

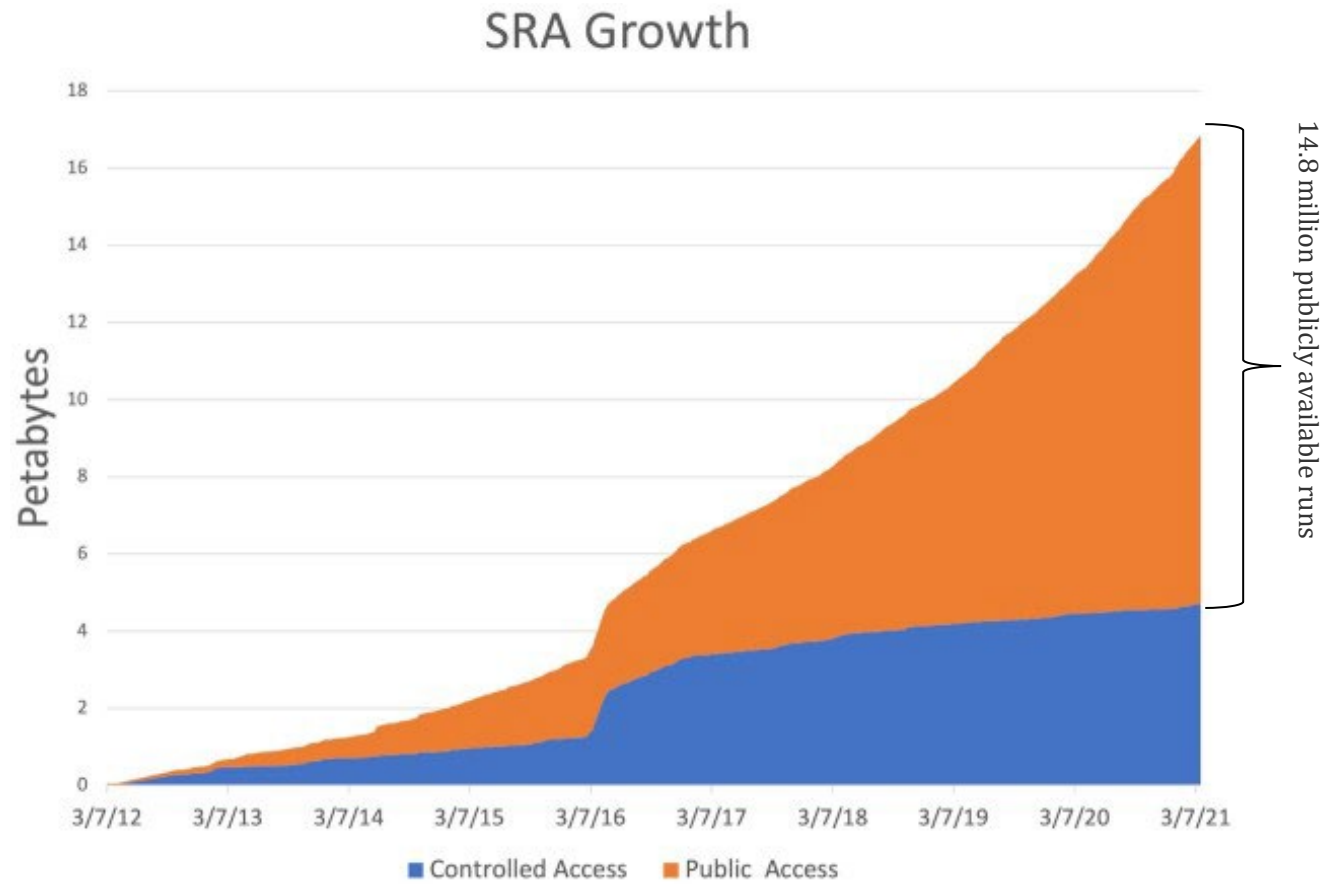
# WGS repositories and data sharing

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

