Mobilisation of SARS-CoV-2 data

Data brokering for Swedish SARS-CoV-2 data submissions to ENA

2022-05-19

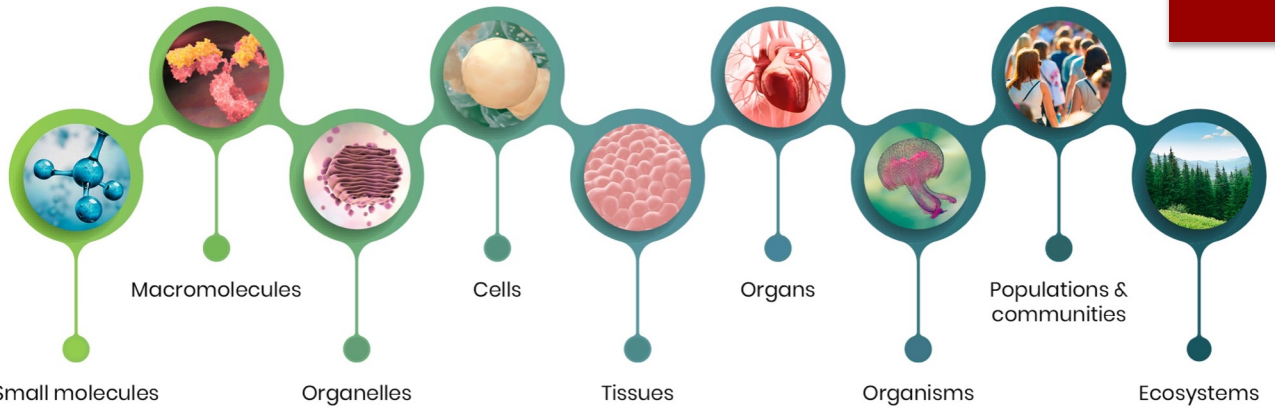
Parul Tewatia

Data Centre ScilifeLab

parul.tewatia@scilifelab.se

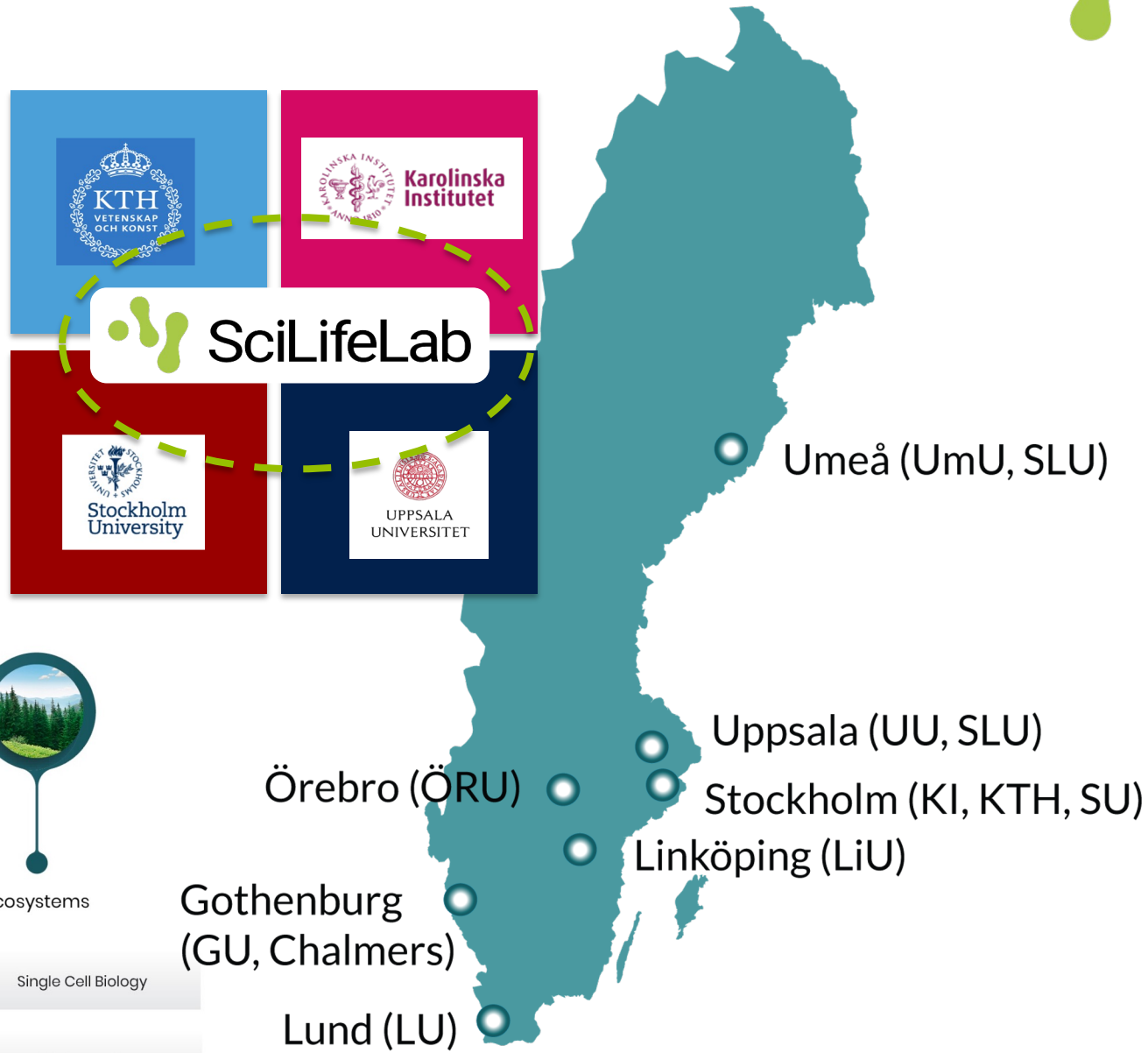
SciLifeLab

- An institution for **advancing molecular biosciences** and a **research infrastructure**
