquoise**, then an entry for each specimen is mandatory for that field, even if only to declare why the information is missing. For all other fields, we strongly encourage data entry, but it is not mandatory if it has not been collected.

Changes to Uploaded Sample Metadata

COPO has a version history. In case of need for update please reach out to manifest managers (ERGA pilot: pilot@erga-biodiversity.eu, ERGA-BGE: r.monteiro@leibniz-lib.de). COPO users can update their own manifests in case of changes. Please see the following documentation: https://copo-docs.readthedocs.io/en/latest/updates/samples.html/. Not all fields can be updated, users should contact EI.COPO@earlham.ac.uk for further guidance. For a list of fields which can be updated, see: https://copo-project.org/api/sample/updatable_fields/ERGA/.

Photographs of Specimen or Sample

Every submitted specimen should be accompanied by a photograph as for now according to standards for taxonomic groups developed in the European ICEDIG and DISSCO projects (for taxon specific recommendations please see https://icedig.eu/content/deliverables) with explicit labeling as described below. Please reach out to samples@erga-biodiversity.eu if you can provide a standard or there is no appropriate standard for your taxon of interest. Images will be deposited in BioImage Archive. **To upload images, first upload a manifest which validates without error. At this point, click the "Upload Images" button. Select the images which correspond to the specimens.**

Please name images using the following format: **SPECIMEN_ID-X.Y** where X is a numerical identifier for the number of photographs you have taken of the same individual, and Y is the file format, e.g., SPECIMEN_ID-1.png and SPECIMEN_ID-2.png for two photos of the same specimen provided. Please use **PNG or JPG** format. **File names must exactly match the SPECIMEN_ID** in order to match photographs to samples automatically.

Vouchers of Specimen or Sample

A submitted specimen must be vouchered in a public scientific collection dedicated to permanent storage and with an accessible voucher catalog. Upon integration in a collection, vouchers will be attributed a collection specific ID that can be recorded in the metadata alongside the collection name. To be properly displayed in INSDC, collections (e.g. museums, herbaria, collections) should register with the NCBI Biocollections database (https://www.ncbi.nlm.nih.gov/biocollections) by contacting gb-admin@ncbi.nlm.nih.gov. To confirm if the collection is already in the NCBI Biocollections or to look for the correct institution and collection codes, the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/) may be used, or the database can be looked up here: https://ftp.ncbi.nih.gov/pub/taxonomy/Cowner_dump.txt (Institutions) https://ftp.ncbi.nih.gov/pub/taxonomy/Ccode_dump.txt (Collections). Ideally, in addition to a physical voucher, tissue and/or cells and/or DNA should be deposited in a public Biobank/frozen repository, ideally a member of GGBN. If voucher material cannot be derived from the same specimen as the genome data, e.g. due to size limitations, the voucher can also be another specimen of the same species from the same population, i.e. use a proxy voucher. Ideally, the voucher is barcoded to show its genetic similarity with the submitted specimen. When even this proxy vouchering is not possible as the species is, for example, very rare or problematic to sample, we ask that digital images are recorded prior to destructive sampling and submitted in lieu of physical samples (digital voucher, see above Photographs of Specimen). If sample providers do not have access to collections and/or biobanks, reach out to samples@erga-biodiversity.eu to be directed to an appropriate ERGA partner facility for vouchering.

ERGA sample manifest roadmap

The manifest is divided into eleven theme blocks covering different aspects of metadata acquisition.

Mandatory fields are marked in **bold** in the table below.

Block 1: Sample submission information including specimen identifier and tube/well identifiers, as well as information on the sample coordinator
(columns A to F)
Block 2: Taxonomic information including species name, family and common name
(columns G to O)
Block 3: Biological information of the sample including lifestage, sex, and organism part
(columns P to T)
Block 4: Details of the submitting GAL and the associated organisational codes
(columns U and V)
Block 5: Data on the collector, collection event, and collection localities
(columns W to AR)
Block 6: Information on taxonomic identification, taxonomic uncertainty and risks
(columns AS to AW)
Block 7: Details of the tissue preservation event
(columns AX to BD)
Block 8: Information on DNA barcoding
(columns BE to BI)
Block 9: Information on Biobanking and Vouchering (columns BJ to BU)
Block 10: Information on regulatory compliances, Indigenous rights, traditional knowledge and permits
(columns BV to CI)
Block 11: Additional information including a free text field to house other important sample notes
(columns CJ and CM)

A	B	C	D	E	F	G
TUBE_OR_WELL_ID	SPECIMEN_ID	PURPOSE_OF_SPECIMEN	SAMPLE_COORDINATOR	SAMPLE_COORDINATOR_AFFILIATION	SAMPLE_COORDINATOR_ORCID_ID	ORDER_OR_GROUP

H	I	J	K	L	M	N
FAMILY	**GENUS**	**TAXON_ID**	**SCIENTIFIC_NAME**	TAXON_REMARKS	INFRASPECIFIC_EPITHET	CULTURE_OR_STRAIN_ID
O	**P**	**Q**	**R**	**S**	**T**	**U**
COMMON_NAME	**LIFESTAGE**	**SEX**	**ORGANISM_PART**	SYMBIONT	RELATIONSHIP	**GAL**
V	**W**	**X**	**Y**	**Z**	**AA**	**AB**
GAL_SAMPLE_ID	**COLLECTOR_SAMPLE_ID**	**COLLECTED_BY**	**COLLECTOR_AFFILIATION**	**COLLECTOR_ORCID_ID**	**DATE_OF_COLLECTION**	TIME_OF_COLLECTION
AC	**AD**	**AE**	**AF**	**AG**	**AH**	**AI**
COLLECTION_LOCATION	**DECIMAL_LATITUDE**	**DECIMAL_LONGITUDE**	LATITUDE_START	LONGITUDE_START	LATITUDE_END	LONGITUDE_END
AJ	**AK**	**AL**	**AM**	**AN**	**AO**	**AP**
HABITAT	DEPTH	ELEVATION	ORIGINAL_COLLECTION_DATE	ORIGINAL_GEOGRAPHIC_LOCATION	ORIGINAL_DECIMAL_LATITUDE	ORIGINAL_DECIMAL_LONGITUDE
AQ	**AR**	**AS**	**AT**	**AU**	**AV**	**AW**
DESCRIPTION_OF_COLLECTION_METHOD	**DIFFICULT_OR_HIGH_PRIORITY_SAMPLE**	**IDENTIFIED_BY**	**IDENTIFIER_AFFILIATION**	**IDENTIFIED_HOW**	**SPECIMEN_IDENTITY_RISK**	**MIXED_SAMPLE_RISK**
AX	**AY**	**AZ**	**BA**	**BB**	**BC**	**BD**
PRESERVED_BY	**PRESERVER_AFFILIATION**	**PRESERVATION_APPROACH**	PRESERVATIVE_SOLUTION	**TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION**	**DATE_OF_PRESERVATION**	**SIZE_OF_TISSUE_IN_TUBE**
BE	**BF**	**BG**	**BH**	**BI**	**BJ**	**BK**
TISSUE_REMOVED_FOR_BARCODING	**TUBE_OR_WELL_ID_FOR_BARCODING**	**TISSUE_FOR_BARCODING**	**BARCODE_PLATE_PRESERVATIVE**	**BARCODING_STATUS**	**TISSUE_REMOVED_FOR_BIOBANKING**	**TISSUE_VOUCHER_ID_FOR_BIOBANKING**
BL	**BM**	**BN**	**BO**	**BP**	**BQ**	**BR**
PROXY_TISSUE_VOUCHER_ID_FOR_BIOBANKING	**TISSUE_FOR_BIOBANKING**	**BIOBANKED_TISSUE_PRESERVATIVE**	**DNA_REMOVED_FOR_BIOBANKING**	**DNA_VOUCHER_ID_FOR_BIOBANKING**	**VOUCHER_ID**	PROXY_VOUCHER_ID
BS	**BT**	**BU**	**BV**	**BW**	**BX**	**BY**
VOUCHER_LINK	PROXY_VOUCHER_LINK	VOUCHER_INSTITUTION	**REGULATORY_COMPLIANCE**	**ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_RIGHTS_APPLICABLE**	INDIGENOUS_RIGHTS_DEF	ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_PROJECT_ID
BZ	**CA**	**CB**	**CC**	**CD**	**CE**	**CF**
ASSOCIATED_TRADITIONAL_KNOWLEDGE_CONTACT	**ETHICS_PERMITS_REQUIRED**	**ETHICS_PERMITS_DEF**	**ETHICS_PERMITS_FILENAME**	**SAMPLING_PERMITS_REQUIRED**	**SAMPLING_PERMITS_DEF**	**SAMPLING_PERMITS_FILENAME**
CG	**CH**	**CI**	**CJ**	**CK**	**CL**	**CM**
NAGOYA_PERMITS_REQUIRED	**NAGOYA_PERMITS_DEF**	**NAGOYA_PERMITS_FILENAME**	**HAZARD_GROUP**	PRIMARY_BIOGENOME_PROJECT	ASSOCIATED_PROJECT_ACCESSIONS	OTHER_INFORMATION

Detailed instructions for filling in the Sample Manifest

I. The manifest has several tabs. Please only fill in the **Metadata Entry** tab. If you discover a missing attribute in the drop-down menus, new attributes can be suggested to the SSP committee at samples@erga-biodiversity.eu.

