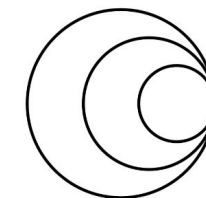


# Module 4: Data Sharing and Interpretation

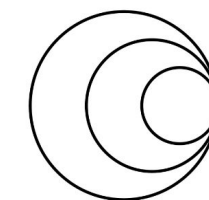
Instructors for Module  
Zahra Waheed, Marcela Suárez-Esquivel, Una Ren



# Week 4 - Greater Meaning and context

## or now I have a genome what to do next?

- Summary of what are being covered
  - Linkage and phylogenetic analysis:
    - Theory and basic concepts
    - Construct your own tree
  - Data interpretation
    - Outbreak investigations
    - Limitations
    - Phylodynamics
    - Visualisation (microreact, nextstrain)
  - Data sharing, introduction to GISAID and ENA



**wellcome**  
**connecting**  
**science**



**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Why genomic surveillance?

- **Additional & independent line of evidence**

- Outbreak investigation
- Effectiveness of mitigation strategies
- Source attribution

- **Understanding disease dynamics**

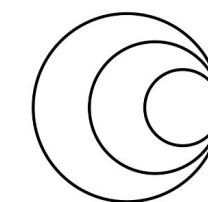
- Introduction: Where / how often?
- Transmission: How fast is it spreading? And how is it spreading?

- **Understanding Diversity**

- Inform vaccine, diagnostic and drug susceptibility changes.
- What is in the environment?
- Which ones are expanding / causing problems?



The Global Genomic Surveillance Strategy  
for Pathogens with Pandemic and Epidemic Potential



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# SARS-CoV-2 pandemic in the genomic era

- **Global effort**

215 countries and territories shared 13,290,083 viral genome sequences from human cases of COVID-19 via GISAID since 10 January 2020. (GISAID, 14/10/2022)

- **Open Science**

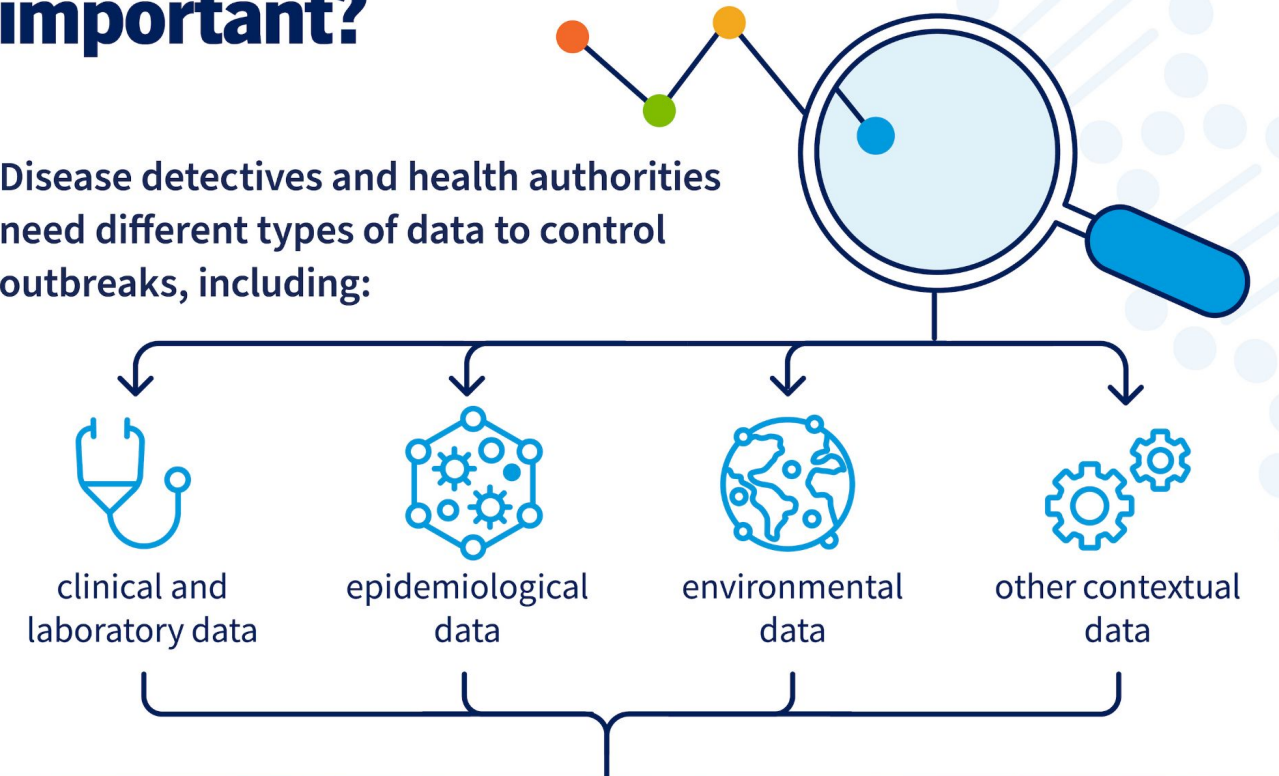
Open tools and protocols drove global surveillance

**Enabled:**

- Identification of variants of concern
- Understand transmission and immune evasion
- Vaccine and diagnostic development

## Why is GENOMIC SURVEILLANCE important?

Disease detectives and health authorities need different types of data to control outbreaks, including:

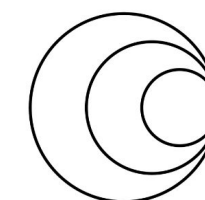


By adding genomic data, they can more quickly understand how a pathogen behaves and how to control it.



This is a powerful tool in public health surveillance.

The Global Genomic Surveillance Strategy for Pathogens with Pandemic and Epidemic Potential



wellcome connecting science



COVID-19 GENOMICS GLOBAL TRAINING

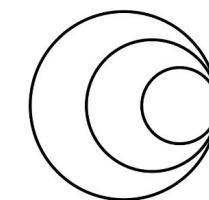
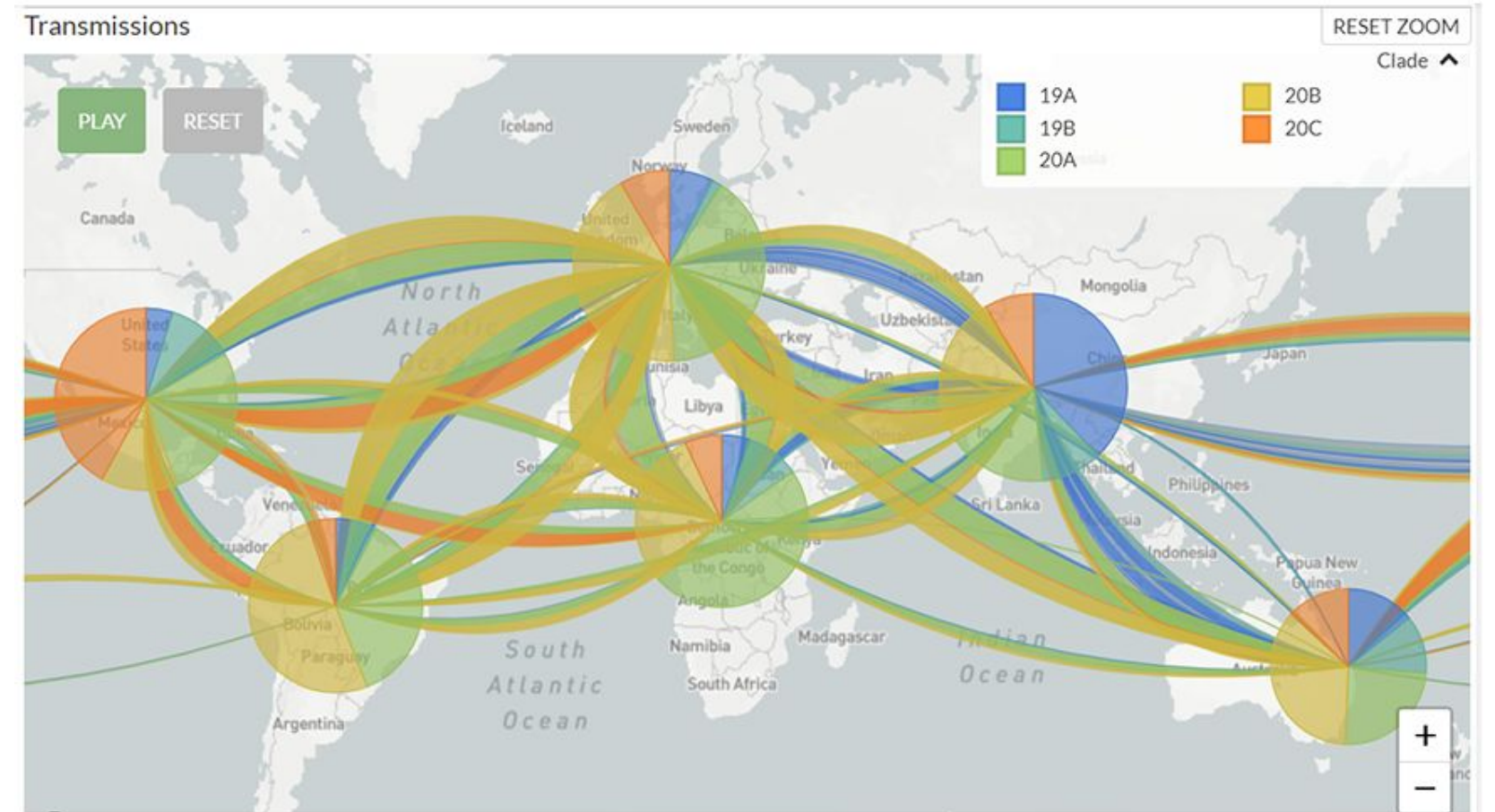
# SARS-CoV-2 genomic epidemiology: phylogenetics takes the spotlight.

- Origin of the virus
- Estimating R0
- Spread
- Identifying variant of concern by careful analysis of phylogeny and transmission
- Understand and advice on mitigation measures
- Outbreak control

Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from **GISAID**

Showing 3743 of 3743 genomes sampled between Dec 2019 and Oct 2020.

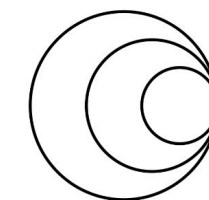


**wellcome**  
**connecting**  
**science**



**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Section 1: Intro to phylogenetics



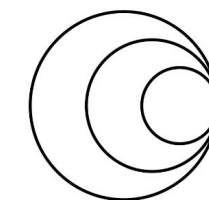
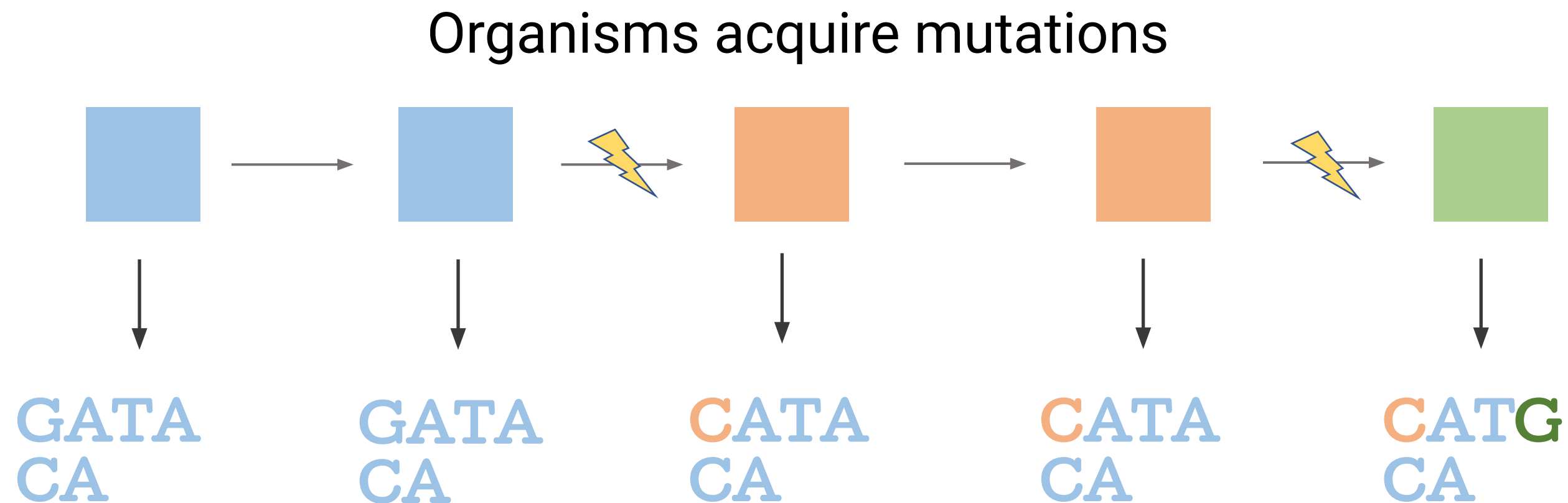
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Section 1: Intro to phylogenetics

- In biology, phylogenetics is the study of the evolutionary history and relationships between or within groups of organisms.

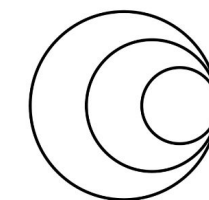
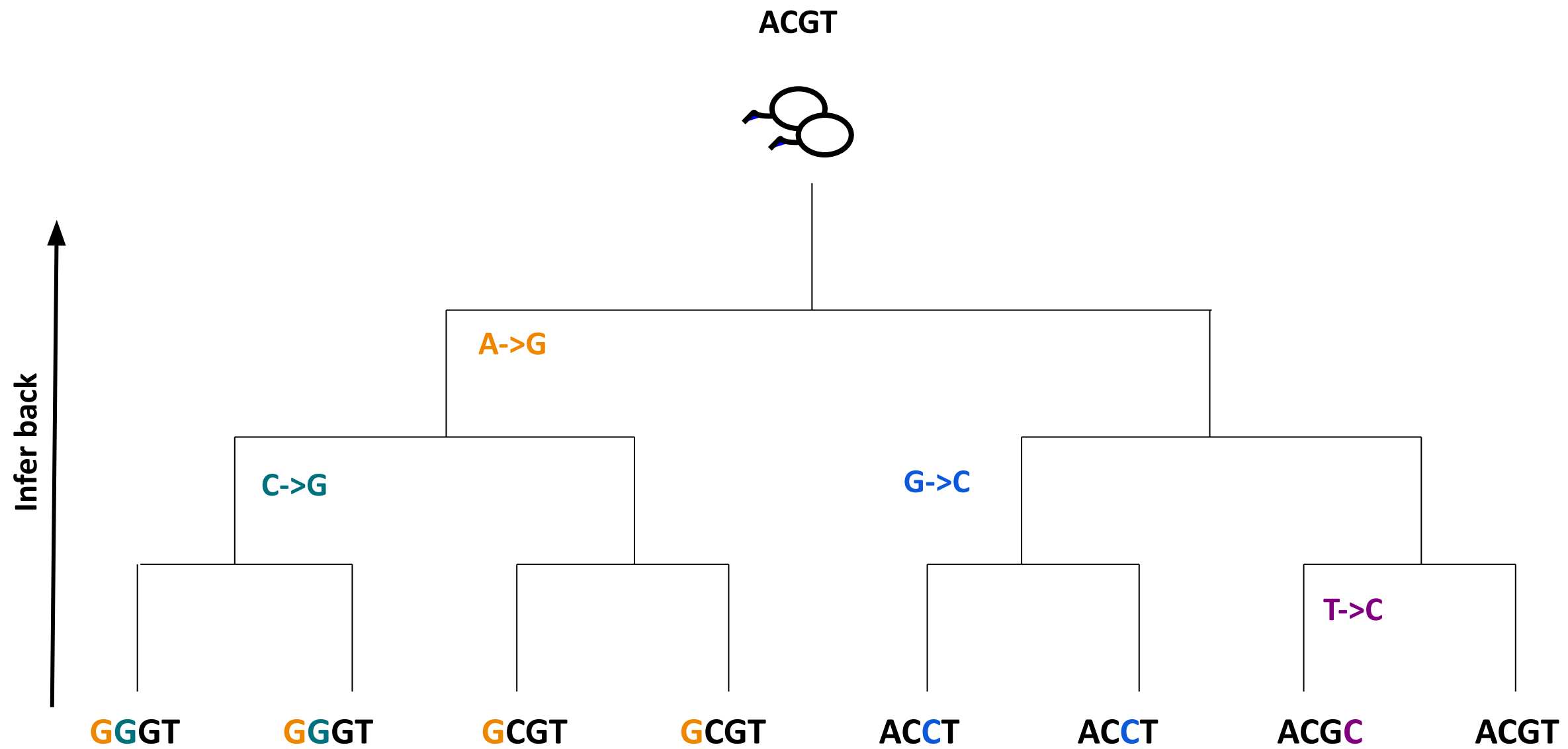


wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

# Mutations tell us about relationships



wellcome  
connecting  
science

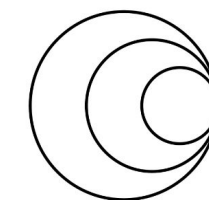
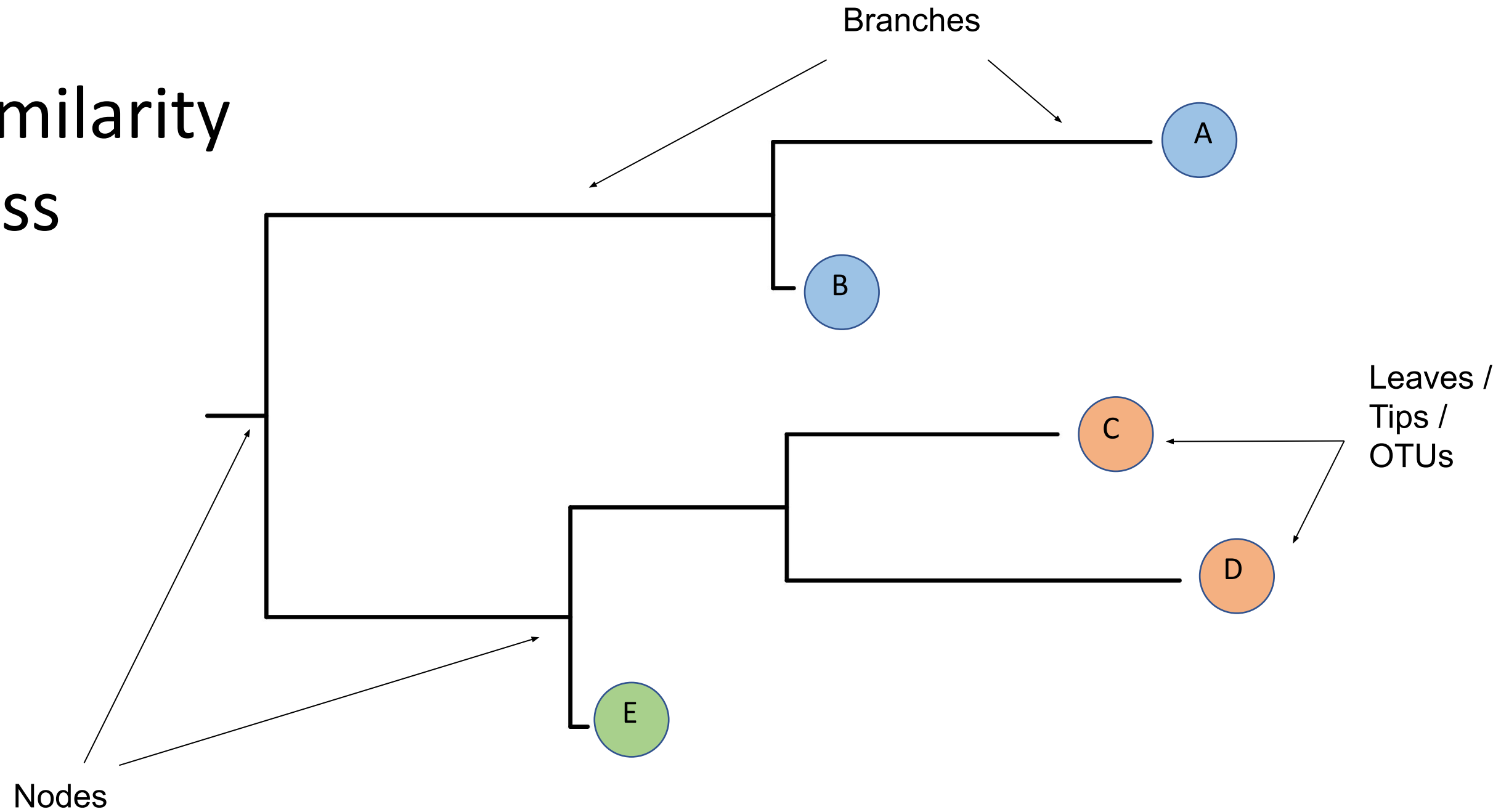


COVID-19  
GENOMICS  
GLOBAL TRAINING



# Phylogenetic trees reveal relationships

Genetic similarity  
Relatedness

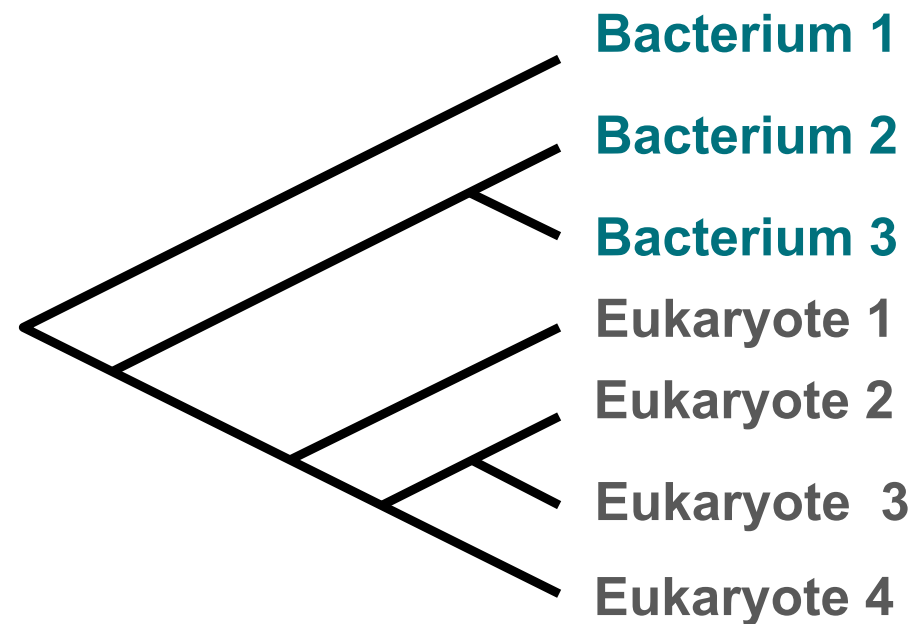


**wellcome**  
**connecting**  
**science**

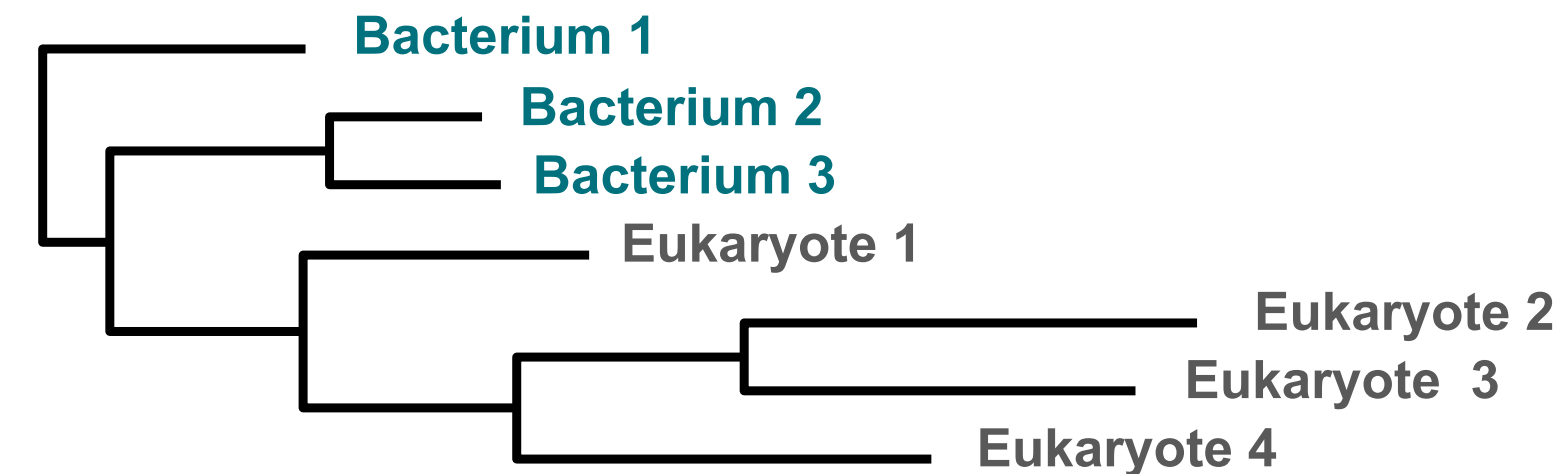


**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Cladograms vs Phylograms

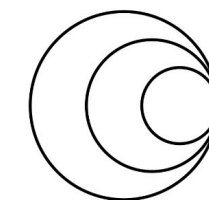


**Cladograms** show branch order (topology) only - branch lengths are meaningless



**Phylograms** show branch order and branch lengths with scale

Absolute measure of divergence  
(e.g. time, SNPs)



