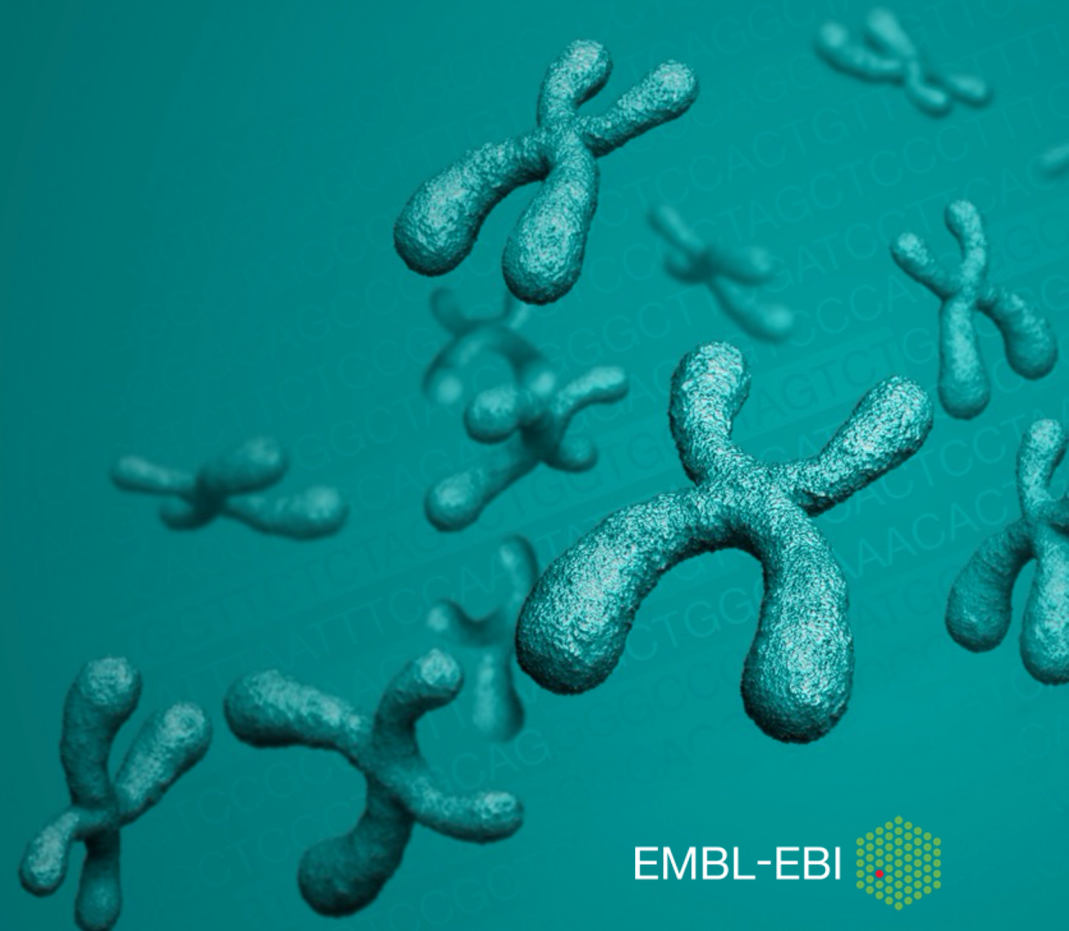


# The European COVID-19 Data Platform

Guy Cochrane



# European COVID-19 Data Platform

The screenshot shows the COVID-19 Data Portal website. The header includes the logo and navigation links: About, Data Hubs, Related resources, Our partners, and Submit data. Below the header is a navigation bar with categories: Sequences, Expression, Proteins, Structures, Compounds, Targets, and Literature. The main content area features a banner with the text "Accelerating research through data sharing" and a grid of data categories, each with a description and a record count:

- Sequences**: Raw and assembled sequences related to the COVID-19 outbreak, including outbreak isolates and records relating to coronavirus biology. Includes extensive sampling information. 22,739 records >
- Expression data**: Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections. 51 records >
- Proteins**: Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors. 138 records >
- Structures**: SARS-CoV-2 protein structures, highlighting important structural features to support the development of treatments and vaccines. 229 records >
- Compounds**: Curated range of SARS-CoV-2 compound activity data. 8 records >
- Targets**: Range of SARS-CoV-2 and SARS-CoV data for drug target identification and prioritisation. 24 records >
- Literature**: Search for the latest literature about SARS-CoV-2. 83,945 publications >
- Related resources**: A range of related resources for studying the SARS-CoV-2 coronavirus and the COVID-19 disease.

There is also a section "About this portal" and a "Share new COVID-19 data" button. The footer includes contact information, a "Support & Feedback" button, and logos for EMBL-EBI, ELIXIR, and the European Commission.

The screenshot shows the EC AV Portal video page for the launch of the EU COVID-19 Data Platform. The page features a navigation bar with "HIGHLIGHTS" and links to Coronavirus, College, #DigitalEU, #EUGreenDeal, Press Conferences, Midday Briefings, and Press corner. The main content area has a large heading: "Statement by Ursula von der Leyen, President of the European Commission, on the launch of the EU COVID19 Data Platform (international sign language version)". Below the heading is a video player showing Ursula von der Leyen speaking with a sign language interpreter. To the right of the video player is a metadata sidebar:

