Recording Sample Metadata for the European Reference Genome Atlas Project*

Sample Manifest Standard Operating Procedure

Version: 2.4

Published Date: June 2022

Authors: Jennifer A. Leonard, Olga Vinnere Petterson, Seanna McTaggert, Ann McCartney, Alice Minotto, Luísa Marins, Torsten Struck, Martin Husemann, Carmela Gissi, Isabelle Florent, Katja Reichel, Seanna McTaggert, Astrid Böhne, and the ERGA 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. Also contained is generic guidance on how to process specimens. Help for taxon-specific SOPs is available from each taxonomic working group (see contacts listed below) to provide guidance on sample processing and regulatory compliance. Guidance on sample submission to sequencing partners should be sought from the sequencing facility (see list of sequencing partners below). Guidance on submission of vouchering and biobanking material should be sought from the respective collection facility (see list of partners below). Suggestions to this manifest should be sent to the sampling committee SSP samples@erga-biodiversity.eu. The submission of a completed manifest is mandatory for every ERGA genome.

***This sampling manifest builds on the work of the Darwin Tree of Life (DToL) sampling committee, in particular of Mara Lawniczak and Robert Davey. The ERGA consortium thanks DToL for allowing ERGA to adopt their sampling metadata manifest to ERGA's needs and for document sharing. All changes to this document only apply to ERGA not to DToL.**

Preamble: To be able to register a sample/specimen and its metadata for ERGA, the submitting person (most often identical to the species ambassador) 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 from samples and to embed these sequences into best practices in scientific research and the landscape of biodiversity science. To do this we must adhere to correct, legal and ethical physical handling of the specimens, and correct collation of rich metadata describing the specimens. This SOP contains specific instructions for filling in the metadata manifest. ERGA will not access and process 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 an ethical code of conduct. The legal responsibility for acquiring samples remains with the species' ambassador(s). By submitting the sampling manifest and providing information on compliance with sampling permits, the ambassador 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 species ambassador. Where necessary and applicable, material transfer agreements can be issued and signed by sequencing and collection facilities and the species ambassador; the oversight of this process lies with the species' ambassador(s).

Additional SOPs: Additional related SOPs are forthcoming on how to prepare samples for different taxonomic groups, which helps to assure delivery of high-quality samples that are more likely to be transformed into high quality genomes. As of now, help for actual sampling and vouchering can be requested from taxon chaperons of the ERGA SSP committee taxonomic focus groups as listed below.

Future plans for this SOP: This SOP will be reviewed on a quarterly basis by the manifest task force of the SSP committee to incorporate feedback from the community. Metadata are currently collected manually by the species ambassador using a defined spreadsheet, referred to as the ERGA SAMPLE MANIFEST V2.4. This document will allow 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. Finally, 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 under development.

Raising issues: 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 Samples Committee at **samples@erga-biodiversity.eu** or reach out on Keybase, Team erga.listserv, channel #Committee_SSP. 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.

Table 1 Document History

Major 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 of one field into two. SPECIMEN_IDENTITY_RISK previously included misidentification of specimens,	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

		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* 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*	

Completing the Sample Manifest: Overview

Scope of this document

Specific guidance on preparing samples is not covered by this SOP. Please contact the chaperon for the specific taxonomic group you are working on.

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 also allows the laboratory team to resample the same individual specimen (thus 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, a single specimen split across ten tubes would result in each of those ten tubes having the same ERGA SPECIMEN_ID. This unique ERGA SPECIMEN_ID has three critical functions: identifying the species ambassador that holds responsibility for the specimen, tracking an individual sample's status and declaring the genetic uniqueness of the specimen.