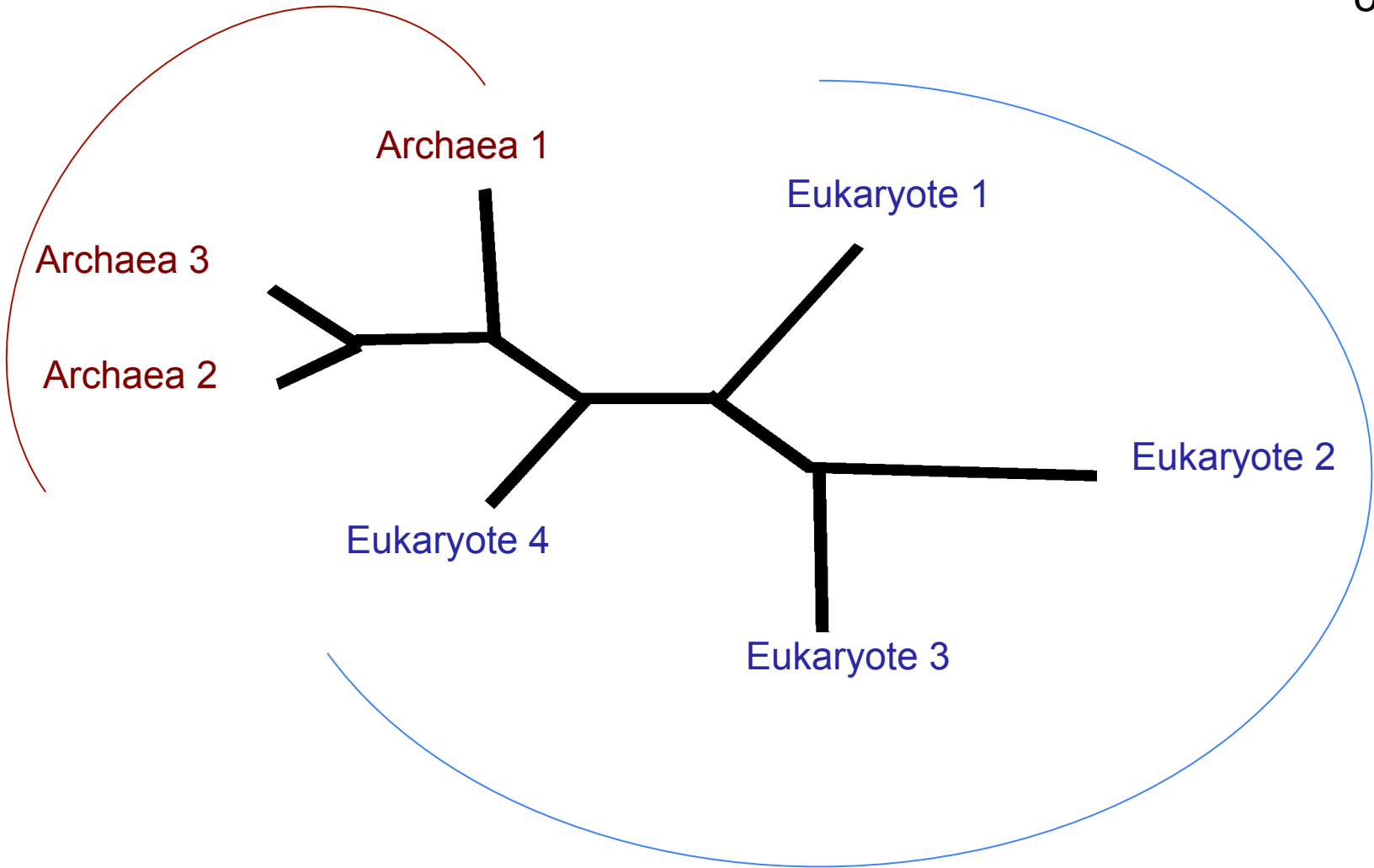
**wellcome  
connecting  
science**



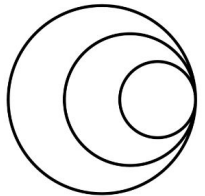
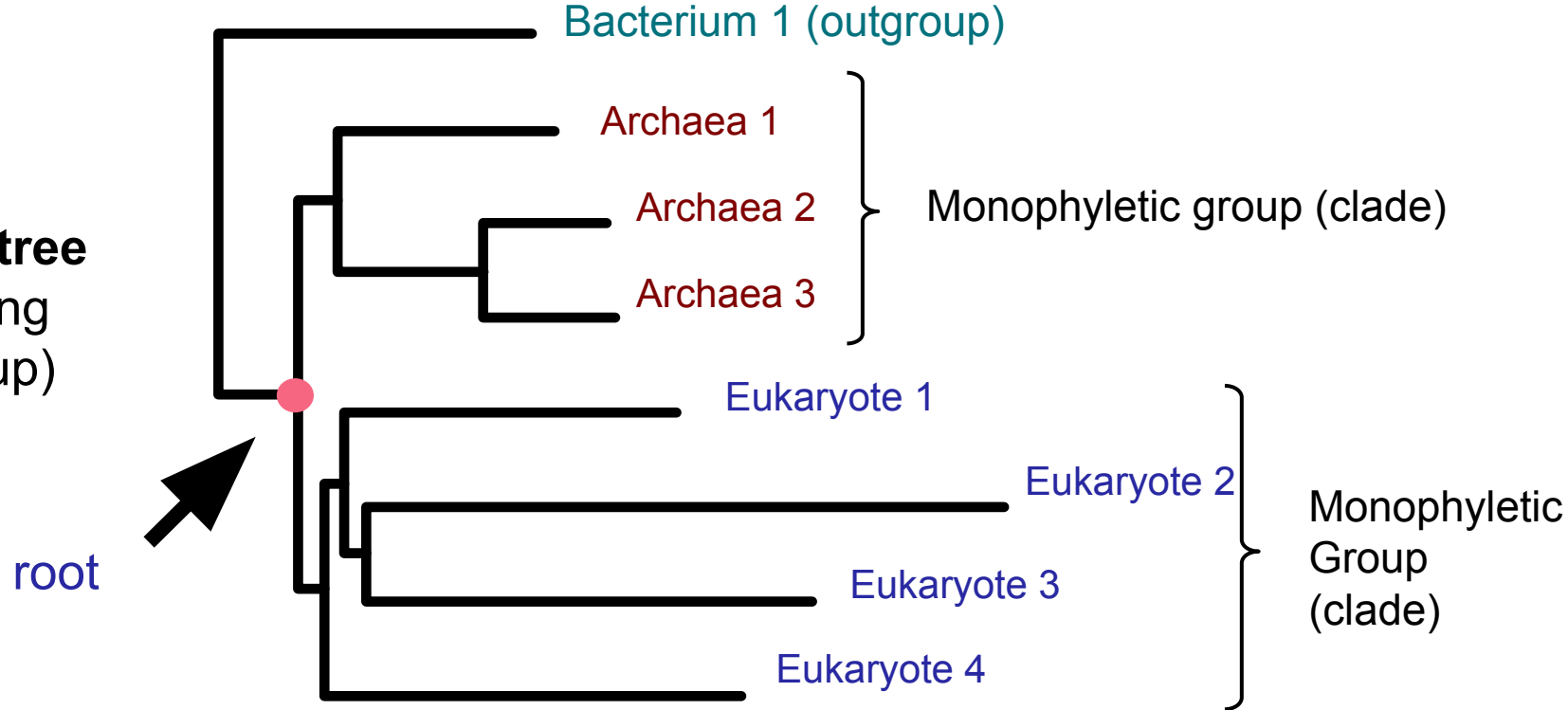
**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Rooted and Unrooted trees

Unrooted tree

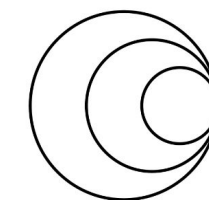


Rooted tree  
(by using  
outgroup)



# Where to root a tree?

- Midpoint or Outgroup
  - Check what other people in the field are doing and define outgroup
  - Include published references in phylogeny, choose midpoint root and check to see where the published sequences cluster
  - **If in doubt** start with midpoint root and work from there



**wellcome**  
**connecting**  
**science**



**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Building a phylogenetic tree

Identify protein, DNA or RNA sequences of interest

- Fasta format file of concatenated sequences

Multiple sequence alignment

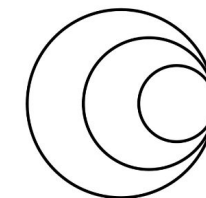
- ClustalX, Muscle, Mafft

Construct phylogeny

- PHYML, RAxML, IQ-Tree, FastTree

View and edit tree

- Figtree



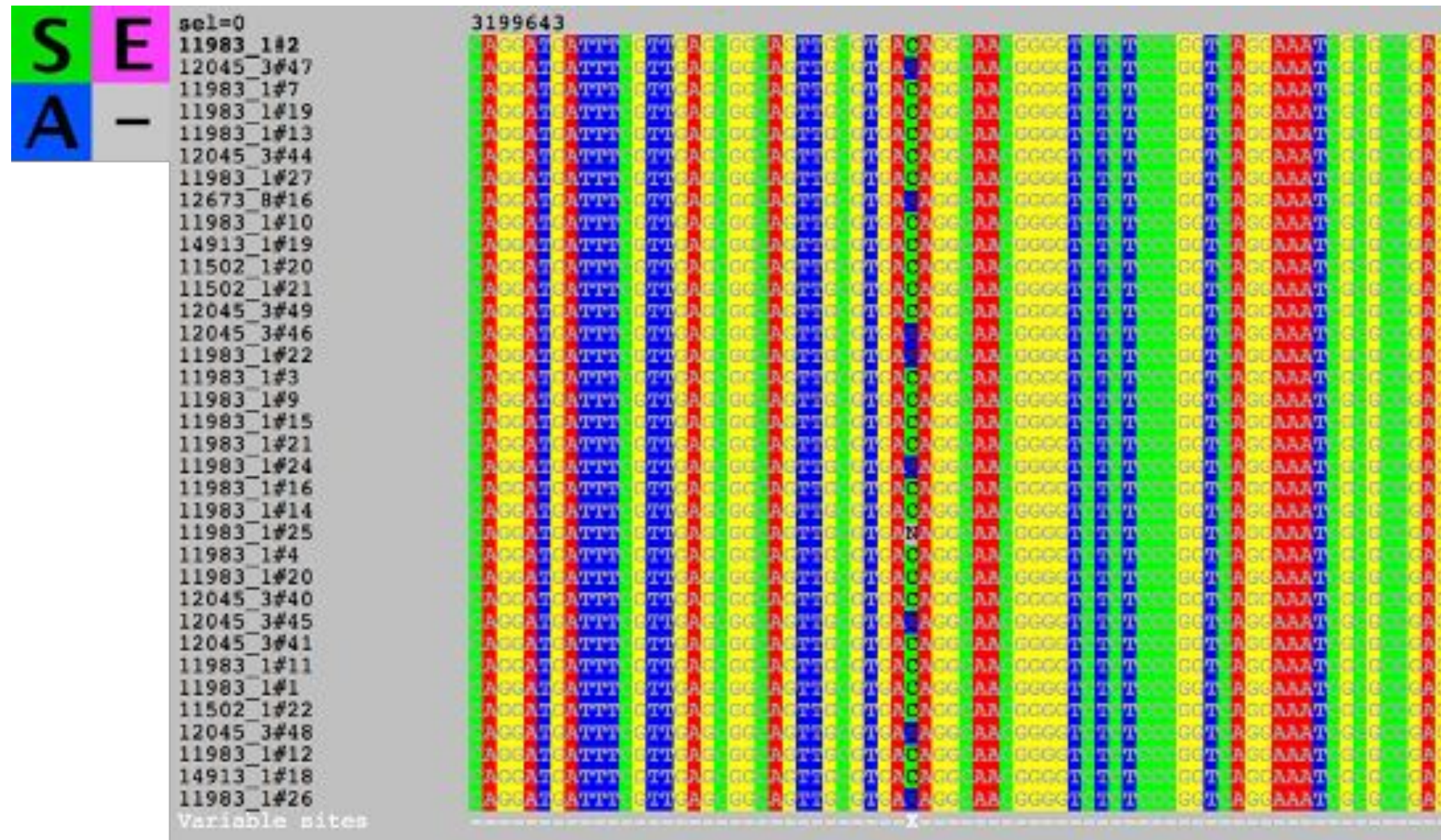
**wellcome  
connecting  
science**



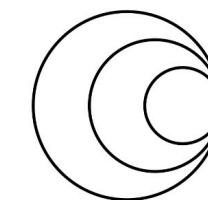
**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Multiple sequence alignment (MSA)

MSA is best hypothesis of **positional homology** between bases/amino acids of different sequences



This is perhaps most important step!!



wellcome  
connecting  
science

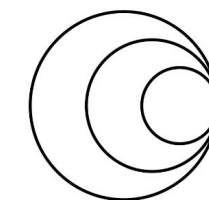


COVID-19  
GENOMICS  
GLOBAL TRAINING

# Constructing a phylogenetic tree

Method	Data used	Tree search	Evolutionary Model
<b>Distance</b>	Pairwise distance	Simple algorithm	Can be complex
<b>Parsimony *</b>	All sites	Mainly hill climbing	Simple
<b>Maximum likelihood *</b>	All sites	Hill climbing	Can be complex
<b>Bayesian inference *</b>	All sites (+ other info)	MCMC	Can be very complex

\* attempt to find the BEST tree



**wellcome  
connecting  
science**

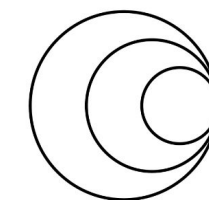


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Constructing a phylogenetic tree

Method	Data used	Tree search	Evolutionary Model
<b>Distance</b>	Pairwise distance	Simple algorithm	Can be complex
<b>Parsimony *</b>	All sites	Mainly hill climbing	Simple
<b>Maximum likelihood *</b>	All sites	Hill climbing	Can be complex
<b>Bayesian inference *</b>	All sites (+ other info)	MCMC	Can be very complex

\* attempt to find the BEST tree



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**



# Constructing a phylogenetic tree

Method

Data used

Tree search

Evolutionary Model

**Distance**

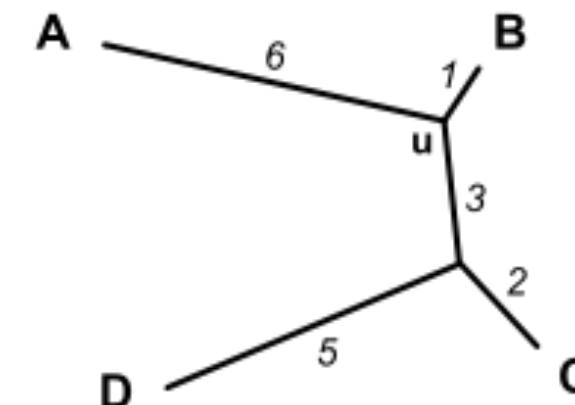
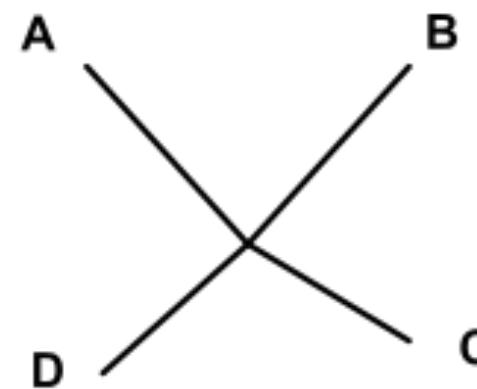
Pairwise distance

Simple algorithm

Can be complex

	A	B	C	D
A	0	7	11	14
B	7	0	6	9
C	11	6	0	7
D	14	9	7	0

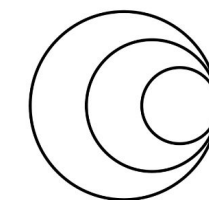
Distance matrix



# Constructing a phylogenetic tree

Method	Data used	Tree search	Evolutionary Model
<b>Distance</b>	Pairwise distance	Simple algorithm	Can be complex
<b>Parsimony *</b>	All sites	Mainly hill climbing	Simple
<b>Maximum likelihood *</b>	All sites	Hill climbing	Can be complex
<b>Bayesian inference *</b>	All sites (+ other info)	MCMC	Can be very complex

\* attempt to find the BEST tree

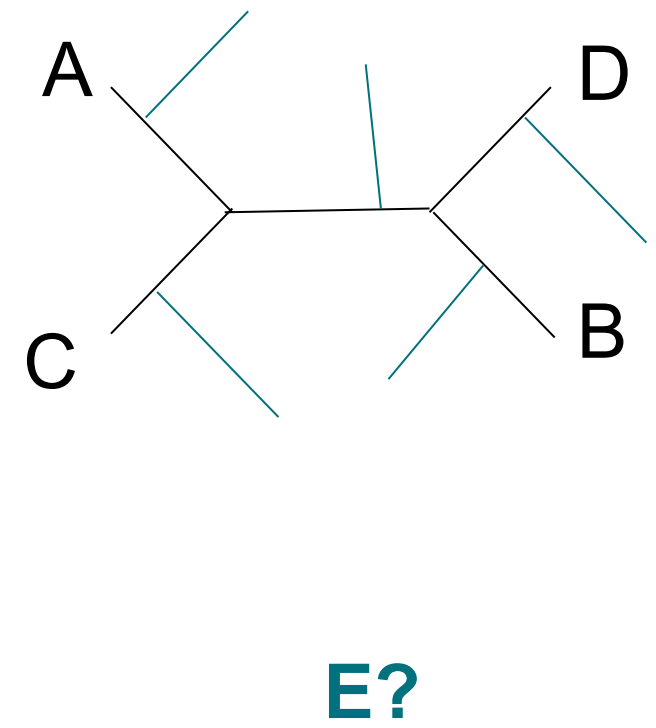
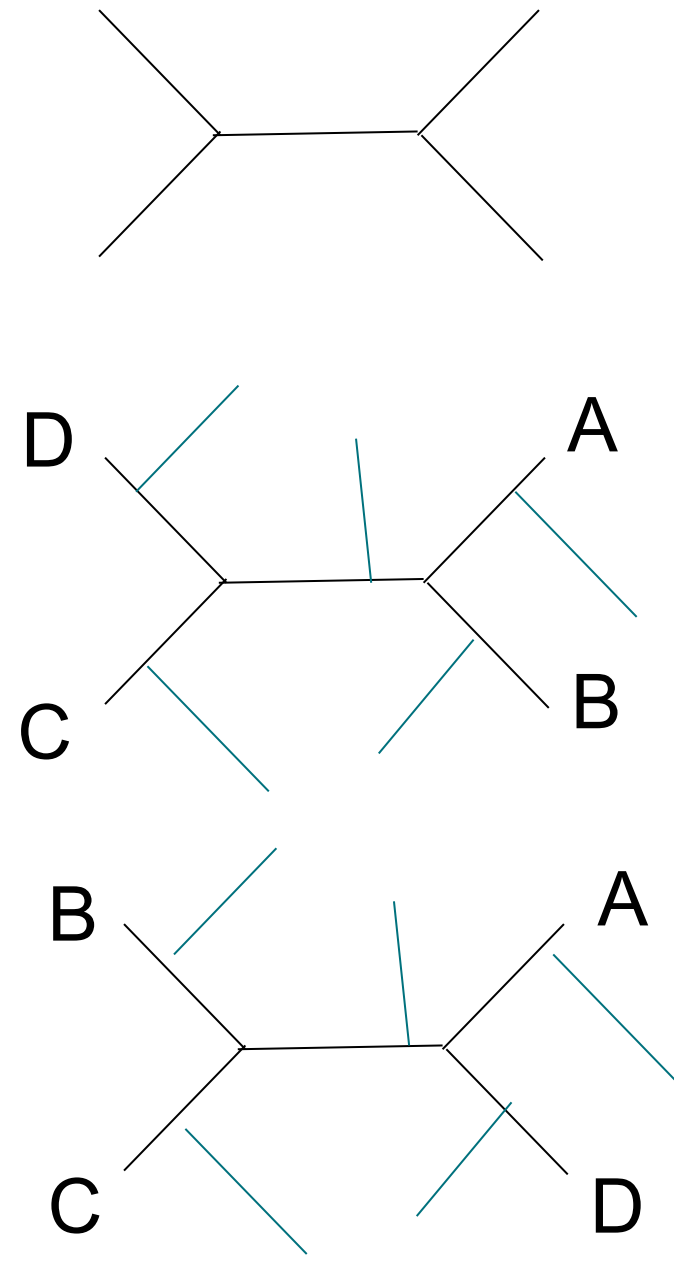


**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

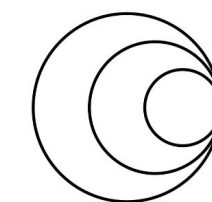
# Tree searching algorithms



Possible number of trees for  $n$  taxa

No. taxa	No. unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10395
80	$2.18 \times 10^{137}$

**$2n-3$**  possibilities to root the tree  
( $10 - 3 = 7$  for 5-taxon)

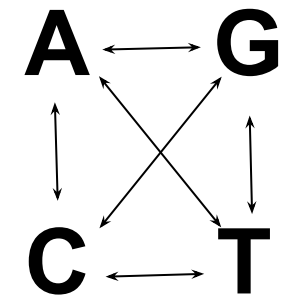
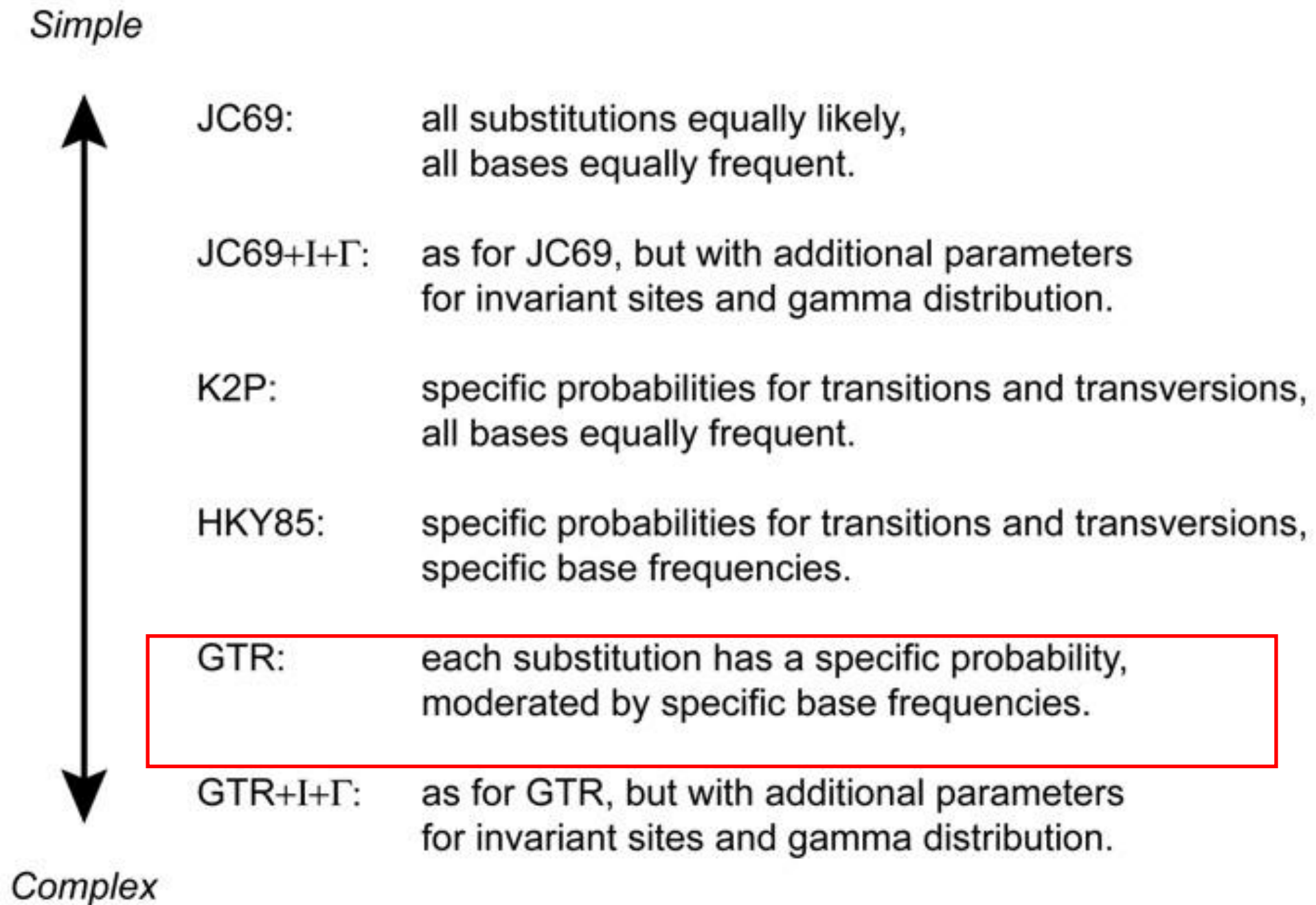


**wellcome  
connecting  
science**

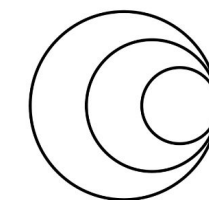


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Maximum likelihood evolutionary models

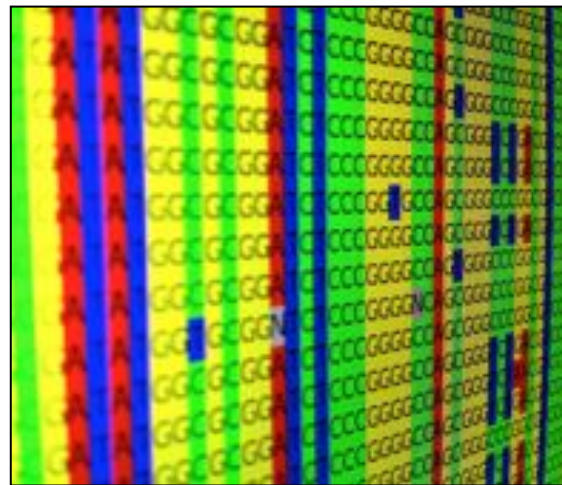


4 equilibrium base  
frequency parameters and  
6 substitution rate  
parameters

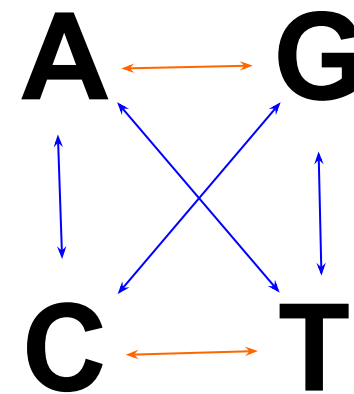


# Maximum likelihood phylogenetic models maximize the probability of achieving ...

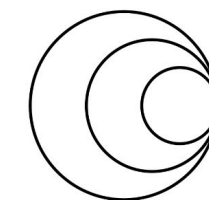
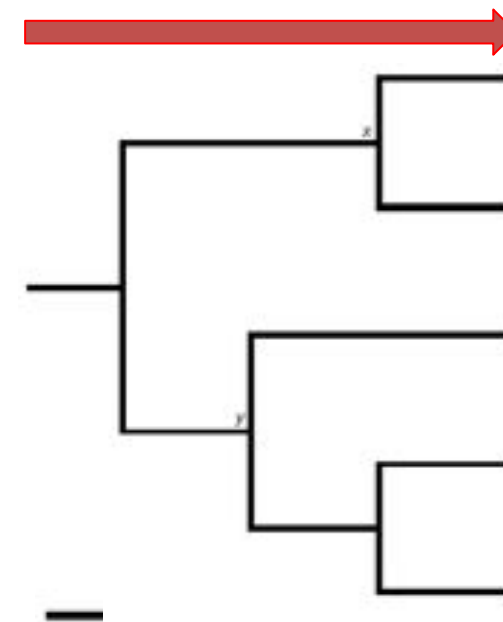
these data...



... if this happens...



