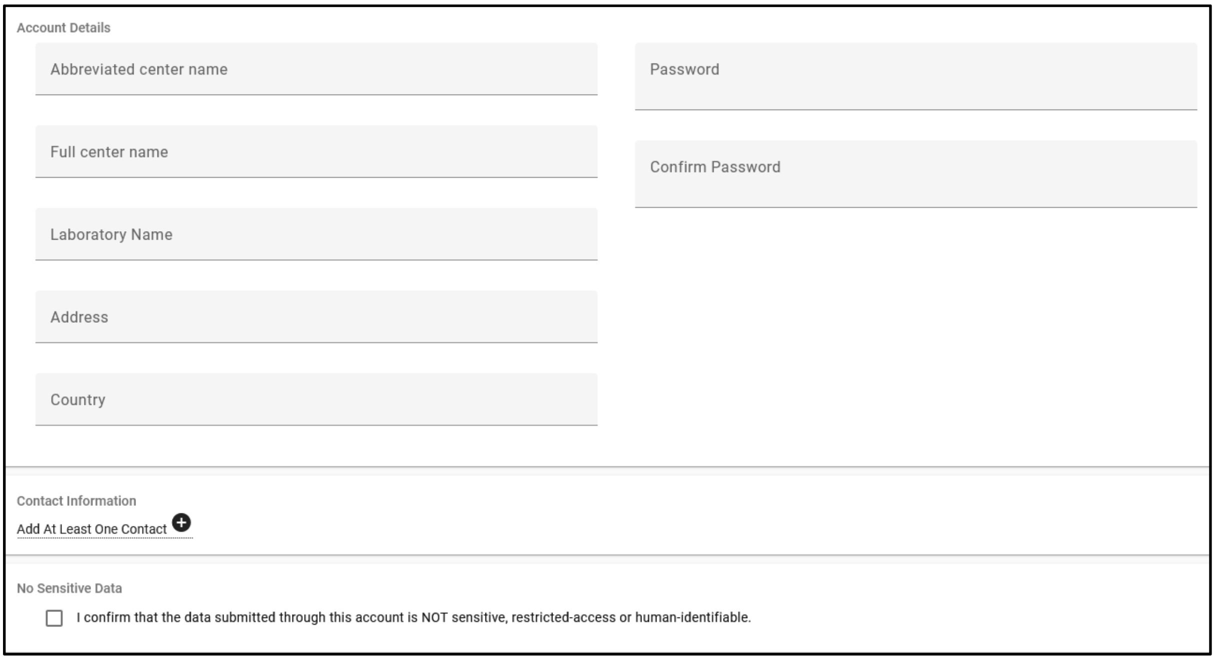
PHENET EMPHASIS ENA European Nucleotide Archive

ENA Data Submission Practical

In this exercise, you will have opportunity to submit a set of read data to the test version of the ENA Interactive Submission service.
You will need a Webin account to use for this exercise. Register it here, with a valid email address:

Webin Submissions Portal



Account Details

Abbreviated center name

Full center name

Laboratory Name

Address

Country

Password

Confirm Password

Contact Information

Add At Least One Contact ⊕

No Sensitive Data

□ I confirm that the data submitted through this account is NOT sensitive, restricted-access or human-identifiable.

Make a note of your account ID and the password.

It will be helpful to review how the metadata model works. You can find an overview of this here:

Metadata Model

Throughout this exercise, you are asked to use the ENA test service, which allows you to make submissions which will be removed after 24 hours. If you're not sure if you're in the test server, you can check the address in your browser which will begin 'wwwdev' rather than 'www':

https://wwwdev.ebi.ac.uk/ena/submit/webin/

If at any point you believe you have accidentally submitted to the real submission, service, please inform the instructor or send a message including accession numbers via ENA support form:

