Metadata handling: Part II

Marco Moretto
marco.moretto@fmach.it

Fondazione Edmund Mach
Research and Innovation Centre
Computational Biology Unit

5th February 2020

FONDAZIONE EDMUND MACH

METHADA 2020
TRAINING SCHOOL IN TRANSCRIPTOMIC METADATA HANDLING & DATA ANALYSIS
COST Action CA17111 INTEGRAPE

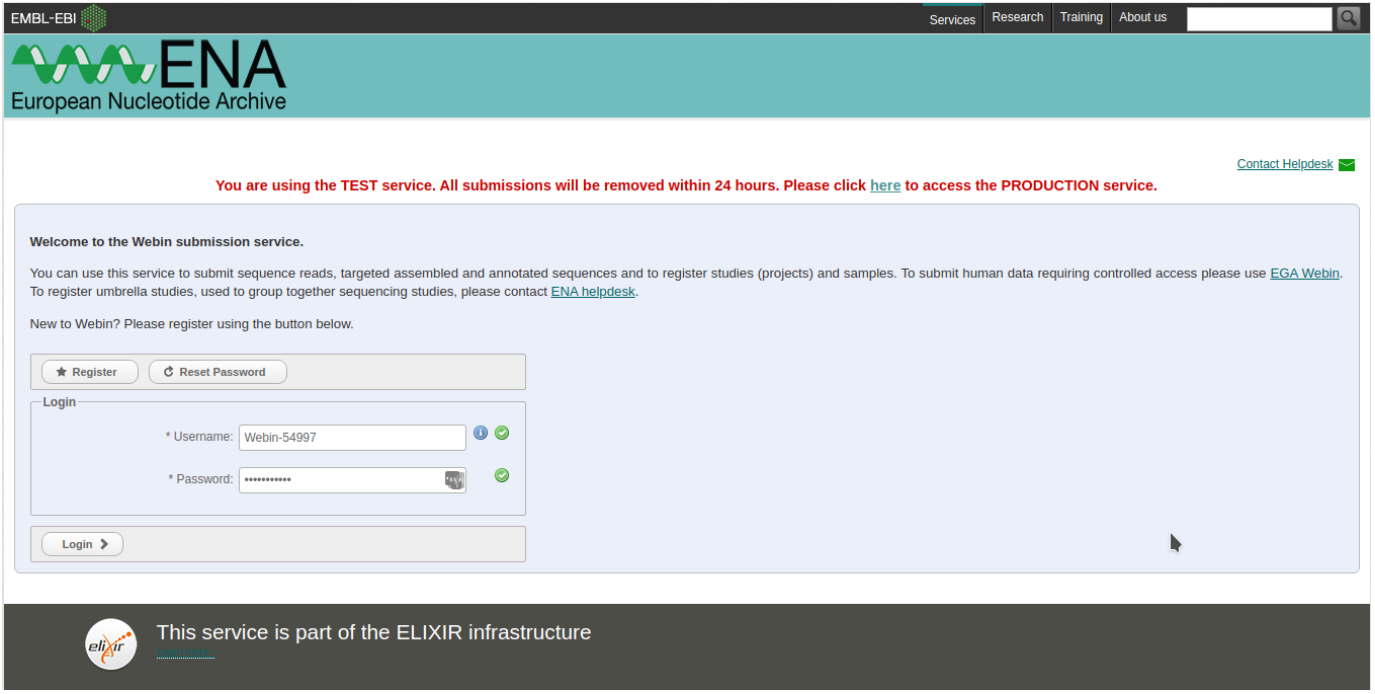
Data submission: EBI-ENA

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

Useful links

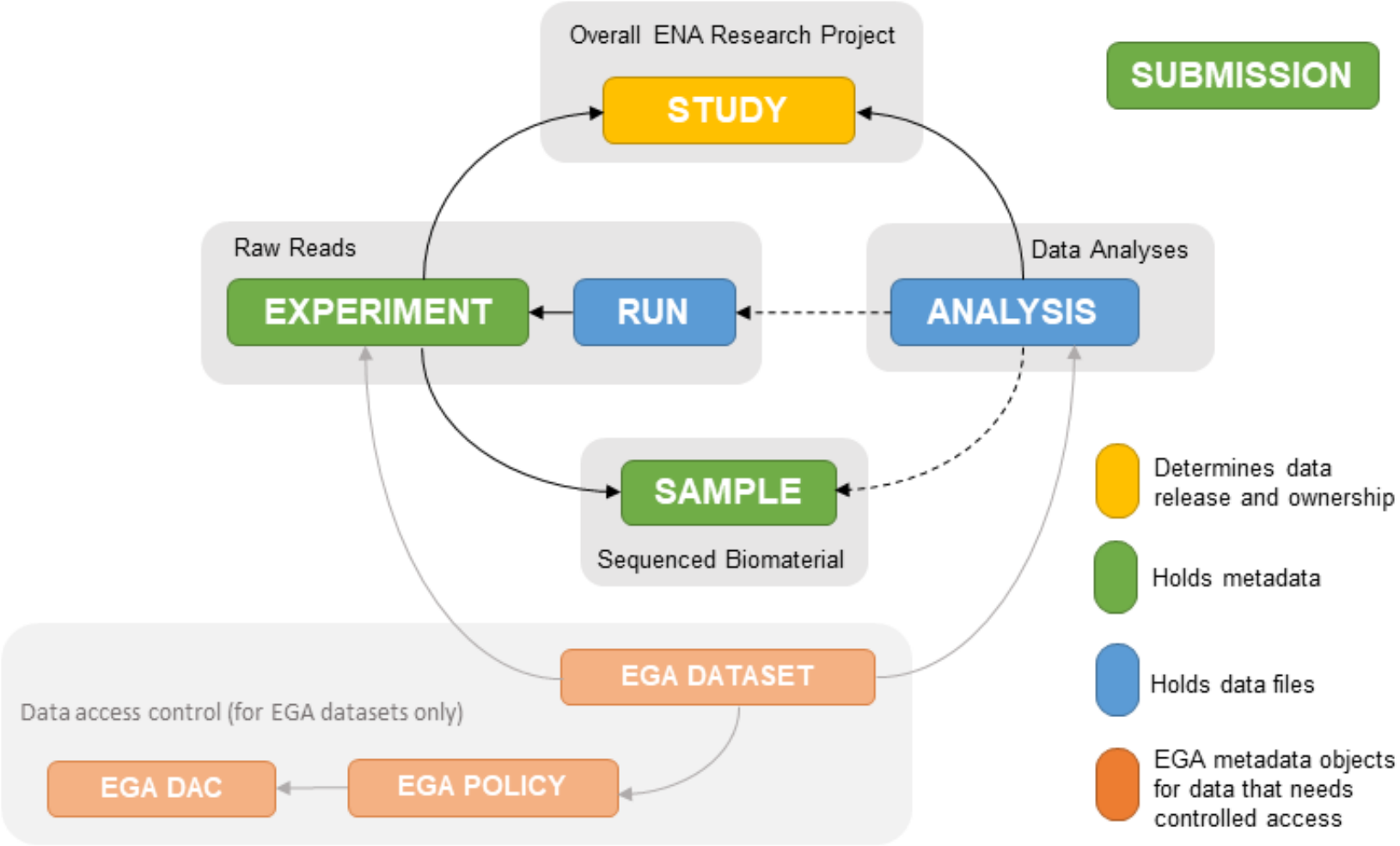
- `https://ena-docs.readthedocs.io/en/latest/`
- `https://wwwdev.ebi.ac.uk/ena/submit/sra`