- Activities at all major Swedish universities

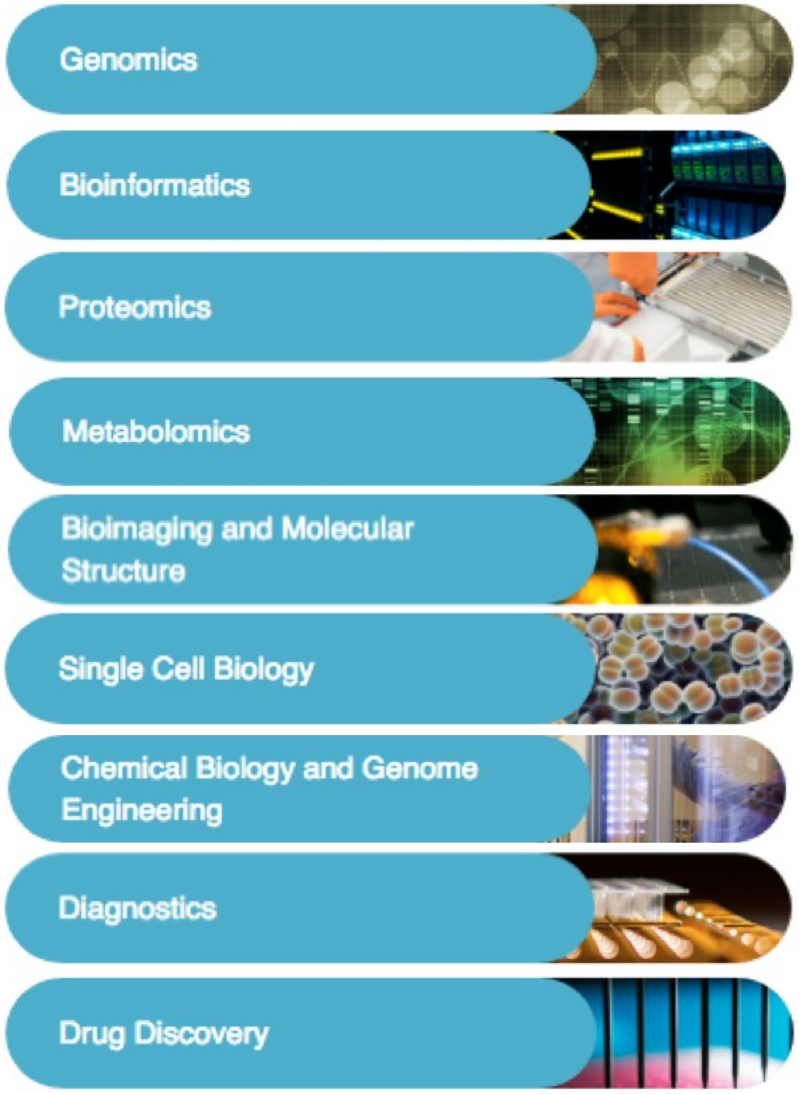


Genomics | Spatial Omics | Proteomics | Metabolomics | Single Cell Biology

Bioimaging and Molecular Structure | Chemical Biology and Genome Engineering | Drug Discovery | Diagnostics | Bioinformatics