... over this tree



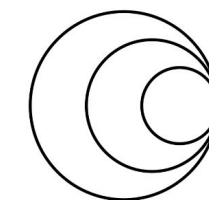
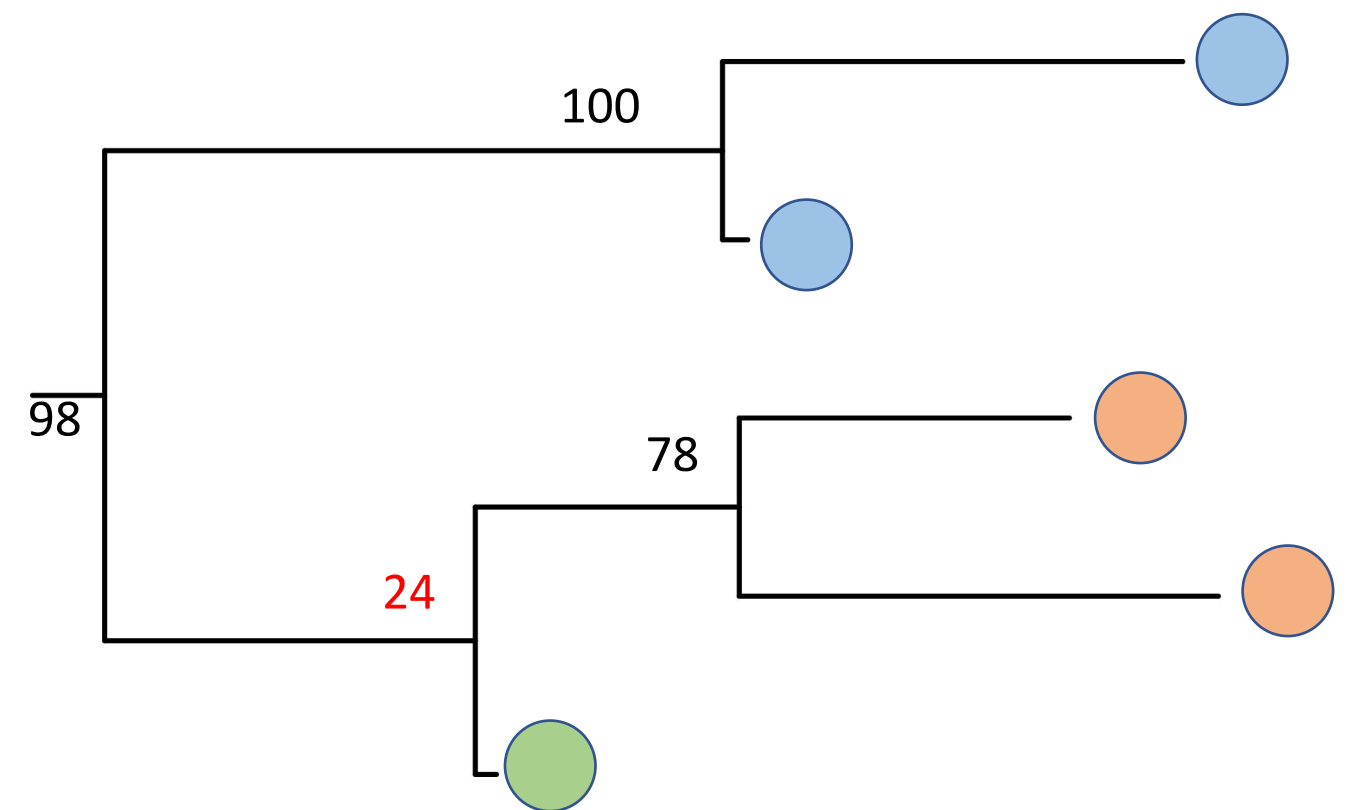
**wellcome**  
**connecting**  
**science**



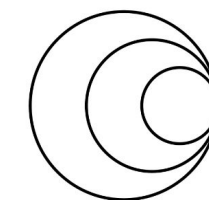
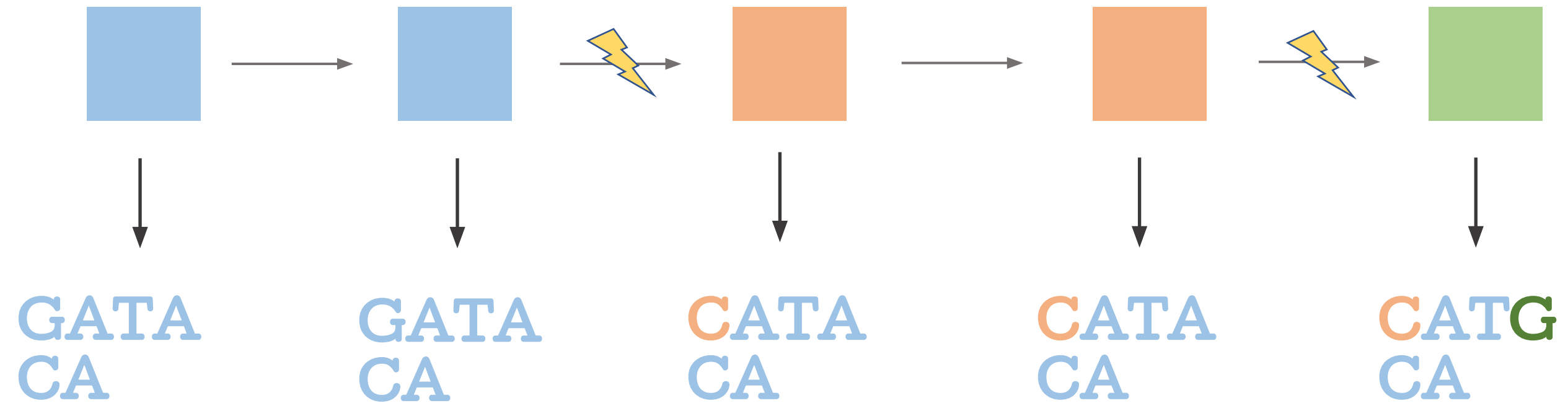
**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Bootstrapping

- Bootstrapping is a way to produce a confidence measure in the topology relationships found in a phylogenetic analysis
- X number of bootstraps (resampled replicates) are created of your input data (MSA)
- Typically run 100 – 1,000 bootstraps for ML analysis
- These are commonly used as a measure of support for these branches and are represented as a number on each tree branch



# Pathogens mutate as they transmit

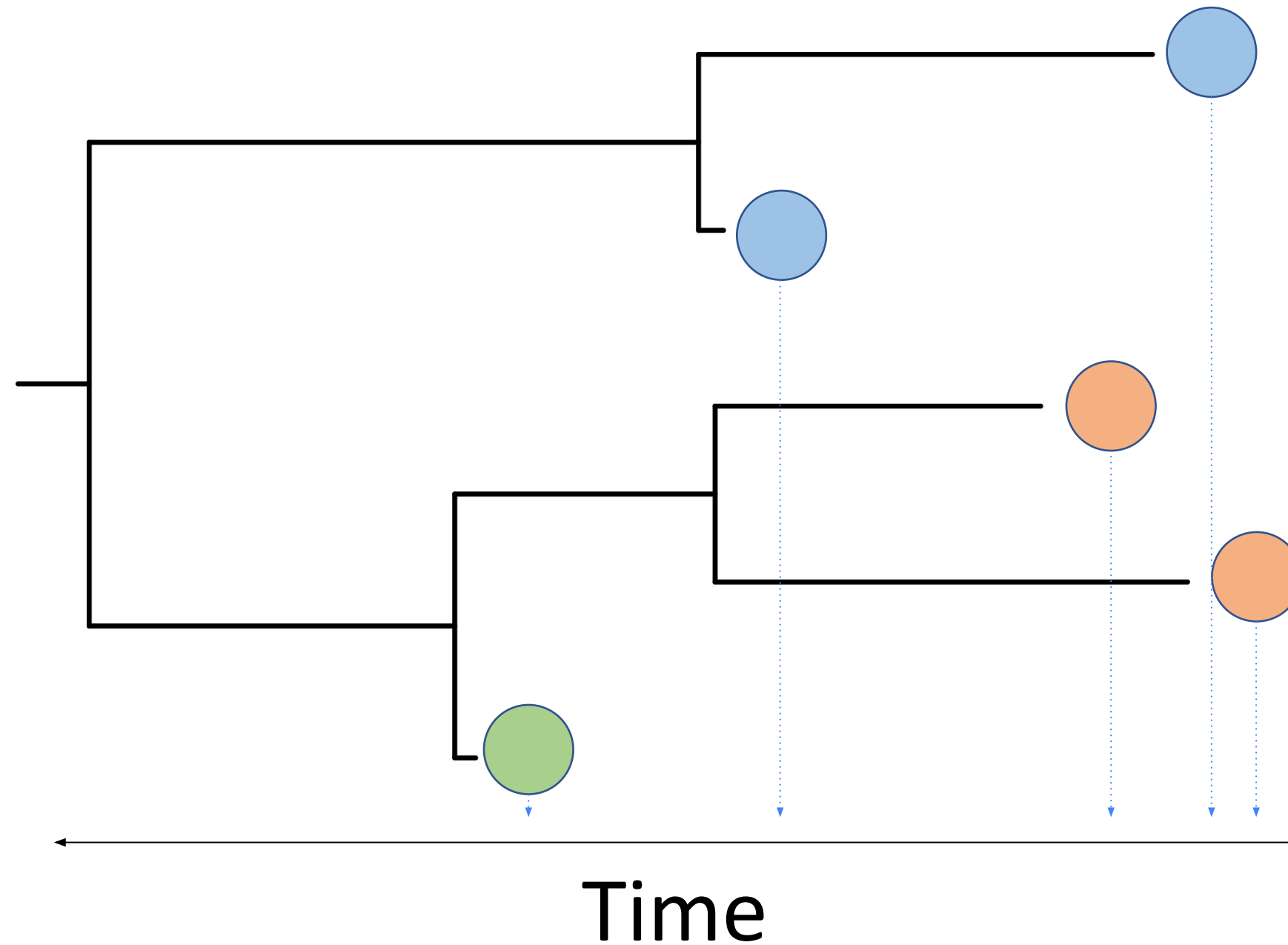


**wellcome**  
**connecting**  
**science**

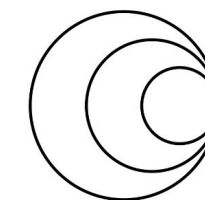


**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Trees reveal timing



Typically use BEAST to generate



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**



# Section 2: Interpreting phylogenetic analysis

Some Resources:

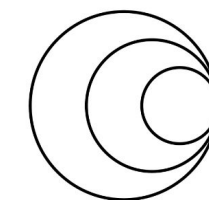
Understanding Evolutionary Trees, <https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-008-0035-x>

How to interpret the phylogenetic trees, <https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html>

Interpretation of Whole-Genome Sequencing for Enteric Disease Surveillance and Outbreak Investigation, <https://www.liebertpub.com/doi/10.1089/fpd.2019.2650>

Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9028907/>

Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9241095/>



**wellcome  
connecting  
science**

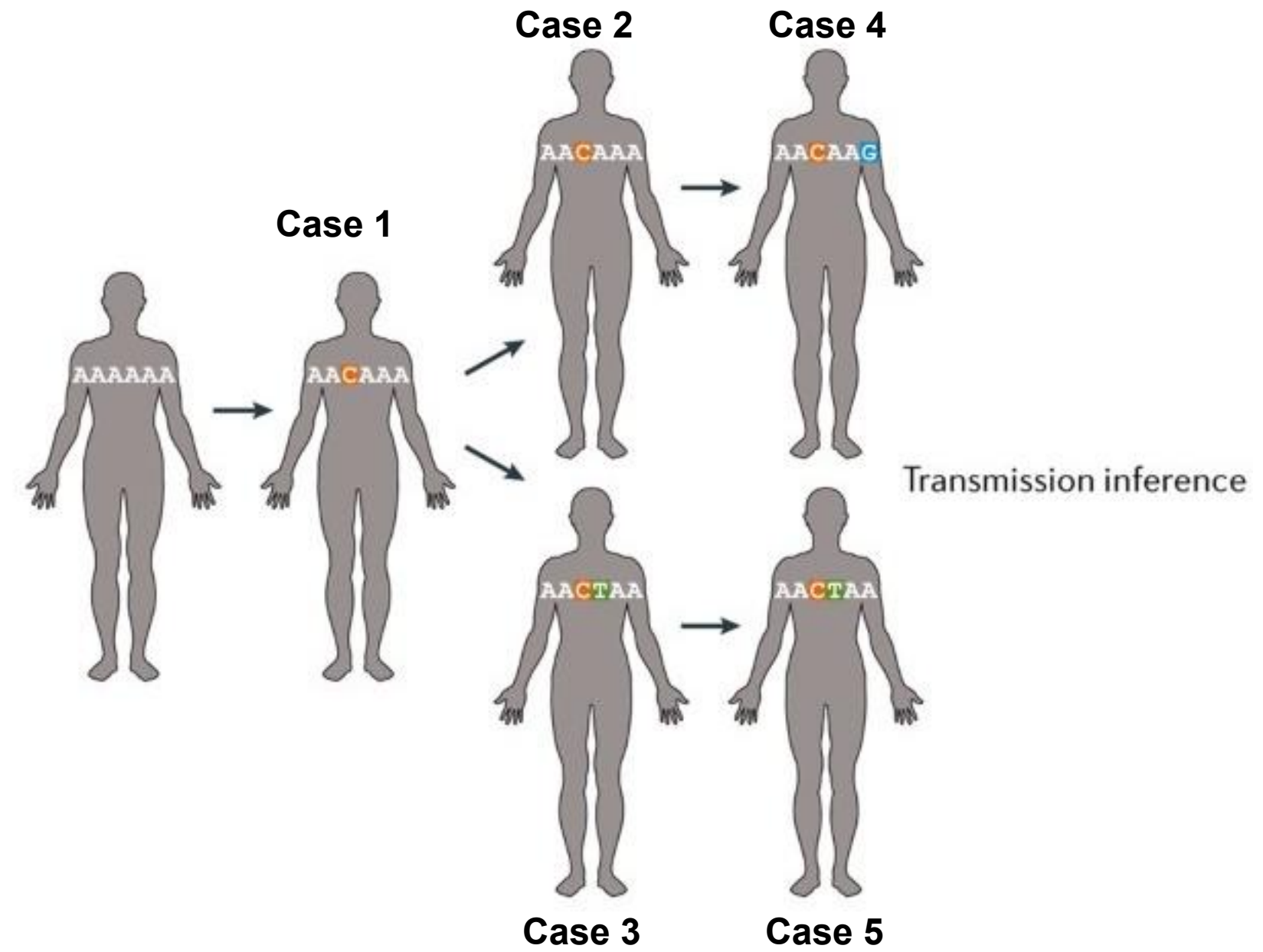


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Sequence relatedness can be used to infer transmission

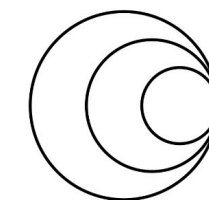
Virus replicate, random mutations occur

Key assumption is closer in sequence means share a more recent ancestor



Nature Reviews | Genetics

<https://www.nature.com/articles/nrg.2017.88>

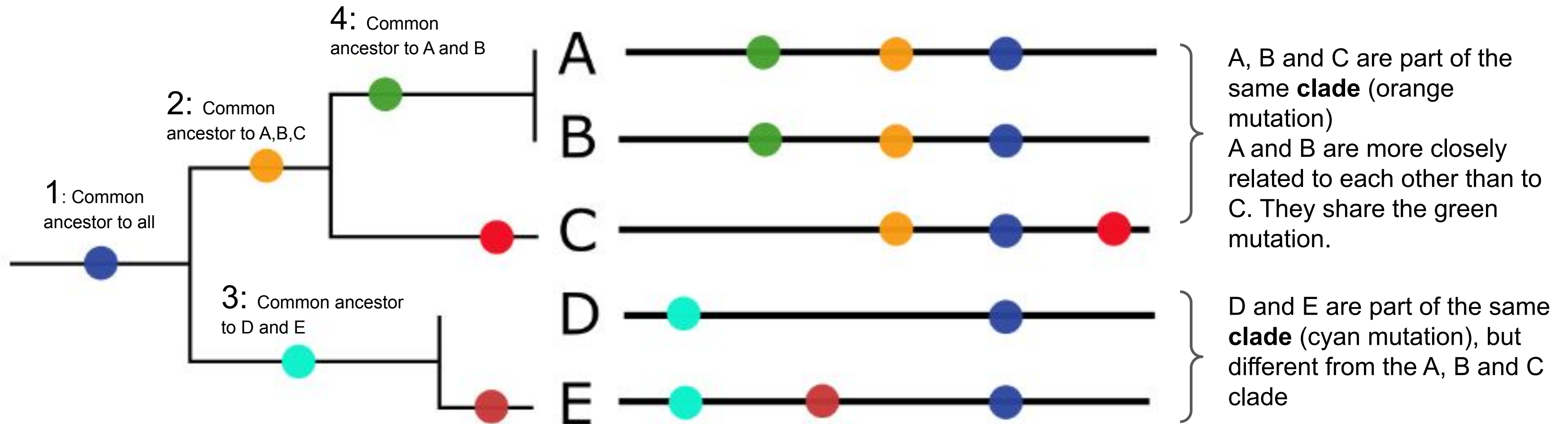


wellcome  
connecting  
science

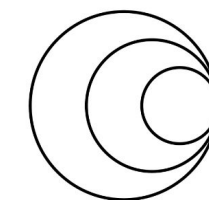


COVID-19  
GENOMICS  
GLOBAL TRAINING

# Phylogeny can be used to generate hypothesis about transmission



Trevor Bedford, <https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html>



wellcome  
connecting  
science

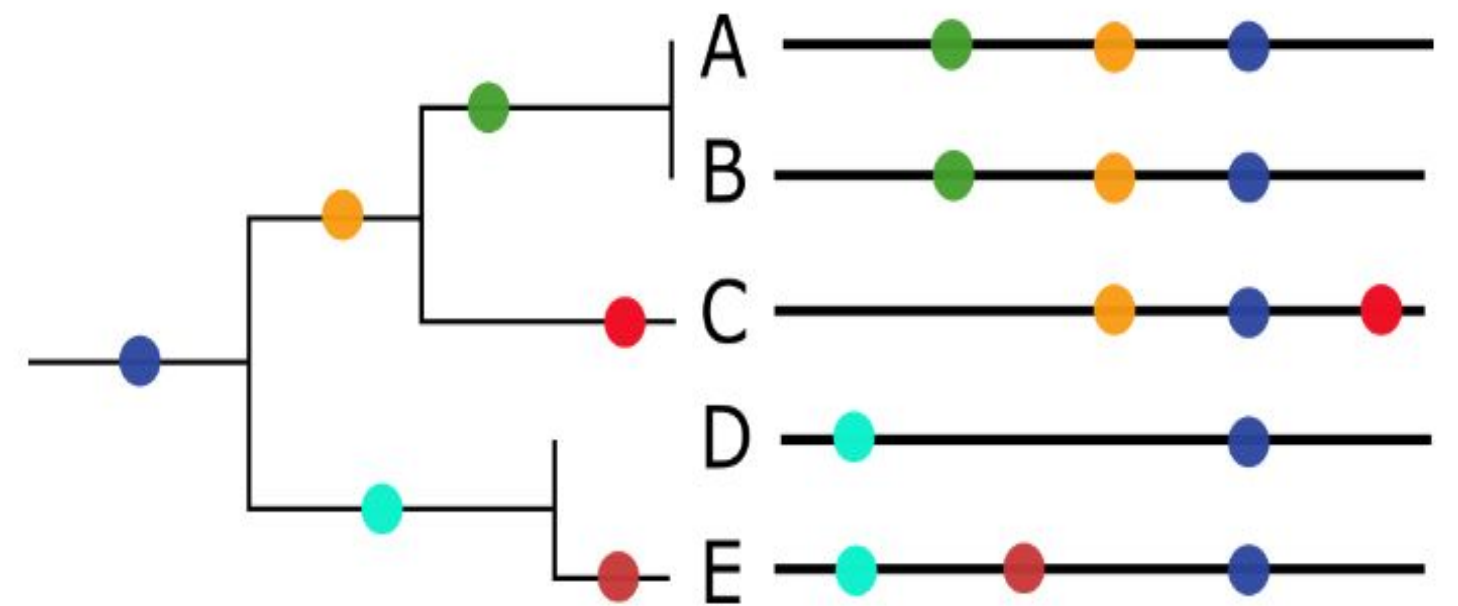


COVID-19  
GENOMICS  
GLOBAL TRAINING

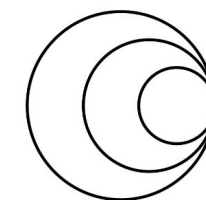
# Phylogeny can be used to generate hypothesis about transmission

For example, we could hypothesize that A and B are part of the same transmission event. **But** we cannot distinguish direct transmission between A and B or they were infected by the same individual.

Epidemiological information can support, reject or refine model of transmission.



<https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html>



wellcome  
connecting  
science

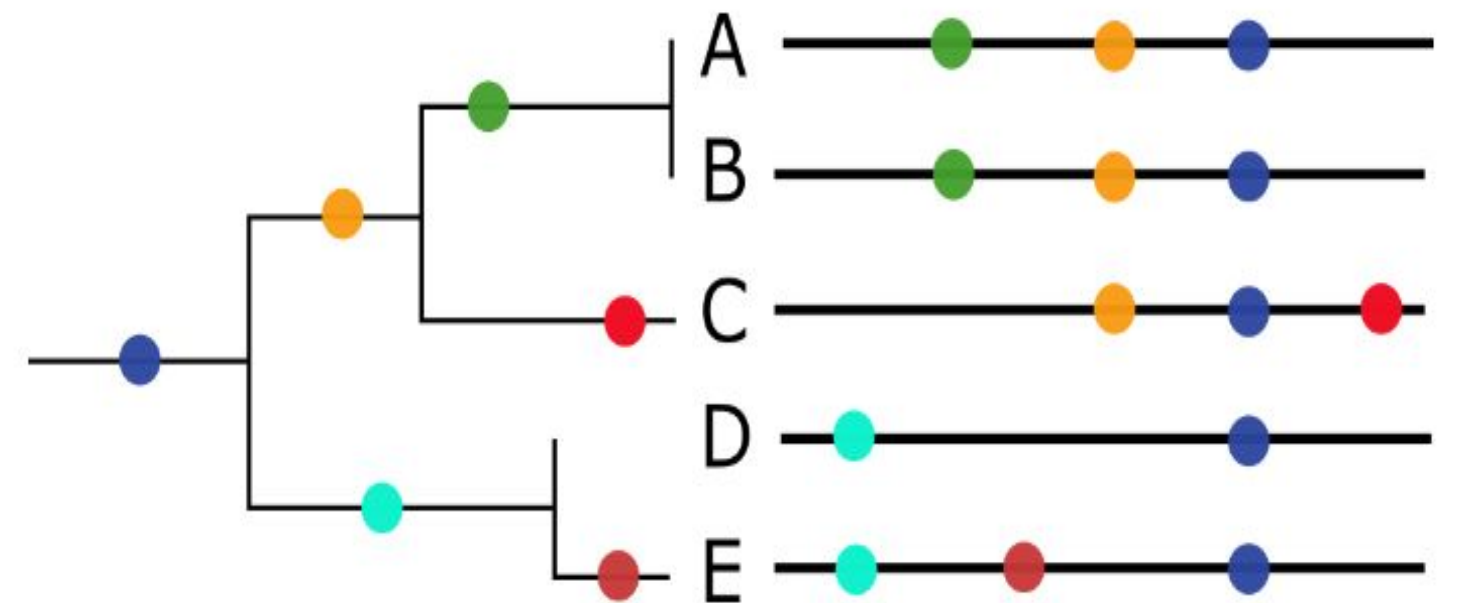


COVID-19  
GENOMICS  
GLOBAL TRAINING

# Phylogenetics can assist epidemiological investigations related to outbreaks

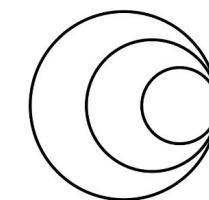
By refining outbreak by ruling in or out individuals

By generating hypothesis regarding transmission



<https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html>

Now a couple of examples



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Example 1: inflight transmission?

Genomic Evidence of In-Flight Transmission of SARS-CoV-2 Despite Predeparture Testing  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7920679/>

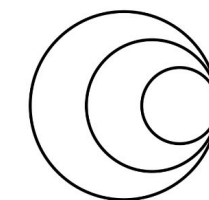
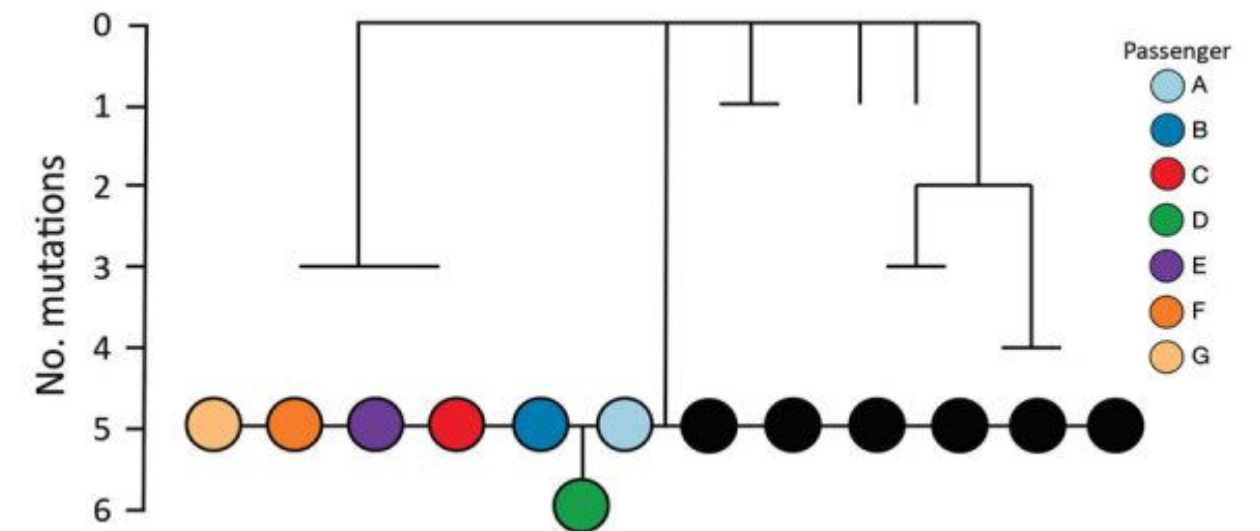
Dataset\_1\_flight

Follow notebook to analyse this dataset

7 passengers on the same flight tested positive after arriving in New Zealand.

**Question:** Are these cases linked? If so where did transmission take place?

You will find that all sequences are indistinguishable part from that of passenger D who has an additional SNP



wellcome  
connecting  
science



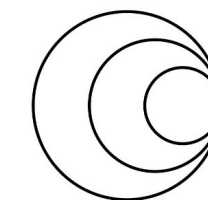
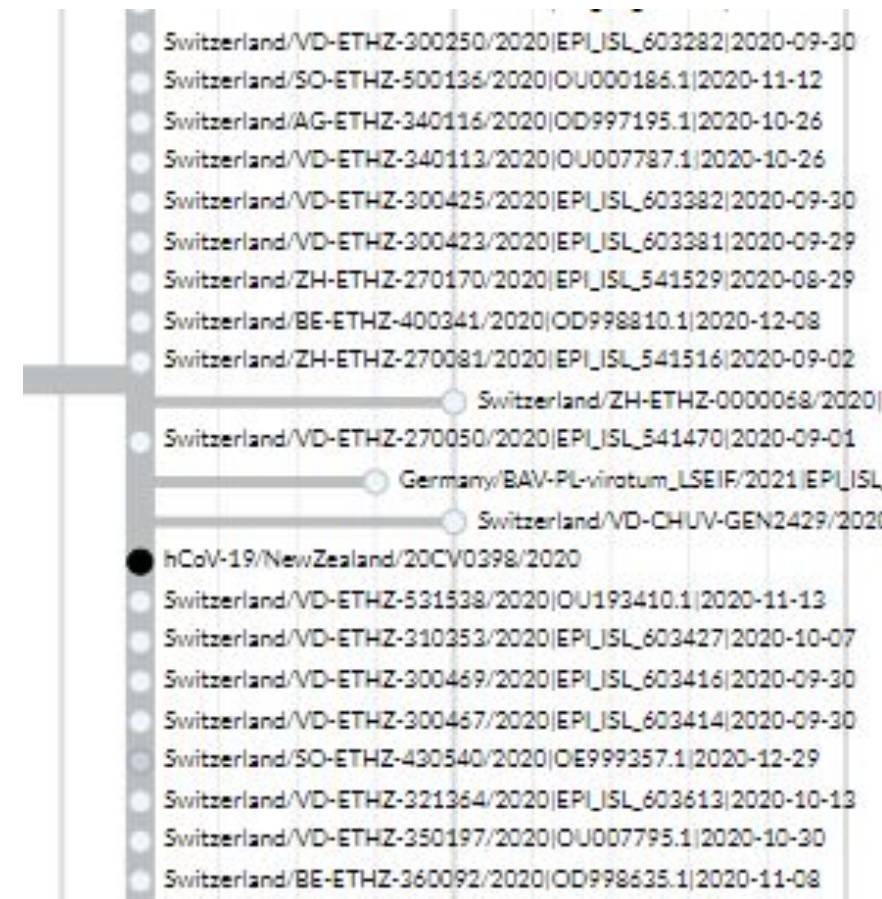
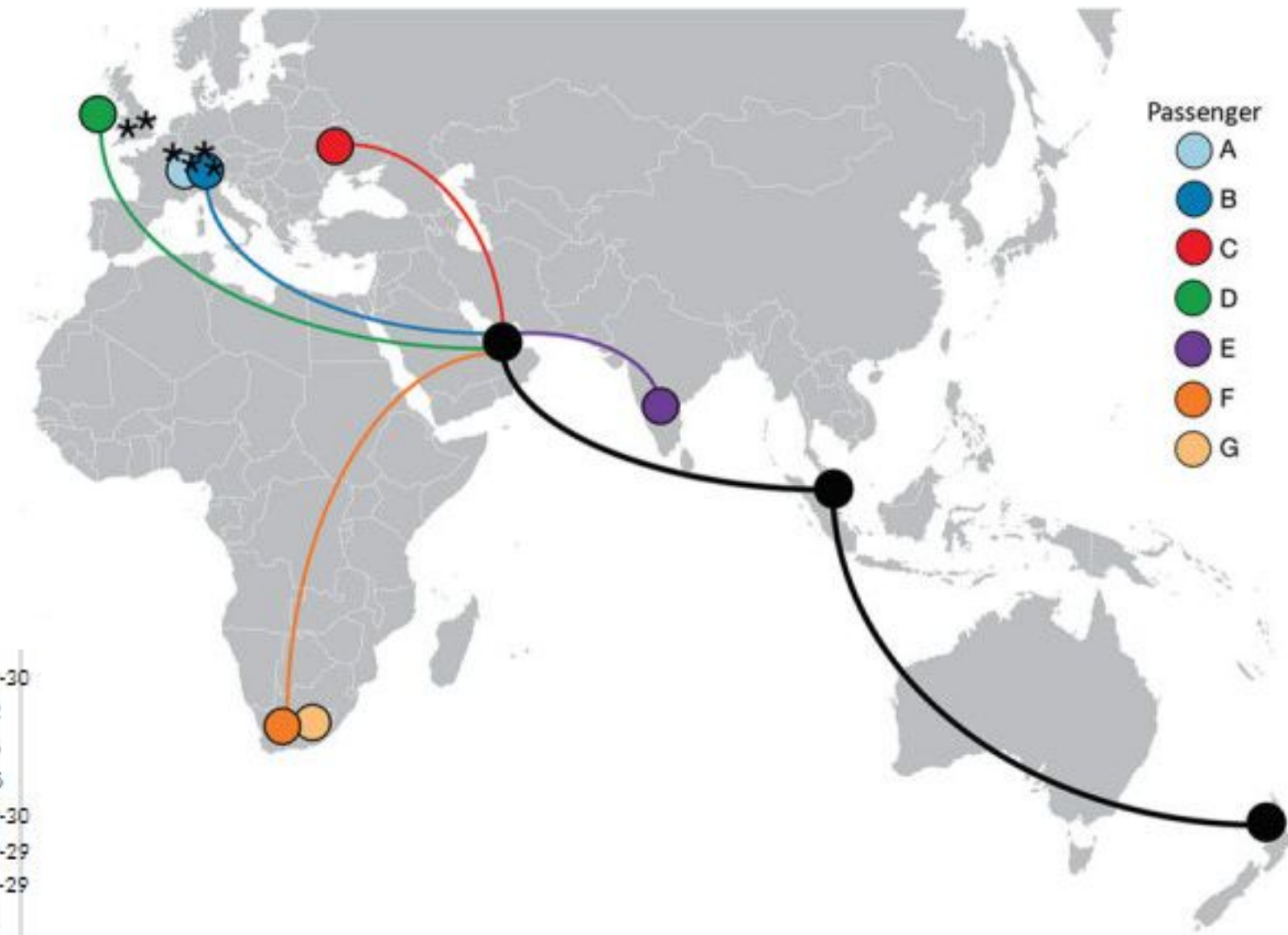
COVID-19  
GENOMICS  
GLOBAL TRAINING

# Example 1: inflight transmission?

Since these cases involved passengers on an international flight, we want to investigate whether there are any international linkages.