Data submission: EBI-ENA

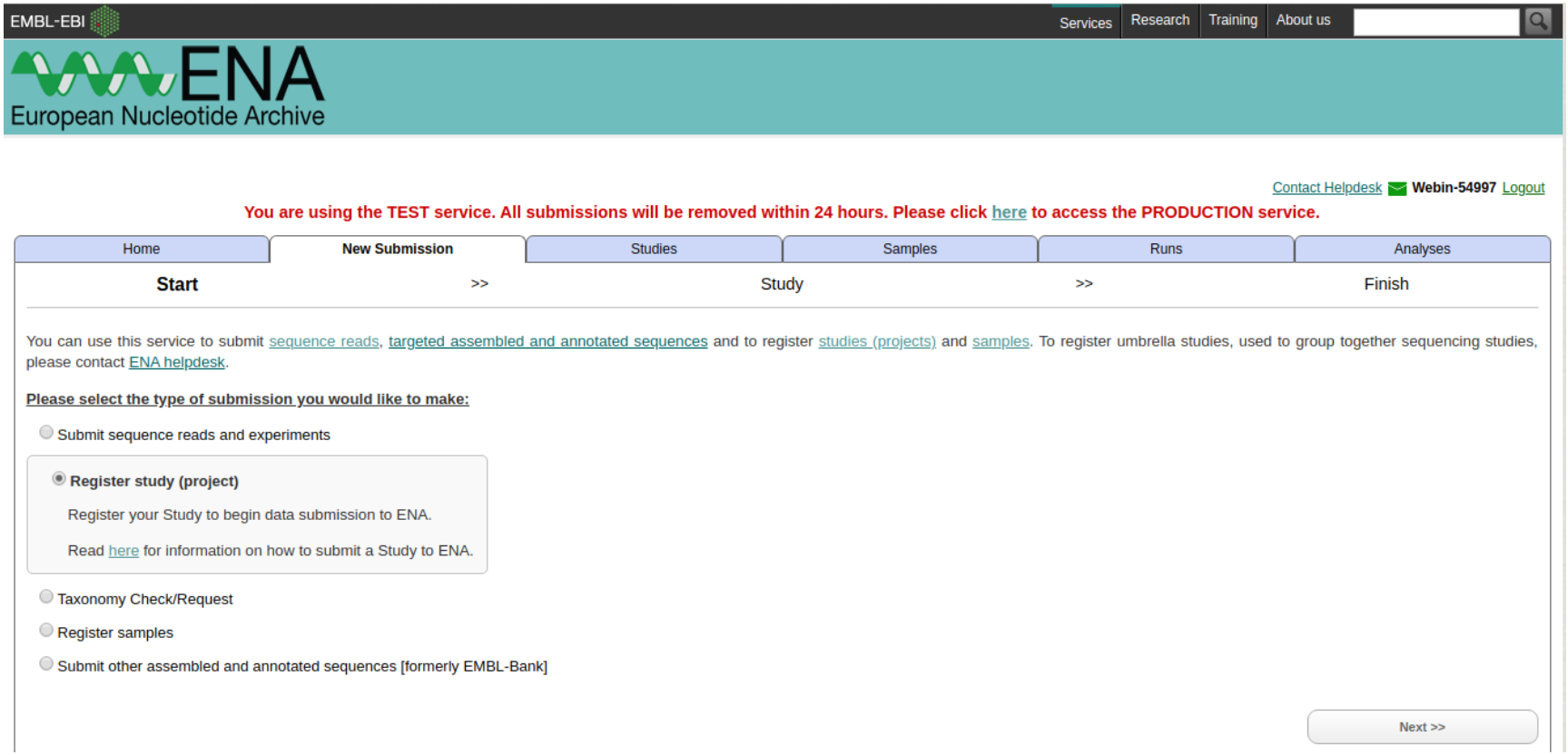


username: Webin-54997
password: Methada2020

Data submission: EBI-ENA



Data submission: EBI-ENA



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

- [x] sample_10
- [ ] sample_11
- [ ] sample_12
- [ ] sample_13
- [ ] sample_14
- [ ] sample_15

1-6 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_el36_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: EL_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 helpdesk.

