## Practical guide to amplicon nucleotide sequence uploads to the European Nucleotide Archive (ENA) for the UK Crop Microbiome Cryobank (UKCMCB) project

04/03/2025

sue.jones@hutton.ac.uk, nicola.holden@sruc.ac.uk

### Introduction

This document details how the amplicon fastq formatted sequence files were uploaded to ENA for the UKCMCB project; after we had registered our UKCMCB project with ENA **(PRJEB58189)**. The UKCMCB project has its own data catalogue (agmicrobiomebase.org) which links multiple data types together (including the sequence data) using a unique microbiome (m) identifier. Further details on some aspects of the upload process are available on the UKCMCB project GitHub site: https://github.com/HuttonICS/agmicrobiomebase.

Generic information on ENA file uploads is available in documentation on the ENA website here: https://ena-docs.readthedocs.io/en/latest/submit/fileprep/upload.html.

Our UKCMCB project upload process has two main parts summarised in Figure 1.

(1) Sequence file (fastq) upload to ENA using the ENA ftp site after calculating and storing the md5 hashes for these files

(2) Creation and upload of ENA checklists: first samples and then fastq data. For samples there are both plants (child) and soils (parent), and we needed to show the relationship between the two using a custom made BioSamples spreadsheet.

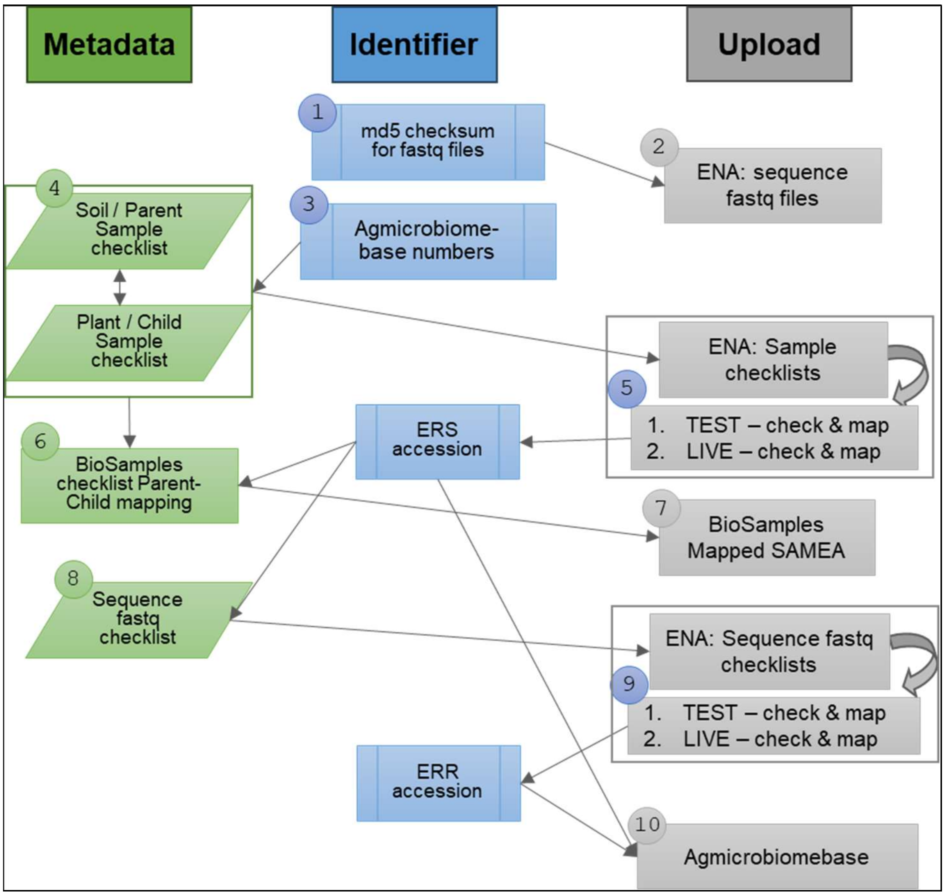


***Figure 1**: Flowchart for ENA amplicon sequence upload process for the UKCMCB project*

## Important tips to remember before starting ENA sequence uploads

1. Sequence uploads to any public repository takes time, and requires care, organisation and attention to detail. Meta data for sequence files is essential for data to be found within a repository and re-used. You can never have enough metadata for samples.
2. ENA checklist spreadsheets are edited in MS Excel. However, to upload the spreadsheet they need to be saved as a TSV file. Take care in the file conversion and ensure that extra column headings that may have been added to help complete the spreadsheet are not included in the TSV file.
3. Be consistent in naming files and directories if you have multiple samples. It is important to have the same file name structure for each crop and soil sample, otherwise it's hard to keep track of the essential files as the upload process progresses.
4. ENA has a TEST and a LIVE database.
    a. URL TEST: https://wwwdev.ebi.ac.uk/ena/submit/webin/
    b. URL LIVE: https://www.ebi.ac.uk/ena/submit/webin/
5. All upload steps should be tested on the TEST database first. Once data had been uploaded to the LIVE database, data files and associated spreadsheets cannot be removed. The uploads can be cancelled but the data within the cancelled spreadsheets is still visible to subsequent uploads. This means that unique identifies cannot be reused making for additional complexities.
6. All the checklists and files were created in a common MS Sharepoint / Teams site. Everything was first created in a TEST folder area (e.g. ENA-Upload-16S-TEST) and tested using the ENA TEST portal site only. Once success on the TEST portal was confirmed, they were then copied across to a LIVE directory (e.g. ENA-Upload-16S-LIVE) and upload conducted using the ENA LIVE portal. However, no files were submitted to the ENA LIVE portal without a two people checking the spreadsheets and TEST data confirmations.

