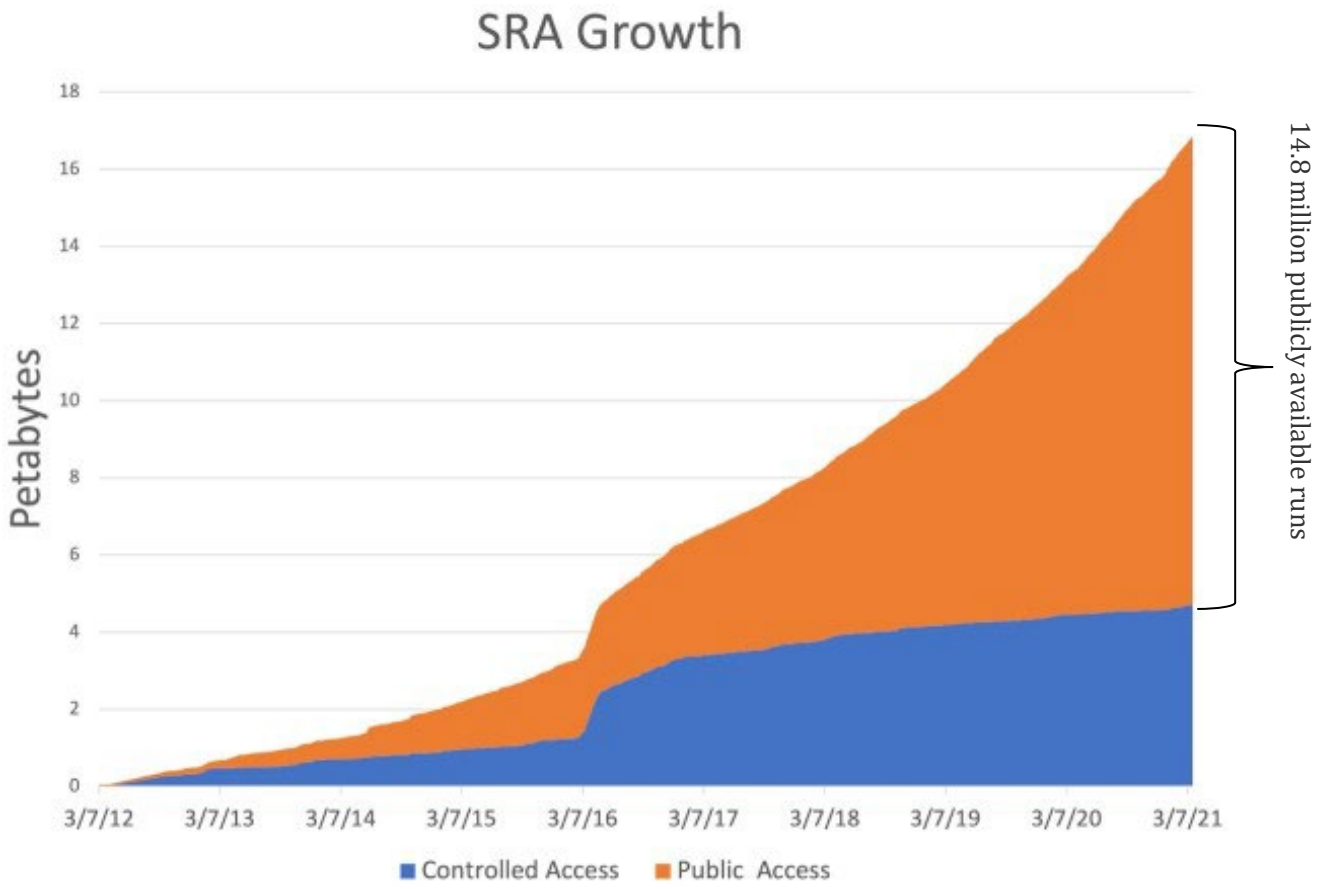
WGS repositories and data sharing

FWD AMR-RefLabCap
Workshop
12 March 2024
Faisal Khan, PhD
(fakh@food.dtu.dk)

Outline

- Whole Genome Sequencing (WGS) repositories
- Databases within GenBank
- Accessing WGS data
- Sharing WGS data

Tsunami of WGS data in last decade

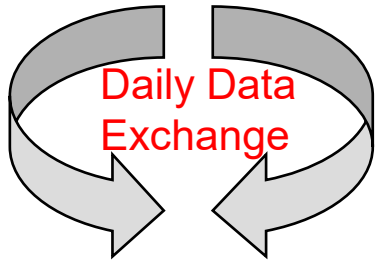


Bacterial Sequence repositories

- Almost all sequences are submitted to International Nucleotide Sequence Database Collaboration