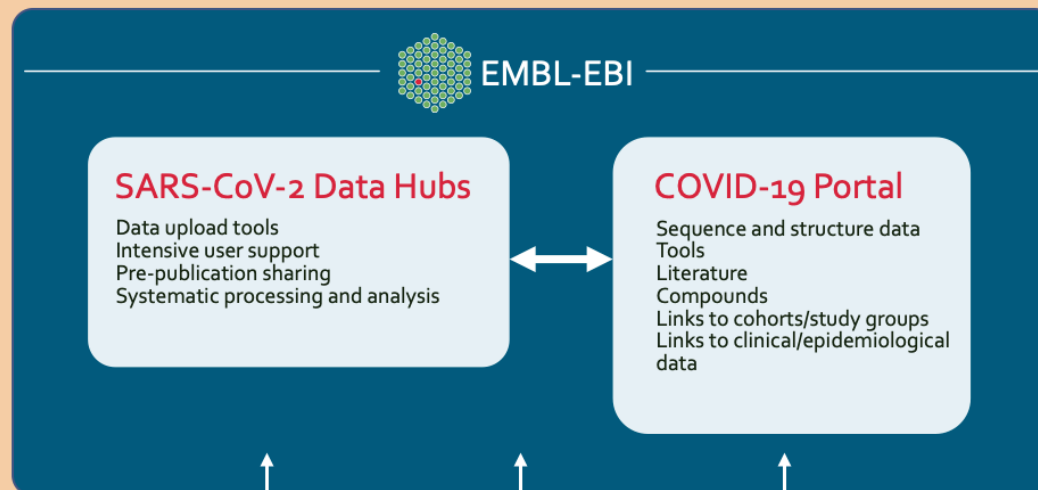
- ID: I-189639
- Type: Complete speech
- Date: 20/04/2020
- Location: Brussels - EC/Berlaymont
- Tag: [Research and development](#), [Medical treatment](#), [Public health](#), [Data Sharing](#), [Epidemic](#), [Crisis Management](#), [An economy that works for people](#), [<Political priority VDL>](#), [Coronavirus](#), [COVID-19](#)
- Personalities: [Ursula von der Leyen](#)
- Language: [Sign](#)
- Views: 39303

Below the video player, there is a text block: "On 20 April 2020, Ursula von der Leyen, President of the European Commission, recorded a video message at the European Commission in Brussels, Belgium, on the launch of the EU COVID19 Data Platform." and another text block: "On this platform, researchers will be able to store, exchange and analyse a wide range of knowledge about the".

<https://audiovisual.ec.europa.eu/en/video/I-189639>

<https://www.covid19dataportal.org/>

# European COVID-19 Data Platform



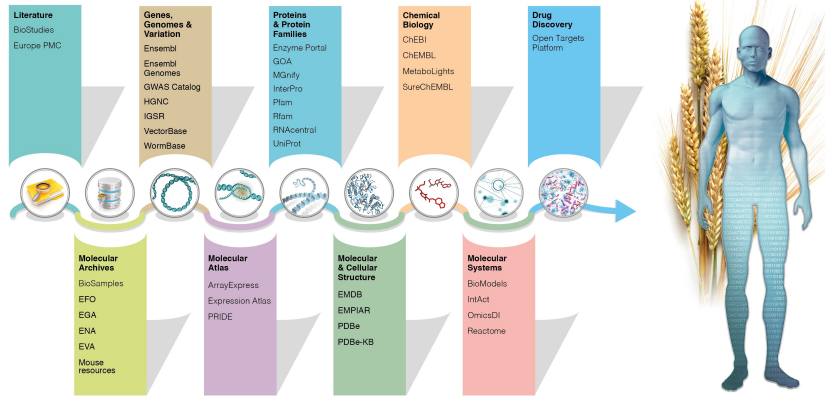
Clinical/epidemiological data where infrastructures do not exist

Sequence Data

Data links



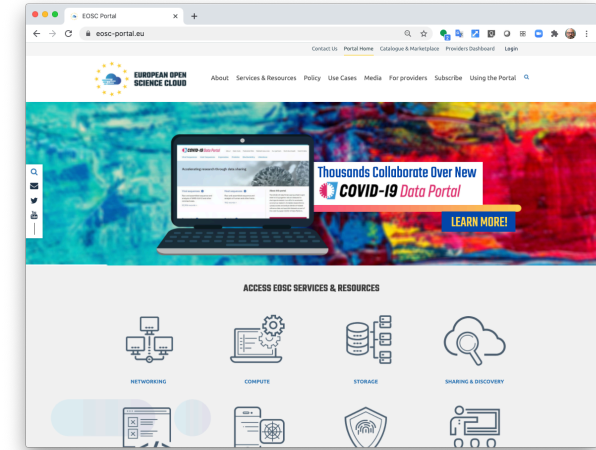
# Foundations



ELIXIR Observers



<https://elixir-europe.org/>



<https://www.eosc-portal.eu/>

EMBL-EBI  <https://www.ebi.ac.uk/>



Erasmus Medical Centre, the Netherlands



National Institute for Public Health and the Environment  
Ministry of Health, Welfare and Sport  
National Institute for Public Health and the Environment (RIVM), the Netherlands

DTU  
Technical University of Denmark (DTU)



Universitätsklinikum Heidelberg, Germany



Eötvös Loránd University, Hungary

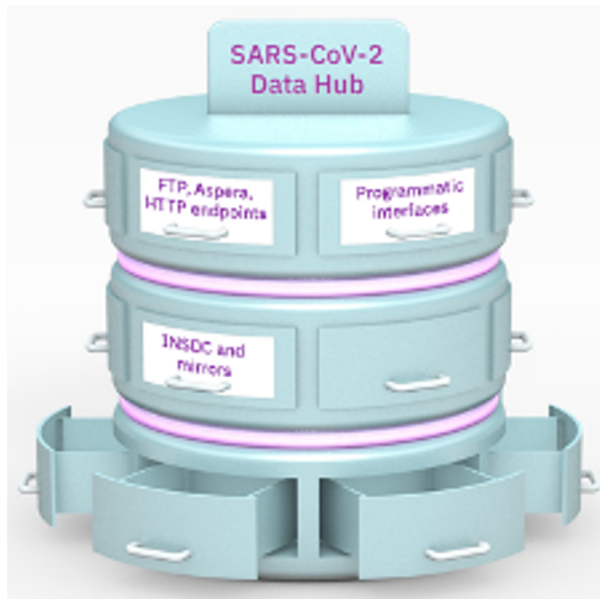
Amid *et al.* (2019) The COMPARE Data Hubs. *Database: the Journal of Biological Databases and Curation*, 01 Jan 2019, 2019 <http://doi.org/10.1093/database/baz136>

