EBI Metagenomics portal: Submitting metagenomics data to the European Nucleotide Archive

Hubert Denise [1]

- DNA & RNA
- Beginner
- 0.5 hour

This tutorial provides you with a step-by-step guide for submitting metagenomics data to the European Nucleotide Archive (ENA) in order for it to be analysed by the EBI Metagenomics resource.

Learning objectives:

- Be able to submit your metagenomics data to the ENA
- Find out why EBI Metagenomics requires you to archive your data in the ENA
- Evaluate why it is important to provide accurate contextual metadata with your metagenomic sequence data

What is Metagenomics?

Metagenomics is an umbrella term, covering a range of study types that use high-throughput DNA sequencing to characterise microbial systems. These include whole-genome shotgun [2] (WGS)-sequenced metagenomic and metatranscriptomic [3] studies, as well as amplicon [4]-based approaches which target specific marker genes (often rRNA [5] genes).

The EBI Metagenomics portal

The EBI Metagenomics portal [6] is a free to use analysis service for metagenomics data. In order to have your data analysed by EBI Metagenomics, your sequence data have to be archived in the European Nucleotide Archive [7] (ENA).

This tutorial will explain why you have to follow this procedure and will then guide you through the process of archiving your metagenomic data in the ENA.

Find out more

Learn more about the EBI Metagenomics portal and ENA in our Train online quick tours.
Why should I submit my metagenomics data to ENA?

If you only stored your sequence data on your computer, the wider scientific community would not be able to access it, therefore it couldn't be used for data mining and discovery. This is why most scientific journals, and many funding bodies, require you to submit your sequence data to a public repository. By doing so, the sequences will be archived and will receive recognised accession numbers.

In addition, Next-Generation Sequencing (NGS) output files have become very large, and many institutions cannot ensure safe long-term storage of such large datasets. The large file sizes also complicate data sharing (e.g. with collaborators), as it can take a very long time to transfer data.

**How long would it take to share your data?**

Imagine that for your project on the human gut metagenome, you sequenced 130 samples and obtained 130 FASTQ files. Each of these files contained on average 28,000,000 sequences and had a compressed size of ~2.5 Gb.

The full data set would require ~ 320GB of storage in compressed form. This represents about 400 CDs, 68 DVDs or a third of the storage capacity of a modern hard drive.

From your computer, it would take well over eight hours to transfer the data through a fast ethernet connection (100 Mbit/sec) but less than 2 hours to download it from ENA using their FTP server.

**Submitting pre-publication data**

The ENA is a specialist resource with capacity to securely store and share sequencing data. Pre-publication data can be kept confidential until published. Once public, data is made accessible for downloading and mining, benefiting the whole scientific community.

**What type of metagenomic data can I submit to ENA?**

The following types of data can be submitted to the ENA and are subsequently analysed by EBI Metagenomics:

- whole genome sequences;
- metatranscriptomic sequences;
- amplicon metagenomic sequences;
- genome assembly data.

ENA accepts data in all NGS formats including:

- Illumina;
- LS454;
- SOLiD;
- Ion Torrent;
- Oxford Nanopore.

FASTA files are accepted, although formats containing quality information such as: FASTQ, SFF, SOLiD and FAST5 are preferred as they allow EBI Metagenomics to perform a complete quality control, ensuring results are uniformly analysed.

Find out more about different types of NGS file formats [16].

**Preparing your samples for submission**

- In order to prepare your samples for submission, you must trim the sequencing adapters [17] and, eventually, de-multiplex [18] the reads.
- In the case of patient data, it is your responsibility to ensure that all human sequences are removed and no sensitive information (allowing identification of patients) is provided.
- Sequencing data for EBI Metagenomics analysis must be submitted with its associated metadata [19].

**What are metadata and why are they so important?**

Metadata are the in-depth, controlled description of the sample that your sequence was taken from. Essentially, the ‘what, where, how, and when’ of your study from collection to sequence generation, plus contextual data such as environmental conditions (latitude, longitude, temperature) or clinical observations.

