Pandemic Research Infrastructure (PaRI)

“ The one-year PaRI project focused on facilitating Nordic research on pandemics and especially the COVID-19 pandemic”

1 nov 2020–31 okt 2021

Presented by Wolmar Nyberg Åkerström, wolmar.n.akerstrom@uu.se

www.neic.no/pari

PaRI

neic

A Nordic e-infrastructure project

- **6 partners** from Denmark, Estonia, Germany, Norway and Sweden
- **3 observers** from Finland, Norway and Sweden
- Liaised with **Nordic & European activities**
- Diverse and active reference group

EOSC-Life

EOSC NORDIC

elixir CONVERGE

PRACE

COVID-19 Data Portal

Heilsa

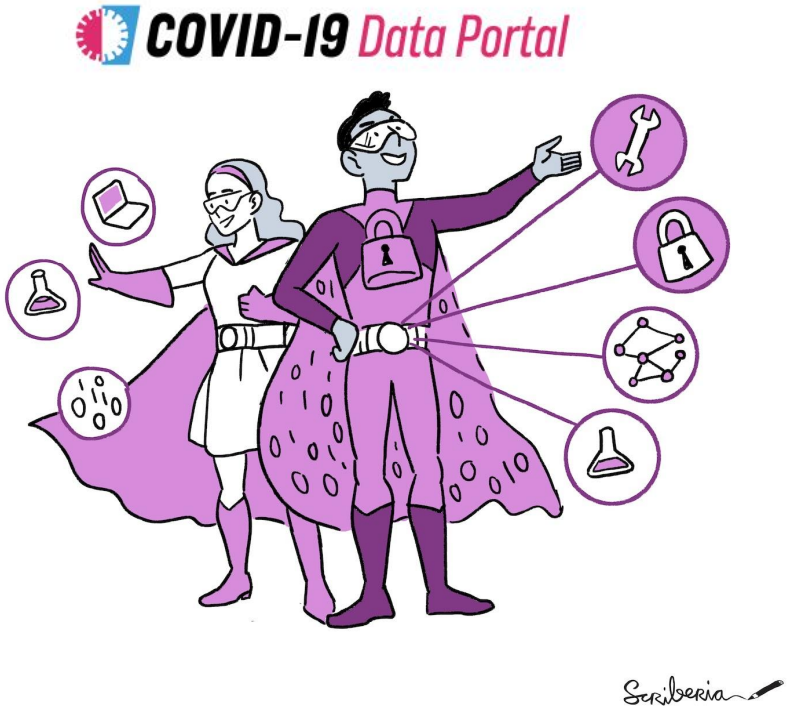
Enable easy and seamless access

WP1 Data management

Support for data sharing

Access to data resources

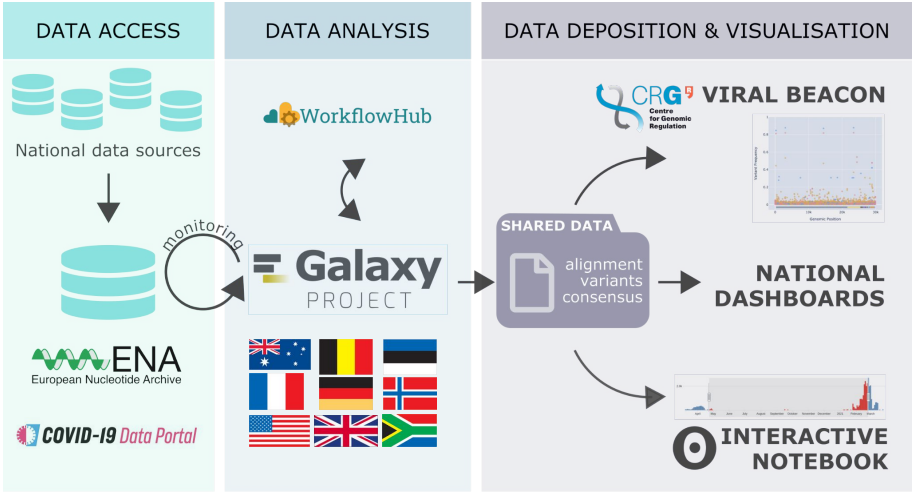
Index of infrastructure services



Enable easy and seamless access

WP1 Data management
- Support for data sharing
- Access to data resources
- Index of infrastructure services

