

## Mobilisation of SARS-CoV-2 data

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

2022-05-19

Parul Tewatia

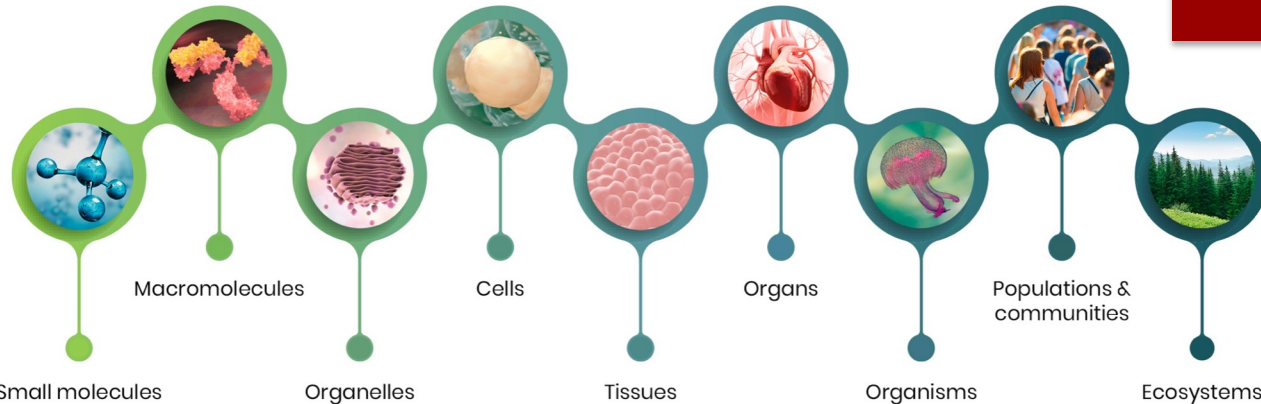
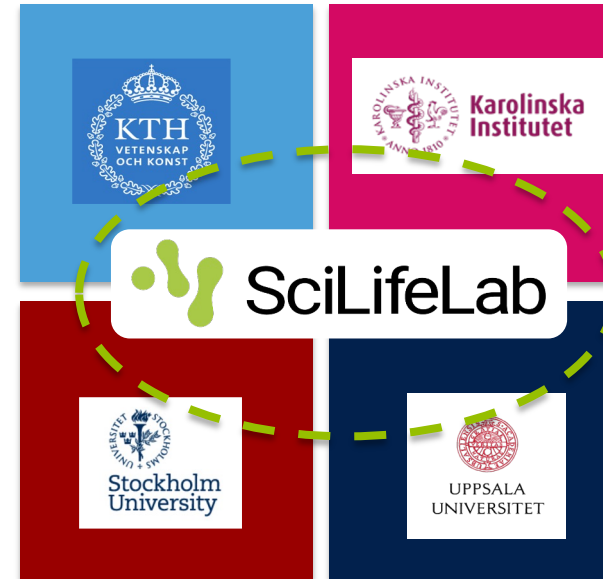
Data Centre ScilifeLab

[parul.tewatia@scilifelab.se](mailto: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

Gothenburg (GU, Chalmers)

Lund (LU)

Örebro (ÖRU)

