The European COVID-19 Data Platform

Nadim Rahman
European Nucleotide Archive (ENA), EMBL-EBI

rahman@ebi.ac.uk

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

The question is not *if* the next pandemic will happen but *when*

Recent outbreaks:

- 2002-2004 SARS (Severe Acute Respiratory Syndrome)
- 2009 Swine Flu
- 2011 Germany E. coli O104:H4
- 2013-2016 Western African Ebola virus
- 2015-2016 Zika virus

BY-COVID is developing the infrastructure to deal with the next pathogen.

Funded under HORIZON-INFRA-2021-EMERGENCY-01: FAIR and open data sharing in support to European preparedness for COVID-19 and other infectious diseases

TIME magazine, May 2017

MAY 15, 2017
TIME
WARNING:
WE ARE NOT READY FOR THE NEXT PANDEMIC
SCIENCE KNOWS HOW TO FIGHT AN OUTBREAK— BUT POLICY STILL GETS IN THE WAY BY BRYAN WALSH
HOW TO KEEP THE WORLD SAFE BY BILL GATES

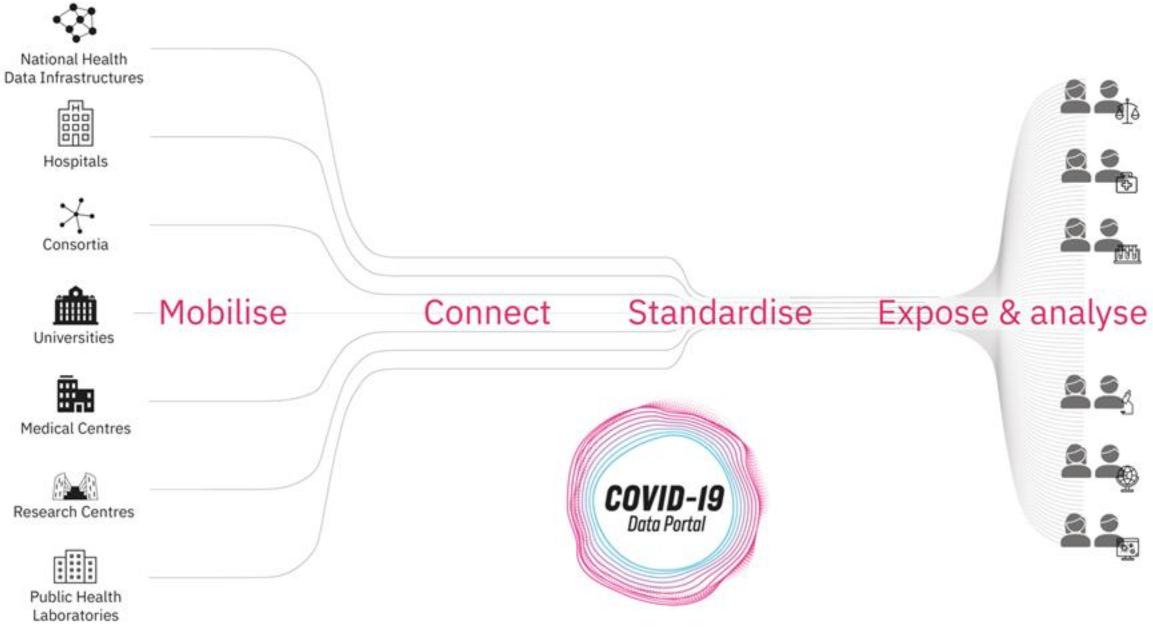
BY-COVID in a nutshell

BY-COVID
beyond covid

Total budget	€12 million
Organisations	53
Countries	19
Duration	36 months
Coordination	elixir

European COVID-19 Data Platform

Obj 3: Establish a sustainable and federated infrastructure enabling open sharing of scientific results



BY-COVID is one of the Horizon Europe projects that supports the operation of the **European COVID-19 Data Platform**

The **COVID-19 Data Portal contains >14 million viral sequences and >1million open scientific articles related to COVID-19.** Since its launch in April 2020, the Portal received 8.8M visits from 500k unique visitors by April 2023.