# Outline

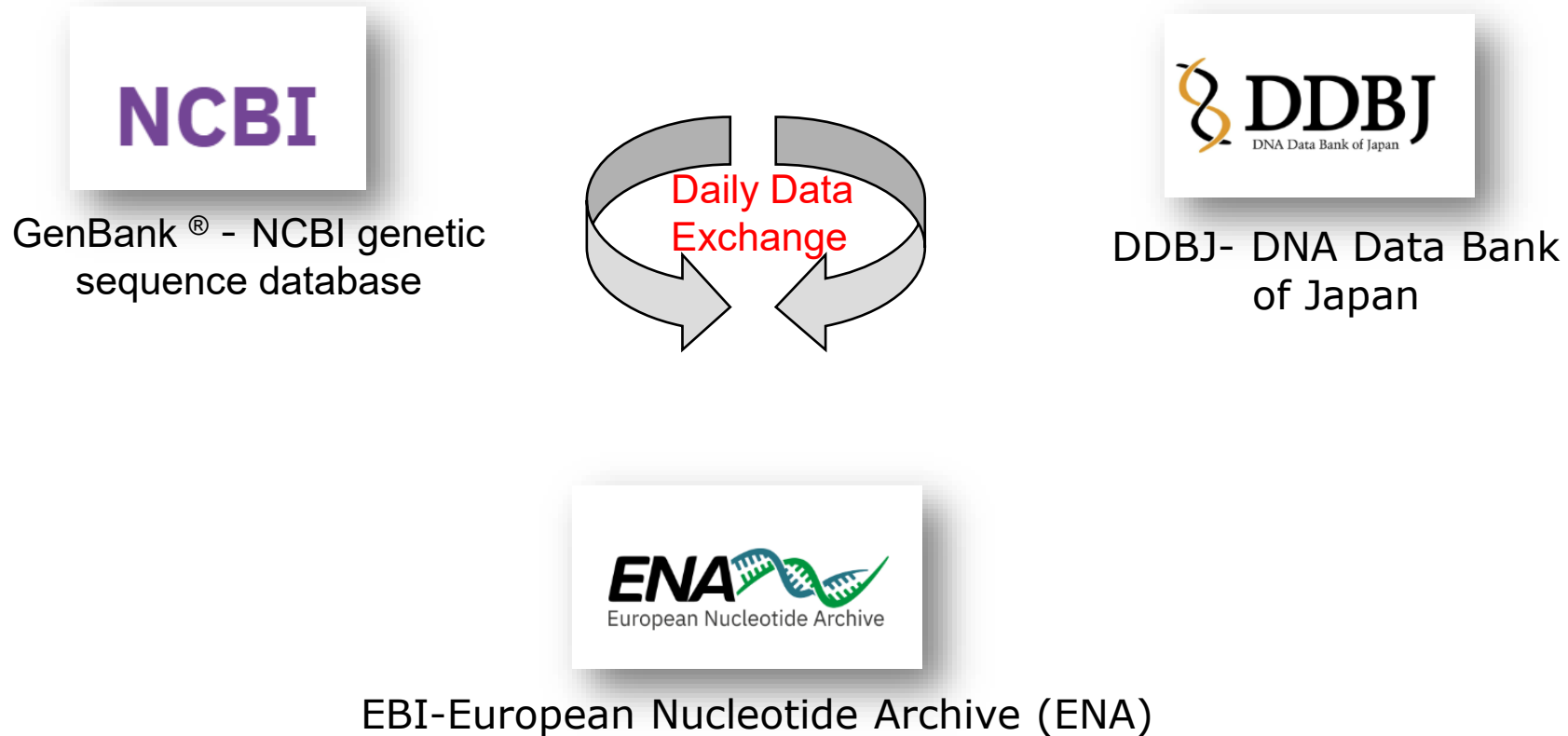
- Whole Genome Sequencing (WGS) repositories
- Databases within GenBank
- Accessing WGS data
- Sharing WGS data