EMBL-EBI 

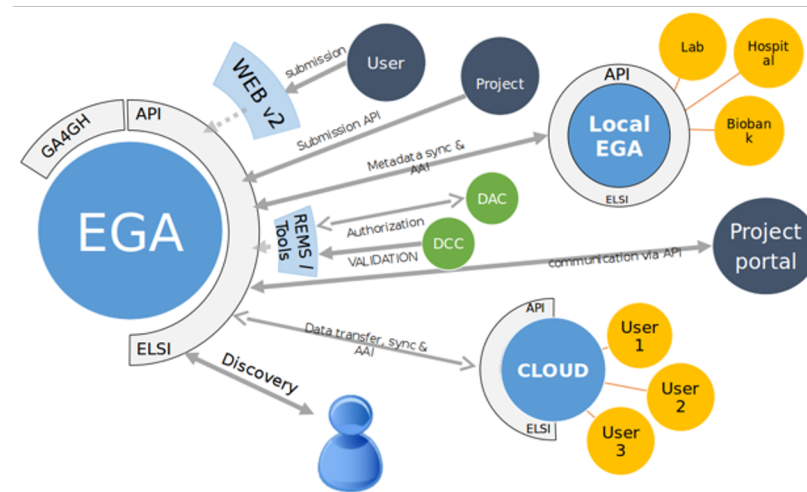
# European COVID-19 Data Platform

diverse data – rapid sharing – degrees of completeness – synergy and cross-fertilisation

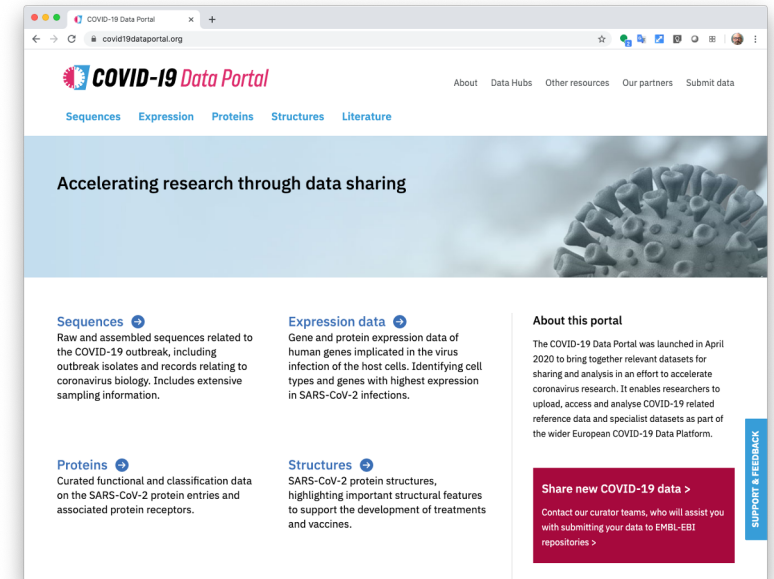
## SARS-CoV-2 Data Hubs



## Federated European Genome-phenome Archive



## COVID-19 Data Portal



<https://www.covid19dataportal.org/>

Priority I: data integration

# Data integration

## Viral Sequences

All (152,407)  
 Sequences (23,861)  
 Raw reads (65,323)  
 Sequenced samples (50,342)  
 Studies (167)  
 Genes (22)  
 Browser (1)  
 Variants (12,691)

## Host Sequences

All (973)  
 Human studies (controlled access) (3)  
 Human reads (consented for full access) (712)  
 Other species reads (250)  
 Association studies (8)

## Expression

All (69)  
 Gene expression (4)  
 Gene expression experiments (23)  
 Single cell expression (2)  
 Single cell expression experiments (14)  
 Protein expression experiments (26)

## Proteins

All (611)  
 Protein sequences (53)  
 Protein families (126)  
 Protein structures - Knowledge Base (7)  
 Protein structures (260)  
 Electron microscopy density maps (165)

## Biochemistry

All (1,853)  
 Pathways (16)  
 Interactions (1,415)  
 Complexes (24)  
 Compound document (8)  
 Drug targets (390)

## Literature

All (176,427)  
 Coronaviruses (90,722)  
 Diseases (79,381)  
 Related viruses and diseases (2,007)  
 Genes, receptors and antibodies (4,317)

### Organisms

- Mus musculus (154)
- Macaca mulatta (56)
- Chlorocebus aethiops aethiops (16)
- Macaca fascicularis (8)
- Homo sapiens (8)
- Chlorocebus sabaeus (6)
- Severe acute respiratory syndrome coronavirus 2 (2)

**Host sequences**  
 Raw and assembled sequence and analysis of human and other hosts

Search  **Search**

Examples: ACE2, Severe acute respiratory syndrome 2...

Showing 3 of 3 in Host sequences > Human studies (controlled access)

powered by

Accession	Name	Description
EGAS00001004412	Potent neutralizing antibodies against SARS-CoV-2	The COVID-19 pandemic urgently needs therapeutic and prophylactic interventions. Here we report the rapid identification of SARS-CoV-2 neutralizing antibodies by high-throughput single-cell RNA and VDJ sequencing of antigen-enriched B cells from 60 convalescent patients., EGA study EGAS00001004412
EGAS00001004419	SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells	The SARS-CoV-2 pandemic affecting the human respiratory system severely challenges public health and urgently demands for increasing our understanding of COVID-19 pathogenesis, especially host factors facilitating virus infection and replication. SARS-CoV-2 was reported to enter cells via binding to ACE2, followed by its priming by TMPRSS2. Here, w. .... EGA study EGAS00001004419
EGAS00001004481	Single-cell analysis of airway samples identifies immune	To investigate the immune response and mechanisms associated with severe COVID-19, we performed single-cell RNA-seq on nasopharyngeal and bronchial samples from 19 clinically well-characterized patients with moderate or critical

**Data types**  
 All (973)  
 Human studies (controlled access) (3)  
 Human reads (consented for full access) (712)  
 Other species reads (250)  
 Association studies (8)

**Organisms**  
 Homo sapiens (3)

**SUPPORT & FEEDBACK**

**Biochemistry**  
 COVID-19 pathways, interactions, complexes, targets and compounds

Search  **Search**

Examples: ACE2...