WP2 **Analysis platform**
- Custom analysis portal (Galaxy)
- Selection of tools & workflows
- Proof of concept deployment



Enable easy and seamless access

WP1 Data management

Support for data sharing

Access to data resources

Index of infrastructure services

WP2 Analysis platform

Custom analysis portal (Galaxy)

Selection of tools & workflows

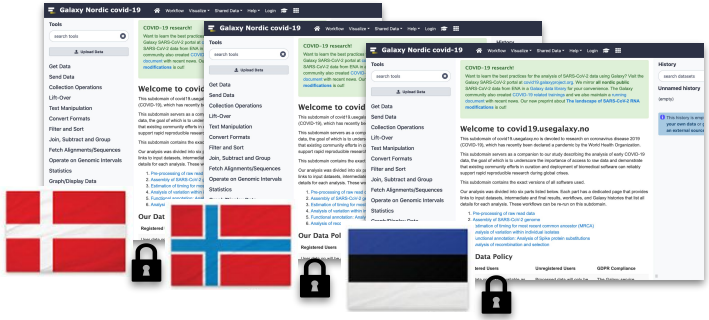
Proof of concept deployment

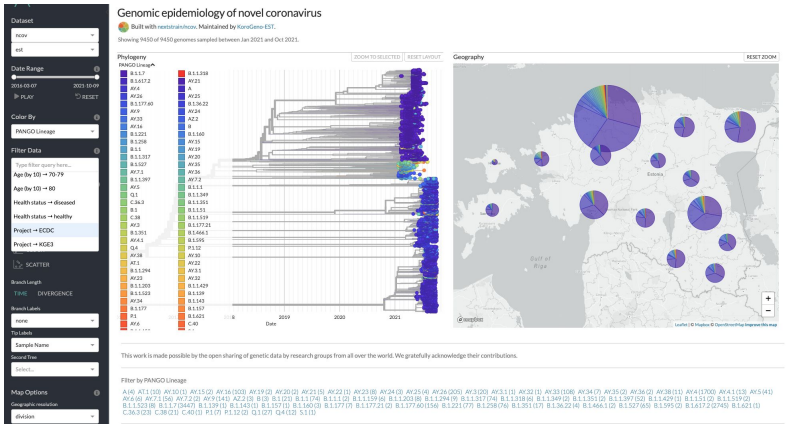
WP3 Scaling the analysis

Workflows for data sharing

Local Galaxy for sensitive data

Nordic pandemic dashboard





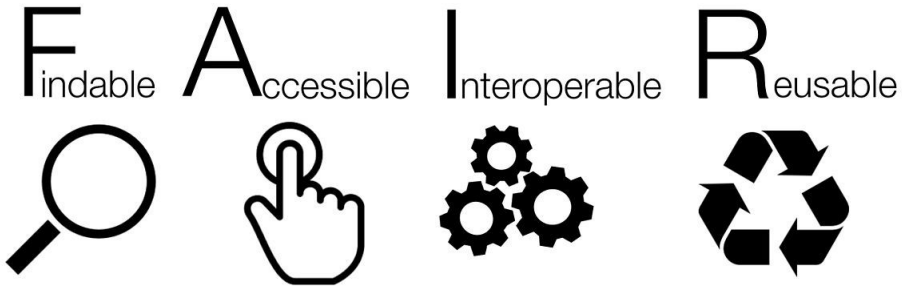
Enable easy and seamless access

WP1 Data management
- Support for data sharing
- Access to data resources
- Index of infrastructure services

WP2 Analysis platform
- Custom analysis portal (Galaxy)
- Selection of tools & workflows
- Proof of concept deployment

WP3 Scaling the analysis
- Workflows for data sharing
- Local Galaxy for sensitive data
- Nordic pandemic dashboard