II. **Information must be entered for all fields below with turquoise bold names**. In the Google spreadsheet version of the manifest, these fields are represented by cells with pink fill. The fill will turn white when an entry has been made to help you identify where mandatory fields still require data. All **mandatory fields** with **turquoise bold names**, even if information is unavailable, must be populated with the appropriate term. The **acceptable missing value terms** follow the INSDC recommendations:

NOT_APPLICABLE = information is inappropriate to report. This can also indicate that the standard itself fails to model or represent the information appropriately.

NOT_COLLECTED= information of an expected format was not given because it has not been collected.

NOT_PROVIDED = information of an expected format cannot be given upon initial manifest submission but a value may be given at a later stage (particularly useful missing information term for VOUCHER_ID, TISSUE_VOUCHER_ID_FOR_BIOBANKING and DNA_VOUCHER_ID_FOR_BIOBANKING)

Fields that are named in **BOLD** without color do not require any entry (e.g., most samples will not have DEPTH information). However, if you have collected the information related to these terms, please enter it to enrich your metadata reporting.

Terms that will have the data released publicly as part of the INSDC record are indicated by "**ENA_submission**" next to the name of the term.

III. **All dates** in the manifest must be formatted consistently as **YYYY-MM-DD** (ISO8601).

IV. In fields that are "free text", please use only the core alphanumeric characters, full stop ".", hyphen "-", underscore "_" and spaces (summarised in coding parlance as " **-_.a-zA-Z0-9**"). Please avoid "|" (the vertical pipe symbol) except where we indicate it should be used to separate elements in a list. **Do not** use "special characters" (such as "#"';:?!@*()[]{}/\,=+", etc.)

Column by column instructions for the Metadata Entry tab.

A. **TUBE_OR_WELL_ID**: Record the individually attributed label on the tube submitted for sequencing. If samples are submitted in plate format, provide the relevant well information here. If tube barcodes are entered, use a barcode scanner in advance of preparing samples to reduce errors. Free text.

B. **SPECIMEN_ID:** (**ENA_submission**) This is the unique identifier that refers to the genetic identity of the supplied material. The ERGA SPECIMEN_ID shall refer to a singular genetic individual. If the same individual specimen is split into several samples submitted in separate tubes, the ERGA SPECIMEN_ID for these samples will be the same. Multiple individuals of a species must be placed in individual tubes, each with a unique ERGA SPECIMEN_ID. If sampling from organisms where distinguishing genetic individuals is difficult (e.g mat-forming species like mosses or bryozoans or colonial ascidians), tease out genetic/clonal individual units (genets) as far as possible, place each in a separate tube with a unique ERGA SPECIMEN_ID. ERGA SPECIMEN_IDs must be standardized, to format ERGA_XY_0123_00001, beginning with the prefix "**ERGA_"** followed by "**SAMPLE COORDINATOR" initials** (up to 10 letters out of A-Z, if this is not possible reach out to samples@erga-biodiversity.eu), followed by **"_"** underscore, **"last four digits of OrcidID"** of SAMPLE COORDINATOR, "_" underscore, **RUNNING NUMBERS** (make sure to use 00001, 00002…). SPECIMEN_IDs must be unique to an individual (e.g., ERGA_XY_0123_00001 cannot be used again after it has been assigned to a specimen). Free text.

C. **PURPOSE_OF_SPECIMEN**: Select appropriate from the drop-down menu.
   - Use **REFERENCE GENOME** for all specimens/samples of a particular species unless they should be destined for an alternative use only. All samples listed for REFERENCE_GENOME sequencing are assumed to also need DNA barcoding and RNA sequencing; the term "REFERENCE GENOME" encompasses all three things (reference genome, barcoding, RNA sequencing), wherever samples allow. Indicate over RELATIONSHIP if a specimen is solely used for one method (e.g., you use one specimen for HiFi data and another one for scaffolding/RNA sequencing).
   - If a particular tissue is needed solely for RNAseq use "**RNA-SEQUENCING**"
   - If the specimen is intended for population genetics or resequencing use "**RESEQUENCING**.
   - If a particular tissue or specimen is intended for research and development, use "**R&D**". These samples may be used for protocol testing.
   - **DNA_BARCODING_ONLY** is reserved for specimens submitted solely for DNA barcoding (e.g., when the sample is too small to provide material for both reference genome and barcoding and genome paratype/other specimens must be used as proxies; or when the specimen was identified to species level but died before being preserved).
   - If the final intended purpose for the sample is not decided at the time of sample manifest submission, use NOT_PROVIDED

D. **SAMPLE_COORDINATOR:** (**ENA_submission**) For the ERGA pilot, also known as

the ERGA sample ambassador, for the BGE project the Genome Team Coordinator. Enter the name of the person or people who is responsible for the genome project of the sample using all CAPITALS, and separate names with "|" (vertical pipe symbol), e.g., "CAROLUS LINNAEUS | JEAN_BAPTISTE LAMARCK". Free text.

- We note that storage of names with affiliations in a database brings the system under the aegis of the GDPR regulations, and we must ask all involved to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of record).

E. **SAMPLE_COORDINATOR_AFFILIATION**: (**ENA_submission**) Supply the affiliation that is responsible for the genome project of the sample, typically the society or institution of the person(s) specified in the SAMPLE_COORDINATOR field. If multiple people are specified in SAMPLE_COORDINATOR, ensure that their institutional affiliations are separated by a vertical pipe symbol here, match the position in the list of affiliations to the same position in the list of names (e.g., PERSON A | PERSON X | PERSON C will have their affiliations as: (INSTITUTE A | INSTITUTE X | INSTITUTE C). If multiple people are listed but all from the same affiliation, no need to repeat the affiliation. Free text.

F. **SAMPLE_COORDINATOR_ORCID_ID**: (**ENA_submission**) Enter the 16 digits ORCID ID of the person/people responsible for the genome project of the sample in the format 0000-0000-0000-0000. Multiple entries need to be separated by a vertical pipe symbol (e.g., 0000-0000-0000-0000 | 0000-0000-0000-0000). Free text.

G. **ORDER_OR_GROUP**: The taxonomic Order into which the Family is placed or (if this is not defined) the monophyletic group to which the Family or Genus belongs. This should correspond to the taxonomy as represented in the NCBI TaxonomyDB. In case of a disagreement with NCBI TaxonomyDB, please raise this with the NCBI TaxonomyDB curators (see below). Free text.

H. **FAMILY**: The taxonomic Family into which the Genus is placed. This should correspond to the taxonomy as represented in the NCBI TaxonomyDB. In case of a disagreement with NCBI TaxonomyDB, please raise this with the NCBI TaxonomyDB curators (see below). Free text.

I. **GENUS**: The taxonomic Genus to which the Species belongs. This should correspond to the taxonomy as represented in the NCBI TaxonomyDB, and with the generic component of TAXON_ID. In case of a disagreement with NCBI TaxonomyDB, please raise this with the NCBI TaxonomyDB curators (see below). Free text.

J. **TAXON_ID**: (**ENA_submission**) A valid NCBI TAXON_ID to the species or where applicable subspecies level is mandatory in order to submit data to public repositories. The species name in the manifest must be identical to that listed in the "current name" box in the Taxonomy Browser for that species. If this is not the case, you must write to ena-bge@ebi.ac.uk to request the change.
If there is another taxon database for your group, e.g., EukRef, LSIDs, please fill in the NCBI TAXON_ID, and use the TAXON_REMARKS field to specify the taxon database and the ID/accession/URL.

