



EUROPEAN REFERENCE GENOME ATLAS

Guidelines on Data Submission v1.5



Guidelines on data submission

-
- Version 1.0 July 19, 2024: Chiara and Tom wrote the document
 - Version 1.1 July 31, 2024: Chiara added a few sentences on the BioSample registration section to address comments from sequencing centers
 - Version 1.2 February 18, 2025: Chiara and Tom changed the text according to Kerstin's comments
 - Version 1.3 March 13, 2025: Chiara updated her email address
 - Version 1.4 July 21, 2025: Chiara and Tom updated the FAQs
 - Version 1.5 February 09, 2026: Chiara added the locus tag prefix in section 8.2
-

Table of Contents

1. Structural organisation of Umbrella BioProjects	2
2. Important terminology	3
3. Register a submission account	3
4. ToLID registration	4
5. Register a BioSample	4
5.1 Via COPO	4
5.2 Via ENA	4
6. Create sequencing data and genome assembly BioProjects	5
7. Submitting and publishing sequencing reads	8
7.1 Prepare your data for submission	8
7.2 Upload files to ENA	8
7.3 Submit Raw Reads Interactively	9
7.4 Submit Raw Reads with Webin-CLI	9
7.5 Submit Raw Reads Programmatically	10
8. Submitting and publishing genome assemblies (and annotations)	10
8.1 Prepare your data for submission	10
8.1 Chromosome-level assembly	10
8.2 Assemblies with protein-coding sequence annotation	12
8.3 Other assemblies	13
9. Group BioProjects under a Species Umbrella BioProject	13
APPENDIX	16
1. Create an ERGA Umbrella BioProject	16
FREQUENTLY ASKED QUESTIONS (FAQs)	18

2. Important terminology

Here, we clarify a few terms that will be used throughout these guidelines:

- **Primary assembly:** a representative assembly of the genetic material from one haplotype of the target genome. This should ideally also include assemblies for any mitochondrial or plastid sequences;
- **ToLID:** a unique identifier for an individual of a species sampled for genome sequencing (e.g. xgPerPere3 for the third registered individual of *Peregriana peregra*);
- **ToLID prefix:** the part of the ToL ID that is made of (1) one lower case letter for the high-level taxonomic rank and one lower case letter for the clade, (2) one upper, two lower case letters for genus, (3) one upper, three lower case letters for species (e.g. iIAbaGlen for *Abagrotis glenni*). For historical reasons this differs for vertebrates, which contains one letter for the high-level taxonomic rank and one upper, two lowercase letters for the species only (e.g. aRanPir for *Rana pirica*);
- **Taxon ID:** a stable, unique numerical identifier of a biological taxon (e.g., 9606 for *Homo sapiens*).
- **Scientific name (or species name):** the binomial name formally used to name species. It is made of two parts, the first - the *generic name* - identifies the *genus* to which the species belongs, the second part - the *specific name* - distinguishes the species within the genus (e.g., *Homo sapiens* for human).

3. Register a submission account

Before you can submit data to ENA, you must register a Webin submission account. To do so, please navigate to the [Webin Portal](#) and click the 'Register' button. Fill out the boxes, giving details of the group and center for which you are submitting. Always make sure your account's center name is correct before you perform a submission.

NOTE: The center name under which you register your Webin account is the one that will be displayed in the BioProject "Center Name" section once data is submitted. The center name can be changed on a later stage. However, this change will be applied irrevocably to all submissions made from the account. If you need to submit data for various sequencing centers, you may want to [submit a request](#) to the ENA for a broker account.

4. ToLID registration

To help monitor the progress of biodiversity genomics projects across the globe, the Earth Biogenome Project recommends assigning a unique Tree of Life Identifier (ToLID) to your specimen. This can then be used to identify samples, sequencing projects and assemblies once available in public repositories as well as coordinate between EBP-affiliated genome projects. If you are registering samples via COPO (**see 5.1**), ToLIDs will be automatically assigned to your samples. If you are submitting sample metadata to ENA manually (**see 5.2**), you can register a ToLID for your samples via id.tol.sanger.ac.uk/, which can then be included in the BioSample registration.