It is essential to describe your samples with such data in order to carry out a meaningful comparison with other samples or projects. For this to happen, all submitters must use a common set of terms to ensure that the vocabulary used to define metadata is constrained.

**Using a controlled vocabulary to describe metadata**

To help describe data, ENA uses sets of controlled vocabularies such as the Environmental Ontology [20] (ENVO), which can be accessed on the ENVO page [21] (Figure 1).
Surely there is only one way to describe where a sample of lake water comes from!

Here are some of the terms that could be used to describe it:

- fresh water lake - lake - inland sea
- pond - lentic water - millpond
- inland body of water - loch - lentic habitat
- lakelet - still water - mere ....

Can you see the confusion that could arise by not using a controlled vocabulary? However, by using ENVO: freshwater lake (ENVO:00000021), there is no ambiguity.

The classification is hierarchical and would be:

'Environmental: Geographic feature: Hydrographic feature: Water body: Lake: Freshwater lake' for these samples.
Providing the required metadata for your submission to ENA

In order to support submission, ENA has implemented the 'Minimum Information about any (X) Sequence' (MiXS [22]) standard developed by the Genomic Standards Consortium [23] (GSC).

The GSC has defined a number of checklists corresponding to different biomes. Each checklist contains all mandatory fields required to accurately provide the metadata describing a project associated with this biome [24]. You can also provide optional parameters.

Submitting pre-publication data

As mentioned earlier, you can keep the pre-publication data you have submitted to ENA confidential for up to two years in the first instance. If you want EBI Metagenomics to analyse data held confidentially, you have to provide an explicit agreement that EBI Metagenomics can access the data, in line with ENA's data access policies. You can provide this agreement while registering on the EBI Metagenomics portal. If you are already a registered ENA user, you can provide your agreement while submitting your data via the EBI Metagenomics portal [25].

The EBI Metagenomics analysis results for confidential data are held in the private section of the website - you will need to log in (using the same credentials as on the ENA website) to access them.

Considerations before submitting metagenomic data to ENA

Before submitting your metagenomic data to ENA, you should ensure that you have all the required information to fill out the relevant GSC MiXS checklist [22] for your submission.

All checklists require:

- sampling date and location*;
- sample nature;
- the amount of material sampled;
- investigation type.

(*: if you do not want to indicate the precise location for any reason, please indicate at least the country coordinates).

You must also ensure that data are compressed prior to submission. CRAM [26] and gzip [27] formats are accepted by ENA.

The submission process

ENA have developed an online tool, Webin [28], to help you submit your data. While efforts have been made to make the submission straightforward, it still requires time.

The step-by-step guide described in the next section will take you through the whole submission process (which takes approximately 1-2 hours depending on whether or not you have used Webin before). However, it is possible to split the submission into four separate steps [29] that could be performed at different times, if more convenient. This method will be explained after the complete procedure.

How to submit metagenomics data to ENA in one session
This section will show you step-by-step how to submit your data using ENA's webin submission service.

1. Log into the ENA webin tool [30]

You can access the ENA Webin submission service from the ENA ‘Submit & Update' page or via the EBI Metagenomics ‘Submit data’ pages (Figure 2).
Submitting and updating data

We offer a number of services through which data (including updates) can be submitted to the European Nucleotide Archive (ENA). These technologies provide options appropriate for the scale and frequency of submission, the expertise and capacity of the submitter and the nature of the data to be transferred. The choices below lead users most directly to the appropriate submission route.

- Submit read data
- Submit assembled sequence and/or annotation (No partial or complete assemblies)
- Submit genome assemblies (contigs/scaffolds/chromosomes)
- Email ENA helpdesk

Submit & Update
- Data formats
- Uploading data files
- Reads
- Sequences
- Genome assemblies
- Taxonomy
- Sample checklists
- Environmental
- Epigenomic
- Species BARCODE
- Metadata model
- Register submission account
- Programmatic XML submissions
- Programmatic tabulated submissions

Popular
- Submit and update
- Sequence submissions
- Genome assembly submissions

Submit data

We provide a free service for submission of raw metagenomics sequence data and associated metadata to the European Nucleotide Archive (ENA) and analysis by EBI Metagenomics.

