Uploading sequence files to SRA data database (using ftp)

Note: This process can take ages. Your submission is autosaved as you go along the tabs so you can exit. You incomplete submission should then appear on your sra submissions page. If you are logged in, all your submissions (part completed or completed) can be found here: <https://submit.ncbi.nlm.nih.gov/subs/sra/>

*You can log into NCBI using you Google account to avoid making a new account you’ll never remember the password to!*

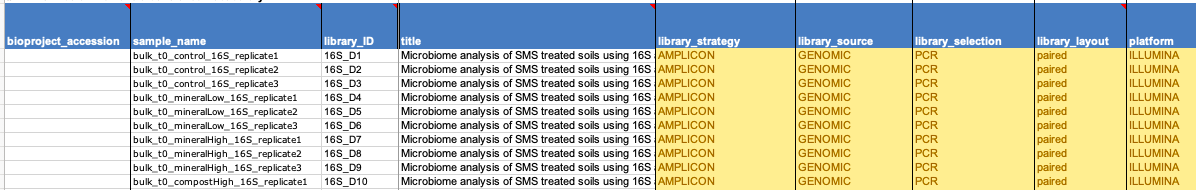
## Set up

There are 2 files needed to during the submission process (apart from sequence files):

1. SRA\_metadata file – information on sequencing strategy & file names
2. MiMarks survey file – information on sample collection & metadata

Here are some tips for getting these two files ready & correct otherwise the whole process can be needlessly painful.

### SRA metadata file:



* This gives information on actual sequencing, namely what fastq files are associated with what sample name and how they were sequenced.
* The SRA metadata file can be found here:

<https://www.ncbi.nlm.nih.gov/core/assets/sra/files/SRA_metadata_SP.xlsx>

*tip: hover over red triangle in upper corner of column headers for information*

* **Bioproject\_accession** column: leave empty for now.
* **Library\_ID** column: each entry must be unique
* **Sample\_name**: while it doesn’t bring an error on this file if not unique, it will when you make survey file later. So make each entry unique eg adding replicate1,2 or 3 if replicates were used to make them unique (as in example image above)
* **Title**: choose short title to describe project
* **Design description**: most important info on experiment (primers used; no of samples etc)
* **Filename**: For paired reads R1 fastq file name can go under column header ‘**filename’** and R2 fastq file names can go under column ‘**filename2’**

*Tip: the easiest way to get all your fastq file names without doing by hand:*

* + Put all fastq files to be uploaded in one directory,
  + install (or copy) mothur.exe into the directory
  + in your terminal, navigate to that directory (using cd nameOfDir)
  + run mothur (./mothur)
  + then run command:

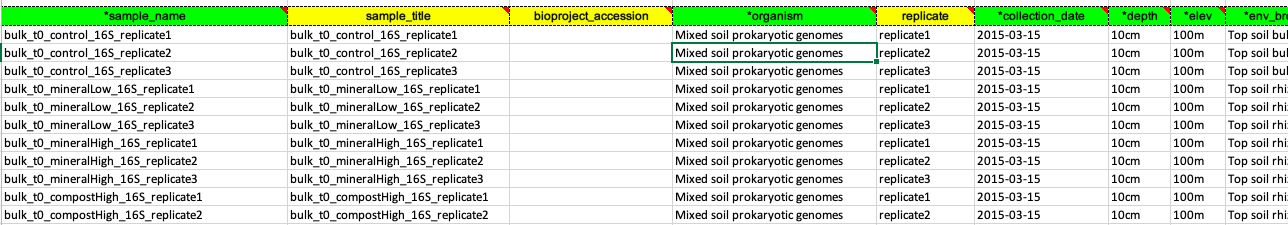
make.file(inputdir=., type=fastq, prefix=stability)

* + This will output a file called stability.files.
  + Ignore the first column and 2nd and 3rd columns should be a list of all the paired reads mothur found in your directory. Copy and paste these columns into ‘filename’ and ‘filename2’ columns of SRA\_metadata sheet.

When you reach the ‘Sample Type’ tab of the sample submission process, you will be asked what kind of sample and sequencing was done. This will help you end up with the right survey file to input information about each of your samples. For a 16S or 18S rRNA soil data set, this will be a MiMarks.survey.soil file, where MiMarks is for marker genes (ie 16S or ITS amplicon datasets). For metagenome data, a MiMs file will be downloaded instead.

### MiMarks survey file

Here is example of the top of a completed MiMarks survey file



The purpose of this file is to give the sampling or experimental conditions associated with each sample name.