- TAXON_IDs can be looked up based on the species at the following links:

https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
or
https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi
or
https://www.ebi.ac.uk/ena/taxonomy/rest/scientific-name/"organismname", where the species name should be entered instead of "organism name" (e.g. https://www.ebi.ac.uk/ena/taxonomy/rest/scientific-name/Trechus%20terceiranus)

- TAXON_IDs are suitable for use if they qualify for
  a) species level (ENA "rank" : "species")
  b) subspecies level (ENA "rank" : "subspecies")
  c) ENA accepts them as submittable (ENA "submittable" : "true")

- If no TAXON_ID exists, or a credible TAXON_ID exists that likely is a synonym of the species name the collector or submitter would use (through differential usage, error or lack of currency of the NCBI TaxonomyDB), please write to ena-bge@ebi.ac.uk, providing the full name, authority, and publication for the chosen name where possible. If required (e.g., newly described species, species missing from NCBI taxonomy), a new TAXON_ID should be available within 14 days. The final species name on submission of the data to INSDC will be the one associated with the TAXON_ID in NCBI TaxonomyDB.
- If a TAXON_ID exists but the taxonomy is not resolved to the species level, please request a placeholder_ID from ena-bge@ebi.ac.uk using a unique identifier after the genus name. The new placeholder_ID should be available within 14 days. Informal names are described at https://ena-docs.readthedocs.io/en/latest/faq/taxonomy_requests.html#unidentified-novel-organisms
- When a sample is provided that requires DNA barcoding before a species ID is possible, please provide the lowest taxonomic rank identification as possible (ORDER_OR_GROUP, FAMILY, GENUS) and leave SCIENTIFIC_NAME blank. Please place comments on what the specimen is likely to be in TAXON_REMARKS.

K. **SCIENTIFIC_NAME**: (**ENA_submission**) The latin binomial/combined genus and species name with a space in between. Free text.
   - See TAXON_ID above if you or the taxonomic expert have substantive issues with the species name present for the taxon in the NCBI TaxonomyDB.
   - Any changes to SCIENTIFIC_NAME post manifest submission to COPO (due to species re-identification or other taxonomic change), should be requested by email to COPO.EI@ealrham.ac.uk and should include new SCIENTIFIC_NAME and BioSamples accession (other related taxonomic fields will be auto-filled by COPO). If applicable, please include information for the fields COMMON_NAME, TAXON_REMARKS and INSTRASPECIFIC_EPITHET, otherwise these will be overwritten and left **blank**.

L. **TAXON_REMARKS**: Free text to summarize any known issues with the mapping of TAXON_ID to SCIENTIFIC_NAME or add other taxon database identifiers here e.g.,

EukRef. Here you can also comment on STRAIN availability, if the specimen is a representative of a living and accessible strain/colony/culture. If there are no issues, leave this field **blank**.

M. **INFRASPECIFIC_EPITHET**: Where the sample is from a formally named infraspecific taxon, give the infraspecific name here, with prefixes in the following format: ssp. (for subspecies), var. (for variety), cv. (for cultivar), br. (for breed). Entries in this field should reflect organisms that can be found living outside of laboratories (see next attribute for lab strains). If there is no epithet here, leave this field **blank**.

N. **CULTURE_OR_STRAIN_ID:** (ENA_submission) Give the reference ID from the source culture collection, such that the culture accession can be found in the collection's database. This is only relevant if the sequenced material is derived from a living, culturable, named laboratory strain (e.g., *Anopheles coluzzii* N'Gousso strain). This field should not be used to record a variant or type that has been collected anew from the wild: such information should be placed in **OTHER_INFORMATION**. Leave this field **blank** if it is not relevant. Free text.

O. **COMMON_NAME**: Vernacular name. If any guidelines for vernacular names exist (e.g., birds, reptiles), their adoption is recommended. Multiple names of multiple languages can be entered by separating names with a | (vertical pipe) character. English common names, if available, should be entered first. If you are unsure of or the species has no vernacular name leave this field **blank**. Free text.

P. LIFESTAGE: (ENA_submission) The life stage of the specimen from which the sample was derived. Use the drop-down menu or look at the available terms on the second tab of the manifest to complete. Field with controlled vocabulary.
   - If these do not fit your taxa, contact samples@erga-biodiversity.eu. Please enter **NOT_PROVIDED** if your proposal for a lifestage term has not yet been accepted.

Q. SEX: (ENA_submission) The sex of the specimen from which the sample was derived as determined by determined by morphological examination of the specimen or strong inference (e.g., the species is from a clade that is always hermaphroditic/monoecious). Use the drop-down menu. If the sex of the organism is not known, use **NOT_COLLECTED**. The sex may be determined at a later date using the genome sequence data, but this will be captured in a different field. Field with controlled vocabulary.

R. ORGANISM_PART: (ENA_submission) Description of the exact tissue(s) in the tube/well. Accurate information is important for downstream analyses on the symbiome, chromosomal diminution, RNAseq, etc. There is a tab in the Manifest that defines the terms that can be used for ORGANISM_PART listing definitions for the full tissue, but pieces of that tissue are acceptable (e.g., LUNG is defined as 'the lung of a vertebrate', but a small piece of lung is usually expected). Field with controlled vocabulary.
   - If applicable, combine tissues by entering multiple terms from the ontology using the | (vertical pipe) symbol (e.g., for head + abdomen of an insect enter "HEAD | ABDOMEN"). When using multiple body parts, there will be a data validation error that arises in the excel metadata sheet, but these can be ignored as long as the spelling and capitalization of the terms is identical to the provided list. This

will not cause a validation error in COPO. If you are filing in the manifest in excel, you may need to change your field encoding/settings to fill in several terms instead of choosing from the drop-down menu of single terms.

- If the tissue or organism part you are providing is not present in the drop-down menu, choose the best generic category (these start with **) and add the name of the tissue that you have put into the tube in the "OTHER_INFORMATION" free text field. Please email to samples@erga-biodiversity.eu to request necessary additions.
- If the sample is shipped as a DNA or RNA extract, select the tissue from which this was extracted and add further information in the OTHER_INFORMATION field regarding quality, quantity, etc. Note that any shipment of DNA should be discussed in advance with the involved GAL.

S. **SYMBIONT**: This is to indicate whether the sample contains a known symbiont (i.e. you have metadata for it and a species/subspecies level and ENA-submittable TAXON ID). Select "TARGET" if only the "host" metadata is known OR if it is a symbiont-only culture. Thus the default entry for this row should be "TARGET" (if this field is left blank, it will be autofilled as "TARGET" on submission). ONLY select "SYMBIONT" if you have a known symbiont mixed with the "TARGET" AND you have a species/subspecies level identification supported by a valid taxon ID for this symbiont. Where this is the case, the "TARGET" row should be duplicated by copying and pasting it below to create a new row; The term "SYMBIONT" should then be selected in the new row, and then the following fields amended to reflect the symbiont data:
ORDER_OR_GROUP, FAMILY, GENUS, TAXON_ID,
SCIENTIFIC_NAME, TAXON_REMARKS, INFRASPECIFIC_EPITHET,
CULTURE_OR_STRAIN_ID, COMMON_NAME, LIFESTAGE, SEX

The default entry for "ORGANISM_PART" for symbionts should be "WHOLE ORGANISM"; it will be auto-corrected to this on submission. Where there is no explicit species-level specific information for the symbiont available (including a valid taxon ID), then no additional symbiont row should be added, and instead any information on the symbiont should be included in the "OTHER_INFORMATION" column of the "TARGET" row. If the presence of a symbiont is known or likely, but its exact taxonomy is unknown, leave SYMBIONT blank and set MIXED_SAMPLE_RISK to Yes. Field with controlled vocabulary.

T. **RELATIONSHIP:** (**ENA_submission**) Free text field to permit declaration of any known parental, child, or sibling relationship between the specimen and any other specimens that are submitted for the ERGA project, OR to declare if the specimen is a "barcode exemplar" for another specimen.

- If there are known genetic relationships between submitted specimens, please concisely state the relationship: "Full sibling to SPECIMEN_ID1", "Mother to SPECIMEN_ID2", "Maternal half sibling to SPECIMEN_ID1, SPECIMEN_ID2, and SPECIMEN_ID3", or "Trio child of SPECIMEN_ID1 and SPECIMEN_ID2". If knowledge of the relationships is not confident but suspected, do not add anything here and instead add this information to the "OTHER_INFORMATION" field (e.g., "suspected full or half sibling to SPECIMEN_ID2").
- If the specimen is acting as a barcoding exemplar or if it is used for a

complementary sequencing method because the entire organism must be used for (one method of) reference genome sequencing and it is not possible to take a sample for DNA barcoding (e.g., midges from the same swarm where one is submitted for sequencing and 5 are submitted individually for DNA barcoding), then add "barcode/additional sequencing exemplar for SPECIMEN_IDx" and insert the SPECIMEN_ID for the specimen that is going for reference genome sequencing, potentially without its own DNA barcoding.

