Metagenomics Summer School:

The European Nucleotide Archive

Find this tutorial here:

goo.gl/raVcFt

Introduction

The European Nucleotide Archive (ENA) is part of the European Bioinformatics Institute (EBI) and it is a repository for experimental sequence data. The data in the ENA is submitted by the scientific community usually as supporting evidence to research projects and publications. Because modern sequencing technologies produce data that is so rich in content many institutes, companies, and even small scale researchers are accessing data from ENA to carry out their own analysis or to carry out comparison studies with their own data. The ENA is also a platform for large collaborations to share data where some groups will collect and archive source samples, others may provide the sequencing data, and others still can add analysis results.

Sources of sequence data in the ENA.

Source material/samples (sometimes represented in ENA as **sample** objects):
> Isolated (and if microbe, cultured) organism sequencing
> Environmental sample sequencing

Experimental strategies in the ENA, sometimes represented as **experiment** objects:
> Next generation sequencing: shotgun, amplicon, RNAseq, other types of pre sequencing selection
> Sanger capillary multi pass sequencing (usually from cloned DNA)

Data types in the ENA
> raw output from NGS: BAM, FastQ, SFF. Files (or pairs of files) are wrapped in **run** objects.
> sequences assembled and annotated from short reads to create whole genomes, whole chromosomes, or just a large set of contigs. Annotation can be added.
> sequences of specific genes/regions. For example, a gene that codes for a protein, an rRNA gene, barcode gene. These sequences could be assembled from short reads, or they could have been cloned and sequenced with Sanger capillary sequencing (high accuracy but low throughput).
> interpreted data of ENA run objects. A typical workflow based on NGS read files can result in variant call files, BAM alignment files, OTU and count tables, and others. It is possible to add these into the same study and join them to the source run object and sample object.

ENA in Metagenomics

The European Nucleotide Archive plays a part in 2 typical metagenomics work streams.

Work Stream 1: Create your own metagenomics project and webpage to share your data

Part 1 of this tutorial involves publishing a metagenomics study online so that it can be shared and analysed. Putting your data in ENA as early as possible will open up the experimental design and secure its reproducibility. If sufficiently annotated, read data archived with the ENA will undergo taxonomic and functional analysis by the EBI Metagenomics portal. This could even be completed before the end of this course.

ENA Metadata Objects

Groups that are sequencing environmental samples can represent their source samples in the ENA as 'sample objects' and then attach data to them. The sample objects are searchable because the ENA offers a powerful sample filter on its website, and then data is downloadable for anyone to analyse (after it is made public).

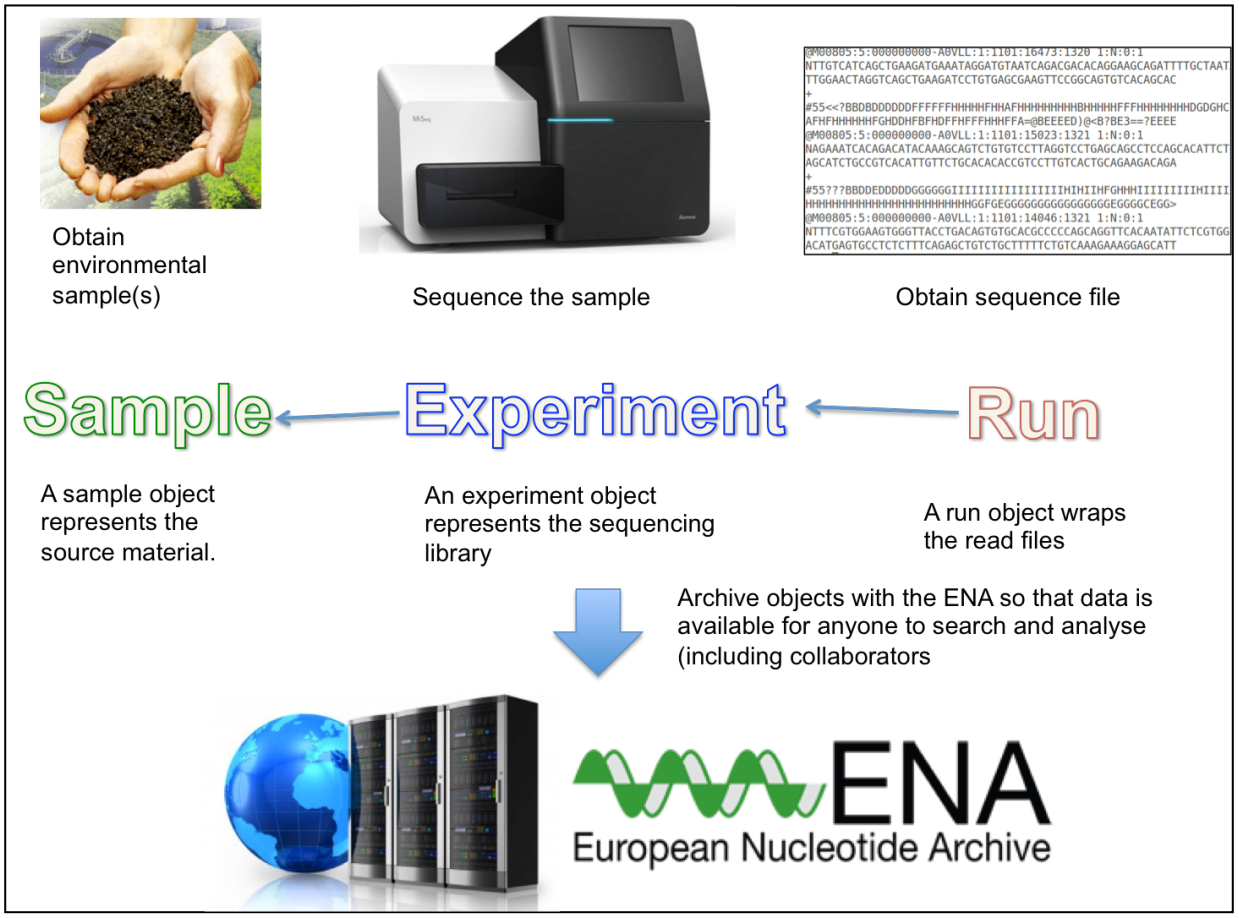


Figure 1: Work stream 1. Store NGS files in ENA worldwide archive. Attach it to sample objects so it is fully searchable

This stream may also involve the data owners or collaborators running analysis pipelines, the results of which can be linked to the sample objects. This means that anyone who finds the sample, can find the data, and can find the analysis results.

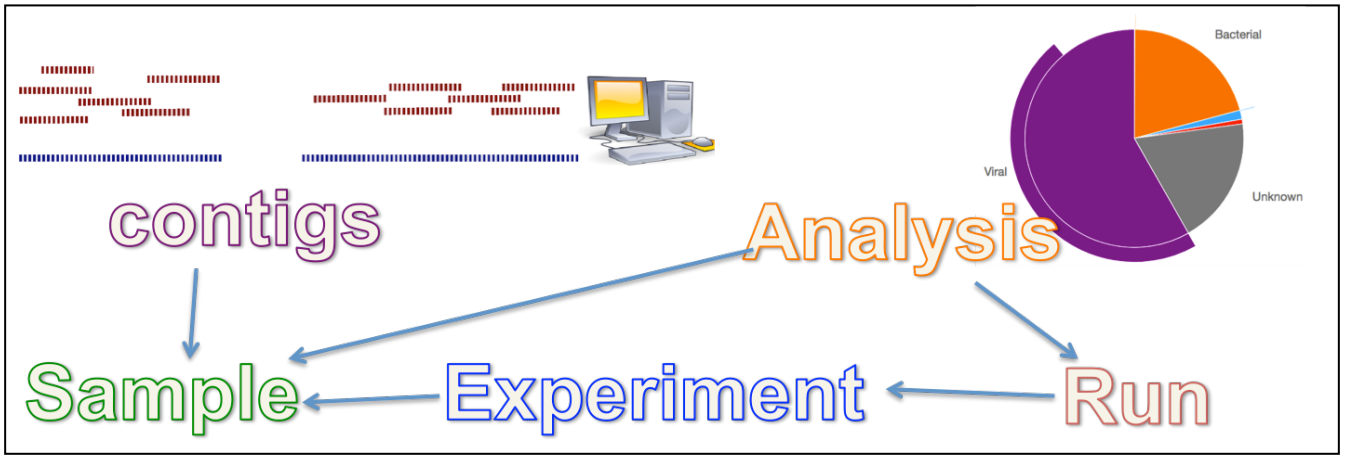


Figure 2: Work stream 1. You can add your own analysis files and assemblies. You can point these to the same sample objects that the raw data is connected to. Your study is now navigable!

Each object (sample, run, analysis etc …) archived with the ENA will obtain its own accession which can be used to access the data online directly. Accessions are used in publications to point to data directly. With sufficient annotation objects can also be discovered without an accession, as you will find later in the workshop.

The value of submitting your project to the ENA

Representing a project online in the ENA allows for instant access by anyone including collaborators and reviewers. All your data is backed up and the ENA object-based data model enables very fast interpretation of the study. For example, you can see how many samples are collected and what the difference is between the samples. You can find all read files that have been sequenced from each sample (already demultiplexed) and instantly find analysis files that result from that read data.

Many free online analysis services such as Taxonomer and Galaxy can stream data directly from the ENA so you do not have to upload data multiple times to multiple services and you do not have to repeat metadata entry for each service. The Metagenomics portal at the European Bioinformatics Institute will provide free analysis for all metagenomics based projects that are archived with the ENA and you can obtain the results before your study is public because of our single sign in agreement with the Metagenomics team.

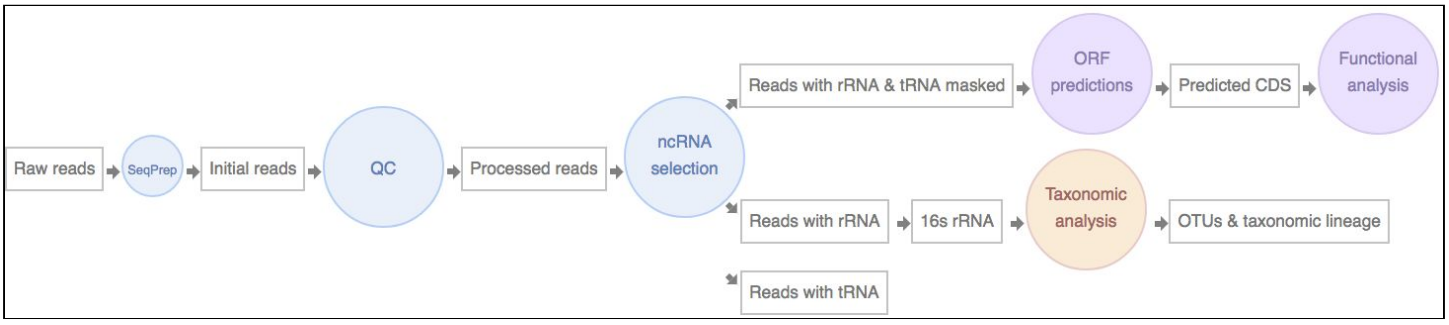


Figure 3: Free services for taxonomic analysis and functional analysis available for highly annotated ENA submission

Work Stream 2: Browsing the ENA for the purpose of 3rd party data analysis

The ENA is a worldwide repository for sequence data. Many research projects, analysis pipelines and publications can be carried out without the expense of collecting samples and NGS sequencing. Large international collaborations are contributing millions of samples and their sequencing output to the ENA, for instance the American Gut Project and Tara Oceans. There are also hundreds of thousands of smaller scale research groups supplying samples and sequencing data to the ENA.