Focusing on pandemic data



Enable easy and seamless access

WP1 Data management
- Support for data sharing
- Access to data resources
- Index of infrastructure services

WP2 Analysis platform
- Custom analysis portal (Galaxy)
- Selection of tools & workflows
- Proof of concept deployment

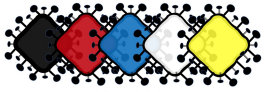
WP3 Scaling the analysis
- Workflows for data sharing
- Local Galaxy for sensitive data
- Nordic pandemic dashboard

Focusing on pandemic data

Findable Accessible Interoperable Reusable

Combining Nordic & EU resources

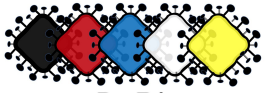
➔ No development of analysis tools
➔ No transfer of data ownership
➔ No operational services
➔ No new hardware



Sensitive pandemic data

Sensitive personal data can be directly *or* indirectly linked to an individual, such as

- ❏ Patient information
- ❏ Personal and phenotypic information
- ❏ Pathogen genomic information with fragments of human genetic information
- ❏ Sample and process information



Sensitive pandemic data

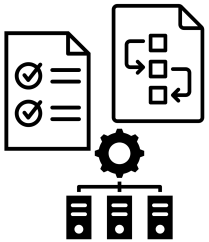
Sensitive personal data can be directly *or* indirectly linked to an individual, such as

- ❑ Patient information
- ❑ Personal and phenotypic information
- ❑ Pathogen genomic information with fragments of human genetic information
- ❑ Sample and process information

- **Ethics approval and/or contracts** required for use and for sharing, including partners, service providers and data sharing platforms.
- Except when it's no longer possible to link the data to an individual.

What agreements & guidance can be prepared for the next outbreak?

FAIR by design for pandemic data



Study & data design

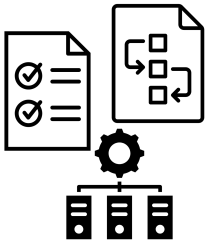
Procedures

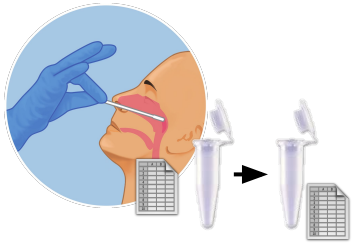
data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, …

Information system

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com

FAIR by design for pandemic data





Study & data design | Sampling & specimen collection | Sample preparation | Sample analysis & data generation

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, …

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, …

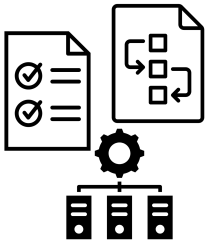
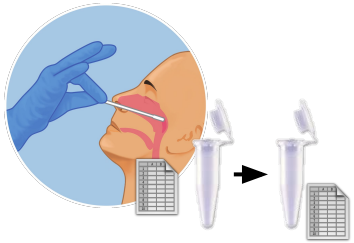
Information system

Sample workspace

Biobank

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com

FAIR by design for pandemic data

Study & data design | Sampling & specimen collection | Sample preparation | Sample analysis & data generation | Data processing to prepare inputs for analysis | Data analysis

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, …

Information system

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, …

Sample workspace

Biobank

Data and computational workflows

digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data, …

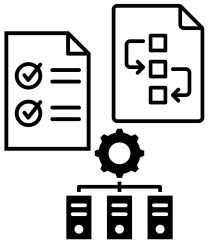
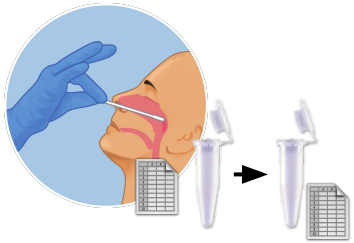
Data delivery