Raw data from **105 countries and regions, 216 unique centres.**

Background

- Sit on top of existing infrastructure at EMBL-EBI
- Includes 3 main components:

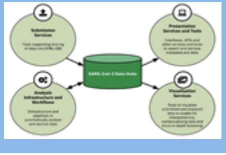
COVID-19 Data Portal

Interface for COVID-19 life sciences data

Publication DOI: 10.1093/nar/gkab417

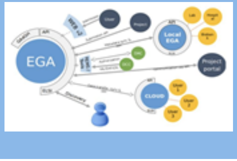
SARS-CoV-2 Data Hubs

Tools to support submission, analysis, visualisation and presentation of COVID-19 sequence data in the COVID-19 Data Portal*


Submission Services
Presentation Services and Tools
SARS-CoV-2 Data Hubs
Analysis Infrastructure and Workflows
Visualisation Services

Federated European Genome-phenome Archive (FEGA)

Controlled-access sharing of human COVID-19 biomolecular and phenotypic data, to present in the COVID-19 Data Portal



* **Preprint**: https://www.biorxiv.org/content/10.1101/2023.04.19.537514v2

European COVID-19 Data Platform

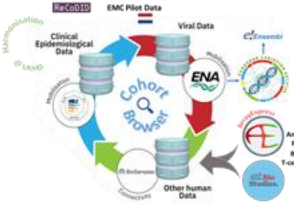
Access

- 25M records across molecular, literature, imaging and social science, backed up by a network of 11 national Data Portals

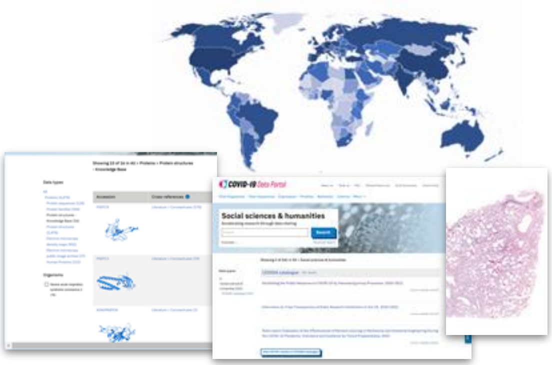
Data Hubs

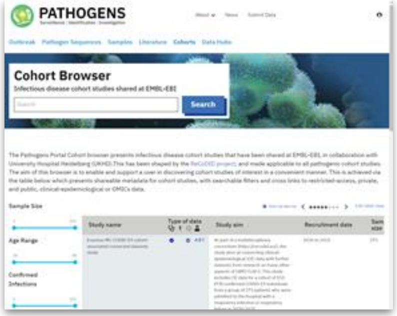
- Mobilisation (>6M) and systematic analysis (>4M) data sets from SARS-CoV-2 isolates from 121 countries
- 15 VEO public health reports

Human data



- Demonstrated protocols to link between sensitive research subject and pathogen data, leveraging (federated) European infrastructure





Cohort Browser

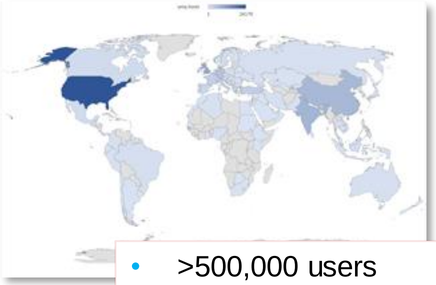
Infectious disease cohort studies shared at EMBL-EBI

The Pathogens Portal Cohort browser presents infectious disease cohort studies that have been shared at EMBL-EBI, in collaboration with University Hospital Heidelberg (UKHD). This has been shaped by the ReCoDID project, and made applicable to all pathogenic cohort studies. The aim of this browser is to enable and support a user in discovering cohort studies of interest in a convenient manner. This is achieved via the table below which presents shareable metadata for cohort studies, with searchable filters and cross links to restricted-access, private, and public, clinical-epidemiological or OMICs data.