Science for Life Laboratory: National infrastructure with 10 platforms, 40 facilities



Bioinformatics

- Support, Infrastructure and Training (G, Li, Lu, S, U, Um)
- Compute and Storage (U)
- BioImage Informatics (U, S)
- AIDA Data Hub (Li)

Genomics

- National Genomics Infrastructure (S, U)
- Ancient DNA (U)
- Microbial Single Cell Genomics (U)

Clinical Genomics

- Clinical Genomics Göteborg
- Clinical Genomics Linköping
- Clinical Genomics Lund
- Clinical Genomics Stockholm
- Clinical Genomics Umeå
- Clinical Genomics Uppsala
- Clinical Genomics Örebro

Metabolomics

- Swedish Metabolomics Centre (Um)
- Exposomics (S)

Clinical Proteomics and Immunology

- Autoimmunity and Serology Profiling (S)
- Translational Plasma Profiling (S)
- Proximity Proteomics (U)
- Mass Cytometry (Li, S)
- Proteogenomics (S)
- Glycoproteomics (G)

Single Cell and Spatial Biology

- Eukaryotic Single Cell Genomics (S)
- Spatial Proteomics(S)
- In Situ Sequencing (S)
- National Resource for Mass Spectrometry Imaging (U)
- Advanced FISH Technologies (S)
- Spatial Transcriptomics (S)

Cellular and Molecular Imaging

- Advanced Light Microscopy (S, G, Um)
- Cryo-EM (S, Um, G, Lu, U)

Integrated Structural Biology

- Swedish NMR Centre (G)
- Structural Proteomics (Lu)

Functional Biology and Target Discovery

- Chemical Biology Consortium Sweden (S, Um)
- Chemical Proteomics (S)
- Center for CRISPR-based Functional Genomics (S)
- Genome Engineering Zebrafish (U)

Drug Discovery and Development

- ADME (Absorption, Distribution, Metabolism, Excretion) of Therapeutics (UDOPP) (U)
- Biochemical and Cellular Assay (S)
- Biophysical Screening and Characterization (U)
- Human Antibody Therapeutics (Lu, S)
- In Vitro and Systems Pharmacology (U)
- Medicinal Chemistry – Hit2Lead (S)
- Medicinal Chemistry – Lead Identification (U)
- Protein Expression and Characterization (S)

SciLifeLab and Wallenberg National Program for Data-Driven Life Science

Knut och Alice Wallenbergs Stiftelse

SciLifeLab

WALLENBERG CENTRES FOR MOLECULAR MEDICINE

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

WACQT | Wallenberg Centre for Quantum Technology

CHALMERS

GÖTEBORGS UNIVERSITET

Karolinska Institutet

KTH VETENSKAP OCH KONST

LINKÖPINGS UNIVERSITET

LUNDS UNIVERSITET

SLU

Stockholms universitet

UMEÅ UNIVERSITET

Naturhistoriska riksmuseet

UPPSALA UNIVERSITET

Three dimensions of SciLifeLab

Research environment

Approx. 190 affiliated research groups

- Environment and climate change
- Farming and forestry
- Evolution and biodiversity
- Gene editing
- Biofuels and biomaterials
- Microbiology and microbiome
- Drugs and biomedicine
- Healthcare and aging

Infrastructure

Service to ~ 1400 Swedish researchers annually (2020)

- Bioinformatics
- Cellular and molecular imaging
- Clinical diagnostics
- Single cell biology
- Genomics
- Chemical biology and gene editing
- Drug development
- Proteomics and metabolomics