### Scope

The upload process is illustrated with our 16S amplicon files (Spring Wheat (SW) 16 S). The same procedure was used for each crop and is also applicable for other amplicon (e.g. ITS-2) or sequence files (e.g. shotgun metagenome). Obviously, the directory structures described in this upload process are specific to our High-Performance Cluster (HPC), but their inclusion gives an idea of how we organised our project data.

### Step 1. Preparation of fastq files and checksums on local HPC

On HPC the fastq files are located within a directory structure related to plates which originated from the sequencing service. Copy fastq files for wheat into a separate upload directory:
/202109_bbr_microbiome/ENA-Upload-16S/WH-wheat/fastq
Create md5sum checklist

```
md5sum * > md5sum-WH-16S.txt
```

This generates a single file with md5sum numbers required for the fastq file checklist (step 8)

### Step 2. Upload of fastq files to ENA site

Log onto ENA ftp site from the directory with fastq files: requires webin user name and passwd. Using "prompt" removes the need to press return after each upload and mput copies multiple local files.

```
ftp webin2.ebi.ac.uk
webin username, passwd
prompt
mput *.fastq.gz
```

### Step 3: Agmicrobiomebase unique identifier

For the agmcirobiomebase data catalogue we created a series of project partner spreadsheet templates unique to the project. These enabled the different project partners to submit their own data to the data catalogue. An example of one of these spreadsheet templates is available as part of the agmicrobiomebase catalogue. This section refers to these unique partner spreadsheet templates.

Generate unique microbiome sequence ms000000 numbers in agmicrobiomebase template. This is a prerequisite for the ENA BioSamples checklist submission.

i. open SRUC-Hutton partner template document
ii. navigate to correct crop / soil combination pre-populated with microbiome (m) numbers
iii. automatically generate microbiome sequence (ms) numbers corresponding to selected sequence type (e.g. 16S amplicon)
iv. save a date-stamped copy of the file – use this for the later stage of uploading to agmicrobiomebase (Step 10).

### Step 4: ENA Samples checklist creation and submission

Preparation of samples checklist to register samples in ENA:
We have generated crop plant spreadsheets (child) and soil spreadsheets (parent) for upload.

| PlantChild Checklist Dir | SoilParent Checklist Dir |
|---|---|
| PlantChild_CO_Bulk<br>PlantChild_SW_Wheat<br>PlantChild_FB_Bean<br>PlantChild_SB_Barley<br>PlantChild_SO_Oats<br>PlantChild_OR_Rape<br>PlantChild_SU_Beet | SoilParent_CL_BO<br>SoilParent_CL_YO<br>etc |

### Plant/Child sample checklist

The checklists used are in the environmental section of the ENA register samples checklist

- Template = GSC MIxS plant associated - ERC000020

The No-Plant Control, Bulk Soil is described with the same template type, within the same group.