5. Register a BioSample

5.1 Via COPO

For researchers producing genomes that will contribute to the efforts of the European Reference Genome Atlas (ERGA), we recommend registering a BioSample through the submission of an ERGA manifest in [COPO](#). For more information on how to submit an ERGA manifest in COPO, please refer to this [page](#).

NOTE: all ERGA-BGE species **must be registered via COPO** before arriving at a sequencing center. The registration via COPO will automatically assign a BioSample that you can retrieve either in the [COPO Dashboard](#) or in the [ERGA-GTC Tracking Tool](#). Use these BioSamples in the following steps.

5.2 Via ENA

As an alternative, it is possible to register a BioSample directly in ENA. Log in to the [Webin Portal](#) and select the 'Register Samples' button. Click the 'Download spreadsheet to register samples' button and select the most appropriate checklist group. In the context of ERGA, we recommend using either of these two checklists: **ERC000011** (ENA default sample checklist) or **ERC000053** (Tree of Life Checklist). For more information on how to complete the template spreadsheet, refer to the ENA guidelines [here](#).

NOTE: When registering a BioSample, you are not submitting any data; instead, you are providing sample metadata.

Once filled in, return to the 'Register Samples' interface in the Webin portal and this time expand the 'Upload filled spreadsheet to register samples' option. Use the 'Browse' button to find the spreadsheet you wish to submit, then click the 'Submit Completed Spreadsheet' button.

At this point, your samples will be validated and if accepted, they will have accession numbers associated with them. If there are errors with the information you have entered, these will be reported to you. This process might take some time, from a couple of hours to a few days, depending on the number of samples provided in the spreadsheet.

6. Create sequencing data and genome assembly BioProjects

To register a BioProject (or study) which houses the sequencing data and genome assemblies, you need to create a [submission.xml](#) and a [project.xml](#) file. In this example, we recommend renaming the [project.xml](#) file to create the raw sequencing data BioProject as [genome_sequencing.xml](#), whereas that for the (primary) genome assembly as [primary_assembly.xml](#). Sections **highlighted in yellow** will need to be changed depending on the species you are working on and the Affiliated Project you want the data to be linked to. To create a BioProject, make sure to have the following information:

- The species scientific name;
- The species ToL ID prefix;
- The name of the sequencing center that generated the raw data;
- The name of the Affiliated Project (e.g., ERGA-BGE, DToL, etc).

You can check if your species has a Taxon ID in the [NCBI Taxonomy database](#). If it does not exist, you can request a Taxon ID through [ENA](#).

You can check if your species has a ToL ID in the [Tree of Life Identifiers](#) website. If it does not exist, you can request a new one on the same website.

[submission.xml](#)

```
<?xml version="1.0" encoding="UTF-8"?>
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <ADD/>
    </ACTION>
    <ACTION>
      <HOLD HoldUntilDate="<yy-mm-dd>"/>
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

To create the BioProject for the raw sequencing data, use the following XML:

genome_sequencing.xml

```
<PROJECT_SET>
  <PROJECT alias="<affiliated-project>-<tolid_prefix>-study-rawdata-<yyyy-mm-dd>"
broker_name1="<broker name>" center_name="<center name>">
    <NAME><tolid_prefix></NAME>
    <TITLE><scientific name>, genomic and transcriptomic data</TITLE>
    <DESCRIPTION>The description goes here.</DESCRIPTION>
    <SUBMISSION_PROJECT>
      <SEQUENCING_PROJECT/>
    </SUBMISSION_PROJECT>
  </PROJECT>
</PROJECT_SET>
```

To register the BioProject, submit the [submission.xml](#) and [genome_sequencing.xml](#) as:

