Population Genomics Sample Metadata for the BGE Project*

Population Genomics Sample Manifest Standard Operating Procedure

Version: v1.1, Adapted for use with the ERGA Sample Manifest Standard Operating Procedure v2.5.

Published Date: XX.11.2023

ERGA Sample Manifest Standard Operating Procedure 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, Astrid Böhne, and the ERGA Consortium

Adapted by: João Pedro Marques, José Melo-Ferreira, João Pimenta, Maria J. Ruiz-López, Leif Andersson, Angelica Crottini

Correct, ethical, and comprehensive recording of sample metadata is critical to the long-term utility of the genomics work in the BGE project: these metadata will link genome data to their origins. This Population Genomics Sample Manifest Standard Operating Procedure (SOP) is to be used with V2.5 of the ERGA Manifest available at https://github.com/ERGA-consortium/ERGA-sample-manifest. Read this SOP in full before completing the Population Genomics Sample Manifest as it contains detailed guidance on how to record metadata. The submission of a completed manifest is mandatory for BGE-associated genomic data.

***This sampling manifest adapts the ERGA Sampling Manifest which builds on the work of the Darwin Tree of Life (DToL) sampling committee, in particular of Mara Lawniczak and Robert Davey. We thank ERGA and DToL for allowing adapting their sampling metadata manifest to the BGE Population Genomics needs and for document sharing. All changes to this document only apply to BGE-ERGA population sampling and not to DToL or ERGA.**

Preamble: To be able to register a sample/specimen and its metadata for BGE, the submitting person (most often identical to the Case Study or Sampling Coordinator) must 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: Biodiversity Genomics Europe (BGE) aims to use and develop genomic tools in biodiversity research from biological 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. It contains specific instructions for filling in the metadata manifest for sampling in the framework of the Case Studies developed under BGE's Work Package 11 (WP11), as well as the pollinator population sampling for genomic analyses under Work Package 12 (WP12). BGE 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 Case Study Coordinator(s) for WP11 or the Sampling Coordinator for WP12, hereafter designated "Coordinator". By submitting the sampling manifest and providing information on compliance with sampling permits, the Coordinator guarantees that the sample in question can be legally processed and sequenced, and has been sampled in compliance with all applicable rules. The responsibility for the oversight of all legal compliance remains with the Coordinator. Where necessary and applicable, material transfer agreements can be issued and signed between the parties involved in sample processing.

Use and Future plans for this SOP: This SOP is intended for use with V1.1 of the Population Genomics Manifest (Based on the v2.5 ERGA Sample Manifest). It is planned that the input of population samples is fully implemented within the next version of the ERGA Manifest. Metadata are currently collected manually by the Coordinator using a defined spreadsheet, referred to as the **BGE-ERGA_PopGenomicsManifest_v1.1.xlsx.** 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: Please refer to the original ERGA Sample Manifest V2.5 SOP for standing issues. Please raise specific issues by emailing the BGE-WP11 at bge-wp11@erga-biodiversity.eu with the subject indicating "Population Genomics Sample Manifest". 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	2023-09-28	first version	João Pedro Marques, José Melo-Ferreira, João Pimenta, Maria J. Ruiz-López, Leif Andersson and Angelica Crottini
1.1	2023-11-07	General document maintenance C.PURPOSE_OF_SPECIMEN: replacement of term “SHORT_READ_SEQUENCING “ with “RESEQUENCING”. AD.DECIMAL_LATITUDE and AE.DECIMAL_LONGITUDE: addition of term NOT_COLLECTED Update AO.ORIGINAL_DECIMAL_LATITUDE and AP.ORIGINAL_DECIMAL_LONGITUDE text. Extended explanation to CE. SAMPLING_PERMITS_DEF and CL. ASSOCIATED_PROJECT_ACCESSIONS, corrected naming in SOP for the latter Update to J.TAXON_ID text	João Pedro Marques, José Melo-Ferreira

Completing the Sample Manifest: Overview

Scope of this document

Specific guidance on preparing samples is not covered by this SOP.

Submission of samples for sequencing is also not covered by this SOP.