https://www.ebi.ac.uk/ena/browser/support

Content of this Practical

The Scenario

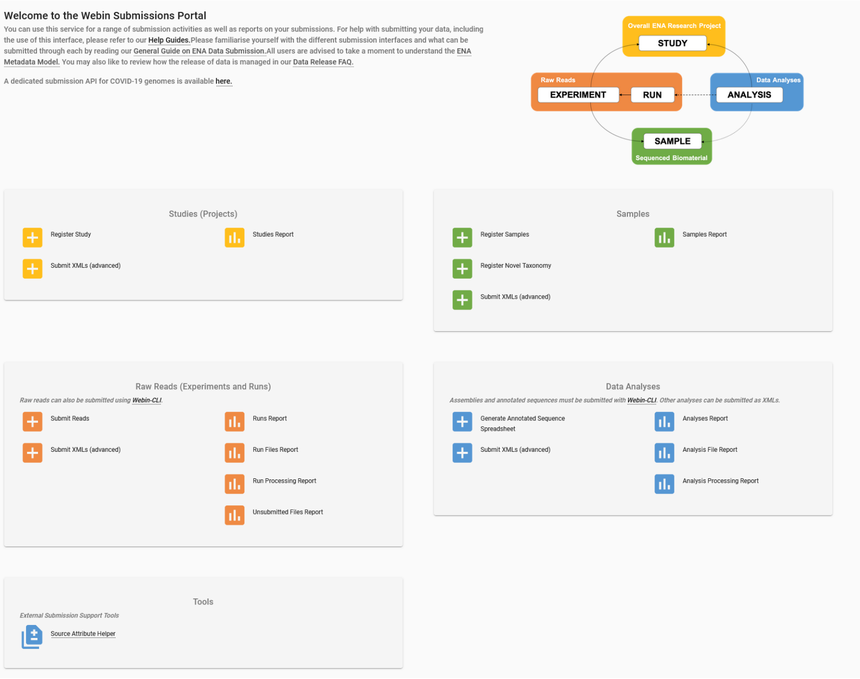
You will be submitting a set of read data derived from barley to the ENA test submission service. This is identical to the production submission service, but submissions are retained for <24 hours. It's therefore perfect for training settings like this, with one exception: it is not linked with BioSamples' equivalent test interface, so we won't be able to use the samples you have registered there on this occasion. For a real submission, you should do so, but in this case we will register a sample directly through ENA.

We will be using a small test data set consisting of 10 barley genotypes that were sequenced using Restriction Digest approach (GBS) within the BRIDGE project (https://doi.org/10.1038/s41588-018-0266-x). You can find it on github: https://github.com/PBR/elixir-fondue-datathon/tree/master/test_data_set/ENA
Download the 10 .fastq.gz files and the metadata_bridge_exercise_reads.tsv and metadata_bridge_exercise_samples.tsv. Make sure to click on the button saying '**download raw file**'.

The Study

To begin, you should register a study. Recall that a study describes the purpose of the work you have done, groups other objects beneath it, and controls when the data becomes public. A study is required for all submissions to ENA.

1. Log in to the Webin test submission service with your account ID and the password:
    https://wwwdev.ebi.ac.uk/ena/submit/webin/



Welcome to the Webin Submissions Portal

You can use this service for a range of submission activities as well as reports on your submissions. For help with submitting your data, including the use of this interface, please refer to our Help Guides.Please familiarise yourself with the different submission interfaces and what can be submitted through each by reading our General Guide on ENA Data Submission.All users are advised to take a moment to understand the ENA Metadata Model. You may also like to review how the release of data is managed in our Data Release FAQ.