```
[1] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@genome_sequencing.xml"
"https://wwwdev.ebi.ac.uk/ena/submit/drop-box/submit/"
```

Where the 'Username' is your Webin account and 'Password' is the password to your Webin account. The command will return a [receipt.xml](#) containing a submission ID, the created BioProject ID (PRJEB###), and details of whether the submission was a success or not (success=true).

The receipt will look like:

receipt.xml

```
<RECEIPT receiptDate="2023-10-13T15:52:20.250Z" submissionFile="submission.xml"
success="true">
  <PROJECT accession="PRJEB###"
alias="<affiliated-project>-<tolid_prefix>-study-rawdata-<yyyy-mm-dd>" status="PRIVATE"
holdUntilDate="yy-mm-dd">
    <EXT_ID accession="ERP###" type="study"/>
  </PROJECT>
  <SUBMISSION accession="ERA###" alias="SUBMISSION-13-10-2023-15:52:20:180"/>
  <MESSAGES>
    <INFO>All objects in this submission are set to private status (HOLD).</INFO>
  </MESSAGES>
  <ACTIONS>ADD</ACTIONS>
  <ACTIONS>HOLD</ACTIONS>
</RECEIPT>
```

If you are happy with the submission you can run the command [1] on the production service, as:

```
[2] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@genome_sequencing.xml"
"https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

¹ The broker name is necessary only if your Webin account is a broker account, otherwise you can remove it.

To release the BioProject to the public, you can remove the 'HoldUntilDate' in the [submission.xml](#). Alternatively, you can create another XML file, called [hold_date.xml](#), specifying the new release date of the BioProject ID, as:

[hold_date.xml](#)

```
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <RELEASE target="PRJEB####" />
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

NOTE: Be sure that the BioProject ID in the [hold_date.xml](#) file matches that in the [receipt.xml](#).

Once you have the [hold_date.xml](#), you can use the [curl](#) command again as below:

```
[3] curl -u Username:Password -F "SUBMISSION=@hold_date.xml" "https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

To create the genome assembly BioProjects, you will need as many XML files as there are assemblies to submit (primary/alternate/cobiont/hap1/hap2/etc). For this step, you can use the [submission.xml](#) and [hold_date.xml](#) files described above.

In this example, we will create two BioProjects, one for the primary ([primary_assembly.xml](#)) and one for the alternate haplotype ([alternate_assembly.xml](#)).

NOTE: It is important that aliases are unique and meaningful.

[primary_assembly.xml](#)

```
<PROJECT_SET>
  <PROJECT alias="<affiliated-project>-<tolid_prefix>_primary--<yyyy-mm-dd>"
broker_name="<broker name>" center_name="<center name>">
    <NAME><tolid_prefix><NAME>
    <TITLE><scientific name>, primary genome assembly, <tolid></TITLE>
    <DESCRIPTION>The description goes here.</DESCRIPTION>
    <SUBMISSION_PROJECT>
      <SEQUENCING_PROJECT/>
    </SUBMISSION_PROJECT>
  </PROJECT>
</PROJECT_SET>
```

[alternate_assembly.xml](#)

```
<PROJECT_SET>
  <PROJECT alias="<affiliated-project>-<tolid_prefix>_alternate--<yyyy-mm-dd>"
broker_name="<broker name>" center_name="<center name>">
    <NAME><tolid_prefix></NAME>
```

```
<TITLE><scientific name>, alternate genome assembly, <tolid></TITLE>
<DESCRIPTION>The description goes here.</DESCRIPTION>
<SUBMISSION_PROJECT>
  <SEQUENCING_PROJECT/>
</SUBMISSION_PROJECT>
</PROJECT>
</PROJECT_SET>
```

As before, to register the BioProjects run:

```
[4] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@primary_assembly.xml"
"https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

```
[5] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@alternate_assembly.xml"
"https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

As before, take note of the [receipt.xml](#) files, which detail the created BioProjects ID and whether the submissions were successful. To release the BioProjects and make them visible, create a [hold_date.xml](#) file for the primary (e.g., [hold_date_primary.xml](#)) and alternate assembly (e.g., [hold_date_alternate.xml](#)). Lastly, run command [3] replacing the [hold_date.xml](#) with the appropriate XML file.