# Tsunami of WGS data in last decade



# Bacterial Sequence repositories

- Almost all sequences are submitted to [International Nucleotide Sequence Database Collaboration](#)



# 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

BioProject   [Create alert](#) [Advanced](#) [Browse by Project attributes](#)

Display Settings: ▾

**Gammaproteobacteria** Accession: PRJDB10842

**Japan Antimicrobial Resistant Bacterial Surveillance on Gram-negative Rods: JARBS-GNR**

This project is for the genomes of third-generation cephalosporin- and carbapenem-resistant Gram-negative bacterial isolates collected from the Japan Antimicrobial Resistant Bacterial Surveillance (JARBS). [More...](#)

Accession	PRJDB10842
Data Type	Genome sequencing
Scope	Multiisolate
Organism	<b>Gammaproteobacteria</b> [Taxonomy ID: 1236] Bacteria; Pseudomonadota; Gammaproteobacteria
Submission	Registration date: 4-Nov-2021 <b>Antimicrobial Resistance Research Center, National Institute of Infectious Diseases</b>
Relevance	Medical

# 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

BioSample  Advanced

---

Full ▾

**Japan Antimicrobial Resistant Bacterial Surveillance isolate JBCBAAD-19-0056**

Identifiers      BioSample: [SAMD00499521](#); SRA: DRS299167

Organism          [Escherichia coli](#)  
cellular organisms; Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia

Package            [Pathogen: clinical or host-associated; version 1.0](#)

Attributes

<b>sample name</b>	JBCBAAD-19-0056
<b>collected by</b>	AMRRC, NIID
<b>collection date</b>	2020-04-08
<b>geographic location</b>	<a href="#">Japan:Gifu</a>
<b>host</b>	Homo sapiens
<b>host disease</b>	missing
<b>isolation source</b>	urine from catheter
<b>latitude and longitude</b>	<a href="#">35.391149 N 136.722199 E</a>
<b>strain</b>	JBCBAAD-19-0056

# 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., **SRX23903934**
    - 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

Summary ▾ 200 per page ▾

View results as an expanded interactive table using the RunSelector. [Send results](#)

Send to: ▾

Filters: [Manage F](#)

Choose Destination

- File
- Clipboard
- Collections
- BLAST
- Run Selector

Download 131 items.

Format

RunInfo ▾

Create File

- 1.
- 2.
- 3.
- 4.

### Links from BioProject

Items: 131

[Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN79](#)

1. 1 ILLUMINA (NextSeq 500) run: 1M spots, 303.3M bases, 137.4Mb downloads  
Accession: SRX10701983

[Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN515566](#)

2. 1 ILLUMINA (NextSeq 500) run: 2.5M spots, 739.5M bases, 332.9Mb downloads  
Accession: SRX10701982

[Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN752317](#)

3. 1 ILLUMINA (NextSeq 500) run: 1.1M spots, 336.6M bases, 152.6Mb downloads  
Accession: SRX10701981

[Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN752325](#)

4. 1 ILLUMINA (Illumina MiSeq) run: 1.3M spots, 749.5M bases, 456.5Mb downloads

(131)

	Klebsiella pne	
	PRJNA72548	9
		131
	SRR1434801	080
		131
	Japan Antimic	
	Surveillance i	2



# Ways to retrieve fastq files from SRA

- NCBI provides a tool to extract and download sequences 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

**SRX10701983: Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN752309**  
 1 ILLUMINA (NextSeq 500) run: 1M spots, 303.3M bases, 137.4Mb downloads

**Design:** cDNA extracted with MORIO DNeasy UltraClean sequencing libraries prepared with KAPA HyperPlus pcr-free

## Illumina WGS of Klebsiella pneumoniae subsp pneumoniae str. MRSN752309

[Metadata](#)
[Analysis](#)
[Reads](#)
[Data access](#)
[FASTA/FASTQ download](#)