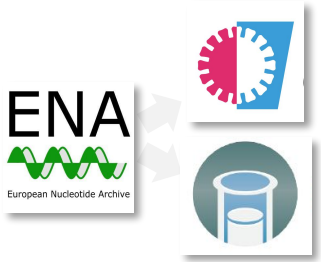
Digital workspace

Local data archive etc

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com

FAIR by design for pandemic data

Study & data design	Sampling & specimen collection	Sample preparation	Sample analysis & data generation	Data processing to prepare inputs for analysis	Data analysis	Communicating results

Procedures
data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, …

Information system

Biosamples and instruments
populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, …

Sample workspace

Biobank

Data and computational workflows
digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data, …

Data delivery

Digital workspace

Outputs
publications,
data,
tools,
workflows,
reports,
dashboards, …

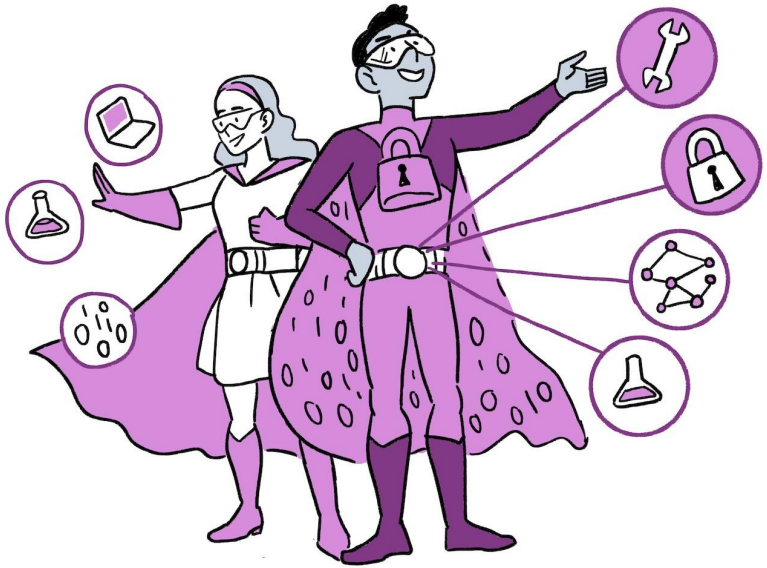
Local data archive etc

Research databases

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com

Support for data sharing

- **Guidance to projects, labs and other organisations** producing or commissioning viral sequencing data
- **What constitutes a good and ultimately (re)usable** pathogen genome data record with a Nordic perspective?
- **National COVID-19 Data Portals** as channels to get support



COVID-19 Data Portal

Access to data resources

- **COVID-19 data portal (ENA)**
  Open & direct access to assembled genomes and raw reads
- **SARS-CoV-2 Contextual**
  Manually curated metadatabase with links to sequences
- **Federated EGA**
  Controlled access to sensitive data stored nationally

Access to data resources

- **COVID-19 data portal (ENA)**
  Open & direct access to assembled genomes and raw reads
- **SARS-CoV-2 Contextual**
  Manually curated metadatabase with links to sequences
- **Federated EGA**
  Controlled access to sensitive data stored nationally
- **GISAID EpiCoV™**
  Contract/license-based access to assembled genomes
- **ECDC TESSy/EpiPulse**
  Controlled access to surveillance and reporting coordinated by national public health authorities
- **National registers**
  Often controlled access to health data, statistics, etc.