Great! Now you have all the BioProjects to which you can link your raw sequencing data and genome assemblies and/or annotations.

7. Submitting and publishing sequencing reads

7.1 Prepare your data for submission

All files submitted to ENA need to be in the appropriate format to be accepted. For a comprehensive overview of the data format accepted by ENA, please refer to the [Accepted Read Data Formats](#). All files must be compressed with [gzip](#) or [bgzip2](#) and they must have their [MD5 checksum](#) registered in lower case letters. You can obtain this file by running either [md5](#) or [md5sum](#).

7.2 Upload files to ENA

You must upload your data files into your private Webin file upload area at EMBL-EBI before you can submit the files through the Webin submission service. You can upload files to ENA in different ways:

1. Through the [Webin File Uploader](#): this is the **most user-friendly approach**. However, you first need to download the Webin File Uploader from the [Webin Portal](#).
2. Through the command line FTP Client (suitable for Linux and Mac users). Windows users can use FileZilla instead.
3. Through Aspera ascp command line: this is the **simplest and quickest way** to upload files to ENA. To use Aspera, first download it from [here](#).

Then run: `ascp -QT -l300M -L- <file(s)> Username@webin.ebi.ac.uk.`

For more information on uploading files to ENA, please refer to [this page](#).

Once you have uploaded and registered the raw reads, the Webin will report two unique accession numbers for each read submission. The first starts with ERR## and is called the Run accession. The other starts with ERX## and is called the Experiment accession.

At this stage of the process, you can proceed with the submission of the raw reads. For this step, you can choose from three options.

7.3 Submit Raw Reads Interactively

This might be the **simplest approach** to submit raw reads. For this, log in to the [Webin Portal](#) and complete these three steps: (1) Select and customise a read submission template spreadsheet (2) Fill out the template spreadsheet, (3) Validate and submit the template spreadsheet. For more information on this procedure, refer to [this page](#).

7.4 Submit Raw Reads with Webin-CLI

You can submit reads using the Webin command line submission interface that you can download [here](#). For this, you will need to prepare a manifest file, as explained [here](#).

This file takes the form of a tab-delimited two columns text file, which details the name of the sequencing files, the metadata for how the library was constructed and sequenced as well as linking to the BioSample and BioProjects created above:

[manifest.txt](#)

```
STUDY Study accession or unique name (e.g., PRJEB###)
SAMPLE Sample accession or unique name (e.g., ERS###)
NAME Unique experiment name
PLATFORM See permitted values. Not needed if INSTRUMENT is provided
INSTRUMENT See permitted values
INSERT_SIZE: Insert size for paired reads
LIBRARY_NAME: Library name (optional)
LIBRARY_SOURCE: See permitted values
LIBRARY_SELECTION: See permitted values
LIBRARY_STRATEGY: See permitted values
DESCRIPTION: free text library description (optional)
FASTQ Single file in FASTQ format
FASTQ Single file in FASTQ format
...
FASTQ Single file in FASTQ format
```

NOTE: Replace 'FASTQ' with 'CRAM' or 'BAM' if you submit a file in CRAM or BAM format.

Once you have prepared the manifest file, you can run:

```
[6] java -jar ./webin-cli-6.7.2.jar -context=reads -manifest=manifest.txt -username=Webin-#### -passwordFile=pwd.txt  
-submit -ascp
```

Where `ascp` must be in your `$PATH`. You can provide the password either in a file (e.g., `pwd.txt`) or as an argument. See `java -jar ./webin-cli-6.7.2.jar -help` for more information.

7.5 Submit Raw Reads Programmatically

You can submit reads by providing an experiment and run XML files. For more information about this procedure, refer to [this page](#).

8. Submitting and publishing genome assemblies (and annotations)

8.1 Prepare your data for submission

All files submitted to ENA need to be in the appropriate format to be accepted. For a comprehensive overview of the data format accepted by ENA, please refer to the [Accepted Genome Assembly Data Formats](#) page.