Detailed process: Populate the ENA samples checklist:

i. navigate to Agmicrobiome base numbering file and obtain specific entries for a single crop set, e.g. SW
ii. enter taxID for host plant using Latin name
iii. enter new msXXX numbers and Pot ID sample name
iv. enter Sample description
v. Reformat date from MS xls default to ISO form YYYY-MM-DD
vi. Save file as .tsv

### Soil/Parent sample checklist

The same process is carried out for the soil samples, although each checklist only has a single entry to describe the soil metadata.

- Template = GSC MIxS soil - ERC000022

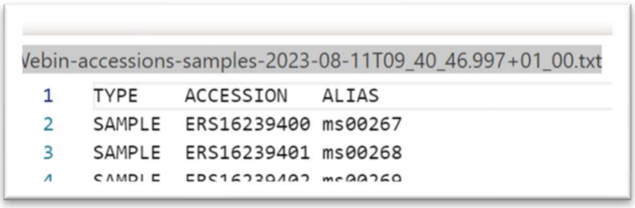
### PCR Control sample checklist

The sample process is carried out for PCR controls, with the default ENA checklist

- default sample checklist - ERC000011 for positive PCR controls

### Step 5: ENA Samples checklist upload and check

Uploading a completed samples spreadsheet generates ERS numbers that are added to the fastq file spreadsheet (see Step 6). The upload process is in two parts, first into the TEST portal to check that the checklist is correct, then to the LIVE portal. The upload process generates a txt file with the ERS numbers.



```
Vebin-accessions-samples-2023-08-11T09_40_46.997+01_00.txt
TYPE     ACCESSION   ALIAS
SAMPLE   ERS16239400 ms00267
SAMPLE   ERS16239401 ms00268
SAMPLE   ERS16239402 ms00269
```

*Figure 2: Example output text file from samples spreadsheet upload.*

Detailed process:

i. Upload the checklist into the TEST portal
ii. Download the Webin ERS accession number file and xml receipt. Save to TEST dir
iii. Confirm the accessions match the msXXX alias numbers as expected
iv. If confirmed, then upload checklist into the LIVE portal
v. Download the Webin ERS accession number file and xml receipt. Save to LIVE dir
vi. Confirm the accessions match the msXXX alias numbers as expected

### Step 6: Parent-Child relationship

Once the uploads are completed (Step 5), we can then map the Parent (soil) and Child (crop) samples to describe which Parent samples relate to which Child samples.
Use the ERS and corresponding SAMEA numbers to map to each other.
The No-Plant Control, Bulk Soil has the same type of relationship to the source Soil/Parent checklist

### Step 7: Upload Parent-Child SAMEA numbers

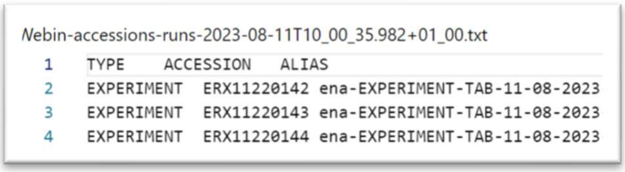
Submit this map to BioSamples. For our project this required a custom spreadsheet. ENA sample submission generates SAMEA BioSample numbers that were used in the parent/child mapping spreadsheet. This custom spreadsheet is available as part of the agmicrobiomebase catalogue

### Step 8: ENA fastq checklist creation

The fastq checklist is uploaded into the Webin section 'runs-experiment' to register raw read files into ENA. To prepare the fastq file checklist use the template for read submission "*Submit paired reads using two Fastq files*".
A read submission checklist is needed for each crop. The ERS numbers generated in Step 5 are added to this spreadsheet. The fastq checklist spreadsheet also requires the forward and reverse file fastq file names and the md5sum values. The md5dsum values were generated in Step 1.

