Metadata and submitting sequences

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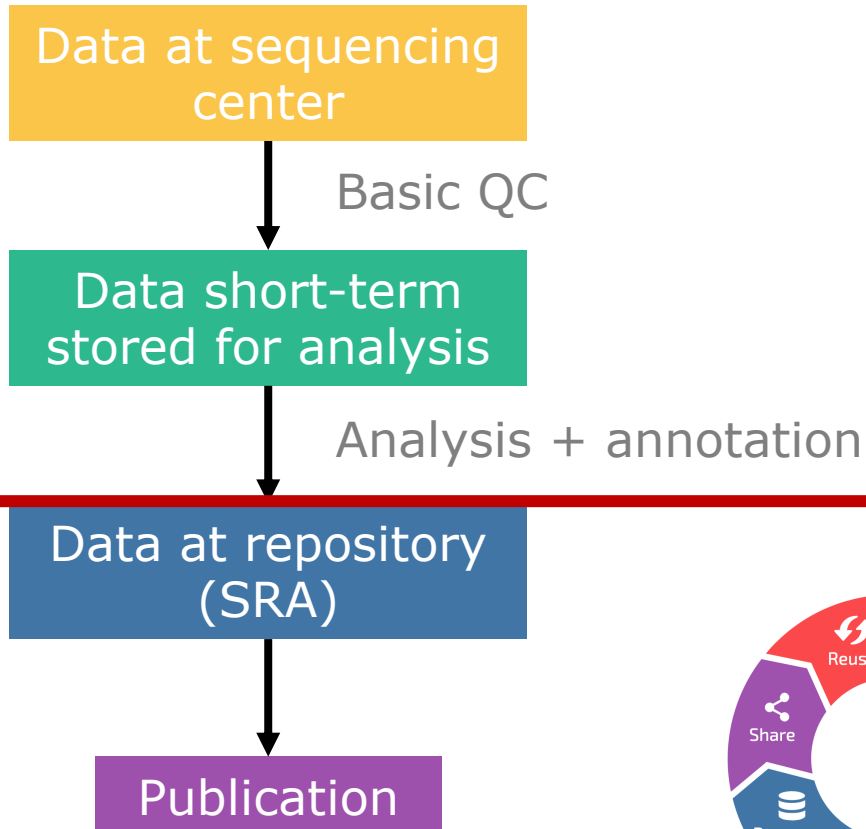
Question

Data life cycle

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

Sequencing data life

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

- <u>Use</u> your data management plan

- Plan data submission as part of the experiment

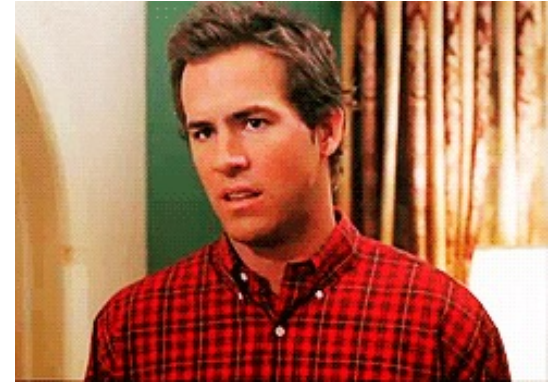- Start the submission soon in the project

How?