Showing 15 of 390 in Biochemistry > Drug targets

powered by

Evidence	Target	COVID-19 association score
Evidence for HMGR in COVID-19	HMGR	1.000
Evidence for TLR7 in COVID-19	TLR7	1.000
Evidence for CSF2RA in COVID-19	CSF2RA	1.000
Evidence for PTGS2 in COVID-19	PTGS2	1.000
Evidence for DPP4 in COVID-19	DPP4	1.000
Evidence for IFNAR2 in COVID-19	IFNAR2	1.000

**Data sources**  
 All (1,853)  
 Pathways (16)  
 Interactions (1,415)  
 Complexes (24)  
 Compound document (8)  
 Drug targets (390)

**SUPPORT & FEEDBACK**

## Data sources

All (138)  
Protein sequences (41)  
Protein families (97)

## Keywords

- Reference proteome (41)
- Membrane (30)
- 3D-structure (23)
- Transmembrane helix (22)
- Transmembrane (22)
- Host-virus interaction (21)
- Host membrane (20)
- Viral immunoevasion (12)
- Host Golgi apparatus (12)
- Phosphoprotein (11)

## Organisms

- Severe acute respiratory syndrome-related coronavirus (15)
- Severe acute respiratory syndrome coronavirus 2 (14)

Accession	Description	Organism	Gene
Q9BYF1	Angiotensin-converting enzyme 2	Homo sapiens	ACE2
Q10589	Bone marrow stromal antigen 2 BST-2	Homo sapiens	BST2
Q92499	ATP-dependent RNA helicase DDX1	Homo sapiens	DDX1
P09958	Furin	Homo sapiens	FURIN
P52292	Importin subunit alpha-1	Homo sapiens	KPNB1
P20701	Integrin alpha-L	Homo sapiens	ITGA1
Q8N3R9	MAGUK p55 subfamily member 5	Homo sapiens	MPP1
Q99623	Prohibitin-2	Homo sapiens	PHB2
P35232	Prohibitin	Homo sapiens	PHB1
O43765	Small glutamine-rich tetratricopeptide repeat-containing protein alpha	Homo sapiens	SGTA
P84022	Mothers against decapentaplegic homolog 3 MAD homolog 3 Mad3 Mothers against DPP homolog 3 hMAD-3	Homo sapiens	SMA3
O15393	Transmembrane protease serine 2	Homo sapiens	TMP22
P59632	ORF3a protein	Human SARS coronavirus	

Intuitive results views and navigation

Showing 15 of 2,147 in Nucleotide Sequences

Downloads

## Data types

All (25,148)  
Sequences (2,147)  
Assemblies (101)  
Raw reads (15,373)  
Sequenced samples (7,490)  
Studies (37)

<input type="checkbox"/>	Accession	Collection
<input type="checkbox"/>	MT081067	Jan 31, 2020
<input type="checkbox"/>	MN938384	Jan 10, 2020
<input type="checkbox"/>	MN982287	Jan 10, 2020

<https://www.covid19dataportal.org/>

Showing 12 of 83,350 in Literature

## Sections

All (83,350)  
Corona viruses (42,835)  
Diseases (35,711)  
Related viruses and diseases (1,662)  
Genes, receptors and antibodies (3,142)

Corona viruses 42,835 results

Emerging novel coronavirus (2019-nCoV)-current scenario, evolutionary perspective based on genome analysis and recent developments.

Malik YS Sircar S, Bhat S, Sharun K, Dhama K, Dadar M, Tiwari R, Chaicumpa W  
The veterinary quarterly 40 (1) 68-76 Dec 1, 2020  
PMC7054940

Hypothesis for potential pathogenesis of SARS-CoV-2 infection-a review of immune changes in patients with viral pneumonia.

Lin L, Lu L, Cao W, Li T  
Emerging microbes & infections 9 (1) 727-732 Dec 1, 2020  
PMC7170333

Different longitudinal patterns of nucleic acid and serology testing results based on disease severity of COVID-19 patients.

Yongchen Z, Shen H, Wang X, Shi X, Li Y, Yan J, Chen Y, Gu B  
Emerging microbes & infections 9 (1) 833-836 Dec 1, 2020  
PMC7241531

View all 42,835 results in Corona viruses

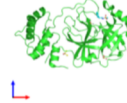
Literature

## Data sources

All (162)  
Protein structures (137)  
Electron microscopy density maps (25)

PanDDA analysis group deposition -- Crystal Structure of COVID-19 main protease in complex with

Accession: 5r7y  
HYDROLASE/HYDROLASE INHIBITOR x-ray diffraction entry at resolution 1.6



Rendered structures

PanDDA analysis group deposition -- Crystal Structure of SARS-CoV-2 main protease in complex with

Accession: 5r80  
HYDROLASE/HYDROLASE INHIBITOR x-ray diffraction entry at resolution 1.9

## European Virus Bioinformatics Center Coronavirus tools

Source: European Virus Bioinformatics Center

The European Virus Bioinformatics Center is curating a list of bioinformatics tools designed explicitly for SARS-CoV-2 and coronaviruses, covering the routine detection of SARS-CoV-2 infection, the reliable analysis of sequencing data, the tracking of the COVID-19 pandemic and evaluation of containment measures, the study of coronavirus evolution, the discovery of potential drug targets and development of therapeutic strategies.

## OpenAire for COVID-19

Source: OpenAire

OpenAire aims to aggregate COVID-19 records (publications-data-software-other research outcomes) and link them together to provide a single access point for their discovery and navigation.

## data-against-covid

Source: data-against-covid

data-against-covid is a community of volunteer data scientists and bioinformaticians who offer their assistance to the wider scientific community for better understanding of COVID-19/SARS-CoV-2 data analysis.

## COVID-19 Workflows Hub

Source: The University of Manchester

Use COVID-19 Workflows Hub to find cheminformatics workflows such as the virtual screening of the SARS-CoV-2 main protease

COVID-19 Data Portal

About Data Hubs Related resources Our partners Submit data

Sequences Expression Proteins Structures Compounds Targets Literature

## Related resources

### Resource type

All (49)  
Databases and atlases (22)  
Source: PDBs  
Computing support (7)  
Standards for data sharing (6)  
Eliver publications (5)  
Eliver activity and events (4)  
Other European projects (3)

Databases and atlases 22 results

Molecular structures of SARS-CoV-2

Access all the molecular structure data including all observed ligand binding sites and protein-protein interaction residues. The resource allows to easily identify important structural features to support the development of treatments and vaccines.