Part 2 of this tutorial looks at searching for samples of interest according to specific variables and factors, reporting the data files that are associated with them and then downloading the data files so that they can applied to analysis pipelines that will be covered in later sessions of this metagenomics summer school.

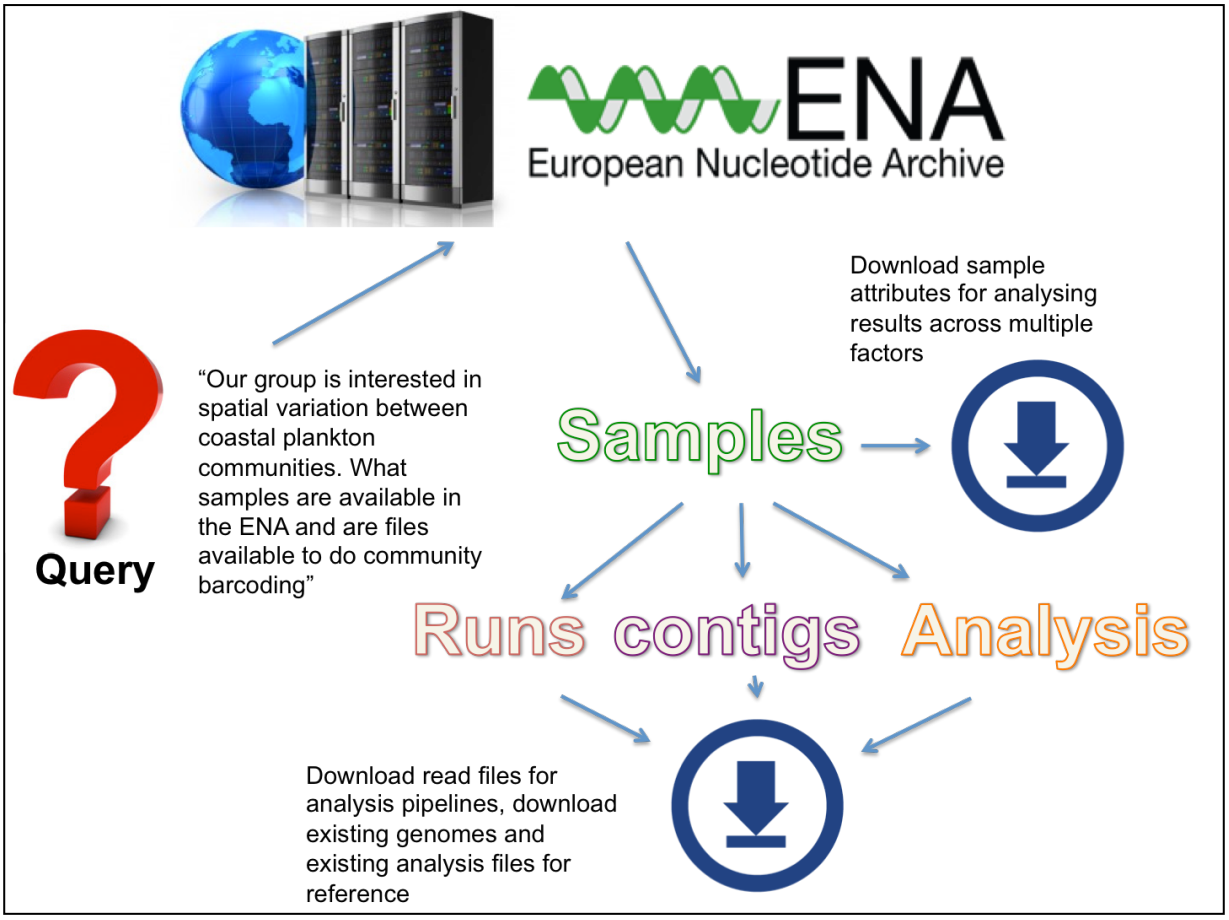


Figure 4: Typical workflow using the ENA: Search, download, analyse

Part One: Representing a metagenomic study in the ENA

The Meta Data Model

Understanding the ENA metadata model is required so that you can represent all elements in a study in a way that they can be easily searched and navigated. ENA meta data 'objects' represent real life objects.

Sample Object

An ENA sample for example, has fields like 'isolation source', 'location', 'collection date'. If you have collected a sample from an old book in a library in Bari, Italy on 05-Jun-2017 and you create a sample in the ENA with those fields correctly annotated then your ENA sample can be instantly discovered if anyone is searching for these criteria. If there is data connected to the sample in ENA then that data instantly has context and it can be included in comparison studies against other book samples that may not be yours!

Experiment Object

This represents the library solution. Multiple libraries can be created from a source sample. For example, in one case you could amplify 16S rRNA genes before sequencing, in another case you can prepare a shotgun sequencing library for the Illumina platform, and in another case still you could prepare a shotgun library for the 454 platform with the interest of combining longer 454 reads with shorter Illumina reads to enable contig assembly. These are 3 libraries from the same sample. A library/experiment object 'points' to its source sample so that it is possible to filter on sequencing platform, selection method, molecule type (RNA/DNA) etc ….

Run Object

This object wraps a [pair of] read file[s]. ENA will add some more things when the read files are processed like number of reads in the file and the number of base pairs. This means that it is even possible to filter search results by how large the read files are. A run object 'points' to a library/experiment object. It is not common but many runs can point to a one experiment object. This could reflect technical replicates or a deep coverage sequencing experiments where one library is used on multiple sequencing lanes.

The run does not point to the source sample directly but it does point to an experiment, which in turn points to the source sample so the connection is made and the run can only have one source sample (demultiplexed) because a run object has one experiment pointer only, while the experiment has one sample pointer only.

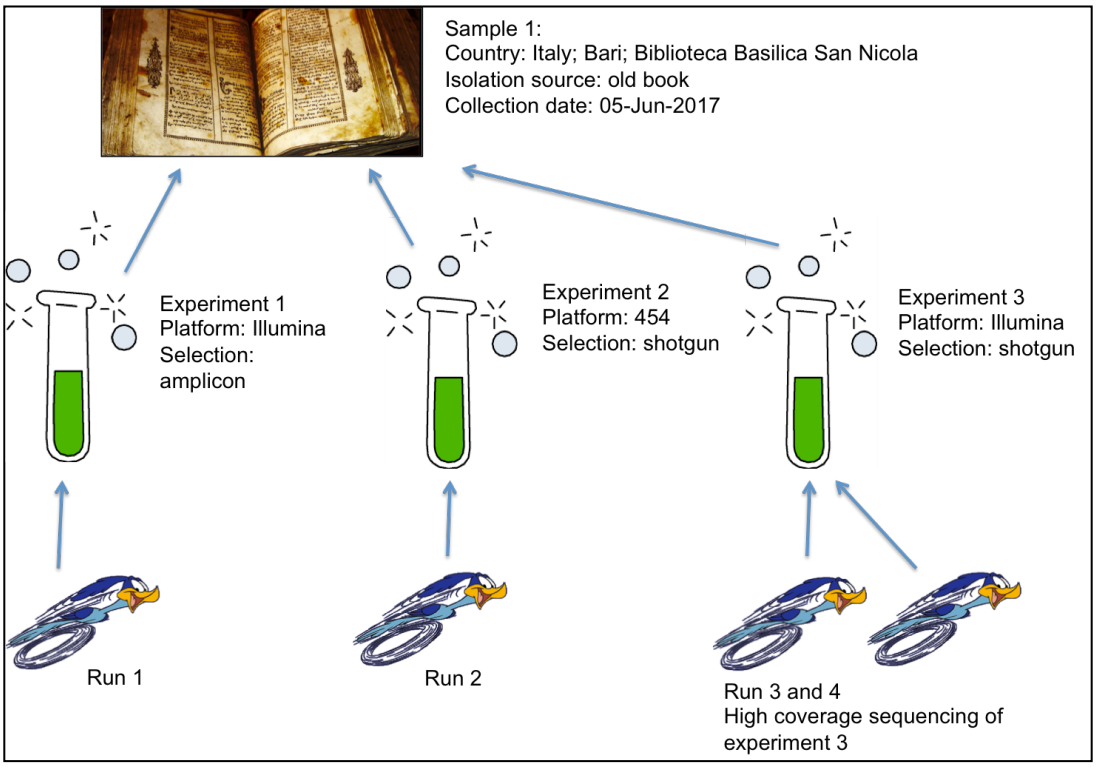


Figure 5: Different metadata objects are submitted to ENA. Because they are connected it is possible to navigate a study easily

The way that metadata objects are connected allows for fast interpretation of a study. It can be easy to gage the scale of sample collection, the test environments, the sequencing strategies and the quality of the read data as easily as (or more easy than) reading the journal publication.

Assembled contigs and the analysis object

If your raw reads are assembled into contigs you can submit the contig set. The contigs can be annotated with features such as CDS and RNA. A contig set or higher level (scaffold set) genome requires a source sample object to be registered in the ENA before it can be submitted. An analysis object is like a run object because its main function is to wrap a file and connect it to other objects in the ENA. Instead wrapping a raw read file (as with run object) the analysis can wrap most types of interpreted files like VCF, OTU/tab, aligned BAM. Analysis objects are quite flexible. They should point to the source of the analysis where appropriate, so they will often be connected to sample objects or more accurately run objects. Unlike other objects they can point to multiple objects of the same type.for instance, if you have taken reads from several runs and aligned them to a reference sequence, the resulting aligned BAM file can point to all source run objects.

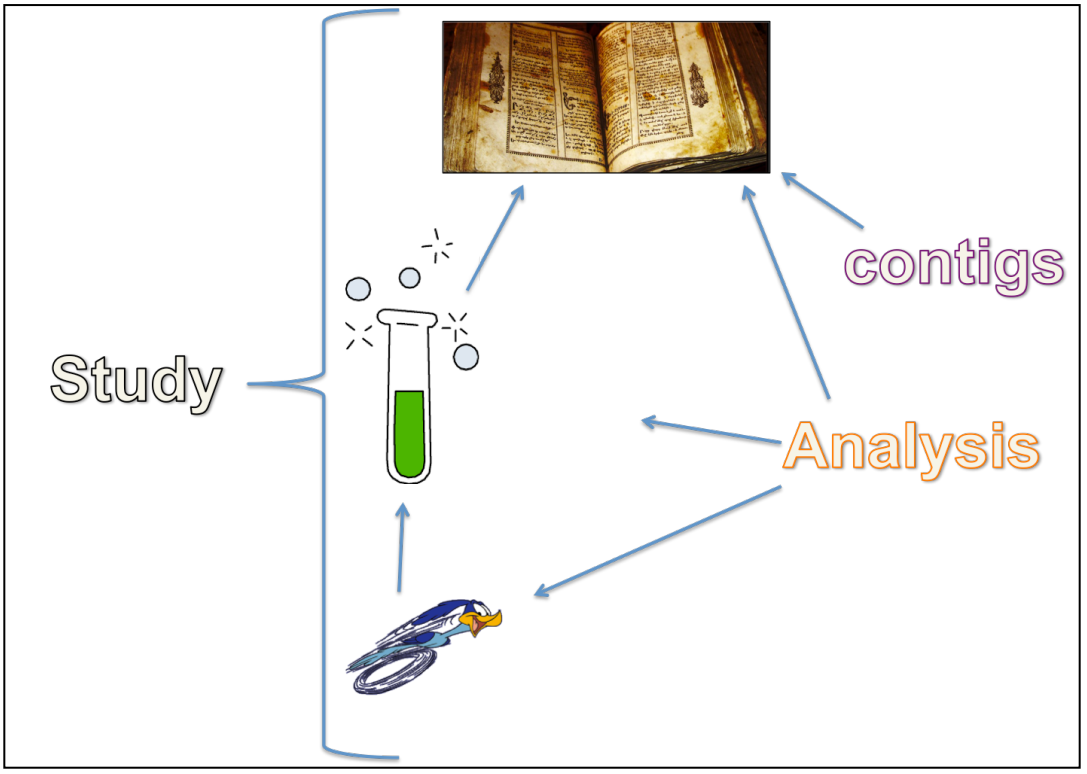
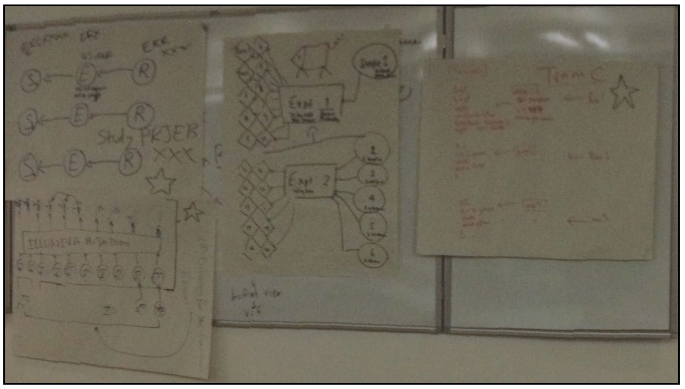