Each ERGA specimen must be linked to a standardized, unique ID that begins with the prefix **ERGA_** followed by **SPECIES AMBASSADOR INITIALS** (up to 10 letters out of A-Z, if this is not possible please reach out to **samples@erga-biodiversity.eu**), followed by **underscore** followed by the **last four digits of the SPECIES AMBASSADOR's ORCID ID**, **underscore**, and **running numbers** (e.g. if you as ambassador register more than one sample make sure to use 01, 02…). SPECIMEN_IDs must be unique to an individual (e.g., ERGA_XY_1234_01 cannot be used again after it has been assigned to a specimen). SPECIMEN_IDs must follow the format described above.
Furthermore, Genome Acquisition Labs GALs (the partners or companies performing the actual genome sequencing) will attribute identifiers to each SPECIMEN. Examples of the formats of the ERGA partners and companies are provided in the table below for their internal sample tracking. Such identifiers need to be provided from the Sequencing partners to the ambassadors.

Table 2: Genome Acquisition Labs GAL Specimen Codes

GAL	Code Model	Number of digits	Contact Person Email address*
SANGER INSTITUTE (SAN)	SAN0000000	7	Nancy Holroyd, neh@sanger.ac.uk
EARLHAM INSTITUTE (EI)	EI_00000	5	Seanna McTaggart, seanna.mctaggart@earlham.ac.uk
CENTRO NACIONAL DE ANÁLISIS GENÓMICO (CNAG)			Project Manager, projectmanager@cnag.crg.eu
SCILIFELAB (SCI)			Olga Vinnere Pettersson, olga.pettersson@igp.uu.se
WEST GERMAN GENOME CENTRE (WGGC)			Peter Nürnberg, nuernberg@uni-koeln.de Antonella Succurro, a.succurro@uni-bonn.de
NGS COMPETENCE CENTER TÜBINGEN (NCCT)			Nicolas Casadei, Nicolas.Casadei@med.uni-tuebingen.de
DRESDEN-CONCEPT (DRC)			Sylke Winkler, winkler@mpi-cbg.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
GIGA-GENOMICS CORE FACILITY UNIVERSITY OF LIEGE (GIGA)			Wouter Coppieters, wouter.coppieters@uliege.be
SVARDAL LAB, ANTWERP (SVL)			Hannes Svardal, hannes.svardal@uantwerpen.be
LEIBNIZ INSTITUTE FOR THE ANALYSIS OF BIODIVERSITY CHANGE, MUSEUM KOENIG, BONN (LIB)	NA	NA	Lars Podsiadlowski, l.podsiadlowski@leibniz-lib.de
INDUSTRY PARTNER (IP)	NA	NA	as provided by sequencing company

* As of March 2022, reach out to samples@erga-biodiversity.eu if your GAL is not in the list, extension request can be integrated with next version release

Other "_ID"s

A sample can represent a set of specimens as well as multiple parts of the same specimen, and so 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. The same is true of the GAL_SAMPLE_ID. For example, 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 strongly urge sample providers to complete metadata collection and upload before commencing sequence 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 species ambassador/genome team.

Manifest Validation Process

Choose whether you prefer to use the Sample Manifest from the google spreadsheet or another option (e.g., epicollect, ARCGIS). We recommend that you retain a copy in Excel (XLS/XLSX) or Google spreadsheet form so as not to lose the data validation given the likelihood that further edits will be required, even after upload to COPO.

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). 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.

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. If mismatches are found, this will require the submitter to examine the mismatches and determine the nature of the problem. Please read the guidance on TAXON_ID below carefully as you should be able to ensure that each TAXON_ID precisely and accurately matches a species name in advance of

submitting your manifest. There are too many possibilities to enumerate them all here, but three of 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. More information on how to fix these issues is below in the discussion of the TAXON_ID field.

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. If any other 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 mal formatted or missing information.

Once this process is complete and every sample has a TAXON_ID together with complete metadata, the manifest is considered to be “validated”.
When a validated manifest is submitted, each sample will be allocated a “ToLID” that reflects both the species and the SPECIMEN_ID (i.e., the genetic identity of the sample). The ToLID is created by the Sanger institute and tracks the submitted samples through the sequencing process and acts as assembly names when the data are submitted to the INSDC databases at the end. 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.