### Download for Experiment SRX10701983

<input type="checkbox"/> Accession	Total Bases	Spots	
		Total	Filtered
<input checked="" type="checkbox"/> SRR14348003	303.3Mbases	1.0M	

**Filter Runs**

Search by sub-sequence,

[What can the filter be applied to?](#)

**Download**

Filtered
  Clipped
  FASTA
  or
  FASTQ

# 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/fastq files
3. Download

Project: PRJDB6407

Carbapenem-resistant Enterobacteriaceae (CRE) are spreading throughout the world. The resistant organisms are already endemic in many Asian countries. Through support from AMED, we sequenced the genomes of CRE isolated in Asian countries, providing the basis for understanding the epidemiology of CRE and their resistance mechanisms. This information may contribute to the establishment of preventive measures, and to the development of novel drugs and detection/diagnostic systems for CRE infections.

**Organism:** [Enterobacteriaceae](#)  
**Secondary Study Accession:** DRP004495  
**Study Title:** AMED CRE Consortium: Carbapenem-resistant Enterobacteriaceae in Vietnam  
**Center Name:** Department of Infectious Diseases  
**Study Name:** Enterobacteriaceae  
**ENA-REFSEQ:** N  
**PROJECT-ID:** 494657  
**ENA-FIRST-PUBLIC:** 2018-10-09  
**ENA-LAST-UPDATE:** 2023-05-19

Read Files

Show Column Selection

Download report: [JSON](#) [TSV](#)

Get download script

Download selected files

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Generated FASTQ files:		Submitter
						FTP	FTP	
PRJDB6407	SAMD00112427	DRX117711	DRR124908	573	Klebsiella pneumoniae	<input checked="" type="checkbox"/>	DRR124908_1.fastq.gz	N/A
						<input checked="" type="checkbox"/>	DRR124908_2.fastq.gz	
PRJDB6407	SAMD00112428	DRX117712	DRR124909	573	Klebsiella pneumoniae	<input type="checkbox"/>	DRR124909_1.fastq.gz	N/A
						<input type="checkbox"/>	DRR124909_2.fastq.gz	
PRJDB6407	SAMD00112430	DRX117714	DRR124911	573	Klebsiella pneumoniae	<input type="checkbox"/>	DRR124911_1.fastq.gz	N/A
						<input type="checkbox"/>	DRR124911_2.fastq.gz	

# Another alternative- Galaxy

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

The screenshot shows the Galaxy Europe interface with the following steps highlighted:

- 1. Search fasterq**: The search bar in the Tools panel contains the text 'fasterq'.
- 2. select fasterq**: The tool 'Faster Download and Extract Reads in FASTQ format from NCBI SRA' is selected in the Tools panel.
- 3. Select input type=downloaded list**: The 'select input type' dropdown is set to 'List of SRA accession, one per line'.
- 4. Upload SRR list**: The 'sra accession list' field contains the file path '54: SRR\_list.txt'.
- 5. Run tool**: The 'Run Tool' button is highlighted.

# Searching for specific AMR Pathogens

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

Matched Isolates													
Page 1 of 16733   Records per Page 20   Choose columns   Download   Show all AMR genotypes   Expand all   Cross-browser selection   Displaying 1 - 20 of 334649													
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) <b>blaNDM</b> 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	
AD_0560		PDT000023140.1	2020-12-24			clinical	PDS000108275.2	1	n/a	SAMN11052002		Complete (22)	