NCBI

GenBank ® - NCBI genetic sequence database



DDBJ
DNA Data Bank of Japan

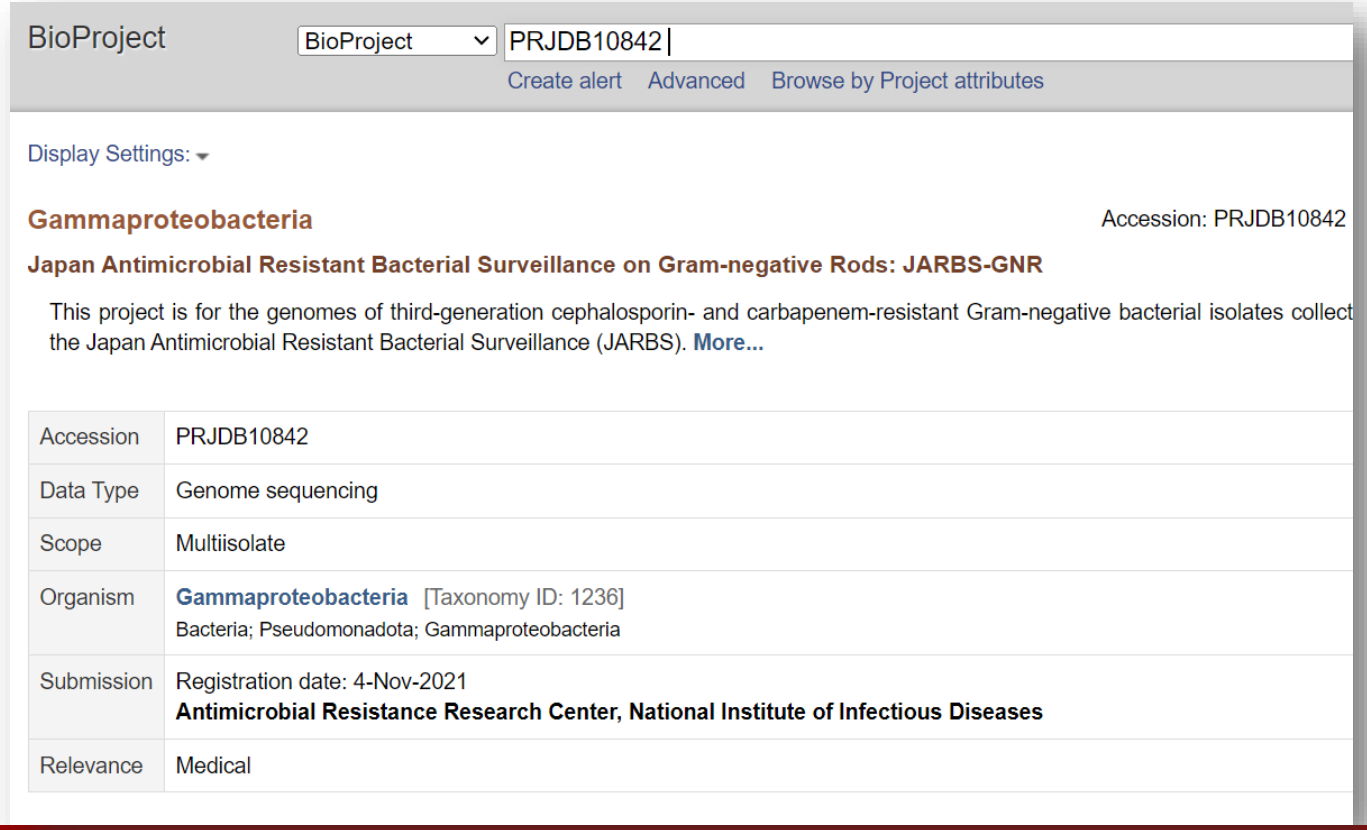
DDBJ- DNA Data Bank of Japan

ENA
European Nucleotide Archive

EBI-European Nucleotide Archive (ENA)

