

# Metadata handling: Part II

Marco Moretto  
marco.moretto@fmach.it

Fondazione Edmund Mach  
Research and Innovation Centre  
Computational Biology Unit

5th February 2020



# Data submission: EBI-ENA

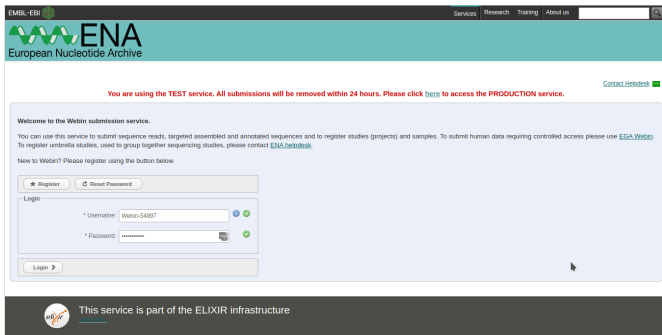


*The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing*

## Useful links

- <https://ena-docs.readthedocs.io/en/latest/>
- <https://wwwdev.ebi.ac.uk/ena/submit/sra>

# Data submission: EBI-ENA



The screenshot shows the EBI-ENA (European Nucleotide Archive) Webin submission service interface. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below this is the EBI-ENA logo and the text 'European Nucleotide Archive'. A warning message states: 'You are using the TEST service. All submissions will be removed within 24 hours. Please click [here](#) to access the PRODUCTION service.' Below the warning, there is a 'Contact Helpdesk' link. The main content area is titled 'Welcome to the Webin submission service.' and provides instructions on how to use the service. It includes a 'New to Webin? Please register using the button below.' section with 'Register' and 'Reset Password' buttons. Below this is a 'Login' section with fields for 'Username' (containing 'Webin-54997') and 'Password' (containing 'Methada2020'). There are also 'Login' and 'Forgot Password' links. At the bottom, there is a footer with the EBI logo and the text 'This service is part of the ELIXIR infrastructure'.

EMBL-EBI

Services Research Training About us

ENA  
European Nucleotide Archive

You are using the TEST service. All submissions will be removed within 24 hours. Please click [here](#) to access the PRODUCTION service.

[Contact Helpdesk](#)

Welcome to the Webin submission service.

You can use this service to submit sequence reads, targeted assembled and annotated sequences and to register studies (projects) and samples. To submit human data requiring controlled access please use [EGA Webin](#). To register umbrella studies, used to group together sequencing studies, please contact [ENA helpdesk](#).

New to Webin? Please register using the button below.

Register Reset Password

Login

\* Username: Webin-54997

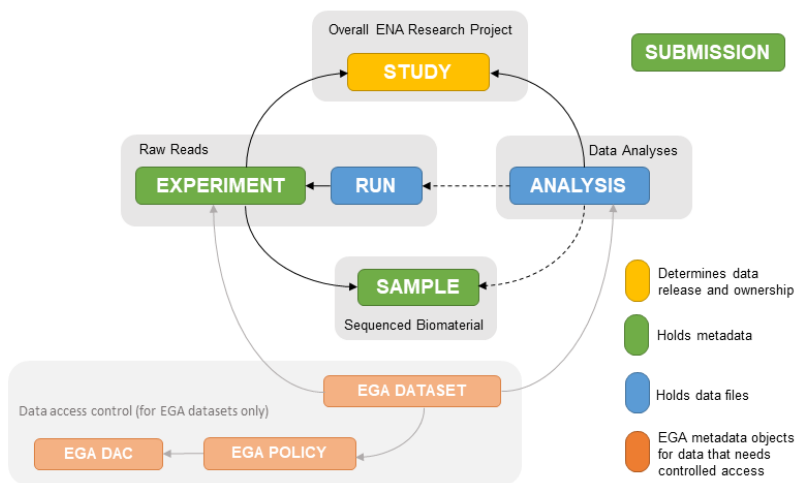
\* Password: Methada2020

Login

This service is part of the ELIXIR infrastructure

username: Webin-54997  
password: Methada2020


# Data submission: EBI-ENA




# Data submission: EBI-ENA

EMBL-EBI

Services Research Training About us

**ENA**  
European Nucleotide Archive

[Contact Helpdesk](#)  [Webin-54997](#) [Logout](#)