If you have data that you wish to have analysed, you need an ENA submitter account that has been registered with EBI Metagenomics. This allows us to track your submitted data and ensures that we have consent to access it for analysis.

With a valid ENA submitter account, you can submit your data directly using the ENA WebIN tool, which will help you describe your metadata and upload your sequence data.

Once your reads are uploaded to the ENA, the EBI Metagenomics team will assess and perform the analysis, which is done in several steps. You will receive an email once the analysis starts and another when the analysis of all samples is complete (view figure for more details of the submission and analysis process). The analysis time is dependent on the number of samples submitted and requests by other submitters at the time. If your samples are private, you will need to log in to the EBI Metagenomics homepage to be able to view the results of the analysis.
Figure 2 Access to the ENA Webin submission service from the ENA (A) and EBI Metagenomics (B) websites.

Starting a new submission

2) To start a completely new submission, select 'Submit sequence reads and experiments' (Figure 3).

You will then be guided through the process which, sequentially, involves:

a) uploading your sequence data;

b) providing study level information;

c) choosing a MiXs sample checklist;

d) providing sample metadata;

e) linking the sequencing data to the sample metadata.

We will go through these steps in more detail over the next few pages.
Uploading the sequence data

Navigate to the folder containing your sequence data in the Webin Uploader window, then select the files (these must be compressed in gzip or CRAM format) you would like to upload (Figure 4).
Figure 4 Using the ENA Webin sequence Uploader to upload your sequencing data.

Providing study level information

Creating a new study in ENA Webin

Providing study level information: after running the Webin Uploader, you will be redirected to the ‘Study submission’ page where you will create your study and enter study-level metadata (Figures 5 and 6).
Figure 5 Creating a new study in ENA Webin.

Providing study-level metadata in ENA Webin
Figure 6 Providing study-level metadata in ENA Webin.

Choosing a MIxS sample checklist

Choose the most relevant MIxS sample checklist [22] for your sample(s). The checklist will define the parameters you have to provide to accurately describe your samples and the environment from which they were collected.
Figure 7 Select one of the 15 GSC MiXS checklist available in ENA Webin.

Providing sample metadata

Organism details
Entering sample metadata starts with providing basic details which are common to all your samples (Figure 8).

- You will then be asked for the ‘Organism details’ related to your sample(s) - this is where you will indicate that they are metagenomic samples. Use the text search box to look for the metagenome taxonomy identifier related to your sample (Figure 8). NB: this field has to be filled even if your samples are not host-associated.
- For host-associated samples, do not enter the host organism but instead use the corresponding metagenome (e.g. ‘human metagenome’ for human metagenomic samples). For non-host-associated...
samples, there are defined metagenomes for most cases (e.g.: 'soil metagenome', 'marine metagenome', 'freshwater metagenome', 'air metagenome' or simply 'metagenome').

- Finally, in the bottom right-hand corner of the page, you can provide the metadata relevant to the checklist you chose - mandatory fields are labelled with a red icon. You can expand each section by clicking on the '+' sign in front or using the 'Expand' button.

NB: To ensure metadata are consistent, drop-down menus are provided where appropriate.

Adding samples

- Once you have filled in the checklist attributes, click 'Next' to go to the samples page. You can add as many samples as you like. Each one will require a unique sample title. The checklist metadata will be automatically associated with all these samples. You can edit the attributes for a particular sample on this page by selecting the corresponding sample checkbox.
- To make this process quicker, particularly if you have a lot of samples, ENA Webin allows you to download the attributes for all the samples in a tab-separated file. The file can easily be edited in software tools such as Excel and LibreOffice Calc. After editing, you can upload it back into ENA by clicking 'Previous' - this takes you back to the GSC MiXS checklist ([Figure 7](#)). Be aware that the date format may have been altered if you used Excel.
- Once you have updated the metadata for your samples, click 'Next' to associate your samples to your sequencing run data.

Linking the sequencing data to the sample metadata

In order to link your sequence data to the sample metadata, you have to provide information about the sequencing that took place. This can be provided manually (as show in Figure 10) or by uploading a pre-filled spreadsheet for all your samples. The spreadsheet template is available to download on the page.