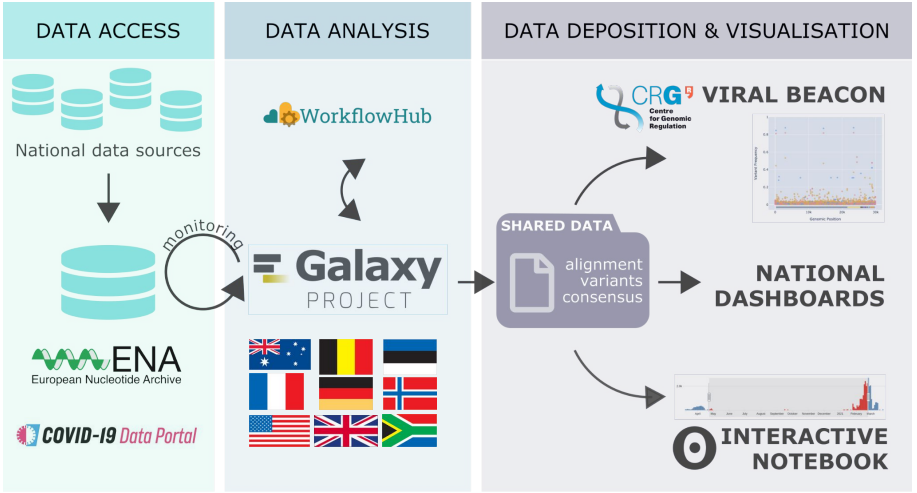
Access to data resources

- **COVID-19 data portal (ENA)**
  Open & direct access to assembled genomes and raw reads
- **SARS-CoV-2 Contextual**
  Manually curated metadatabase with links to sequences
- **Federated EGA**
  Controlled access to sensitive data stored nationally
- **GISAID EpiCoV™**
  Contract/license-based access to assembled genomes
- **ECDC TESSy/EpiPulse**
  Controlled access to surveillance and reporting coordinated by national public health authorities
- **National registers**
  Often controlled access to health data, statistics, etc.

What data should a fully operational pandemic research infrastructure cover?

Analysis platform using Galaxy

- Open platform connecting **data**, computational **workflows**, **visualisations** and other **services**
- Supports **reproducible**, **and transparent** computational analysis
- Community committed to **improving tools & workflows**



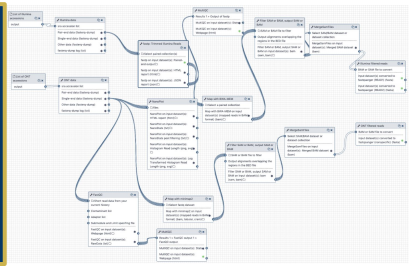
Galaxy · Galaxy EUROPE · Galaxy AUSTRALIA

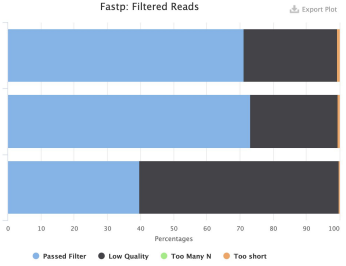
EOSC-Life · elixir · Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

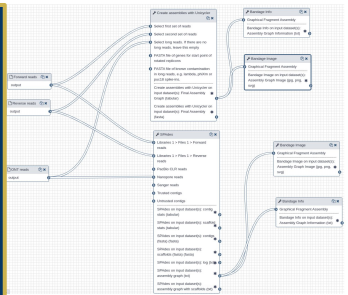
What can be done to promote FAIR for workflows?