To submit a genome assembly to ENA, you must provide some metadata to describe your research project. This helps make your data reusable and searchable. ENA recognises three assembly levels: chromosome, scaffold, contig.

8.1 Chromosome-level assembly

This is the **highest level** of assembly and it includes assembled chromosomes. To submit a chromosome-level assembly, you need the following files (all should be zipped):

- 1 manifest file
- 1 FASTA file OR 1 [flat file](#)
- 1 [chromosome list file](#)
- 0-1 [unlocalised list files](#)
- 0-1 [AGP files](#)

NOTE: Include any organelles (mito/chloroplast/etc) as a chromosome in the FASTA file of the primary assembly.

For more information about the files, please refer to [this page](#).

The manifest file is a tab-delimited file which lists the BioProject and BioSample to which the assembly should be linked, the files to be submitted and the methods by which the assembly was produced. Please be as detailed as possible listing all software and versions as well as all sequencing types used. The assembly and chromosome list files should be gzipped.

[genome_manifest.txt](#)

STUDY Study accession (e.g., PRJEB###)
 SAMPLE Sample accession (e.g., ERS###)
 ASSEMBLYNAME Unique assembly name, user-provided
 ASSEMBLY_TYPE 'clone or isolate'
 COVERAGE The estimated depth of sequencing coverage
 PROGRAM The assembly program, or comma-separated list of programs
 PLATFORM The sequencing platform, or comma-separated list of platforms
 MINGAPLENGTH Minimum length of consecutive Ns to be considered a gap
 MOLECULETYPE 'genomic DNA', 'genomic RNA' or 'viral cRNA'
 DESCRIPTION Free text description of the genome assembly - *optional*
 RUN_REF: Comma separated list of run accession(s) - *optional*
 FASTA Name of assembly in FASTA format (zipped)
 CHROMOSOME_LIST Chromosome list file (zipped)

The chromosome list file is a tab-delimited file which takes the following format:

[chromosome_list.txt](#)

```
chr_1 1 Linear-Chromosome
chr_2 2 Linear-Chromosome
chr_3 3 Linear-Chromosome
chr_4 4 Linear-Chromosome
chr_5 5 Linear-Chromosome
chr_6 6 Linear-Chromosome
...
chr_21 21 Linear-Chromosome
chr_X X Linear-Chromosome
chr_Y Y Linear-Chromosome
mito_ctg000001c MT circular-chromosome Mitochondrion
chloroplast_ctg000002c CT circular-chromosome Chloroplast
```

Where the first column corresponds to the name of the sequence in your submitted assembly, second column is either the name/number of the chromosome or an abbreviation of a plastic, and the third and fourth columns give information about the topology and nature of the chromosome.

NOTE: Make sure that the name of the sequences in your submitted assembly matches that of the first column in the chromosome list file.

Once you have all your files ready, you can validate, upload, and submit them using the Webin command line submission interface or Webin-CLI. First, run the Webin-CLI validation command, specifying your credentials and the path to your manifest file, as below:

```
[7] webin-cli -username Webin-XXXXX -password YYYYYYY -context genome -manifest genome_manifest.txt -validate
```

Where 'Webin-XXXXX' is your Webin account and 'password' is the password to your Webin account.

Second, run the Webin-CLI submission command as:

```
[8] webin-cli -username Webin-XXXXX -password YYYYYYY -context genome -manifest genome_manifest.txt -submit
```

If both command lines are successful, an analysis (ERZxxxxxx) accession number is immediately assigned and returned to the submitter by the Webin-CLI. The purpose of the ERZ accession number is for the submitter to be able to refer to their submission within the Webin submission service.