Figure 6: assembled contigs can be added to a project. They reference a source sample. Interpreted files like OTU tables can be added using the analysis object. The analysis object is quite flexible - it has many pointers so you can quickly find the source raw data used in the original analysis pipeline.

Task 1: Map your study as it would be organised in the ENA

Figure 5 above demonstrates how a NGS study is represented in the ENA. It is put to together from a series of connectable metadata objects. This architecture enables a study to be built by different account owners and easily updated or added to. Pay attention to the direction of the arrows in the figure. An experiment points TO a sample but not the other way around. This means that once a source sample/physical material is represented in the ENA as a sample object, additional libraries/experiments, runs/NGS files, contigs and analysis results can be added without modifying the original sample or owning it. The data model works for large collaborations but also for small scale studies.

In groups of 2 or 3 people pick a study that one of you is working on and think about how many source samples are involved in the study, what kind of libraries are required, and what runs/NGS files will be generated. Also consider what extra analysis files can be added and then draw the objects on a piece of paper and connect them with arrows as in figure 5 and 6. A study in the ENA must be annotated. Add some basic annotation to the source samples - particularly how they are different from each other (time collected, treatment applied, location/distance). The experiment object should be annotated with the sequencing platform and if it is set up for paired or single end sequencing. The experiment object should also include the what is being selected for - for example exome sequencing, RNA seq, 16S amplicon etc …



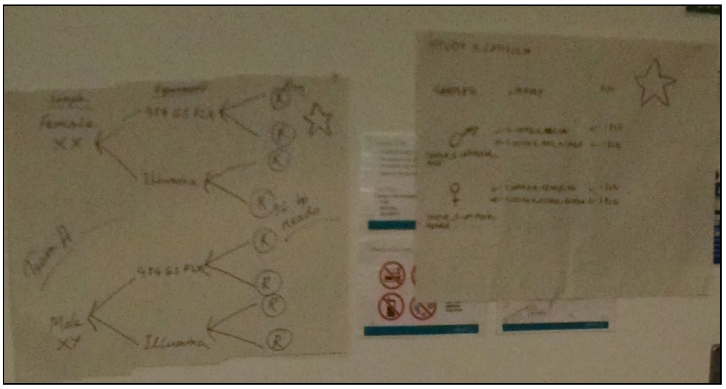


Figure 7: first step towards publishing your study is to be able to reflect the experimental design accurately using ENA's platform

Task 2: Navigating the Sample Domain and Read Domain in ENA

As discussed, data in the ENA is organised using metadata objects. It is organised in a similar way internally, as a large relational database (we have sample tables, run tables, experiment tables …) but is necessary to extend this concept into 'domains' to reflect how the data is indexed for searching and browsing online. It is not a relational database on the public web front, but instead a grid based warehouse that enables fast searching and filtering.

The Sample domain

All sample objects in the ENA have a unique accession that looks like this **ERS914738** or this **SAMEA3607589**. The second one is EBI-wide/cross-services so that other resources can attach metadata without the sample existing multiple times in the EBI (The European Bioinformatics Institute is home of resources including ENA, Array Express/Expression Atlas, Metagenomics portal, Ensembl, TrEMBL/UniProt, ChEMBL and many more). The first type of accession is the original ENA accession, maintained for legacy reasons. Both types are valid and both will be seen in publications to reference real life source material. As in real life, a source sample should not exist more than once in the ENA/EBI.

To take a look at the sample, or any other ENA object in fact, append its accession to the URL
http://www.ebi.ac.uk/ena/data/view/
Use your browser to navigate there.
This is the sample page. It is rendered from this XML
Compare the **attributes** tab on the sample page with the <SAMPLE_ATTRIBUTES> block in the XML.

While the sample page looks better to a human than the XML page, many services that are downstream to the ENA parse large amounts of ENA data to incorporate the into their own analysis pipelines and portals. XML format is easily read into a database or a report and if the accession is already known, many pipelines downstream of the ENA can be completely automated. Objects are stored internally in the ENA as XML format as well.

The sample attributes provide all the context for any data that is attached to the sample. The submitter of the sample has applied these themselves and there are ENA recommendations that you will discover when you create a sample in the ENA later.

It is up to the user to parse the XML into something that they can plug into their analysis pipeline. There is an example workflow at the end of this tutorial where we have used R to parse sample XMLs into the analysis results. The ENA can report some of the fields that are in the XML (so you maynot need to write a parser). These are the indexed fields. An indexed field means that the sample attribute can be searched and filtered against. Sample ERS914738 is very thoroughly annotated and contains attributes that are not indexed by the ENA, but there is some overlap. This URL uses the string *display=report* to ask the ENA server for all fields in sample ERS914738 that ENA currently indexes (you can find this list in the documentation)

http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22sample_accession=%22SAMEA3607589%22%22&result=sample&fields=sample_accession,accession,secondary_sample_accession,bio_material,cell_line,cell_type,collected_by,collection_date,country,cultivar,culture_collection,description,dev_stage,ecotype,environmental_sample,first_public,germline,identified_by,isolate,isolation_source,location,mating_type,serotype,serovar,sex,submitted_sex,specimen_voucher,strain,sub_species,sub_strain,tissue_lib,tissue_type,variety,tax_id,scientific_name,sample_alias,checklist,center_name,depth,elevation,altitude,environment_biome,environment_feature,environment_material,temperature,salinity,sampling_campaign,sampling_site,sampling_platform,protocol_label,project_name,host,host_tax_id,host_status,host_sex,submitted_host_sex,host_body_site,host_gravidity,host_phenotype,host_genotype,host_growth_conditions,environmental_package,investigation_type,experimental_factor,sample_collection,sequencing_method,target_gene,ph,broker_name,sample_title&display=report

Try putting the above query in your browser. The result will be *tab delimited* which means that it can be viewed in a spreadsheet program. I have loaded it into the Google spreadsheet program here. Many fields are missing but you can see where the submitter has applied some ENA indexed sample attributes by chance. If you look back at the XML or the attributes tab you will also find additional attributes that are not recognised by the ENA reporter.

Can you answer these questions:

1. What is the sex of the person that this sample is representing?
2. How do we know that this sample is a human sample?
3. What season was the sample collected in?
4. What type of sample is this?
5. Find one attribute in the XML that is not available as an ENA report
6. What weight in kg is the subject?
7. Using the geolocation field, what is the closest town to where this sample was collected?

The Read Domain

Let's stick with sample ERS914738.
Check the <SAMPLE_LINKS> and answer these questions:

1. Does the sample have any experiment objects using it?
2. If so, how many and what are their accessions?
3. Does the sample have any run objects connected?
4. If so, how many and what are their accessions?

According to figure 5 above, a sample does not have its own pointers (run-> experiment-> sample). If the sample contained information about the experiments using it then everytime a new experiment was added, the sample would need to be updated. This is more difficult if multiple collaborators are contributing to a project, all with separate submission accounts. In fact as it is stored in the ENA the sample has no connections to other objects, but database tables can be joined. When the sample is indexed for the purpose of being displayed on the ENA browser then the joins are done and the related objects and added to the XML so that the browser can display these connections. In other words, the <SAMPLE_LINKS> block is not part of the original sample. It is added when it is published. Browser indexing happens daily so edits to the data and the connections between the objects will usually appear in 24 hours.

According to the model **run -> experiment -> sample**, this sample should have at least one experiment pointing to it (find it in the <SAMPLE_LINKS> block). The experiment object will have an accession that looks like ERXNNNNN. Go to the experiment object by adding the accession to the URL
http://www.ebi.ac.uk/ena/data/view/
From here, load the XML version of the experiment object by adding "&display=xml" to the end of the URL. Can you find the sample pointer? Pointers must be applied correctly when submitting a study to the ENA so it can be useful to know what they look like.

The sample pointer looks similar this:

```
<SAMPLE_DESCRIPTOR accession="ERS914738" >
```

The experiment is a special object because apart from representing the library preparation created from the source sample for sequencing on an NGS platform, it is also responsible for holding a study together. This is because it has a study pointer as well a sample pointer:

```
<STUDY_REF accession="ERP012803" >
```

The sample and run objects do not connect to the study directly. They can only be affiliated to a study if there is an experiment object to make the connection. This is how groups and organisations can download large number of studies to present them in their own portals or to analyse the raw data- the pointers between the XMLs allow the data to be stored in a relational database or a series of tables that are connected. This is also why submitters should make accurate connections so that their study can be easily interpreted.

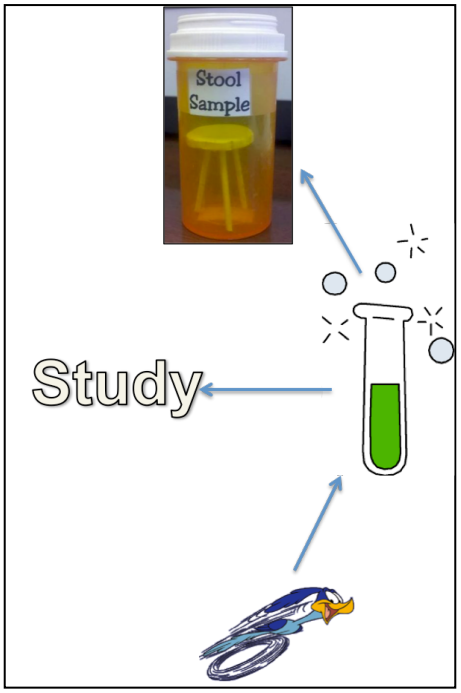


Figure 6: The experiment object points to the study. The run and sample are associated to the study via the experiment object.

The experiment holds information about the library preparation. Can you answer these questions:

1. Is this a shotgun library or an amplicon library?
2. What instrument platform was used to carry out sequencing?
3. Is library for paired or single end sequencing?