**You are using the TEST service. All submissions will be removed within 24 hours. Please click [here](#) to access the PRODUCTION service.**

Home	New Submission	Studies	Samples	Runs	Analyses
Start	>>	Study	>>	Finish	

You can use this service to submit [sequence reads](#), [targeted assembled and annotated sequences](#) and to register [studies \(projects\)](#) and [samples](#). To register umbrella studies, used to group together sequencing studies, please contact [ENA helpdesk](#).

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

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

Next >>

# Data submission: EBI-ENA

Home	New Submission	Studies	Samples	Runs	Analyses
------	----------------	---------	---------	------	----------

Start

>>

Sample

>>

Finish

You can use this service to submit [sequence reads](#), [targeted assembled and annotated sequences](#) and to register [studies \(projects\)](#) and [samples](#). To register umbrella studies, used to group together sequencing studies, please contact [ENA helpdesk](#).

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

- ☐ Submit sequence reads and experiments
- ☐ Register study (project)
- ☐ Taxonomy Check/Request
- ☒ **Register samples**  
Register Samples to give context to your data.  
Read [here](#) for information on how to submit your Samples to ENA.
- ☐ Submit other assembled and annotated sequences [formerly EMBL-Bank]

Next >>

 [Restart Submission](#)

# Data submission: EBI-ENA

Home


New Submission


Studies

Samples

Runs













Analyses

Start  >> Sample >> Finish


 Please add samples to the submission. Multiple samples can be created by increasing the number by the add button

+ Add

1 samples

<input checked="" type="checkbox"/>	sample_10		
<input type="checkbox"/>	sample_11		
<input type="checkbox"/>	sample_12		
<input type="checkbox"/>	sample_13		
<input type="checkbox"/>	sample_14		
<input type="checkbox"/>	sample_15		

1 of 6



 Please submit by clicking the **Submit** Button. Alternatively, download your data as a spreadsheet using the **Download Spreadsheet** button. Once you have filled the spreadsheet please restart the submission process and upload the spreadsheet using the **Upload Completed Spreadsheet** button.



Download Spreadsheet



< Previous Sample

Next Sample >


Basic Details


\* Unique Name: sample\_10  



\* Title: berry\_ei36\_1  



Description:  



Organism Details

 If your organism is not found please go [here](#)



Search: 



\* Tax Id: 29760  

\* Scientific Name: Vitis vinifera  

Common Name: wine grape  

Part and developmental stage of organism

dev\_stage: EI\_36  

organism part: berry fruit  

<< Previous

Submit

# Data submission: EBI-ENA

Home	New Submission	Studies	Samples	Runs	Analyses
Start	>> Study	>> Sample	>> Run	>> Finish	

You can use this service to submit [sequence reads](#), [targeted assembled and annotated sequences](#) and to register [studies \(projects\)](#) and [samples](#). To register umbrella studies, used to group together sequencing studies, please contact [ENA helpline](#).

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

- ☒ Submit sequence reads and experiments

We recommend that Fastq, BAM, and CRAM read files are submitted using [Webin-CLI](#).

When using this interface instead of [Webin-CLI](#), raw sequences must be [uploaded](#) in one of the supported [data formats](#) before they can be submitted. All data submitted in a single submission will be associated with the same study. Data for different studies must be submitted in separate submissions. The study and the sequenced samples can be either pre-registered or registered during the submission process. Please note that each individual study and sample should be registered only once. In addition, you will be asked to provide information about the sequencing libraries and instruments. Please quote the study accession number (ERP\*) when citing data submitted to ENA.

Read [here](#) for more information on how submit your raw reads to ENA.

- ☐ Register study (project)
- ☐ Taxonomy Check/Request
- ☐ Register samples
- ☐ Submit other assembled and annotated sequences [formerly EMBL-Bank]

The first step of your submission is to [upload data files](#). Data files can be uploaded using FTP or Aspera, or using the [Webin File Uploader](#). If you have already uploaded your data files into your Webin upload area please proceed directly to the next step. Please note that unsubmitted files that are older than 2 months will be deleted as explained in our [Fair Use Policy](#).