- If there is no relationship to note, this field can be left **blank**.

U. GAL: (ENA_submission) Use the drop-down menu to select the Genome Acquisition Lab (GAL) responsible for this sample. If your GAL is not available, select "**Other_ERGA_Associated_GAL**" and send a request to integrate your GAL for the next manifest release to samples@erga-biodiveristy.eu. Your sample(s) need to be attributed to a GAL for manifest upload. Field with controlled vocabulary.

V. GAL_SAMPLE_ID: (ENA_submission) This is the unique name assigned to the sample by the GAL, matching entries in Table 1. This is a free text field, do not use spaces or special characters (e.g., #, !, ^, *) . It is fine for the GAL_SAMPLE_ID to be the same as the COLLECTOR_SAMPLE_ID and the SPECIMEN_ID if warranted. GALs may maintain their own registers of GAL_SPECIMEN_IDs for the project. Please ensure with your GAL that you do not use IDs that have already been used, and stick to the format required by the GAL. Free text.

W. COLLECTOR_SAMPLE_ID: Unique name assigned to the sample by the COLLECTOR or COLLECTOR_AFFILIATION. **Do not use spaces or special characters**, other than hyphens and underscores ("-" and "_") i.e do not use #, !, ^, *, etc. Free text.

- If you split a single specimen across multiple tubes (see SPECIMEN_ID), please consider what kind of information you want in your unique sample names for this. E.g., if the specimen is a butterfly with SPECIMEN_ID = ERGA_SAI_1234_01, and you put the head in one tube and the thorax in another, your COLLECTOR_SAMPLE_IDs might reflect this with one tube called ERGA_SAI_1234_01-h and the other called ERGA_SAI_1234_01-t. Likewise, the COLLECTOR_SAMPLE_ID may be the name given to a collection consisting of a 'clump' from a mat-forming species, which may then be subdivided into different specimen tubes, each given a unique SPECIMEN_ID.

X. COLLECTED_BY: (ENA_submission) Enter the name of the person or people who collected the sample using all CAPITALS, and separate names with "|" (vertical pipe symbol), e.g., "CAROLUS LINNAEUS | JEAN_BAPTISTE LAMARCK". Free text.

- We note that storage of names with affiliations in a database brings the ERGA system under the aegis of the GDPR regulations, and we must ask sample coordinators, GALs, and collaborators to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of record). The sample coordinator is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

Y. COLLECTOR_AFFILIATION: (ENA_submission) Supply the affiliation that is

responsible for the collected specimen, typically the affiliation of the person(s) specified in the COLLECTED_BY field. If multiple people are specified in COLLECTED_BY, ensure that their affiliations are separated by a vertical pipe symbol. Position in the list of affiliations should match the person in the same position in the list of names (e.g., PERSON A | PERSON X | PERSON C will have their affiliations as: (INSTITUTE A | INSTITUTE X | INSTITUTE C). If multiple people are listed but all from the same affiliation, no need to repeat the affiliation. Free text.

Z. **COLLECTOR_ORCID_ID:** (**ENA_submission**) Enter the 16 digits ORCID ID of the person or people who is responsible for the collection of the sample in the format 0000-0000-0000-0000. If more than a single entry is specified separate by a vertical pipe symbol (e.g. 0000-0000-0000-0000 | 0000-0000-0000-0000). Free text.

AA. **DATE_OF_COLLECTION**: (**ENA_submission**) The date the sample was collected, with year, month, and day specified (**YYYY-MM-DD**). Provide maximum possible precision (i.e., you can also only provide YYYY-MM or YYYY if you do not have DD or DD-MM).

- If the specimen is from a zoo, botanic garden, culture collection or similar and has a known origin elsewhere, please note this information in ORIGINAL_COLLECTION_DATE, ORIGINAL_GEOGRAPHIC_LOCATION and ORIGINAL_DECIMAL_LATITUDE & ORIGINAL_DECIMAL_LONGTITUDE and only include here the date at which a sample was taken.

AB. **TIME_OF_COLLECTION**: (**ENA_submission**) Time of day of sample collection in 24-hour clock format, with hours and minutes separated by colon e.g., 13:35, 04:53, etc. This should be in GMT/UTC. This field may be particularly relevant for RNAseq. Leave this field **blank** if the time was not recorded.

- If the specimen is from a zoo, botanic garden, culture collection or similar and has a known origin elsewhere, please note this information in **ORIGINAL_COLLECTION_DATE**, **ORIGINAL_GEOGRAPHIC_LOCATION**, **ORIGINAL_DECIMAL_LATITUDE** and **ORIGINAL_DECIMAL_LONGTITUDE** and only include here the exact time at which a sample was taken.

AC. **COLLECTION_LOCATION**: (**ENA_submission**) General description of the location where the specimen or sample was sampled for genome sequencing. Start with the geographical origin of the sample country as defined by the country or sea in agreement with INSDC country list (look up accepted country names here https://www.insdc.org/country.html), also include more specific locations (e.g., “Barton’s Pond”) ranging from least to most specific and separated by “|” pipe character, e.g., “United Kingdom | East Anglia | Norfolk | Norwich | University of East Anglia | UEA Broad”. It is important to give the name of the site here if possible. Free text.

- If the specimen is from a zoo, botanic garden, culture collection or similar and has a known origin elsewhere, please note this information in **ORIGINAL_COLLECTION_DATE, ORIGINAL_GEOGRAPHIC_LOCATION**, **ORIGINAL_DECIMAL_LATITUDE** and **ORIGINAL_DECIMAL_LONGTITUDE** and only include here information about the location of the specimen at the time from which a sample was taken (e.g., “London Zoo”, “Millennium Seed Bank”, etc).

AD. **DECIMAL_LATITUDE**: (**ENA_submission**) The geographic location where the

specimen or sample was taken in decimal degrees, between -90 and 90. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees). Provide maximum possible precision. Using 3 decimal places gives a location accurate to 111 meters, using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters (http://wiki.gis.com/wiki/index.php/Decimal_degrees).

- If the specimen is from a zoo, botanic garden, culture collection or similar and has a known origin elsewhere, please note this information in **ORIGINAL_GEOGRAPHIC_LOCATION and ORIGINAL_DECIMAL_LATITUDE & ORIGINAL_DECIMAL_LONGTITUDE** and **only** include here the coordinates of information about the location of the specimen at the time from which a sample was taken (e.g., the coordinates of “London Zoo”, “Millennium Seed Bank”, etc).
- Only provide if **LATITUDE_START** and **LATITUDE_END** are set to “NOT_COLLECTED”.
- If not known, use **NOT_COLLECTED**

AE. **DECIMAL_LONGITUDE**: (**ENA_submission**) The geographic location where the specimen or sample was taken in decimal degrees, between -180 and 180. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees). Provide maximum possible precision. Using 3 decimal places gives a location accurate to 111 meters, using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters (http://wiki.gis.com/wiki/index.php/Decimal_degrees).

- If the specimen is from a zoo, botanic garden, culture collection and has a known origin elsewhere, please note this information in **ORIGINAL_GEOGRAPHIC_LOCATION and ORIGINAL_DECIMAL_LATITUDE & ORIGINAL_DECIMAL_LONGTITUDE** and **only** include here the coordinates of information about the location of the specimen at the time from which a sample was taken (e.g., the coordinates of “London Zoo”, “Millennium Seed Bank”, etc).
- Only provide if **LONGITUDE_START** (AG) and **LONGITUDE_END** (AI) are set to “NOT_COLLECTED”.
- If not known, use **NOT_COLLECTED**

AF. **LATITUDE_START**: (**ENA_submission**) Only fill in if your sample was collected in a transect and cannot be attributed to a single point location. The geographic location where the collection transect started in decimal degrees, between -90 and 90. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees). Provide maximum possible precision. Using 3 decimal places gives a location accurate to 111 meters, using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters (http://wiki.gis.com/wiki/index.php/Decimal_degrees).

- Only provide if **DECIMAL_LATITUDE** is set to “NOT_COLLECTED”.