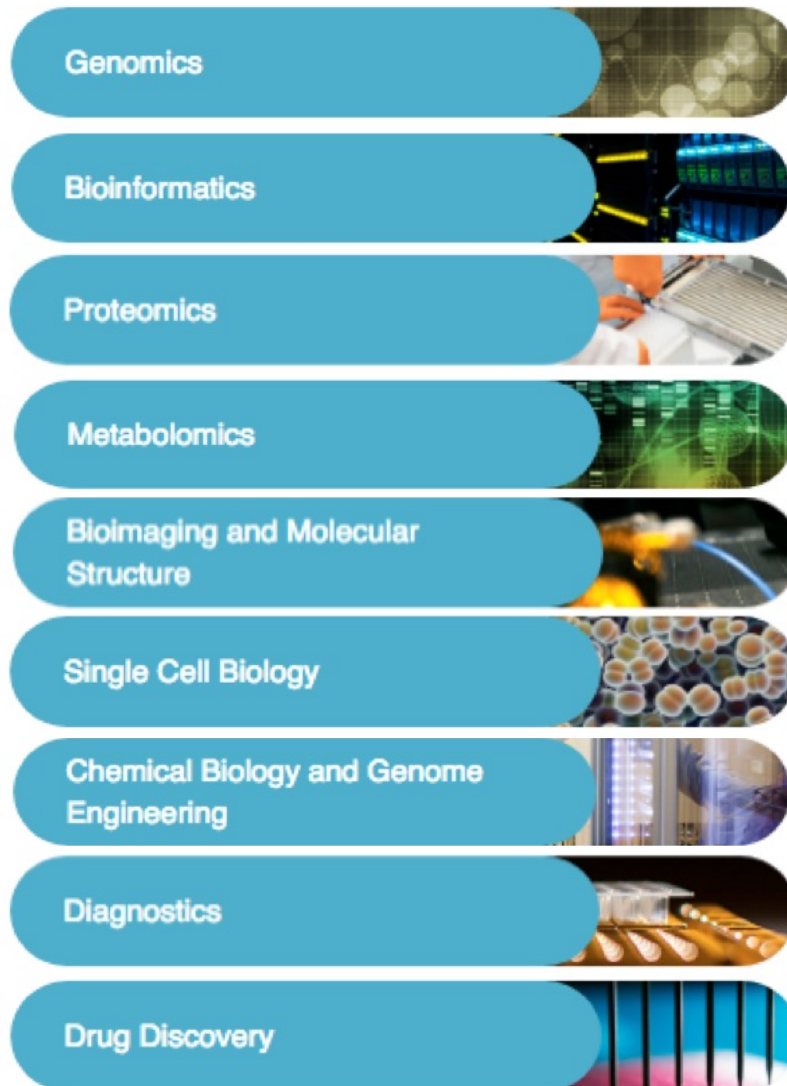
Uppsala (UU, SLU)

Stockholm (KI, KTH, SU)

Linköping (LiU)

Umeå (UmU, SLU)

