How to submit NGS sequencing data to the European Nucleotide Archive (ENA)

Interfaculty Bioinformatics Unit
University of Bern

Overview of required information

You need to provide information on the

1) **Study** = title and abstract for the study

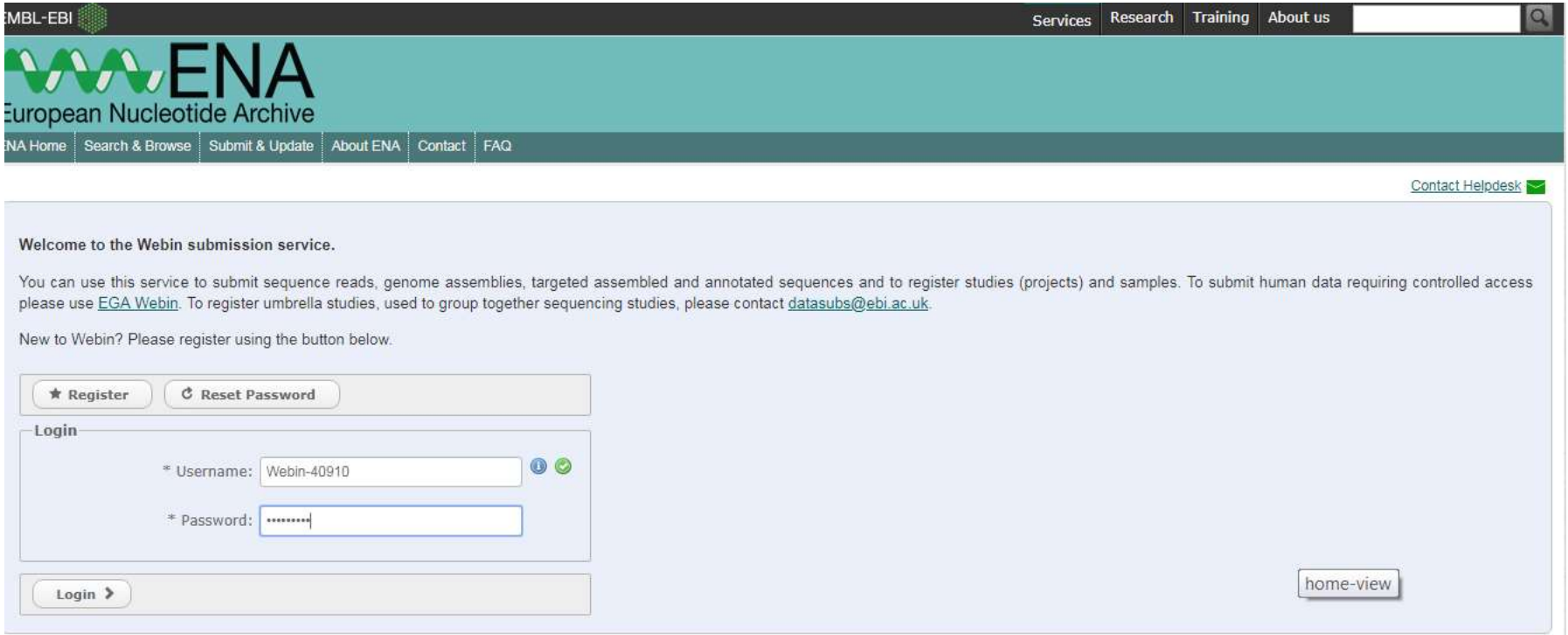
2) **Samples** = information on the biological samples

3) **Experiment** = details on how the sequencing was done

Steps 2 and 3 can either be done manually for each sample or you can download spreadsheets which you can complete and upload. This tutorial explains only this second option which is much easier as soon as you have multiple samples.

1) Create an account and log in

https://www.ebi.ac.uk/ena/submit/sra/#home



2) Select «New Submission» tab

2) Pre-register the study

Please select the type of submission you would like to make:

- ○ Submit sequence reads and experiments
- ◉ **Register study (project)**
  - Register your Study to begin data submission to ENA.
  - Read here for information on how to submit a Study to ENA.
- ○ Taxonomy Check/Request
- ○ Register samples
- ○ Submit other assembled and annotated sequences [formerly EMBL-Bank]

Complete the required fields and press «Submit»

Please specify the release date of your study:
This is when your study will be made public.

23-Dec-2020

Please provide a short name for the study:

Please provide a short descriptive title for the study: (*)

Please provide an abstract to describe the study in detail: (*)

Please provide attributes to add a deeper description of the study:

Tag	FieldType

Add

Please provide PubMed IDs of publications you want to associate with the study:
(numeric value)

PubMed IDs

Add

For genome assembly projects only: In this study, will you provide functional genome annotation? (*)
PLEASE ANSWER WITH YES IF YOU HAVE ANNOTATION: Locus tag prefixes are only associated to studies providing functional genome annotation.