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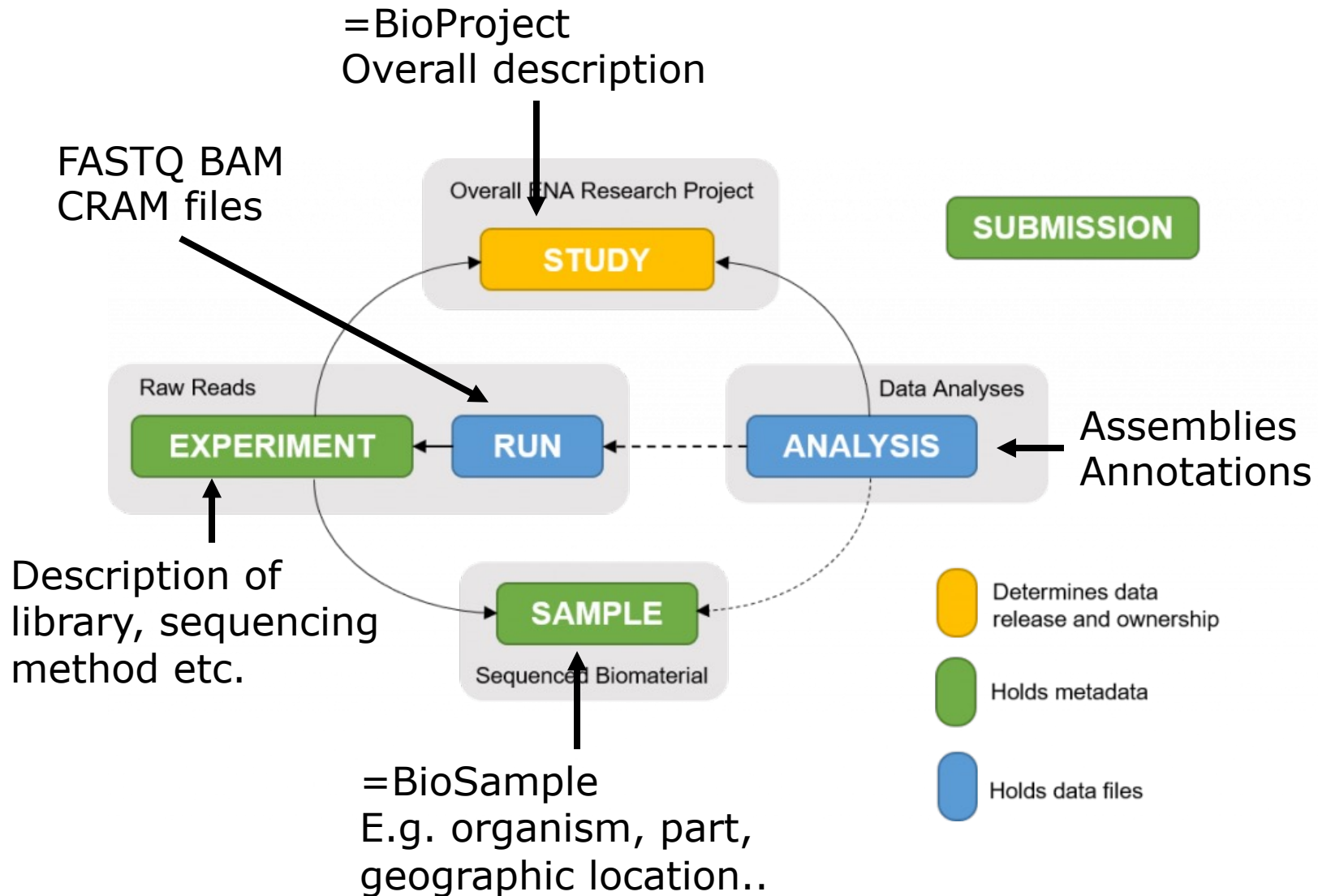 - Describe (metadata)
  - Upload

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

Description of
library, sequencing
method etc.

Assemblies
Annotations

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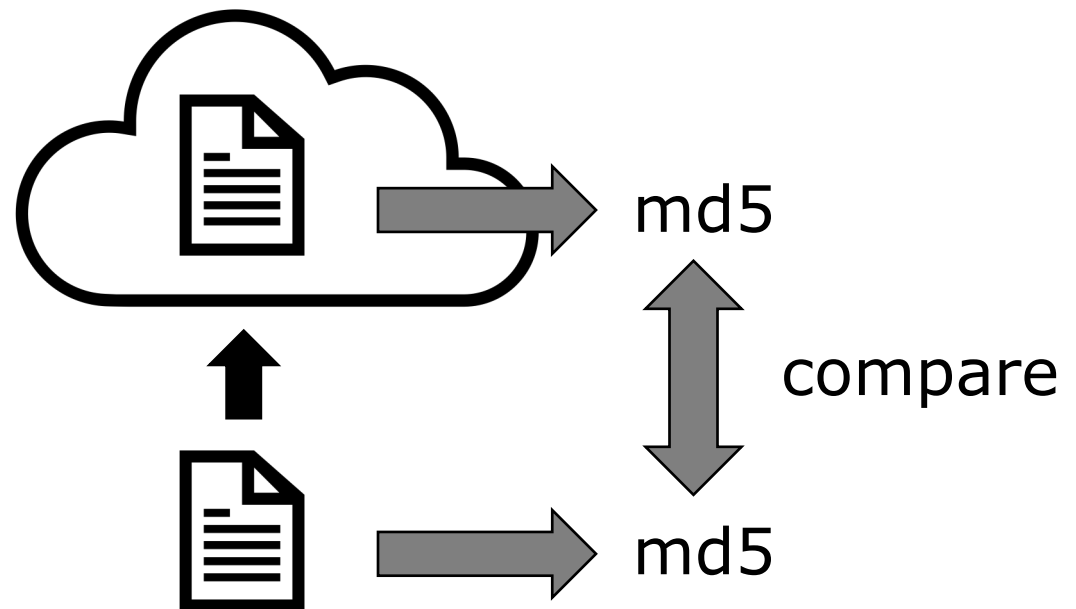geographic location..

8

Transferring data

- Sequence data files are LARGE
- Possible protocols:
  - Webin File Uploader (GUI)
  - 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

md5

compare

md5

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