8.2 Assemblies with protein-coding sequence annotation

Submissions of annotations must be performed with an assembly via an EMBL flat file format. To generate this, we recommend using [emblmygff3](#), which can be installed via BioConda and takes the genome assembly fasta and protein-coding gff3 file as input and produces a .embl file. This file contains all of the information from the fasta and gff3 file, so only one file is submitted alongside the chromosome list file as detailed above. Submission follows the same procedure via the [webin-cli](#), with the only exception that the manifest file now includes the "FLATFILE" entry and should detail the methods for genome annotation as well as assembly.

NOTE: When submitting an annotation, a **locus tag prefix** is required. Locus tags are identifiers applied systematically to every gene in a sequencing project that are meant to give an unambiguous name to every gene. A locus tag prefix can be registered when a project is registered. For more information, please visit [this page](#). If you did not add a prefix when you registered your project, it is possible to update the project with a prefix later on. For more information, visit [this page](#).

[annotated_genome_manifest.txt](#)

STUDY Study accession (e.g., PRJEB###)

SAMPLE Sample accession (e.g., ERS###)

ASSEMBLYNAME Unique assembly name, user-provided
 ASSEMBLY_TYPE 'clone or isolate'
 COVERAGE The estimated depth of sequencing coverage
 PROGRAM The assembly program, or comma-separated list of programs
 PLATFORM The sequencing platform, or comma-separated list of platforms
 MINGAPLENGTH Minimum length of consecutive Ns to be considered a gap
 MOLECULETYPE 'genomic DNA', 'genomic RNA' or 'viral cRNA'
 DESCRIPTION Free text description of the genome assembly - *optional*
 RUN_REF: Comma separated list of run accession(s) - *optional*
 FLATFILE Fast file in EMBL format
 CHROMOSOME_LIST Chromosome list file (zipped)

Once you have the files and manifest ready, run:

[9] `webin-cli -username Webin-XXXXX -password YYYYYYY -context genome -manifest annotated_genome_manifest.txt -validate`

[10] `webin-cli -username Webin-XXXXX -password YYYYYYY -context genome -manifest annotated_genome_manifest.txt -submit`

8.3 Other assemblies

If you are not working on a chromosome-level assembly, you will not need to submit a chromosome list file. In this case, you will need to submit just a zipped FASTA file through the manifest, removing the final line ("CHROMOSOME_LIST") from the manifest.

9. Group BioProjects under a Species Umbrella BioProject

Now that you have created a BioProject for the sequencing data and genome assembly, and now that the raw reads and genome assembly have been submitted to ENA, you can proceed with grouping these BioProjects under a Species Umbrella BioProject. To create an Umbrella BioProject, you will need your Webin username and password and two XML files, one for the umbrella project itself ([umbrella.xml](#)) and one for the submission to ENA ([submission.xml](#)).

[umbrella.xml](#)

```
<PROJECT_SET>
  <PROJECT alias="<affiliated-project>-<tolid_prefix>-study-umbrella--<yyyy-mm-dd>"
broker_name="<broker name>" center_name="<center name>">
    <NAME><tolid_prefix></NAME>
    <TITLE><scientific name></TITLE>
    <DESCRIPTION>The description goes here.</DESCRIPTION>
    <UMBRELLA_PROJECT/>
  </PROJECT>
</PROJECT_SET>
```

submission.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <ADD/>
    </ACTION>
    <ACTION>
      <HOLD HoldUntilDate="<yy-mm-dd>" />
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

To create the umbrella project, use the following command:

```
[11] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@umbrella.xml"
"https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

As usual, you should receive a receipt, including the ID of the newly created BioProject, like so:

receipt.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="receipt.xsl"?>
<RECEIPT receiptDate="2023-11-21T15:32:39.571Z" submissionFile="submission.xml"
success="true">
  <PROJECT accession="PRJE#####"
alias="<affiliated-project>-<tolid_prefix>-study-umbrella--<yyyy-mm-dd>" status="PRIVATE"
holdUntilDate="2025-11-20Z"/>
  <SUBMISSION accession="ERA#####" alias="SUBMISSION-21-11-2023-15:32:39:468"/>
  <MESSAGES>
    <INFO>All objects in this submission are set to private status (HOLD).</INFO>
  </MESSAGES>
  <ACTIONS>ADD</ACTIONS>
  <ACTIONS>HOLD</ACTIONS>
</RECEIPT>
```

NOTE: It is important to make a note of the BioProject ID (PRJE#####), as we will need this in later steps.