- ○ Yes
- ◉ No

<< Previous

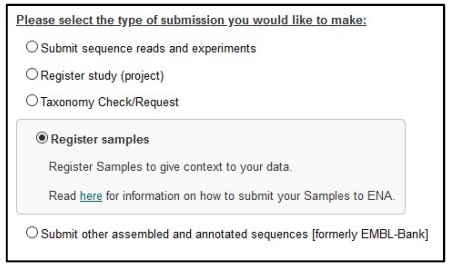
Submit

Restart Submission

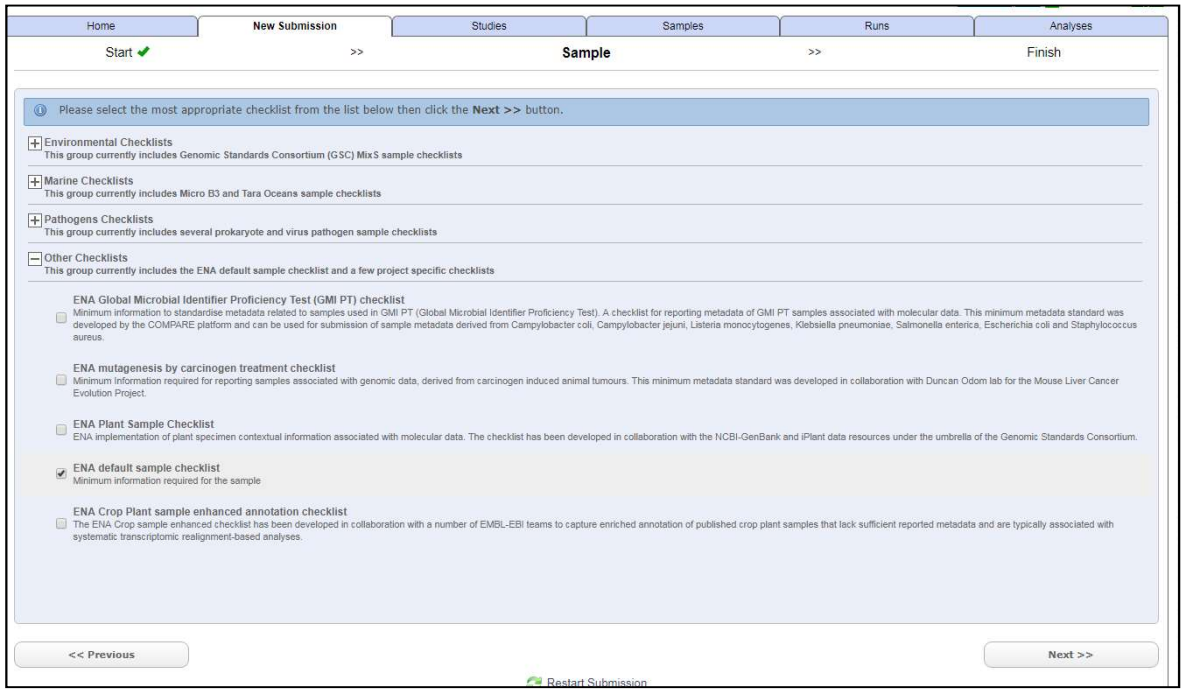
3a) Register the samples - Download template

This page explains how to download a template for the sample spreadsheet. You can skip this step if you already have a template, e.g. from a previous submission

Back on the «New Submission» page:



Select Checklist >



(May vary between projects)

3a) Register the samples - Download template

Here, you can specify various optional fields that you could provide if available, e.g. sex, tissue type etc.

Contact Helpdesk Webin-40910 Logout

Home | New Submission | Studies | Samples | Runs | Analyses

Start >> Sample >> Finish

Please select any additional optional fields. Mandatory and recommended fields are selected by default. You may add any of the optional fields. You may also add custom fields.

Filter fields...

Add your own custom fiel + Add

Part and developmental stage of organism

- cell_type - *optional*
  cell type from which the sample was obtained
- dev_stage - *optional*
  if the sample was obtained from an organism in a specific developmental stage, it is specified with this qualifier
- germline - *optional*
  the sample described presented in the entry has not undergone somatic genomic rearrangement as part of an adaptive immune response; it is the unrearranged molecule that was inherited from the parental germline
- tissue_lib - *optional*
  tissue library from which sample was obtained
- tissue_type - *optional*
  tissue type from which the sample was obtained

0 of 30 fields selected

+ Expand − Collapse

When you have selected the fields click the **Next >>** button to begin entering your data. Alternatively, download a template spreadsheet using the **Download Template Spreadsheet** button. Once you have filled the spreadsheet please restart the submission process and upload the spreadsheet using the **Upload Completed Spreadsheet** button.