The importance of "SPECIMEN_ID"

The 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 SPECIMEN_ID also allows the laboratory team to resample the same individual specimen 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 SPECIMEN_IDs, even if they are all from the same species. However, a single specimen split across ten tubes would result in each of those ten tubes having the same SPECIMEN_ID. This unique SPECIMEN_ID has three critical functions: identifying the Coordinator that holds responsibility for the specimen, tracking an individual sample's status and declaring the genetic uniqueness of the specimen.

Each specimen must be linked to a standardized, unique ID that begins with the prefix **ERGA_** followed by **COORDINATOR INITIALS** (up to 10 letters out of A-Z, if this is not possible please reach out to bge-wp11@erga-biodiversity.eu), followed by **underscore** followed by the **last four digits of the COORDINATOR's ORCID ID**, **underscore**, and **running numbers** (e.g. if you as coordinator register more than one sample, including samples with purposes other than Population Genomics analyses, make sure to use 00001, 00002…). 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.

Other "_ID"s

A sample can represent a set of specimens as well as multiple parts of the same specimen, and so the 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. 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 two categories

(COLLECTOR_SAMPLE_ID, SPECIMEN_ID). The SPECIMEN_ID is the only ID that is required for a sample to enter the BGE workflow and metadata upload to commence. We strongly urge sample providers to complete metadata collection and upload before commencing sequencing to guarantee a sample adheres to BGE's standards.

Management of COLLECTOR_SAMPLE_ID and their relationship to SPECIMEN_ID is the responsibility of the Coordinator.

Manifest Validation Process

Use the google spreadsheet as an XLS/XLSX (Microsoft Excel format) file for upload to COPO. **Ensure that you upload the manifest to a profile that has "Biodiversity Genomics Europe (BGE)" and "POP_GENOMICS" as associated profile types.** 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.

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.

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 malformatted 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 data are submitted to ENA for release (as part of BioSample, and raw data submissions), the submissions will include all of the fields below indicated by **ENA_submission**. **If the field name is in green, 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. 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 may be vouchered in a public scientific collection dedicated to permanent storage and with an accessible voucher catalog. Vouchering should be coordinated by the Coordinator or another designated person from the institution that keeps ownership of the sample. Upon integration in a collection, vouchers will be attributed a collection specific ID that can be recorded in the metadata (see below, field BG-BP) as well as the collection name. To be properly displayed together with genome sequencing information, collections can 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/biocollections/. Ideally, in addition to a physical voucher, tissue and/or cells and/or DNA can be deposited in a public Biobank/ frozen repository, ideally a member of GGBN. When vouchering is not possible, digital images may be 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 appreciated (see below).

BGE population genomics sample manifest roadmap

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

Mandatory fields are marked in **green** in the table below and Optional in **white**.

Mandatory fields always require filling. If information is absent you need to enter an accepted term (details in parenthesis).**Optional fields** can be left blank if information is absent, in which case a default term is assumed.

Block 1: Sample submission information including specimen identifier and tube/well identifiers, as well as information on the 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 organizational 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 to CM)