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.

Next >>

Restart Submission

Data submission: EBI-ENA

Get MD5 hash for every fastq we want to submit

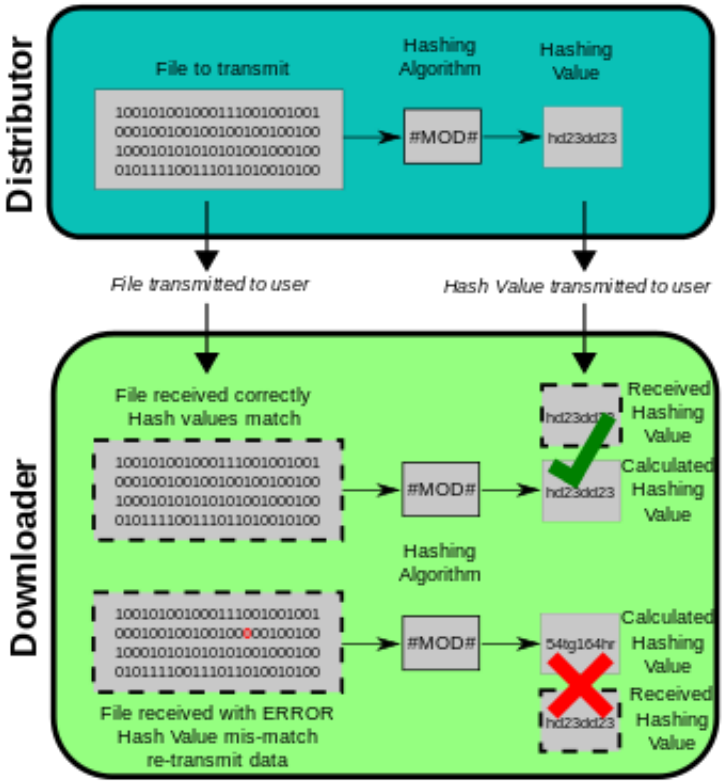
Copy fastq files to ENA via FTP

Data submission: EBI-ENA

What is MD5 hash and why do we need it?

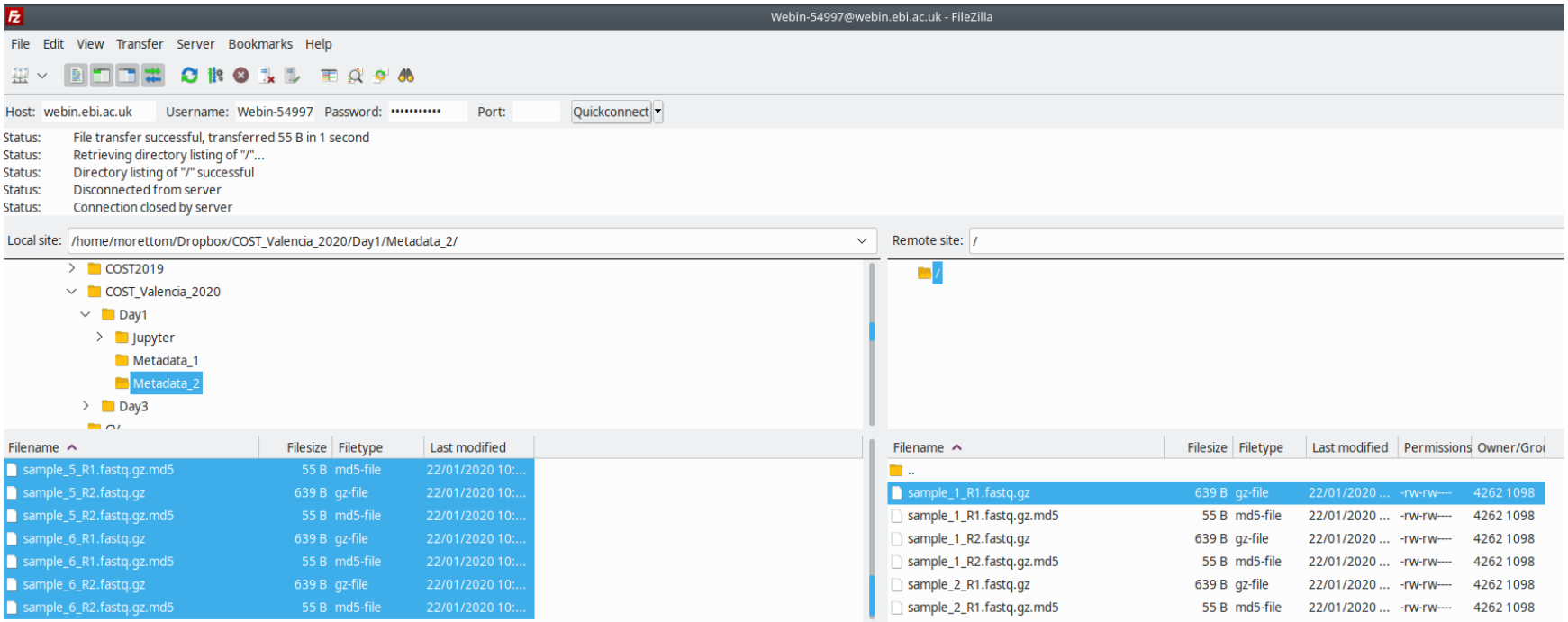
Data submission: EBI-ENA

What is MD5 hash and why do we need it?