To link the BioProjects created before (see **Create sequencing data and genome assembly BioProjects**), you will need to create a new XML, which should be identical to the [umbrella.xml](#) above, but this time it should have additional lines detailing the IDs of the “children” (or BioProjects) you wish to add:

umbrella_modified.xml

```
<PROJECT_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
```

```
<PROJECT alias="<affiliated-project>-<tolid_prefix>-study-umbrella--<yyyy-mm-dd>"
center_name="<center name>">
  <NAME><tolid_prefix></NAME>
  <TITLE><scientific name></TITLE>
  <DESCRIPTION>This project collects the sequencing data and assemblies generated for
<scientific name> by the European Reference Genome Atlas (ERGA,
https://www.erga-biodiversity.eu/) for the Biodiversity Genomics Europe project (BGE,
https://biodiversitygenomics.eu/).</DESCRIPTION>
  <UMBRELLA_PROJECT/>
  <RELATED_PROJECTS>
  <RELATED_PROJECT>
    <CHILD_PROJECT accession="PRJEB#####" />
    <CHILD_PROJECT accession="PRJEB#####" />
  </RELATED_PROJECT>
</RELATED_PROJECTS>
</PROJECT>
</PROJECT_SET>
```

We also require an XML with the modify command ([update.xml](#)), which contains the following text:

[update.xml](#)

```
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <MODIFY/>
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

You can then update the Species Umbrella BioProject by using the following command:

```
[12] curl -u Username:Password -F "SUBMISSION=@update.xml" -F "PROJECT=@umbrella_modified.xml"
"https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

Once again, you should receive a receipt, detailing that the Species Umbrella BioProject was correctly modified:

[receipt.xml](#)

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="receipt.xsl"?>
<RECEIPT receiptDate="2023-11-22T08:59:02.441Z" submissionFile="update.xml" success="true">
  <PROJECT accession="PRJE#####"
alias="<affiliated-project>-<tolid_prefix>-study-umbrella--<yyyy-mm-dd>" status="PRIVATE"
holdUntilDate="2025-11-20Z"/>
  <SUBMISSION accession="" alias="SUBMISSION-22-11-2023-08:59:02:256"/>
  <MESSAGES/>
  <ACTIONS>MODIFY</ACTIONS>
</RECEIPT>
```

Finally, to release the project, so that it is published on INSDC and visible to the world, you will need to create the [hold_date.xml](#):

[hold_date.xml](#)

```
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <RELEASE target="PRJE#####" />
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

Where the BioProject ID (PRJE#####") corresponds to that of the Species Umbrella BioProject. As before, submit the XML to ENA using the following command:

[13] `curl -u Username:Password -F "SUBMISSION=@hold_date.xml" "https://www.ebi.ac.uk/ena/submit/drop-box/submit/"`

APPENDIX

1. Create an ERGA Umbrella BioProject

If you wish to link your BioProject to ERGA but find that none of the existing ERGA Umbrella BioProjects are suitable, you will need to create a new Umbrella BioProject. However, before doing this, have a careful look at the currently available ERGA Umbrella BioProjects [here](#).

To register a new Umbrella BioProject, you will need to create two XML files (parts highlighted in yellow need to be changed accordingly):

[submission.xml](#)

```
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <ADD/
    </ACTION>
    <ACTION>
      <HOLD HoldUntilDate="<yy-mm-dd>" />
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

It is good practice to provide a release date for an Umbrella BioProject which does not have any child projects associated to it yet. Any child projects added to the umbrella will have their own release dates independent of the umbrella project.

[umbrella.xml](#)