Recording Population Genomics Sample Metadata for BGE

A	B	C	D	E	F	G	H	I	J	K
TUBE_OR_WELL_ID	SPECIMEN_ID	PURPOSE_OF_SPECIMEN	SAMPLE_COORDINATOR	SAMPLE_COORDINATOR_AFFILIATION	SAMPLE_COORDINATOR_ORCID_ID	ORDER_OR_GROUP	FAMILY	GENUS	TAXON_ID	SCIENTIFIC_NAME
L	M	N	O	P	Q	R	S	T	U	V
TAXON_REMARKS	INFRASPECIFIC_EPITHET	CULTURE_OR_STRAIN_ID	COMMON_NAME	LIFESTAGE	SEX	ORGANISM_PART	SYMBIONT	RELATIONSHIP	GAL	GAL_SAMPLE_ID
W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
COLLECTOR_SAMPLE_ID	COLLECTED_BY	COLLECTOR_AFFILIATION	COLLECTOR_ORCID_ID	DATE_OF_COLLECTION	TIME_OF_COLLECTION	COLLECTION_LOCATION	DECIMAL_LATITUDE	DECIMAL_LONGITUDE	LATITUDE_START	LONGITUDE_START
AH	AI	AI	AK	AL	AM	AN	AO	AP	AQ	AR
LATITUDE_END	LONGITUDE_END	HABITAT	DEPTH	ELEVATION	ORIGINAL_COLLECTION_DATE	ORIGINAL_GEOGRAPHIC_LOCATION	ORIGINAL_DECIMAL_LATITUDE	ORIGINAL_DECIMAL_LONGITUDE	DESCRIPTION_OF_COLLECTION_METHOD	DIFFICULT_OR_HIGH_PRIORITY_SAMPLE
AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC
IDENTIFIED_BY	IDENTIFIER_AFFILIATION	IDENTIFIED_HOW	SPECIMEN_IDENTITY_RISK	MIXED_SAMPLE_RISK	PRESERVED_BY	PRESERVER_AFFILIATION	PRESERVATION_APPROACH	PRESERVATIVE_SOLUTION	TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION	DATE_OF_PRESERVATION
BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN
SIZE_OF_TISSUE_IN_TUBE	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	PROXY_TISSUE_VOUCHER_ID_FOR_BIOBANKING	TISSUE_FOR_BIOBANKING	BIOBANKED_TISSUE_PRESERVATIVE
BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY
DNA_REMOVED_FOR_BIOBANKING	DNA_VOUCHER_ID_FOR_BIOBANKING	VOUCHER_ID	PROXY_VOUCHER_ID	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	CG	CH	CI	CJ
ASSOCIATED_TRADITIONAL_KNOWLEDGE_CONTACT	ETHICS_PERMITS_REQUIRED	ETHICS_PERMITS_DEF	ETHICS_PERMITS_ FILENAME	SAMPLING_PERMITS_ REQUIRED	SAMPLING_PERMITS_DEF	SAMPLING_PERMITS_ FILENAME	NAGOYA_PERMITS_REQUIRED	NAGOYA_PERMITS_DEF	NAGOYA_PERMITS_FILENAME	HAZARD_GROUP
CK	CL	CM								
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.

II. **Information must be entered for all fields below with green bold names**. 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 may 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. If these fields are left blank, a default “missing value term” may be assumed (indicated for each field below).

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 Coordinator 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 SPECIMEN_ID refers to a singular genetic individual. If the same individual specimen is split into several samples submitted in separate tubes, the 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 SPECIMEN_ID. Each specimen must be linked to a standardized, unique ID that begins with the prefix ERGA_ followed by COORDINATOR INITIALS (up to 10 letters out of A-Z, if this is not possible please reach out to bge-wp11@erga-biodiversity.eu), followed by underscore followed by the last four digits of the COORDINATOR's ORCID ID, underscore and RUNNING NUMBERS (make sure to use 00001, 00002…). SPECIMEN_IDs must be unique to an individual (e.g., ERGA_XY_1234_001 cannot be used again after it has been assigned to a specimen). SPECIMEN_IDs must follow the format described above.

C. **PURPOSE_OF_SPECIMEN:** As specimens will be intended for population genetics via short read resequencing, please use "RESEQUENCING".

D. **SAMPLE_COORDINATOR:** (**ENA_submission**) Also known as the Case study or Sampling Coordinator. Enter the name of the person or people who is responsible for the case study (WP11) or sampling (WP12) 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 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 indicated in the SAMPLE_COORDINATOR field. 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-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 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.
  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) ENA accepts them as submittable (ENA “submittable” : “true”)
  c) the species name qualifies as binomial (ENA “binomial” : “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 taxonomy), 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 taxonomy browser), 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. 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.

- 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). If the information is unknown use **NOT_COLLECTED**.

- 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.

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:

i. 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 or BGE 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 a BGE partner but a commercial provider, leave the field blank and "**INDUSTRY_PARTNER**" will be autofilled.

V. **GAL_SAMPLE_ID:** (**ENA_submission**) This is the unique name assigned to the sample by the GAL. If it is not applicable, it can be the same as TUBE_OR_WELL_ID or COLLECTOR_SAMPLE_ID. GAL_SAMPLE_ID 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. Please ensure you do not use IDs that have already been used, and if available that you stick to the format required by the GAL. If left **blank** it will default to **COLLECTOR_SAMPLE_ID** (see W).