As with the sample, some experiment attributes are indexed, to save the user from parsing the XMLs in simpler case scenarios and to allow searches based on experiment/library criteria. So we can create a report from them:
http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22experiment_accession=%22ERX1152774%22%22&result=read_run&fields=secondary_study_accession,experiment_accession,instrument_platform,instrument_model,library_name,library_layout,library_strategy,library_source,library_selection,experiment_title,experiment_alias&display=report

The above URL includes
experiment_accession="ERX1152774"
&display=report
And most experiment fields that are currently indexed.

Because the study is known to the experiment we can also find all experiments in one study:
http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22secondary_study_accession=%22ERP012803%22%22&result=read_run&fields=secondary_study_accession,experiment_accession,experiment_title&display=report

As you can tell, this is a very large study!

We are now left with the run object, which should have a pointer to the experiment if we are to believe the **run -> experiment -> sample** idea. Can you find the run from the experiment XML (<EXPERIMENT_LINKS>
block) or the one of the previous reports and then navigate to the run XML? When you have found the run XML try to find the block that represents the run pointing to the experiment that it is a sequencing run of.

When compiling internet search tables for the browser, the experiment and run attributes are put in the same table, so we can use the same report source as above (denoted by *&result=read_run*) and also mix and match run and experiment attributes. Here are a few run attributes that are indexed and therefore searchable as criteria:

http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22run_accession=%22ERR1072624%22%22&result=read_run&fields=run_accession,read_count,base_count,first_public,run_alias,fastq_bytes,fastq_md5,fastq_ftp,submitted_bytes,submitted_md5,submitted_ftp,submitted_format&display=report

This report is more easily viewed as a spreadsheet.
Most of these fields, such as read count, base count and file size are added after the original file is submitted as a run object. The main purpose of the run object is to store the location of the file for download but it also stores additional details about the file.
Every submitted file is processed and converted into a standardised fastq file (even if the original file is not a FastQ). This standardised fastq is useful for downstream analysis pipelines because they do not need to account for the different types of files that are submitted. All standardised files will follow the the same specification regardless of where the original file comes from. This is the difference between fields 'submitted' and 'fastq'. For example "submitted_bytes" refers to the file size of the original submitted file and "fastq_bytes" is the file size of the standardised fast q that ENA has created from the original file.

The ENA is set up for 2 different types of users. Large scale users can search for accessions and then generate reports of the metadata for those accessions and also get a table of file stats such as md5, file size and the ftp location of the file. They can then download the files for insertion into their pipelines and most of this can be done automatically with minimal manual intervention.
The second type of user is smaller scale, who is interested in one or two studies and is happy to navigate the ENA browser and click to download the various elements of a study.

Can you answer these questions about run ERR1072624?

1. What is the total read count?
2. What is the difference in size between the submitted file and the fastq standardised file?
3. When did the file first become public?

Task 3: Navigating other important domains in the ENA

As mentioned in figure 6, assembled contigs can be added to a project and interpreted files like OTU tables can be added using the analysis object.

The Analysis Domain

Here is an example of an analysis object wrapped around an interpreted data file. In this case the file is a gene table:
ftp://ftp.sra.ebi.ac.uk/vol1/ERA413/ERA413485/tab/TARA_072_DCM_0.22-3.gene.tsv
The analysis object is like a run object because its job is to hold details about a file. In the case of a run object that file is a raw data file but in the case of an analysis object it can be almost any type of file that has been created from the raw data. In the metagenomics community these files are typically results from taxonomy analysis and functional analysis of sequence data.

This analysis file was created from 2 paired read files. The same read files were assembled into contigs as part of a related analysis pipeline. This contig set is also archived in the ENA. Can you answer these questions:

1. The analysis object has 2 source runs (ERRXXXXXX). What are they?
2. The 2 run objects are generated from the same sample. What is the sample (ERSXXXXXX)? You don't have to navigate to the runs to find this because the analysis has a sample pointer as well as run pointers so you can find the source sample immediately.
3. An analysis object, like an experiment object, must point to 1 study. What study (ERPXXXXXX) does this analysis belong to?

The Assembly domain

Here is the assembly related to the previous predicted gene table.
http://www.ebi.ac.uk/ena/data/view/GCA_001040105
The ENA allows multiple levels of assembly. The first level is the contig set where raw reads are assembled into contiguous sequences to create a set. Contigs can then be put in order to create longer sequences with gaps. These are called scaffolds. Scaffolds can further be joined into a chromosome and chromosomes can be collected together to make a full genome. These genomes are the basis of annotated reference genomes that appear in genome databases such as Ensembl.

In an environmental sample there will be no single organism so assemblies beyond contig sets are rare for the moment. A contig set is accessioned with a 4 character prefix, and then each sequences is numbered as part of that set. Assembly GCA_001040105 has a contig set. The prefix is CERB01 ('01' refers to the version). From the same webpage, click on WGS Sequence Set: CERB01000000. You can now see that the full range is CERB01000001-CERB01346087 so there are 346087 sequences in this set.

Try navigating to the 50th sequence by adding its accession 'CERB01000050' to the URL
http://www.ebi.ac.uk/ena/data/view/
Now click on the View: TEXT link or add "&display=text" to the end of the URL. This is a typical sequence file in the ENA. Source sample information is embedded in the same file as the sequence itself, in the 'source' feature. Representing this information as a sample object is a newer concept that is being adopted in sequence files as well, so this sequence does actually point to a sample object as well.

Can you answer these questions:

1. Can you find the sample pointer?
2. Can you find the study pointer?
3. How many base pairs does this sequence have?
4. This is the 50th sequence in the full assembled set. Now navigate to the 100th and find out how many base pairs it has.

The Study Domain

The study/project domain is mostly regarding the study object which has already been discussed. Like the sample object, the study object does not point to any other objects. It is up to the other objects to point to it. This means that data can be added to projects without having to update them or even own them. This enables easy collaborations and sharing of data. The study that contains the assembly and analysis mentioned in the previous section is a good example of a large collaboration where many groups have added content over a several year period:
http://www.ebi.ac.uk/ena/data/view/PRJEB7988
Click on the 'navigation' tab. This study contains analysis and assemblies. Whereas this one is part of the same extended study and contains mostly read data:
http://www.ebi.ac.uk/ena/data/view/PRJEB4419
Look at the navigation tab. Note how the samples are shared - the same samples are being used by elements in separate studies. It us up to the submitter how they choose to organise a study - they can put all elements in the same one or they can split the elements across multiple studies (as long as the source samples are not duplicated). One common reason to split a project is because the ENA project is the unit of publication and different elements may be released to the public domain at different times so these can go in separate projects according to their publication status.

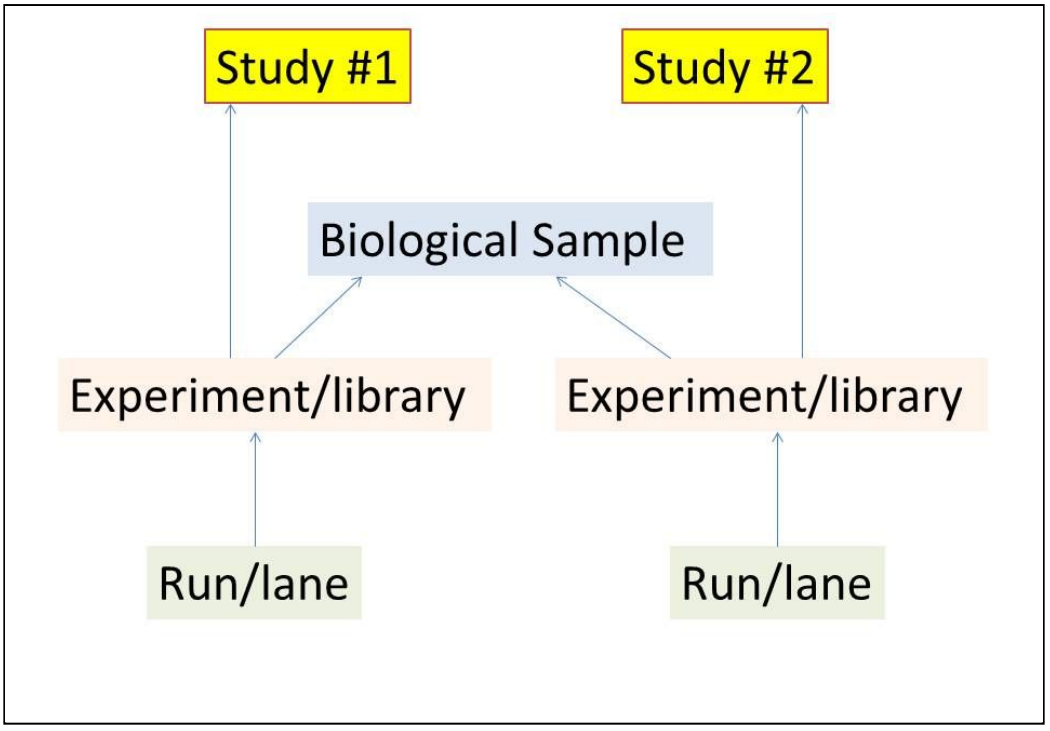


Figure 7: The ENA datamodel allows data to be split across studies without having to duplicate the source sample, which should only exist once, as it does in real life.

Large collaborations can use an 'umbrella' study to group together several children projects:
http://www.ebi.ac.uk/ena/data/view/PRJEB402
Click on the 'component projects' tab to see all the child projects.
An umbrella study should not have any data directly attached, and children studies should not be parents to other studies so the hierarchy can not be more than 2 levels. This is to avoid scenarios becoming too complicated and hard to interpret for ENA users.

Here is another study from the large Tara Oceans expedition. This study contains single cell sequencing results. Because it is single cell, an assertion can be made about the identity of the organism. For example, sample SAMEA3727694 has organism name Stramenopiles sp. TOSAG23-6 instead of an environmental sample/metagenomic classification as we have seen before. Because the the reads are known to be from a single organism a higher level assembly can be created. The contigs set is here but some of these contigs have been joined to make scaffolds, like this one.

Can you answer these questions?

1. How many different data types have been added to the single cell sequencing study PRJEB6603?
2. Scaffold LT635936 has been created from contigs from assembly set with accession prefix 'FQSH01'. Can you tell how many contigs are used to build LT635936 and which ones they are from FQSH01?
3. How big is scaffold LT635936?

Task 4: Submitting a study to the ENA

It is important to register your source samples and create a study in the ENA as early as possible, even if you do not have sequenced data yet. Representing the study in the ENA will also help to organise your workflows and the EBI Metagenomics portal (featuring later in the summer school) can provide taxonomic and functional analysis (and soon they can provide assemblies) as soon you add your sequenced data. Data submitted to the ENA is not instantly public - you will apply a release date (which can be extended at any time) which must expire before the data is moved to the public domain.

Basic Scenario

Some simplified/reduced size files are available to do a practice submission. This will help you to realise the steps involved so you can begin your own study in the ENA as soon as possible. Consider this scenario: You are studying food metagenomes and have a sample of bread dough from a bakery business in Bari. You have prepared a shotgun library for paired end sequencing on an Illumina MiSeq. You have obtained a set of paired fastq files which you have used to assembled a metagenome/contig set. You have also done taxonomic analysis on the fastq files and you will deposit a tab delimited OTU table which assigns each read to a taxonomic classification.