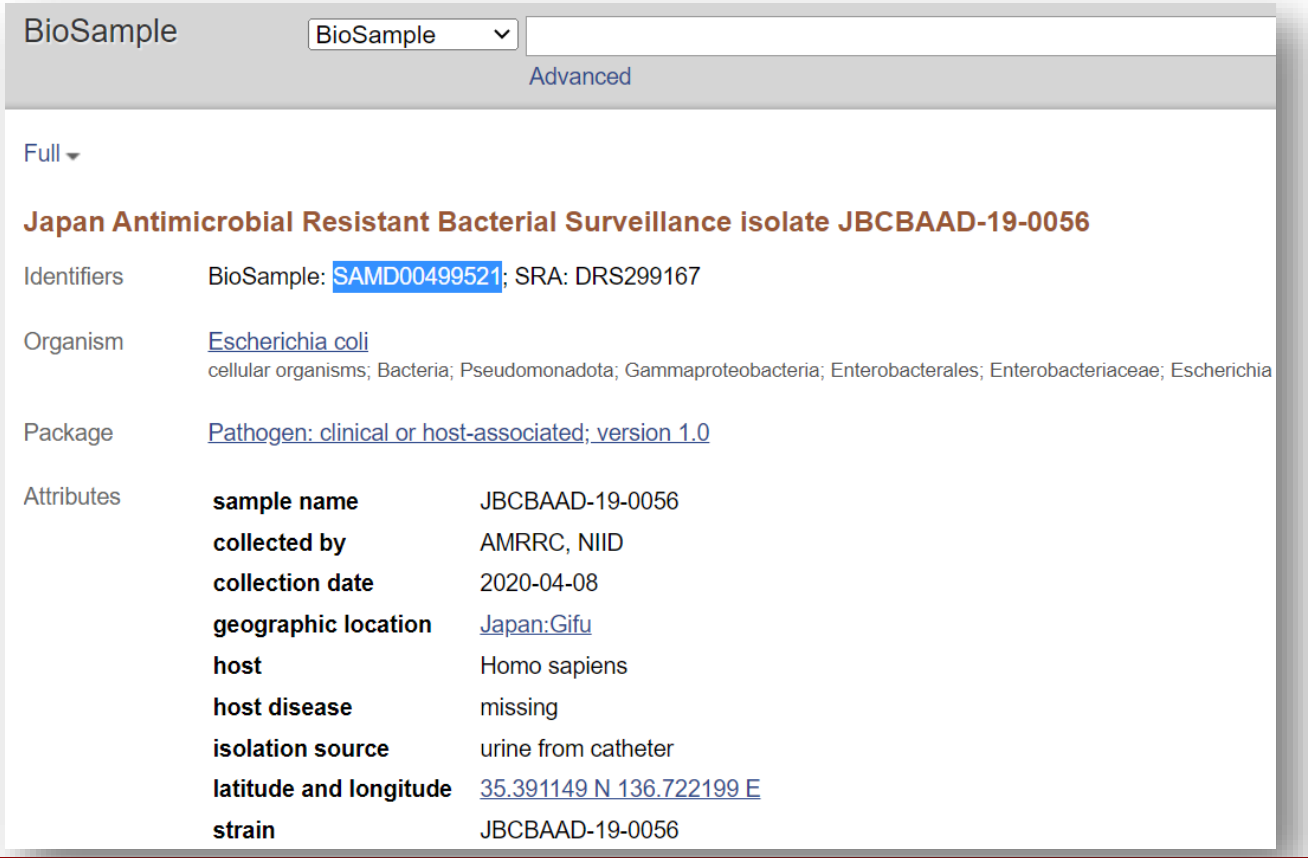
Databases within GenBank

- **BioProject-** collection of biological data related to a sequencing project
  - BioProject accession always starts with **PRJ**…. e.g., PRJNA271013
  - Contains description of the study/publication and number of samples



Databases within GenBank

- **BioSample-** contains description of biological source material (sample/isolate)
  - BioSample accession always starts with **SAM**…. e.g., SAMD00499521
  - Contains description of the isolate: strain name, host, country, collection date



Databases within GenBank

- **Sequencing Run Archive (SRA):** largest repository of raw sequencing data

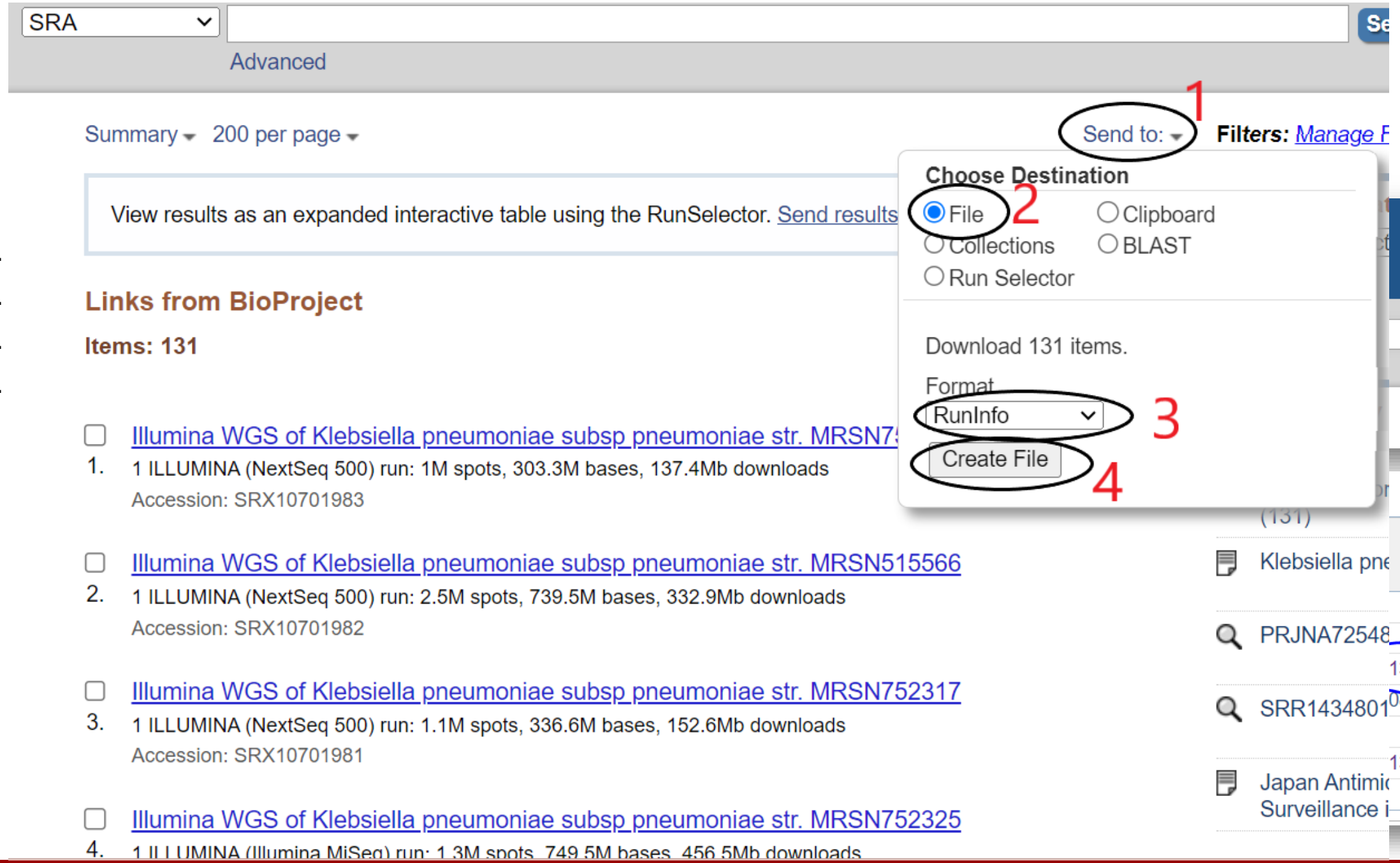
  – Contains raw reads
    - Samples (SRS)>Experiment (SRX)>Sequencing runs(SRR)