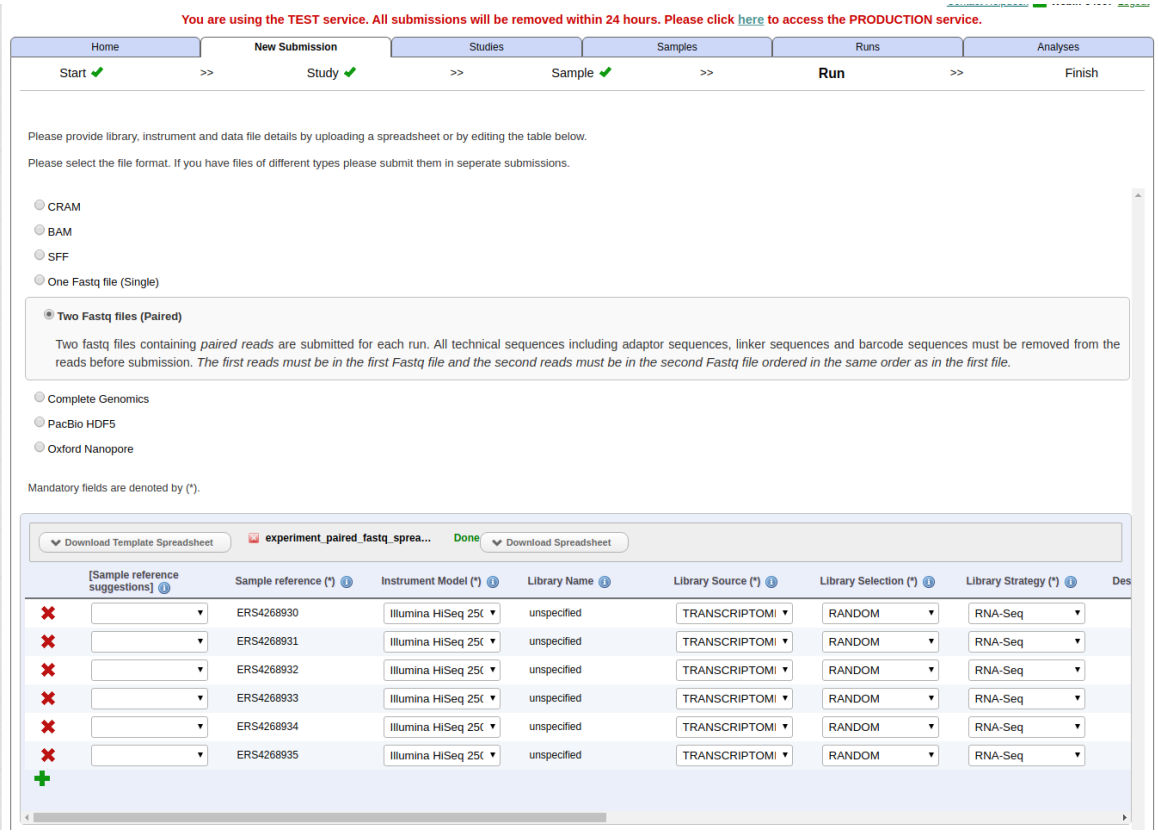


Data submission: EBI-ENA

FTP - Filezilla



Data submission: EBI-ENA



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)