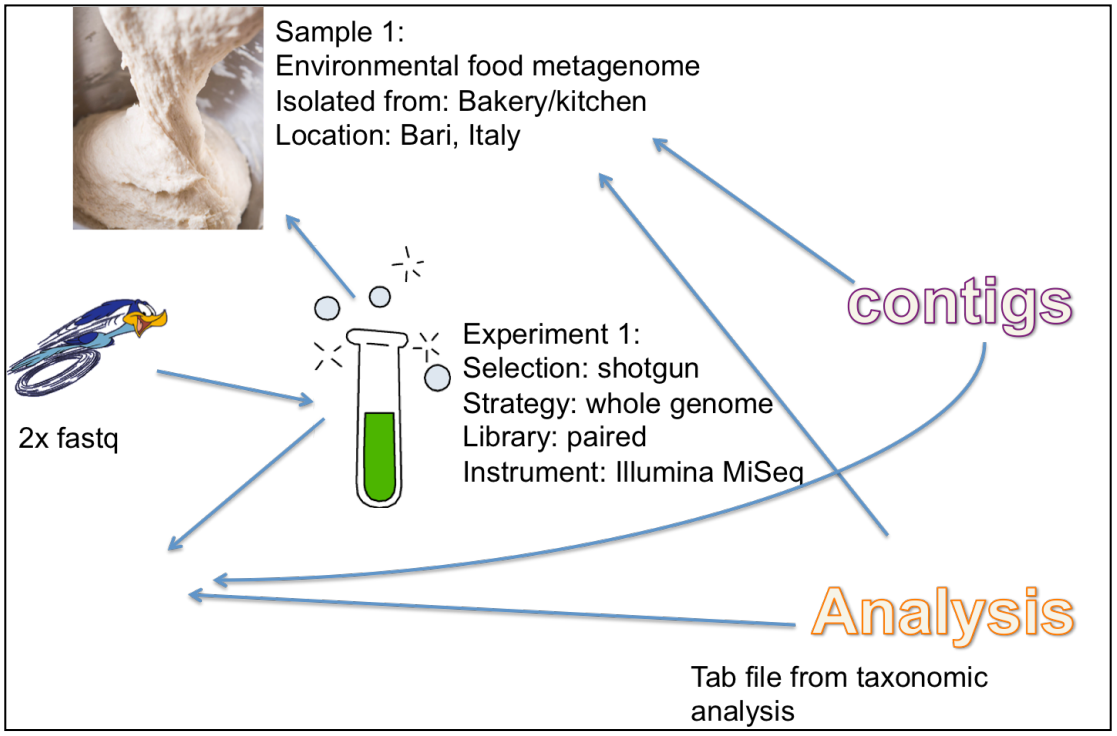
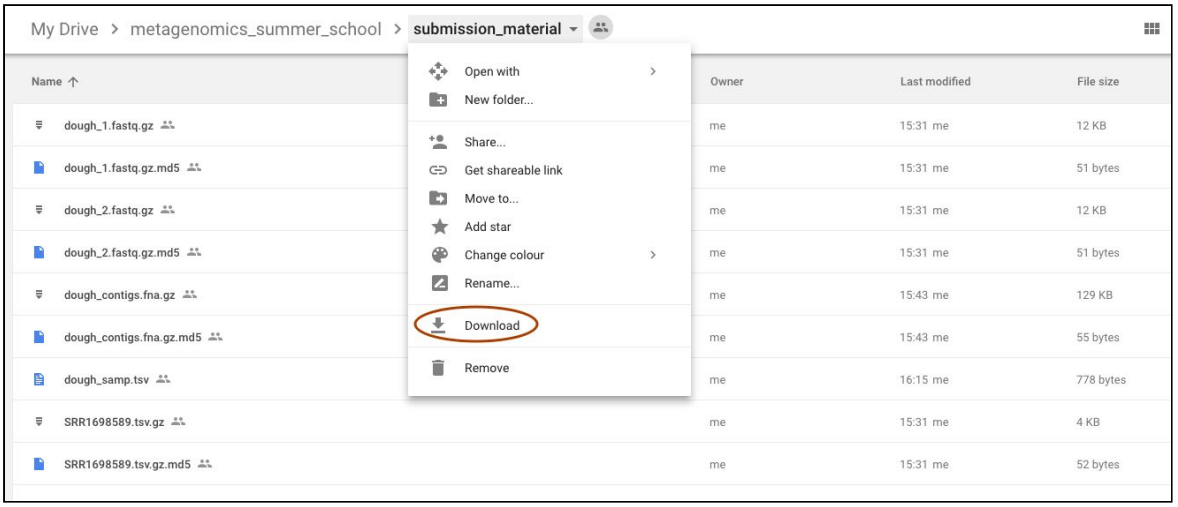


Figure 8: How our practice scenario might look in the ENA

Files for the submission are available here:
https://drive.google.com/drive/folders/0B3aYtilFw3a8Rmt3cjJIVWhTU0E?usp=sharing
You can download the whole folder:



A zip file is created which you can unzip and put somewhere to work from it.

The Study

Begin with the study. Go to the submission service:

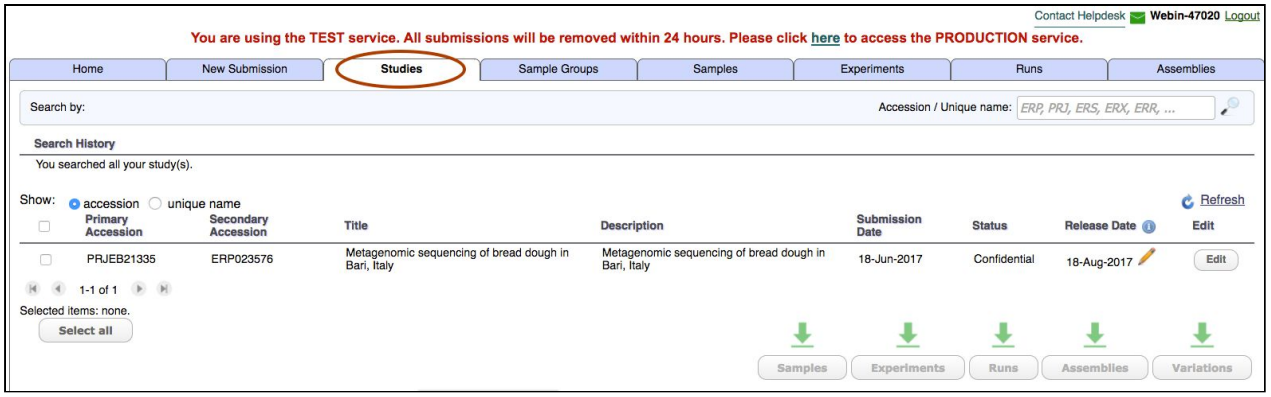
https://wwwdev.ebi.ac.uk/ena/submit/sra/#submissions
This is a test version of the submission service. A submission here will not persist in the ENA database. The test service is useful for trial and error when you come to submit your own study. Also when submitting programmatically using the REST API (may be you are going to implement an automatic submission pipeline for your institute) you can send all your tests to the ENA test server.

There are a number of workshop accounts available for logging in. It is important to try to pick an account that is not being used by someone else so that every object that you submit is unique to the account. The following document is editable, please pick a free account and add you name.
https://docs.google.com/document/d/1R7-cezIhyW4s0KOoIIn97rQTgMqFZnSJV423PlQ2j-4/edit?usp=sharing

The password on all the accounts is "**ws2017**".

1. Copy the 'webin' username you have selected and log in to the submission system.
2. Click the 'New Submission' tab
3. Select the 'Register study (project)' radial and click 'next' (bottom right of screen)
4. Note the first field is the hold date mentioned previously. Only the study needs a hold date. All objects that use the study will inherit the same hold date.
5. The field "Please provide a short name for the study" is important because you will use this to reference your study later when you link objects to it. Call it "**dough_study**" for consistency with this workshop.
6. For your own study take time to add a detailed abstract, title, description, and journal article. If you have not published a journal article yet you can edit/add to most of these fields later. For this practice session a brief entries are fine. For example "Metagenomic sequencing of bread dough in Bari, Italy" can be added to most fields.
7. Click 'submit' when all mandatory fields have some content. Click 'confirm' when prompted.

Your ENA study is now ready to build on! You should see it by clicking on the "Studies" tab. What is the accession number for your study? You can use this in your publication (when you do a real submission for your own project).



The Sample

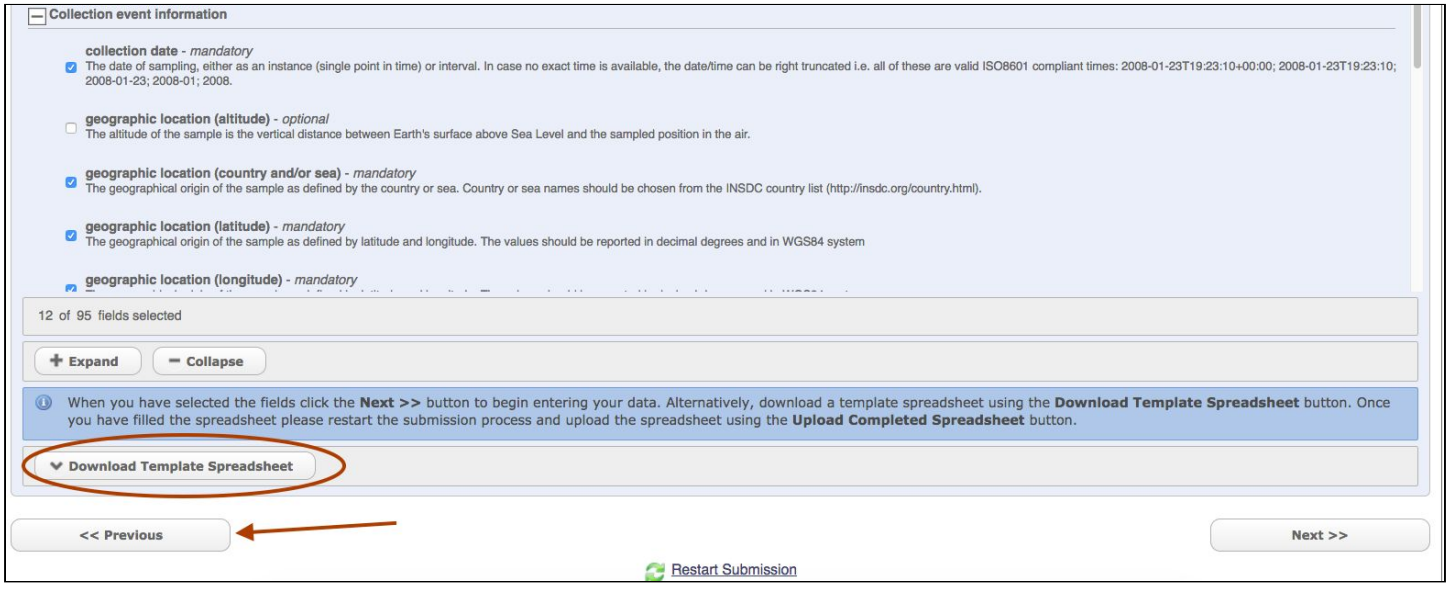
The sample is probably the most important object because it gives the read data and analysis data context and allows it to be comparable. Also most searching and filtering is done on sample data so a well annotated sample will increase exposure of your dataset to the community. From the sample someone can then find the study and any publication information that you have included.