AG. **LONGITUDE_START**: (**ENA_submission**) Only fill in if your sample was collected in a transect and cannot be attributed to a single point location. The geographic location where the collection transect started in decimal degrees, between -180 and 180. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees). Provide maximum possible precision.

Using 3 decimal places gives a location accurate to 111 meters, using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters (http://wiki.gis.com/wiki/index.php/Decimal_degrees).

- Only provide if **DECIMAL_LONGITUDE** is set to "NOT_COLLECTED"

AH. **LATITUDE_END**: (**ENA_submission**) Only fill in if your sample was collected in a transect and cannot be attributed to a single point location. The geographic location where the collection transect ended in decimal degrees, between -90 and 90. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees). Provide maximum possible precision. Using 3 decimal places gives a location accurate to 111 meters, using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters (http://wiki.gis.com/wiki/index.php/Decimal_degrees).

- Only provide if **DECIMAL_LATITUDE** is set to "NOT_COLLECTED".

AI. **LONGITUDE_END**: (**ENA_submission**) Only fill in if your sample was collected in a transect and cannot be attributed to a single point location. The geographic location where the collection transect ended in decimal degrees, between -180 and 180. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees). Provide maximum possible precision. Using 3 decimal places gives a location accurate to 111 meters, using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters (http://wiki.gis.com/wiki/index.php/Decimal_degrees).

- Only provide if **DECIMAL_LONGITUDE** is set to "NOT_COLLECTED"

AJ. **HABITAT**: (**ENA_submission**) Comments about the location, habitat or substrate, e.g. "*damp mossy ground in moderate shade*". We recommend using terms from the ENVO ontology. If the specimen is from a zoo or botanic garden, you can add its original habitat to "OTHER_INFORMATION". Here, only capture its habitat at the time of collection (e.g. "reptile cage at Zoo XYZ"). If substrate is living and there is a chance that it is included in the sample, add this to the SYMBIONT category, differentiating between the two reporting guidelines depending on the availability of a species-level identification and taxon ID for the substrate. Free text.

AK. **DEPTH**: (**ENA_submission**) Depth below water body surface or earth surface in sediment or soil, supplied in meters. This is not the absolute depth of the water body. Do not supply the unit, e.g., use 200 for 200 m below sea level, 100-200 for 100-200 m range below sea level, etc. Leave this field **blank** if the depth was not recorded or it is not an applicable field. Provide maximum possible precision.

AL. **ELEVATION:** (**ENA_submission**) Altitude above sea level, supplied in meters. Do not supply the unit, e.g., use 200 for 200 m above sea level, 100- 200 for 100-200 m range above sea level, etc. Please supply elevation of water surface for inland water bodies. Leave this field **blank** if the elevation was not recorded or it is not an applicable field. Provide maximum possible precision.

AM. **ORIGINAL_COLLECTION_DATE**: (**ENA_submission**) If the specimen is from a zoo, botanic garden, culture collection and has a known date of collection **from a known origin elsewhere** (e.g., the wild), please record the date here in as much detail as

possible, with year, month and day specified (**YYYY-MM-DD**). This information is important for regulatory compliance checks. Leave this field **blank** if it is not applicable.

AN. **ORIGINAL_GEOGRAPHIC_LOCATION**: (ENA_submission) If the specimen is from a zoo, botanic garden, culture collection and has a **known origin elsewhere**, please record the general description of the original location here. Start with the country (look up accepted country names here https://www.ebi.ac.uk/ena/browser/view/ERC000053), also include more specific locations (e.g., "Barton's Pond") ranging from least to most specific and separated by vertical pipes, e.g., "United Kingdom | East Anglia | Norfolk | Norwich | University of East Anglia | UEA Broad" when available. It is important to give the name of the site here if possible. This information is important for regulatory compliance checks. Leave this field **blank** if it is not applicable.

AO. **ORIGINAL_DECIMAL_LATITUDE**: (ENA_submission) The geographic location where the specimen or sample was originally taken in decimal degrees, between -90 and 90. This field only applies to specimens that are from a zoo, botanic garden, culture collection or have a known origin elsewhere to the current location. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees).

AP. **ORIGINAL_DECIMAL_LONGITUDE**: (ENA_submission) The geographic location where the specimen or sample was originally taken in decimal degrees, between -180 and 180. This field only applies to specimens that are from a zoo, botanic garden, culture collection or have a known origin elsewhere to the current location. We advise that locations are specified to a minimum of 3 decimal places (https://en.wikipedia.org/wiki/Decimal_degrees).

AQ. DESCRIPTION_OF_COLLECTION_METHOD: (ENA_submission) A detailed as possible description of the sample collection methods, e.g., "*caught with fiber net within densely wooded area, and immediately placed into the collection container*".

AR. DIFFICULT_OR_HIGH_PRIORITY_SAMPLE: Drop-down menu to flag species/samples that are difficult to collect (rare/rare in target area), difficult to be integrated in genome generation process (e.g., hard to get good quality HMW DNA) or high priority to push through sequencing as agreed upon with the ERGA consortium for any reason. Field with controlled vocabulary.

AS. IDENTIFIED_BY: (ENA_submission) Enter the name of the person or people who identified the sample to species level. Use ALL CAPs, and separate names with | (vertical pipe symbol), e.g., "CAROLUS LINNAEUS | JEAN-BAPTISTE LAMARCK".

- We note that storage of names with affiliations in a database brings the ERGA system under the aegis of the GDPR regulations, and we must ask sample coordinator, GALs, and collaborators to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of record). The sample coordinator is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO. Free text.

AT. IDENTIFIER_AFFILIATION: (ENA_submission) Affiliation that is responsible for the

collected specimen. This is typically the affiliation of the person(s) specified in the IDENTIFIED_BY field. If multiple people are specified in IDENTIFIED_BY, ensure that their institutional affiliations are separated by a vertical pipe symbol. Position in the list of affiliations should match the person in the same position in the list of names (e.g., “Person A | Person X | Person C” will have their affiliations as: “Institute A | Institute X | Institute C”. If multiple people are listed but all from the same affiliation, no need to repeat the affiliation. Free text.

AU. **IDENTIFIED_HOW**: Indicate what method(s) were used to identify the specimen to the nominal species (e.g., morphology, ITS barcoding). Include reference to an authoritative key if possible. If the identification is by a taxon expert, note that here and ensure the name of that person is in the IDENTIFIED_BY column. Free text.

AV. **SPECIMEN_IDENTITY_RISK**: **Y/N** field to indicate if there is any risk that the SPECIMEN_ID provided does not reflect the species names it has been submitted under (e.g., species that is part of a species complex; group where it can be difficult to be certain of species identity and/or species boundaries). Please make every effort to ensure this field is “N” if possible (e.g., consulting with taxonomic experts, using DNA barcoding to confirm species identity).

AW. **MIXED_SAMPLE_RISK**: **Y/N** field to indicate if there is any risk that the SPECIMEN_ID provided does not reflect a single genetic entity of the target species. Please make every effort to ensure this field is “N” if possible (e.g., by taking single strands of clumpy organisms or parts of the host that are most likely to reflect a single genetic entity).

AX. **PRESERVED_BY**: Name of person that carried out the preservation, supplied in CAPITALS. Multiple preserver names should be separated by a | character. Free text.

- We note that storage of names with affiliations in a database brings the ERGA system under the aegis of the GDPR regulations, and we must ask sample coordinators, GALs and collaborators to agree to their data being stored in COPO and to those data being propagated to secondary databases (including ENA and the final collections of records). The sample coordinator is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

AY. **PRESERVER_AFFILIATION**: Supply the affiliation that is responsible for the collected specimen. This is typically the affiliation of the person(s) specified in the PRESERVED_BY field. If multiple people are specified in PRESERVED_BY, ensure that their institutional affiliations are also separated by a vertical pipe symbol. Position in the list of affiliations should match the person in the same position in the list of names (e.g., Person A | Person X | Person C will have their affiliations as: (Institute A | Institute X | Institute C). If multiple people are listed but all from the same affiliation, there is no need to repeat the affiliation. Free text.

AZ. **PRESERVATION_APPROACH**: Free text field specifying e.g., snap frozen, dry ice, ethanol/dry ice slurry, in RNALater, lyophilised, air dried, etc.