Things that usually bring errors:

1. All green columns must be completed
2. The list of samples put under ‘sample\_name ’ (and ‘sample\_title’ - just put same name in each) must exactly match with ‘sample\_name’ in SRA metadata file that you already made
3. Each row of a the survey file must be unique. Ie two diff samples cant have exactly the same information across all the columns
   1. For example: if cDNA and DNA were done for same samples, reflect that in the sample name, then in column ‘organism’ put ‘mixed genomes’ for DNA samples or ‘mixed transcriptomes’ for cDNA samples.
   2. E.g: if ITS and 16S were both done on same samples, put Prokaryotic genomes in column ‘organism’ for 16S and put ‘Fungal genomes’ for ITS.
   3. **TIP: if you have biological/technical replicates where the rows really are the same for say 3 samples (3 replicates), insert a column called ‘replicates’ after the ‘organism’ row and include replicate1, replicate2, replicate3 (as shown column 5 in example file image above)**

## Uploading files

Once above steps are completed, you will end up at tab 7 which is where files are uploaded. In general this will be quicker using a compressed folder containing all your sequence files. To do this (compress them) go to the folder containing all your sequence files (eg fastq files) and tar them.

For example in ubuntu, open the terminal and navigate (cd) to the folder below your folder (directory) containing all your sequence files. Use tar to compress:

tar -czvf MyFastqFiles.tar.gz /FolderOfFastqFiles

MyFastqFiles.tar.gz = name I want to give my compressed tar.gz file

/FolderOfFastqFiles = name of dir/folder that contains all my fastq files.

-c: **C**reate an archive.

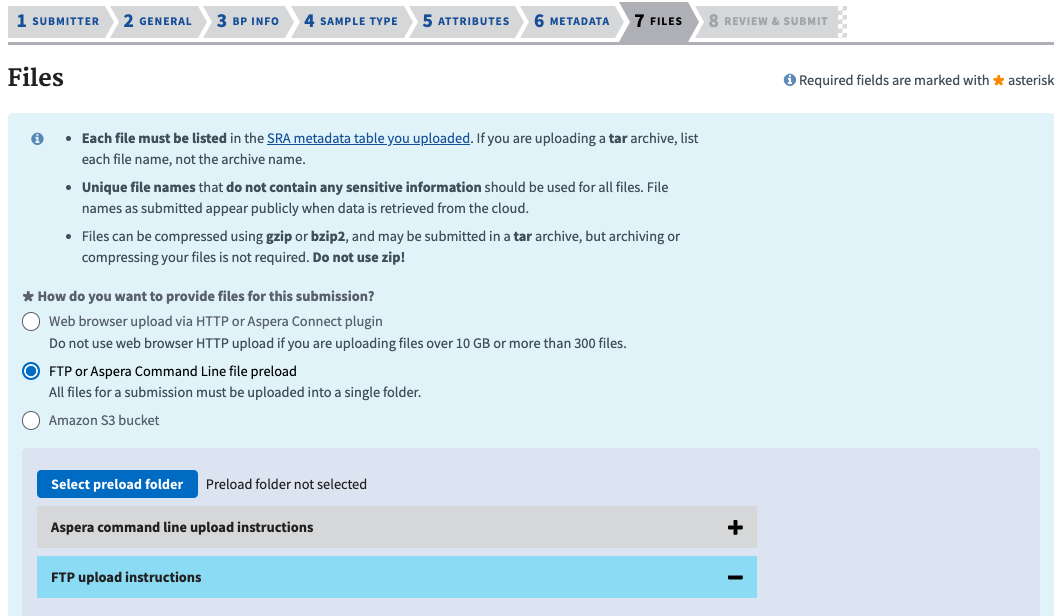
-z: Compress the archive with g**z**ip.

-v: Display progress in the terminal while creating the archive, also known as “**v**erbose” mode. The v is always optional in these commands, but it’s helpful.

-f: Allows you to specify the **f**ilename of the archive

Note: As it says in the instructions on the submission page, do not use zips.

The submission page will look like this:

****

Below is detailed the only way I could get the file upload to work (for file >2 GB)

When you get to Files page, chose FTP or Aspera command line file preload (middle option)

Then click the + sign beside ‘FTP upload instructions’ to open up that menu. A list of instructions will appear

We’ll go through those instructions after some set up steps.