Can use this website to find linkages to genomes submitted to GISAID:  
<https://genome.ucsc.edu/cgi-bin/hgPhyloPlace>

Find closely related genomes from Switzerland, two of the passengers are from Switzerland.



wellcome  
connecting  
science

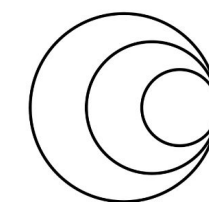
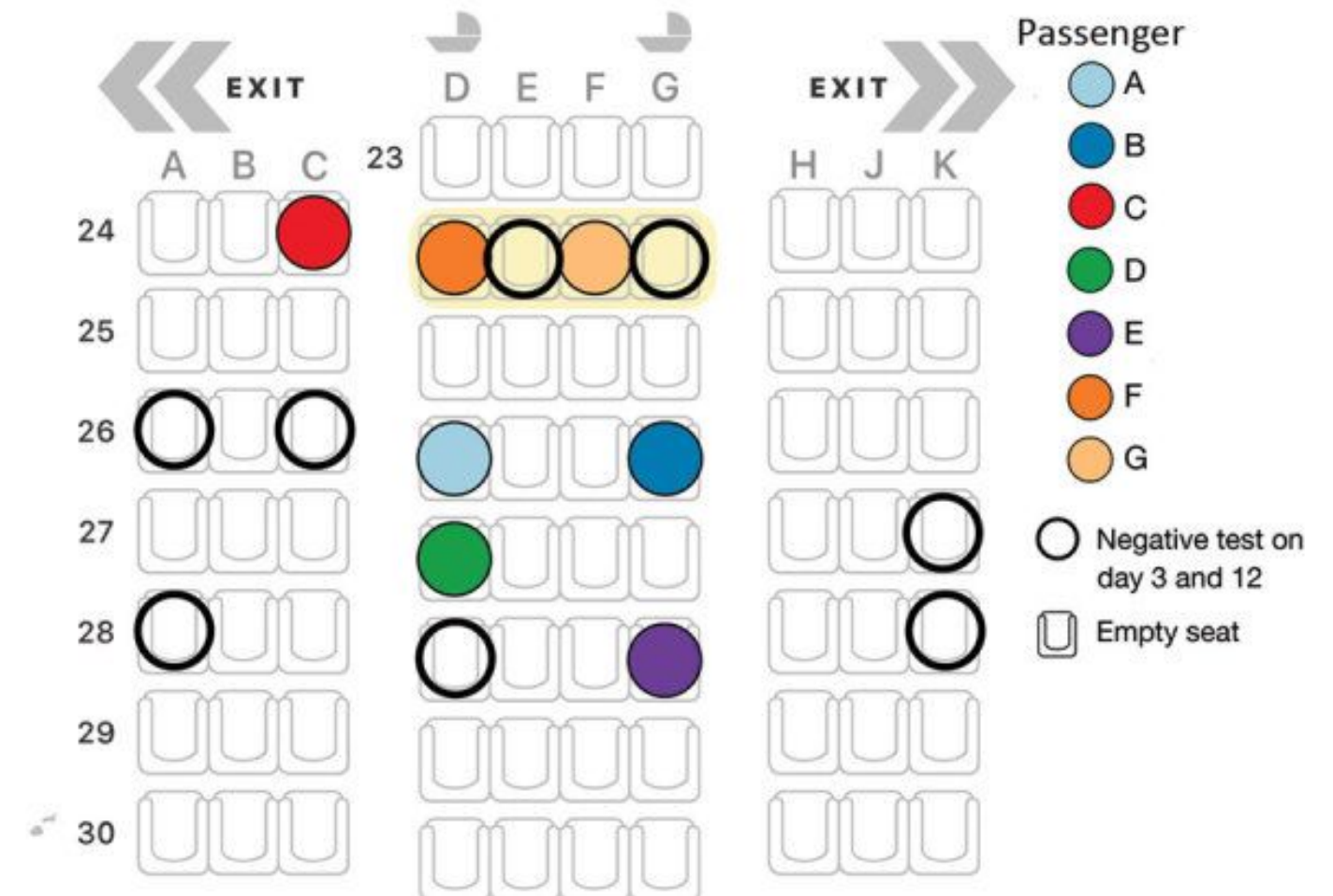


COVID-19  
GENOMICS  
GLOBAL TRAINING

# Example 1: inflight transmission?

Why likely inflight transmission?

- a) Unlikely happened after arrival in New Zealand because the 7 passengers traveled to different hotels on different buses.
- b) Passengers arrived from different countries and did not interact at connecting airports that we know of.
- c) Sitting closely to each other on the flight.



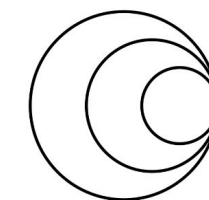
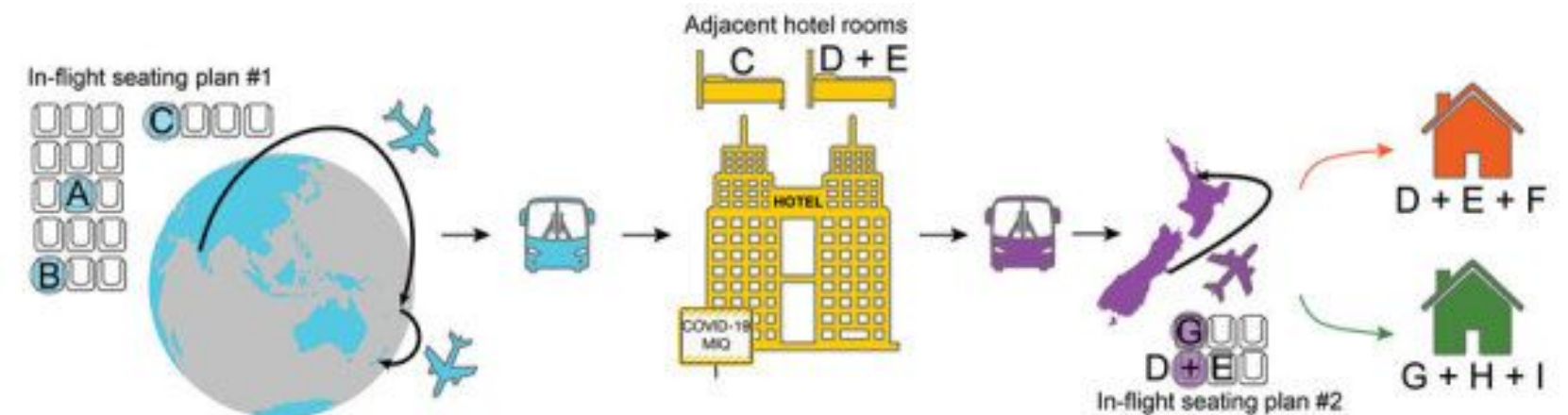
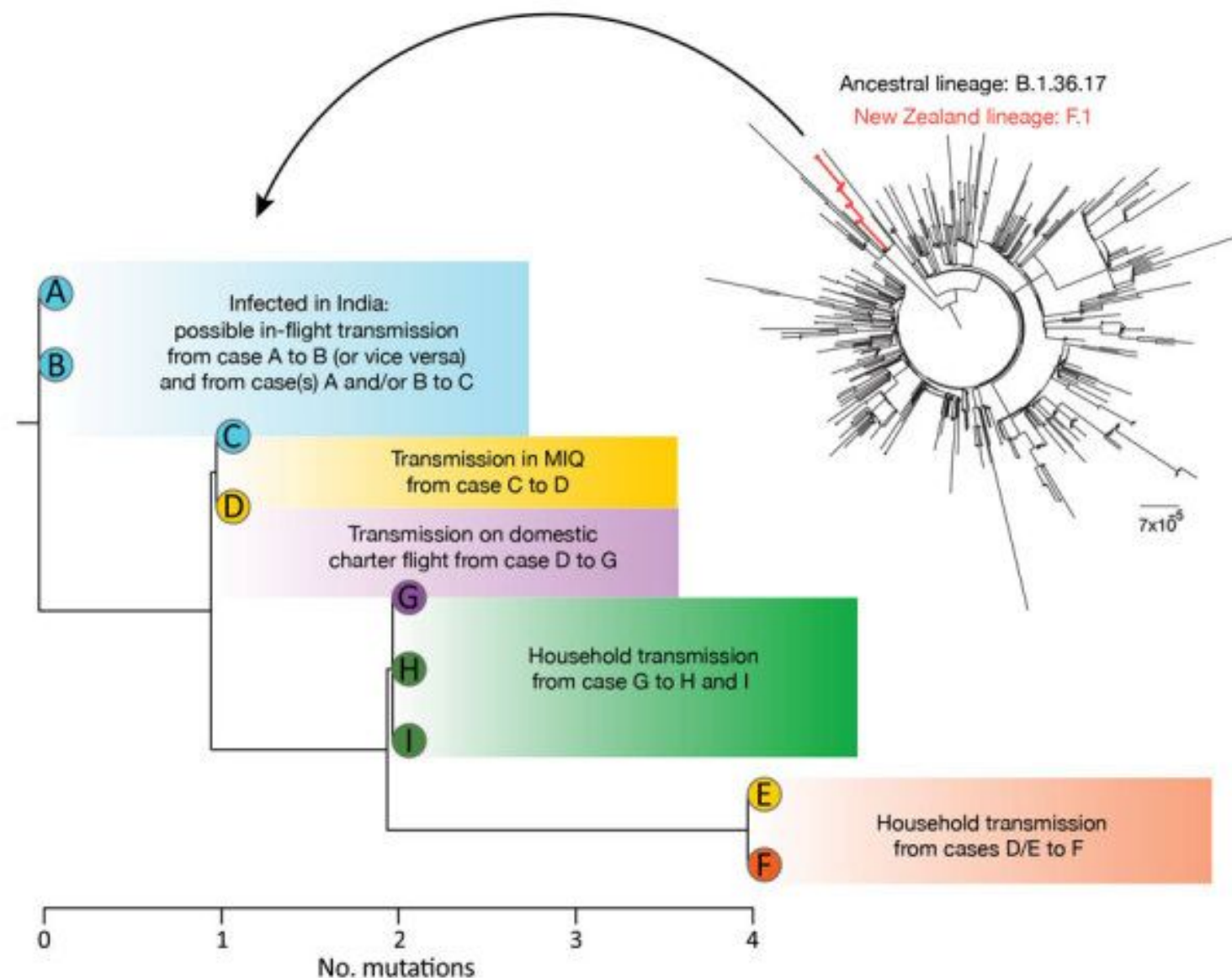


# Example 2: following an incursion

Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 during Border Quarantine and Air Travel, New Zealand (Aotearoa), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8084504/>

Dataset\_2\_hotel

Follow notebook to analyse this dataset



wellcome  
connecting  
science

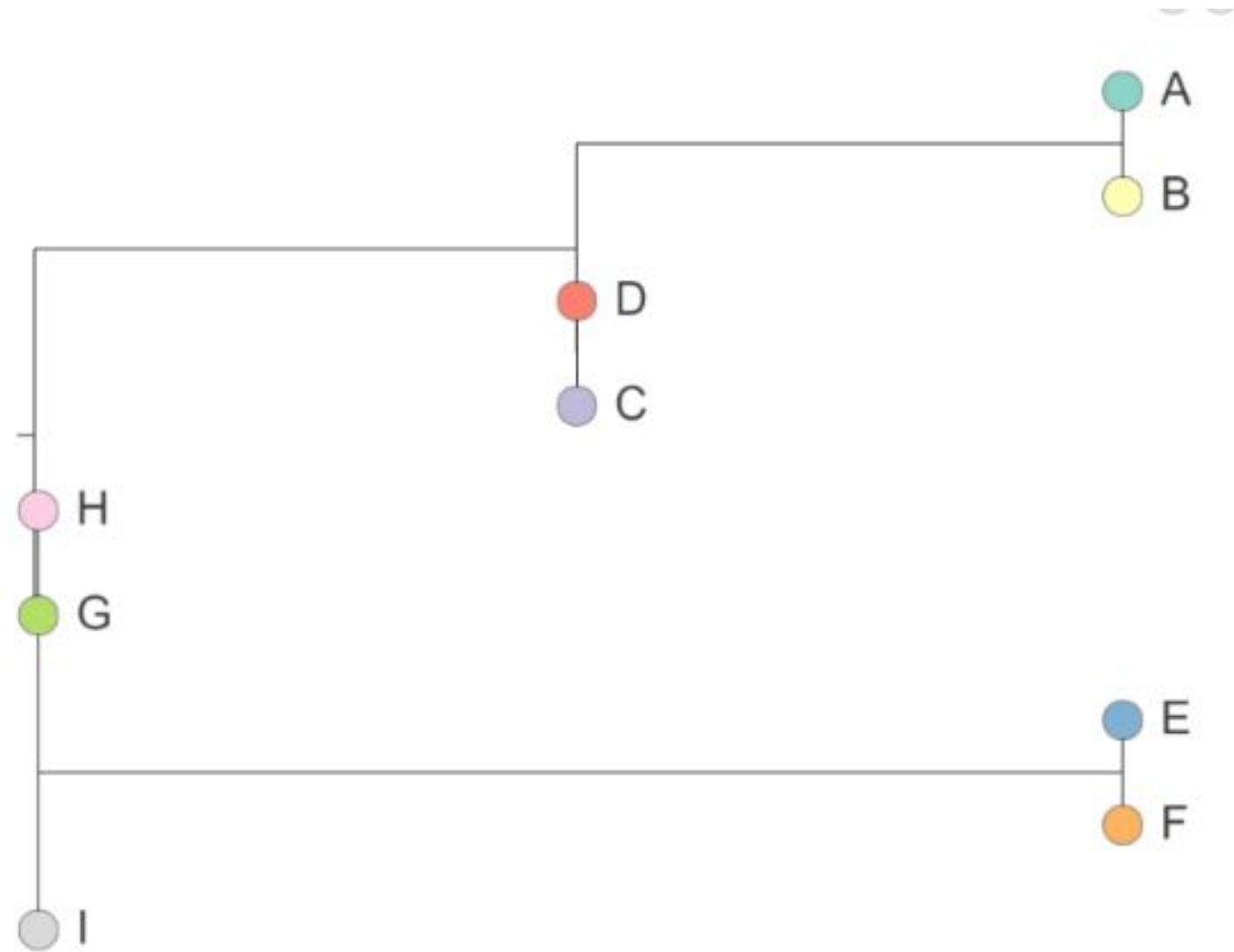


COVID-19  
GENOMICS  
GLOBAL TRAINING

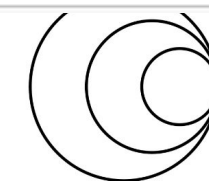
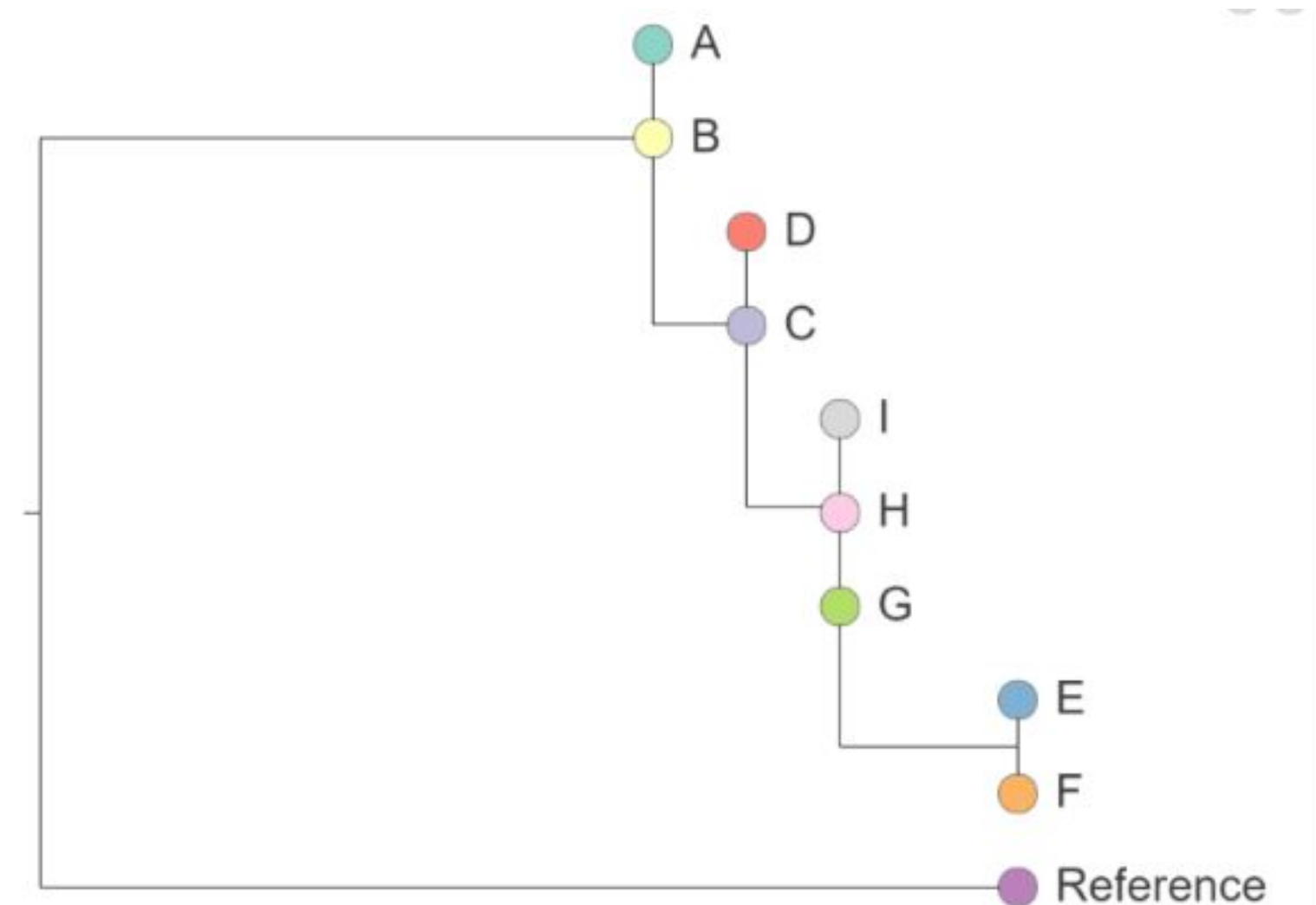
# Example 2: following an incursion

Highlights the importance of rooting and using an outgroup: try rooting your tree to A or B in notebook and see the effect

No outgroup



Rooted using reference

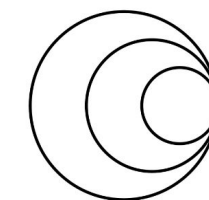
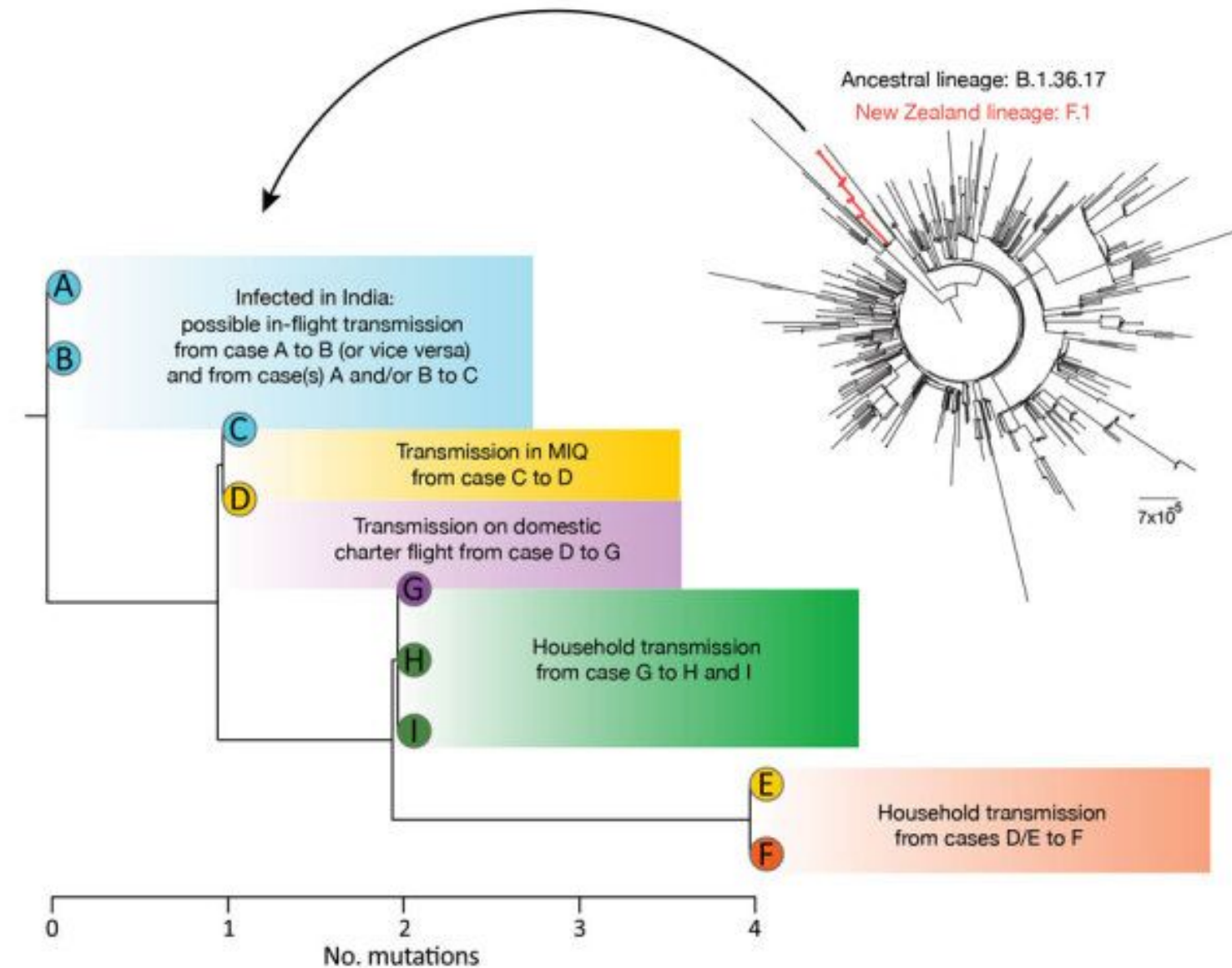


# Example2: following an incursion

With just 9 cases, you can see it is already quite difficult to follow.