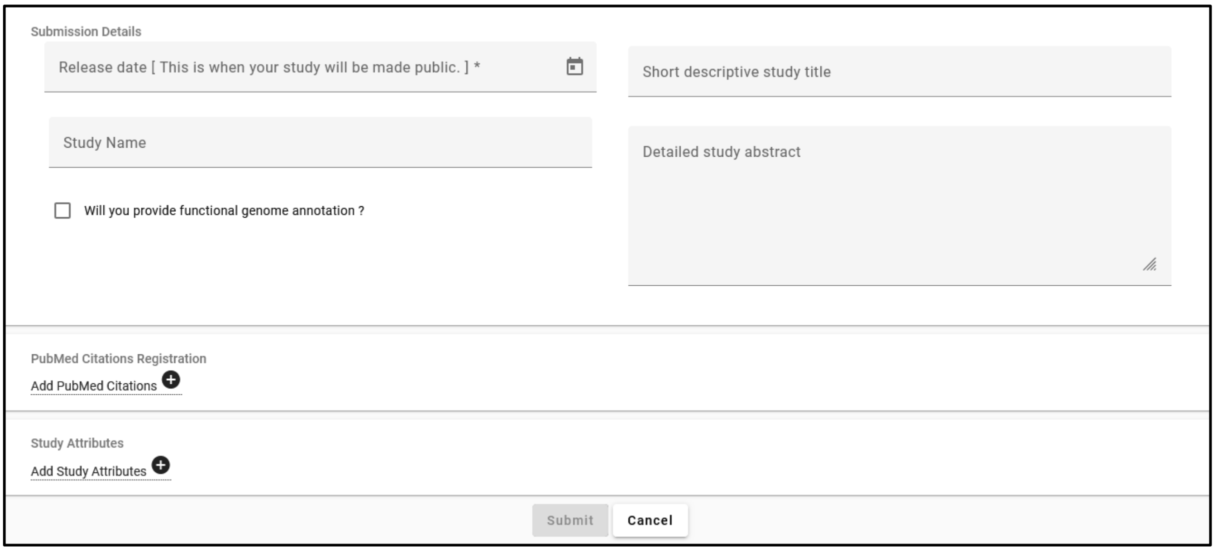
A dedicated submission API for COVID-19 genomes is available here.

Raw reads can also be submitted using Webin-CLI.

Assemblies and annotated sequences must be submitted with Webin-CLI. Other analyses can be submitted as XMLs.

External Submission Support Tools

2. Find the '**Studies (Projects)**' tile and the '**Register Study**' button. Click it to see the study registration interface:



Submission Details

Release date [This is when your study will be made public.] *

Short descriptive study title

Study Name

Detailed study abstract

□ Will you provide functional genome annotation ?

PubMed Citations Registration

Study Attributes

3. The '**Short descriptive study title**' field should be filled in with something brief and meaningful, e.g.:

```
barley_study_2021
```

4. You should take time to provide a descriptive title and informative abstract for your own studies, but these can be edited later if needed. For now, use as your **'Study Name' and 'Detailed study abstract'**:

    GBS Study of Barley from <Your Town/Lab>

5. When you have completed all required fields, click '**Submit**' and then confirm.

6. Now navigate to the '**Studies Report**' tile to see the study you just registered. You might need to refresh the page!
   Make a note of its accession numbers, which will resemble:

   ERP###### and **PRJEB######**

   For a real submission, these would be the numbers you would cite in any publications involving the data

The Sample

The next step is to register the sample, which will give other users essential context for the sequence data you are submitting. The sample describes the source biological material of your sequencing work.
As discussed above, **samples are best submitted through BioSamples**. Since that's not possible with the test interface, we will instead register a sample directly through ENA.

In ENA, samples are required to conform to a checklist of values. Checklists define a set of mandatory and recommended values for a given type of sample. It is recommended that you look at these early and make sure you collect all required metadata items for the type of sample you will be registering.
The selection of available checklists can be browsed at:

https://www.ebi.ac.uk/ena/browser/checklists

1. Return to the Webin Submissions Portal.

2. Find the '**Register Samples**' button under the '**Samples'** tile and click it.

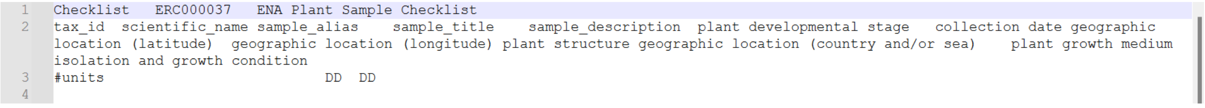
3. You must choose an appropriate checklist of values to be provided for your sample: click '**Download spreadsheet to register samples**' and select '**Other Checklists**' group to browse checklists of this type

4. Select the checklist named '**ENA Plant Sample Checklist**'.

5. Submitters now have the option of including additional fields in their checklist. It is not necessary to include any additional fields but you can take this opportunity to see which fields are included by default, what requirements they have, and what else is available. There are also recommended fields that you should fill, but in the frame of this practical we will **deselect all of them**.