## Uploading using ftp at the command line

First we need to get ftp sorted

* Windows and ubuntu, ftp should be installed by default
* Mac: have to install ftp as doesn’t come with mac anymore. Can install with the following command in terminal:

brew install tnftp

Now to uploading files:

1. open terminal, use cd to navigate to the directory containing your files for upload (ie the folder containing your fastq.tar.gz file made in earlier steps)
   * + if not in the right directory the upload won’t work as it won’t be able to find your data
2. Invoke ftp using following command (in unix anyway, windows might be different):

ftp -i

1. Now we will more or less follow the instructions that appear on the submission page:
2. Establish an FTP connection using ‘open’ and the credentials below:  
   open ftp-private.ncbi.nlm.nih.gov

Once you do this, it will ask you for your ‘name’

Enter the Username that appears in point 5) of instructions (copy & paste)

Once you press enter, it will ask for the password. Again, copy & paste what is written after ‘Password’. Note: the password text will not appear in the terminal. The terminal will look exactly the same after your password has been pasted in, but trust that it is there. Press enter

1. Username: xxxxx  
   Password: xxxx

Now you need to navigate to your username uploads folder in ncbi, by entering the following command (where xxx will be name of your own folder as in instructions on submission page):

1. Navigate to your account folder:  
   cd uploads/xxxx\_gmail.com\_xxx

Now you are in your folder. Each submission needs data to be put in a new subfolder within your folder. Use the command below (mkdir) to make this new subfolder (aka directory). Give it a name meaningful to you. Example below is new\_folder but can be anything (eg SoilData2019)

1. Create a subfolder (required!) with a meaningful name:  
   mkdir new\_folder

Now you have successfully made the folder, you need to navigate to it using cd as in example below. IF YOU ARE NOT IN THIS SUBFOLDER WHEN YOU START THE UPLOAD YOU WILL NOT BE ABLE TO FIND YOUR DATA LATER!!!!

1. Navigate to the target folder you just created:  
   cd new\_folder

Now use put command to upload. For example, if you have a tar.gz file, it would be

put fastaFiles.tar.gz

1. **Copy your files into the target folder**:  
   put Sequences\_file\_name

NOTE: If at any time there is a message in your terminal saying anything about ‘Timeout’ or ‘no connection’ it means ftp is no longer connected to ncbi and you need to start again from step 4).

If everything goes correctly after **put** command, the following message will appear in your terminal: (where file name will change of course)

Opening BINARY mode data connection for fastaFiles.tar.gz

A percentage process will show in the terminal along with an estimate of upload time -> DO NOT close you laptop or the connection until the upload is 100 % finished!

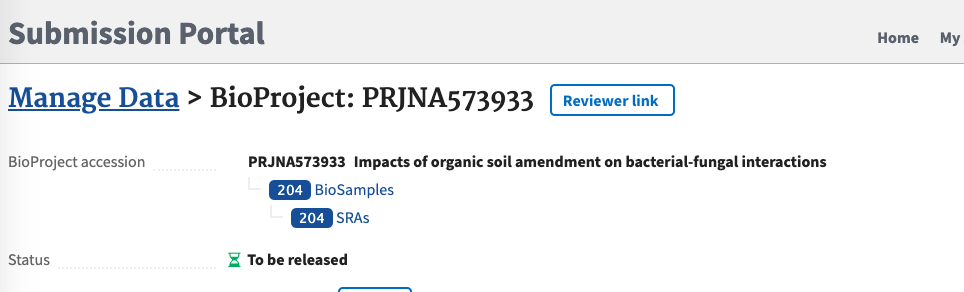
1. Now, once you press the blue box back on ncbi submission page saying ‘Select Preload Folder’ a pop up should appear and your folder will eventually appear there (no stress if not there yet, see step 12).
2. Once ftp upload is complete (it will say 100 % complete in terminal) then you can close everything (terminal and submission page) even if your Preload folder hasn’t appeared yet. Your submission will not be lost -> it will be saved (as incomplete) and you can find it at any time (as long as you are logged in) on

<https://submit.ncbi.nlm.nih.gov/subs/sra/>

1. After a few hours, go back to your submission (link in step 11) and back to file upload page, click the blue box of ‘select preload folder’ and the folder you made in step 7) should **finally** be showing there. Select that folder.
2. Click submit and the next step will be uncompressing your tar.gz file (on the ncbi database) so that they can check the file names match exactly with those given in the sra\_metadata file.
3. If the files are all fine, submission will be finished.
4. You’ll receive an email from SRA if anything is weird (missing files or anything).

Once everything is sorted, you’ll be able to get a ‘reviewer link’ – because your files should be made private until submission (this will be an option at start of submission). To get reviewer link, log in and go to <https://submit.ncbi.nlm.nih.gov/subs/sra/>

Then choose ‘Manage Data’ and select bioproject you need link, the page it takes you to should look like this:



Click ‘reviewer link’ to get it. That goes in your paper under the submission section so they reviewers can see you really did make them available! Then, when paper accepted that will be changed and the submission should automatically become public, but if not do by hand by entering DOI of paper in relevant section (forgotten where but its simple enough).

## Handy links:

good instructions for ftp

<https://www.ncbi.nlm.nih.gov/sra/docs/submitfiles/>

other stuff

<https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/#log-in-to-the-sra-submission-por>