Here is a interactive visualisation you can follow:

<https://microreact.org/project/5ELv2rXSKKeZ8XZCFXq9Ug-dataset2hotel#3sul-unnamed-view>



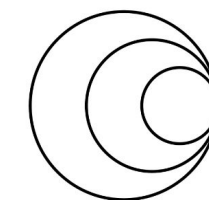
wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

# The two examples highlight:

- Importance of careful epidemiological investigation
- Need background or ancestral (basal) genomes to properly orient the tree
- Importance of sharing data



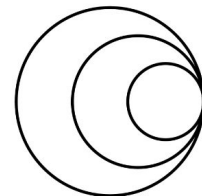
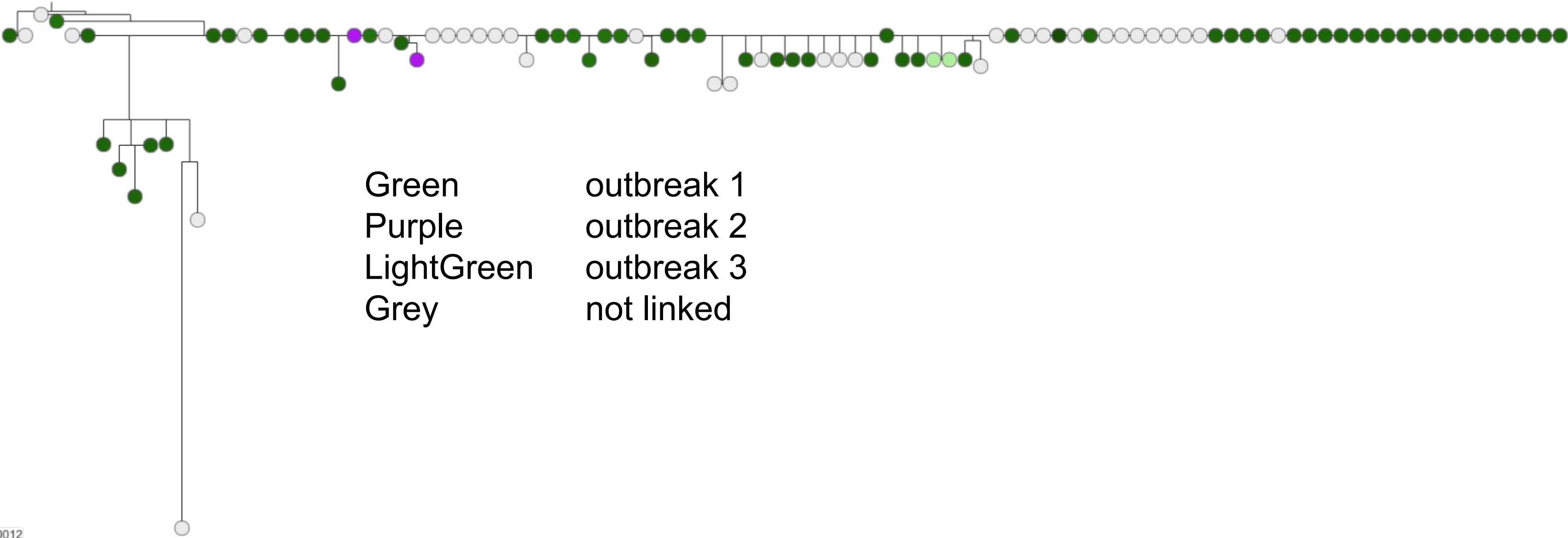
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Considerations:

Relative **low mutation** rate, epidemiology is especially important for outbreak detection, establishing linkages, and define outbreaks



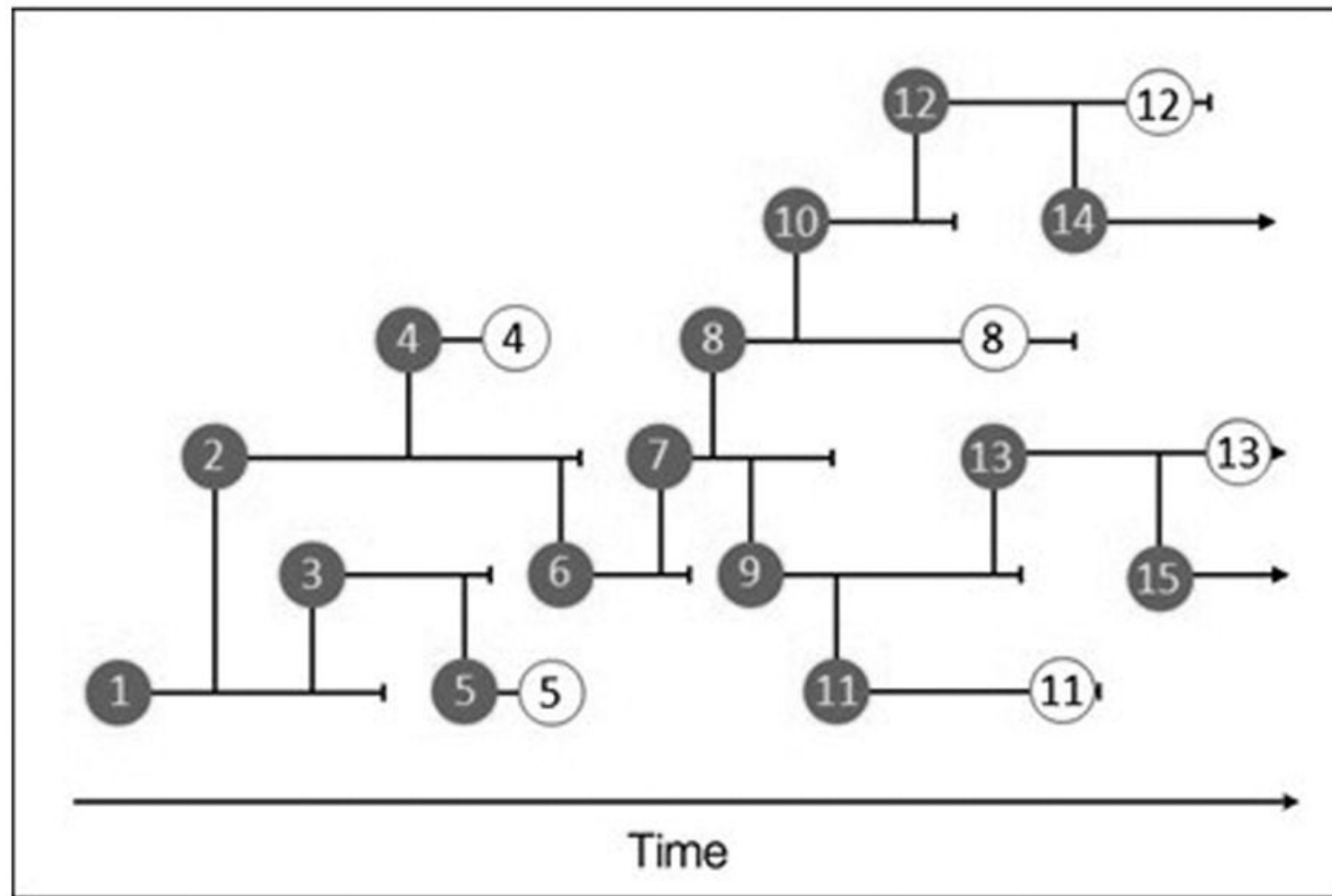
**wellcome**  
**connecting**  
**science**



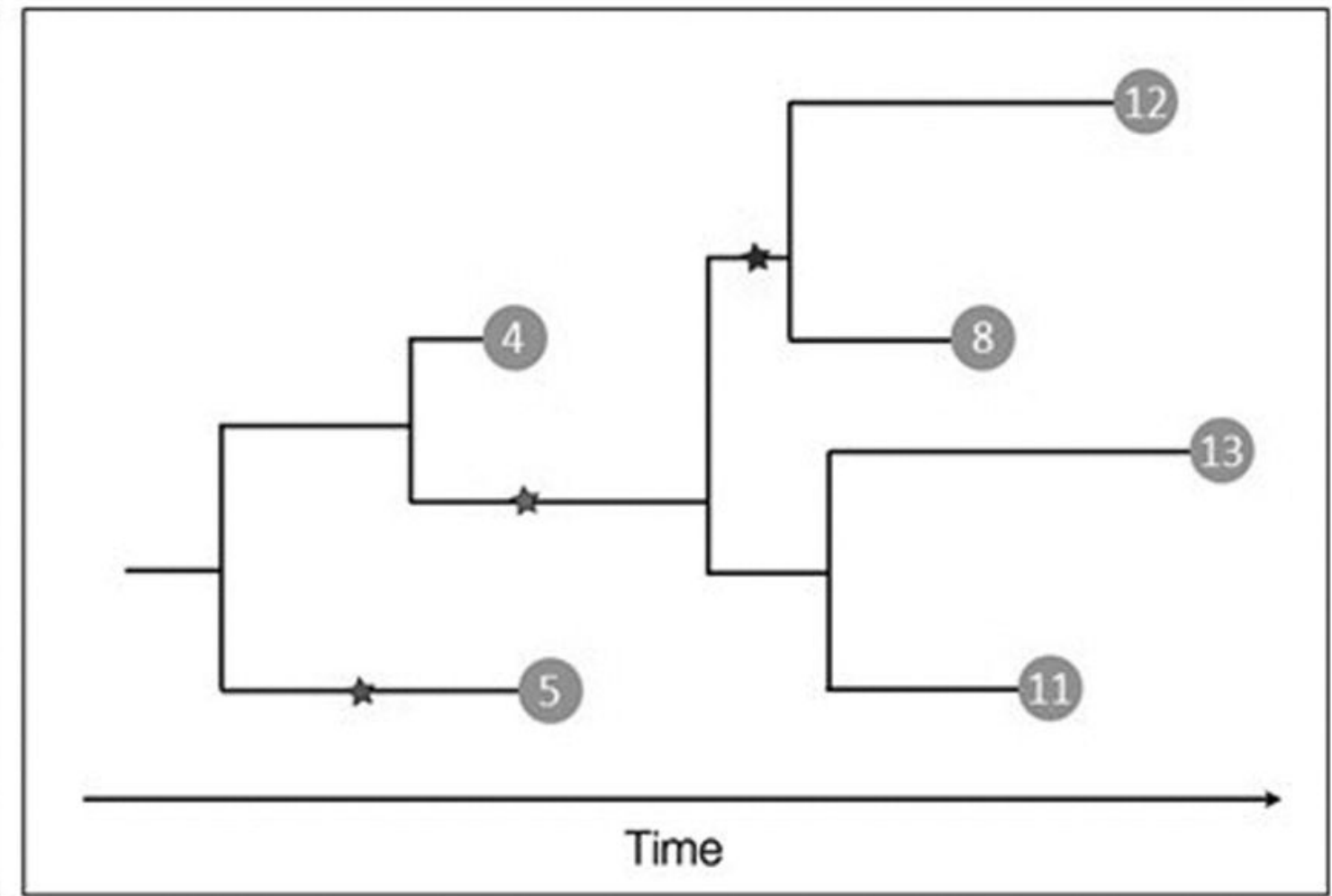
**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

# Considerations:

**missing transmission events** means cause-effect and direction cannot be certain



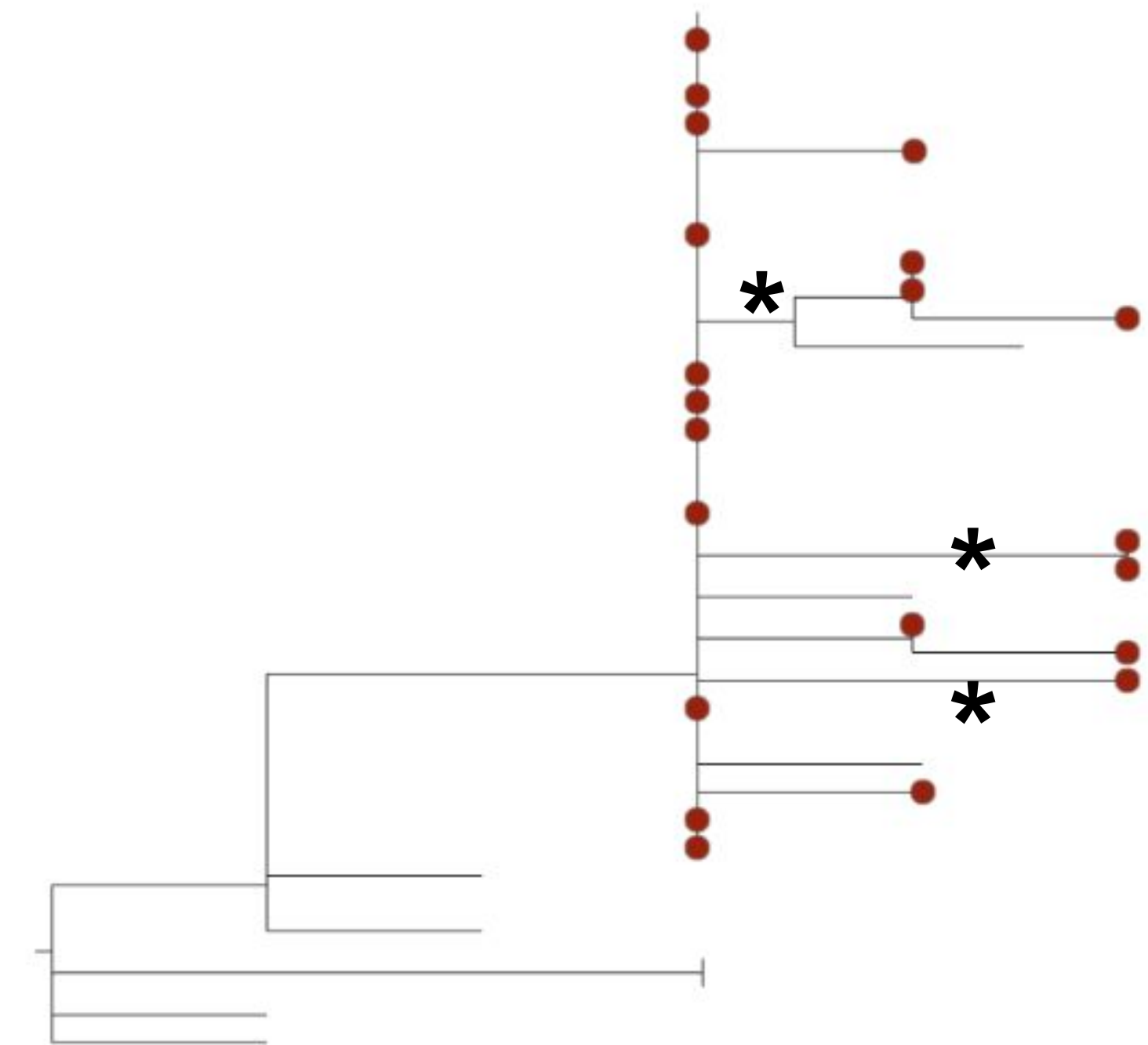
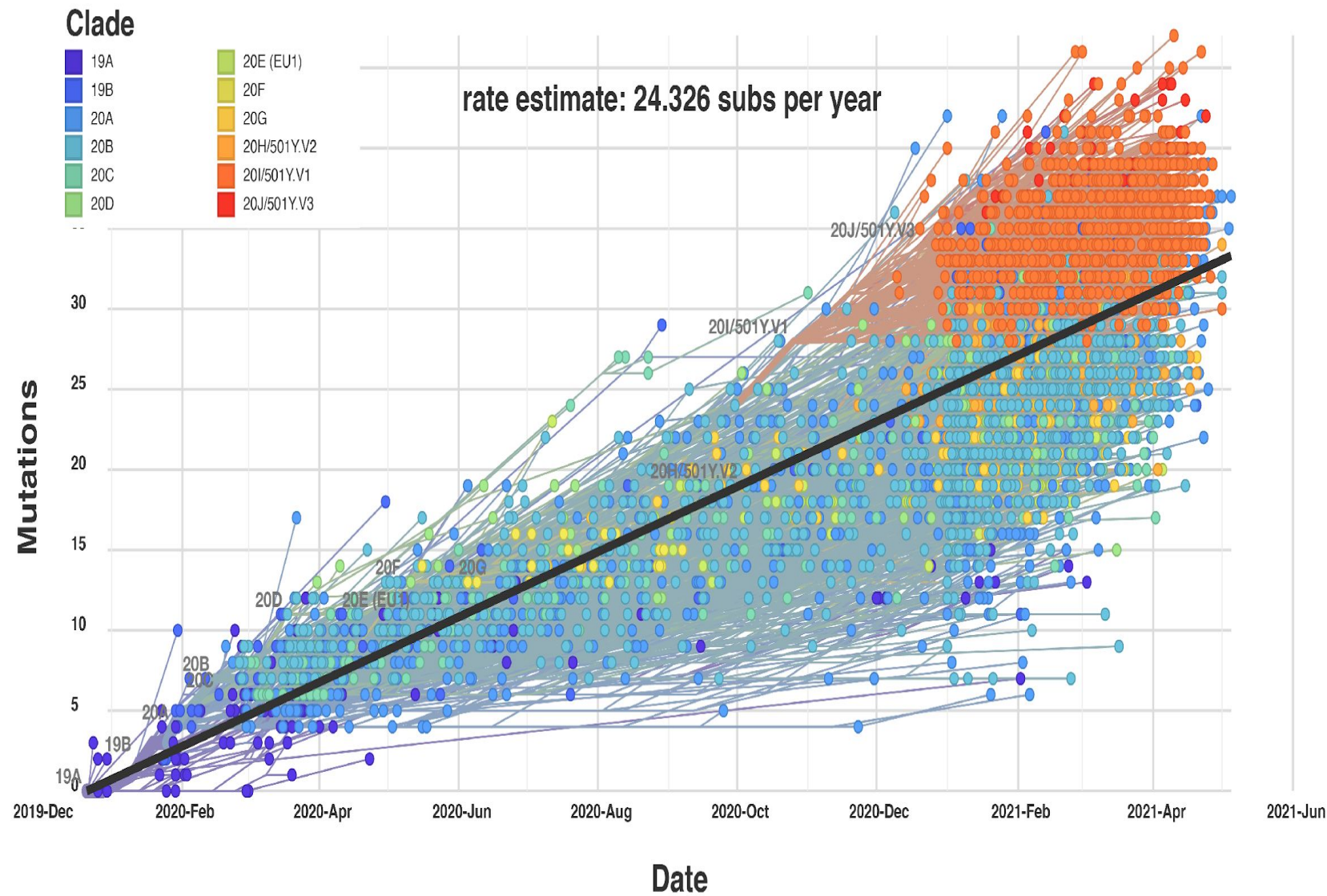
Actual transmission tree  
15 infections  
6 samples



Phylogenetic tree  
Based on 6 sequences

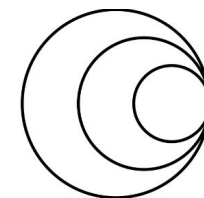
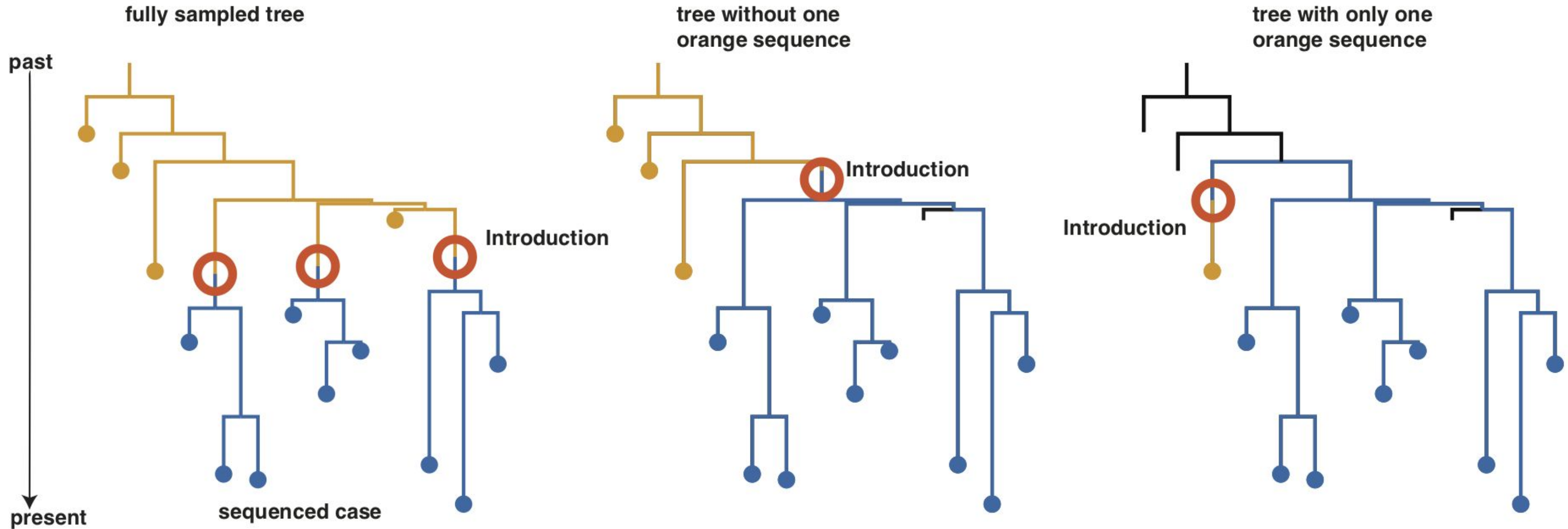
# Signature of missing events

Longer than expected branches over a short amount of time can be a signal for missing events



# Considerations:

**Poor sampling** means cautions is needed when interpreting geographical origin and number of introductions



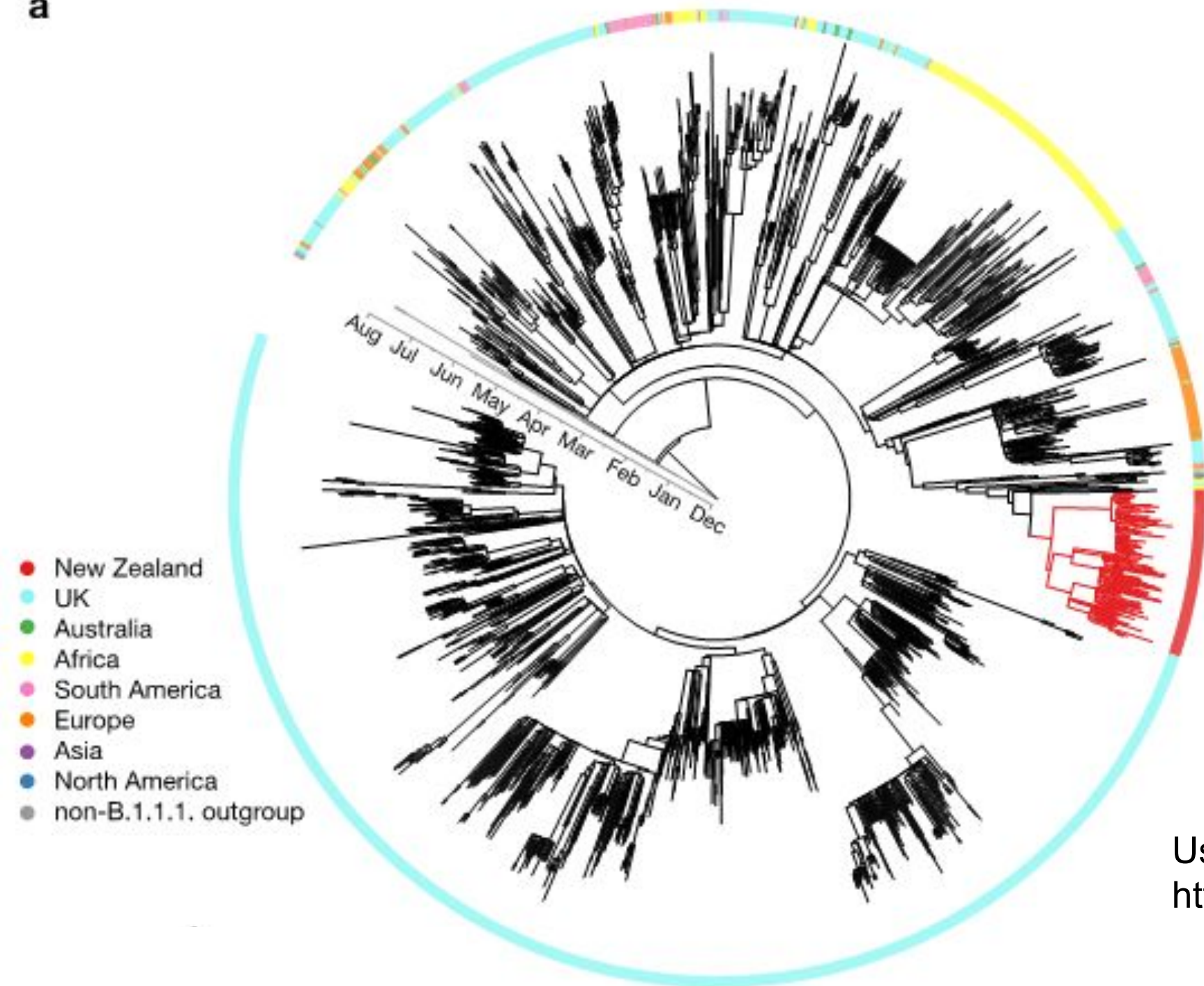


# Considerations:

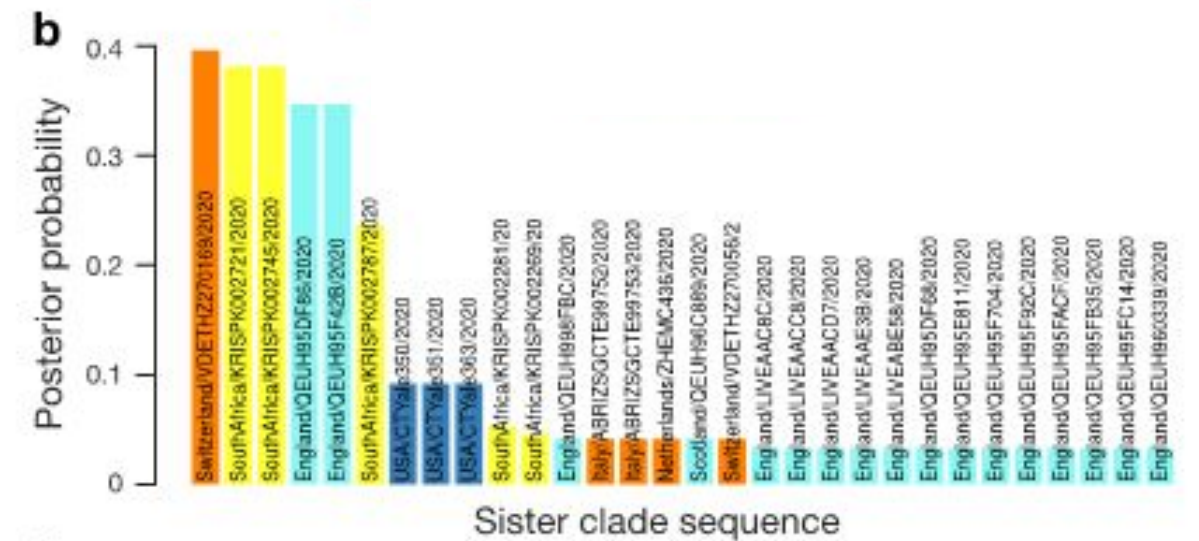
**Poor sampling** means cautions is needed for interpreting geographical origin

Biases in international data

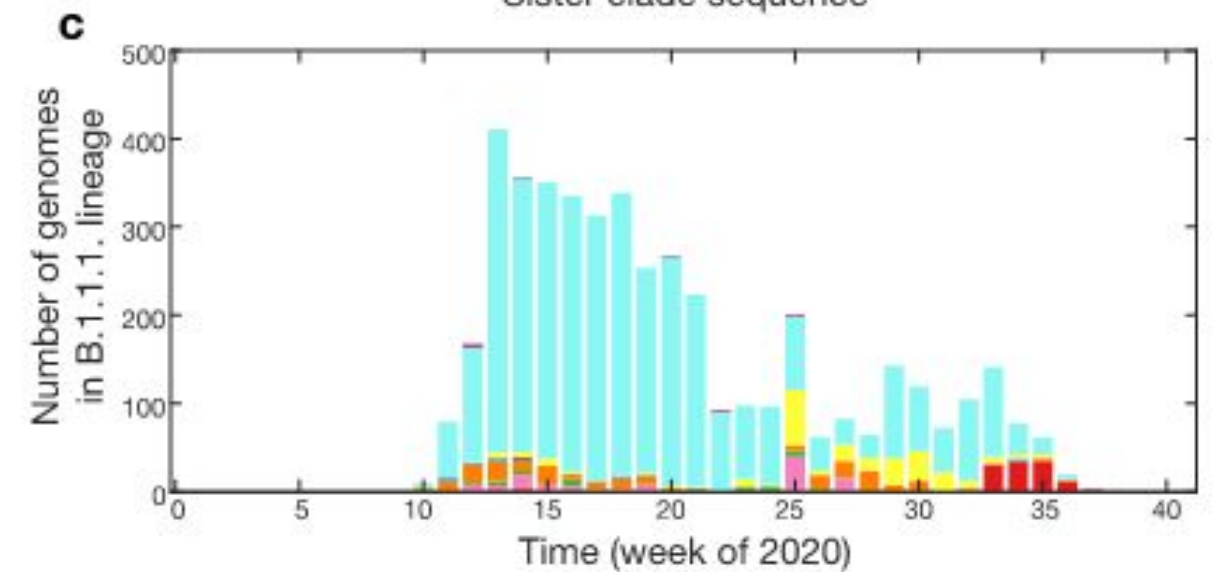
a



b



c



Use of Genomics to Track Coronavirus Disease Outbreaks, New Zealand  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8084492/>