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 (for the ERGA pilot: pilot@erga-biodiversity.eu, in the future to samplemanifest@erga-biodiversity.eu). Any updates or changes to any fields for uploaded specimens should be sent as an email request to EI.COPO@earlham.ac.uk specifying the BioSamples accession, the field to update and the new value. For taxonomic changes, only the BioSamples accession and the new SCIENTIFIC_NAME is needed to update the taxonomy of a sample/specimen. COPO will produce a pipeline to update metadata for uploaded samples (see visual COPO documentation for more information on manifest submission and process updates).

Vouchers of Specimen or Sample

Whenever possible, a submitted specimen must be vouchered in a public scientific collection dedicated to permanent storage and with an accessible voucher catalog. Ideally, in addition to a physical voucher, tissue or cells and/or DNA should be deposited in a public Biobank/ frozen repository, ideally a member of GGBN. If this is not possible, such as for small species like most invertebrates or fungi, 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 (e.g., the deep sea, endangered species), we ask that digital images are recorded prior to destructive sampling and submitted in lieu of physical samples. Taxon specific vouchering information should be found in taxon specific sampling SOPs. When possible, photographs are generally appreciated (see below). 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.

Photographs of Specimen or Sample

Photography instructions will be found in taxon-specific SOPs (under development); using a scale measurement is encouraged. Every submitted specimen should ideally 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 appropriate pdfs here 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. Please upload your image to ERGA's B2Drop instance on Nextcloud in which all documents associated with the registered specimen will be stored. As of now for each SPECIMEN_ID, there will be a directory that will be communicated to you in ERGA's reserved and secured space on a B2Drop instance on Nextcloud in which all documents associated with the registered specimen will be stored. Images need to be dropped to the subdirectory called "IMAGES".

In preparation for linking images to metadata, 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. When uploading photographs, please use PNG or JPG format.

File names must exactly match the SPECIMEN_ID in order to match photographs to samples automatically.

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 ambassador (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 T and U)
Block 5: Data on the collector, collection event, and collection localities (columns W to AO)
Block 6: Information on taxonomic identification, taxonomic uncertainty and risks (columns AP to AT)
Block 7: Details of the tissue preservation event (columns AU to BA)
Block 8: Information on DNA barcoding (columns BB to BF)
Block 9: Information on Biobanking and Vouchering (columns BG to BP)
Block 10: Information on regulatory compliances, Indigenous rights, traditional knowledge and permits (columns BQ to CA)
Block 11: Additional information including a free text field to house other important sample notes (columns CB and CC)

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**	GRID_REFERENCE	**HABITAT**	DEPTH	ELEVATION
AJ	**AK**	**AL**	**AM**	**AN**	**AO**	**AP**
ORIGINAL_COLLECTION_DATE	ORIGINAL_GEOGRAPHIC_LOCATION	ORIGINAL_DECIMAL_LATITUDE	ORIGINAL_DECIMAL_LONGITUDE	**DESCRIPTION_OF_COLLECTION_METHOD**	**DIFFICULT_OR_HIGH_PRIORITY_SAMPLE**	**IDENTIFIED_BY**
AQ	**AR**	**AS**	**AT**	**AU**	**AV**	**AW**
IDENTIFIER_AFFILIATION	**IDENTIFIED_HOW**	**SPECIMEN_ID_RISK**	**MIXED_SAMPLE_RISK**	**PRESERVED_BY**	**PRESERVER_AFFILIATION**	**PRESERVATION_APPROACH**
AX	**AY**	**AZ**	**BA**	**BB**	**BC**	**BD**
PRESERVATIVE_SOLUTION	**TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION**	**DATE_OF_PRESERVATION**	**SIZE_OF_TISSUE_IN_TUBE**	**TISSUE_REMOVED_FOR_BARCODING**	**TUBE_OR_WELL_ID_FOR_BARCODING**	**TISSUE_FOR_BARCODING**
BE	**BF**	**BG**	**BH**	**BI**	**BJ**	**BK**
BARCODE_PLATE_PRESERVATIVE	**BARCODING_STATUS**	**TISSUE_REMOVED_FOR_BIOBANKING**	**TISSUE_VOUCHER_ID_FOR_BIOBANKING**	**TISSUE_FOR_BIOBANKING**	**DNA_REMOVED_FOR_BIOBANKING**	**DNA_VOUCHER_ID_FOR_BIOBANKING**
BL	**BM**	BN	BO	BP	**BQ**	**BR**
VOUCHER_ID	**PROXY_VOUCHER_ID**	VOUCHER_LINK	PROXY_VOUCHER_LINK	VOUCHER_INSTITUTION	**REGULATORY_COMPLIANCE**	**ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_RIGHTS_APPLICABLE**
BS	**BT**	**BU**	**BV**	**BW**	**BX**	**BY**
INDIGENOUS_RIGHTS_DEF	**ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_PROJECT_ID**	ASSOCIATED_TRADITIONAL_KNOWLEDGE_CONTACT	**ETHICS_PERMITS_REQUIRED**	**ETHICS_PERMITS_DEF**	**SAMPLING_PERMITS_REQUIRED**	**SAMPLING_PERMITS_DEF**
BZ	**CA**	**CB**	**CC**			
NAGOYA_PERMITS_REQUIRED	**NAGOYA_PERMITS_DEF**	**HAZARD_GROUP**	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 by raising a request to the SSP committee at **samples@erga-biodiversity.eu**. Please only do this if absolutely required (i.e., no available term is a good proxy, and the absence of the attribute likely to affect many samples).