Lists all E. coli with blaNDM



# Quality Control of sequence data

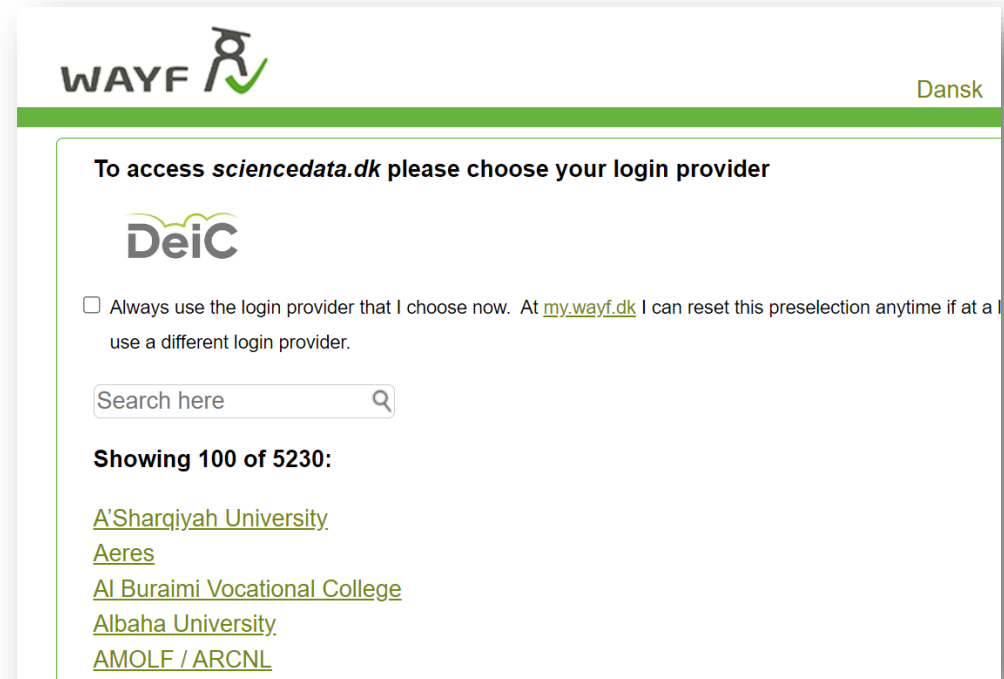
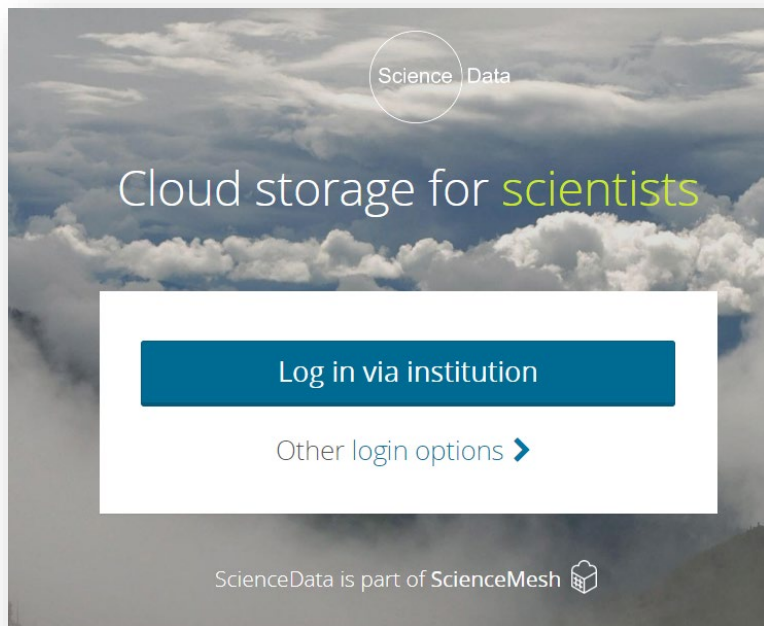
- Always check the quality of publicly available sequence data
- For fastq files (raw reads)
  - Check quality of the reads
  - e.g., with FastQC
- For fasta files (assembled reads)
  - Check the quality of assembly
  - e.g., with QUAST

# Sharing large sequence data- [ScienceData.dk](https://www.science-data.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)

- <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?**