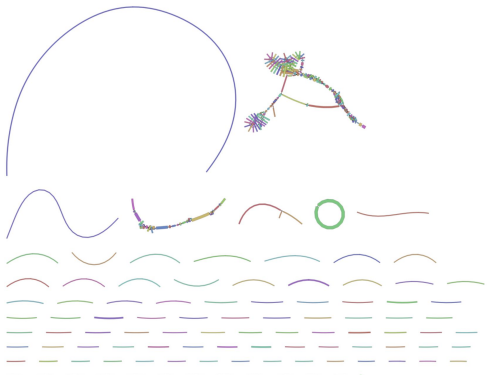
Pre-processing



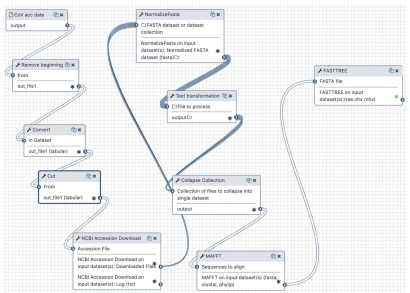


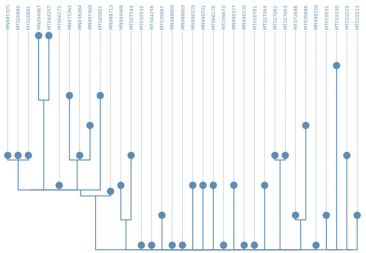
Assembly



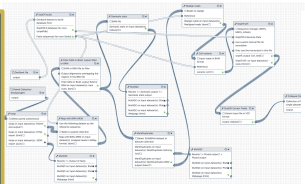


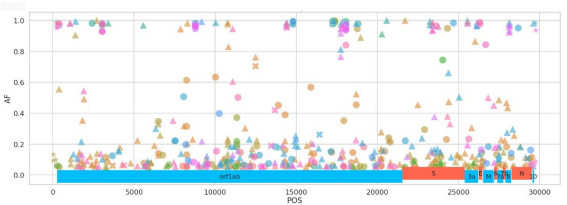
MRCA timing



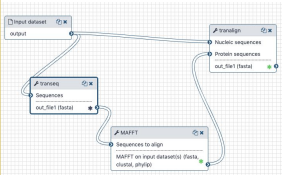


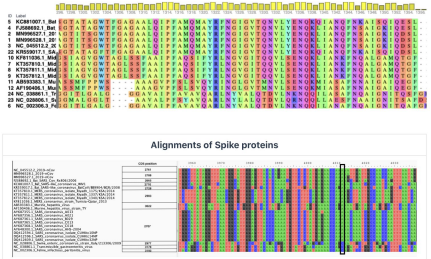
Variation analysis

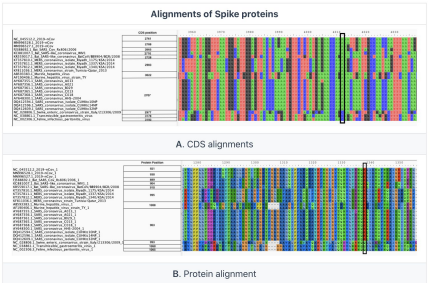




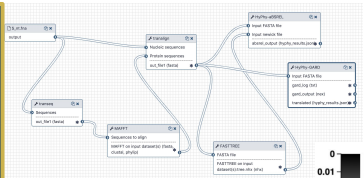
S- analysis

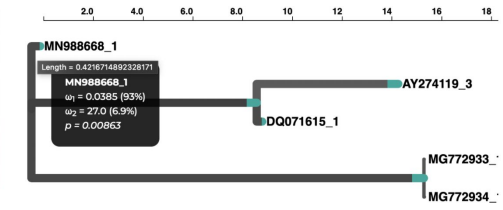






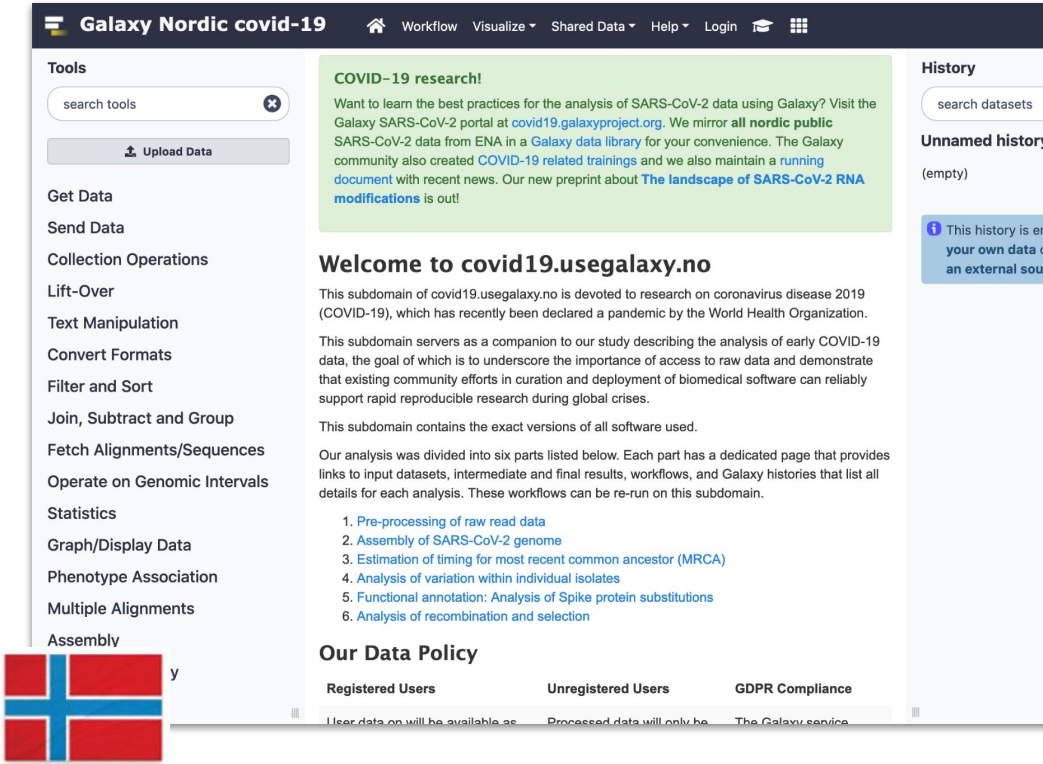
Evolutionary analysis





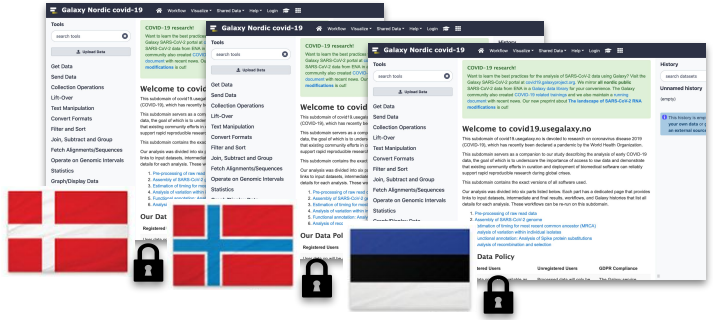
Custom analysis portal (Galaxy)