  – Accession numbers are given based on source database and type of record, e.g., **SRX**23903934

  – The first letter in the accession makes a notation of the source database - SRA, EBI, or DDBJ correspondingly

  – Third letter is given based on the type of data represented
    - Project/Study (e.g., the SRA record associated with a specific BioProject): SRP#, ERP#, or DRP#
    - Sample (e.g.,the SRA record associated with a specific BioSample): SRS#, ERS#, or DRS#
    - Experiment (e.g., the SRA record for a specific experiment or run(s)): SRX#, ERX#, or DRX#
    - Run (e.g., the SRA record for a specific run): SRR#, ERR#, or DRR#

We need run accession numbers to download the sequences


SRA
Advanced
Summary
200 per page
Send to:
1
Filters: Manage F
View results as an expanded interactive table using the RunSelector. Send results
Choose Destination
File
2
Clipboard
Collections
BLAST
Run Selector
Download 131 items.
Format
RunInfo
3
Create File
4
Links from BioProject
Items: 131
1. Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN75
1 ILLUMINA (NextSeq 500) run: 1M spots, 303.3M bases, 137.4Mb downloads
Accession: SRX10701983
2. Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN515566
1 ILLUMINA (NextSeq 500) run: 2.5M spots, 739.5M bases, 332.9Mb downloads
Accession: SRX10701982
3. Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN752317
1 ILLUMINA (NextSeq 500) run: 1.1M spots, 336.6M bases, 152.6Mb downloads
Accession: SRX10701981
4. Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN752325
1 ILLUMINA (Illumina MiSeq) run: 1.3M spots, 749.5M bases, 456.5Mb downloads
(131)
Klebsiella pne
PRJNA72548
9
131
SRR1434801
080
131
Japan Antimi
Surveillance i
2

1.
2.
3.
4.