Study name	Type of data	Study aim	Recruitment date	Sam size

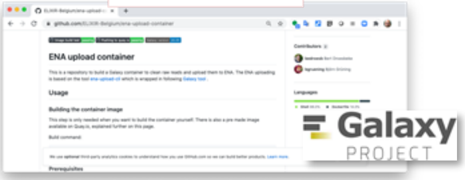
Usage

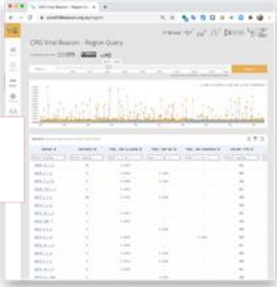
Open Targets



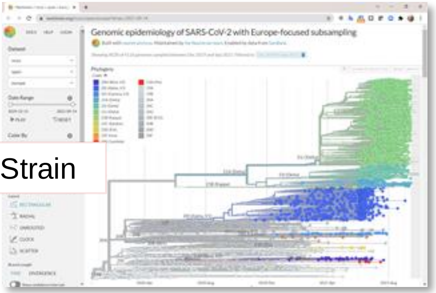
- >500,000 users
- 8.8 million requests

Cloud workflows

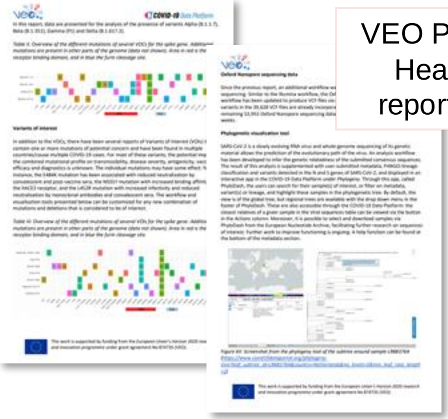




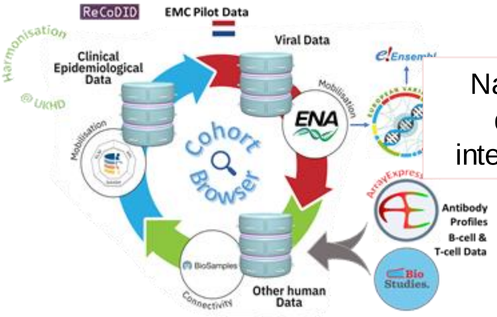
CRG Viral Beacon


Genomic epidemiology of SARS-CoV-2 with Europe-focused subsampling

NextStrain


VEO
COVID-19 Data Platform
Variants of interest
Oxford Nanopore sequencing data
Phylogenetic visualization tool

VEO Public Health reporting

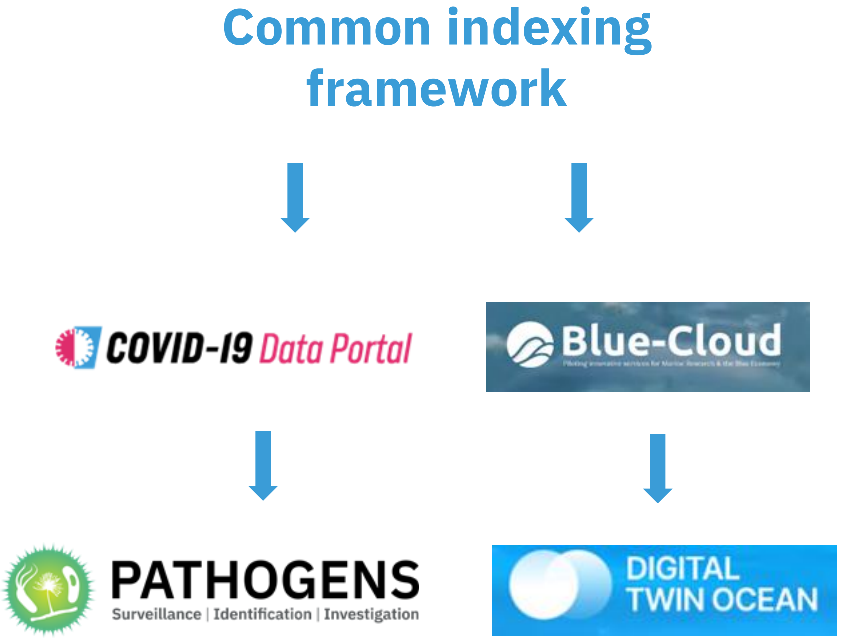


National data integration

Sustainability by design

- Molecular biology components built on existing ELIXIR data resources
    - Resources with institutional commitment and history of sustained activity
    - Typically globally connected through data exchange collaborations
    - Long-term sustainability: Global Biodata Coalition
- Open standards and software
- High level of distribution of expertise and effort