BA. **PRESERVATIVE_SOLUTION**: Free text field specifying the suspension liquid used to preserve the sample, e.g., RNALater, RLT Buffer, DESS. Record volume,

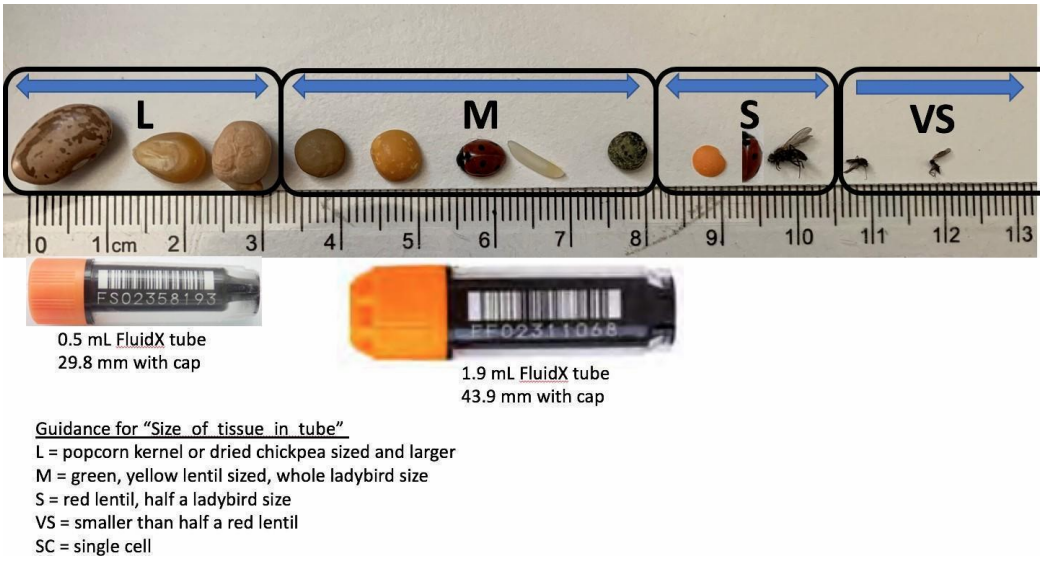
concentration, and type of liquid used here. If no preservative was used, leave **blank**.

BB. TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION: Some organisms may be held living in collection for some time for starvation or other factors. Specify hours without units (e.g., 0.5 for half an hour, 3 for 3 hours).

BC. DATE_OF_PRESERVATION: Date on which the species was preserved in **YYYY-MM-DD** format.

BD. SIZE_OF_TISSUE_IN_TUBE: Select from the drop-down menu how large the sample is in the tube. We aim most often for one lentil-sized piece per tube but sometimes adding more or less tissue than this will be necessary. Field with controlled vocabulary. **Discuss requirements with your GAL, avoid overfilling of tubes and consult taxon-specific SOPs**. Note approximate size of the piece or pellet, using the following shorthand:

- "VS" for very small (< 2 mm)
- "S" for small (~red lentil sized, (3-4mm))
- "M" for medium (~yellow lentil/ladybird sized/5mm)
- "L" for large (>5mm, chickpea/bean sized)
- If the specimen is a single cell, use "SINGLE_CELL"
- If the sample has been shipped as extracted DNA please enter "NOT_APPLICABLE".



BE. TISSUE_REMOVED_FOR_BARCODING: Select from drop-down menu "**Y**" (yes) or "**N**" (no). Instructions for appropriate Molecular Barcoding SOPs have to be arranged by the sample coordinator with the Barcoding partner, noting that barcoding may require materials in specific tube or plate types.

BF. TUBE_OR_WELL_ID_FOR_BARCODING: This is either the well number on a plate OR the barcode/unique identifier on the tube containing the tissue sample if shipped to the same GAL.

BG. TISSUE_FOR_BARCODING: Please select from the drop-down menu what part of the organism was dissected for DNA barcoding (e.g. leg, soft-body tissue etc.). This list is a

repeat of the attributes available for “ORGANISM_PART” with one addition of “DNA_EXTRACT”. Field with controlled vocabulary.

BH. **BARCODE_PLATE_PRESERVATIVE**: Record volume, concentration, and type of preservative/method of preservation used here. Free text.

BI. **BARCODING_STATUS**: Drop-down menu to indicate the status of DNA barcoding at the point of manifest submission. Options are 1) DNA barcoding completed, 2) DNA barcoding to be performed by GAL, 3) DNA barcode exempt, or 4) DNA barcoding failed. Both Option 3 (indirectly) and Option 4 (directly) refer to DNA barcoding sequencing failures. “DNA barcode exempt” is used for taxonomic groups which are known to repeatedly fail for DNA barcode sequencing, or for which barcoding as of yet is not possible. “DNA barcoding failed” means that DNA barcoding was attempted but no barcode was produced. Samples which lack DNA barcodes for either of these reasons will only proceed for genome sequencing if the field SPECIMEN_IDENTITY_RISK has the entry “N”. Field with controlled vocabulary.

BJ. **TISSUE_REMOVED_FOR_BIOBANKING**: ERGA aims to support reference genomes with frozen, biobanked material for future access. In some cases, frozen, biobanked material will need to be made from a specimen that is different from the one being submitted for sequencing (see **BK** and **BL**). If any such material is available to you for your species in question, select “**Y**” from the drop-down menu, else select “**N**”. Instructions for appropriate Biobanking SOPs have to be arranged by the sample coordinator with the Biobanking partner, noting that biobanking may require materials in specific tube or plate types.

BK. **TISSUE_VOUCHER_ID_FOR_BIOBANKING**: (**ENA_submission**) Accession number of frozen, biobanked material from the sequenced specimen. This ID should be prefixed by the name of the institution (institution unique name), followed by the collection code (if available) and the voucher id (**institution_unique_name:collection_code:voucher_id** or **institution_unique_name:voucher_Id**). It refers to a frozen, physical voucher of the specimen that is accessioned and curated into a collection accessible over GGBN (https://www.ggbn.org/ggbn_portal/) or the collection’s webportal. Registered institution and collection codes can also be looked up using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/). If not available to you upon manifest validation but **TISSUE_REMOVED_FOR_BIOBANKING** is “**Y”** you need to use “**NOT_PROVIDED”** as a placeholder, allowing for update at a later time**.** Where there are multiple biobanked vouchers to cite for a given specimen, separate the different Voucher IDs with a “|” pipe symbol. If biobanking material is from a proxy (see below BL), select in **TISSUE_REMOVED_FOR_BIOBANKING** “**Y”** and select here “**NOT_APPLICABLE”.**

BL. **PROXY_TISSUE_VOUCHER_ID_FOR_BIOBANKING:** (**ENA_submission**) In some cases, frozen, biobanked material will need to be made from a specimen that is different than the one being submitted for sequencing (e.g., a midge is too small to provide both a voucher for biobanking and a specimen for sequencing, so another midge from the same swarm may provide a para-genomotype voucher for biobanking). When this is the case, the Proxy Tissue voucher ID for Biobanking should be noted

here. This ID should be prefixed by the name of the institution (institution unique name), followed by the collection code (if available) and the voucher id (**institution_unique_name:collection_code:voucher_id** or **institution_unique_name:voucher_Id**). It refers to a frozen, physical voucher of the specimen that is accessioned and curated into a collection accessible over GGBN (https://www.ggbn.org/ggbn_portal/) or the collection's webportal. Registered institution and collection codes can also be looked up using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/). Where there are multiple biobanked vouchers to cite for a given specimen, separate the different Voucher IDs with a "|" pipe symbol.

BM. TISSUE_FOR_BIOBANKING: Select from the drop-down menu what part of the organism was dissected for biobanking (e.g. leg, soft-body tissue etc.). This list is a repeat of the attributes available for "ORGANISM_PART". Field with controlled vocabulary.

BN. BIOBANKED_TISSUE_PRESERVATIVE: Record volume, concentration, and type of preservative, describe method of preservation used. Free text.

BO. DNA_REMOVED_FOR_BIOBANKING: Select from drop-down menu "**Y**" (yes) or "**N**" (no).

BP. DNA_VOUCHER_ID_FOR_BIOBANKING: (ENA_submission) Accession number of DNA biobanked from the sequenced specimen. This ID should be prefixed by the name of the institution (institution unique name), followed by the collection code (if available) and the material id (**institution_unique_name:collection_code:voucher_id** or **institution_unique_name:voucher_Id**). It refers to a frozen sample of DNA of the specimen that is accessioned and curated into a collection accessible over GGBN (https://www.ggbn.org/ggbn_portal/) or the biobank's webportal. Registered institution and collection codes can also be looked up using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/). If not available to you upon manifest validation but **DNA_REMOVED_FOR_BIOBANKING** is "**Y"** you need to use "**NOT_PROVIDED"** as a placeholder, allowing for update at a later time. Where there are multiple vouchers to cite for a given specimen, separate the different Voucher IDs with a "|" pipe symbol.