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 green fill. The fill will go white when an entry has been made to help you identify where mandatory fields still require data.] For all mandatory fields with **turquoise bold names**, even if information is unavailable, they must be populated with the appropriate term describing why this information is missing. The acceptable missing value terms follow the INSDC recommendations and are as follows :

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 (this may be a 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 an entry describing why the information is missing because we expect that many samples will not have information for these fields (e.g., most samples will not have DEPTH information). However, if you have collected the information related to these terms, please do enter it.

Many terms will have the data released publicly as part of the ENA record. For every field for which this is true, you will find "**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", we ask that you use only the core alphanumeric characters, plus 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. Please **do not** use "special characters" (such as other punctuation and "logical" marks: "#"`;:?!@*()[]{}/\,=+", etc.).

Column by column instructions for the Metadata Entry tab.

A. **TUBE_OR_WELL_ID**: This field should record the individually attributed label of the genome team on the tube submitted for sequencing. If samples are submitted in plate format, provide the relevant well information here. If barcodes are entered, use a barcode scanner in advance of preparing samples to reduce errors – do not enter barcodes manually.

B. **SPECIMEN_ID**: (**ENA_submission**) This is a unique identifier that refers to the genetic identity of the supplied material. It is assumed that the ERGA SPECIMEN_ID refers 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 would be the same. If multiple individuals of a species are sampled (e.g., from the same population), they must be placed in multiple, 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 is possible (e.g., single gametophyte from a moss mat), and place each in a separate specimen tube with a unique ERGA SPECIMEN_ID. Each ERGA specimen must be linked to a standardized, unique ID that begins with the prefix ERGA_ followed by SPECIES AMBASSADOR INITIALS (up to 10 letters out of A-Z, if this is not possible please reach out to **samples@erga-biodiversity.eu**) _underscore and last four digits of OrcidID of SPECIES AMBASSADOR and _underscore RUNNING NUMBERS (e.g. if you as ambassador register more than one sample make sure to use 01, 02…). SPECIMEN_IDs must be unique to an individual (e.g., ERGA_XY_0123_01 cannot be used again after it has been assigned to a specimen). SPECIMEN_IDs must follow the format described above.

C. **PURPOSE_OF_SPECIMEN**: Please select appropriate from the drop-down menu.

- The majority of specimens will be for “REFERENCE_GENOME”. All samples listed for REFERENCE_GENOME sequencing are assumed to also need DNA BARCODING and RNA-SEQUENCING, and the term “REFERENCE GENOME” encompasses all three things (reference genome, barcoding, rna-seq) wherever samples allow. Please use REFERENCE GENOME for all specimens / samples of a particular species unless they should be destined for an alternative use only. This comprises all necessary sequencing methods needed to obtain a reference. Please indicate over RELATIONSHIP if a specimen is solely used for one method required to establish a reference genome (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 please use “SHORT_READ_SEQUENCING.
- If a particular tissue or specimen is intended for research and development, for example as part of an R&D diversity panel, or as part of a preservation trial, please use “R&D”. These samples may not progress to reference genome sequencing and may be used for protocol testing.
- The drop-down option for DNA_BARCODING_ONLY is reserved for those 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, or is otherwise unsuitable for HMW DNA, but the material is valuable for barcoding).