```
<PROJECT_SET>
```

```
<PROJECT center_name="<center name>" alias="<alias>">
  <TITLE><The title goes here></TITLE>
  <DESCRIPTION><The description of the project goes here></DESCRIPTION>
  <UMBRELLA_PROJECT/>
</PROJECT>
</PROJECT_SET>
```

You can submit these files using the following command:

```
[1] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@umbrella.xml"
"https://wwwdev.ebi.ac.uk/ena/submit/drop-box/submit/"
```

Where the 'Username' is your Webin account and 'Password' is the password to your Webin account. This command will return a [receipt.xml](#) containing a submission ID, the created BioProject ID, and details of whether the submission was a success or not (success=true). Only when you are happy with the result of the submission you can use the [curl](#) command specifying the production service:

```
[2] curl -u Username:Password -F "SUBMISSION=@submission.xml" -F "PROJECT=@umbrella.xml"
"https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

If you want to release the Umbrella BioProject, you can remove the 'HoldUntilDate' in the [submission.xml](#). Alternatively, you can create another XML file, called [hold_date.xml](#), specifying in the new release date and BioProject accession, as below:

```
hold_date.xml
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <RELEASE target="PRJEB####" />
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```

Where the target is the BioProject ID created with command [1]. Once you have this file, you can use the [curl](#) command again as:

```
[3] curl -u Username:Password -F "SUBMISSION=@hold_date.xml" "https://www.ebi.ac.uk/ena/submit/drop-box/submit/"
```

FREQUENTLY ASKED QUESTIONS (FAQs)

1. How should I proceed if I have pooled samples?

Virtual samples are needed whenever you work on samples that are pooled prior to sequencing (e.g. for Iso-Seq, or when different tissues might be used for genome sequencing because one extraction didn't yield enough material). When it is time to submit the data, you need to be able to associate many BioSamples to a single data submission. As ENA does not allow this, their workaround is to create a **virtual sample** that is a list of all the real samples that were combined.

How to create a virtual sample:

1. Login to your Webin account and click 'Register Samples';
2. Choose the closest checklist possible and fill it in using ALL the attributes that samples have in common;
3. Under the column 'sample_description' add a description clarifying that it is a virtual sample, e.g. (please edit as appropriate)
"This sample is a virtual sample of assembled raw reads from multiple physical samples of <organism> genome and is composed of <number> physical samples <ERSXXXXX1, ERSXXXXX2, ERSXXXXX3 etc>."
4. Add the following as an user defined attribute referencing the component samples:
TAG: sample composed of
VALUE: <ERSXXXXX1, ERSXXXXX2, etc.>
5. Upload the completed checklist to your Webin account and submit it.

Make sure you replace all the '<>' brackets with the correct accessions of the samples within the group and any other specified values. Then, when you register your run/experiment/assembly with a sample, just reference this virtual grouped sample.

2. Which sample and corresponding ToL ID should I use for assemblies based on several samples?

We advise to name the assemblies after the ToL ID of the sample used for the generation of the long reads.

3. Is there a specific nomenclature for naming the genome assembly in the manifest file?

No, there is no specific nomenclature. However, we recommend naming the assembly following this nomenclature: ToLID.assembly_type.version_number. So, for example, mHomSap34.pri.1 if the ToLID is mHomSap34, this is a "primary assembly" and version one from this sample. The pri can be replaced with alt/hap1/hap2/mat/pat depending on the type of assembly performed. Use the ID of the sample that gave the long read data.

4. Which coverage should I report in the manifest file necessary for the genome assembly submission?

We recommend reporting the coverage of the long reads used to generate the initial contigs.

5. When submitting the genome assembly, I must submit other files (e.g., FASTA, AGP, etc). Is there a specific nomenclature for naming these files?

No, there is no specific nomenclature. These files are never displayed anywhere in ENA and the files themselves are only used for internal processing by ENA, so you choose the naming yourself.

6. I have obtained a genome assembly from two different BioSamples. Which sample accession should I report in the manifest file?

You are expected to report only one BioSample. We recommend using the BioSample of the sample from which the long reads were produced as this is the underlying sequence.