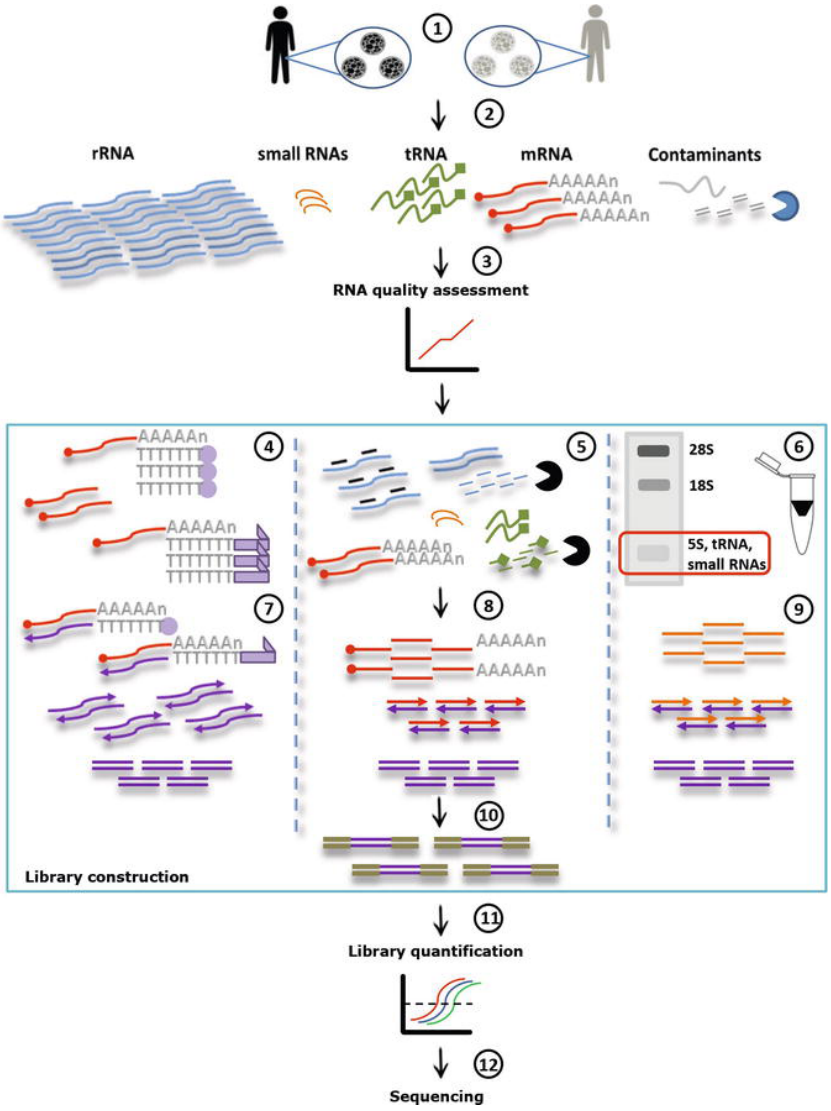
Complex and Complicated

http://gigapan.com/gigapans/17217

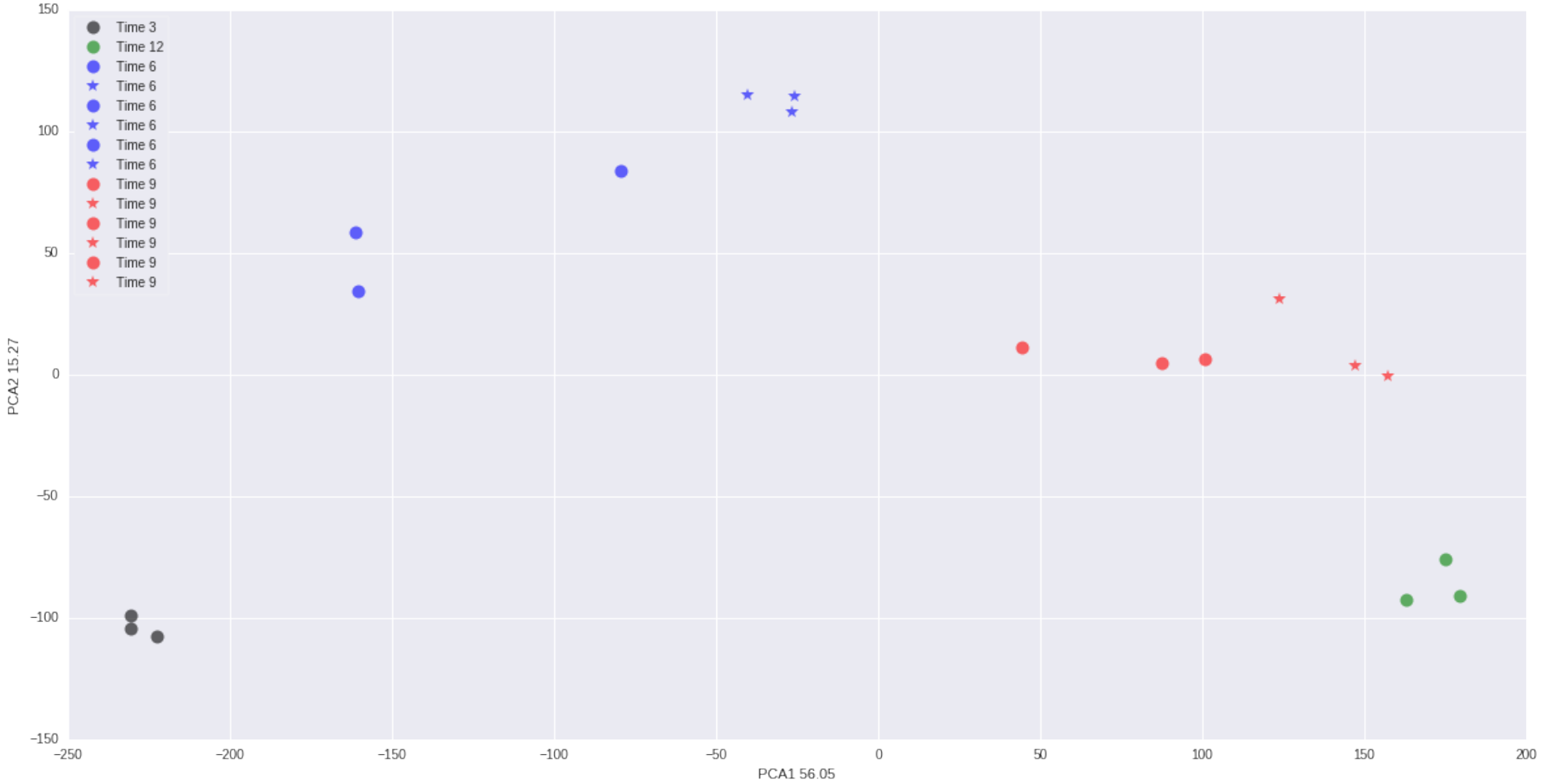
Complex and Complicated

Complex and Complicated

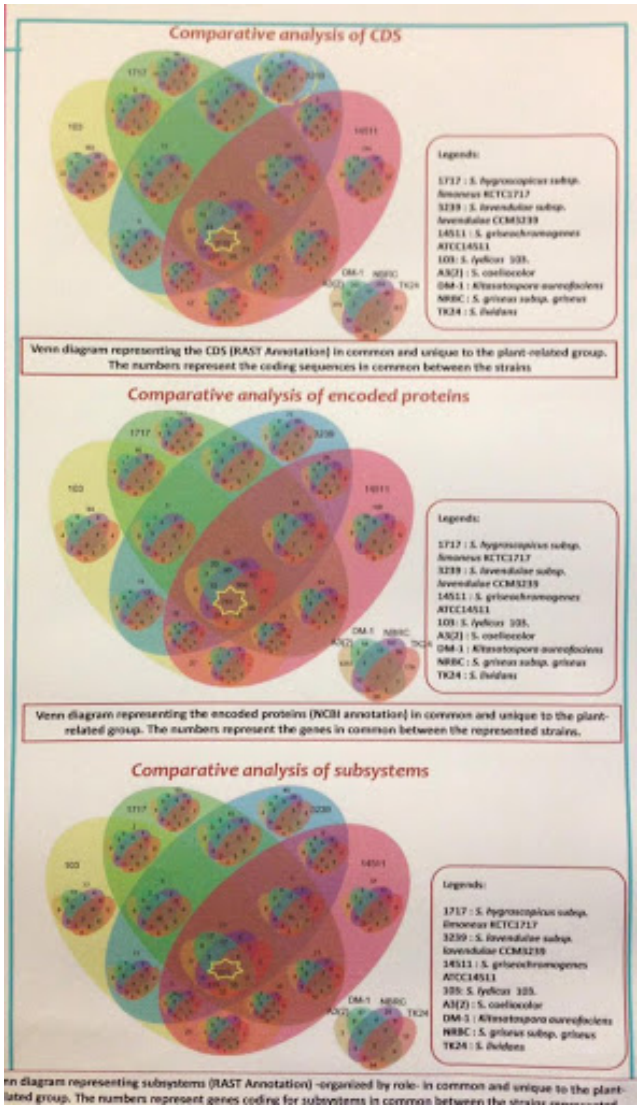