6. Click '**4 - Download spreadsheet template**' and click on the button labelled **'Download TSV template'** when you're ready. This will download the TSV template file to your local computer.

7. Open up this downloaded file and take a look at the structure. (You can open TSV files using any text editor or spreadsheet program.) It should resemble something like this below:



```
1  Checklist   ERC000037   ENA Plant Sample Checklist
2  tax_id  scientific_name sample_alias    sample_title    sample_description  plant developmental stage   collection date geographic
   location (latitude)  geographic location (longitude) plant structure geographic location (country and/or sea)    plant growth medium
   isolation and growth condition
3  #units                           DD  DD
4
```

8. We will first save this template in a new file that we will be working on. Let's call it **'ENA_samples.tsv'.** Each line within this file will represent a single sample that we will submit. For the purpose of this practical we will be submitting 10 samples in total that share many metadata attributes.

9. Starting in **line 4** we will now add the metadata to our samples. We start from left to right (each item separated by the others by tabulator) with the first requested attribute which is '**tax_id**'. That is the NCBI Taxonomy ID for the organism. In this case we are looking for barley (*Hordeum vulgare*), we can look it up using the Taxonomy Browser: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

10. Next item would be the '**scientific_name**'. In our case this is '**Hordeum vulgare**'.

11. The next item is the '**sample_alias**', this will be the sample identifier, for the first sample we will take the prefix of the fastq.gz files: '**HOR 337 BRG**'.

12. After that we have '**sample_title**', this will appear as the title in the sample report later, for the purpose of this practical we will call it: '**BRIDGE PROJECT SSD for HOR 337**'.

13. The next item is '**sample_decription**' and is meant to describe the sample in meaningful details to a human. Therefore we will describe the biological material: '**SSD derived material from a single seed of a single plant belonging to accession HOR 337**'.

14. The next item is '**plant developmental stage**' and should be the developmental stage at the time of sample collection as specified by an ontology term (Plant Ontology). In our case we are looking for the ontology term for 'vascular leaf post-expansion stage' at https://bioportal.bioontology.org/ontologies/PO?p=classes. We do find this under '**PO_0001053**'.

15. The next item is '**collection date**'. This refers to the date the sample was collected with the intention of sequencing, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008. For this practical we will just use the year '**2016**'.

16. The next item is '**geographic location (country and/or sea)**'. The geographical origin of where the sample was collected from, with the intention of sequencing, as defined by the country or sea name. Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html). For this practical this will be '**Germany**'.

17. The next two items will be ‘**geographic location (latitude)**’ and ‘**geographic location (longitude)**’. The values should be reported in decimal degrees and in WGS84 system (also known as the GPS coordinates). For latitude this will be ‘**+51.816**’ and longitude ‘**+11.283**’. These are the GPS coordinates of the IPK Gatersleben in the middle of Germany where this experiment was conducted.

18. The next item is ‘**plant structure**’ and refers to the name of the plant structure that the sample was obtained from (as specified by an ontology term, such as Plant Ontology). We will be using https://bioportal.bioontology.org/ontologies/PO?p=classes again and look for ‘leaf’. We do find this under ‘**PO_0025034**’.

19. The next item is ‘**plant growth medium**’. Specification of the media for growing the plants or tissue cultured samples, e.g. soil, aeroponic, hydroponic, in vitro solid culture medium, in vitro liquid culture medium. This should be represented by an ontology term as specified by the ENVO (Environment Ontology). We will be using https://ontobee.org/ontology/ENVO to find the appropriate term, in our case: Soil. We do find this under ‘**ENVO_00001998**’.

20. The last item is ‘**isolation and growth condition**’. This refers to a publication reference in the form of pubmed ID (pmid), digital object identifier (doi) or url for isolation and growth condition specifications of the organism/material. For this practical this was specified in a publication and we will write the doi in this field ‘**https://doi.org/10.1038/s41588-018-0266-x**’.