Infrastructure already shared across projects


Common indexing framework
COVID-19 Data Portal
Blue-Cloud
PATHOGENS
Surveillance | Identification | Investigation
DIGITAL TWIN OCEAN

- Common indexing framework, supporting
  - Metadata indexing with three tiers of granularity
  - Domain-level classification system to define data "partitions", e.g. "blue domain"
  - Coverage spans ELIXIR Deposition and Core Data Resources and broadly connected via FAIRSharing
  - Coverage goes beyond ELIXIR including for image and social science data with further data resources being added
- Continued development work under both BY-COVID and Blue-Cloud26
- Supports metadata feeds to future initiatives in both domains

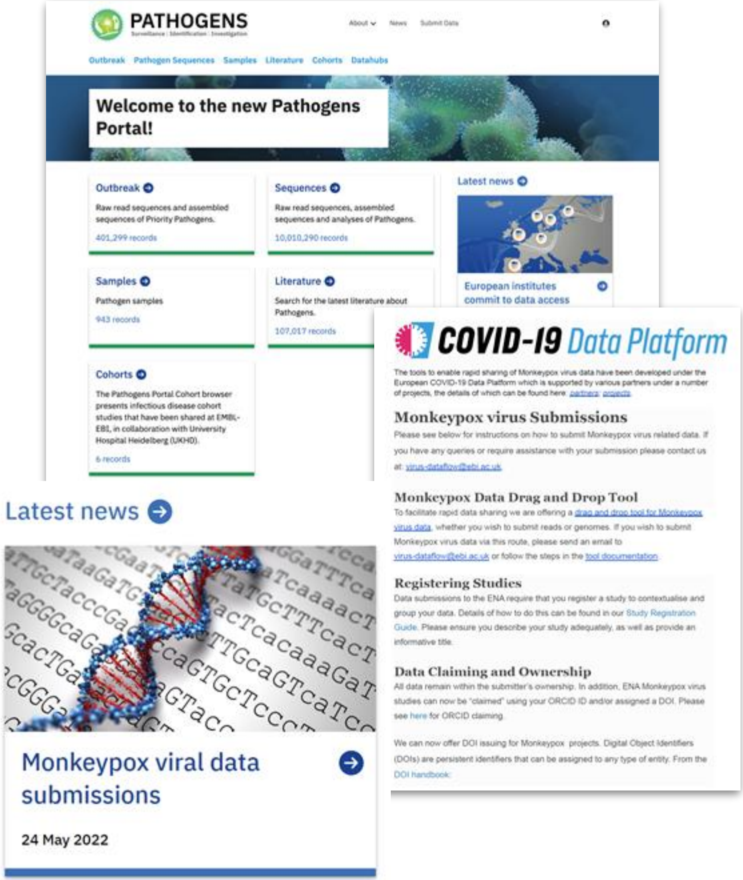
BY-COVID

Blue-Cloud2026

A broader Pathogens Platform

- Scope
  - All pathogen, all disease approach
  - Hosts, vectors and pathogens
  - Antimicrobial resistance
- Preparedness and Outbreaks
  - Pathogens Portal
  - Pathogen Data Hubs - responding to outbreaks as they arise
- Applicability
  - Climate change → disease outbreaks
  - Cancer associated viral/bacterial infection
- Future roles
  - Food security, e.g. plant pathogens
  - Biodiversity loss

https://www.pathogensportal.org/



Welcome to the new Pathogens Portal!

Outbreak

Raw read sequences and assembled sequences of Priority Pathogens.

401,299 records

Sequences

Raw read sequences, assembled sequences and analyses of Pathogens.

10,010,290 records

Samples

Pathogen samples

943 records

Literature

Search for the latest literature about Pathogens.

107,017 records

Cohorts

The Pathogens Portal Cohort browser presents infectious disease cohort studies that have been shared at EMBL-EBI, in collaboration with University Hospital Heidelberg (UKHD).

