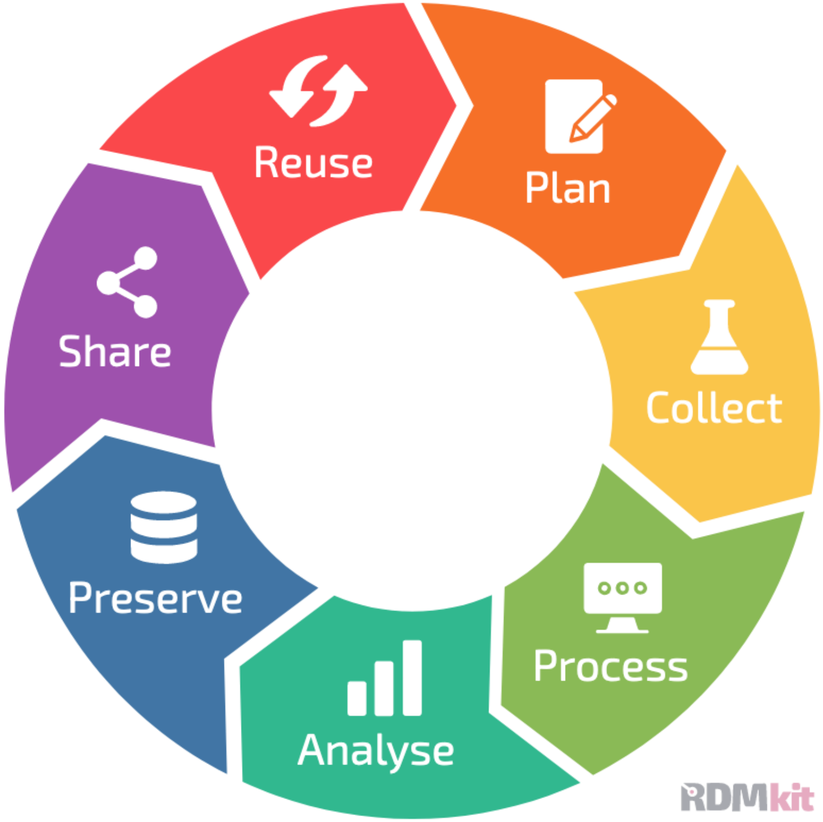
Metadata and submitting sequences

Geert van Geest
Interfaculty Bioinformatics Unit, UniBe
Training group, SIB

Question

Data life cycle

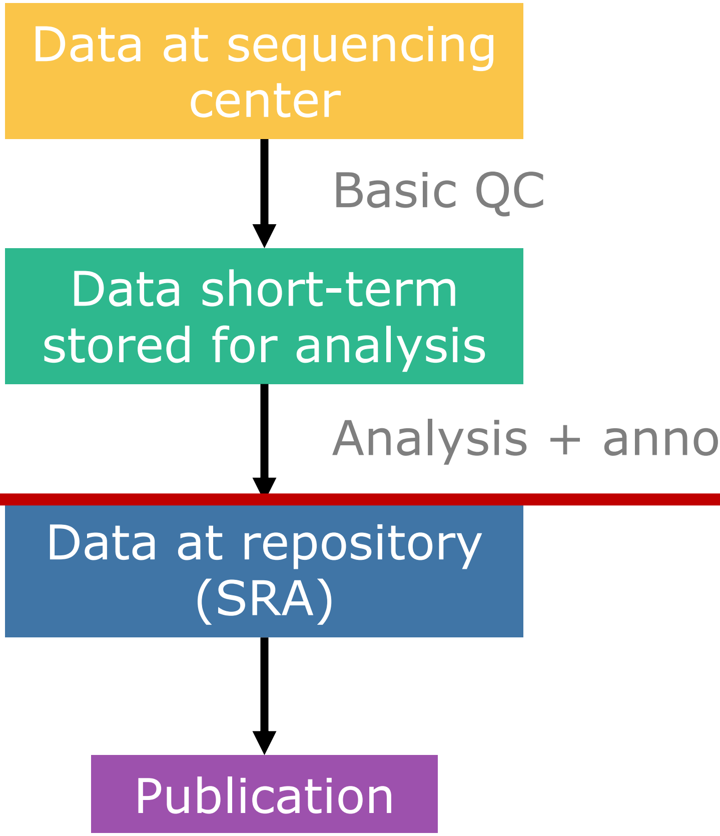


https://rdmkit.elixir-europe.org/

Sequencing data life

DANGER ZONE

- **No standardized metadata** - info gets lost
- **Limited backups** - technical failure leads to data loss