Complex and Complicated



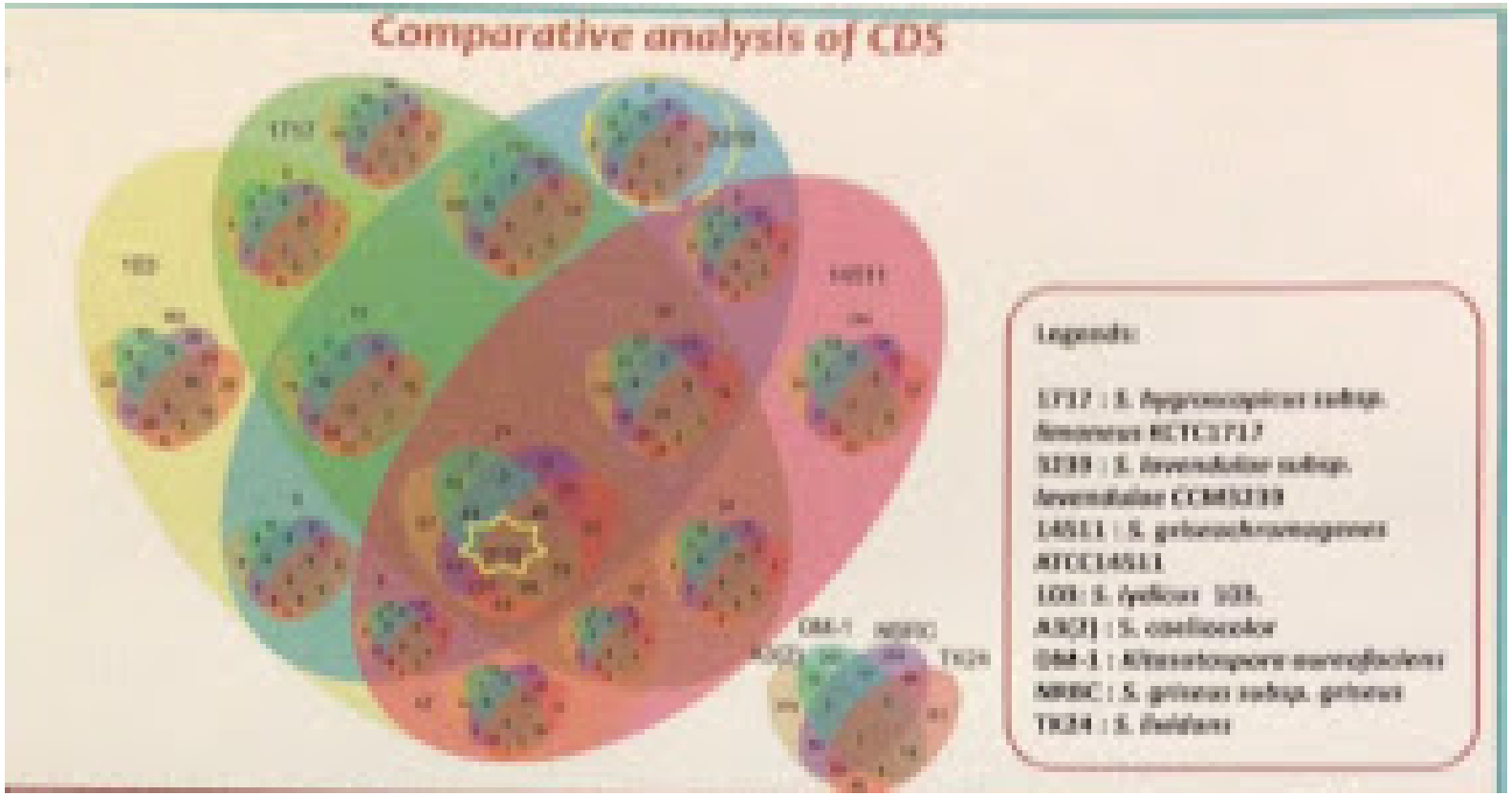
Complex and Complicated



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

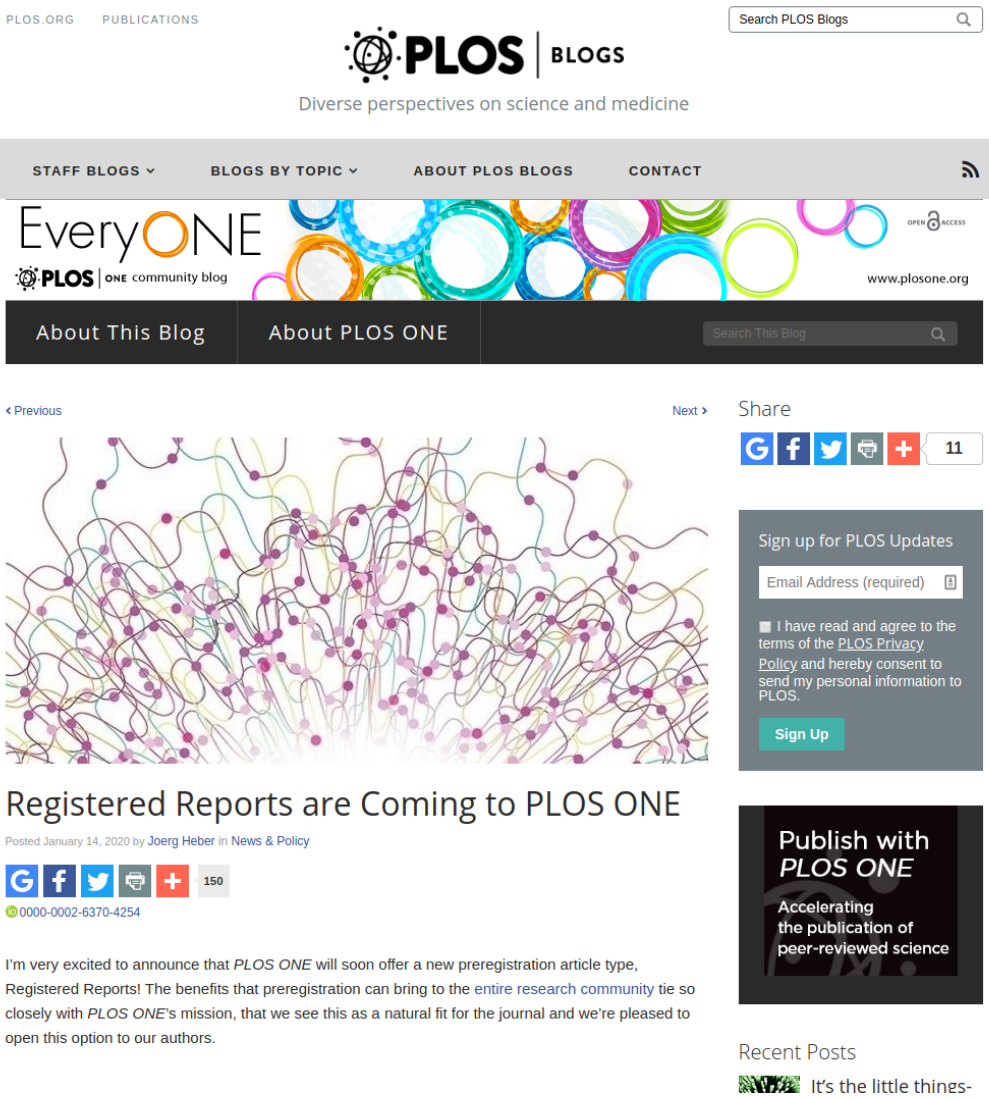