1. Go to the tab 'New Submission'. If your previous submission is still loaded find at the bottom of the page 'To create a new submission please go to New Submission' and select it, then click 'ok' to confirm
2. Choose the 'Register samples' radial and click 'next >>'
3. You have 2 options now. Click on 'Select checklist'. You will find there are a number of different checklists. These are to help you to annotate to high standards. For environmental samples it especially important to provide information about the location, the environment and the timing because

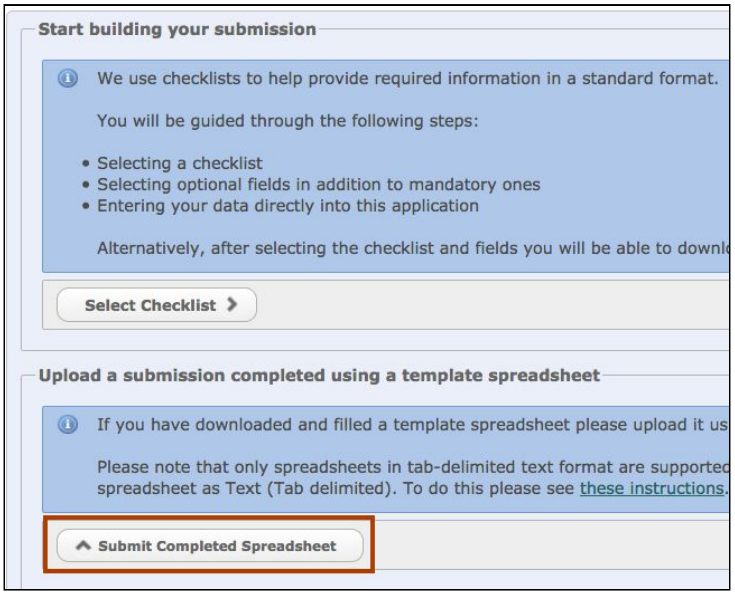
the eventual outcome from submitting to the ENA is that your data will be compared to other similar samples.

4. Expand the ‘Environmental Checklists’ group. You will see checklists for many different types of samples including marin, soil, human gut, human skin etc … There is no food environmental sample so I selected ‘GSC MIxS miscellaneous natural or artificial environment’ for this practical.

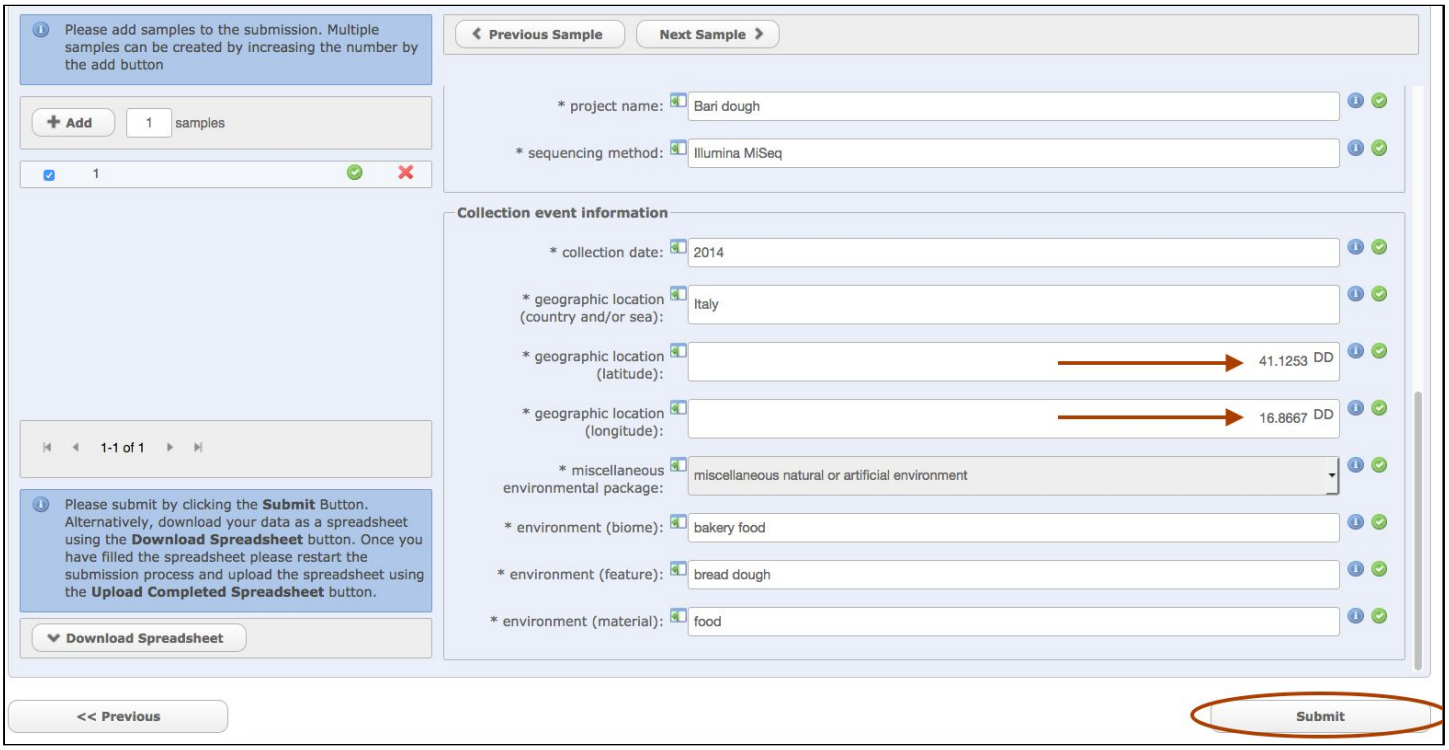
5. When you click next you will find a series of fields. Some are mandatory, some are suggested, and some are available to use but you do not have to. At this point is good practice to download a spreadsheet template of your checklist after you have selected the fields that you want to include. This has already been done for this practical session (the file is in the submission_material directory, it is called ‘dough_samp.tsv’) so instead click ‘<< previous’



6. Now select ‘Submit Completed Spreadsheet’ and navigate to the file ‘dough_samp.tsv’. The sample should load into the webform. You can check what fields are required by the ‘GSC MIxS miscellaneous natural or artificial environment’ checklist. You can also look straight at the spreadsheet using a program like microsoft excel or a simple text editor.



7. Also note that a GPS location is one of the required fields. This most useful for natural environments compared to an artificial environment such as a bakery but it is still an interesting field to apply because searches can be done in the ENA based on very precise locations (example of this later). To complete the sample submission click ‘submit’ and then ‘OK’ to confirm. You should now see your sample in the ‘Samples’ tab! What accession does it have?



The Experiment and the Run objects

In the Webin submission system that we have used so far, the run and experiment objects are created at the same time. It is not possible to assign multiple runs to one experiment for this reason, even though the data model discussed does allow it. There is a REST API submission system available with more flexibility but the Webin service is fine for most user cases. The submission system is used for registering metadata and creating the metadata objects while the read files themselves are not uploaded through the Webin submission system. Each Webin account has an ftp directory which you can upload files to. When registering the run in this step, the submission service will check that the file you are registering exists in the ftp directory for your account before registering the run and experiment objects and delivery accessions.

The read files are in the submission_material directory. They are dough_1.fastq.gz (forward reads) and dough_2.fastq.gz (reverse reads). The files are compressed with gzip tool. This is mandatory because it means the upload will be faster and there is less chance the files will be corrupted during the transfer because they are smaller. You will also see a checksum file for each read file:
dough_1.fastq.gz.md5
dough_2.fastq.gz.md5
These contain md5 checksums - which are the result of a calculation done on the file. These can be uploaded along with the files themselves. The ENA will calculate the checksums after the files are registered. If they do not match the ones that you provided then we know that the files were not uploaded 100% and processing should not continue until you repeat the upload.

If you have access to a linux command line or a mac terminal you can do the following to prepare your files. If you do not have access to a linux or Mac command line it is ok - all accounts in this workshop already have these files uploaded to their ftp directory and md5 sums calculated.

In terminal app, working directory is the submission_material directory downloaded and unzipped during previous session.

Check contents of one of the fastq files:

```
gunzip -c dough_1.fastq.gz | head
```

Create a checksum file of one of the fastq files

```
md5 dough_1.fastq.gz > dough_1.fastq.gz.md5
```

Upload the files to the ftp directory

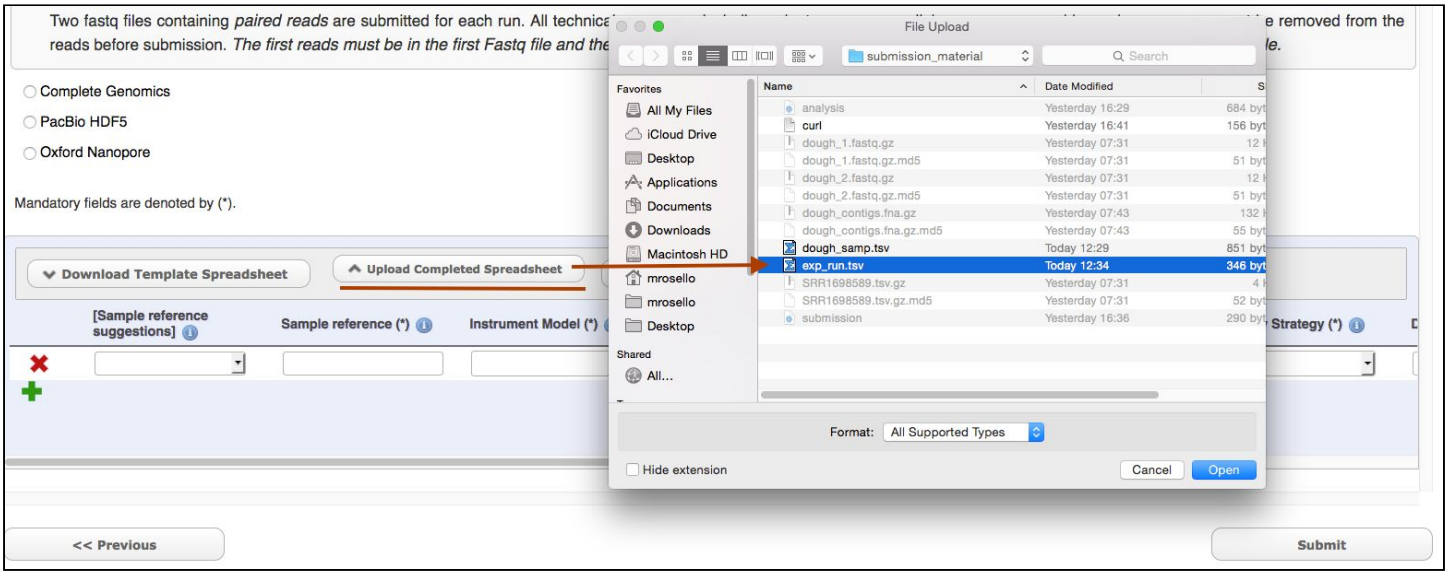
1. Type 'ftp'
2. Type 'open webin.ebi.ac.uk'
3. Enter the username and password associated with your Webin submission account.
4. Type **bin** to use **binary mode**.
5. Type 'ls' command to check the content of your drop box.
6. Type 'prompt' to switch off confirmation for each file uploaded.
7. Use 'mput' command to upload files. Because they all ebgin with 'dough' you can use the wild card '*': mput dough*
8. Use 'bye' command to exit the ftp client.
9. Use 'exit' command to exit the command line interpreter.

It is not always possible to do the above in a workshop environment. The files are already in the ftp directory for each webin account so it is fine to proceed to the experiment and run submission ….