BQ. VOUCHER_ID: (ENA_submission) Accession number of voucher material from the sequenced specimen. The ID should have the following structure: name of the institution (institution unique name) followed by the collection code (if available) and the material id (**institution_unique_name:collection_code:voucher_id** or **institution_unique_name:voucher_Id**)**.** The **Institution unique name** identifies the institution that holds the voucher. The **Collection Code** identifies the collection within the institution. Registered institution and collection codes can also be looked up using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/) or on NCBI FTP (Instittution - https://ftp.ncbi.nih.gov/pub/taxonomy/Cowner_dump.txt and Collection - https://ftp.ncbi.nih.gov/pub/taxonomy/Ccode_dump.tx). The **Voucher ID** is the catalogue number within the collection (e.g., often the physical barcode attached to the specimen or database key for that specimen). Where there are multiple vouchers to cite for a given specimen, separate the different Voucher IDs with a "|" pipe symbol. This field can be updated in COPO at a later date if accession numbers are not

available at the time of sample preparation. In such cases please use "**NOT_PROVIDED"** as a placeholder, allowing for update at a later time.

BR. **PROXY_VOUCHER_ID:** (ENA_submission) In some cases, voucher material will need to be made from a specimen that is different than the one being submitted for sequencing (e.g., a midge is too small to provide both a voucher and a specimen for sequencing, so another midge from the same swarm may provide a para-genomotype voucher). When this is the case, the Proxy Voucher ID should be noted here. The ID should have the following structure: name of the institution (institution unique name) followed by the collection code (if available) and the material id (**institution_unique_name:collection_code:voucher_id** or **institution_unique_name:voucher_Id**). The **Institution Code** identifies the institution that holds the voucher. The **Collection Code** identifies the collection within the institution. Registered institution and collection codes can be looked up using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/) or on NCBI FTP (Instittution - https://ftp.ncbi.nih.gov/pub/taxonomy/Cowner_dump.txt and Collection - https://ftp.ncbi.nih.gov/pub/taxonomy/Ccode_dump.tx). The **PROXY VOUCHER ID** is the catalog number within the collection (e.g. often the physical barcode attached to the specimen or database key for that specimen). Where there are multiple proxy vouchers to cite for the specimen, separate the different Voucher IDs with a "|" symbol. This field can be updated in COPO at a later date if accession numbers are not available at the time of sample preparation. In such cases please use "**NOT_PROVIDED"** as a placeholder, allowing for updates at a later time.

BS. **VOUCHER_LINK:** (ENA_submission)This should contain an actionable link, HTTPS(S) URI, to the specimen that the institution is committed to maintaining for the foreseeable future. The best practice is to follow a standard approach such as adopted by CETAF (https://cetaf.org/resources/best-practices/cetaf-stable-identifiers-csi-2/) but DOI or, handles quoted in their HTTPS form would also be suitable if available. Where there are multiple vouchers for a given specimen, separate the different VOUCHER_LINKs with a "|" symbol.

BT. **PROXY_VOUCHER_LINK:** (ENA_submission)This should contain an actionable link, HTTPS(S) URI, to the specimen that the institution is committed to maintaining for the foreseeable future. The best practice is to follow a standard approach such as adopted by CETAF (https://cetaf.org/resources/best-practices/cetaf-stable-identifiers-csi-2/) but DOI or, handles quoted in their HTTPS form would also be suitable if available. Where there are multiple proxy vouchers for a given specimen, separate the different PROXY_VOUCHER_LINKs with a "|" symbol.

BU. **VOUCHER_INSTITUTION:** (ENA_submission) This should contain an actionable link, HTTP(S) URI, to the record for the voucher institution in a global registry. It is recommended to link to the ROR record for the institution (e.g. https://ror.org/0349vqz63) or the Wikidata record if a ROR isn't available (e.g. https://www.wikidata.org/wiki/Q1807521). This should NOT be a link to the institution's own website. It serves as a backup if the Voucher ID or Voucher Link fields can't be interpreted. It also guarantees a machine readable version of the voucher's location.

BV. REGULATORY_COMPLIANCE: Please select from the drop-down menu "**Y"** (yes), "**NOT_APPLICABLE"** or "**N"** (not known). Note that ERGA will not be able to process further any samples where N is entered.

- Enter "**Y**" if you have affirmed that the necessary regulatory compliance documents have been obtained by the sample coordinator and are available to the sample coordinator and all involved partners including the GAL. These documents need to cover all regulatory compliance including sampling, vouchering, sample transfers, sequencing, and sequence deposition. These may include landowner permission, restricted area (SSSI, Nature Reserve, etc.) permission, BAP, CITES or other endangered species permission, ethical and Home Office Licencing for sampling for specified animals (vertebrates, cephalopods), phytosanitary permissions, veterinary pathogen sampling permissions, etc. These all fall under the SOP categories "**SAMPLING_PERMITS_REQUIRED"** and "**SAMPLING_PERMITS_DEF**"
- If you have determined that no regulatory permissions or documents are required enter "**NOT_APPLICABLE"**.

This is an important "per species" check that ensures that permissions were granted to collect and transfer the specimen for this research purpose. The sample provider should ensure this documentation is obtained, and that copies of the relevant paperwork are shared with the sequencing institution where necessary and as stipulated, for example, by regulations/approvals or licensing authorities.

BW. ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_RIGHTS_APPLICABLE: Mandatory information upon if indigenous rights are applicable to the sample/the species the sample was derived from, select "**Y**" (yes) or "**N**" (no) from drop-down menu. Indigenous rights in this SOP mean Associated Traditional Knowledge and Biocultural Rights DSI. If **"Y"** please register through the Local Context Hub (https://localcontexts.org/) to get a ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_PROJECT_ID.

BX. **INDIGENOUS_RIGHTS_DEF:** Please state which rights (e.g., Associated Traditional Knowledge, Biocultural Rights, DSI) are applicable if **ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_RIGHTS_APPLICABLE** is set to "Y" (yes). Free text.

BY. **ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_PROJECT_ID**: project ID provided by the Local Context Hub (https://localcontexts.org/) upon notice registration. If information is not available, please leave this field blank.

BZ. **ASSOCIATED_TRADITIONAL_KNOWLEDGE_CONTACT**: Provide reference for contact, could be linked to an ORCID ID. Free text.

CA. ETHICS_PERMITS_REQUIRED: Mandatory information upon if an ethics permit is needed to sample/sequence/voucher/biobank the sample/the species the sample was derived from, select "**Y**" (yes) or "**N**" (no) from drop-down menu.

CB. ETHICS_PERMITS_DEF: Free text explaining permits, permit issuing entity and permit number. If ETHICS_PERMITS_REQUIRED is "N", enter "**NOT_APPLICABLE"**. In COPO, an upload field will be triggered if ETHICS_PERMITS_REQUIRED is set to "Y".

All explained permits need to be uploaded in a single (concatenated) pdf named SPECIMEN_ID_ETHICS_PERMITS.pdf

CC. **ETHICS_PERMITS_FILENAME**: Free text indicating the exact file name, if applicable. If ETHICS_PERMITS_REQUIRED is set to "N", enter **NOT_APPLICABLE**.

CD. **SAMPLING_PERMITS_REQUIRED**: Mandatory information upon if sampling permits (according to international and national legislation) are needed to sample/sequence/voucher/biobank the sample/the species the sample was derived from, select "**Y**" (yes) or "**N**" (no) from drop-down menu.

CE. **SAMPLING_PERMITS_DEF**: Free text explaining permits, permit issuing entity and permit number. Separate information on multiple permits by vertical pipe and use the same order as in the concatenated uploaded pdf. If SAMPLING_PERMITS_REQUIRED is set to "N", enter **NOT_APPLICABLE**. In COPO, an upload field will be triggered if SAMPLING_PERMITS_REQUIRED is set to "Y" and all explained permits need to be uploaded in a single (concatenated) pdf named SPECIMEN_ID_SAMPLING_PERMITS.pdf.

CF. **SAMPLING_PERMITS_FILENAME**: Free text indicating the exact file name, if applicable. If SAMPLING_PERMITS_REQUIRED is set to "N", enter **NOT_APPLICABLE**.