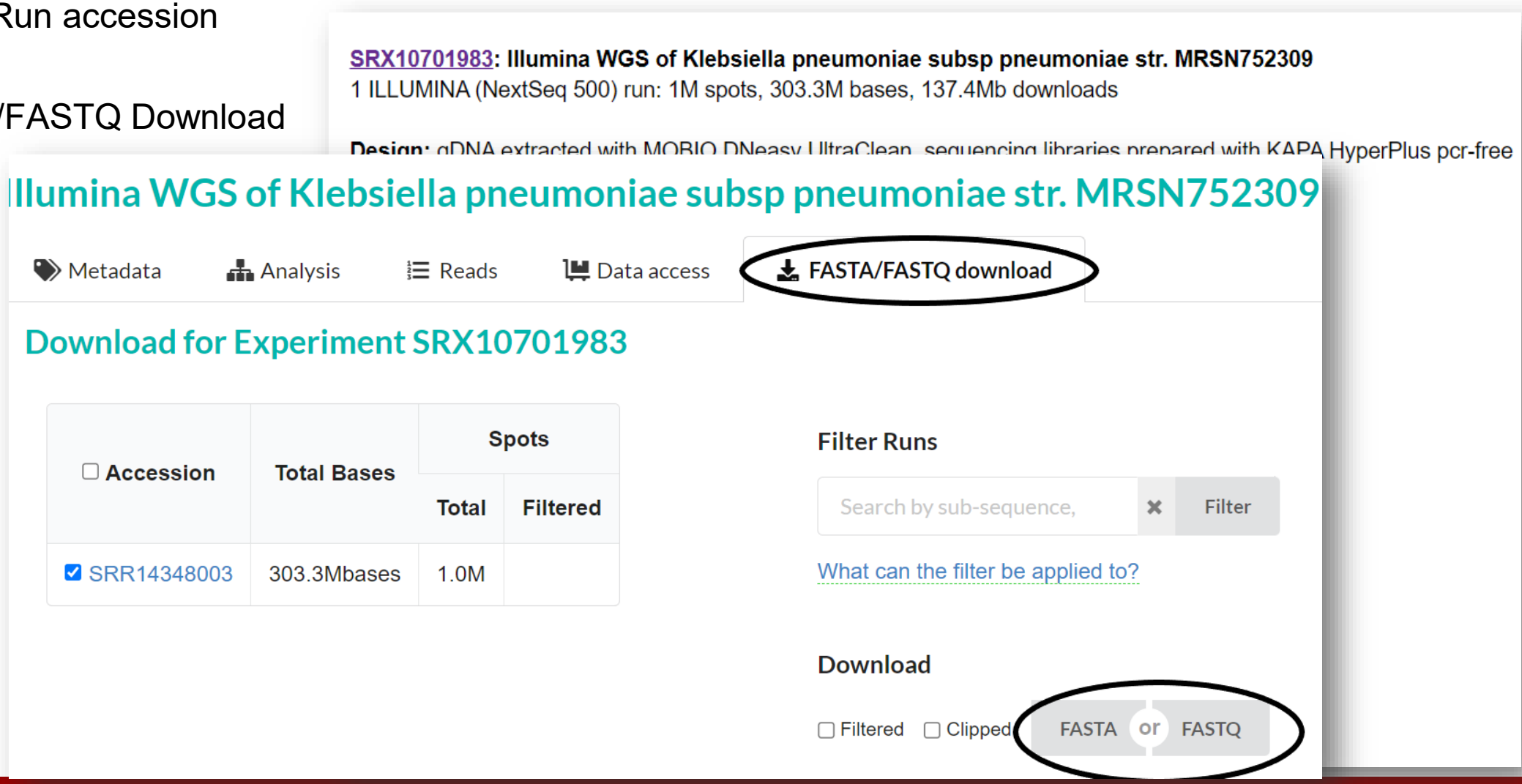
Ways to retrieve fastq files from SRA

- NCBI provides a tool to extract and download sequnces as fastq files
  - SRA-Toolkit (*fasterq-dump)*
  - Commandline-tool that is available for Linux, macOS, and Windows.
  - Can download multiple runs with one command

- Access fastq files via web-based servers
  - NCBI-GenBank
  - ENA Browser
  - Using *fasterq-dump* in Galaxy

Downloading genomic data from SRA database: Using BioProject

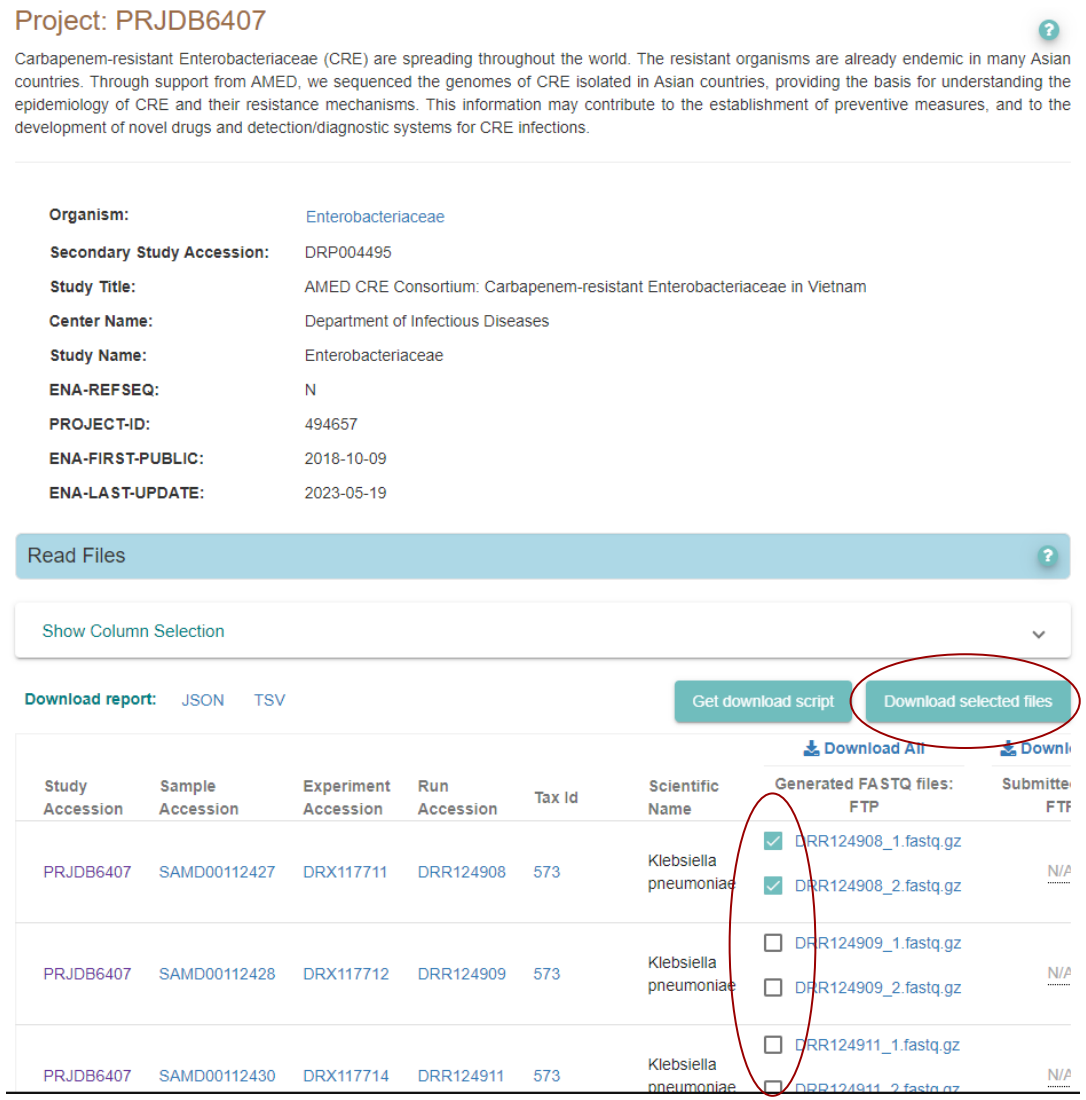
- Click on the Run accession
- Click FASTA/FASTQ Download