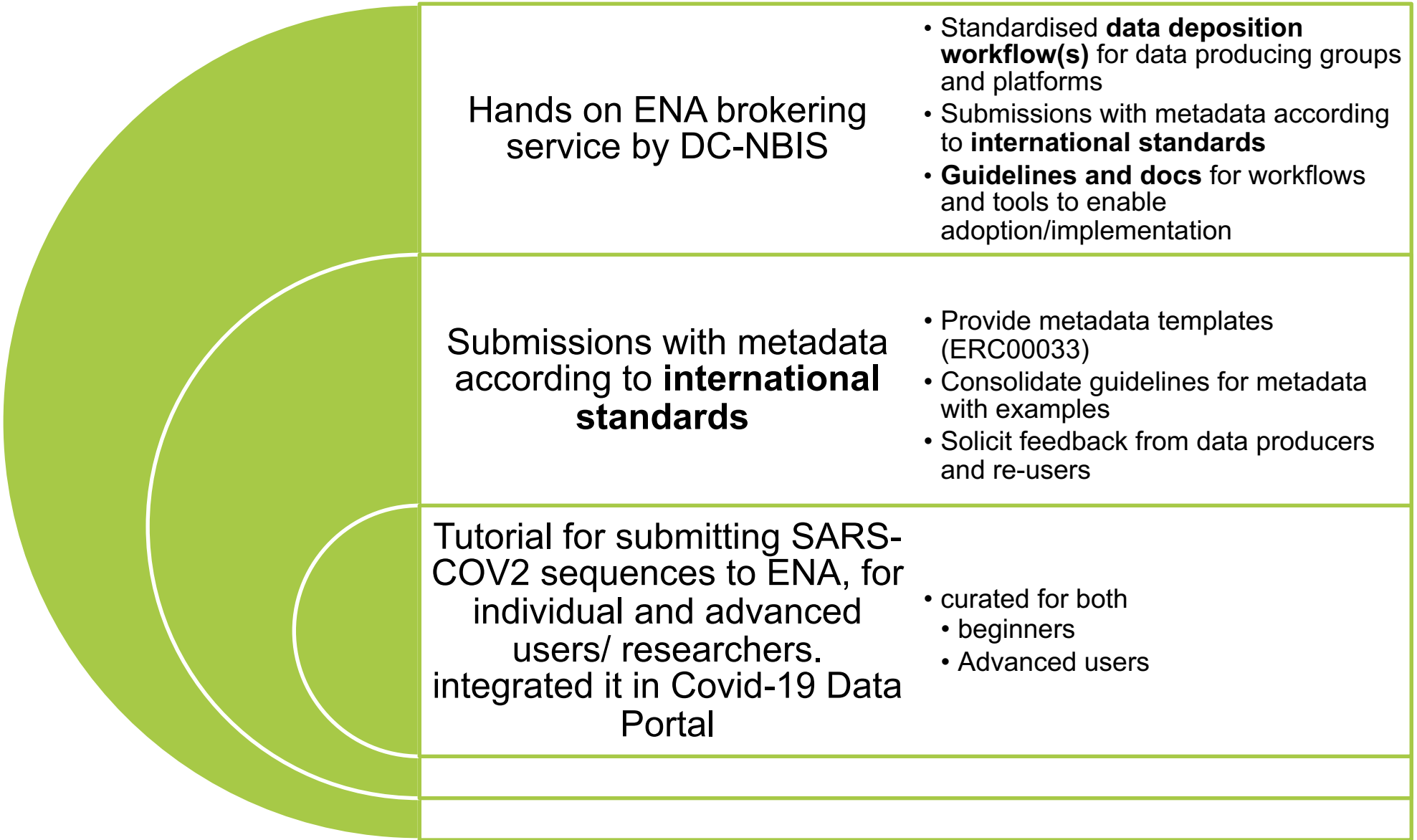
Data-driven life science

3.1 billion SEK, 12-year-program

Putting Sweden at the forefront of data-driven life science research and fostering the next generation of life scientists

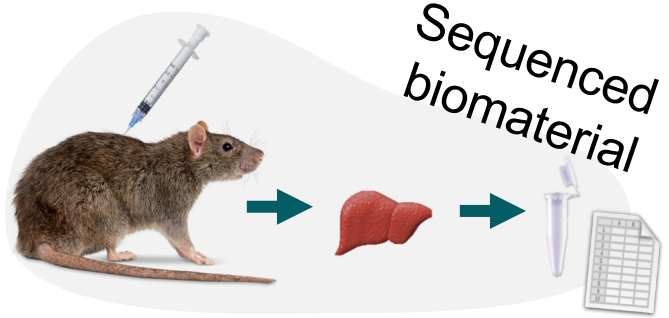
- Four strategic research areas
- Recruiting talent from across the globe
- Academic and industry PhD and postdoc programs
- Sparking collaborations, innovation and interdisciplinary team science
- Building a strong computational and data science base for open, real-time data

SARS-CoV-2 genome data to ENA



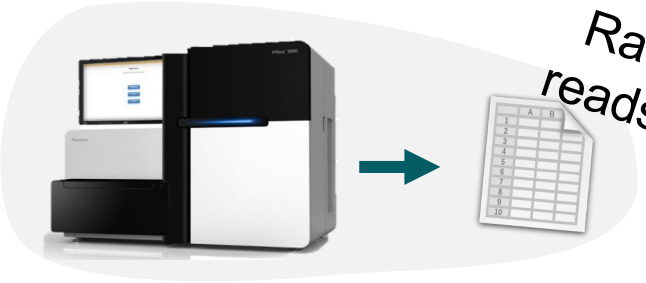
Mapping sequencing steps

SARS-CoV-2 sequencing step	ENA metadata object
Formulation of the scope of the SARS-CoV-2 study	**Study** Describe the focus of the project
Extraction/collection of samples from infected individuals	**Sample** Describe the biological sample and its sequencing conditions
Shotgun sequencing	**Experiment** Describe the sequencing performed with the sample
Reads/Sequences	**Run** Describe the raw data files (add files)
Assembly/Annotation /Analysis	**Analysis** Describe the data analysis and annotation