Once you have completed this step, click on the 'Submit' button to complete your submission. Confirm that you would like to submit - then you're done!
How to submit data in four separate steps

The previous section demonstrated how to submit your data in one go - a process that can take some time. However, Webin also allows you to submit your data in four separate (shorter) steps. These can be completed individually, meaning that you won't lose any data if you log out of Webin in between the steps.

1) Upload your sequencing data as indicated in Figure 4 [32].

2) Create a study (Figure 11, shown below). Click on 'Register study (project)' from the 'New Submission' tab of the Webin tool. Then follow the procedure described in Figures 5 and 6 [33].

3) Enter your sample details and metadata as described in Figures 7 to 9 [31]. Start a Webin session and select 'Register samples' (Figure 11). Take note of the 'Unique Name Prefix' for each of your samples. It might be easier to use the 'Download template' button available on the 'Sample' page.

4) When you are ready to complete your submission, restart a Webin session and select 'Submit sequence reads and experiments' as indicated in Figure 3 [34].

- You should now see the study which you created listed in the 'Study' section. Select it by clicking on the left-hand side check box then click 'Next'.
- As you previously created your samples, click 'Skip' on the 'Sample' page.
- You will be redirected to the association page described in Figure 10 [35]. Select the format of your files. Then create the association by entering the 'Unique Name Prefix' (see step 3 above) in the 'Sample reference' text box.
- Fill in the other information as described in Figure 10 and choose the corresponding file. We recommend that you use the downloadable spreadsheet when you have several samples to associate with sequencing data.

Once completed, click 'Submit' and confirm that you want to complete your submission.
Summary

- EBI Metagenomics is a dedicated resource to analyse metagenomics data.
- Metagenomic sequence data and their descriptive metadata [19] have to be archived in ENA before EBI Metagenomics can analyse them.
- The submission of metagenomics data to ENA involves four steps:
  - uploading metagenomic sequence data;
  - creating an ensemble of samples, referred to as a study;
  - providing metadata by completing a biome [24]-related checklist;
  - linking samples and metadata to the metagenomic sequence data.
- The submission can be performed in one session or in separate sessions.

Quiz: EBI Metagenomics submission

| Questions: | 6 |
| Attempts allowed: | Unlimited |
| Available: | Always |
| Pass rate: | 75 % |
Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on submitting metagenomics data

- ENA has a dedicated helpdesk that can be contacted by clicking on the 'Contact helpdesk' link located on the top right hand corner of all Webin pages (Figure 3) or by datasub[at]ebi.ac.uk (emailing ENA).
- Guidance for all types of submission can also be found on the ENA 'Submit and Update page [36]'. Additionally, guidelines for environment submissions are also available here [37].
- metagenomics-help[at]ebi.ac.uk (EBI Metagenomics support) is also available.

Recommended reading


Contributors

Hubert Denise [1]
EMBL-EBI
Bioinformatician, Metagenomics

Hubert Denise is a Bioinformatician in the Metagenomics group at the European Bioinformatics Institute in Cambridge, UK. He joined EMBL-EBI in 2012. He has a PhD in Molecular Biology and Biochemistry and has been involved in several research projects in the areas of infective diseases and comparative genomics.
Alex Mitchell is co-ordinator for the InterPro and EBI Metagenomics databases at EMBL-EBI. He obtained his DPhil in pharmacology from the University of Oxford, and was previously employed as a molecular biologist at the Institute of Psychiatry. He moved to the University of Manchester to work on protein sequence analysis and functional classification, before joining EMBL-EBI in 2011.
[34] https://www.ebi.ac.uk/training/online/course/metagenomics-portal-submitting-metagenomics-data-e/how-submit-metagenomics-data-ena/starting-
[36] http://www.ebi.ac.uk/ena/submit/read-submission
[37] http://www.ebi.ac.uk/ena/submit/environmental-submissions
[38] http://europepmc.org/abstract/MED/26582919
[39] https://www.ebi.ac.uk/training/online/trainers/mitchell