# 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)
Biolmage 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*

WASP | WALLENBERG AI  
AUTONOMOUS SYSTEMS  
AND SOFTWARE PROGRAM

 **SciLifeLab**

WALLENBERG CENTRES FOR MOLECULAR MEDICINE

WACQT | Wallenberg Centre  
for Quantum Technology



# 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



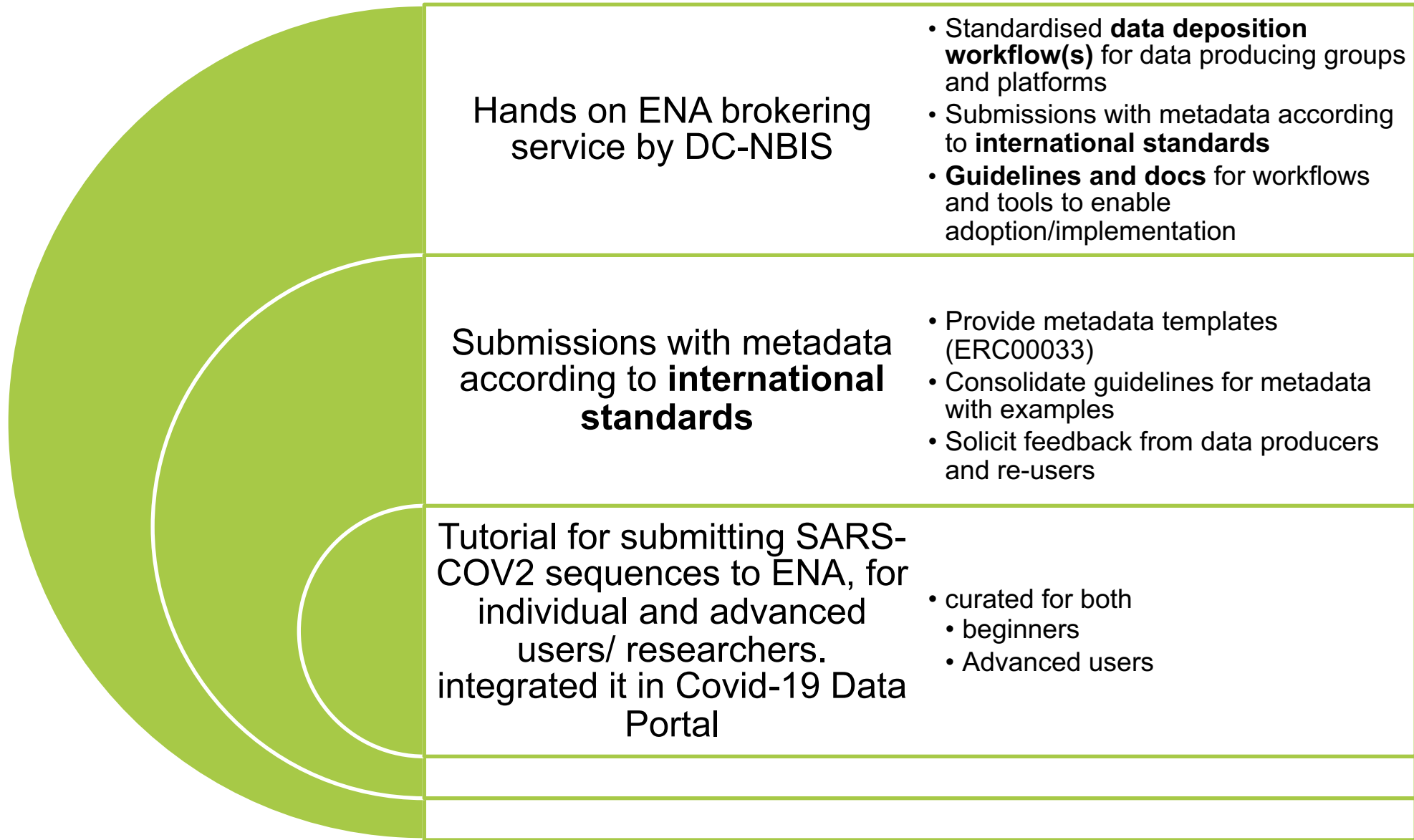
## Data-driven life science

**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

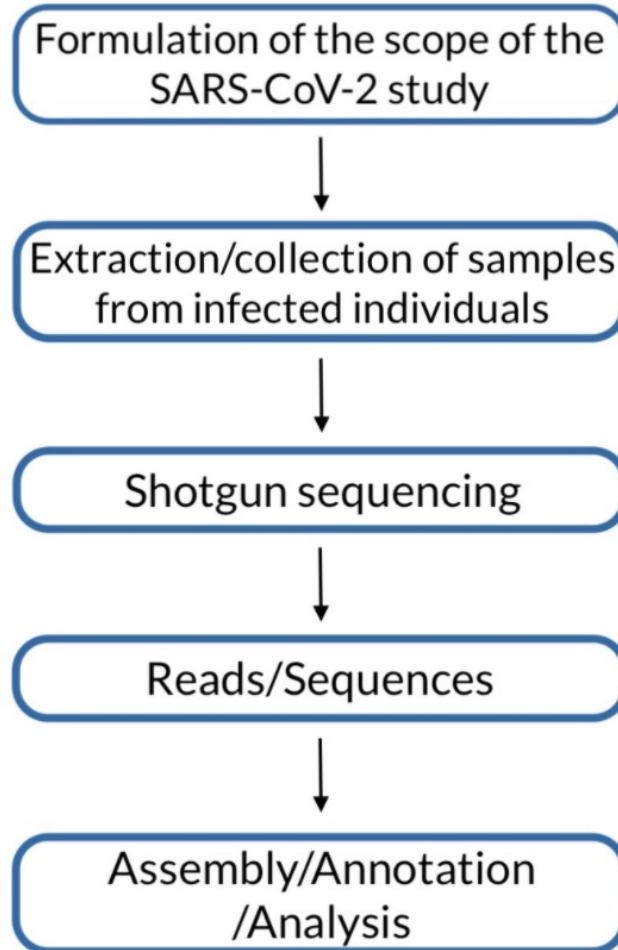




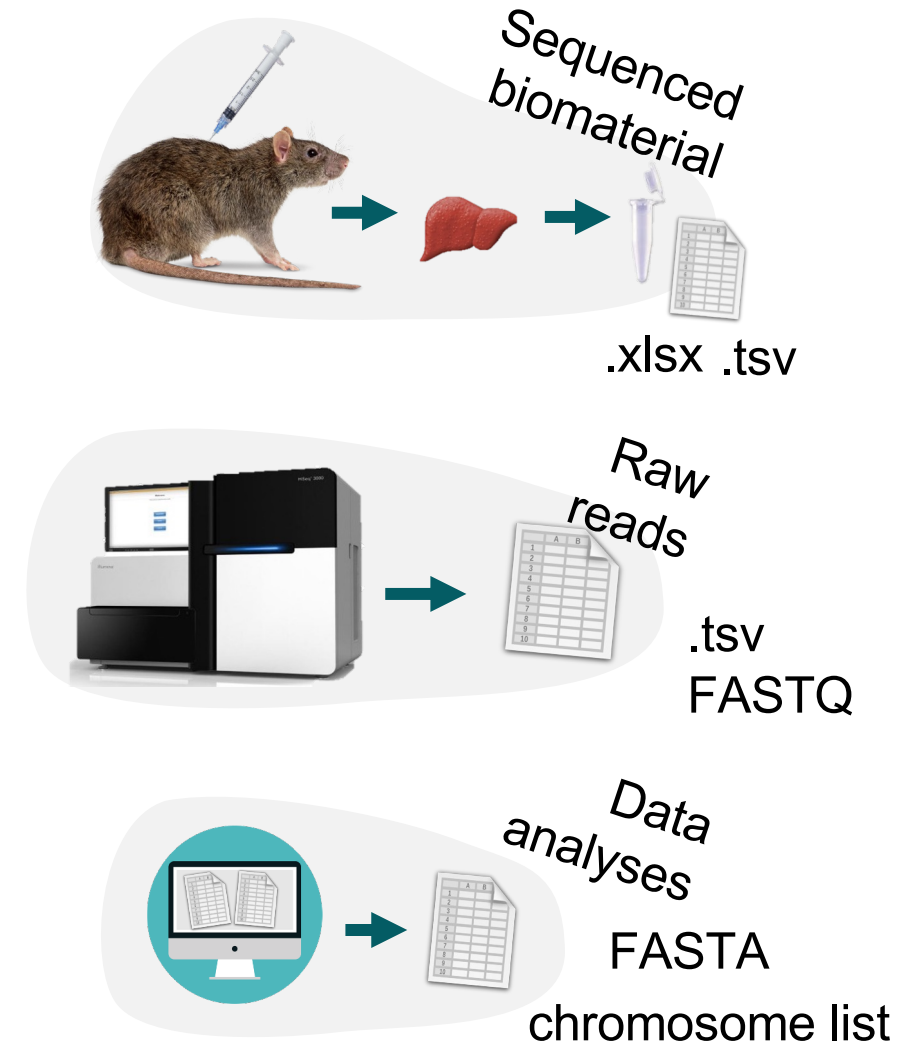
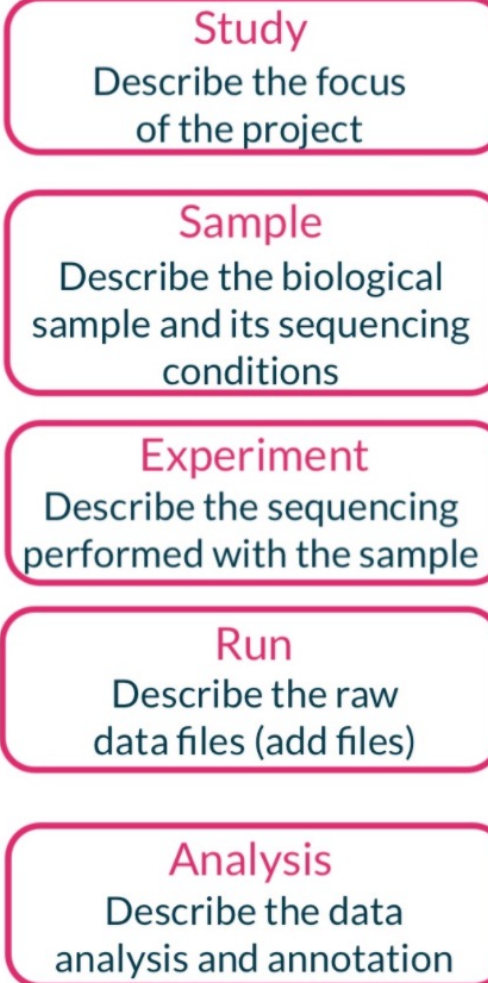
# Mapping sequencing steps