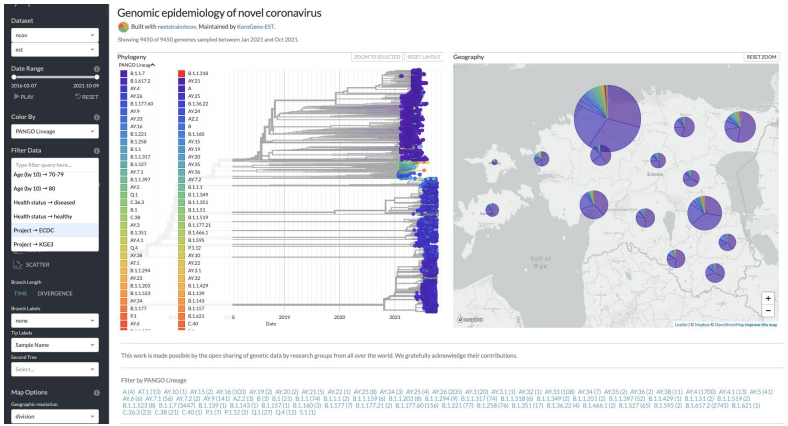
- Convenient access to Nordic SARS-CoV-2 genomic **data and analysis workflows**
- **Managed collection of tools**, versions and documentation
- **Bring your own** data and build your own workflows
- **Automated configuration** to allow replication and sharing across environments



Scaling the analysis

- FAIR data sharing requires planning workflows for **data mapping** and **anonymisation**
- The analysis portal can be sandboxed to **run on secure infrastructures**
- Proof of concept “real-time” dashboard of geographic distribution of virus variants over time





What common data definitions can we agree on across the Nordics?