21. We now repeat steps 9-20 for the other nine samples. In the frame of time, you can take a look at how this should look like in the file ‘**metadata_bridge_exercise_samples.tsv**’. Copy all the information in this file into your own ‘**ENA_samples.tsv**’ and save it.

22. By the time you have completed this, your sample will be well-annotated and understandable to people finding it in the database later. Go back to the Webin Submission Portal and select ‘**Upload filled spreadsheet to register samples**’. Find your ‘**ENA_samples.tsv**’ and upload with the button labelled ‘**Submit completed spreadsheet**’.

23. Now navigate to the 'Samples' tab and select ‘**Samples Report**’ to see the sample you just registered. If you find any errors in the metadata being displayed you can still change this using the ‘Action’ button and selecting ‘Edit sample XML’.
    Make a note of its accession numbers as you will need these later:

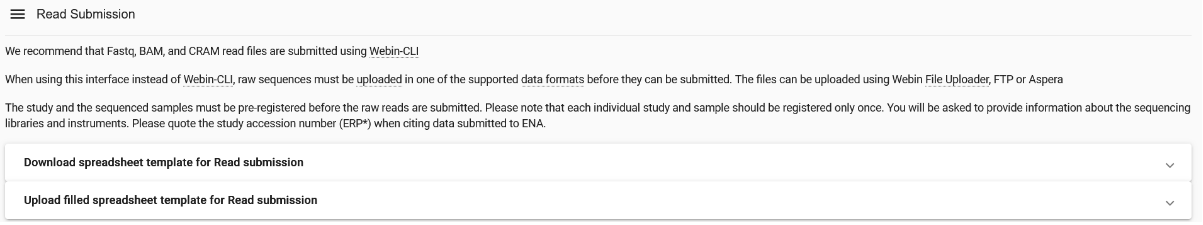
    ERS####### and **SAMEA#######**

The Read Data

Now that study and sample metadata have been registered, it is time to submit the read data you have produced.

1. Return to the Webin Submissions Portal.

2. Find the '**Submit Reads**' button under the '**Raw Reads (Experiments and Runs)'** tile and click it.



Read Submission

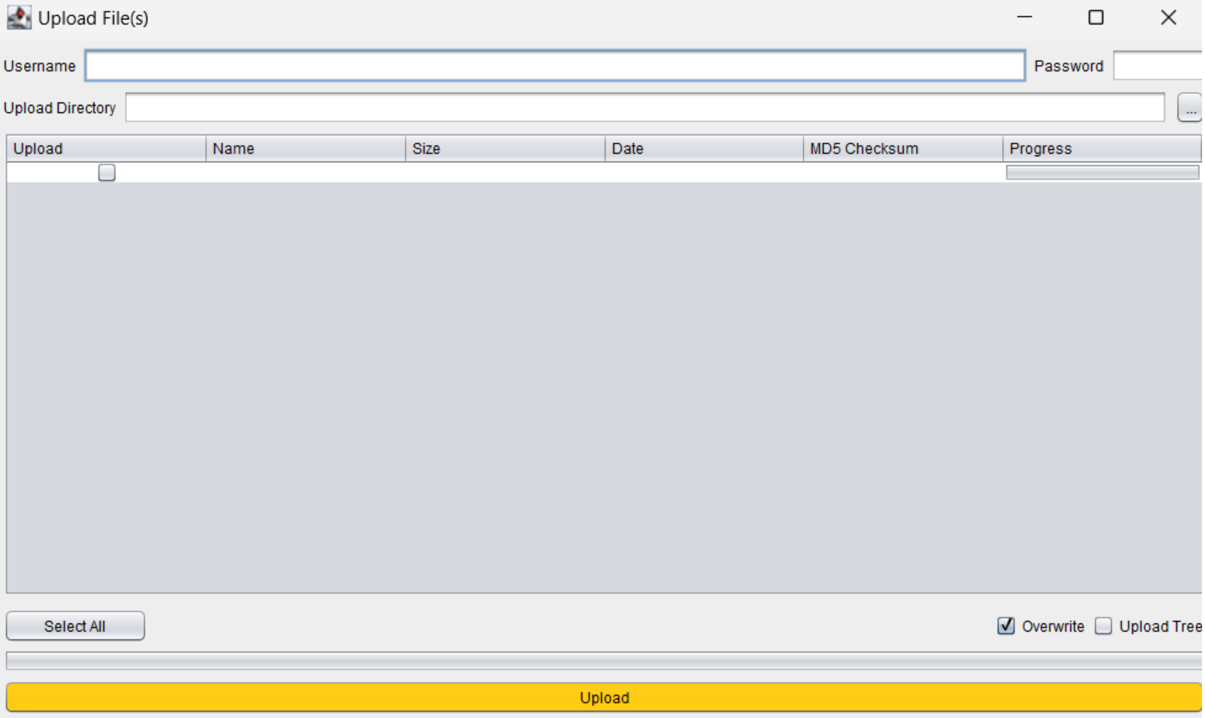
We recommend that Fastq, BAM, and CRAM read files are submitted using Webin-CLI