6 records

Latest news

European institutes commit to data access

Latest news

Monkeypox viral data submissions

24 May 2022

COVID-19 Data Platform

The tools to enable rapid sharing of Monkeypox virus data have been developed under the European COVID-19 Data Platform which is supported by various partners under a number of projects, the details of which can be found here: partners; projects.

Monkeypox virus Submissions

Please see below for instructions on how to submit Monkeypox virus related data. If you have any queries or require assistance with your submission please contact us at: virus-dataflow@ebi.ac.uk.

Monkeypox Data Drag and Drop Tool

To facilitate rapid data sharing we are offering a drag and drop tool for Monkeypox virus data, whether you wish to submit reads or genomes. If you wish to submit Monkeypox virus data via this route, please send an email to virus-dataflow@ebi.ac.uk or follow the steps in the tool documentation.

Registering Studies

Data submissions to the ENA require that you register a study to contextualise and group your data. Details of how to do this can be found in our Study Registration Guide. Please ensure you describe your study adequately, as well as provide an informative title.

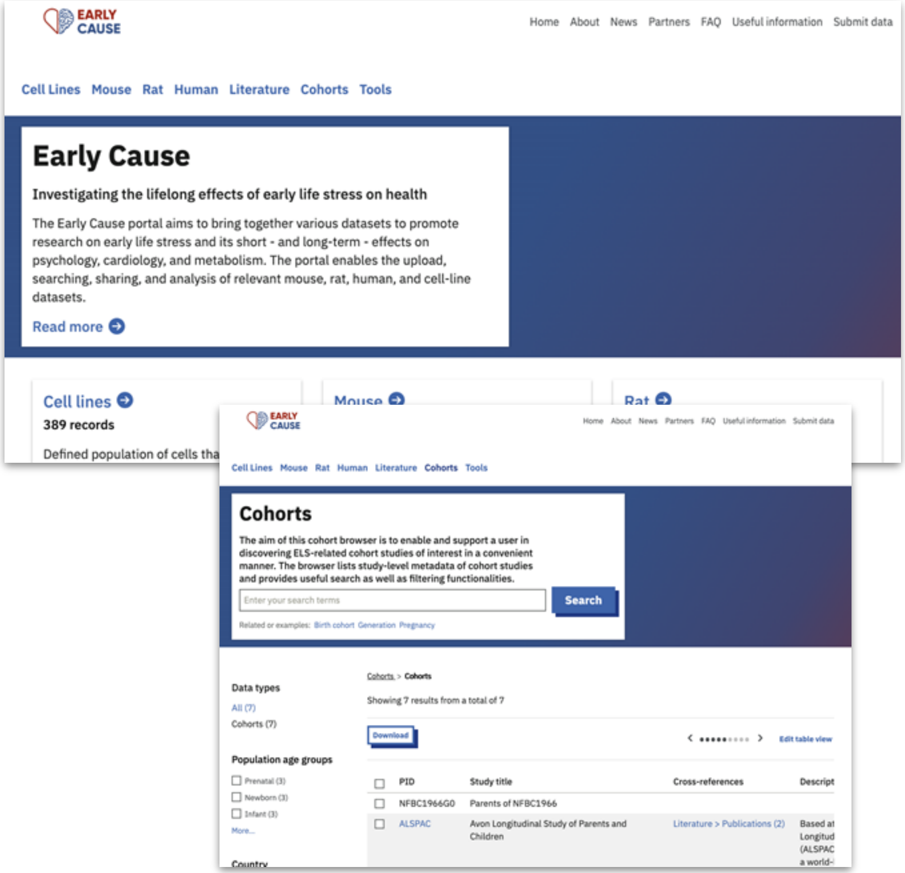
Data Claiming and Ownership

All data remain within the submitter's ownership. In addition, ENA Monkeypox virus studies can now be "claimed" using your ORCID ID and/or assigned a DOI. Please see here for ORCID claiming.