## SARS-CoV-2 sequencing step



## ENA metadata object



# Tutorial for submission to ENA

[https://covid19dataportal.se/support\\_services/tutorial\\_ena/tutorial\\_ena\\_terminology/](https://covid19dataportal.se/support_services/tutorial_ena/tutorial_ena_terminology/)

For submissions:

- 1-100 sequences
- None or limited knowledge of command line

## Tutorial for SARS-CoV-2 genome data submission to ENA

[Home](#) / [Support Services](#) / [Tutorial for SARS-CoV-2 genome data submission to ENA](#)

**Introduction**

[Terminology and Metadata](#)

[Preparation for Submissions](#)

[Select Submission Route](#)

**Submission Route 1**

[Submission Route 2](#)

[Get Help](#)

[FAQs](#)

### About this tutorial

The research community has put considerable effort into research on the SARS-CoV-2 virus and COVID-19. Fast and open access to different data types (societal, molecular, epidemiological, among others) has been key to the swift development and deployment of, for example, preventative measures, tests, vaccines, and treatments for COVID-19. The pandemic has thus further highlighted how important making data open and [FAIR \(Findable, Accessible, Interoperable, Reusable\)](#) is in facilitating research efforts.

Thanks to efforts globally, many SARS-CoV-2 genome sequences have been made openly available in international databases, such as the Global Initiative on Sharing Avian Influenza Data ([GISAID](#)), and the European Nucleotide Archive ([ENA](#)). The ENA is part of the International Nucleotide Sequence Database Collaboration ([INSDC](#)), and also indexes data from the National Centre for the Biotechnology Information ([NCBI](#)) and [DDBJ](#).

Both GISAID and ENA constitute valuable resources, each with distinct relative advantages for those performing research. For example, as of February 2022, GISAID contains more SARS-CoV-2 data from all around the world. Specifically, GISAID contained almost 8 million SARS-CoV-2 sequences, whereas ENA contained around 800,000 sequences. The data in GISAID thus enables more reliable insights to be made into the situation globally. However, GISAID only accepts the consensus sequences of assembled genomes, whilst ENA accepts both consensus sequences and 'raw' sequence data. Further, although the data in GISAID is considered open, access is restricted to individuals with verified accounts, whilst there are no restrictions on who can access the data in ENA. This means that using data from ENA simplifies sharing the data (e.g. between members of your group) and access to the data is less likely to become compromised during a project.