Providing Nordic Added Value

- **Increased capabilities** to support pandemic data sharing and analysis
- Contributed to making **more pathogen genome data** available for research
- Strengthened contributions to related activities **not only in the Nordics**
- Work continues in European projects and acoss partners

EOSC-Life

Heilsa

BY-COVID

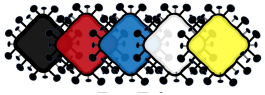
elixir CONVERGE

PRACE

COVID-19 Data Portal

EOSC NORDIC

eosc



Extra slides…

FAIR Pandemic data sharing

Robot: MetaManMachine by Nikola Vasiljevic (2021), CC BY-SA 4.0, doi:10.5281/zenodo.4471098

Background from 'FAIR Principles' by Martínez-Lavanchy, et al (2019), CC-BY 4.0. doi:10.11581/dtu:00000049

Success factors and impact

- ❏ **Timely and rewarding** platform for knowledge exchange between the partners & stakeholders
- ❏ Useful to address **data sharing guidelines and access** from workflows in parallel
- ❏ **Satisfying to see a complete prototype** for managing and executing tools and workflows on Nordic pandemic data sets

- ➔ Increased our **capabilities** to support pandemic data sharing and analysis
- ➔ Contributed to making **more pathogen genome data** available for research
- ➔ Contributed experiences to related activities **not only in the Nordics** but also worldwide

PaRI continues as a NeIC Affiliate

Professional network

- ❑ FAIRification of pandemic related omics data
- ❑ Running and administering COVID-19 workflows in Galaxy
- ❑ Curating pandemic data for visualisations in Nextstrain / Auspice (possibly other visualisation tools)

Outreach

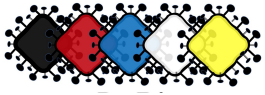
- ❑ Expand the professional network to new members
- ❑ Share experiences with active projects and new partners

Shared tools & workflows

- ❑ Populate a GitHub organisation with tools and workflows from the partners

PaRI affiliate coordinator:
Wolmar Nyberg Åkerström
wolmar.n.akerstrom@uu.se

https://neic.no/affiliates/



EOSC use case: Data & metadata

Access to genome data for the virus that causes COVID-19

PaRI provided guidance and hands-on support to deposit genome data and the information collected with the samples according to the schemas and data specifications supported by two different data sharing platforms.

Genome data is acquired by analysing samples collected from individuals. The samples are collected by a diverse set of organisations ranging from public health surveillance initiatives to research projects focusing on different aspects of the virus or the disease. What information is collected with the samples, how the information collected is structured, and in what detail information can be shared varies. To be widely accessible to the public health and research communities the data should be shared through different data sharing platforms with overlapping but different schemas to structure and encode the information. Access to sufficiently detailed data specifications and data mappings is required when developing mediary tools and resources that enable syntactic and semantic interoperability.

EOSC use case: Tools & workflows

Access to tools and workflows to analyse genome data for the virus

PaRI designed data flows and provisioned/hosted workflow systems for running computational workflows on data from data producing partners as well as data available in public data sharing platforms.

There are well-established data formats for storing genome data and the bioinformatics community has developed a wide array of shared practices, tools and computational workflows to automate some aspects of quality control and data analysis. Adopting/developing a computational workflow includes specifications of the expected input and output data as well as considerations on how to document the process to enable peer-review and reproducibility. If all information required for a workflow is available in a data sharing platform, that information could be mapped to the corresponding input data specification resulting in consistently produced outputs for large collections of data.

EOSC use case: Dashboards

Access to a dashboard showing geographic distribution of virus variants over time

PaRI designed data flows for regional surveillance and hosted a proof of concept for a dashboard based on Auspice/Nextstrain. Curated public data/data from data producing partners. Mapped to a database hosting contextual data for SARS-CoV-2 sequences.

Aggregate data and visualisations presented on regional dashboards can be used to convey information about how a virus/outbreak evolves and spreads and to provide an overview of what data is available for research. The dashboard presents consistent views of underlying data that may have been produced according to varying naming and encoding schemas, such as definitions of geographic regions, age groups, labelling of virus characteristics, etc. A dynamic dashboard requires workflows to be available for mapping information from all the sources of underlying data to the views presented on the dashboard.