We can now offer DOI issuing for Monkeypox projects. Digital Object Identifiers (DOIs) are persistent identifiers that can be assigned to any type of entity. From the DOI handbook:

EarlyCause - early life stress

- Data portal, search functionalities
- Biodata to support investigations into lifelong effects of early-life stress
- https://portal.earlycause.eu/
- Data on organism level or data type (Mouse, Rat, Human, Cell Lines, Literature, Cohorts)
- Reusable infrastructure and framework to bring forward biodata (e.g. soil biodata)



Early Cause

Investigating the lifelong effects of early life stress on health

The Early Cause portal aims to bring together various datasets to promote research on early life stress and its short - and long-term - effects on psychology, cardiology, and metabolism. The portal enables the upload, searching, sharing, and analysis of relevant mouse, rat, human, and cell-line datasets.

Read more

Cell lines
389 records
Defined population of cells tha

Mouse

Rat

Cohorts

The aim of this cohort browser is to enable and support a user in discovering ELS-related cohort studies of interest in a convenient manner. The browser lists study-level metadata of cohort studies and provides useful search as well as filtering functionalities.

Enter your search terms | Search

Related or examples: Birth cohort Generation Pregnancy

Data types
All (7)
Cohorts (7)

Population age groups
□ Prenatal (3)
□ Newborn (3)
□ Infant (3)
More...

Country

Cohorts > Cohorts

Showing 7 results from a total of 7

Download

Edit table view

□	PID	Study title	Cross-references	Descript
□	NFBC1966G0	Parents of NFBC1966		
□	ALSPAC	Avon Longitudinal Study of Parents and Children	Literature > Publications (2)	Based at Longitud (ALSPAC a world-

Thanks!

EMBL-EBI

Ahmad Zyoud
Alexey Sokolov
Amonida Zadissa
Andrew Parton
Andrii Ludin
Andy Yates
Carla Cummins
Claire O'Donovan
Claire Rye
Colman O'Cathail
Craig Russell
Dipayan Gupta
Dylan Spalding
Eloise Stapleton
Gabi Rinck
Galabina Yordanova
Geetika Malhotra
Giselle Kerry
Guy Cochrane
Helen Parkinson
Henning Hermjakob
Jeena Rajan
Jeff Knaggs
Joseph Rosetto
Josephine Burgin
Karoly Erdos
Laura Harris
Mallory Freeberg
Manish Kumar
Marianna Ventouratou
Matt Pearce
Melanie Courtot
Mihai Glont
Milena Mansurova
Nadim Rahman
Nicola Buso
Oana Stroe
Ossama Edbali
Pablo Moreno
Peter Harrison
Peter Walter
Raheela Aslam
Rasko Leinonen
Rodica Petrusevschi
Rodrigo Lopez
Rolf Apweiler
Ross Thorne
Sam Holt
Sandeep Kadam
Sandeep Selvakumar
Sarah Hunt
Senthil Vijayaraja
Simon Kay
Stefan Gutnick Allen
Suran Jayathilaka
Thomas Keane
Timothee Cezard
Tony Burdett
Tracey Mahoney
Vishnu Kadhirvelu
Youngmi Park
Zahra Waheed
Zamin Iqbal

Independent
Robert Petryszak

ELIXIR
Katharina Lauer
Niklas Blomberg

DTU
Jose Luis Bellod Cisneros
Martin Christen Frølund Thomsen
Johanne Ahrenfeldt
Rolf Sommer Kaas
Lukasz Dariusz Dynowski
Frank Aarestrup
Jeffrey Skiby
Judit Szarvas
Camilla Hundahl Johnsen
Rene S. Hendriksen
Martin Koliba
Philip Clausen
Máté Gulyás

ELTE
János Márk Szalai-Gindl
Balint Pataki
Jozsef Steger
Dávid Visontai
Krisztian Papp
Istvan Csabai
Ágnes Becsei
Ákos Gellért
Anikó Mentes
Orsolya Pipek

EMC
Marion Koopmans
Clara Amid
David van de Vijver
Marjolein Poen
Miranda de Graaf
Maarten Hoek
David Nieuwenhuijse
Divyae Prasad
Marie-Catherine Bouquieaux

FLI
Dirk Hoeper
Ariane Belka
Maria Jenckel
Claudia Wylezich
Martin Beer
Anne Pohlmann

RIVM
Dennis Schmitz
Florian Zwagemaker
Annelies Kroneman

Thanks for listening!
Any questions?

Nadim Rahman
rahman@ebi.ac.uk