The aim of this tutorial is to assist researchers in submitting SARS-CoV-2 sequence data to ENA. This should ultimately lead to an increased availability of open data, including 'raw' sequence data. This would not only facilitate greater reproducibility, but also provide more opportunity for reusing the data to address new scientific questions.

### In this section:

- [About this tutorial](#)
- [Learning outcomes](#)
- [Prerequisites](#)
- [Overview](#)
- [References used for this tutorial](#)



# Tutorial for submission to ENA

[https://covid19dataportal.se/support\\_services/tutorial\\_ena/tutorial\\_ena\\_terminology/](https://covid19dataportal.se/support_services/tutorial_ena/tutorial_ena_terminology/)

For submissions:

- More than 100 sequences (batch uploads)
- Good knowledge of command line

## Tutorial for SARS-CoV-2 genome data submission to ENA

[Home](#) / [Support Services](#) / Tutorial for SARS-CoV-2 genome data submission to ENA

**Introduction**

[Terminology and Metadata](#)

[Preparation for Submissions](#)

[Select Submission Route](#)

[Submission Route 1](#)

**Submission Route 2**

[Get Help](#)

[FAQs](#)

### About this tutorial

The research community has put considerable effort into research on the SARS-CoV-2 virus and COVID-19. Fast and open access to different data types (societal, molecular, epidemiological, among others) has been key to the swift development and deployment of, for example, preventative measures, tests, vaccines, and treatments for COVID-19. The pandemic has thus further highlighted how important making data open and [FAIR \(Findable, Accessible, Interoperable, Reusable\)](#) is in facilitating research efforts.

Thanks to efforts globally, many SARS-CoV-2 genome sequences have been made openly available in international databases, such as the Global Initiative on Sharing Avian Influenza Data ([GISAID](#)), and the European Nucleotide Archive ([ENA](#)). The ENA is part of the International Nucleotide Sequence Database Collaboration ([INSDC](#)), and also indexes data from the National Centre for the Biotechnology Information ([NCBI](#)) and [DDBJ](#).

Both GISAID and ENA constitute valuable resources, each with distinct relative advantages for those performing research. For example, as of February 2022, GISAID contains more SARS-CoV-2 data from all around the world. Specifically, GISAID contained almost 8 million SARS-CoV-2 sequences, whereas ENA contained around 800,000 sequences. The data in GISAID thus enables more reliable insights to be made into the situation globally. However, GISAID only accepts the consensus sequences of assembled genomes, whilst ENA accepts both consensus sequences and 'raw' sequence data. Further, although the data in GISAID is considered open, access is restricted to individuals with verified accounts, whilst there are no restrictions on who can access the data in ENA. This means that using data from ENA simplifies sharing the data (e.g. between members of your group) and access to the data is less likely to become compromised during a project.

The aim of this tutorial is to assist researchers in submitting SARS-CoV-2 sequence data to ENA. This should ultimately lead to an increased availability of open data, including 'raw' sequence data. This would not only facilitate greater reproducibility, but also provide more opportunity for reusing the data to address new scientific questions.

### In this section:

- [About this tutorial](#)
- [Learning outcomes](#)
- [Prerequisites](#)
- [Overview](#)
- [References used for this tutorial](#)