D. **SAMPLE_COORDINATOR:** (**ENA_submission**) Also known as the ERGA sample ambassador, 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”.

    - 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**) Free text field to supply the university, institution, or society that is responsible for the genome project of the sample. This is 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 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, no need to repeat the affiliation.

F. **SAMPLE_COORDINATOR_ORCID_ID:** (**ENA_submission**) Enter the 16 digits ORCID ID of the person or people who is responsible for the genome project of the sample. If multiple entries are provided, ensure that they are separated by a vertical pipe symbol.

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 Taxonomy Database. If you or a taxonomist have a disagreement with the taxonomy represented on NCBI Taxonomy Database, please raise this with the NCBI TaxonomyDB curators as described below.

H. **FAMILY**: The taxonomic Family into which the Genus is placed. This should correspond to the taxonomy as represented in the NCBI Taxonomy Database. If you or your taxonomist have a disagreement with the taxonomy represented on NCBI Taxonomy Database, please raise this with the NCBI TaxonomyDB curators as described below.

I. **GENUS**: The taxonomic Genus to which the Species belongs. This should correspond to the taxonomy as represented in the NCBI Taxonomy Database, and with the generic component of the scientific name given below. If you or your taxonomist have a disagreement with the taxonomy represented on NCBI Taxonomy Database, please raise this with the NCBI TaxonomyDB curators as described below.

J. **TAXON_ID**: (**ENA_submission**) A valid NCBI TAXON_ID to the species 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, write to ena-dtol@ebi.ac.uk to request the change in NCBI Taxonomy.

If there is another taxon database for your group, e.g., EukRef, LSIDs, please fill in the NCBI TAXON_ID, and then 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.

- 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 taxonomy), please ask for assistance by writing to ena-dtol@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 taxonomy browser), a new TAXON_ID should be available within 14 days. In the case of conflict, the sample submitter will be contacted and may be required to provide further information. Please note that the final species name on submission of the data to INSDC will be the one associated with the TAXON_ID in NCBI

Taxonomy.

- 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. You may care to 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.

- 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.

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) Please 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.

O. **COMMON_NAME**: Vernacular name. If any guidelines for vernacular names exist (e.g., birds: https://birdsoftheworld.org/bow/species; reptiles: https://ssarherps.org/wp-content/uploads/2014/07/HC_39_7thEd.pdf), 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**.

P. LIFESTAGE: (ENA_submission) The life stage of the specimen from which the sample was derived. This field has a controlled vocabulary: use the drop-down menu or look at the available terms on the second tab to complete. Please note that there

are currently curated attributes for animals, for plants/fungi/macroalgae, and for some protists.

- If these do not fit your taxa, please 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. This field has a controlled vocabulary: 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, so this field should refer solely to the sex as determined by morphological examination of the specimen or strong inference (e.g., the species is from a clade that is always hermaphroditic/monoecious).

R. **ORGANISM_PART**: (**ENA_submission**) A description of the exact tissue(s) in the tube or well. Accurate information here is important for downstream analyses on the symbiome, chromosomal diminution, RNAseq, etc. There is a tab in the Sample Manifest that defines the terms that can be used for ORGANISM_PART. This tab lists 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 - not the whole lung - is expected).

- Please 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 as long as spelling is correct. 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, please 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 also email the ERGA SSP committee at samples@erga-biodiversity.eu to request the necessary additions. We will update attributes quarterly.

- 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-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" (and 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-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:

   a. 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.