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.

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 BGE system under the aegis of the GDPR regulations, and we must ask 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 Coordinator 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. If left **blank** it defaults to **NOT_PROVIDED**.

AA. **DATE_OF_COLLECTION**: (**ENA_submission**) The date the sample was collected, in ISO8601 format (year, month and day, or year and month, or year; **YYYY-MM-DD, YYYY-MM, YYYY**).

AB. **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.

AC. **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 INSDC 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. The number of decimal places can be used to accommodate for precision of the geographic location. For example, using 3 decimal places is accurate for 111 meters, 2 is accurate for 1.11 Km, 1 is accurate for 11.1 Km and zero is accurate for 111 Km (https://en.wikipedia.org/wiki/Decimal_degrees). We advise that locations are specified, when possible, to a minimum of 3 decimal places.

- 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).
- 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. The number of decimal places can be used to accommodate for precision of the geographic location. For example, using 3 decimal places is accurate for 111 meters, 2 is accurate for 1.11 Km, 1 is accurate for 11.1 Km and zero is accurate for 111 Km (https://en.wikipedia.org/wiki/Decimal_degrees). We advise that locations are specified, when possible, to a minimum of 3 decimal places.

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

- Only provide if **DECIMAL_LATITUDE (AD)** is “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. For example, using 3 decimal places gives a location accurate to 111 meters, whereas using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters See http://wiki.gis.com/wiki/index.php/Decimal_degrees.

- Only provide if **DECIMAL_LONGITUDE (AE)** is "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 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. For example, using 3 decimal places gives a location accurate to 111 meters, whereas using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters See http://wiki.gis.com/wiki/index.php/Decimal_degrees.

- Only provide if **DECIMAL_LATITUDE (AD)** is "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 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. For example, using 3 decimal places gives a location accurate to 111 meters, whereas using 4 is accurate to 11.1 meters, and 5 is accurate to 1.11 meters See http://wiki.gis.com/wiki/index.php/Decimal_degrees.

- Only provide if **DECIMAL_LONGITUDE (AE)** is "NOT_COLLECTED"

AJ. **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.

AK. **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.

AL. **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.

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**). YYYY-MM or 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.

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. 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.

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) or difficult to be integrated in genome data generation process (e.g. hard to get good quality DNA).

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 BGE system under the aegis of the GPDR regulations, and we must ask 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 Coordinator is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

AT. **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.

AU. **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.

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. 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).

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. If left **blank**, **NOT_COLLECTED** is assumed.

- We note that storage of names with affiliations in a database brings the BGE system under the aegis of the GDPR regulations, and we must ask 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 Coordinator is asked to seek agreement from all involved collaborators before uploading the metadata sheet into COPO.

AY. **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. If left **blank**, **NOT_COLLECTED** is assumed.

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

BA. **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**.

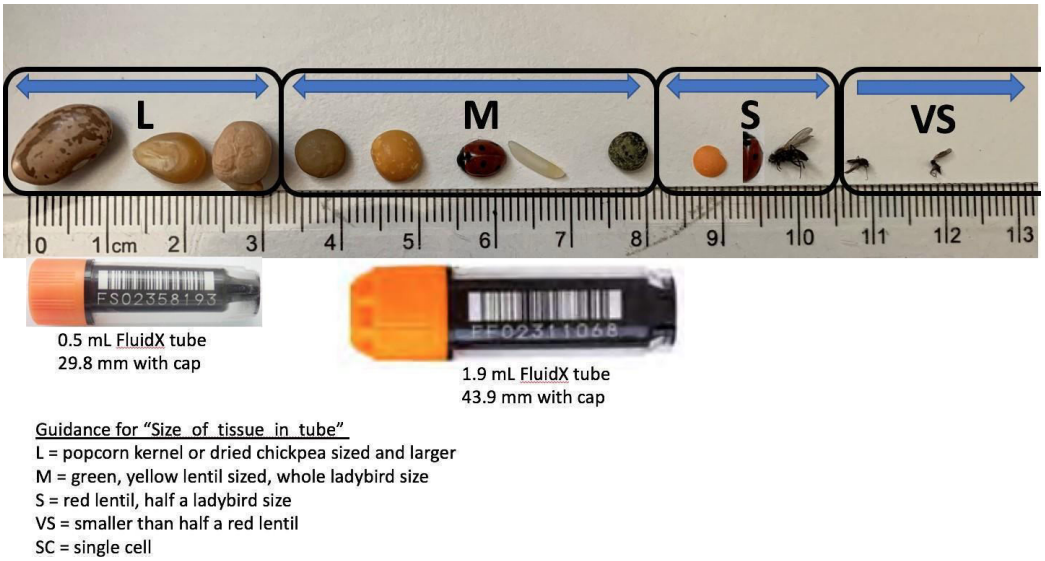
BB. **TIME_ELAPSED_FROM_COLLECTION_TO_PRESERVATION**: some organisms may be held living in collection for a period of time for starvation or other factors. This entry should be specified in hours, but no unit, e.g., 0.5 for half an hour, 3 for 3 hours, etc. If left **blank**, **NOT_COLLECTED** is assumed.