# Brokering Best practices and tools

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

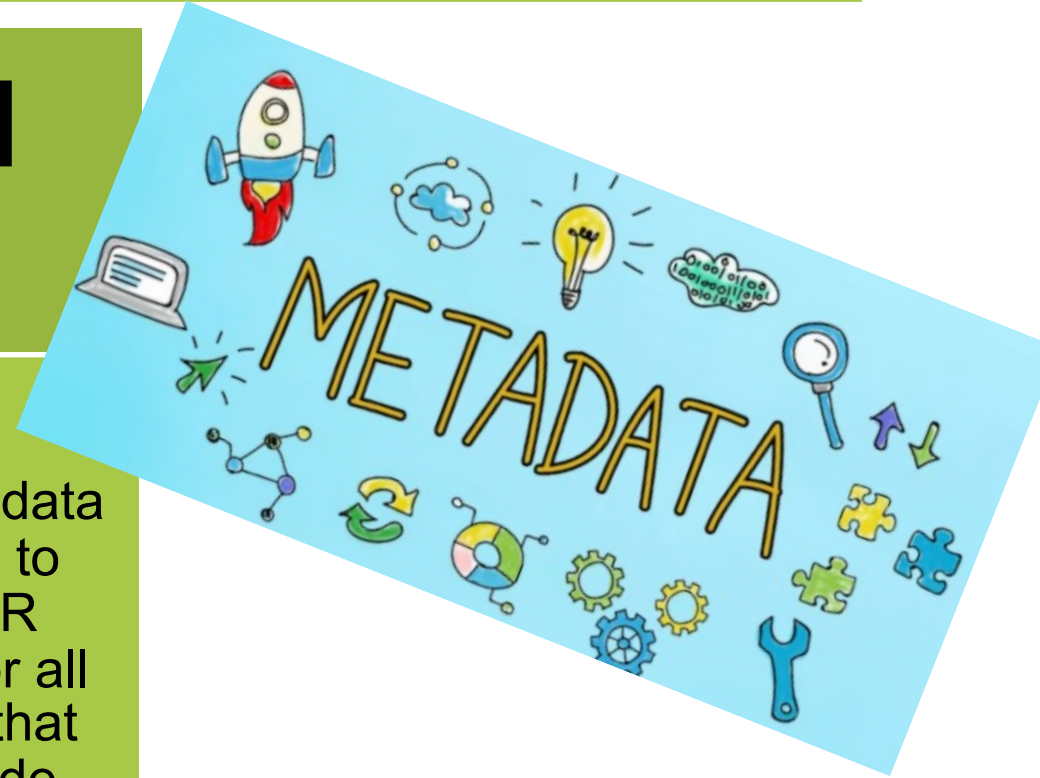


## Prepare the data and metadata

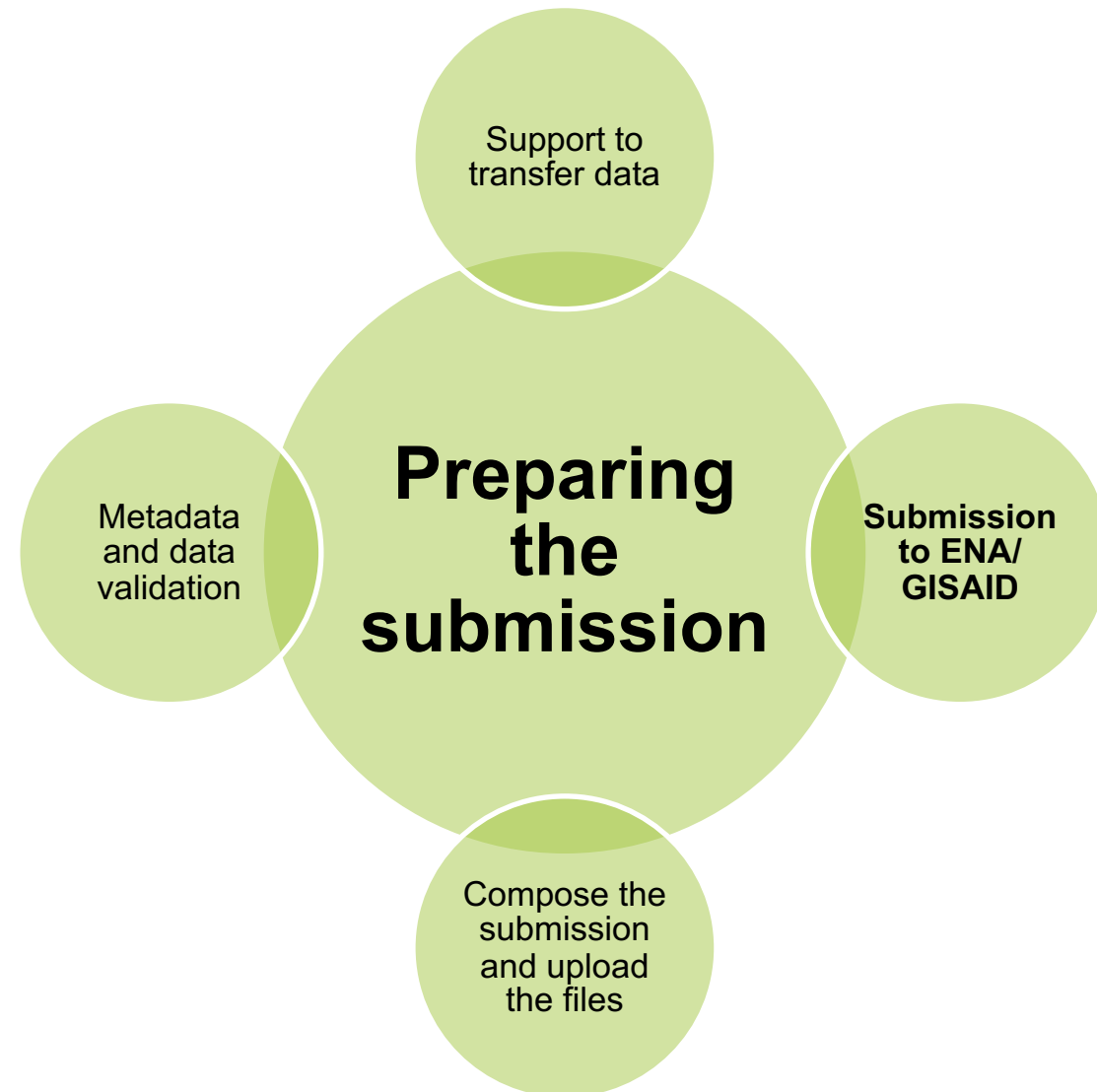
Support to get right metadata from researchers