T. **RELATIONSHIP:** (**ENA_submission**) This is a 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.

V. GAL_SAMPLE_ID: (ENA_submission) This is the unique name assigned to the sample by the GAL. This might include an abbreviation for the GAL and a simple shorthand identifier. This is a free text field, but please do not use spaces or special characters, e.g., #, !, ^, *, etc. 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 if available that you stick to the format required by the GAL. Please see table on page 5 for details. In case the GAL has no established naming system we suggest to use here the abbreviated name of the GAL as listed in Table 2 (in parentheses), followed by a specimen id and a sample/material id (institution code:specimen id:material_id).

W. COLLECTOR_SAMPLE_ID: This is the unique name assigned to the sample by the COLLECTOR or COLLECTOR_AFFILIATION. This is a free text field, but please **do not use spaces or special characters**, other than hyphens and underscores ("-" and "_") i.e do not use #, !, ^, *, etc.

- In some cases, you will be splitting a single specimen across multiple tubes (see SPECIMEN_ID), and you will want to consider what kind of information you want in your unique sample names for this. For example, 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".

- 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 species ambassadors, 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 species ambassador is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

Y. COLLECTOR_AFFILIATION: (ENA_submission) Free text field to supply the university, institution, or society that is responsible for the collected specimen. This is typically the society or institution of the person(s) specified in the COLLECTED_BY field. If multiple people are specified in COLLECTED_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, no need to repeat the affiliation.

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. If more than a single entry is specified ensure that they are separated by a vertical pipe symbol.

AA. DATE_OF_COLLECTION: (ENA_submission) The date the sample was collected, with year, month, and day specified (**YYYY-MM-DD**).

BB. **TIME_OF_COLLECTION**: 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 but it is not mandatory. Leave this field **blank** if the time was not recorded.

CC. COLLECTION_LOCATION: (ENA_submission) General description of the location where the tissue/organism part was sampled for genome sequencing. This should start with the geographical origin of the sample country as defined by the country or sea in agreement with ISNDC country list (look up accepted country names here https://www.insdc.org/country.html), but also include more specific locations (e.g., “Barton’s Pond”) ranging from least to most specific and separated by | 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.

- 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 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).

- 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 **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).

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).

- 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 **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).

AF. **GRID_REFERENCE**: Information to geolocate the sample area, where the specimen or sample was taken at the time (e.g., GRID reference of the field sampling location, or “London Zoo”, “Millennium Seed Bank”, etc). Preferably, the information should be provided with a map or standardized geolocation reference, e.g. OS GRID REF: SP45998 08751. https://osmaps.ordnancesurvey.co.uk/ is useful to map lat-long to grid references. This field is optional and can be left **blank**.

AG. **HABITAT**: (**ENA_submission**) Any 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” but here, please only capture its habitat at the time of collection (e.g. “reptile cage at London Zoo”). 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.

AH. **DEPTH**: (**ENA_submission**) Depth below water body surface or earth surface in sediment or soil, supplied in metres. 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.

AI. **ELEVATION:** (ENA_submission) Altitude above sea level, supplied in metres. 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.

AJ. **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**). YYYY-MM and YYYY is acceptable where further detail is not known. This information is important for regulatory compliance checks. Leave this field **blank** if it is not applicable.

AK. **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. This should start with the country (United Kingdom, or look up other accepted country names here https://www.ebi.ac.uk/ena/browser/view/ERC000053), but 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.

AL. **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).

AM. **ORIGINAL_DECIMAL_LONGITUDE**: (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).

AN. 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”*.

AO. 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.

AP. **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 GPDR regulations, and we must ask species ambassadors, 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 species ambassador is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

AQ. **IDENTIFIER_AFFILIATION**: (**ENA_submission**) Free text field to supply the university, institution, or society that is responsible for the collected specimen. This is typically the society or institution of the person(s) specified in the IDENTIFIED_BY field. If multiple people are specified in IDENTIFIED_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, no need to repeat the affiliation.

AR. **IDENTIFIED_HOW**: Indicate what method(s) were used to identify the specimen to the nominal species (e.g., morphology, ITS barcoding). This is free text and should 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.

AS. **SPECIMEN_ID_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. For example where a species is part of a species complex or 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., by consulting with taxonomic experts and using results from DNA barcoding to confirm species identity).