When using this interface instead of Webin-CLI, raw sequences must be uploaded in one of the supported data formats before they can be submitted. The files can be uploaded using Webin File Uploader, FTP or Aspera

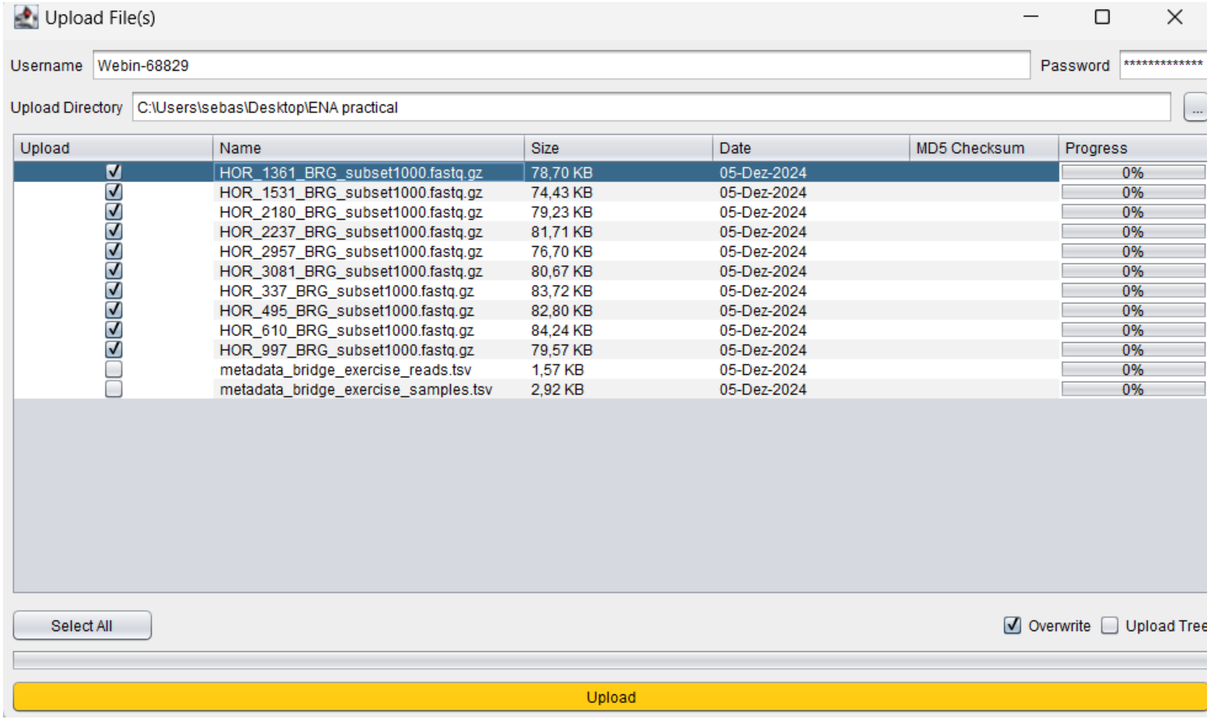
The study and the sequenced samples must be pre-registered before the raw reads are submitted. Please note that each individual study and sample should be registered only once. You will be asked to provide information about the sequencing libraries and instruments. Please quote the study accession number (ERP*) when citing data submitted to ENA.