CG. **NAGOYA_PERMITS_REQUIRED**: Mandatory information upon if a permit in compliance with the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity* is needed for the sample in question/the species the sample was derived from, Select "**Y**" (yes) or "**N**" (no) from drop-down menu.

CH. **NAGOYA_PERMITS_DEF**: Free text explaining permits, permit issuing entity and permit number. If NAGOYA_PERMITS_REQUIRED is set to "N", enter **NOT_APPLICABLE**. In COPO, an upload field will be triggered if NAGOYA_PERMITS_REQUIRED is set to "Y" and all explained permits need to be uploaded in a single (concatenated) pdf named SPECIMEN_ID_NAGOYA_PERMITS.pdf.

CI. **NAGOYA_PERMITS_FILENAME:** Free text indicating the exact file name, if applicable. If column NAGOYA_PERMITS_REQUIRED is set to "N", enter **NOT_APPLICABLE**.

CJ. **HAZARD_GROUP**: EU biological hazard groups 1, 2, 3, and 4 according to Directive 2000/54/EC on the protection of workers from risks related to exposure to biological agents at work with 1: biological agent unlikely to cause human disease; 2: biological agent can cause human disease and might be a hazard to workers, unlikely to spread to community, effective prophylaxis or treatment available; 3: biological agent can cause severe human disease and present a serious hazard to workers; it may present a risk of spreading to the community, usually effective prophylaxis or treatment available; 4: biological agent that causes severe human disease and is a serious hazard to workers; it may present a high risk of spreading to the community; no effective

prophylaxis or treatment available. Please note that any specimens above Hazard Group 1 must be discussed prior to shipping samples. Select from drop-down menu. Field with controlled vocabulary.

CK. **PRIMARY_BIOGENOME_PROJECT:** Indicate if your genome is part of ERGA-Pilot, ERGA-BGE, ERGA-associated or ERGA-Community genome. Field with controlled vocabulary.

CL. **ASSOCIATED_PROJECT_ACCESSIONS**: (**ENA_submission**) List of additional associated Biogenome Projects (e.g. DtOL, VGP). Multiple projects can be entered by separating names/BioProject IDs with a "|" (vertical pipe) character. Field with controlled vocabulary.

CM. **OTHER_INFORMATION**: Free text field for further relevant information to sample managers not captured by the other fields. If there is nothing else to add here, this field should be left **blank**.

Table 2 Document History

Version	*Date*	*Changes*	*Contributors*
1.0	2021-09-17	first version	Mara Lawniczak, Jeena Rajan, Robert P Davey, Seanna McTaggart, Mark Blaxter, Alice Minotto, Felix Shaw, DToL SamplesWG, ERGA SSP committee, ERGA ELSI committee, Ann McCartney, Jennifer A Leonard, Olga Vinnere Petterson, Astrid Böhne
2.4		Change of specimen ID standard and description; extension of GALs, change of pop up window requesting ERGA membership and adherence to code of best practice for sampling Addition of ERGA sample manifest roadmap PURPOSE_OF_SPECIMEN: Addition of R&D as an option in the drop-down menu. *C* SYMBIONT: Update of field description. *S* 'ROOTs' added to column ORGANISM_PART. *R* set COLLECTOR_ORCID_ID and DESCRIPTION_OF_COLLECTION_METHOD as an ENA submission field, matches in ENA to "sample collection device/sample collection method". *Z and AN* ORIGINAL_DECIMAL_LATITUDE; ORIGINAL_DECIMAL_LONGITUDE: Insertion of fields to accommodate original collection lat/long coordinates. *AL and AM* Splitting one field into two. SPECIMEN_IDENTITY_RISK previously included misidentification of specimens, and the possibility that multiple genetic individuals are in the tube. These are now two separate fields (one on misidentification retaining SPECIMEN_ID_RISK, one new field dealing with multiple individuals MIXED_SAMPLE_RISK), each with a Yes/No option. *AS and AT* Addition of a field to indicate whether the barcoding is completed (or if it is a sample exempt from barcoding, or the barcoding failed). BARCODING_STATUS. *BF*	Alice Minotto, Oliver Hawlitschek, Martin Husemann, Luísa Marins, Torsten Struck, Carmela Gissi, Isabelle Florent, Katja Reichel, Seanna McTaggart, Ann McCartney, Jennifer A Leonard, Olga Vinnere Petterson, Astrid Böhne, Joana Pauperio, Josephine Burgin

		TISSUE_VOUCHER_ID_FOR_BIOBANKING and DNA_VOUCHER_ID_FOR_BIOBANKING: map to ENA biomaterial. *BH and BK* Addition of a field to indicate when a proxy voucher has been used. PROXY_VOUCHER_ID. *BM* Addition of three fields VOUCHER_LINK, PROXY_VOUCHER_LINK and VOUCHER_INSTITUTION to provide a link to the actual voucher and voucher institution. *BN, BO, BP* ETHICS_PERMITS_REQUIRED, SAMPLING_PERMITS_REQUIRED, NAGOYA_PERMITS_REQUIRED: change of field names, from “mandatory” to “required”. BV, *BX, BZ* TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_ID: Updated field name and definition. Previously this field required the selection of a Local Contexts Label from a drop-down menu. This is now a field where the Project-ID should be provided. *BT* HAZARD_GROUP: Updated field description and addition of HG4 in the drop-down menu. *CC*	
2.4.2	31.03.2023	Extension of the description section *Vouchers of Specimen or Sample* Implementation of brokering of photographs of specimen/samples to EBI Bioimages Columns VOUCHER_LINK, PROXY_VOUCHER_LINK and VOUCHER_INSTITUTION and TIME_OF_COLLECTION marked as ENA submissions Update of field description for DECIMAL_LATITUDE, DECIMAL_LONGITUDE Addition of fourfields to allow definition of transect geographic location \LATITUDE_START, LONGITUDE_START, LATITUDE_END, LONGITUDE_END Removal of field GRID_REFERENCE since this information is captured in fields DECIMAL_LATITUDE and DECIMAL_LONGITUDE. Extended vocabulary for field BARCODING_STATUS to include pending and scheduled barcoding with “DNA barcoding to be performed by GAL” Added field PROXY_TISSUE_VOUCHER_ID_FOR_BIOBANKING	Luísa Marins, Torsten Struck, Katja Reichel, Seanna McTaggart, Jennifer A Leonard, Olga Vinnere Petterson, Astrid Böhne, Joana Pauperio, Josephine Burgin, Rita Monteiro, Felix Shaw

		matching field PROXY_VOUCHER_ID Added field BIOBANKED_TISSUE_PRESERVATIVE matching field BARCODE_PLATE_PRESERVATIVE Updated contact email for ENA to ena-bge@ebi.ac.uk Added fields PRIMARY_BIOGENOME_PROJECT (with predefined list of ERGA projects), ASSOCIATED_BIOGENOME_PROJECTS (free text to link to associated genome project) Updated GAL contact information	
2.4.3	19.06.2023	Added three fields: ETHICS_PERMITS_FILENAME, SAMPLING_PERMITS_FILENAME, NAGOYA_PERMITS_FILENAME	
2.5	22.09.2023	General document maintenance PURPOSE_OF_SPECIMEN: addition of terms “RESEQUENCING” and NOT_PROVIDED, withdrawal of SHORT_READ_SEQUENCING term DECIMAL_LATITUTE and DECIMAL_LONGITUE: addition of term NOT_COLLECTED ORIGINAL_DECIMAL_LONGITUDE corrected possible values to range from -180 to 180. Extended explanation to SAMPLING_PERMITS_DEF and ASSOCIATED_PROJECT_ACCESSIONS, corrected naming in SOP for the latter Update to TAXON_ID text Updated GAL contact information	
2.5.1	11.03.2025	Updates to the authorship section Update to “Vouchers of Specimen or Sample” section Update GAL dropdown menu with “NOT_PROVIDED” Update to TISSUE_VOUCHER_ID_FOR_BIOBANKING, PROXY_TISSUE_VOUCHER_ID_FOR_BIOBANKING, DNA_VOUCHER_ID_FOR_BIOBANKING, VOUCHER_ID and PROXY_VOUCHER_ID text	

		Update TAXON_ID and SYMBIONT text and implemented subspecies acceptance. Update ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_PROJECT _ID text. Added “ERGA-Community” to PRIMARY_BIOGENOME_PROJECT list.	