Model of the full SARS-CoV-2 proteome

Source: SWISS-MODEL  
SWISS-MODEL, the protein structure homology-modeling server and repository provides modeling of the complete SARS-CoV-2 proteome.

Coronavirus Phytomes

Source: PhytomeDB

A full phylogenomic analysis of 60 coronavirus genomes, including SARS-CoV-2, SARS and MERS. Browse and download gene phylogenetic trees and multiple sequence alignments.

View all 22 results in Databases and atlases

Computing support 7 results

LEOSS

Source: LEOSS

Mandated by the ESCMID Emerging Infections Task Force (EITaF) and supported by the German Infectious Disease Society, LEOSS offers systematic documentation of patient data in an effort to better understand the implications of the new virus on patients.

IFB

Source: IFB

IFB (ELIXIR France) is providing a federated set of high performance compute and cloud resources including national and regional servers.

SIB

Source: SIB

SIB (ELIXIR Switzerland) is providing a ready-to-use slurm workload manager with a scientific software stack via the ExPaSy SIB Portal.

View all 7 results in Computing support

Standards for data sharing 6 results

European Virus Bioinformatics Center Coronavirus tools

Source: European Virus Bioinformatics Center

The European Virus Bioinformatics Center is curating a list of bioinformatics tools designed explicitly for SARS-CoV-2 and coronaviruses, covering the routine detection of SARS-CoV-2 infection, the reliable analysis of sequencing data, the tracking of the COVID-19 pandemic and evaluation of containment measures, the study of coronavirus evolution, the discovery of potential drug targets and development of the specific strategies.

OpenAire for COVID-19

Source: OpenAire

OpenAire aims to aggregate COVID-19 records (publications-data-software-other research outcomes) and link them together to provide a single access point for their discovery and navigation.

data-against-covid

Source: data-against-covid

data-against-covid is a community of volunteer data scientists and bioinformaticians who offer their assistance to the wider scientific community for better understanding of COVID-19/SARS-CoV-2 data analysis.

View all 6 results in Standards for data sharing

Eliver publications 5 results

Impact metrics and Indicators

Nergulis T, Kanellos I, Chatzopoulos S, Karidi DP, Dalamagas T (2020). SIP4COVID19: Impact metrics and indicators for coronavirus related publications (Version 0.1) [Data set] Zenodo. doi: <https://doi.org/10.5281/zenodo.3723282>

SARS-CoV-2 envelope (E) protein

Nawati V, Lionard L, Longhi S, Combet C, Avouacher A (2020). The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) envelope (E) protein harbors a conserved EHE-like motif. bioRxiv 2020.04.09.033522. <https://doi.org/10.1101/2020.04.09.033522>

COVID-19 Disease Map

Osabawami, M, Mazini, A, Gillespie, ME et al. (2020). COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Sci Data 7, 136. <https://doi.org/10.1038/s41597-020-0477-8>

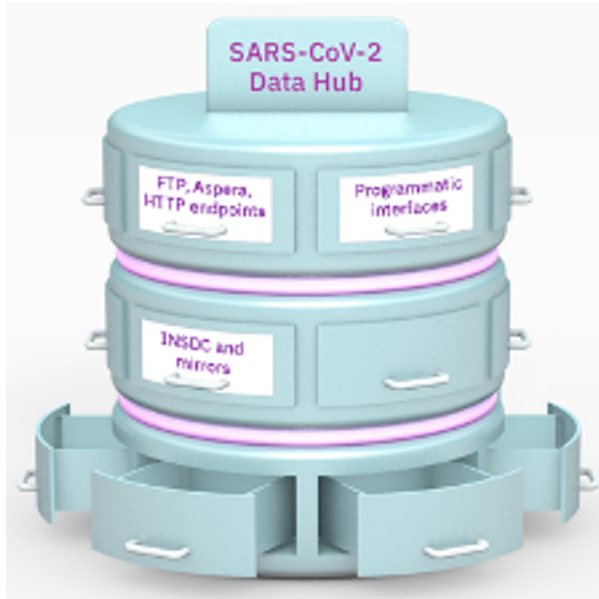
View all 5 results in Eliver publications

Eliver activity and events 4 results



## Priority II: viral data mobilisation

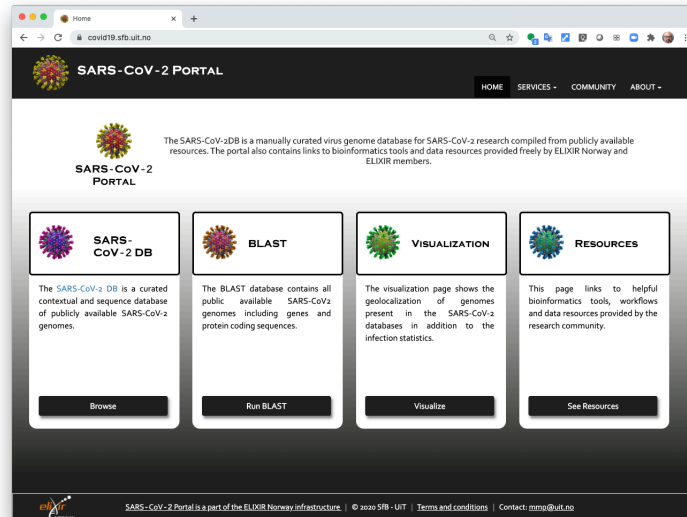
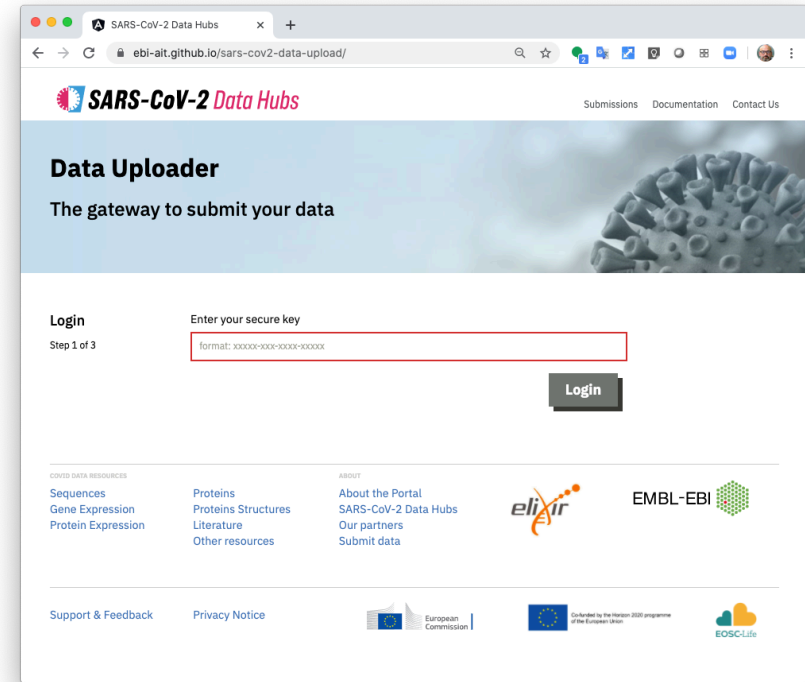
# Viral data mobilization