Download spreadsheet template for Read submission
Upload filled spreadsheet template for Read submission

3. We need to upload our fastq.gz files now to the ENA file system, to do so we can use different routes. For the purpose of this practical we will use the java 'File Uploader'. We need a working java web start installation. https://www.openlogic.com/openjdk-downloads (JDK version 8). After installation (and possibly a restart) click the File Uploader link and open it using the java web start application. It should look like this:



4. Input your username 'Webin-XXXXX' and password, click on the '...' button to select the file directory on your computer where the fastq.gz files are stored.



Upload File(s)

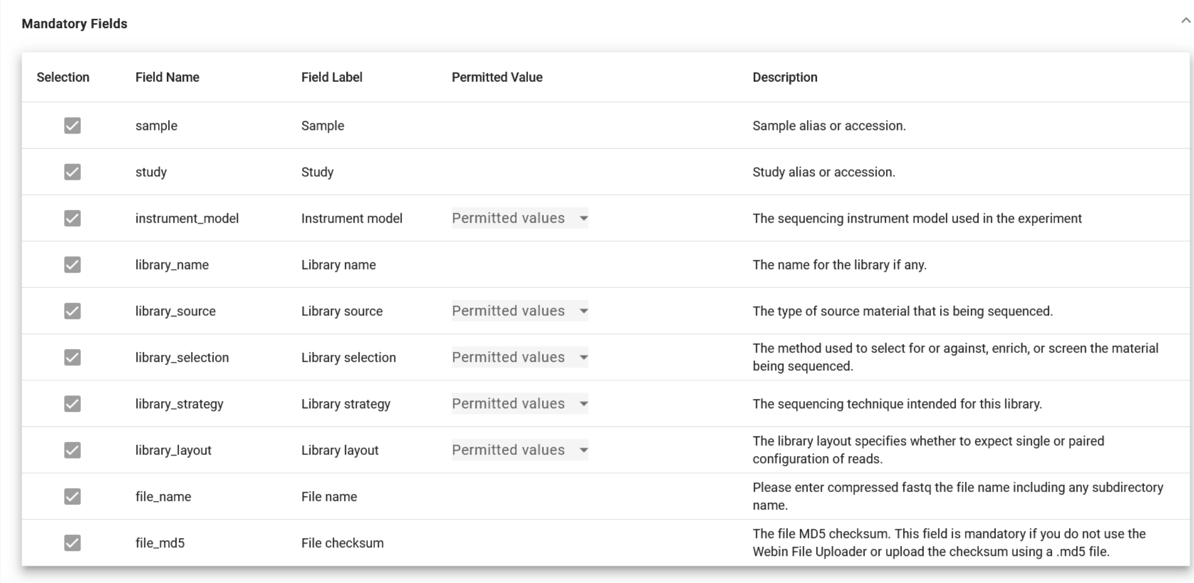
Username Webin-68829 Password *************

Upload Directory C:\Users\sebas\Desktop\ENA practical

Upload	Name	Size	Date	MD5 Checksum	Progress
☑	HOR_1361_BRG_subset1000.fastq.gz	78,70 KB	05-Dez-2024		0%
☑	HOR_1531_BRG_subset1000.fastq.gz	74,43 KB	05-Dez-2024		0%
☑	HOR_2180_BRG_subset1000.fastq.gz	79,23 KB	05-Dez-2024		0%
☑	HOR_2237_BRG_subset1000.fastq.gz	81,71 KB	05-Dez-2024		0%
☑	HOR_2957_BRG_subset1000.fastq.gz	76,70 KB	05-Dez-2024		0%
☑	HOR_3081_BRG_subset1000.fastq.gz	80,67 KB	05-Dez-2024		0%
☑	HOR_337_BRG_subset1000.fastq.gz	83,72 KB	05-Dez-2024		0%
☑	HOR_495_BRG_subset1000.fastq.gz	82,80 KB	05-Dez-2024		0%
☑	HOR_610_BRG_subset1000.fastq.gz	84,24 KB	05-Dez-2024		0%
☑	HOR_997_BRG_subset1000.fastq.gz	79,57 KB	05-Dez-2024		0%
□	metadata_bridge_exercise_reads.tsv	1,57 KB	05-Dez-2024		0%
□	metadata_bridge_exercise_samples.tsv	2,92 KB	05-Dez-2024		0%