AT. **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).

AU. **PRESERVED_BY**: Name of person that carried out the preservation, supplied in

CAPITALS. Multiple preserver names should be separated by a | character.

- 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 species ambassadors, 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 species' ambassador is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

AV. **PRESERVER_AFFILIATION**: Free text field to supply the university, institution, or society that is responsible for the collected specimen. This is typically the society or institution 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.

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

AX. **PRESERVATIVE_SOLUTION**: Free text field specifying the suspension liquid used to preserve the sample, e.g., RNALater, RLT Buffer, DESS. Record the volume, concentration, and type of liquid used here. If no preservative was used, this field should be left **blank**.

AY. **TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION**: some organisms may be held living in collection for a period of time for starvation or other factors. This entry should be specified in hours, but no unit, e.g., 0.5 for half an hour, 3 for 3 hours, etc.

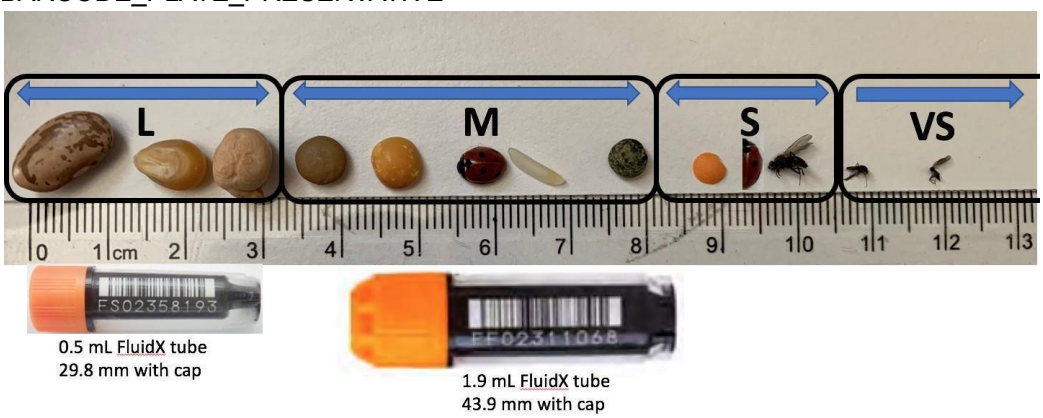
AZ. **DATE_OF_PRESERVATION**: Date on which the species was preserved. Please use **YYYY-MM-DD** format.

BA. **SIZE_OF_TISSUE_IN_TUBE**: Select from drop-down menu how large is the sample in the tube. We aim for one lentil-sized piece per tube but sometimes adding more or less tissue than this will be necessary. Please note the approximate size of the piece or pellet: use the following shorthand:

- "VS" for very small
- "S" for small (~red lentil sized)
- "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"
- Aim for single lentil sized (S or M) pieces in tubes whenever possible. If the sample is L, then wherever possible process this into multiple tubes of S or M sized pieces . See visual guidance below.
- If the sample has been shipped as extracted DNA please enter "NOT_APPLICABLE".

BARCODE_PLATE_PRESERVATIVE



Guidance for "Size_of_tissue_in_tube"
L = popcorn kernel or dried chickpea sized and larger
M = green, yellow lentil sized, whole ladybird size
S = red lentil, half a ladybird size
VS = smaller than half a red lentil
SC = single cell

BB. **TISSUE_REMOVED_FOR_BARCODING**: Select from drop-down menu "**Y**" or "**N**". Instructions for appropriate Molecular Barcoding SOPs have to be arranged by the species ambassador with the Barcoding partner, noting that barcoding may require materials in specific tube or plate types.

BC. **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.

BD. **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".

BE. **BARCODE_PLATE_PRESERVATIVE**: Record the volume, concentration, and type of preservative/method of preservation used here.

BF. **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 barcode exempt, or 3) DNA barcoding failed. Both Option 2 (indirectly) and Option 3 (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 and have been identified by the relevant taxon working group as exempt from the DNA barcoding step. "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_ID_RISK has the entry "N".