Data at sequencing center

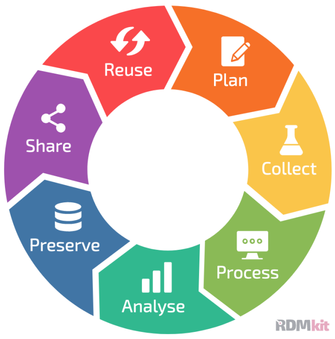
Basic QC

Data short-term stored for analysis

Analysis + annotation

Data at repository (SRA)

Publication



Get out of the danger zone

- Use your data management plan
- Plan data submission as part of the experiment
- Start the submission soon in the project

How?

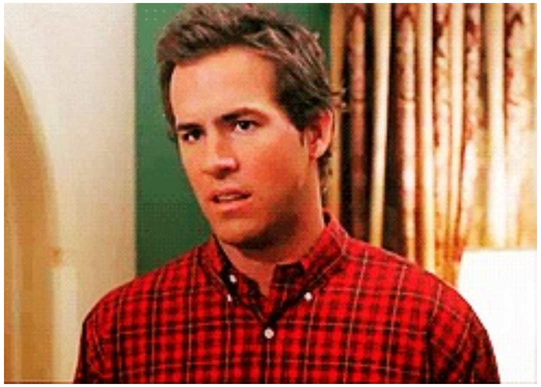
- Submit it to the sequence read archive of one of the INSDC databases:
    - Describe (metadata)
    - Upload