# Phylodynamics: combine phylogeny, epidemiology to uncover hidden patterns

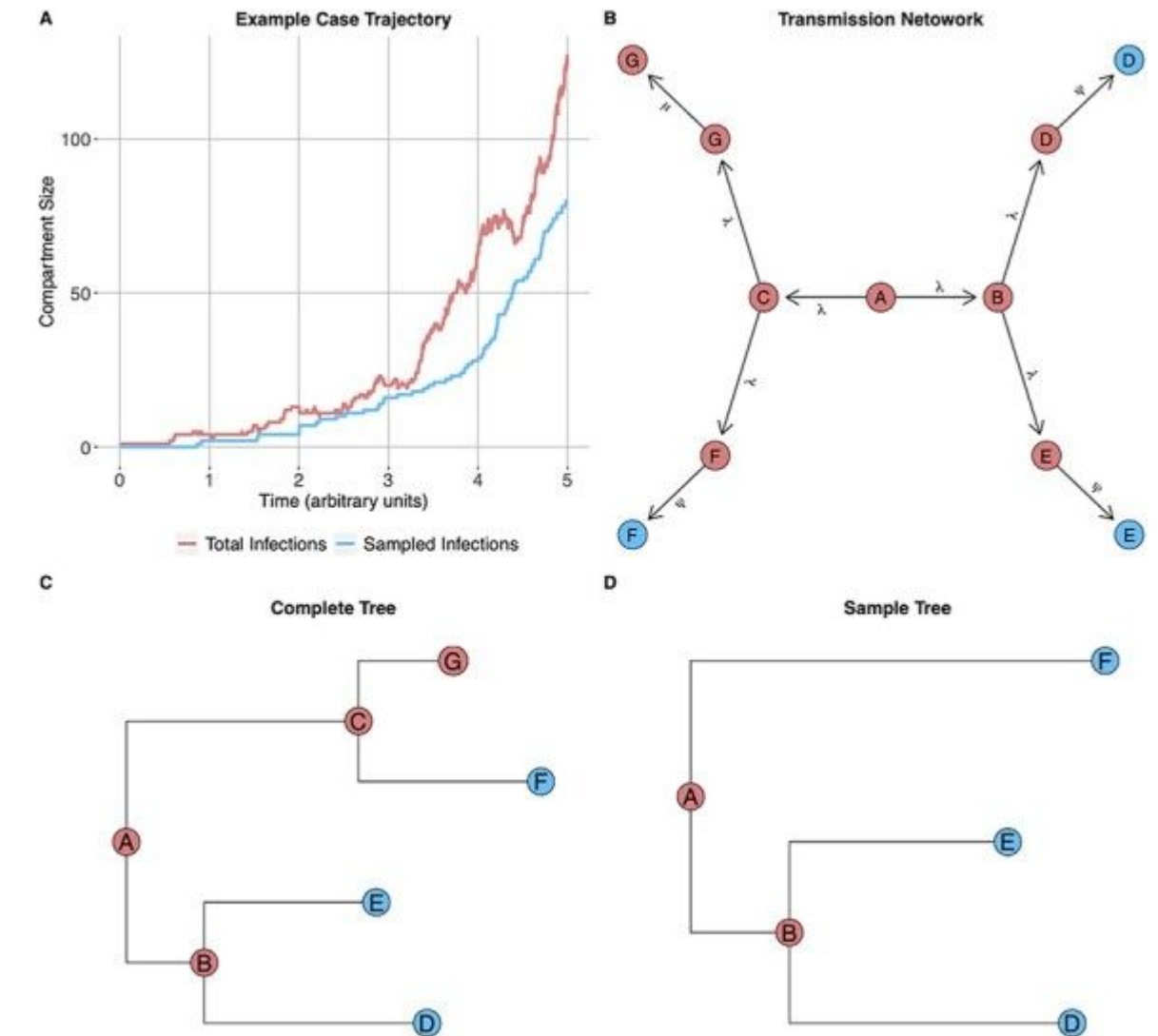
Incorporates model of pathogen epidemiological dynamics, model of evolution via timed phylogeny, and statistic inference to:

- Estimate transmissibility such as  $R_0$  and  $R_e$
- Estimate missing cases and population changes
- Estimate geographical origin and spread

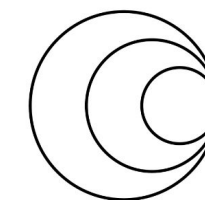
## Resources:

Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9241095/>

Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic, <https://pubmed.ncbi.nlm.nih.gov/35459859/>



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9241095/>



**wellcome  
connecting  
science**



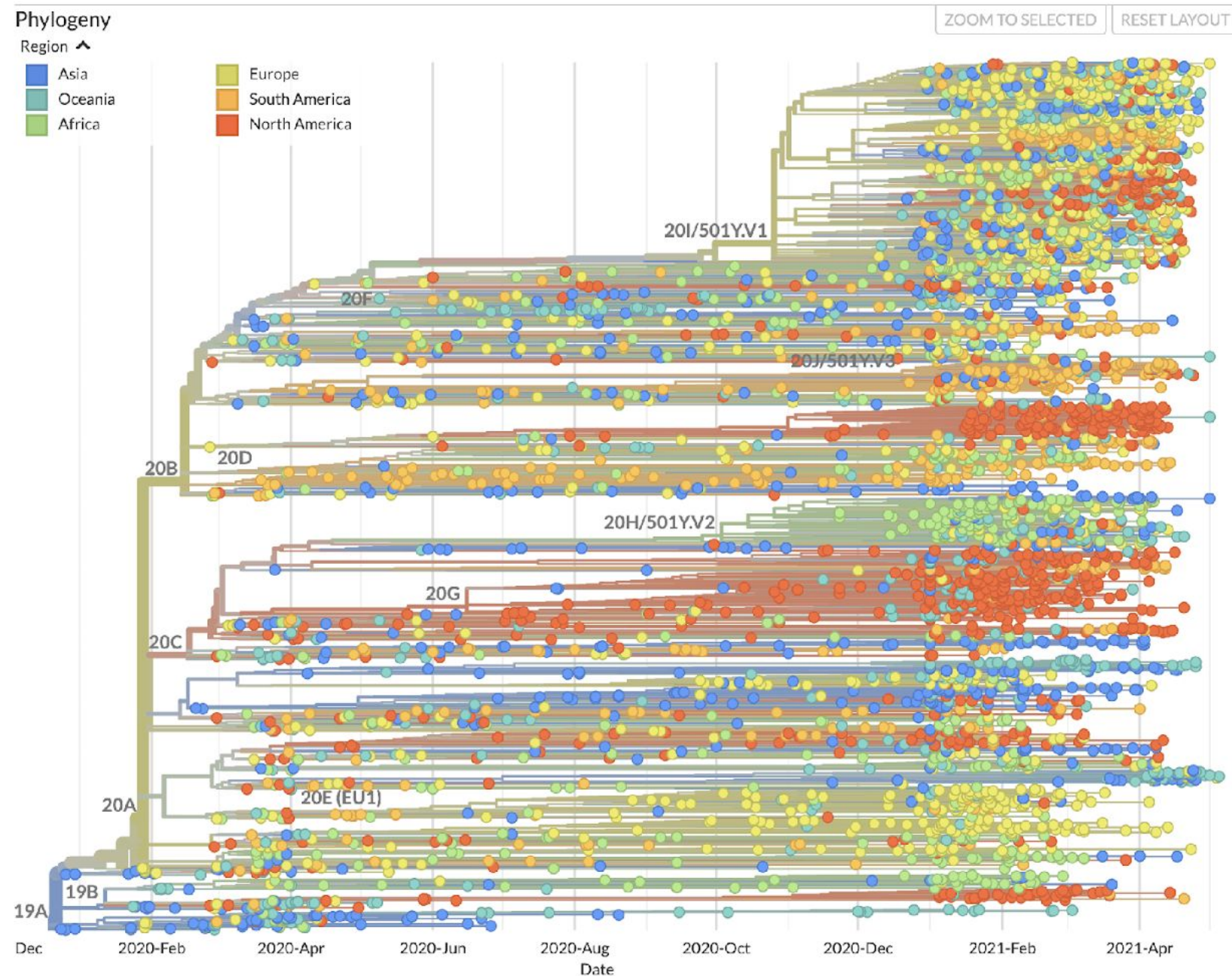
**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# SARS-CoV-2 Phylodynamics

## Genomic epidemiology of novel coronavirus - Global subsampling

Built with [nextstrain/ncov](#). Maintained by the [Nextstrain team](#). Enabled by data from [GISAID](#).

Showing 3825 of 3825 genomes sampled between Dec 2019 and May 2021.



# Genomic epi: visualisation and analysis

<https://microreact.org/>

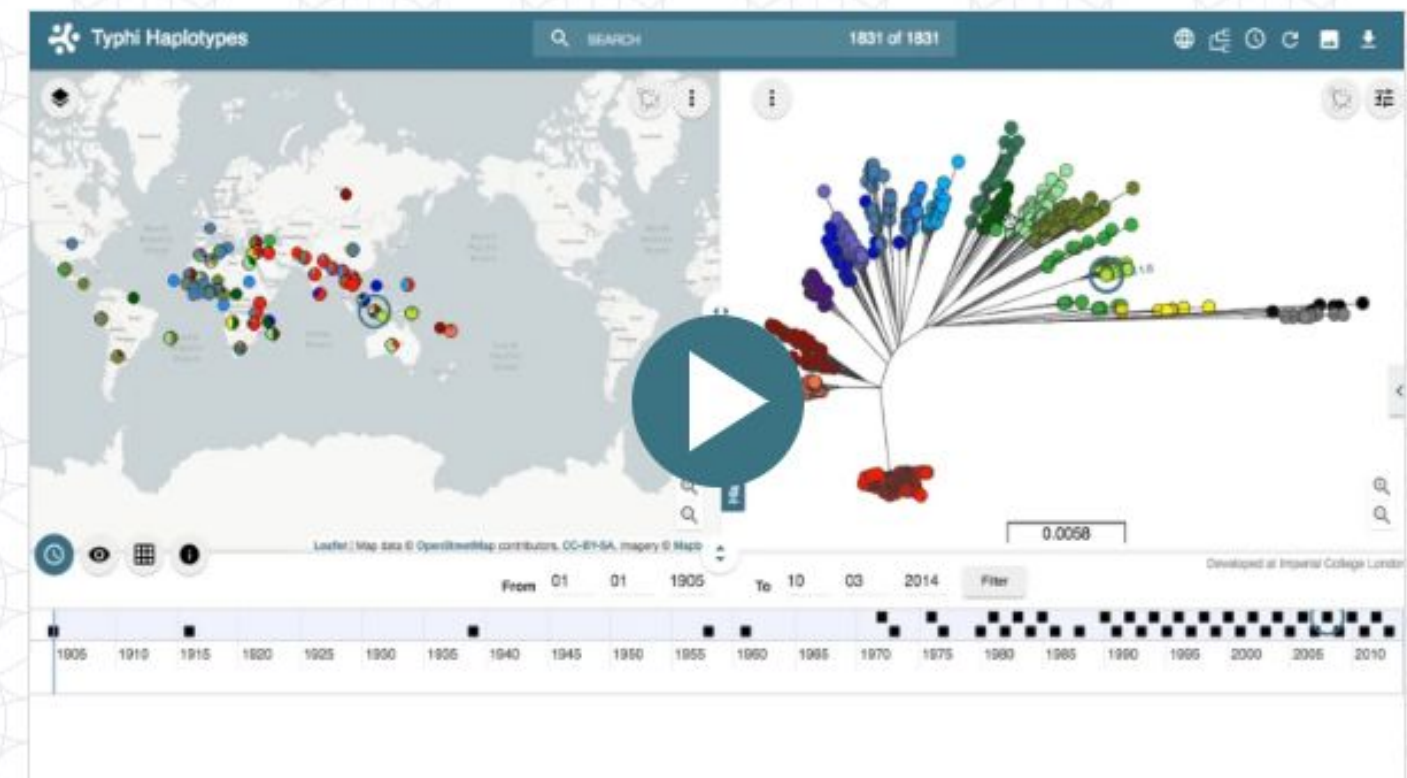
Interactive tree

Annotation

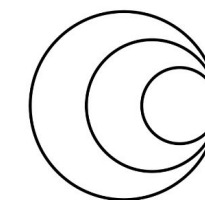
Network diagrammes

Timeline

Open data visualization and sharing for genomic epidemiology



IN PARTNERSHIP WITH



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Genomic epi: visualisation and analysis

Interactive timed phylogeny

Analysis tools such as nextclade and nextalign

Data communication: narrative tool

<https://nextstrain.org/community/narratives/ESR-NZ/GenomicsNarrativeSARSCoV2/aotearoa-border-incursions>



[HELP](#) [DOCS](#) [BLOG](#) [LOGIN](#)

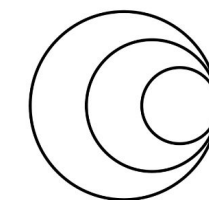
## Nextstrain

### Real-time tracking of pathogen evolution

Nextstrain is an open-source project to harness the scientific and public health potential of pathogen genome data. We provide a continually-updated view of publicly available data alongside powerful analytic and visualization tools for use by the community. Our goal is to aid epidemiological understanding and improve outbreak response. If you have any questions, or simply want to say hi, please give us a shout at [hello@nextstrain.org](mailto:hello@nextstrain.org).

[READ MORE](#)

<https://docs.nextstrain.org/en/latest/learn/interpret/index.html>



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Section 3: Genomic data sharing

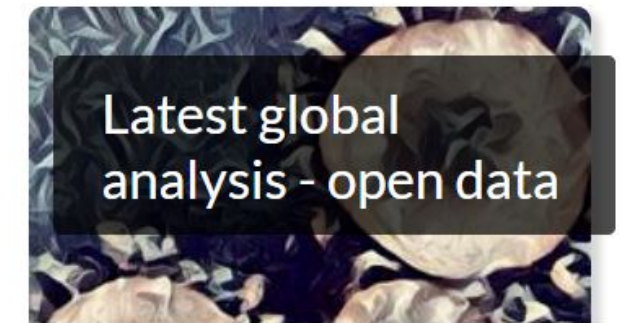
Data sharing is crucial for genomic surveillance and epidemiology

Data sharing enables comparisons between cases

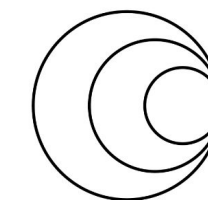
Data sharing informs and drives public health changes

## SARS-CoV-2 (COVID-19)

We are incorporating SARS-CoV-2 genomes as soon as they are shared and providing analyses and situation reports. In addition we have developed a number of resources and tools, and are facilitating independent groups to run their own analyses.



[SEE ALL RESOURCES](#)



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Why share genomic data ?

Data sharing is important for:

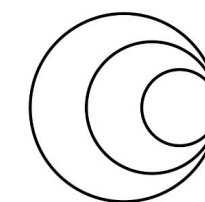
**Reproducibility**

**Adherence to FAIR principles**

**Collaboration**

**Data discovery - e.g. improved cross referencing and data linking**

**Advancing scientific discovery - e.g. vaccine development**



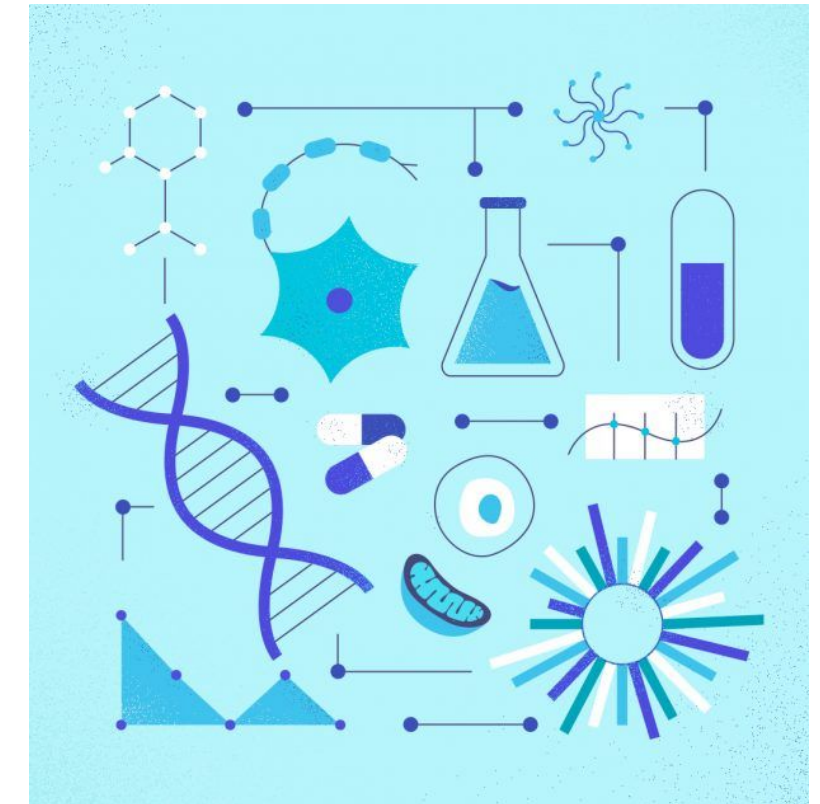
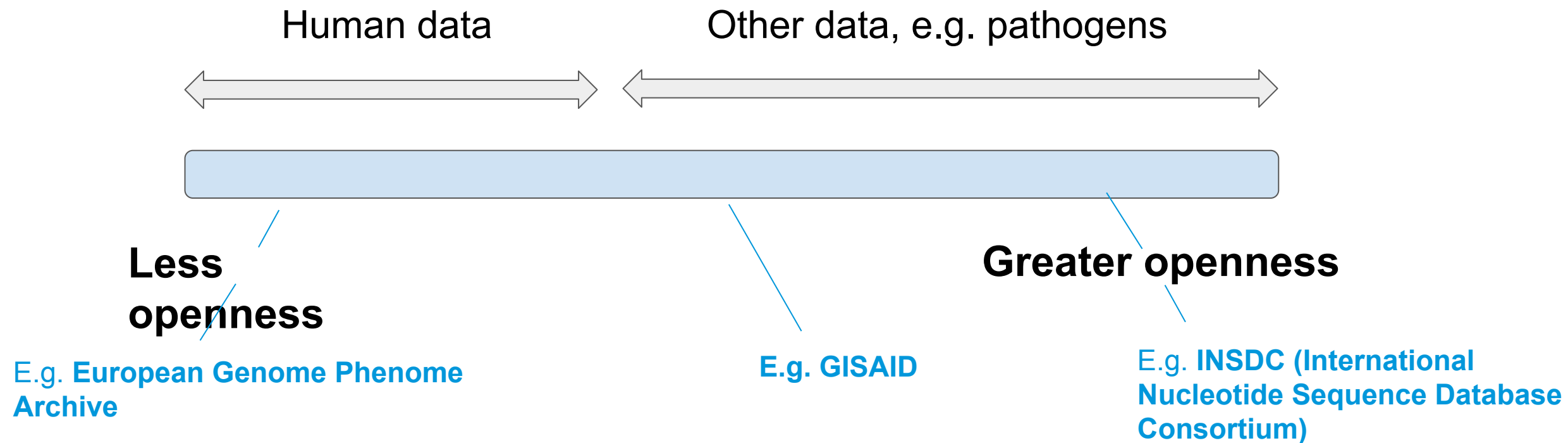
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

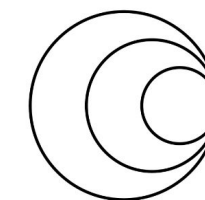
# Genomic databases

Many different open nucleotide sequence repositories, local and international, with different levels of data access:



*Data should be “as open as possible, as closed as necessary”*

Source: European Commission,  
Horizon2020 program



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**



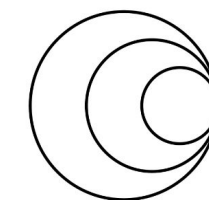
# Genomic data repositories

Public biological data repositories recommended by journals, the WHO, and other life sciences organisations (e.g. ELIXIR):

*SARS-CoV-2 data also shared here*

Data types	Repository options	Data and metadata standards
Raw sequencing data (reads or traces)	INSDC	Browse data and metadata standards endorsed by the Genome Standards Consortium
Annotated sequences	INSDC	Browse data and metadata standards endorsed by the Genome Standards Consortium
Genome assemblies	INSDC GISAID	Browse data and metadata standards endorsed by the Genome Standards Consortium
Sample metadata	INSDC GISAID	Browse data and metadata standards endorsed by the Genome Standards Consortium
Genetic variation data	<a href="#">dbSNP</a> (human variations less than 50bp) <a href="#">dbVar</a> (human variations greater than 50bp) <a href="#">ClinVar</a> (human genotype & phenotype) <a href="#">European Variation Archive (EVA)</a> (all species)	

<https://www.nature.com/sdata/policies/repositories#nuc>



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Sharing SARS-CoV-2 data - GISAID

## Global Initiative on Sharing Avian Influenza Data

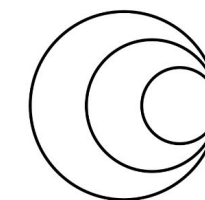
Established in 2008, first SARS-CoV-2 sequence shared in 2020. Now a popular SARS-CoV-2 data sharing platform

International database, but all users must abide by a data access agreement

Assembled sequence submissions only

The screenshot shows the GISAID website interface. The top navigation bar includes the GISAID logo and links for 'About us', 'Database Features', 'Events', 'Collaborations', 'Resources', 'Help', and 'Login'. The main content area is divided into several sections: 'In Focus' featuring a headline 'Omicron discovered on all seven continents' and a 3D structure of the spike protein; 'hMpxV Phylogeny' and 'Lineage comparison' sections with phylogenetic trees and a heatmap; 'Submission Tracker' and 'Tracking Variants' sections with maps and charts; and 'Frequency Dashboards' showing bar charts for hCoV-19, hMpxV, Influenza, and RSV. A 'GISAID Resources' section at the bottom right includes a 'Data Acknowledgement Locator' search box.

<https://gisaid.org/>



wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

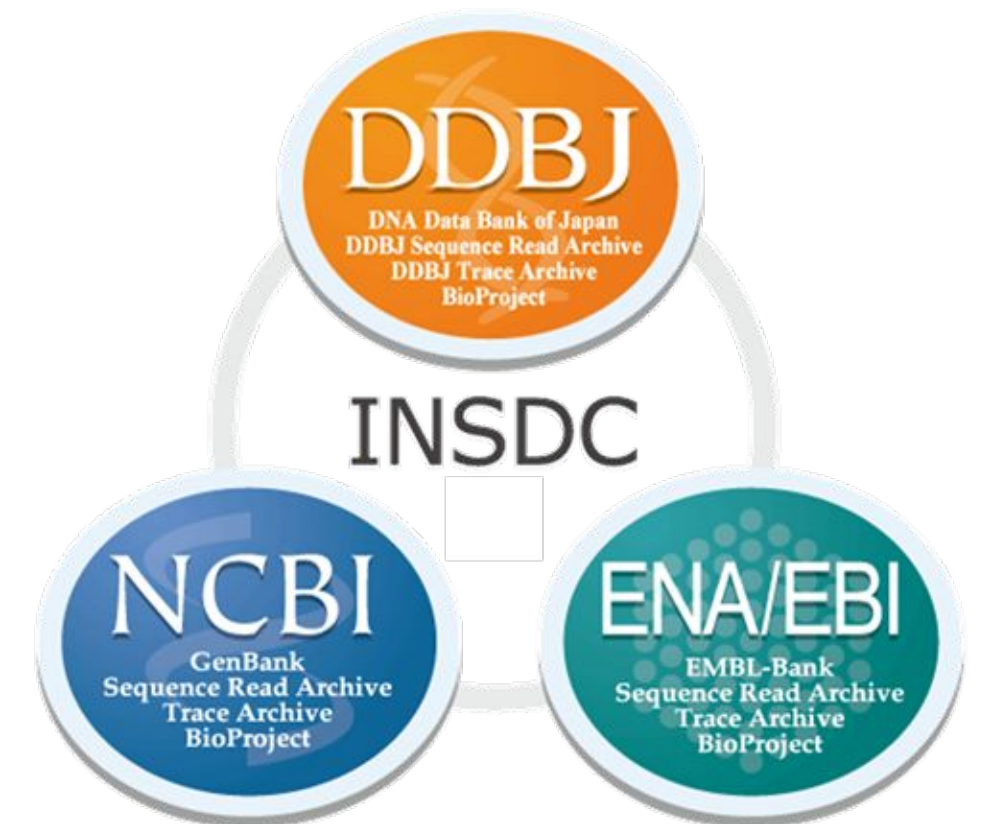
# Sharing SARS-CoV-2 data - ENA



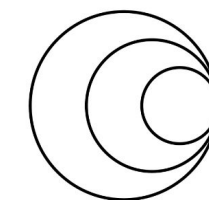
European Nucleotide Archive (European arm of INSDC) - data is mirrored between all 3 nodes

ENA and INSDC established in 1980s to create a central repository for increasing volumes of genetic data

International open access repository covering **raw sequence data**, sequence assembly information and functional annotation for all non-human organisms



International Nucleotide Sequence Database Collaboration (INSDC)



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# What is open access?

Free to deposit and download non-human data and metadata

Users do not need to be verified for data download

No restrictions on re-sharing submitted data, e.g. feeding data in to custom analysis tools

No policy to restrict user access rights

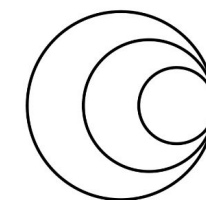
## ***What open access is *\*not\**:***

Records that do not reference original submitting/collecting institutes

All data must become public immediately

## ***Food for thought:***

Please ensure that metadata provided follows data protection laws in your region and data is human read cleaned

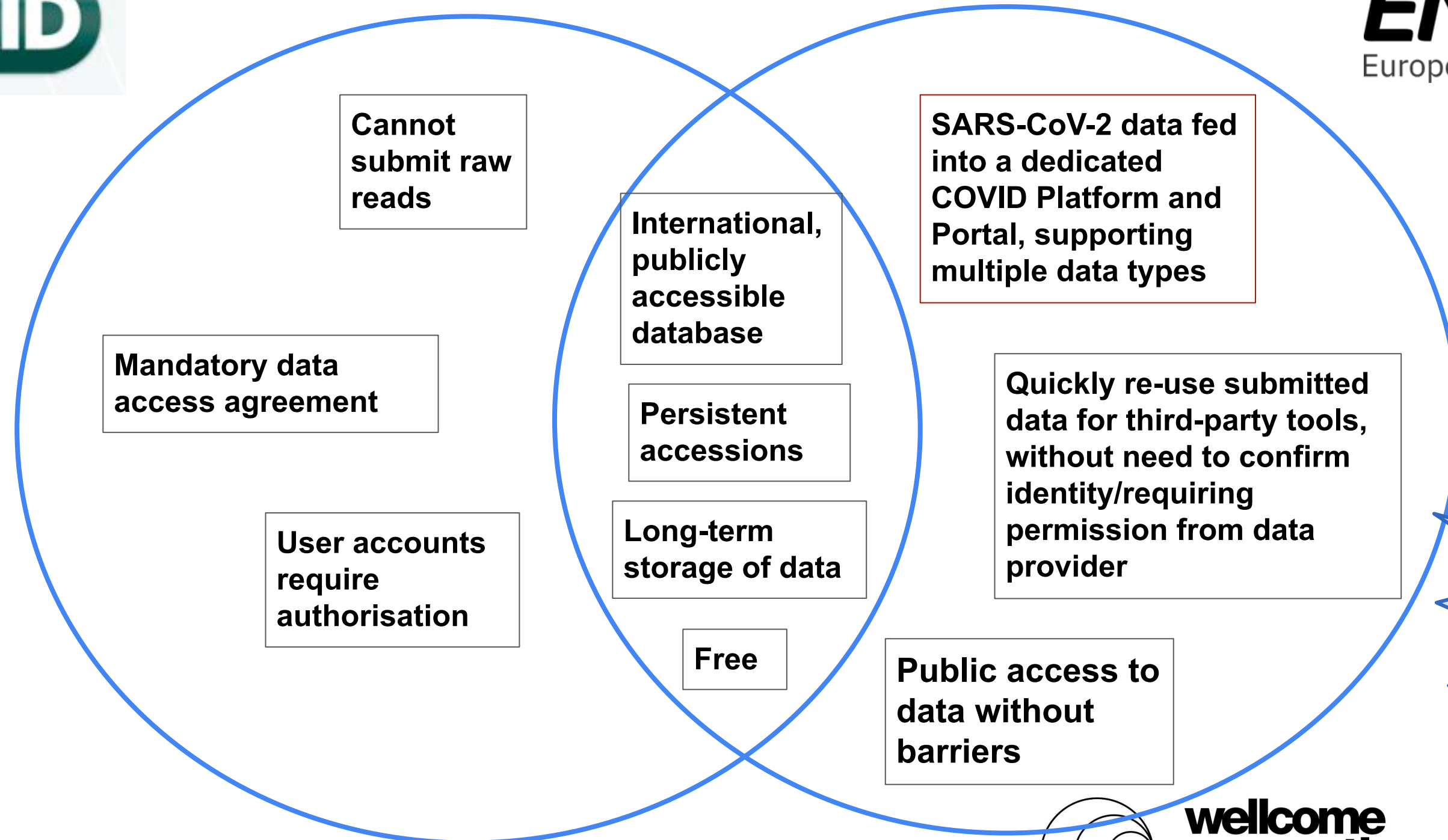


**wellcome  
connecting  
science**

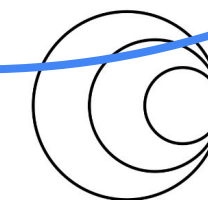


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# GISAID & ENA databases - a comparison



**You should submit data to both!**



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Dual sample submission to GISAID & ENA