Downloading from ENA browser

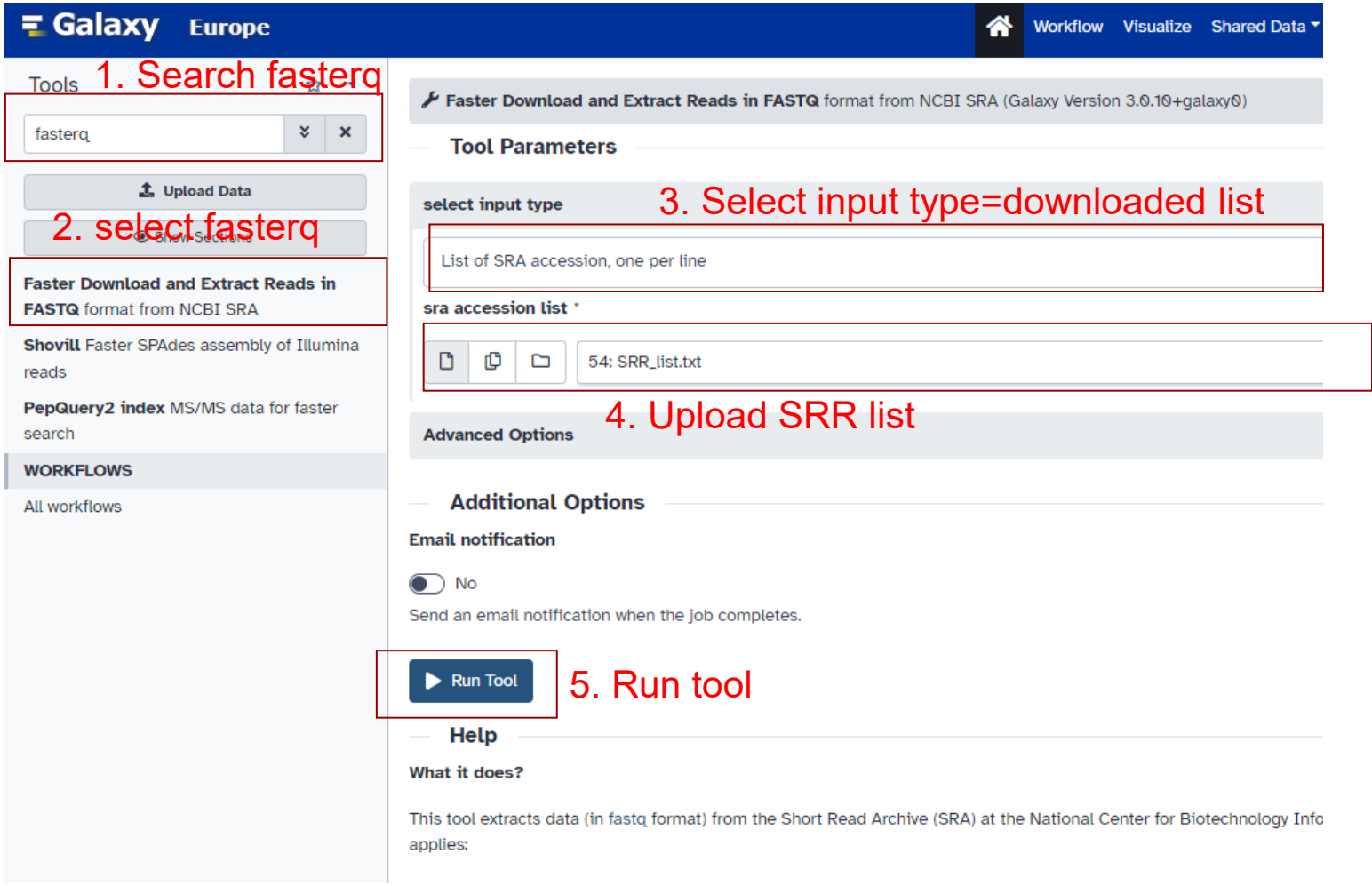
https://www.ebi.ac.uk/ena/browser/search