Download Template Spreadsheet

<< Previous

Next >>

Restart Submission

You may want to go to the next page to look up the taxonomic ID for your species. You will need this in the sample description file (next slide)

3b) Register the samples - Fill in sample spreadsheet

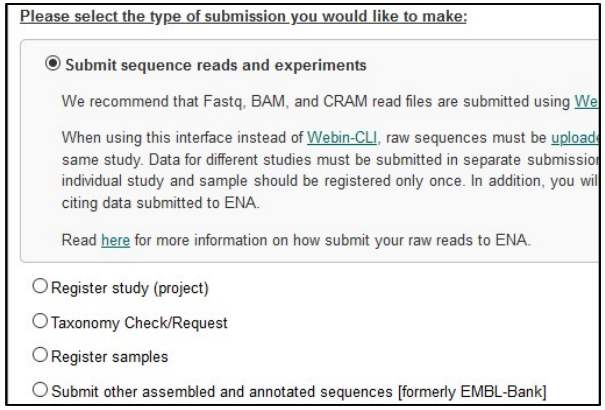
#checklist_accession	ERC000011				
#unique_name_prefix					
sample_alias	tax_id	scientific_name	common_name	sample_title	tissue_type
T1	10090	Mus musculus	house mouse	Triaka ESRP1 mutant mouse (Esrp1<m1Btlr> or Esrp1<Triaka;Triaka>)	intestinal epithelial cells
T2	10090	Mus musculus	house mouse	Triaka ESRP1 mutant mouse (Esrp1<m1Btlr> or Esrp1<Triaka;Triaka>)	intestinal epithelial cells
T3	10090	Mus musculus	house mouse	Triaka ESRP1 mutant mouse (Esrp1<m1Btlr> or Esrp1<Triaka;Triaka>)	intestinal epithelial cells
T4	10090	Mus musculus	house mouse	Triaka ESRP1 mutant mouse (Esrp1<m1Btlr> or Esrp1<Triaka;Triaka>)	intestinal epithelial cells
WT5	10090	Mus musculus	house mouse	wildtype (C57BL/6J)	intestinal epithelial cells
WT6	10090	Mus musculus	house mouse	wildtype (C57BL/6J)	intestinal epithelial cells
WT7	10090	Mus musculus	house mouse	wildtype (C57BL/6J)	intestinal epithelial cells
WT8	10090	Mus musculus	house mouse	wildtype (C57BL/6J)	intestinal epithelial cells

- The sample alias must match the fastq filename. For example, if your fastq file is called T1_R1.fastq.gz, the correct sample alias would be T1.

- Save file as .tsv or .txt

4a) Register the experiments - Download template

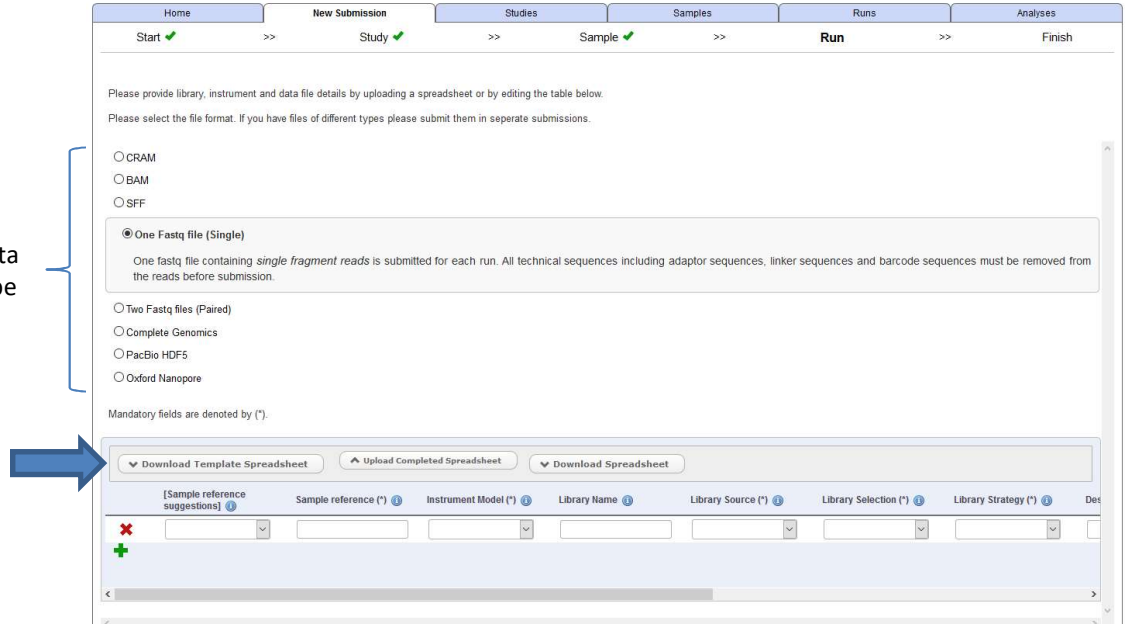
This page explains how to download a template for the experiment spreadsheet. You can skip this step if you already have a template, e.g. from a previous submission