- **‘GISAID Accession ID’** sample attribute links GISAID assemblies to ENA submission, via ENA sample
- **GISAID to ENA xml/xls sample converter**
  - Can use custom GISAID<->ENA field mapping file if desired
  - ‘GISAID Accession ID’ user-defined ENA attribute included by default

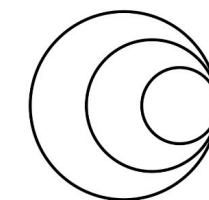
## Input options

```
-h, --help          show a help message and exit
--csv CSV           path to input file (CSV format)
--xls XLS           path to input file (Excel format)
--out OUT           output file name
--outformat (xml|excel) xml or excel output
--taxon TAXON       (optional) taxon name or id of samples (default: detect from GISAID sheet)
--map FILE           (optional) path to custom metadata mapping (default: ./metadata_mapping.tsv)
--sheet SHEET       (optional) name of excel sheet (default: 'Submissions')
```

## Examples

```
# convert GISAID spreadsheet in CSV format to ENA in excel format
gisaid_to_ena.py --csv gisaid.csv --outfile ena.xlsx --outformat excel
# convert GISAID metadata from sheet called 'Samples' to ENA spreadsheet
gisaid_to_ena.py --xls gisaid.xlsx --sheet Samples --outfile ena.xml --outformat xml
# convert using a custom metadata mapping file
gisaid_to_ena.py --xls gisaid.xlsx --outfile ena.xml --outformat xml --map path/to/mapping.tsv
```

[https://github.com/enasequence/ena-content-dataflow/tree/master/scripts/gisaid\\_to\\_ena](https://github.com/enasequence/ena-content-dataflow/tree/master/scripts/gisaid_to_ena)

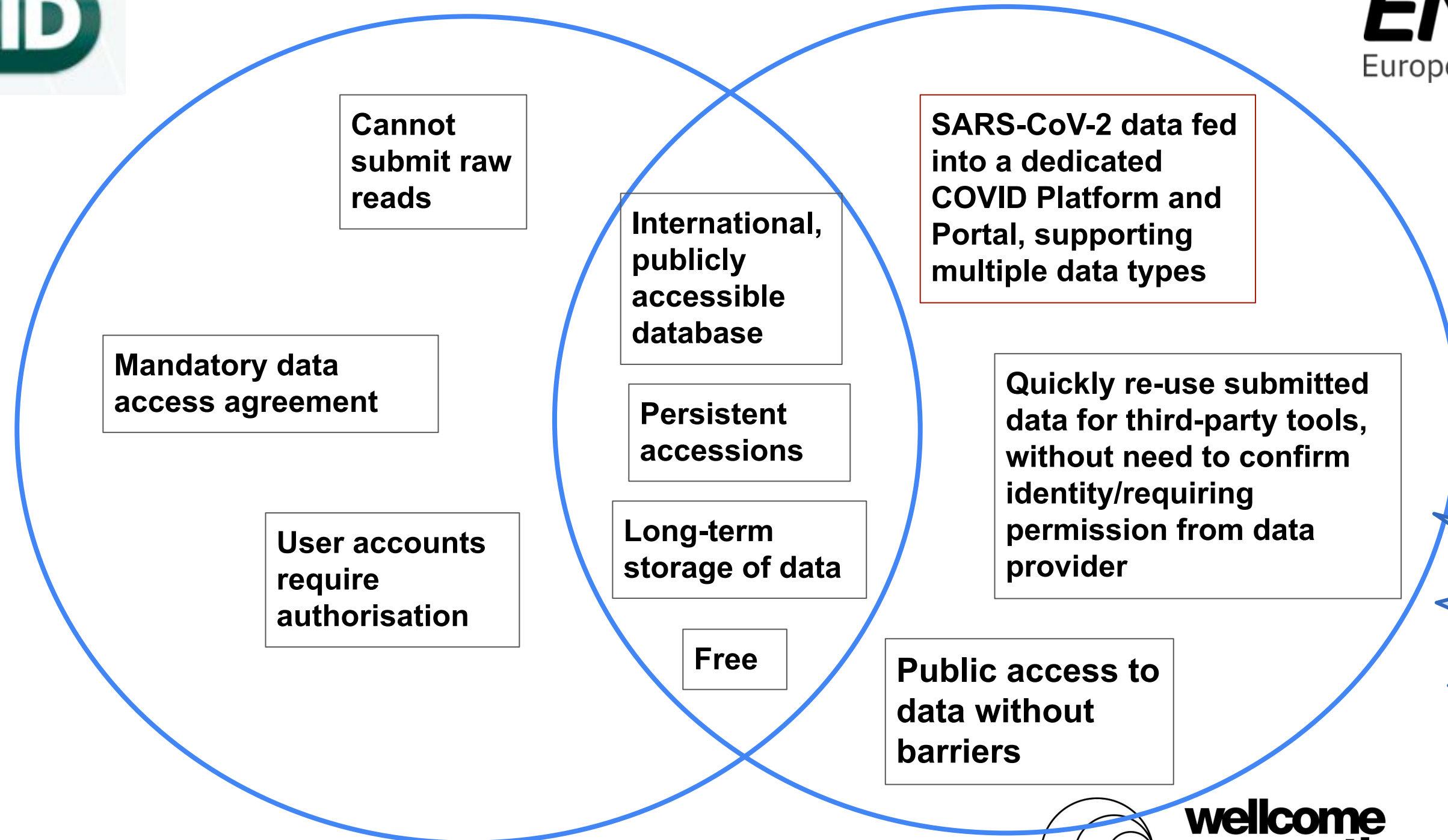


**wellcome  
connecting  
science**

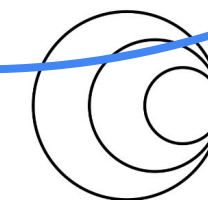


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# GISAID & ENA databases - a comparison



**You should submit data to both!**

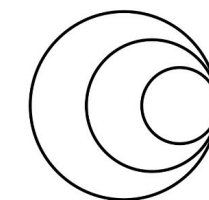


**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Section 4: European COVID-19 Data Platform



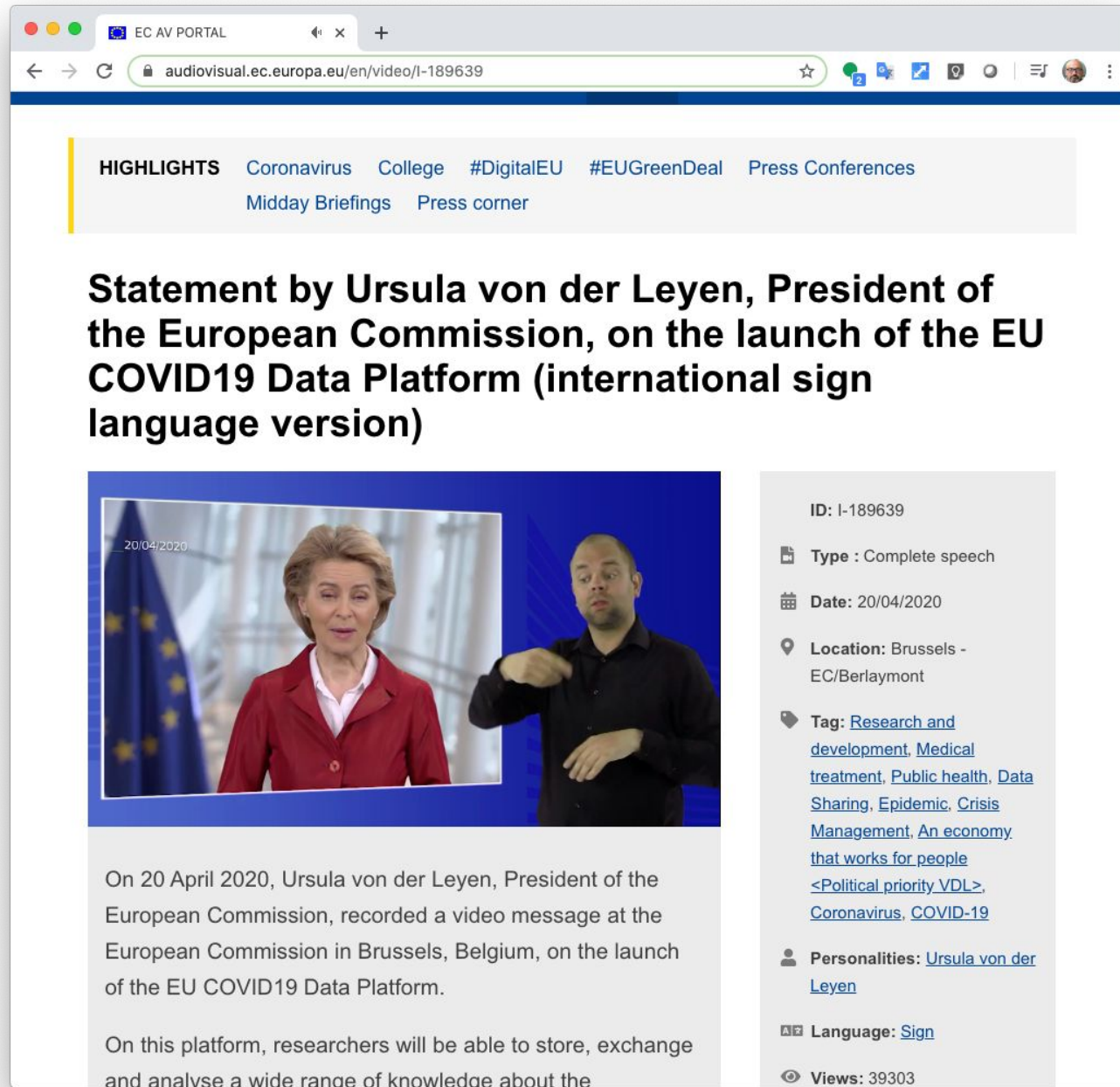
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**



# The European COVID-19 Data Platform



The screenshot shows a web browser window with the URL [audiovisual.ec.europa.eu/en/video/I-189639](https://audiovisual.ec.europa.eu/en/video/I-189639). The page features a navigation bar with links for 'HIGHLIGHTS', 'Coronavirus', 'College', '#DigitalEU', '#EUGreenDeal', 'Press Conferences', 'Midday Briefings', and 'Press corner'. The main content area displays the title '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 title is a video player showing Ursula von der Leyen speaking, with a sign language interpreter on the right. To the right of the video player is a metadata sidebar with the following information: 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, and Views: 39303. Below the video player, there is a short text description: '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. On this platform, researchers will be able to store, exchange and analyse a wide range of knowledge about the'.

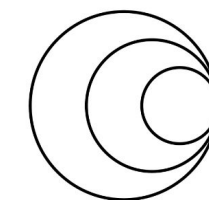
Launched Apr 2020

Open and rapid access to data, tools and workflows

Global data coverage and global access

Collaboration between EMBL-EBI and others

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



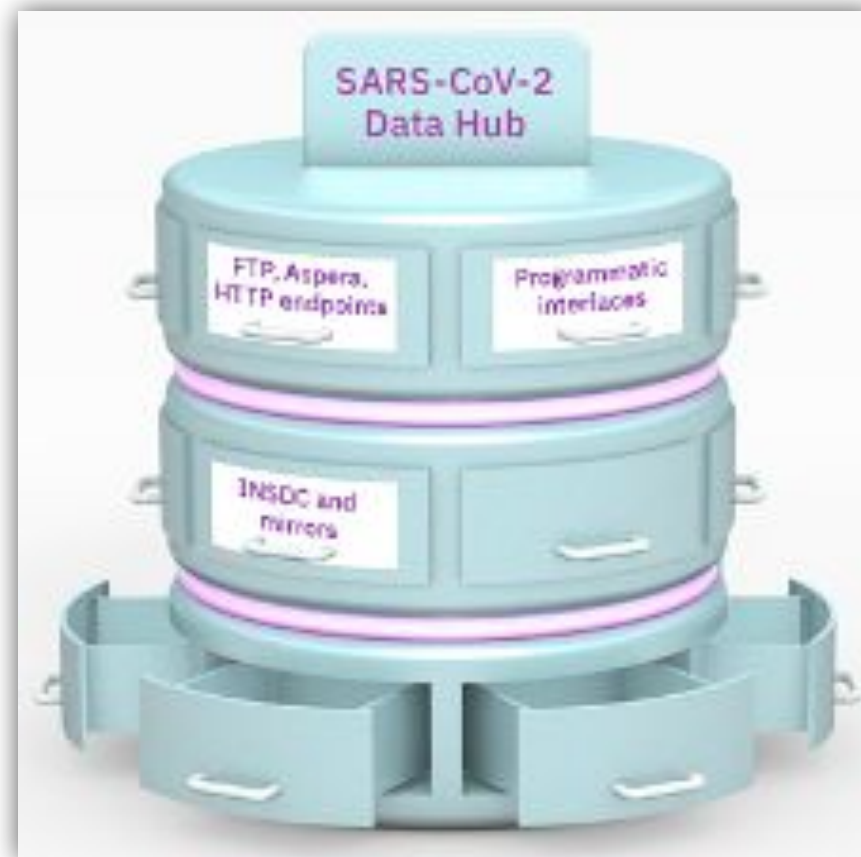
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Components

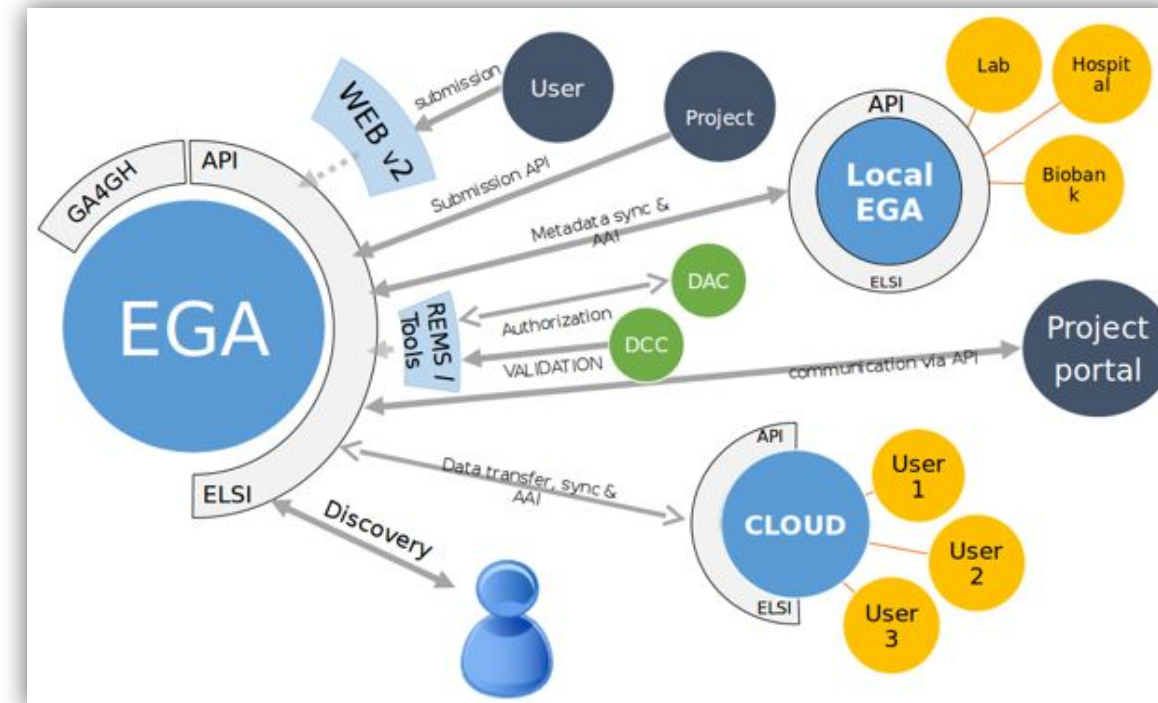
## 1. SARS-CoV-2 Data Hubs



Workspace enabling controlled access sharing of pre-publication sequence data

Tools for data analysis and visualisation

## 2. Federated European Genome-phenome Archive

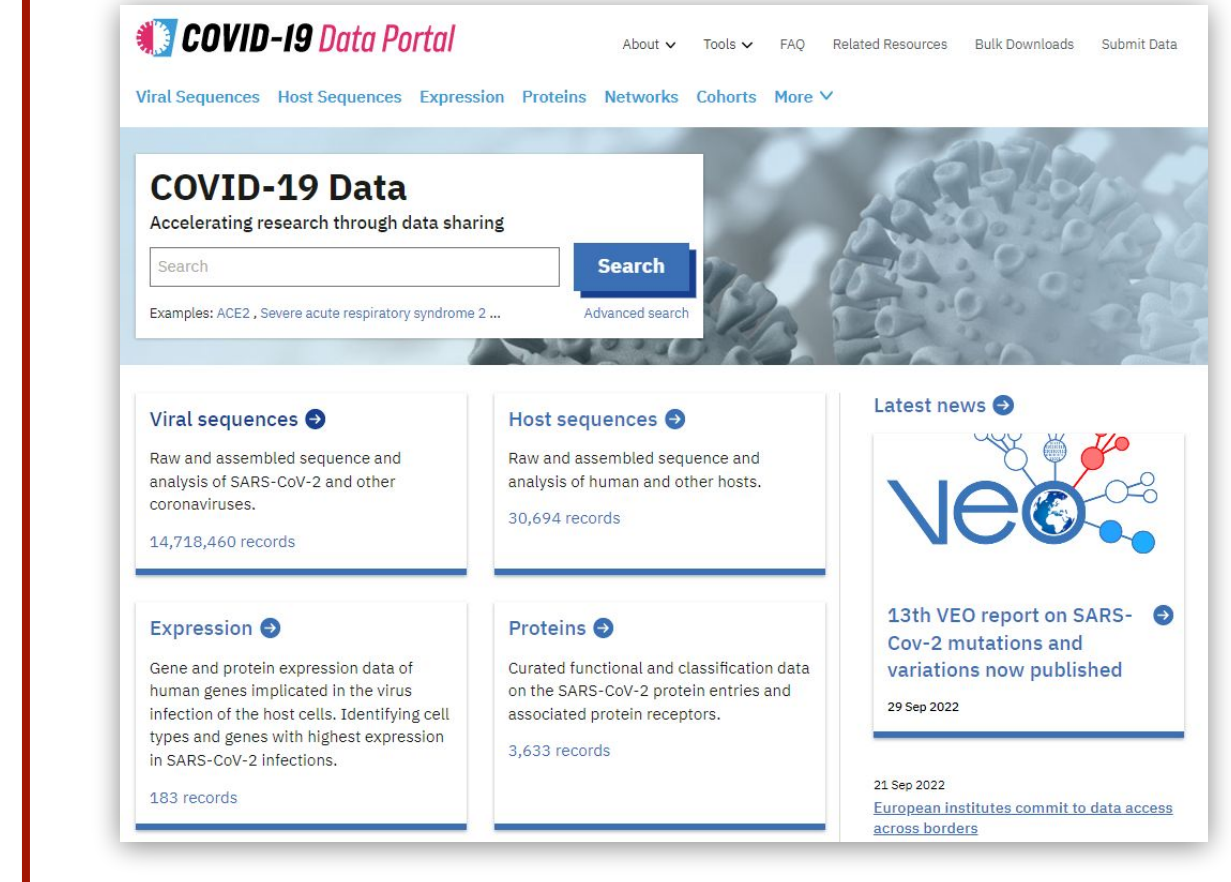


Support for sensitive human data

Restricted/controlled data access

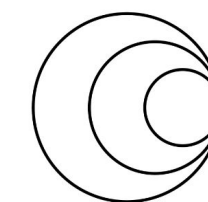
Federated data model

## 3. COVID-19 Data Portal



Central interface presenting a diverse range of COVID-19 related datasets, across [ELIXIR core deposition services](#)

Entry point for data sharing and visualisation tools

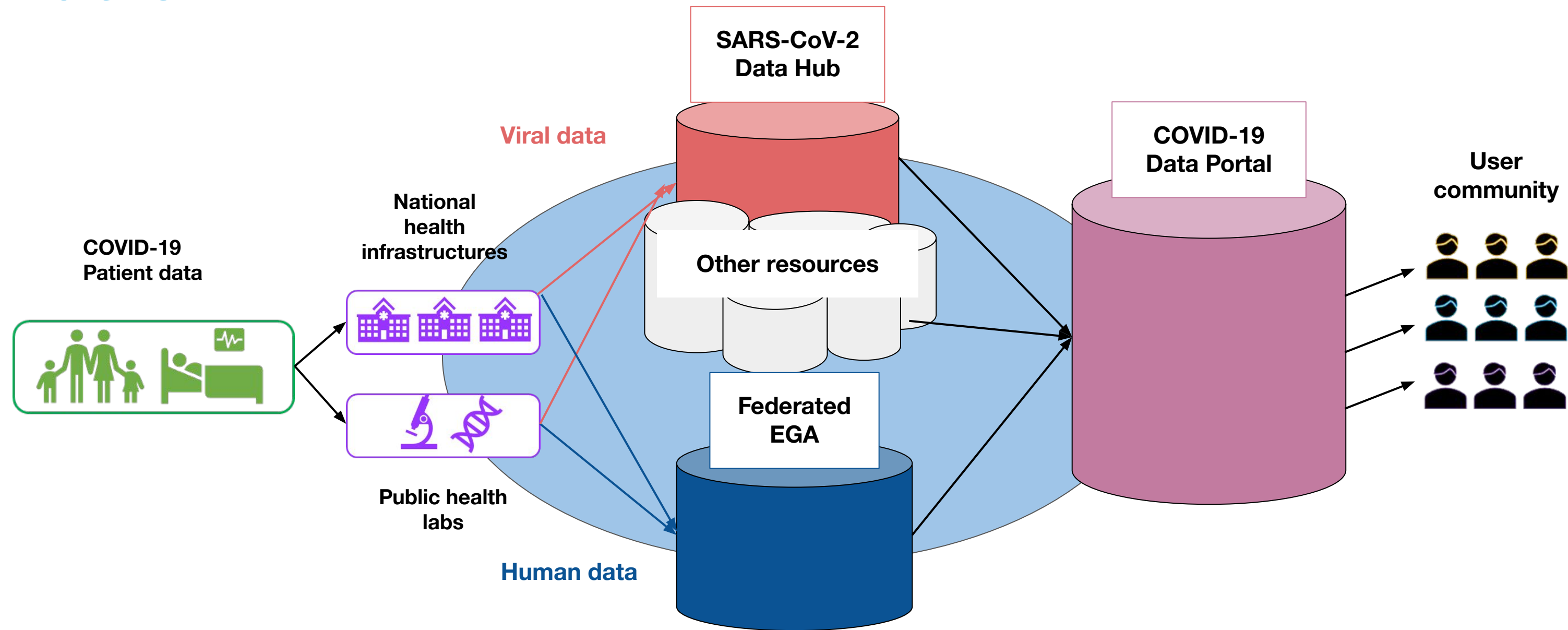


**wellcome  
connecting  
science**

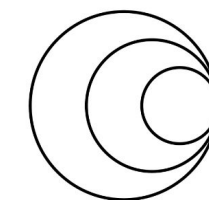


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Data flow through the COVID-19 Data Platform



Adapted from:  
<https://ec.europa.eu/newsroom/rtd/items/700623/en>

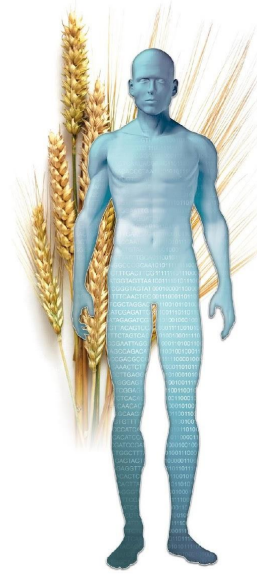
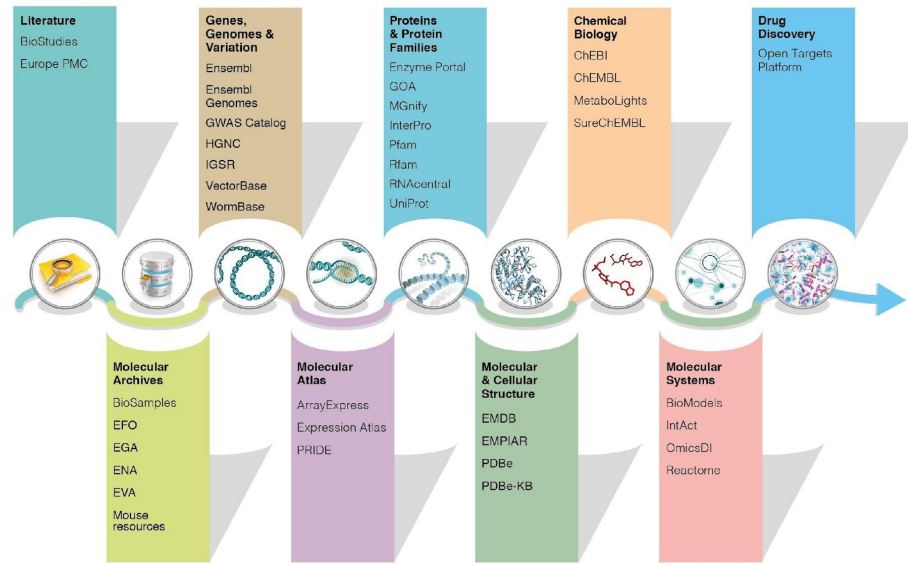


**wellcome**  
connecting  
science



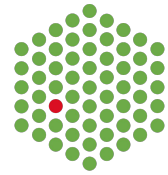
**COVID-19**  
GENOMICS  
GLOBAL TRAINING

# Foundations



<https://www.ebi.ac.uk/>

EMBL-EBI



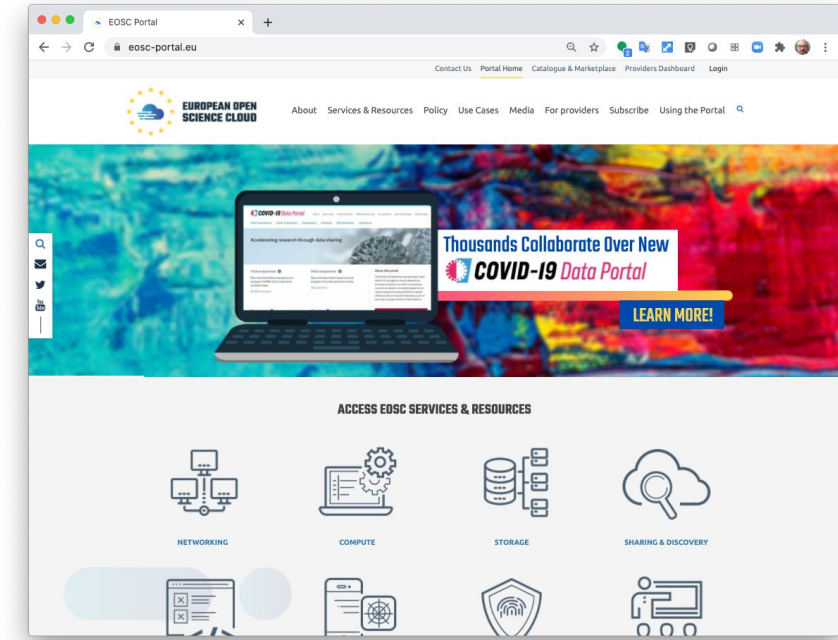
ELIXIR Members



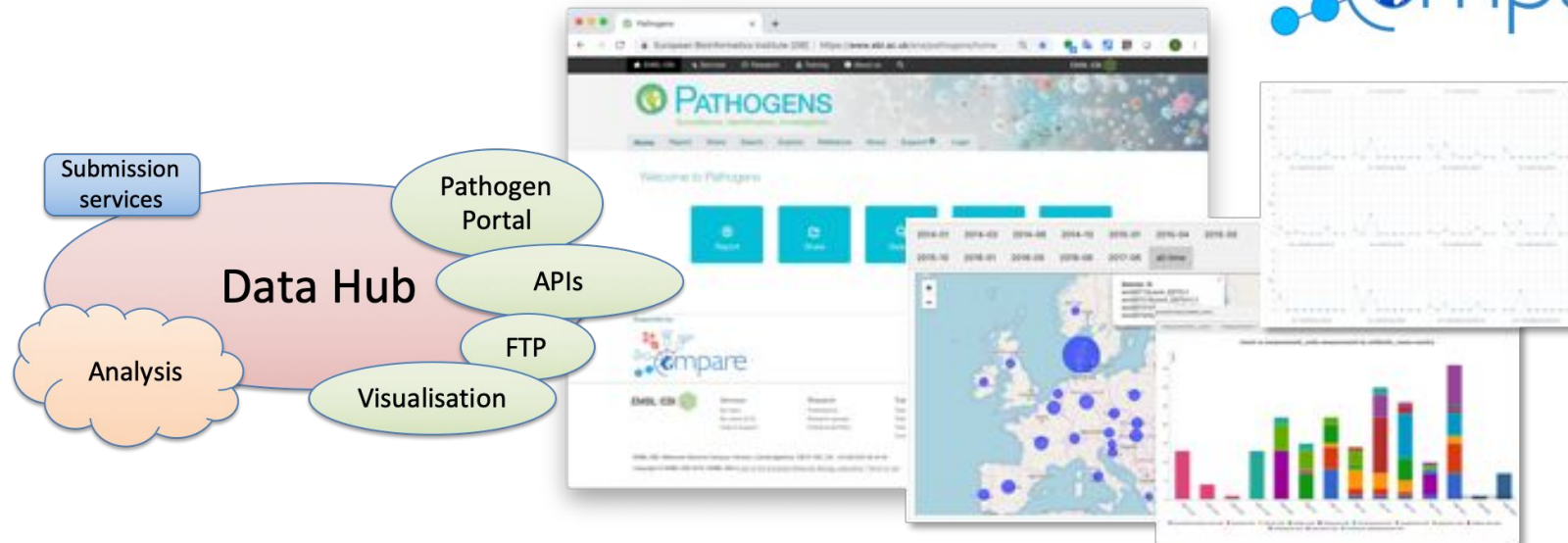
ELIXIR Observers



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



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



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>

Erasmus MC  
Erasmus Medical Centre, the Netherlands

National Institute for Public Health and the Environment (RIVM), the Netherlands

Eötvös Loránd University  
Eötvös Lorand University, Hungary

DTU  
Technical University of Denmark (DTU)