[Restart Submission](#)

Next >>



# Data submission: EBI-ENA

**Get MD5 hash for every fastq we want to submit**

**Copy fastq files to ENA via FTP**

# Data submission: EBI-ENA

**What is MD5 hash and why do we need it?**



# Data submission: EBI-ENA

## FTP - Filezilla

File Edit View Transfer Server Bookmarks Help

Host: webin.ebi.ac.uk Username: Webin-54997 Password: \*\*\*\*\* Port: Quickconnect

Status: File transfer successful, transferred 55 B in 1 second  
Status: Retrieving directory listing of "/...  
Status: Directory listing of "/" successful  
Status: Disconnected from server  
Status: Connection closed by server

Local site: /home/moretton/Dropbox/COST\_Valencia\_2020/Day1/Metadata\_2/ Remote site: /

Local site directory structure:

- COST2019
- COST\_Valencia\_2020
  - Day1
    - Jupyter
      - Metadata\_1
      - Metadata\_2
    - Day3

Local site file list:

Filename	Filesize	Filetype	Last modified
sample_1_R1.fastq.gz.md5	55 B	md5-file	22/01/2020 10:...
sample_1_R2.fastq.gz	639 B	gz-file	22/01/2020 10:...
sample_1_R2.fastq.gz.md5	55 B	md5-file	22/01/2020 10:...
sample_1_R1.fastq.gz	639 B	gz-file	22/01/2020 10:...
sample_1_R1.fastq.gz.md5	55 B	md5-file	22/01/2020 10:...
sample_1_R2.fastq.gz	639 B	gz-file	22/01/2020 10:...
sample_1_R2.fastq.gz.md5	55 B	md5-file	22/01/2020 10:...

Remote site file list:

Filename	Filesize	Filetype	Last modified	Permissions	Owner/Grou
sample_1_R1.fastq.gz	639 B	gz-file	22/01/2020 ...	-rwxrwxrwx	4262 1098
sample_1_R1.fastq.gz.md5	55 B	md5-file	22/01/2020 ...	-rwxrwxrwx	4262 1098
sample_1_R2.fastq.gz	639 B	gz-file	22/01/2020 ...	-rwxrwxrwx	4262 1098
sample_1_R2.fastq.gz.md5	55 B	md5-file	22/01/2020 ...	-rwxrwxrwx	4262 1098
sample_2_R1.fastq.gz	639 B	gz-file	22/01/2020 ...	-rwxrwxrwx	4262 1098
sample_2_R1.fastq.gz.md5	55 B	md5-file	22/01/2020 ...	-rwxrwxrwx	4262 1098

# Data submission: EBI-ENA

You are using the TEST service. All submissions will be removed within 24 hours. Please click [here](#) to access the PRODUCTION service.

Home | New Submission | Studies | Samples | Runs | Analysis

Start ✓ | Study ✓ | Sample ✓ | Run | Finish

Please provide library, instrument and data file details by uploading a spreadsheet or by editing the table below.  
Please select the file format. If you have files of different types please submit them in separate submissions.

☐ CRAM  
☐ BAM  
☐ SFF  
☐ One Fastq file (Single)  
☒ Two Fastq files (Paired)

Two fastq files containing paired reads are submitted for each run. All technical sequences including adaptor sequences, linker sequences and barcode sequences must be removed from the reads before submission. The first reads must be in the first Fastq file and the second reads must be in the second Fastq file ordered in the same order as in the first file.

☐ Complete Genomics  
☐ PacBio HDP5  
☐ Oxford Nanopore

Mandatory fields are denoted by (\*).

