# 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**](https://www.ebi.ac.uk/ena/submit/webin/login)

****

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**](https://ena-docs.readthedocs.io/en/latest/general-guide/metadata.html)

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/**](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**](https://www.ebi.ac.uk/ena/browser/support)

## Content of this Practical

1. [**The Scenario**](#_xnfckmkd7r4o)
2. [**The Study**](#_watil3nq2wi9)
3. [**The Sample**](#_ugqnt37zr4fl)
4. [**The Read Data**](#_wusc13ltxdbg)
5. [**A Note On File Uploads**](#_l29d6b27ih0x)

## 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](https://github.com/PBR/elixir-fondue-datathon/blob/master/test_data_set/ENA/metadata_bridge_exercise_reads.tsv) and [metadata\_bridge\_exercise\_samples.tsv](https://github.com/PBR/elixir-fondue-datathon/blob/master/test_data_set/ENA/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/>



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



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

**barley\_study\_2021**

1. 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>**

1. When you have completed all required fields, click '**Submit**' and then confirm.
2. 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. 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.
2. 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>
3. Next item would be the ‘**scientific\_name**’. In our case this is ‘**Hordeum vulgare**’.
4. 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**’.
5. 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**’.
6. 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**’.
7. 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**’.
8. 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**’.
9. 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**’.
10. 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.
11. 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**’.
12. 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**’.
13. 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**’.
14. 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.
15. 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**’.

1. 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.



1. 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:



1. 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.



1. 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.
2. 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:



1. 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.
2. Open the TSV template like you did with the samples TSV file and save a copy of it as ‘**ENA\_reads.tsv**’.
3. 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\*

1. 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**’
2. Back on the web browser select ‘**Upload filled spreadsheet template for Read submission**’ and upload ‘**ENA\_reads.tsv**’
3. 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>