- metadata submission files can be used during the analysis;
- (should) makes you think ahead of time about biological question;

Complex and Complicated

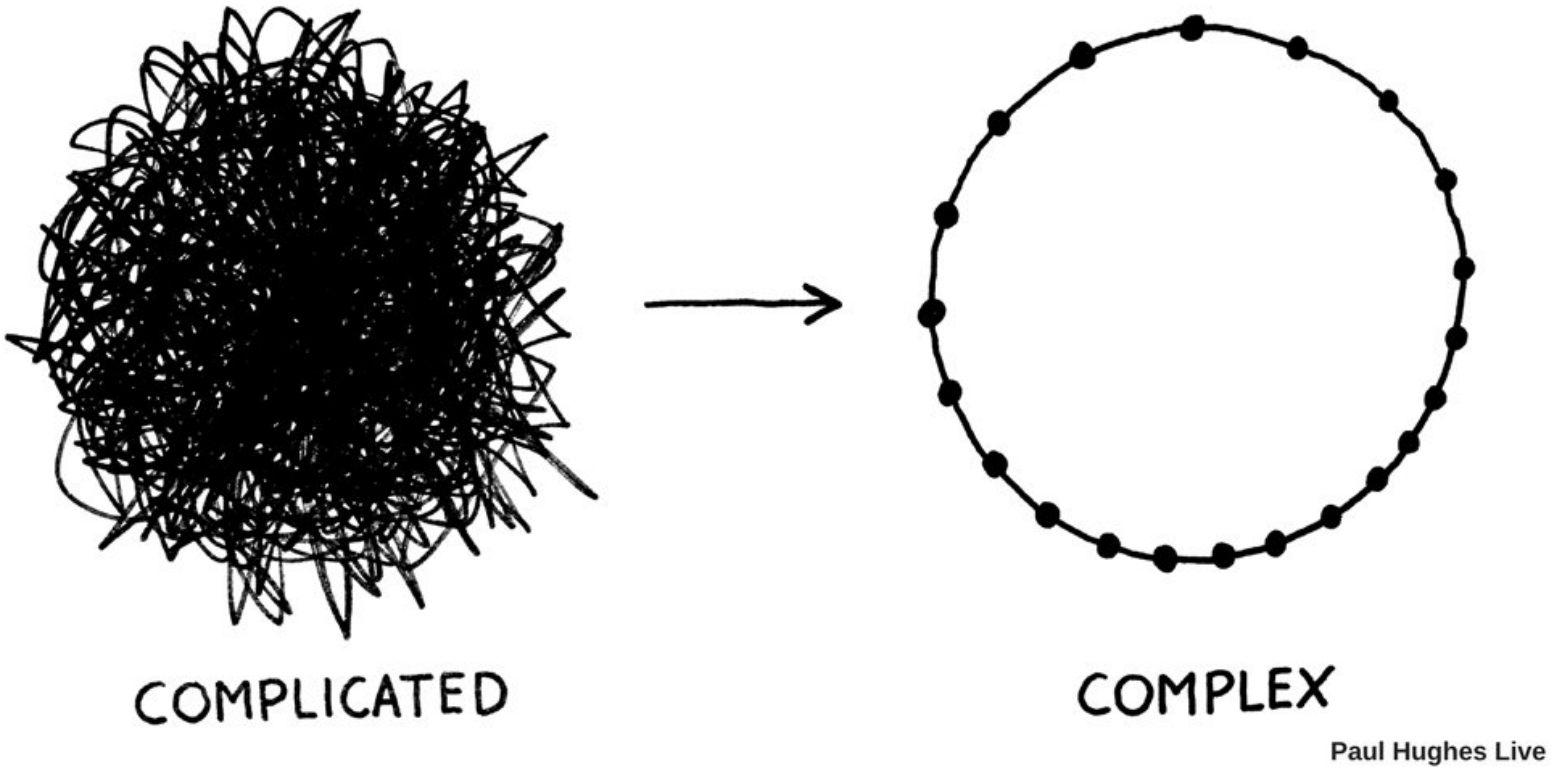
Advantages of early metadata submission/handling