Select the study you preregistered in step 2

Next >>

Skip >>



Data type

Select the correct data type and download template

4b) Register the experiments - Fill in spreadsheet

sample_alias	instrument_model	library_source	library_selection	library_strategy	library_construction_protocol	insert_size	forward_file_name	forward_file_md5	reverse_file_name	reverse_file_md5
T1	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	350	T1_R1.fastq.gz	a122c9eafb484e97073aae05fedf5620	T1_R2.fastq.gz	dce8368de0ca1ddab8fa49283ad213d7
T2	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	350	T2_R1.fastq.gz	d8f9e1e3fc417a4c9a44152c99ebbed8	T2_R2.fastq.gz	ff138de09b6f94f3f0c4c497bab43aa3
T3	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	340	T3_R1.fastq.gz	826ad86910dee4108dfc1bb8b14e6938	T3_R2.fastq.gz	f35f238559e74bda181a17aa832e644e
T4	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	340	T4_R1.fastq.gz	b66152260d7fd7c3a5481f58c43e5573	T4_R2.fastq.gz	7e4a97c04286e81ab9a5adc7149ac318
WT5	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	370	WT5_R1.fastq.gz	a319acfd4410e1f00b35bb3f84dea12b	WT5_R2.fastq.gz	93b9078910be3075ad6f6ec78b9f29dd
WT6	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	360	WT6_R1.fastq.gz	a7133b26d8cb5466e2464501e2fe3f12	WT6_R2.fastq.gz	3cf40b077a1b0b7ca5883cdd9a2d09e5
WT7	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	350	WT7_R1.fastq.gz	750bbb87b9fbc23bf73f10901e098a97	WT7_R2.fastq.gz	0435b510d3435440e33abfefe51dbf3c
WT8	Illumina HiSeq 2500	TRANSCRIPTOMIC	Oligo-dT	RNA-Seq	TruSeq Stranded mRNA	360	WT8_R1.fastq.gz	66b467abe1cea822ada15756249a8a57	WT8_R2.fastq.gz	bf1feac97431d3755222c6ace9c14335

Files and associated md5sums

Important: The file names must match the sample_alias in the samples spreadsheet (step 3b)

For sequencing data produced by NGSP Bern and analysed by IBU:

Columns in brown: This information is available in the LIMS.

Columns in red: Contact IBU for this information.

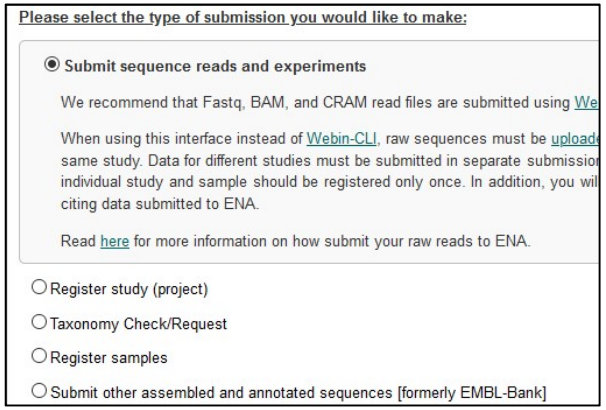
5) Upload the fastq files

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

Wait until the sample upload is complete before going to the next step!

For sequencing data analysed by IBU: Contact IBU to do the file upload for you. We will let you know as soon as it is complete and you can continue with the final step of the submission.

6) Complete the submission



Select the study you preregistered in step 2

Next >>

^ Submit Completed Spreadsheet

Upload the **SAMPLE** spreadsheet from step 3

Next >>

^ Upload Completed Spreadsheet

Upload the **EXPERIMENT** spreadsheet from step 4

Submit

The samples listed in the spreadsheets should be automatically matched to the uploaded fastq files as long as the sample IDs are consistent