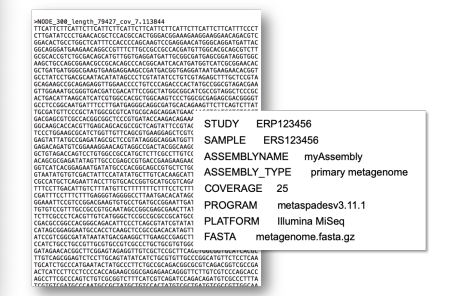
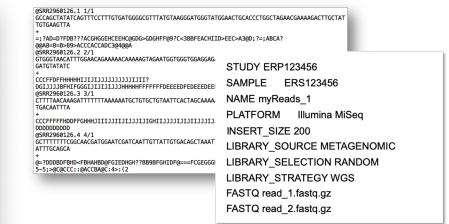
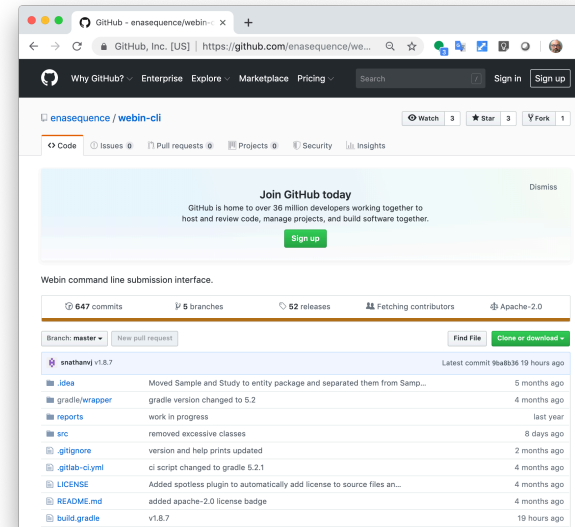
- Raw viral sequence data key to tractable viral variation calling
- Viral variation informs the study of the biology, transmission, spread of the virus
- Required to inform interventions, treatments, vaccine design, etc.
- Ambition to mobilise raw sequence data at scale to support systematic analysis and intercomparability

# Uploader and user support effort

- Tools and services to support data upload
  - SARS-CoV-2 Uploader
  - Existing programmatic and interactive tools
- Extensive support desk
- European campaign
- Standards and compliance support
- Curation, harmonization and update cycles

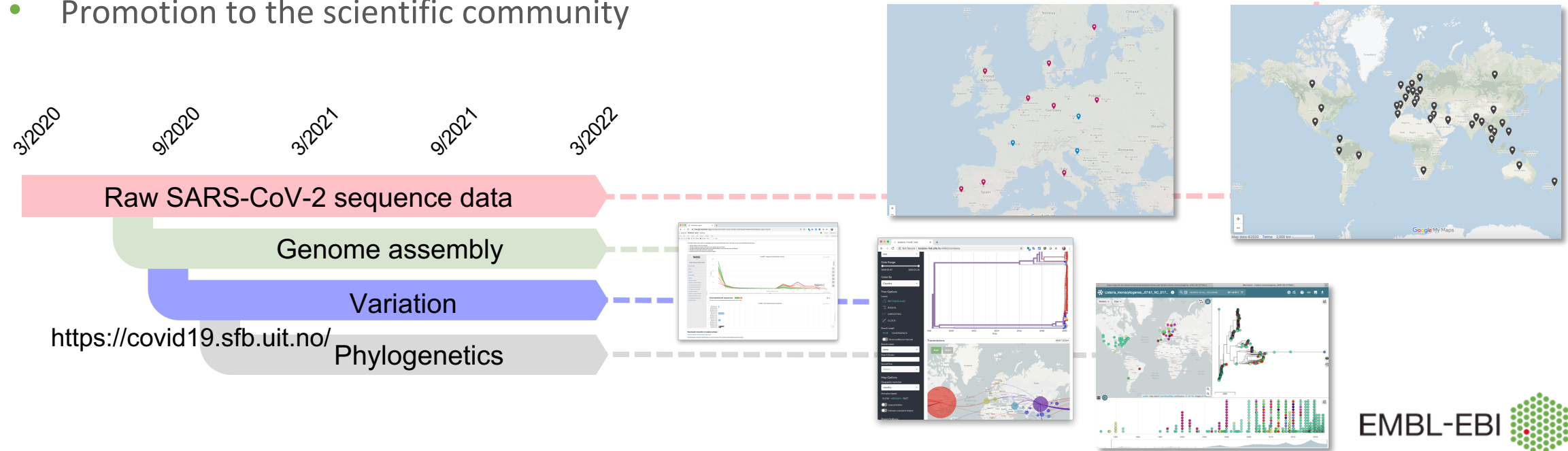
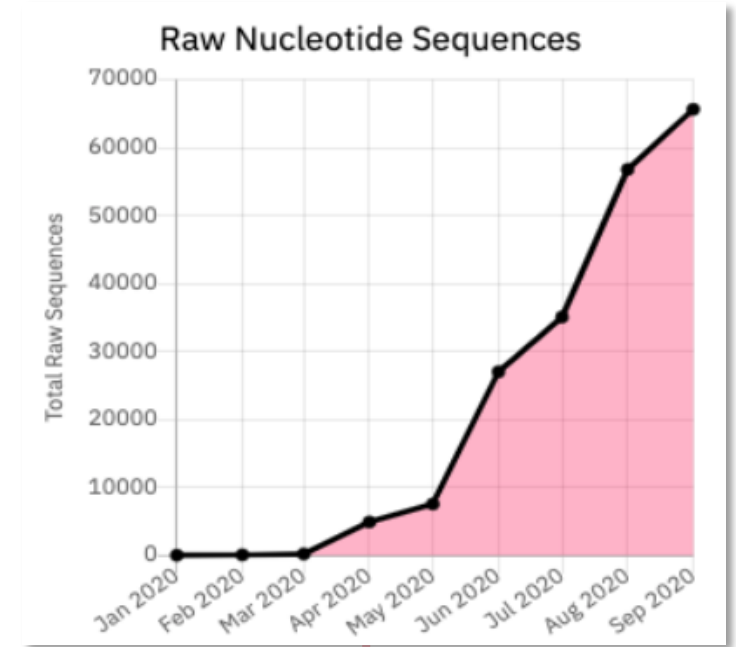