1. Search for the PRJ, SAM, or SRA accession number
2. Select the fastq/fasta files
3. Download



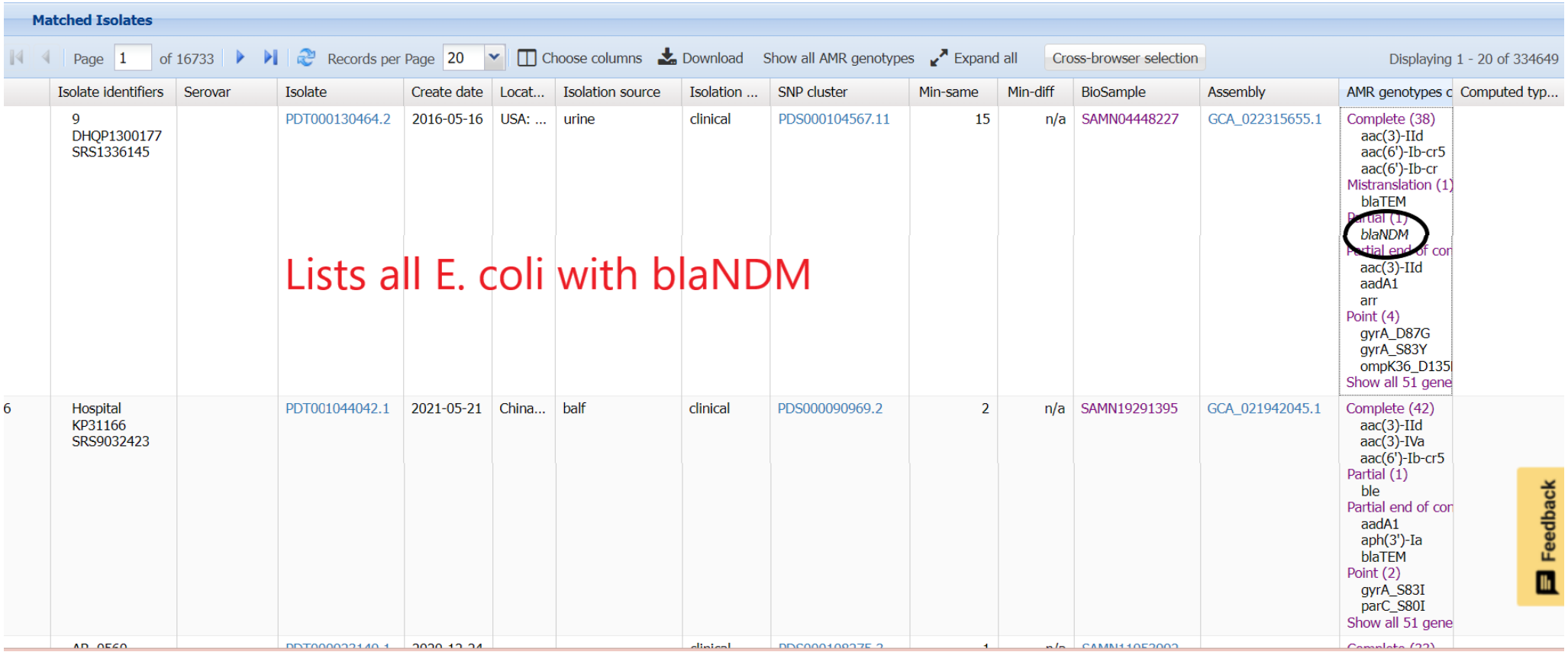
Another alternative- Galaxy

- https://usegalaxy.eu/
- 250GB disk space
- Hundreds of bioinformatics tools available



Searching for specific AMR Pathogens

- NCBI Pathogen Detection (https://www.ncbi.nlm.nih.gov/pathogens/)



	Isolate identifiers	Serovar	Isolate	Create date	Locat...	Isolation source	Isolation ...	SNP cluster	Min-same	Min-diff	BioSample	Assembly	AMR genotypes c	Computed typ...
	9 DHQP1300177 SRS1336145		PDT000130464.2	2016-05-16	USA: ...	urine	clinical	PDS000104567.11	15	n/a	SAMN04448227	GCA_022315655.1	Complete (38) aac(3)-IId aac(6')-Ib-cr5 aac(6')-Ib-cr Mistranslation (1) blaTEM Partial (1) *blaNDM* Partial end of cor aac(3)-IId aadA1 arr Point (4) gyrA_D87G gyrA_S83Y ompK36_D135 Show all 51 gene	
6	Hospital KP31166 SRS9032423		PDT001044042.1	2021-05-21	China...	balf	clinical	PDS000090969.2	2	n/a	SAMN19291395	GCA_021942045.1	Complete (42) aac(3)-IId aac(3)-IVa aac(6')-Ib-cr5 Partial (1) ble Partial end of cor aadA1 aph(3')-Ia blaTEM Point (2) gyrA_S83I parC_S80I Show all 51 gene	