1. Go back to the Webin submission system and go to the 'New Submission' tab.
2. Choose the the 'Submit sequence reads and experiments' radial and click 'Next >>'
3. The study you selected should be present but you may have to click 'Refresh' at the top of the table

4. Select the study by ticking the box on the left hand side and then click 'Next >>'
5. Find the 'skip' button and click it. You have already registered the sample.

6. Choose the ‘Two Fastq files (Paired)’ radial. You will see various fields that are required. There is also an option to ‘Download Template Spreadsheet’ so that you can fill in your runs and experiments offline. This has already been done for you, the file is called ‘exp_run.tsv’ and it is in the submission_material directory.
7. Select ‘Upload Completed Spreadsheet’ and choose the file ‘exp_run.tsv’



8. The fields should load. Fields like insert size, library strategy, library source, library selection and instrument and platform details should be filled in accurately because all these can be filter criteria when searching the ENA. Also note the ‘Sample reference’ column and recall that the experiment has a pointer to the sample. When we submitted the sample we called it ‘dough_sample’ so now we are simply referencing the sample that we created in the previous step. Click ‘Submit’ and ‘OK’ to confirm.

You should now have an experiment object and a run object. You can check the ‘Experiments’ and ‘Runs’ tabs respectively. What are their accessions?

For all objects submitted up to now there will be an ‘Edit XML’ or an ‘Edit’ option. Click on this to see how how the tab file/webform was converted into XML. Can you answer these questions?

1. What pointers does the experiment have?
2. What pointers does the run have?
3. What pointers does the sample have?
4. The run is pointing to the sample with alias “dough_sample”. How would you go about changing it to a different sample?
5. All data and objects are submitted under the same study. How would you change it so that the run, sample, and experiment come under a different study?

Adding a contig set

In the submission_material directory you will find a file called ‘dough_contigs.fna.gz’ this is a fasta file of contigs not unlike you may have after running your own assembly pipeline on your raw data. You can archive the assembly with the ENA.

If you are at a Mac or Linux command line, check the contents of the file

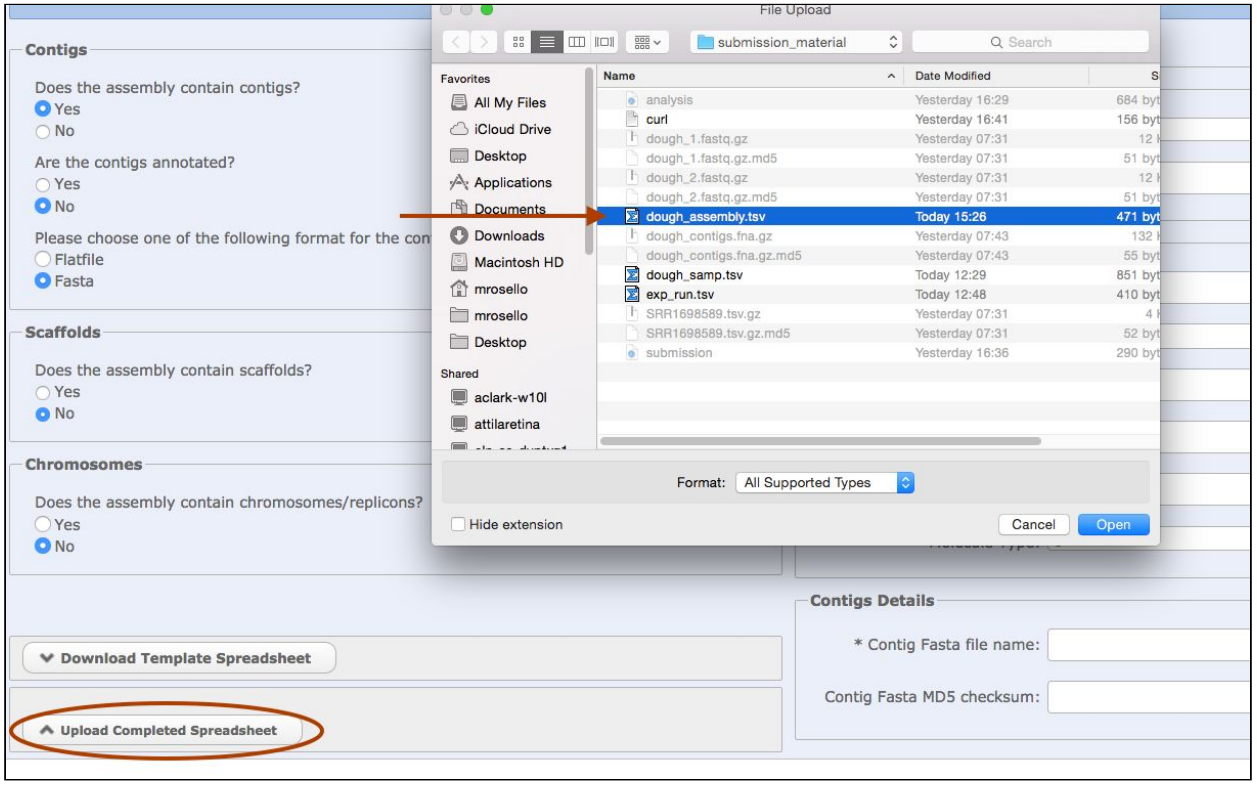
```
gunzip -c dough_contigs.fna.gz | head

gunzip -c dough_contigs.fna.gz | grep ">"
```

This set only has 5 contigs, the purpose of the file is to learn how to submit it and add it to the sample and study that you have already created.

1. In Webin go back to the 'New Submission' tab and select the 'Submit genome assemblies' radial before clicking 'Next >>'
2. Select the existing study that you submitted earlier, tick the box, and click 'Next >>'
3. Find the 'skip' button as we already have a source sample that has been registered.
4. You will find a questionnaire in the left panel. Select these options:

Does the assembly contain contigs? **Yes**
Are the contigs annotated? **No**
Please choose one of the following format for the contigs submission **No**
Does the assembly contain scaffolds? **No**
Does the assembly contain chromosomes/replicons? **No**

5. On the left hand panel you will find some fields to apply. As with previous submissions there is an option to download a spreadsheet template to fill in offline and then re upload it. In the submission_material directory you will find a file called 'dough_assembly.tsv' which is pre prepared. Select 'Upload Completed Spreadsheet' and navigate to 'dough_assembly.tsv'



6. All the fields should be completed. The file 'dough_contigs.fna.gz' is already in the ftp directory for the account you are using. When you used ftp a few steps ago you used wild card '*' to upload all files beginning with 'dough' so this file and its md5 sum file will have been included. If you did not use ftp in the previous steps it is fine because all files were uploaded to each workshop account ftp directory in advance.
7. Click 'Submit' and then 'OK' to confirm. You should now have an assembly registered. Check the 'Assemblies' tab (you may need to use the 'Refresh' option. You will find an accession that looks like ERZXXXXX. This is actually an analysis accession because the analysis object is used to register assembly submissions. For assemblies do not use the analysis accession in any publication. Wait until the contigs are processed and converted and you will be emailed with the prefix accession range and

you will also obtain an accessions that looks like GCAXXXXXXX which reoesents the whole assembly (including scaffolds and chromosomes) and you can use this to publish your research.

Submit a tab file as an analysis object

In the ssubmission_material directory you will find a file SRR1698589.tsv.gz representing the results from a taxonomic profiling analysis done on the run that you submitted previously. The file lists each read and assigns the highest level taxonomic classification that is possible based on the public database. This file is a much shortened version so that it is fast to upload and submit. You will obtain files like this and other files from your analysis pipelines or you may have them provided as a service, for instance by the EBI Metagenomics team. To create file SRR1698589.tsv.gz I used Taxonomer tool which is very fast and if you have submitted your run to the ENA you do not need to upload the read file, you can simply provide the run accession. The EBI Metagenomics team will also analyse your read files providing you have submitted them to the ENA and you have applied sufficient annotation to the sample object (as per this tutorial).

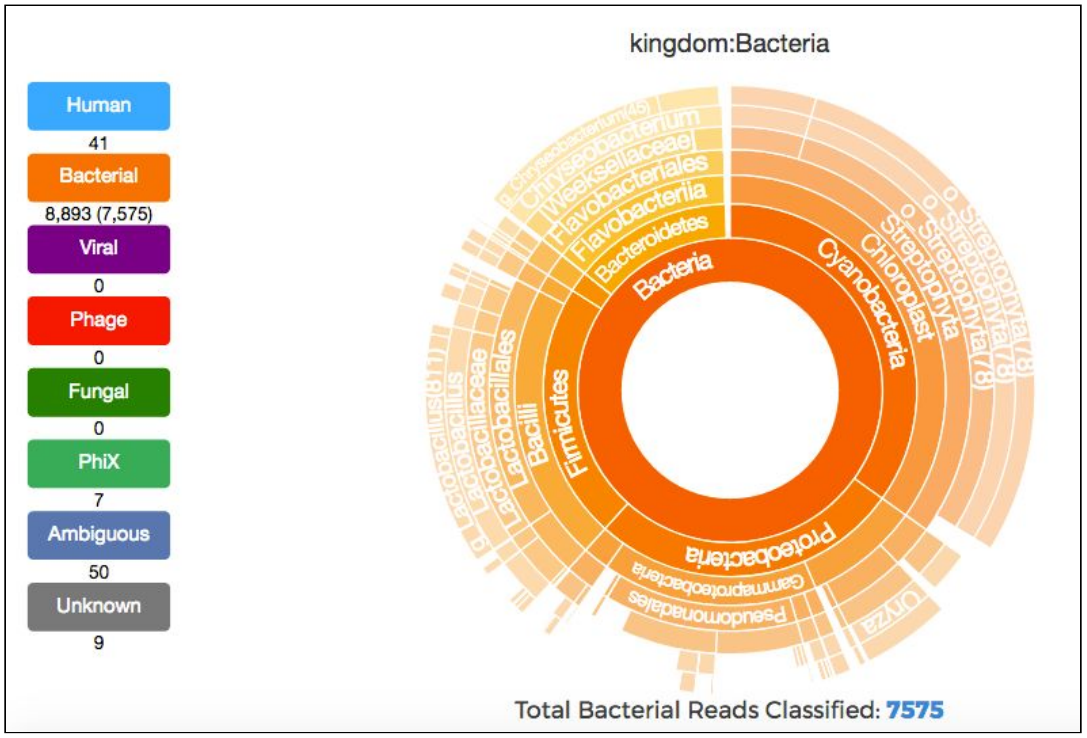


Figure 9: After this 3 day workshop you will have further means of producing analysis files which you can then archive back into the ENA with your original raw data to supplement your publication

The submission service can not be used to submit analysis files as analysis objects but you can use the ENA REST API. You will need a Linux or Mac command line to do the following. If you do not have access you can skip it.

1. Check the file contents.

```
gunzip -c SRR1698589.tsv.gz | head -n20
```

2. You can use the ftp to upload the file to ftp directory as you did for the experiment and run step (se above). But to save time the file and its md5 sum file have already been put on the ENA ftp servers for all the workshop accounts
3. View the analysis.xml file that is in the submission_material directory. Note how the study pointer is not complete (`<STUDY_REF accession=""/>`) whereas the sample pointer is using the alias that we applied to the sample, earlier (`<SAMPLE_REF refname="dough_sample"/>`)

```
cat analysis.xml
```

4. During the study submission you may recall filling out a web form. There was no option to provide an alias for the study but you now know the accession. Check the ‘Studies’ tab in Webin and look for the *PRJEBXXXX* accession number.