Sequenced biomaterial

.xlsx .tsv



Raw reads

.tsv
FASTQ

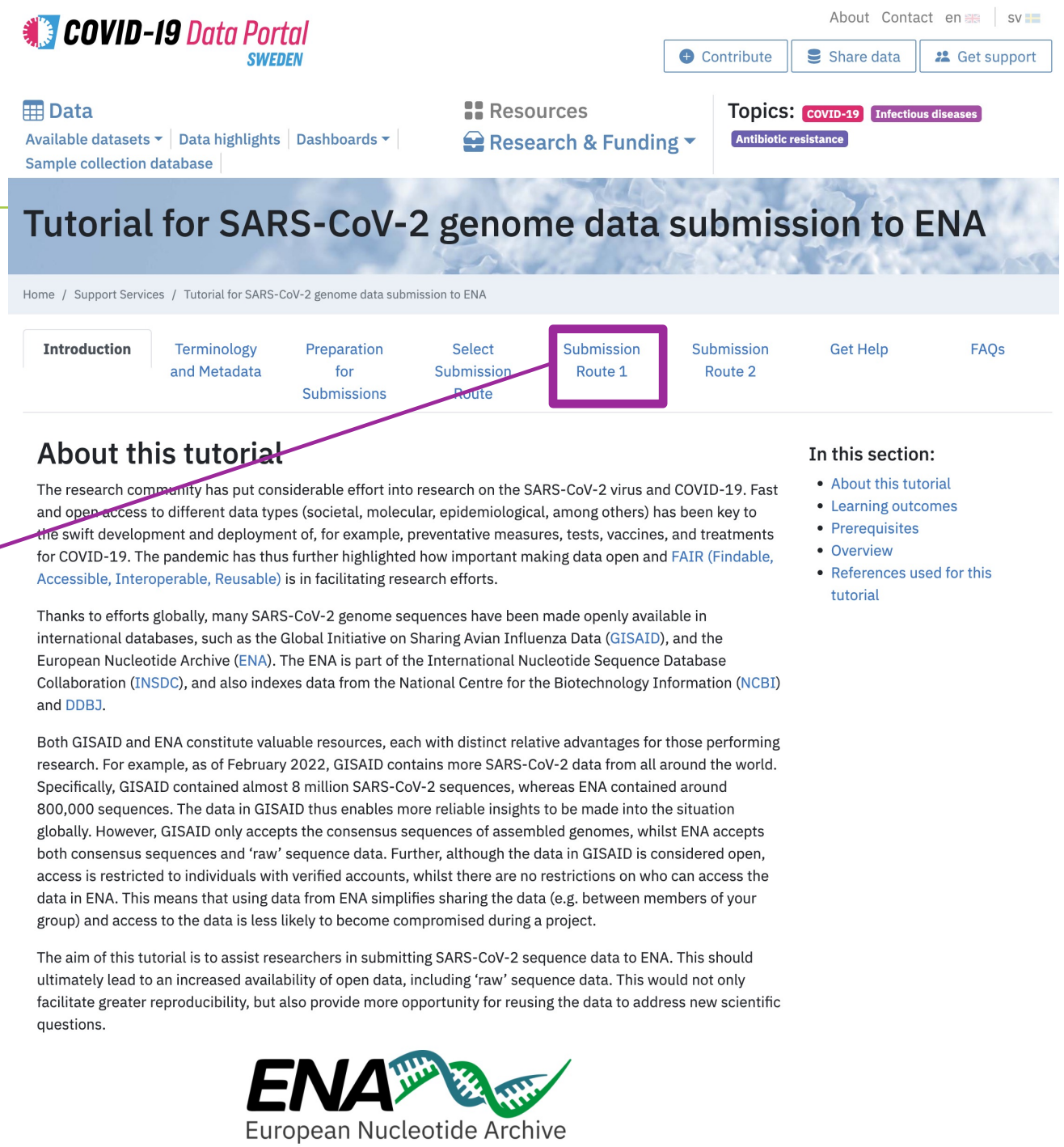
Data analyses

FASTA
chromosome list

Tutorial for submission to ENA

https://covid19dataportal.se/support_services/tutorial_ena/tutorial_ena_terminology/