<https://covid19.sfb.uit.no/>



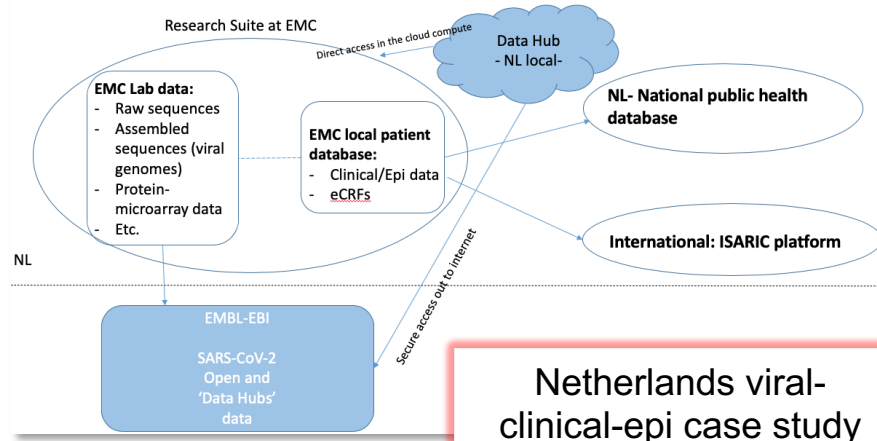
# Status at 6-month point

- 65,000 sequenced isolates from 38 countries and 300 institutions
- 16 national SARS-CoV-2 Data Hubs supporting 70% of data
- Curated metadata
  - primer sets
  - sample descriptions with ELIXIR-Norway
- 3 computational workflows (from Erasmus MC and RIVM)
- Assembly product currently in QC, soon to be published
- Promotion to the scientific community



# Further work streams

## COVID-19 Data Case Study with Netherlands (EMC)

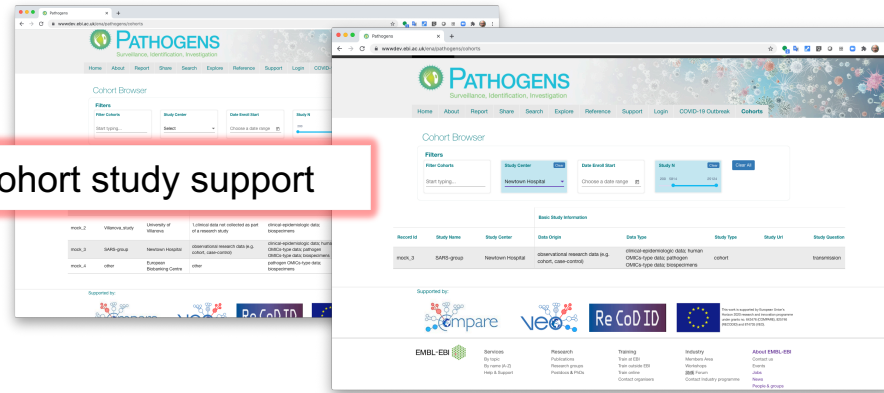


Netherlands viral-clinical-epi case study

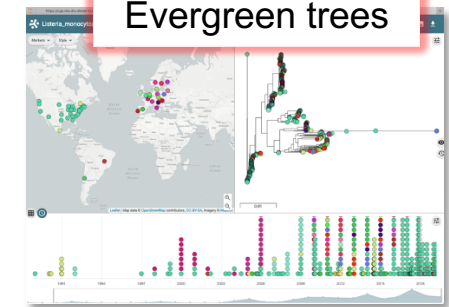
Serology data standards

- Exploratory work
- Data standards
- Minimal reporting requirements
- Deposition database
- Support for
  - interoperability across methods
  - comparator studies
  - engaged communities

Cohort study support

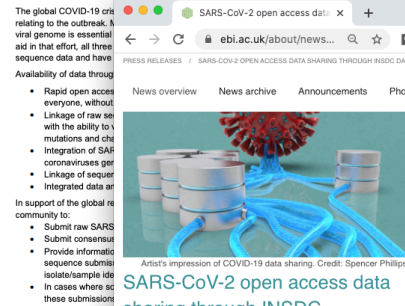


Evergreen trees



## INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19

Guy Cochrane (EMBL-EBI), Sina Karsch-Mizrachi (NCBI-NLM/NIH) and Masaroni Anita (DDBJ) on behalf of the International Nucleotide Sequence Database Collaboration (INSDC):  
 The databases of the International Nucleotide Sequence Database Collaboration (INSDC; <http://www.insdc.org>) capture, organise, preserve and present nucleotide sequence data as part of the open scientific record. INSDC member institutions – the EMBL European Bioinformatics Institute (EMBL-EBI), the NIG DNA Data Bank of Japan (NIG-DBJ) and the National Library of Medicine's National Center for Biotechnology Information at NIH (NCBI) – are committed to the continued delivery of this critical element of scientific infrastructure.