- 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



Complex and Complicated



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

Complex and Complicated

python

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

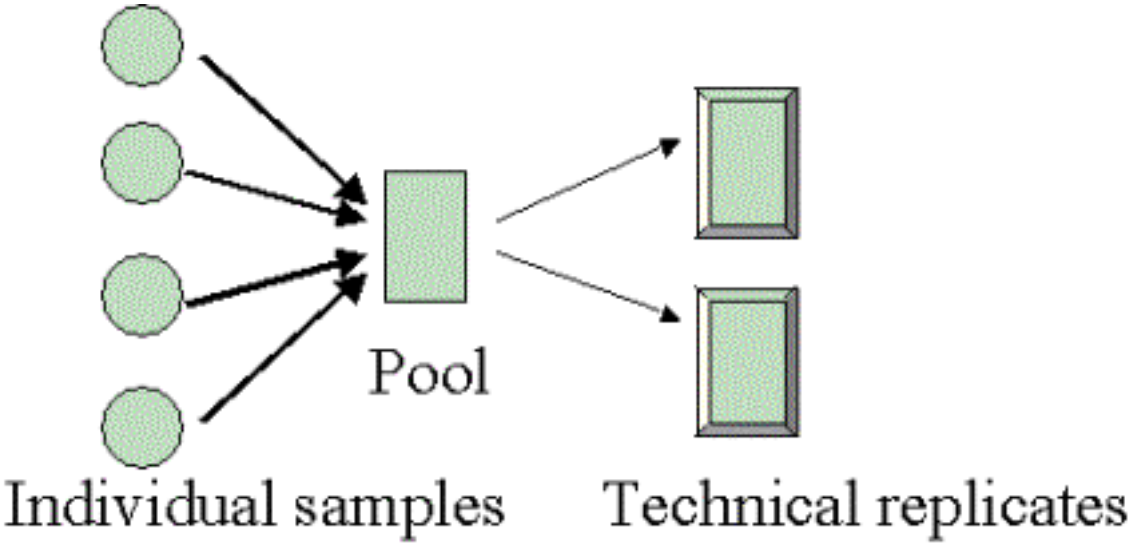
Complex and Complicated

python

- 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

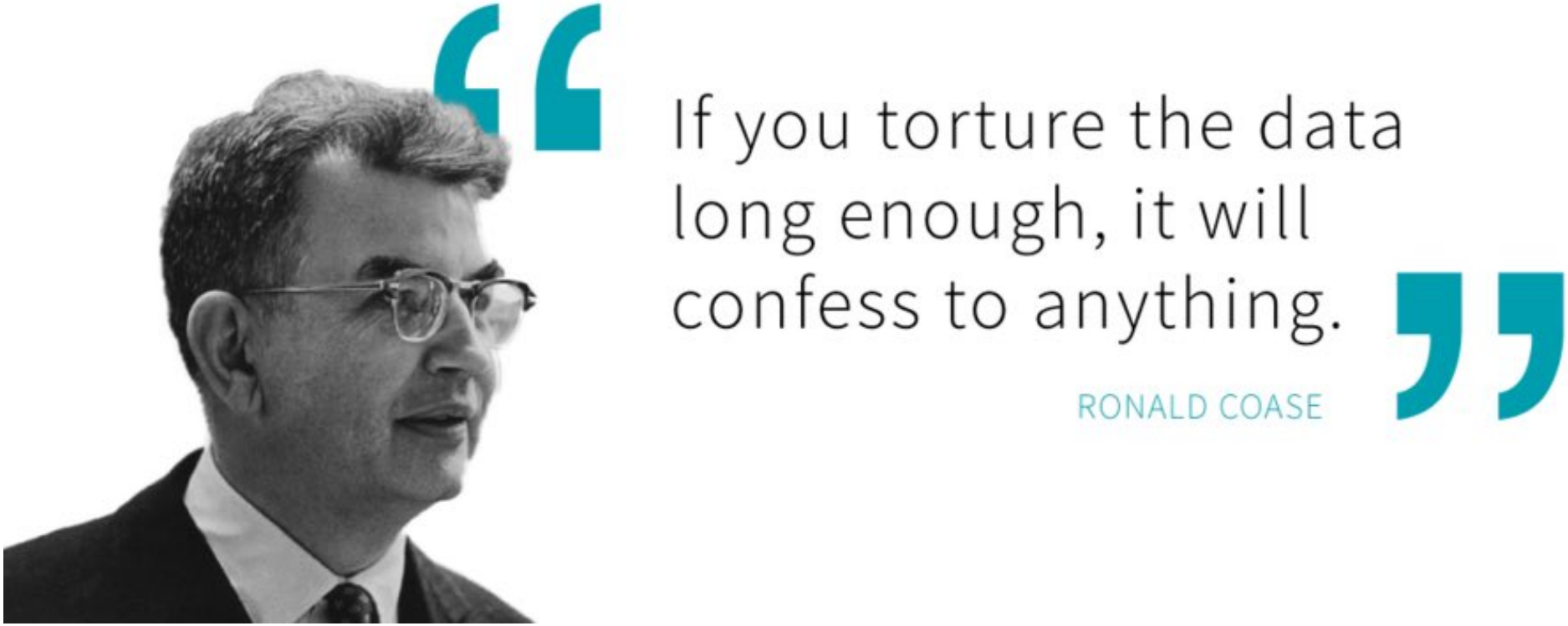
Biological vs Technical replicates



Complex and Complicated

Biological vs Technical replicates

Complex and Complicated



Complex and Complicated

Garbage In Garbage Out

Your analysis is as good as your data