For submissions:
- 1-100 sequences
- None or limited knowledge of command line

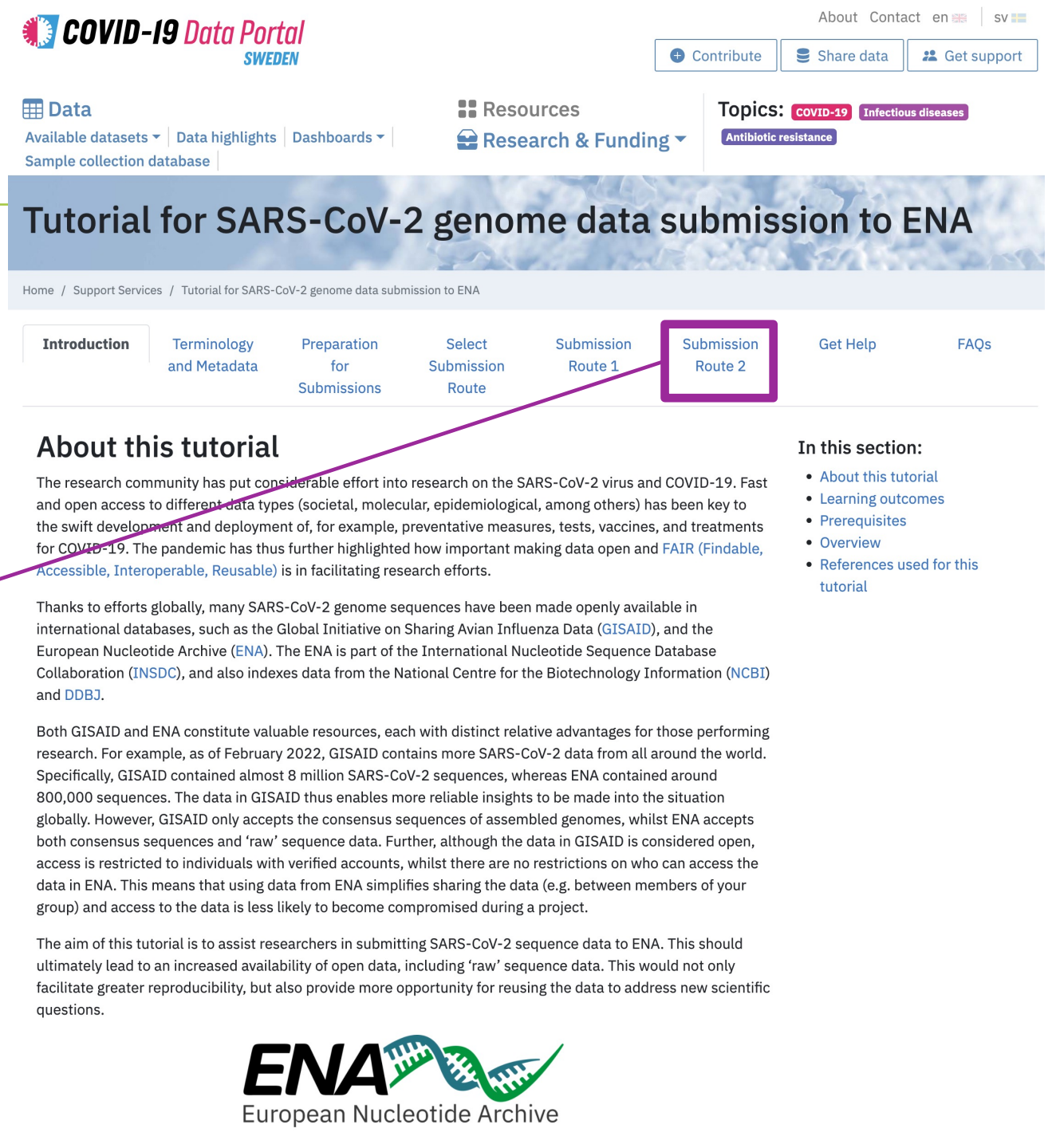


Tutorial for submission to ENA

https://covid19dataportal.se/support_services/tutorial_ena/tutorial_ena_terminology/

For submissions:

- More than100 sequences (batch uploads)
- Good knowledge of command line

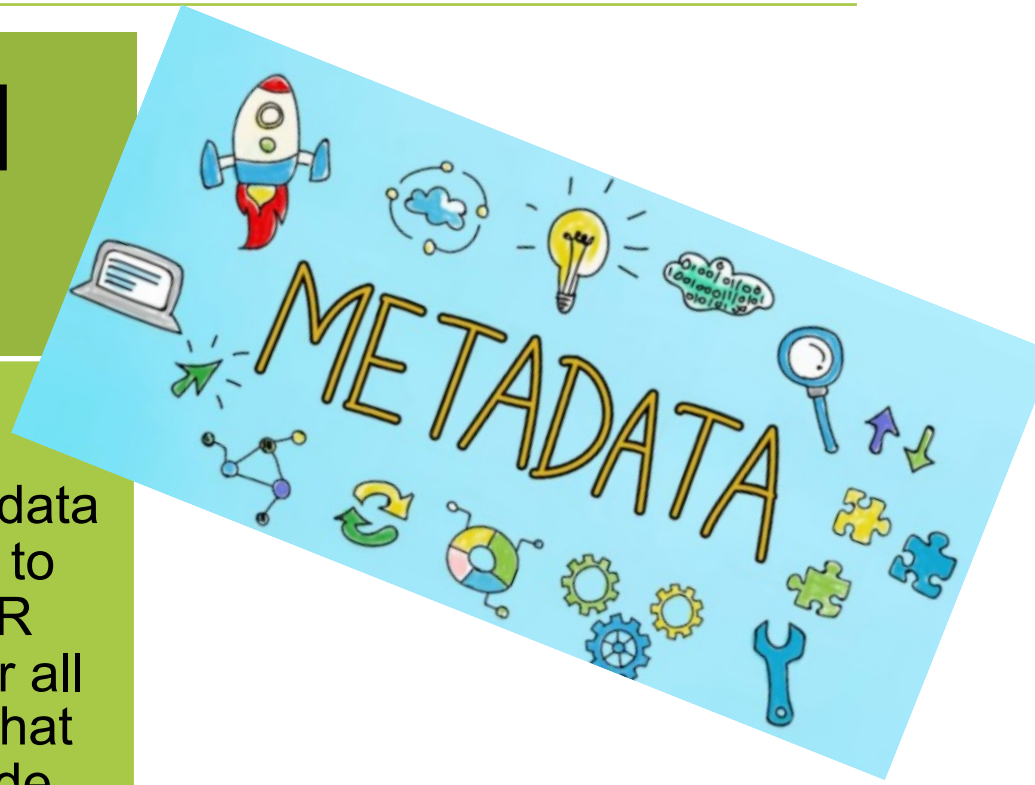


Brokering Best practices and tools

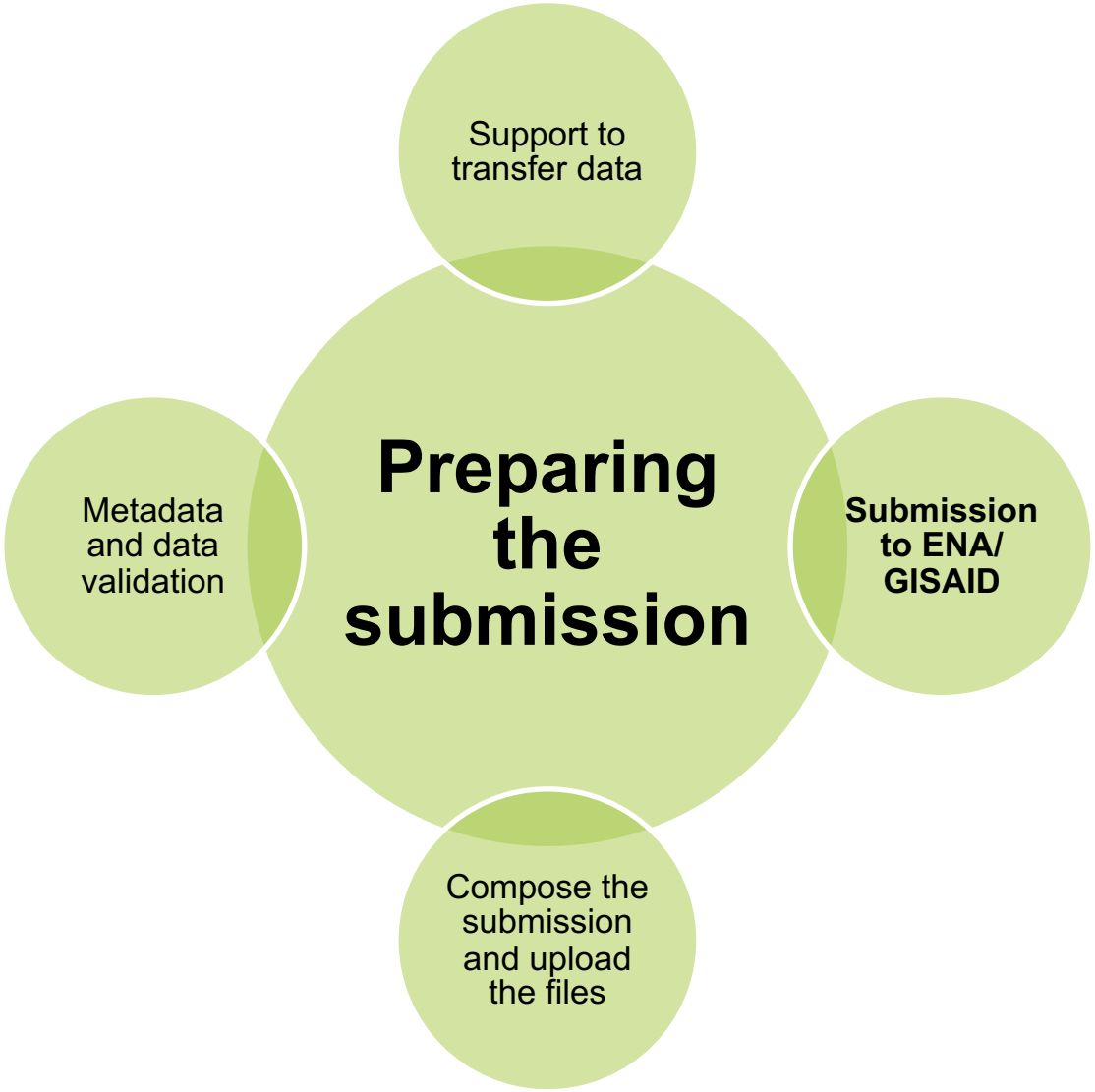
We must solve the problem of metadata. Solving the problem of metadata will never be easy.

Prepare the data and metadata

Support to get right metadata from researchers	Metadata schema- Create the metadata and identifiers according to FAIR principles Ontology / Controlled Vocabulary within attributes	Encourages data generators to follow FAIR principles for all data types that can be made machine actionable.

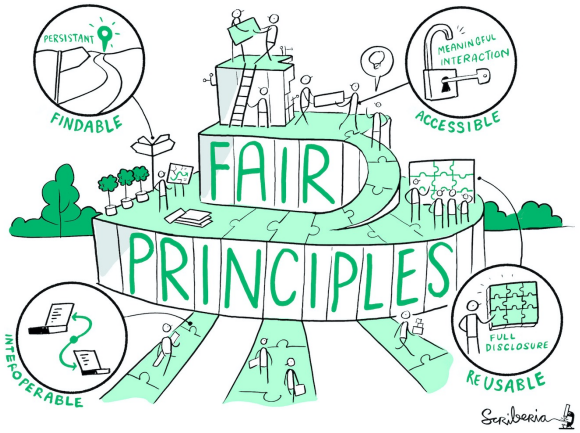


Brokering Best practices and tools…contd



Lessons learnt

- **Have data agreements in place**