DDBJ
DNA Data Bank of Japan

NIH NLM
National Center for
Biotechnology Information

ENA
European Nucleotide Archive

Metadata

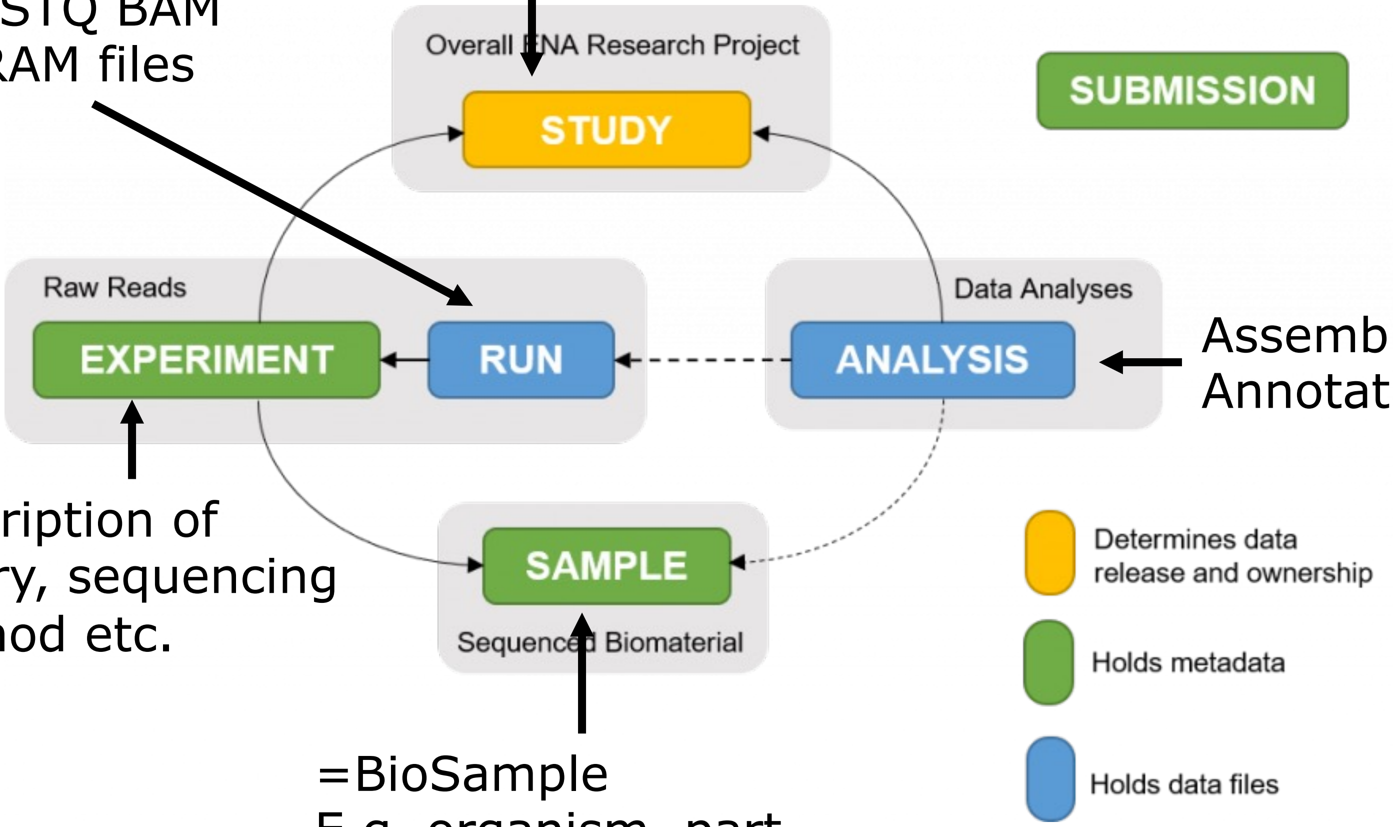


- Annoying:
  - **standardized** description of **unique** data, samples and experiments
  - Everything is already in the **paper**!
- But essential:
  - Enables finding
  - Enables re-use
- Data is useless without context!

ENA database

=BioProject
Overall description

FASTQ BAM
CRAM files



Assemblies
Annotations

Description of
library, sequencing
method etc.

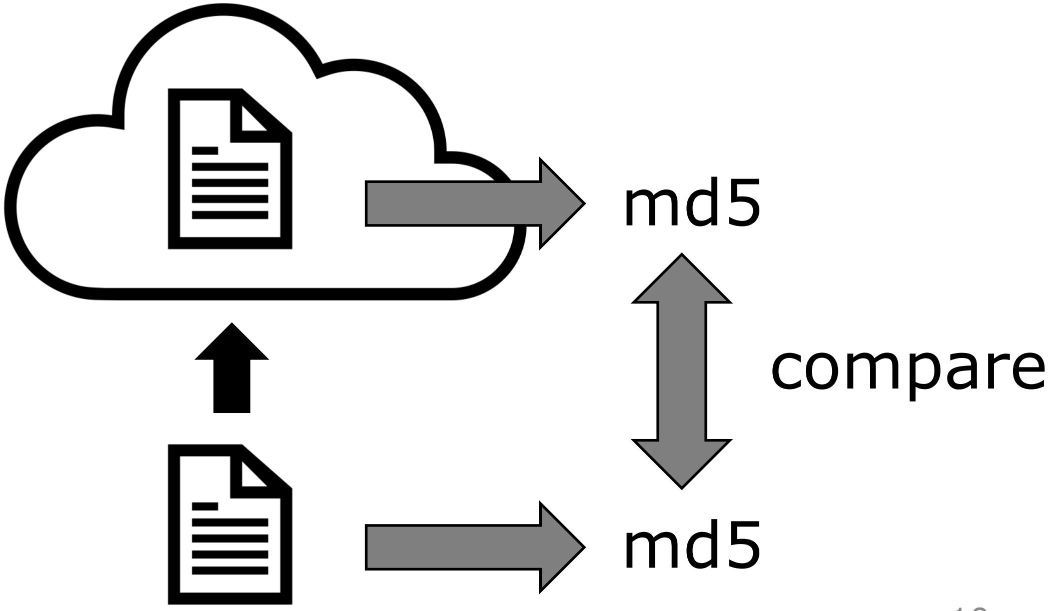
=BioSample
E.g. organism, part,
geographic location..

Transferring data

- Sequence data files are LARGE
- Possible protocols:
  - Webin File Uploader (GUI)
  - Webin CLI
  - File Transfer Protocol (FTP)
    - Command line
    - FileZilla
  - Aspera

md5 checksums

- Timeouts and failed transfers are common
- FASTQ files can become useless
- Solution: md5 checksum - changes if file changes



Exercises

- Create metadata entries on project, samples and experiment
- Upload sequence files to SRA

- We use the test environment during submission!

https://**wwwdev.**ebi.ac.uk/ena/browser/home