Quality Control of sequence data

- Always check the quality of publicly available sequence data

- For fastq files (raw reads)
  - Check quality of the reads
  - e.g., with FastQC

- For fasta files (assembled reads)
  - Check the quality of assembly
  - e.g., with QUAST

Sharing large sequence data- ScienceData.dk

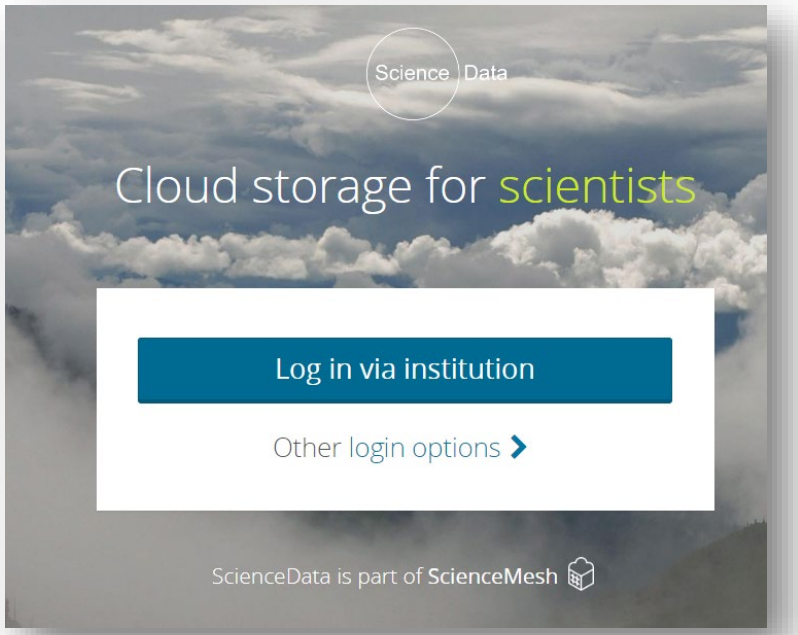
- WGS data can be hundreds of GBs in size
  - Cannot be shared via email
  - Need to be hosted on a cloud storage platform

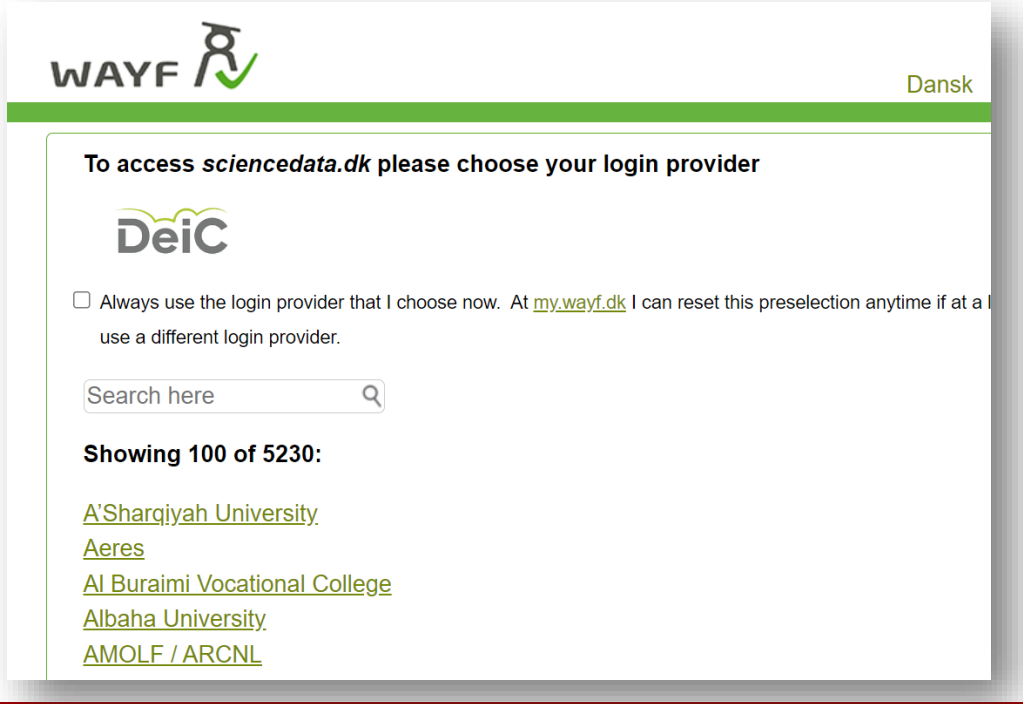
- Provided by the Technical University of Denmark (DTU)
  - Mainly for Danish research institutes
  - Approx. 5200 institutes around the world can access ScienceData (eduGAIN member?)
  - 200GBs of free storage
  - Share large files with others via weblink

Sharing large sequence data- ScienceData.dk

- https://sciencedata.dk/

- Log-in via institution
- Search your institution or country





Sharing large sequence data- EUDAT

- European Collaborative Data Infrastructure (EUDAT)

- B2DROP, the EUDAT's Personal Cloud Storage Service (https://b2drop.eudat.eu)
  - Access through institution?
  - Free 20GBs storage
  - Possibility to share data via weblink

Thank you!

Questions?