Download Template Spreadsheet | experiment paired\_fastq\_spa... Done | Download Spreadsheet

	[Sample reference suggestions]	Sample reference (*)	Instrument Model (*)	Library Name (*)	Library Source (*)	Library Selection (*)	Library Strategy (*)	Details
✗		ERR5428802	Illumina HiSeq 250	unspecified	TRANSCRIPTOM	RANDOM	RNA-Seq	
✗		ERR5428803	Illumina HiSeq 250	unspecified	TRANSCRIPTOM	RANDOM	RNA-Seq	
✗		ERR5428802	Illumina HiSeq 250	unspecified	TRANSCRIPTOM	RANDOM	RNA-Seq	
✗		ERR5428803	Illumina HiSeq 250	unspecified	TRANSCRIPTOM	RANDOM	RNA-Seq	
✗		ERR5428804	Illumina HiSeq 250	unspecified	TRANSCRIPTOM	RANDOM	RNA-Seq	
✗		ERR5428805	Illumina HiSeq 250	unspecified	TRANSCRIPTOM	RANDOM	RNA-Seq	
+								

# NCBI - SRA and GEO



- <https://www.ncbi.nlm.nih.gov/>
- GEO and SRA
- Two way to search for experiments SRA
  - <https://www.ncbi.nlm.nih.gov/sra/>
  - <https://trace.ncbi.nlm.nih.gov/Traces/study/>
- minimum information is not enforced
- looking for transcriptomics experiments only might be difficult (transcriptomic, gene expression, RNA, tRNA, miRNA, ...)
- data model and ids hierarchy might be difficult to follow (GEO - SRA)