BG. **TISSUE_REMOVED_FOR_BIOBANKING**: Select from drop-down menu "**Y**" or "**N**". Instructions for appropriate Biobanking SOPs have to be arranged by the species ambassador with the Biobanking partner, noting that biobanking may require materials in specific tube or plate types.

BH. **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 code), followed by the collection code and the voucher id (institution code:collection code:voucher_id) and 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 on NCBI Biocollections (https://ftp.ncbi.nih.gov/pub/taxonomy/biocollections/). If not available to you upon manifest validation but **TISSUE_REMOVED_FOR_BIOBANKING** is **Y** you need to use **NOT_PROVIDED.**

BI. **TISSUE_FOR_BIOBANKING**: Please 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".

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

BK. **DNA_VOUCHER_ID_FOR_BIOBANKING**: (**ENA_submission**) Accession number of DNA biobanked from the sequenced specimen. This ID should be prefixed by the acronym of the institution, followed by the collection code and the material id (institution code:collection code:material_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 on NCBI Biocollections (https://ftp.ncbi.nih.gov/pub/taxonomy/biocollections/). If not available to you upon manifest validation but **DNA_REMOVED_FOR_BIOBANKING** is **Y** you need to use **NOT_PROVIDED.**

BL. **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 code) followed by the collection code (if available) and the voucher id (institution_code:collection_code:voucher_id). More specifically, the

Institution Code identifies the institution that holds the voucher. It should be a widely used acronym for the institution or the full name if short. The **Collection Code** identifies the collection within the institution. Registered Institution and collection codes can be looked up on NCBI Biocollections (https://ftp.ncbi.nih.gov/pub/taxonomy/biocollections/). 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 "|" 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.

BM. **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 code) followed by the collection code (if available) and the voucher id (institution_code:collection_code:voucher_id). More specifically, the **Institution Code** identifies the institution that holds the voucher. It should be a widely used acronym for the institution or the full name if short. The **Collection Code** identifies the collection within the institution. Registered Institution and Collection codes can be looked up on NCBI Biocollections (https://ftp.ncbi.nih.gov/pub/taxonomy/biocollections/). The (proxy) **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 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 update at a later time.

BN. **VOUCHER_LINK:** 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/). 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.

BO. **PROXY_VOUCHER_LINK:** 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.

BP. **VOUCHER_INSTITUTION:** 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.

BQ. **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 species ambassador and are available to the species ambassador 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, indigenous rights 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 (for example where the sample is from a long-established culture) please 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.

BR. **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.

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

BT. **ASSOCIATED_TRADITIONAL_KNOWLEDGE_OR_BIOCULTURAL_PROJECT_ID**: project ID provided by the Local Context Hub (https://localcontexts.org/) upon notice registration.

BU. **ASSOCIATED_TRADITIONAL_KNOWLEDGE_CONTACT**: Free text allowed, provide reference, could be linked to an ORCID ID.

BV. **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.

BW. **ETHICS_PERMITS_DEF**: Free text explaining permits, permit issuing entity and permit number. If the previous column says no, enter NOT_APPLICABLE.

An upload field will be triggered if column BS is set to "Y" and all explained permits need to be uploaded in a single (concatenated) pdf named SPECIMEN_ID_ETHICS_PERMITS.pdf.

BX. **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.

BY. **SAMPLING_PERMITS_DEF**: Free text explaining permits, permit issuing entity and permit number. If the previous column says no, enter NOT_APPLICABLE.

An upload field will be triggered if column BU is set to "Y" and all explained permits need to be uploaded in a single (concatenated) pdf named SPECIMEN_ID_SAMPLING_PERMITS.pdf.

BZ. **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.

CA. **NAGOYA_PERMITS_DEF**: Free text explaining permits, permit issuing entity and permit number. If the previous column says no, enter NOT_APPLICABLE. -

An upload field will be triggered if column BW is set to "Y" and all explained

permits need to be uploaded in a single (concatenated) pdf named SPECIMEN_ID_NAGOYA_PERMITS.pdf.

CB. **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 the drop-down menu.

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