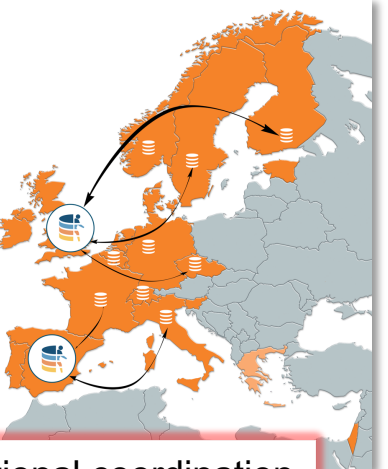
Promotion of FAIR practice

<https://www.ebi.ac.uk/about/news/announcements/sars-cov-2-open-access-data-sharing-through-insdc-databases>



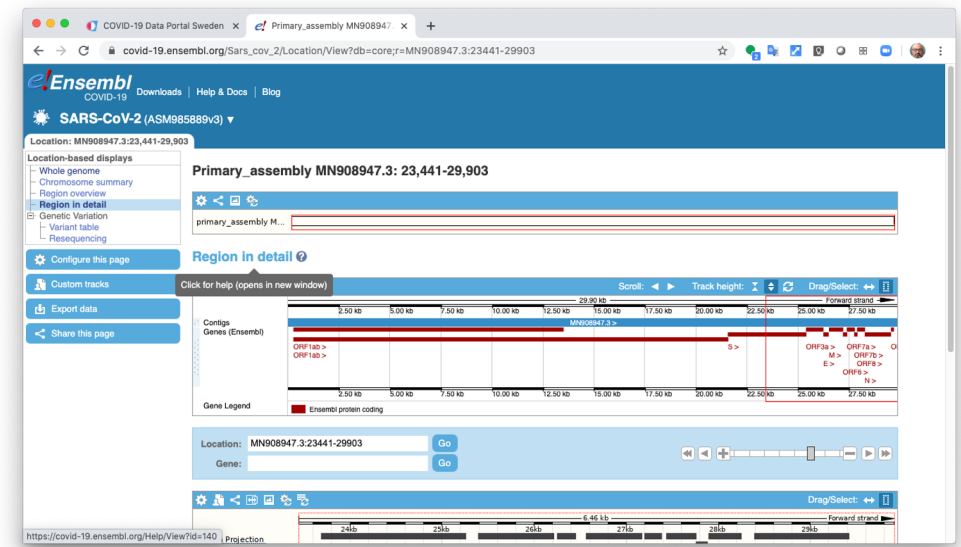
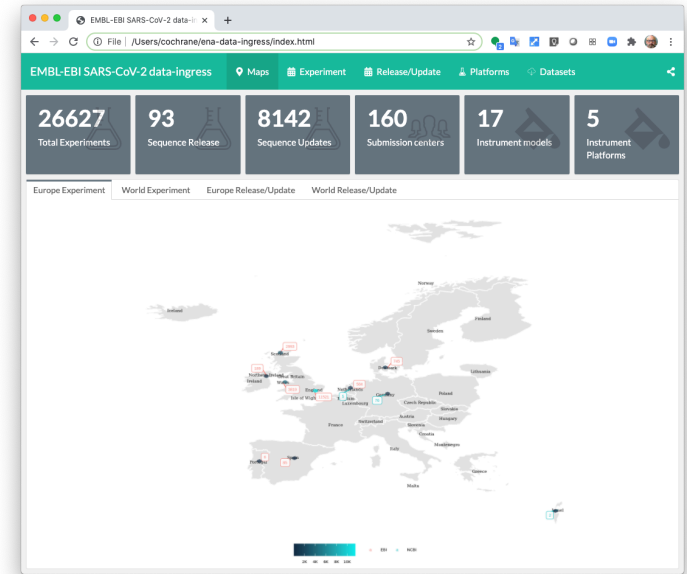
National coordination network & FEAGA

<https://www.covid19dataportal.se/>



# Future

- Broader and richer data
- Further tools for data exploration, analysis and visualization
- Phylogenetic integration
- Discovery of cohorts and their data
- Connected viral-human data sets



# Engagement

- Access data and services via the COVID-19 Data Platform
- Access tools and support for viral data management, analysis and sharing
- Share cohort information and connected data sets
- For national and regional coordinators
  - Explore options for national FEGA activities
  - Join our international stakeholder group
- Contact [cochrane@ebi.ac.uk](mailto:cochrane@ebi.ac.uk)

The screenshot shows the COVID-19 Data Portal website. The header includes the site name and navigation links: About, Data Hubs, Related resources, Our partners, and Submit data. Below the header is a navigation menu with categories: Sequences, Expression, Proteins, Structures, Compounds, Targets, and Literature. The main content area features a banner with the text 'Accelerating research through data sharing' and a background image of a virus particle. Below the banner are several data category cards, each with a description and a record count:

- Sequences**: Raw and assembled sequences related to the COVID-19 outbreak, including outbreak isolates and records relating to coronavirus biology. Includes extensive sampling information. 22,739 records >
- Expression data**: Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections. 51 records >
- Proteins**: Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors. 138 records >
- Structures**: SARS-CoV-2 protein structures, highlighting important structural features to support the development of treatments and vaccines. 229 records >
- Compounds**: Curated range of SARS-CoV-2 compound activity data. 8 records >
- Targets**: Range of SARS-CoV-2 and SARS-CoV data for drug target identification and prioritisation. 24 records >
- Literature**: Search for the latest literature about SARS-CoV-2. 83,945 publications >
- Related resources**: A range of related resources for studying the SARS-CoV-2 coronavirus and the COVID-19 disease.

On the right side, there is a section titled 'About this portal' with a description of the portal's mission and a red button labeled 'Share new COVID-19 data >'. Below this, there is a 'Support & Feedback' button. At the bottom of the page, there are links for 'Support & Feedback' and 'Privacy Notice', along with logos for the European Commission, the European Union, and EMBL-EBI.

<https://www.covid19dataportal.org/>