5. Add the *PRJEBXXXX* accession to the study pointer in the analysis object (change the 'xxxx; for the actual accession!):

Mac:

```
sed -i '.orig' 's/accession=""/accession="PRJEBxxxx"/' analysis.xml
```

Linux:

```
sed -i 's/accession=""/accession="PRJEBxxxxx"/' analysis.xml
```

6. Take a look at the submission.xml file that is also in the submission_material directory. This simply marks the submission event and should accompany any submission that is being carried out via the REST API.

```
cat submission.xml
```

7. We will use a command line program called cURL to send the submission xml and the analysis xml to ENA test server so that it can be registered and accessioned. This command references the 2 xml files, the Webin account (change the 'XXXX' to the Webin account number you are using (from here), and the password (which is set to 'ws2017' for all workshop accounts)

```
curl -k -F "SUBMISSION=@submission.xml" -F "ANALYSIS=@analysis.xml"
"https://www-test.ebi.ac.uk/ena/submit/drop-box/submit/?auth=ENA%20Webin-XXXX%2
0ws2017"
```

8. After running the cURL command you will receive a receipt a bit like the one below. If you see the sting `success="true"` then you will also find an accession number for the new analysis object that you have created. You can use this accession to reference your analysis table when you publish your research. It will also be visible under the study on the ENA website when the study goes live.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="receipt.xsl"?>
<RECEIPT receiptDate="2017-06-19T16:06:12.224+01:00"
submissionFile="submission.xml" success="true">
     <ANALYSIS accession="ERZ458147" alias="dough_taxonomy" status="PRIVATE"/>
     <SUBMISSION accession="ERA961805" alias="dough_taxonomy"/>
     <MESSAGES>
          <INFO>All objects in this submission are set to private status
(HOLD).</INFO>
          <INFO>Submission has been committed.</INFO>
     </MESSAGES>
     <ACTIONS>ADD</ACTIONS>
     <ACTIONS>HOLD</ACTIONS>
</RECEIPT>
```

Task 5: The ENA Browser

The bread dough study from the previous section is based on an actual study in the ENA. The one that you submitted will not become public but you can see this one to have an idea of what the result would be.

The table at the bottom of the page is run centric - for each run in the study (only one in this case) you will find another row in the table and because the table is using the read domain the available columns are mostly run and experiment fields. You can view different fields using the 'Select columns' option and you can also download it as a report using the 'download results in TEXT'

File reports

A lot of ENA users are searching and downloading in bulk. The file report service provides a tab delimited run centric reports. These are like the tables that appear at the bottom of ENA webpages but they are parseable (with a bash script for example).
So to generate a report for the bread dough study:
http://www.ebi.ac.uk/ena/data/warehouse/filereport?accession=PRJNA268304&result=read_run&fields=run_accession,fastq_ftp,fastq_md5,fastq_bytes
The above report provides the ftp location of the files and their md5sum so that you can check the integrity of the files after they have been downloaded. All these actions can be scripted. For instnace, the below command uses cURL to download a file report for the bread dough study (into a file called report.txt).

```
curl
"http://www.ebi.ac.uk/ena/data/warehouse/filereport?accession=PRJNA268304&result=read_run&fields=run_accession,fastq_ftp,fastq_md5,fastq_bytes" -o
"report.txt"
```

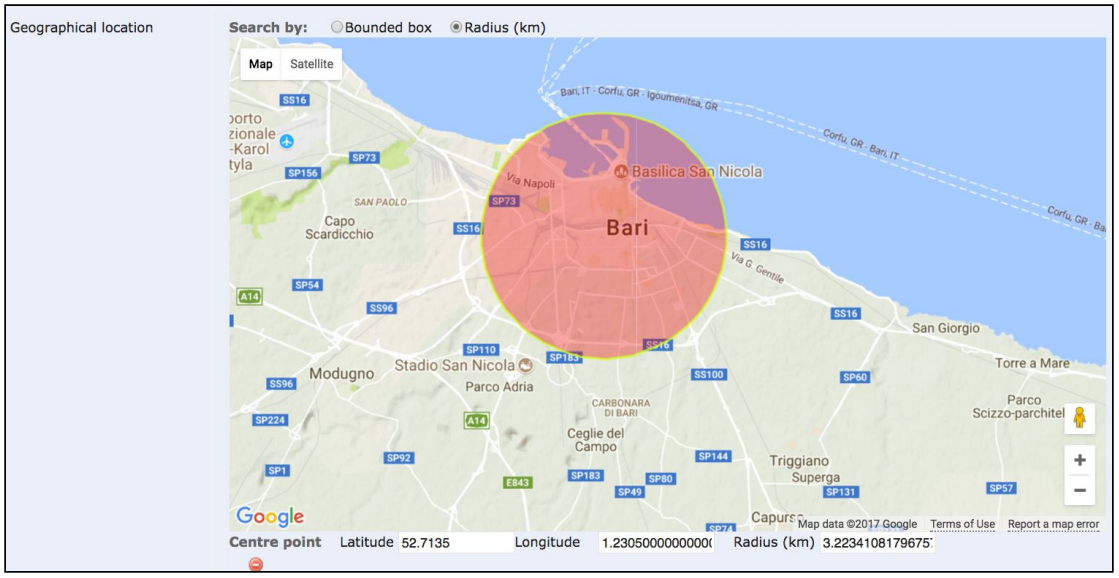
A script can be used to parse the report and download each file. For instance here is he cURL command for downloading the forward read file from the bread dough study into a file called forward.fq.gz.

```
curl "ftp.sra.ebi.ac.uk/vol1/fastq/SRR169/009/SRR1698589/SRR1698589_1.fastq.gz"
-o "forward.fq.gz"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  1  236M    1 4480k    0     0   158k      0  0:25:27  0:00:28  0:24:59  218k
```

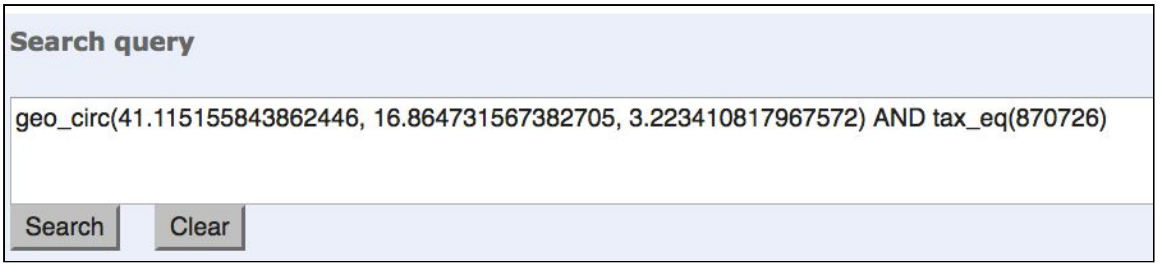
Advanced Search: Sample domain

The main purpose of the ENA is to archive data that is searchable. Most searches are based on the sample domain because this is the point of comparison and context for the data. Recall when submitting the sample object in the previous section we applied a GPS location. While there are many fields to search by, location is very common one.

The ENA advanced search portal is here. Lets try to find similar samples to the one that we submitted. Where it says ‘Select domain:’ choose the **sample** radial. Then type ‘food metagenome’ in the **Taxon name** box and draw a circle around Bari in Italy.



Scroll back up to the top of the page and click ‘search’



This creates a URL that looks a bit like this:
http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22**geo_circ**(**41.115**,%20**16.864**,%20**4**)%20**AND**%20**tax_eq**(**870726**)%22&**domain=sample**
The URL is sending instructions to the ENA data warehouse to search all sample GPS locations that fall within a radius of 4km from GPS point 41.115,16.864 AND that have a tax id of 870726 (food metagenome). You can see it is possible to do searches by building specific URLs like this.

Paste the URL above into your browser and press enter. How many samples are found that are similar to the one that you submitted earlier today? You can change the URL slightly to make it produce a report:
http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22geo_circ(41.115,%2016.864,%204)%20AND%20tax_eq(870726)%22&result=sample&fields=sample_alias&display=report
This is how many ENA users run regular searches for data that is of interest to them. They can run a search like this regularly and parse the results to see if any new samples have been added since the previous time. Then they can use the file reporting service mentioned above to find all the files connected to the new samples. Next they will down load the standardised versions of the files and apply them to their analysis pipelines. This can all be done automatically and programmatically.

Paste the URL above into your browser. Can you find the sample that you submitted (bread dough)? What is the accession?

Advanced Search: Read domain

Recall the American Gut Project discussed earlier in the tutorial. The ENA data warehouse can be used to apply filtering based on most fields. You can do this using the GUI but you can also build the URL yourself as mentioned previously. For example, using the read domain this time (instead of the sample domain) we can expect to search or filter by fields related to the experiment and run objects. Fields such as platform information, library details, file size etc ….

Here is a query that looks at file size and read count:
http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22secondary_study_accession=%22ERP012803%22%22&result=read_run&fields=run_accession,read_count,fastq_bytes,instrument_platform&display=report

You will see that some files are very small and you would not like to include them in your analysis. It is possible to use logical operators AND, OR and NOT in these data warehouse queries. So for instance, you can specify a filter for all runs that have files with read counts that are at least 24000:
http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22secondary_study_accession=%22ERP012803%22%20AND%20read_count>24000%22&result=read_run&fields=run_accession,read_count,fastq_bytes,instrument_platform&display=report

Note the ‘read_count>24000’ clause in the second URL.

Can you answer this:

How many runs have been removed from the original list after apply the filter? You can do this on the mac or Linux command line:

```
curl
"http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22secondary_study_access
ion=%22ERP012803%22%22&result=read_run&fields=run_accession,read_count,fastq_by
tes,instrument_platform&display=report" -o "ERP012803.txt"

curl
"http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22secondary_study_access
ion=%22ERP012803%22%20AND%20read_count>24000%22&result=read_run&fields=run_acce
ssion,read_count,fastq_bytes,instrument_platform&display=report" -o
"ERP012803_filter.txt"

wc -l ERP012803.txt

wc -l ERP012803_filter.txt
```

The above 4 commands downloads both reports to 2 different files and then does a line count on each file to see the difference.

Try this:
Try changing the URL to look for a very small run and then try downloading the read files for that run using cURL. If the run is very small it should not take long to download!

Taxonomy Portal

Because most metagenomics studies can not provide organism classifications to the archived data because no organisms are isolated (they are usually environmental samples) you may think that ENA taxonomy portal can not be made use of. But actually there are a number of special taxonomic classifications that are available to assign to environmental samples. This means that you can use ENA's taxon portal, which is part of the data warehouse/advanced search:
http://www.ebi.ac.uk/ena/data/warehouse/search?portal=taxon

Here is a list of some taxonomic classififcations available for assigning to metagenomics samples:
http://www.ebi.ac.uk/~mrosello/FAQs/env_tax_nodes.txt

Try adding 'food metagenome' to the search box (make sure the 'taxon' radial is selected) and clicking 'search'

Then, when it is displayed click on the '870726' tax id and look in the 'portal' tab
Can you answer these questions?

1. How many sample objects have been registered with this tax id?
2. How many protein coding sequences (coding domain) have been assigned this tax id?

3. Go to one of the coding sequences and find what product it is. I found product “proteolysis tag peptide encoded by tmRNA Weiss_koree_KAC155” from this record.

Explore the other domains in the ENA!

You will notice that there are many interesting domains that have been used to categorise and sort data in the ENA. This tutorial has only brushed the surface. Go ahead and do some searches across some of the other domains that are available from here.