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

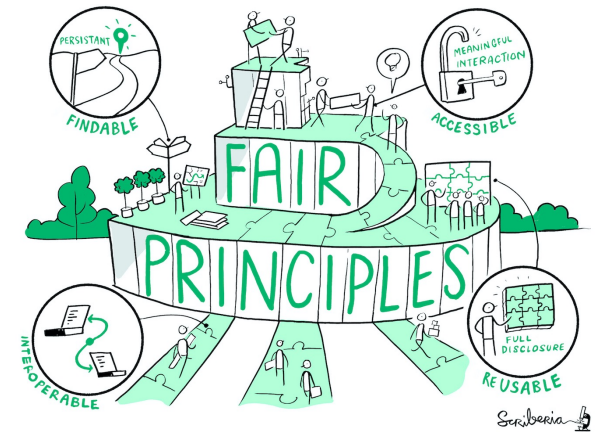
Encourages data generators to adhere to appropriate (meta)data standards.

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



**Johan Rung**  
Head of Data Centre



**Hanna Kultima**  
Data manager, coordinator



**Anna Asklöf**  
Data Steward



**Wolmar Nyberg  
Åkerstöm**  
Data Steward, NBIS



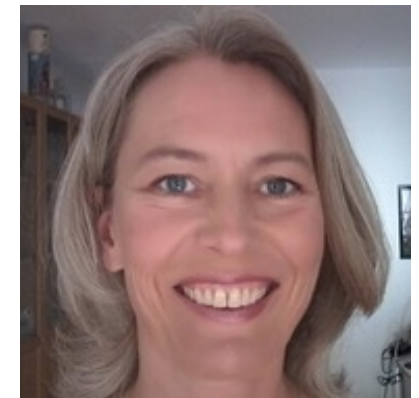
**Arnold Kochari**  
Webmaster



**Katarina Öjefors Stark**  
Coordinator



**Liane Hughes**  
Data Engineer



**Yvonne Kalberg**  
Data Steward, NBIS



**QUESTIONS?**