# Complex and Complicated

`http://gigapan.com/gigapans/17217`

# Complex and Complicated

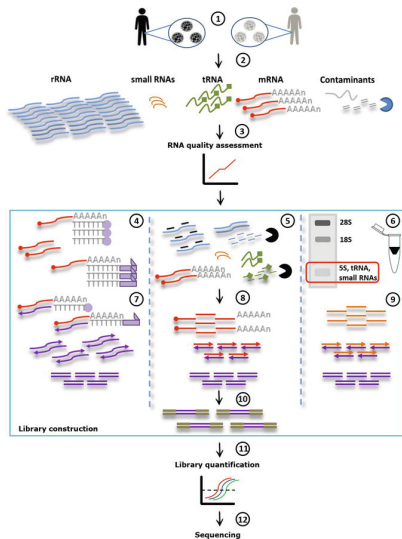




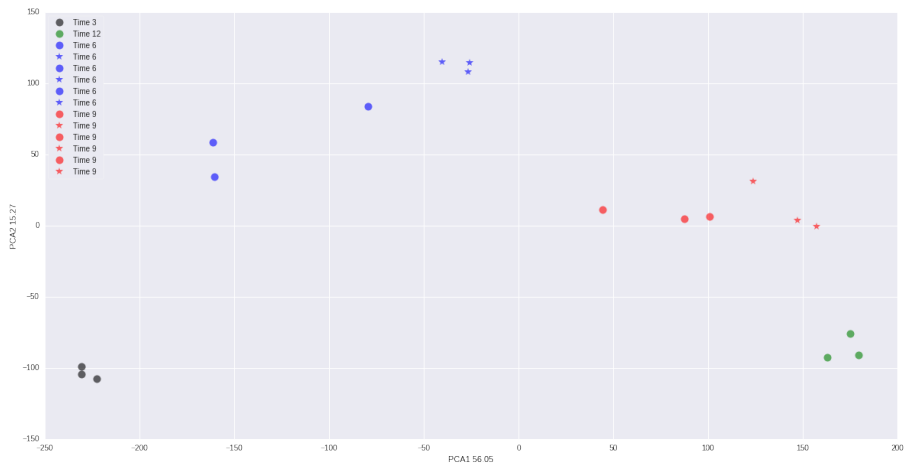
# Complex and Complicated



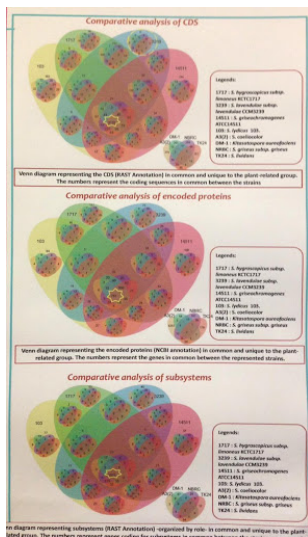
# Complex and Complicated



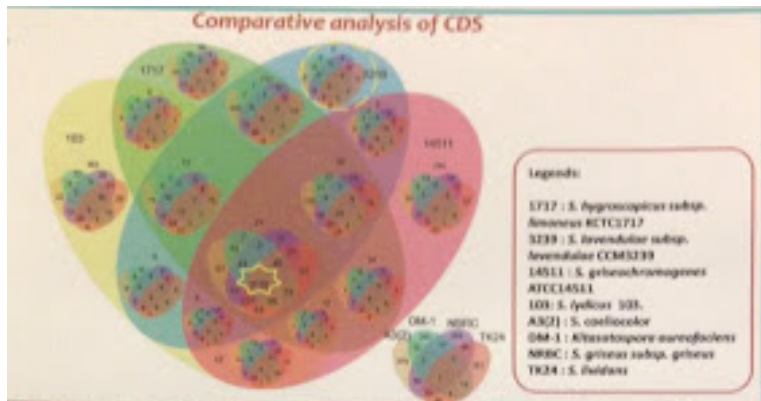
# Complex and Complicated



# Complex and Complicated



# Complex and Complicated



# Complex and Complicated

## Advantages of early metadata submission/handling

- metadata submission files can be used during the analysis;

# Complex and Complicated

## Advantages of early metadata submission/handling

- metadata submission files can be used during the analysis;
- (should) makes you think ahead of time about biological question;

# Complex and Complicated

## Advantages of early metadata submission/handling

- metadata submission files can be used during the analysis;
- (should) makes you think ahead of time about biological question;
- submission has to be done anyway;



# Complex and Complicated

PLOS.ORG PUBLICATIONS



Diverse perspectives on science and medicine

Search PLOS Blogs

STAFF BLOGS

BLOGS BY TOPIC

ABOUT PLOS BLOGS

CONTACT



EveryONE

PLOS ONE community blog

ONE

www.plosone.org

About This Blog

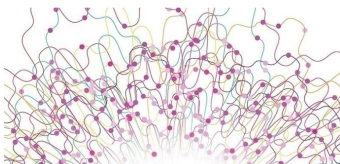
About PLOS ONE

Search This Blog

« Previous

Next »

Share



Sign up for PLOS Updates

Email Address (required)

I have read and agree to the terms of the PLOS Privacy Policy and hereby consent to send my personal information to PLOS.

Sign Up

## Registered Reports are Coming to PLOS ONE

Posted January 14, 2020 by Joerg Heiser in News & Policy



0000-0002-6370-4254

I'm very excited to announce that PLOS ONE will soon offer a new preregistration article type, Registered Reports! The benefits that preregistration can bring to the entire research community tie so closely with PLOS ONE's mission, that we see this as a natural fit for the journal and we're pleased to open this option to our authors.

Publish with  
PLOS ONE

Accelerating  
the publication of  
peer-reviewed science

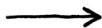
Recent Posts

It's the little things-

# Complex and Complicated



COMPLICATED



COMPLEX

Paul Hughes Live

Complex refers to the number of components in a system, whereas, complicated refers to the level of difficulty of something.

# Complex and Complicated



- Simple is better than Complex
- Complex is better than Complicated

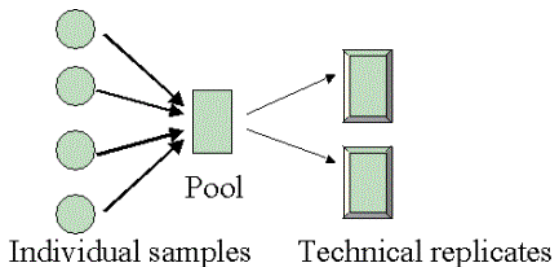
# Complex and Complicated



- Simple is better than Complex
- Complex is better than Complicated
- A simple experiment with lot of replications is easier to **interpret**
- A simple experiment with lot of replications is more **reliable**
- A simple experiment with lot of replications is easier to **reproduce**

# Complex and Complicated

## Biological vs Technical replicates



# Complex and Complicated

## Biological vs Technical replicates





# Complex and Complicated

## Garbage In Garbage Out



Your analysis is as good as your data