  - Encourages data generators to adhere to appropriate (meta)data standards.
  - Open communication about anonymization
  - Encourage data generators to follow FAIR principles for all data types that can be made machine actionable
  - Create the metadata and identifiers according to FAIR principles
- Have enough storage before hand



SciLifeLab Covid-19 Portal team and NBIS

SciLifeLab

Johan Rung

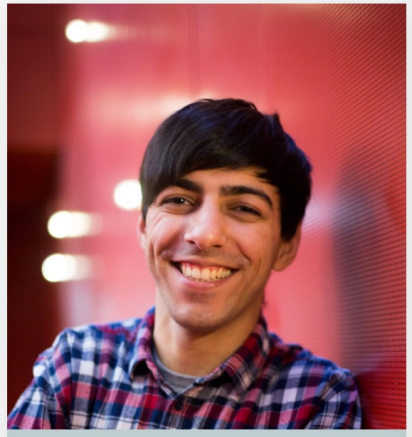
Head of Data Centre

Hanna Kultima

Data manager, coordinator

Anna Asklöf

Data Steward



Arnold Kochari

Webmaster

Katarina Öjefors Stark

Coordinator

Liane Hughes

Data Engineer

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Wolmar Nyberg Åkerstöm

Data Steward, NBIS

Yvonne Kalberg

Data Steward, NBIS

QUESTIONS?