5. Select the 10 fastq.gz files that you want to upload and click the button on the bottom called ‘**Upload**’. Note the md5 checksum that you get when you upload the data, this will be needed for the Read Submission.

6. Return to the web browser where you are still having the Read Submission page open. Click on the button that is called ‘**Download spreadsheet template for Read submission**’. For this practical we will be using ‘**Submit single reads using Fastq files**’, see below the mandatory fields:



Mandatory Fields

Selection	Field Name	Field Label	Permitted Value	Description
☑	sample	Sample		Sample alias or accession.
☑	study	Study		Study alias or accession.
☑	instrument_model	Instrument model	Permitted values ▾	The sequencing instrument model used in the experiment
☑	library_name	Library name		The name for the library if any.
☑	library_source	Library source	Permitted values ▾	The type of source material that is being sequenced.
☑	library_selection	Library selection	Permitted values ▾	The method used to select for or against, enrich, or screen the material being sequenced.
☑	library_strategy	Library strategy	Permitted values ▾	The sequencing technique intended for this library.
☑	library_layout	Library layout	Permitted values ▾	The library layout specifies whether to expect single or paired configuration of reads.
☑	file_name	File name		Please enter compressed fastq the file name including any subdirectory name.
☑	file_md5	File checksum		The file MD5 checksum. This field is mandatory if you do not use the Webin File Uploader or upload the checksum using a .md5 file.

7. Some of the fields (instrument_model, library_source, library_selection, library_strategy and library_layout) only have a few permitted values, to familiarize yourself with them click on the permitted values drop down menu. Click on the ‘**Download TSV template**’ button.

8. Open the TSV template like you did with the samples TSV file and save a copy of it as ‘**ENA_reads.tsv**’.

9. We will now fill this with information about the sequencing reads. This will in principle done like outlined below.

```
sample: enter the BioSample identifier here*

study: enter the Study accession here*

instrument_model : Illumina HiSeq 2500

library_name : barley_library_1

library_source : GENOMIC

library_selection : Restriction Digest

library_strategy : GBS

library_layout : SINGLE

file_name : enter the file name here* (.fastq.gz)

file_md5 : enter the file md5 here*
```

10. In the frame of time you can use the prefilled file titled ‘**metadata_bridge_exercise_reads.tsv**’ and fill the sample and study fields with your information. Save the file as ‘**ENA_reads.tsv**’

11. Back on the web browser select ‘**Upload filled spreadsheet template for Read submission**’ and upload ‘**ENA_reads.tsv**’

12. If you manage to complete your submission, visit the 'Raw Reads (Experiments and Runs)' tab and select ‘**Runs Report**’ to review your submission.

Possible Bugs

Unfortunately, the Webin submission system is not entirely without bugs.

If you receive an error saying ‘File Not Found’:

1. Click the 'Download spreadsheet' button
2. Open the file this downloads
3. You will find the information you entered is present, but the column for your filenames is empty; fill in the filenames as appropriate
4. Return to the submission interface and use the 'Upload spreadsheet' button to provide your amended submission, which should now be accepted

A Note On File Uploads

If you were able to complete the submission, you saw that the final step involved specifying the names of the sequence file to be submitted:

HOR_1361_BRG_subset1000.fastq.gz

All Webin submission accounts come with a space on the ENA server to upload files to. In most cases, submitters must upload their files to this before they are able to submit.

If you're interested in learning more about how to upload your files as well as how to prepare them for submission, please see the below page:

https://ena-docs.readthedocs.io/en/latest/submit/fileprep/upload.html