BC. **DATE_OF_PRESERVATION**: Date on which the species was preserved. Please use **YYYY-MM-DD** format. If left **blank**, **NOT_COLLECTED** is assumed.

BD. **SIZE_OF_TISSUE_IN_TUBE**: Select from the drop-down menu how large is the sample in the tube. If left **blank**, **NOT_COLLECTED** is assumed. 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



BE. **TISSUE_REMOVED_FOR_BARCODING**: Select from drop-down menu "Y" or "N". If left

blank, **N** is assumed.

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. If left **blank**, **NOT_APPLICABLE** is assumed.

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". If left **blank**, **NOT_APPLICABLE** is assumed.

BH. **BARCODE_PLATE_PRESERVATIVE:** Record the volume, concentration, and type of preservative/method of preservation used here. If left **blank**, **NOT_APPLICABLE** is assumed.

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 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_IDENTITY_RISK has the entry "N". If left **blank**, **DNA_BARCODING_TO_BE_PERFORMED_GAL** is assumed.

BJ. **TISSUE_REMOVED_FOR_BIOBANKING:** Select from drop-down menu "Y" or "N". Instructions for appropriate Biobanking SOPs have to be arranged with the Biobanking partner, noting that biobanking may require materials in specific tube or plate types. If left **blank**, **N** is assumed

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 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.** If left **blank**, **NOT_APPLICABLE** is assumed.

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 30 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 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/) or using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/). Where there are multiple vouchers to cite for a given specimen, separate the different Voucher IDs with a "|" symbol. If it is not the case, leave the field **blank**.

BM. **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". If left **blank**, **NOT_APPLICABLE** is assumed.

BN. **BIOBANKED_TISSUE_PRESERVATIVE:** Record the volume, concentration, and type of preservative/method of preservation used here. If left **blank**, **NOT_APPLICABLE** is assumed.

BO. **DNA_REMOVED_FOR_BIOBANKING**: Select from drop-down menu "**Y**" (yes) or "**N**" (no). If left **blank**, **N** is assumed.

BP. **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/) or 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.** If left **blank**, **NOT_APPLICABLE** is assumed.

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 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/) or using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/). 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. If left **blank**, **NOT_PROVIDED** is assumed.

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 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/) or using the ENA Source Attribute Helper API (https://www.ebi.ac.uk/ena/sah/api/). 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.

BS. **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.

BT. **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.

BU. **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.

BV. **REGULATORY_COMPLIANCE:** Please select from the drop-down menu Y (yes), NOT_APPLICABLE or N (not known). Note that the Coordinator 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 Coordinator and are available to the 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 (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.

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

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

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

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 the previous column says no, enter NOT_APPLICABLE. An upload field will be triggered if column BV is set to "Y" and 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

column CA says NO, 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 column CD says NO, 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 the previous column says no, enter NOT_APPLICABLE. An upload field will be triggered if column BZ 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 CG says NO, 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 the drop-down menu.

CK. **PRIMARY_BIOGENOME_PROJECT:** Indicate if your genome is part of ERGA-Pilot, ERGA-BGE or an ERGA-associated genome (select ERGA-associated).

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 not captured by the other fields. If there is nothing else to add here, this field should be left **blank**.