### Step 9: ENA fastq checklist upload and check

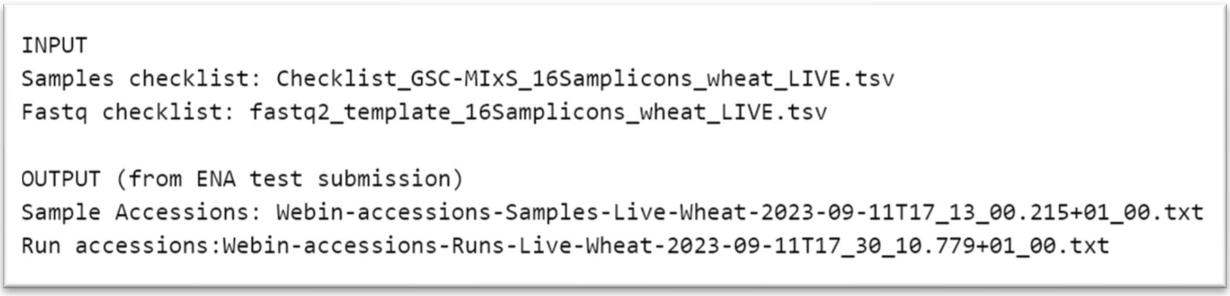
Uploading the fastq raw reads spreadsheet to ENA generates a txt file with ERX numbers.



```
Nebin-accessions-runs-2023-08-11T10_00_35.982+01_00.txt
TYPE       ACCESSION    ALIAS
EXPERIMENT ERX11220142  ena-EXPERIMENT-TAB-11-08-2023
EXPERIMENT ERX11220143  ena-EXPERIMENT-TAB-11-08-2023
EXPERIMENT ERX11220144  ena-EXPERIMENT-TAB-11-08-2023
```

*Figure 3: Example output text file from raw reads spreadsheet upload.*

For each crop we have generated four essential files as part of the upload to the LIVE ENA database:



```
INPUT
Samples checklist: Checklist_GSC-MIxS_16Samplicons_wheat_LIVE.tsv
Fastq checklist: fastq2_template_16Samplicons_wheat_LIVE.tsv

OUTPUT (from ENA test submission)
Sample Accessions: Webin-accessions-Samples-Live-Wheat-2023-09-11T17_13_00.215+01_00.txt
Run accessions:Webin-accessions-Runs-Live-Wheat-2023-09-11T17_30_10.779+01_00.txt
```

*Figure 4: Example of four key ENA/Webin text files generated after the fastq raw reads spreadsheet upload.*

Detailed process:

i. Map the ERS numbers to the fastq checklist using the accession files generated in Step 5. Do not simply copy/paste – it is critical to ensure that the correct ERS accession numbers correspond to the correct file names. The only way to do this is via the msXXX alias number. Use an appropriate script or reproducible/scalable method for mapping
ii. Once the mapping has been validated, clean checklist sheet to a .tsv extension and upload into ENA TEST portal.
iii. Download Webin ERX and ERR accession number file and xml receipt and save to TEST dir.
iv. Confirm the ERR accessions map as expected, check all the way back to msXXX alias number
v. If confirmed, then upload fastq raw checklist into LIVE portal
vi. Download Webin ERX and ERR accession number file and xml receipt and save to LIVE dir.
vii. Confirm the accessions match the msXXX alias numbers as expected

### Step 10: AgmicrobiomeBase partner template upload

The ENA accession numbers (ERS, ERX, ERR) are then mapped back to the SRUC-Hutton partner template for upload to Agmicrobiomebase. Note: the template upload replaces all current content, so the complete dataset needs to be uploaded each time.

Detailed process

i. start with new partner template that has NO previously uploaded information
ii. copy all the new information for the following fields:
    - Microbiome Sequence Id = mXXX
    - Microbiome Id (Alias) = msXXX
    - Sequence type = 16S / ITS / Shotgun metagenome
    - Pot Id V1: ignore – *legacy field*
    - Pot Id V2 = 6-letter code
    - Sample type = rhizosphere / soil / control
    - Study Accession Identifier = PRJ

- Sample Accession Identifier = ERS
- Experiment Accession Identifier = ERX
- Run Accession Identifier = ERR

The partner agmicrobiomebase templates are automatically integrated as part of the creation of the PowerBI data visualisation that appears on the agmirobiomebase.org data webpage (https://agmicrobiomebase.org/data/).