UNIVERSITÄTSKLINIKUM HEIDELBERG  
Universitätsklinikum Heidelberg, Germany



HORIZON 2020  
THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION



WELCOMING  
connecting science



COVID-19 GENOMICS GLOBAL TRAINING

# The COVID-19 Data Portal

Ease of access to a variety of COVID-19 related data types

E.g viral sequences, gene expression, protein structure, biological pathways, imaging data, literature, and more

Tools for data search and retrieval:

1. COVID Portal Advanced Search and API
2. Bulk Downloader tool

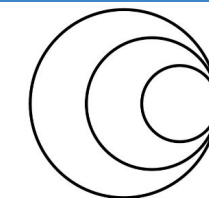
Tools for data visualization:

1. CoVEO variant browser
2. Phylogeny tree built from consensus sequences

The screenshot shows the COVID-19 Data Portal homepage. At the top, there is a navigation bar with links for 'About', 'Tools', 'FAQ', 'Related Resources', 'Bulk Downloads', and 'Submit Data'. Below this is a secondary navigation bar with links for 'Viral Sequences', 'Host Sequences', 'Expression', 'Proteins', 'Networks', 'Cohorts', and 'More'. The main content area features a search bar with a 'Search' button and a list of data categories. To the right of the search bar, there is a box with the text 'Findable Accessible Interoperable Reusable' and corresponding icons. Below the search bar, there is a table with the following data:

Type of Data	Total
Viral Sequences	14,927,387
Host Sequences	30,713
Expression	226
Proteins	3775
Biochemical networks	7801
Imaging	39
Literature	835, 297

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



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

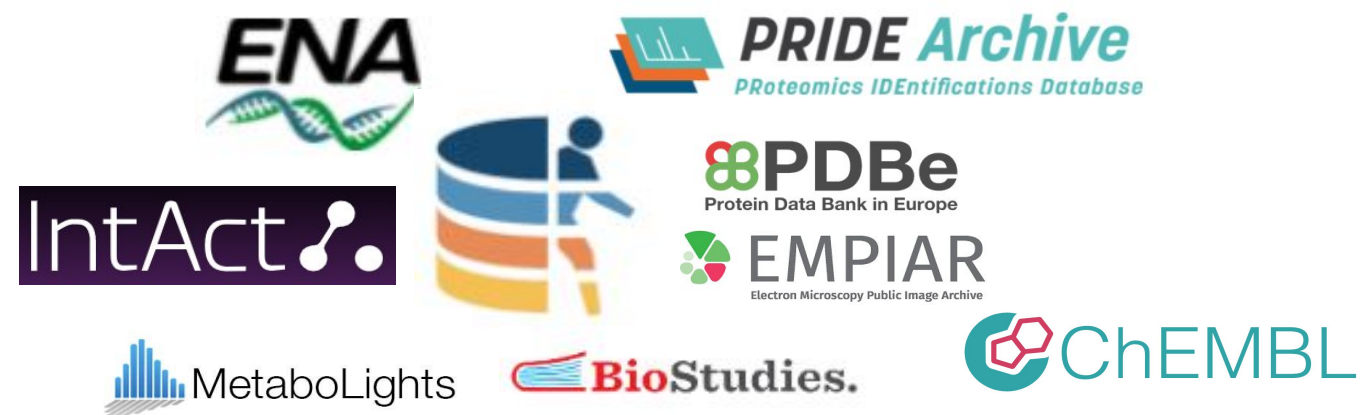
# The COVID-19 Data Portal - data submission

Data is not *submitted to the Portal itself...*

Data submission wizard (new!)  
guides users to the appropriate resource  
to submit their COVID-19 dataset

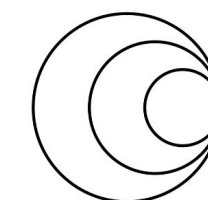
Spans 19 different datatypes

Different resources = different data  
submission methods



The screenshot shows the 'Submit new data' page of the COVID-19 Data Portal. The page features a navigation bar with links for 'About', 'Tools', 'FAQ', 'Related Resources', 'Bulk Downloads', and 'Submit Data'. Below the navigation bar, there are links for 'Viral Sequences', 'Host Sequences', 'Expression', 'Proteins', 'Networks', 'Cohorts', and 'More'. The main content area is titled 'Submit new data' and includes a question: '1 What kind of data would you like to submit?'. Below this question are five buttons: 'Non-human data', 'Human data', 'Combination of human and non-human data', 'Multiple datatypes from the same species', and 'Non-biological data'. Below the buttons, there is contact information: 'To enquire on how to collaborate on the European COVID-19 platform: [ecovid19@ebi.ac.uk](mailto:ecovid19@ebi.ac.uk). For further questions on sharing your data on the COVID-19 Data Portal: [virus-dataflow@ebi.ac.uk](mailto:virus-dataflow@ebi.ac.uk).' The footer contains three columns of links: 'COVID DATA RESOURCES' (Viral Sequences, Host Sequences, Expression, Proteins, Networks, Samples, Cohorts, Imaging, Social sciences & humanities, Literature), 'TOOLS' (Bulk Downloads, Submit Data, API Documentation, Phylogenetic Tree, CoVEO Explorer), and 'ABOUT' (About the Portal, News, Partners, Related Resources, FAQ, Data Statistics). The footer also includes logos for 'eliXir', 'EMBL-EBI', 'European Commission', 'Co-funded by the Horizon 2020 programme of the European Union', and 'EOSC-Life'.

<https://www.covid19dataportal.org/submit-data>



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

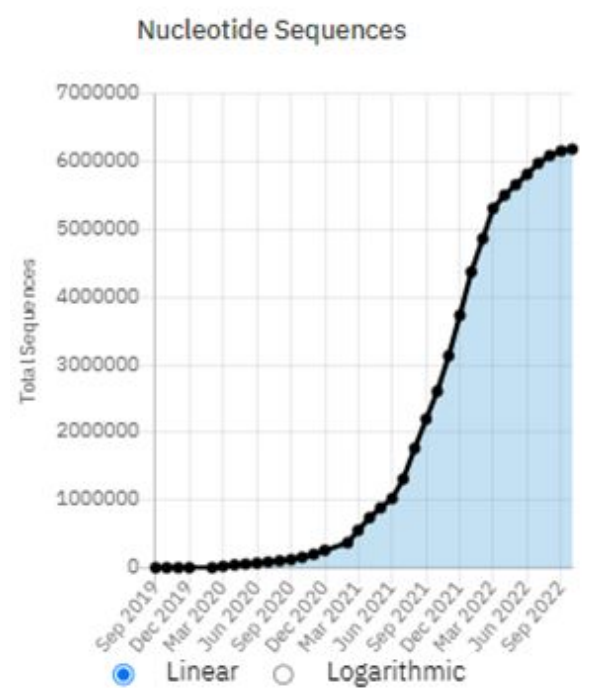
# The ENA & SARS-CoV-2

SARS-CoV-2 viral sequences are archived in ENA, and then fed into the COVID-19 Data Portal:

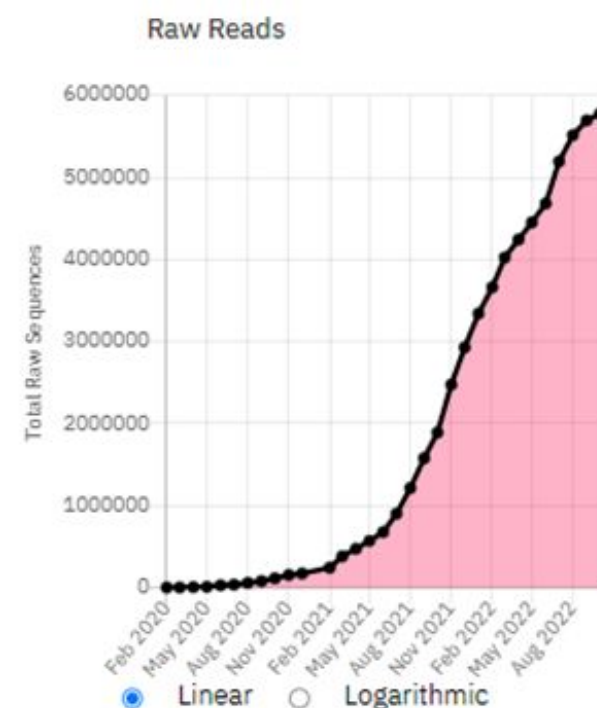


SARS-CoV-2 submissions now make up a quarter of all ENA raw read data

SARS-CoV-2 data submitted from >90 countries...

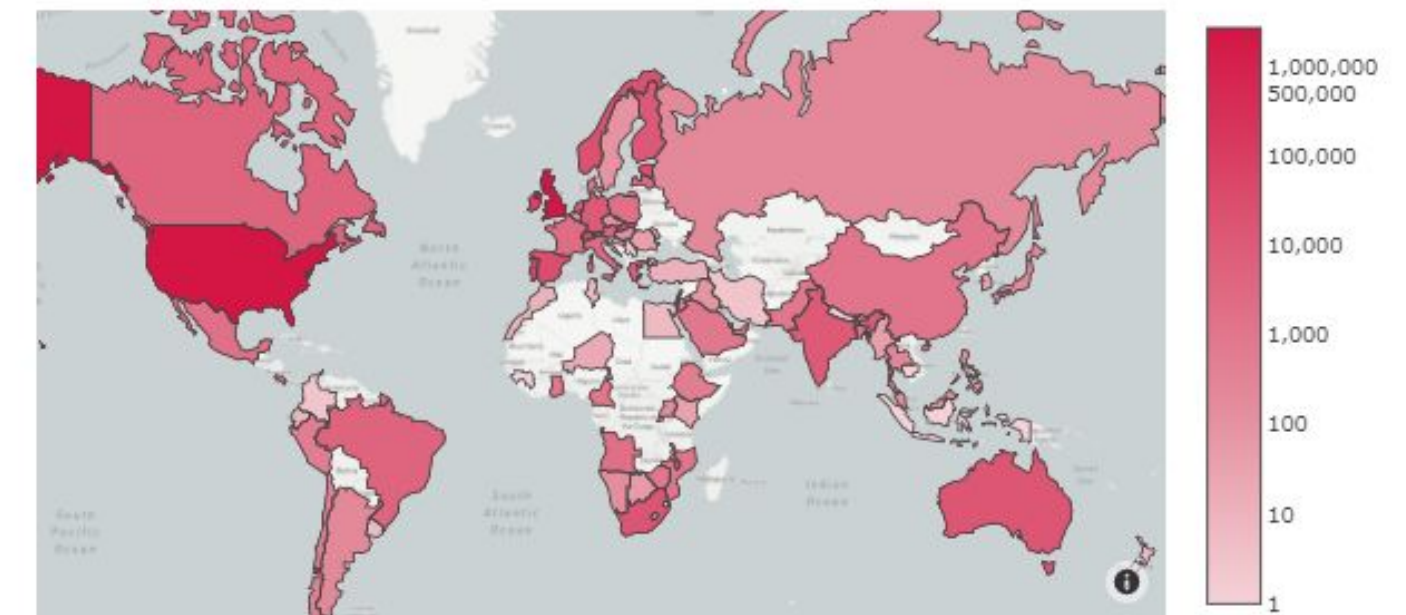


As of 1 November 2022  
**6,186,608**



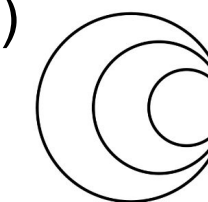
As of 1 November 2022  
**5,779,741**

Raw reads submitted by Country



...on both an individual (e.g. hospitals, labs) and national level (i.e public health authorities, national research institutes)

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



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# CABANA Project

Focus was on actively engaging Latin American countries to help them submit data to the ENA / INSDC

Leveraged network of contacts to mobilise SARS-CoV-2 data from Brazil, Mexico, Costa Rica and Argentina this year to the COVID-19 Data Portal

e.g.

<https://www.ebi.ac.uk/ena/browser/view/PRJEB53987>

ENA SARS-CoV-2 Training workshop delivered at the 2022 International Society for Computational Biology conference (ISCB) this month



**CABANA**  
Capacity building for bioinformatics in Latin America

Home About Workshops Research secondments Train the trainer eLearning Webinars News Contact us

Change language: [English](#) [Spanish](#) [Log in](#)

## What is CABANA?

- CABANA is a capacity strengthening project for bioinformatics in Latin America.
- It aims to accelerate the implementation of data-driven biology in the region by creating a sustainable capacity-building programme focusing on three challenge areas – communicable disease, sustainable food production and protection of biodiversity.
- CABANA is orchestrated by an international consortium of ten organisations - nine in Latin America and one in the UK.
- CABANA is funded by the Global Challenges Research Fund (GCRF) - part of the UK Aid Budget – from October 2017 to December 2021.

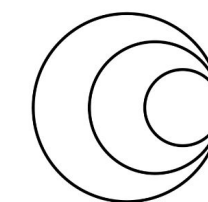
[Join the CABANA mailing list](#) [Learn more...](#)

### CABANA challenges

The project will enable research and deliver training to address three challenges:

- [Communicable disease](#) >
- [Sustainable food production](#) >
- [Protection of biodiversity](#) >

**Workshops** **Research Secondments** **Train the Trainer** **eLearning Resources**



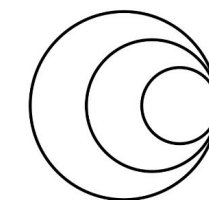
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**



# Section 5: Submitting SARS-CoV-2 data to the ENA



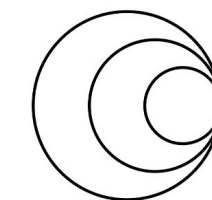
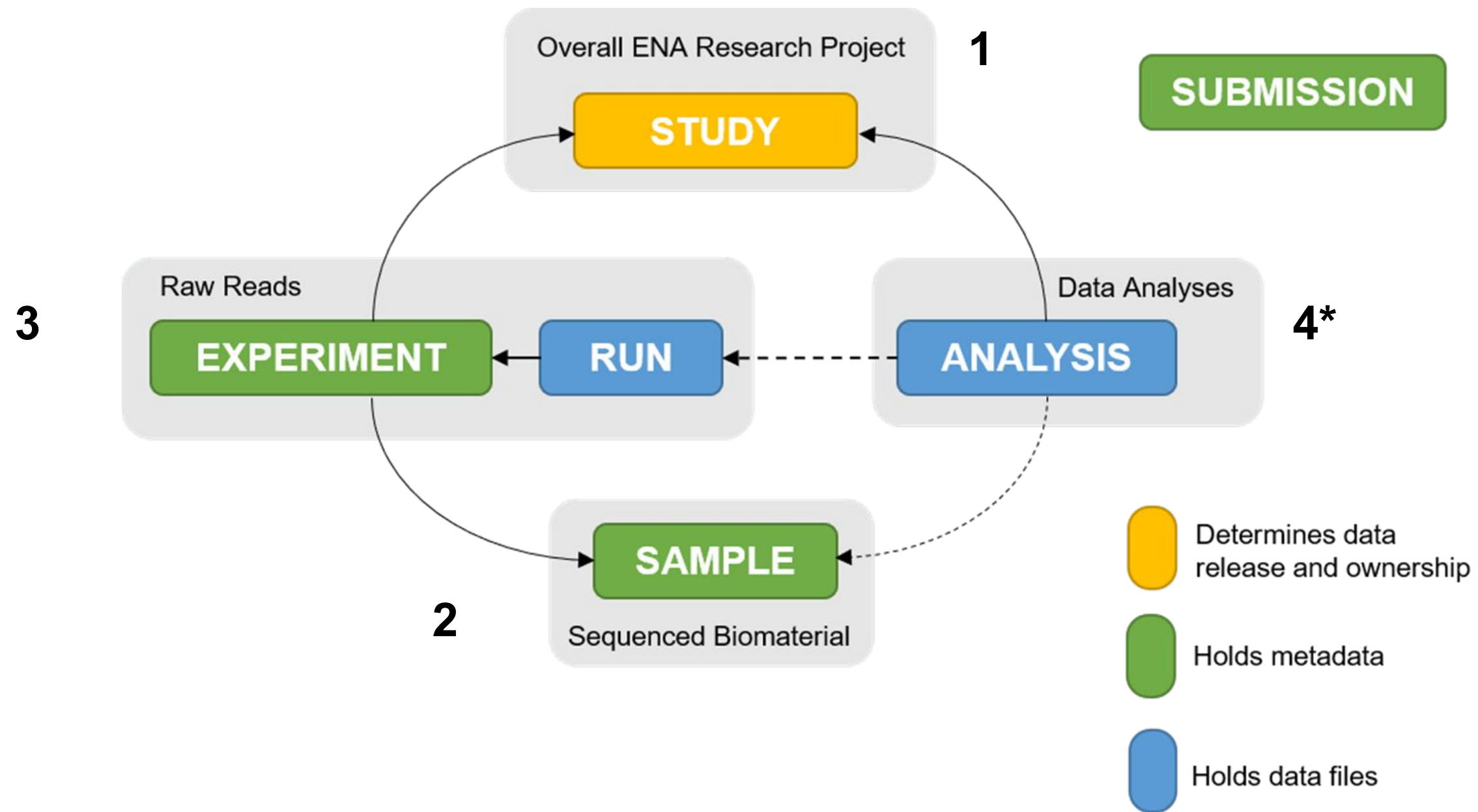
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Submitting SARS-CoV-2 data to the ENA

## The ENA Metadata Model



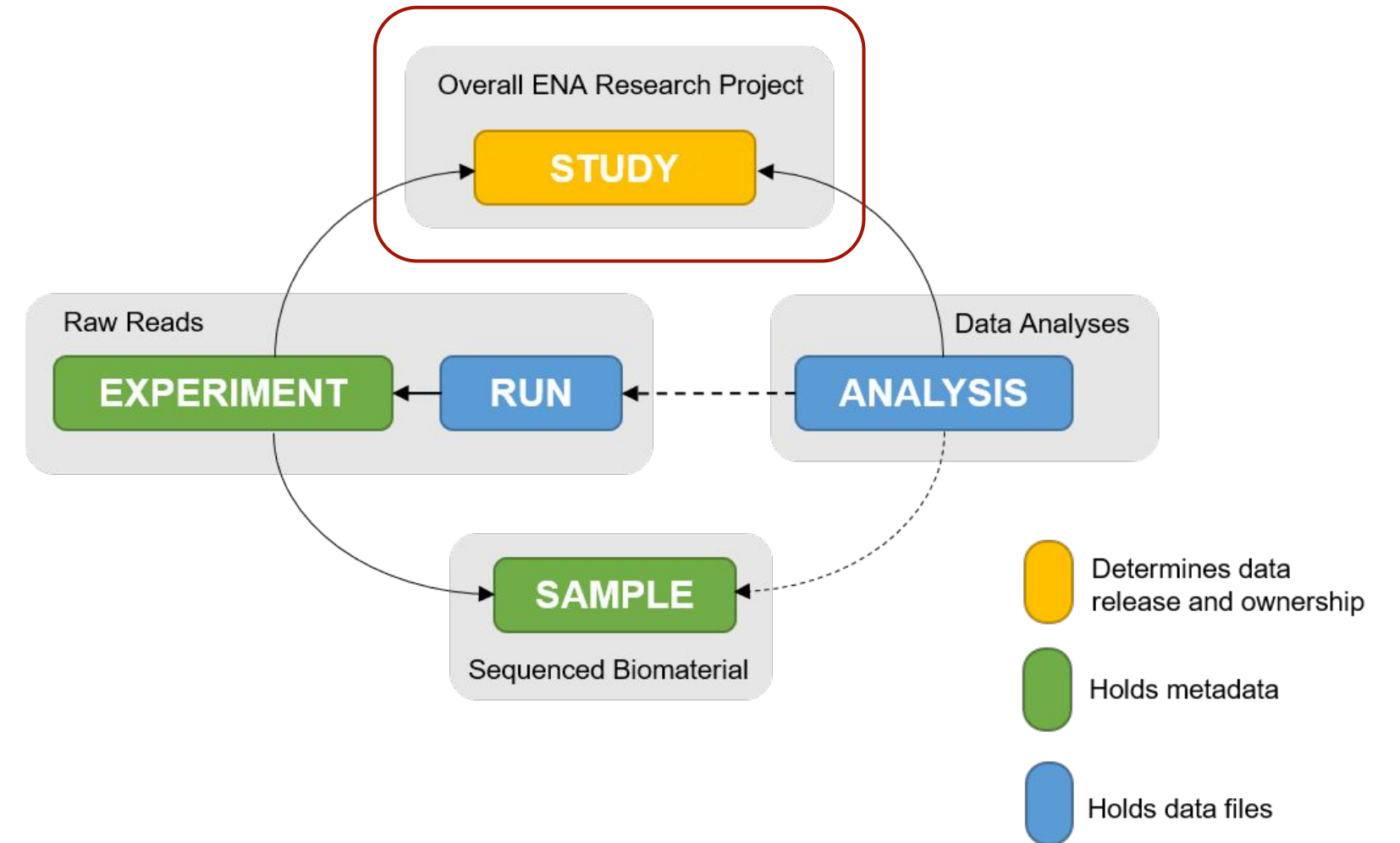
wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

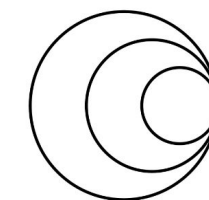
# ENA Metadata Model: Study

- Binds together related samples/runs/analyses
- Accessions like 'PRJEB\*' and 'ERP\*'
- Should be referenced in publications
- Example metadata
  - Title & description
  - Taxonomy, where applicable
  - Affiliations (e.g. submitter, centre name)
  - Release date



<https://www.ebi.ac.uk/ena/browser/view/PRJEB53987>

View > XML



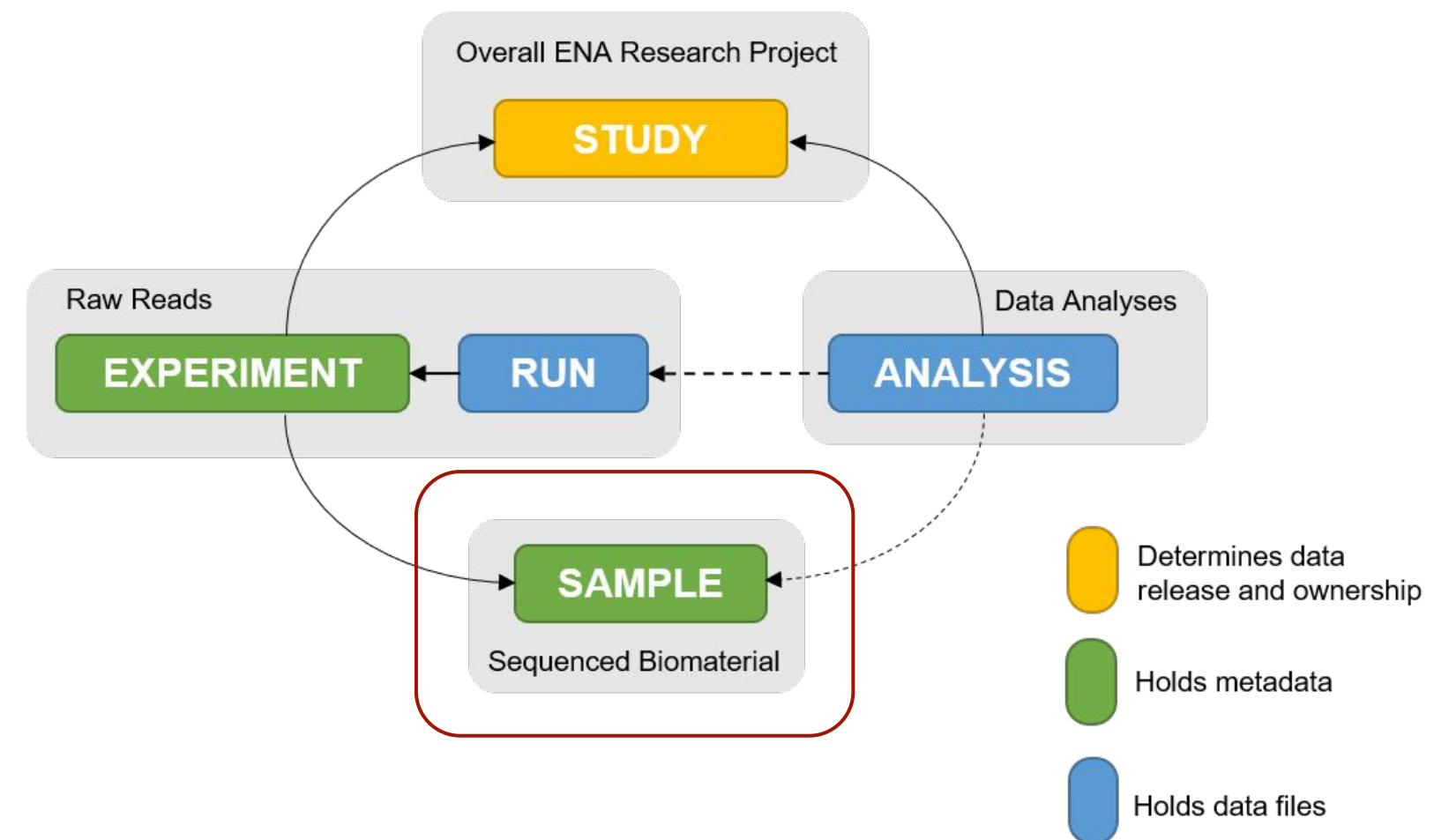
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

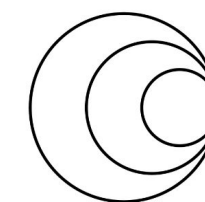
# ENA Metadata Model: Sample

- Description of sequenced biomaterial, e.g. SARS-CoV-2 virus
- Accessions like 'SAME\*' and 'ERS\*'
- Example metadata
  - Taxonomy
  - Collection date and location
  - Host/Lab host information, e.g. age, sex, disease outcome
  - Checklist: e.g. ERC000033
- Custom sample fields supported



<https://www.ebi.ac.uk/ena/browser/view/SAMEA110587357>

View> XML



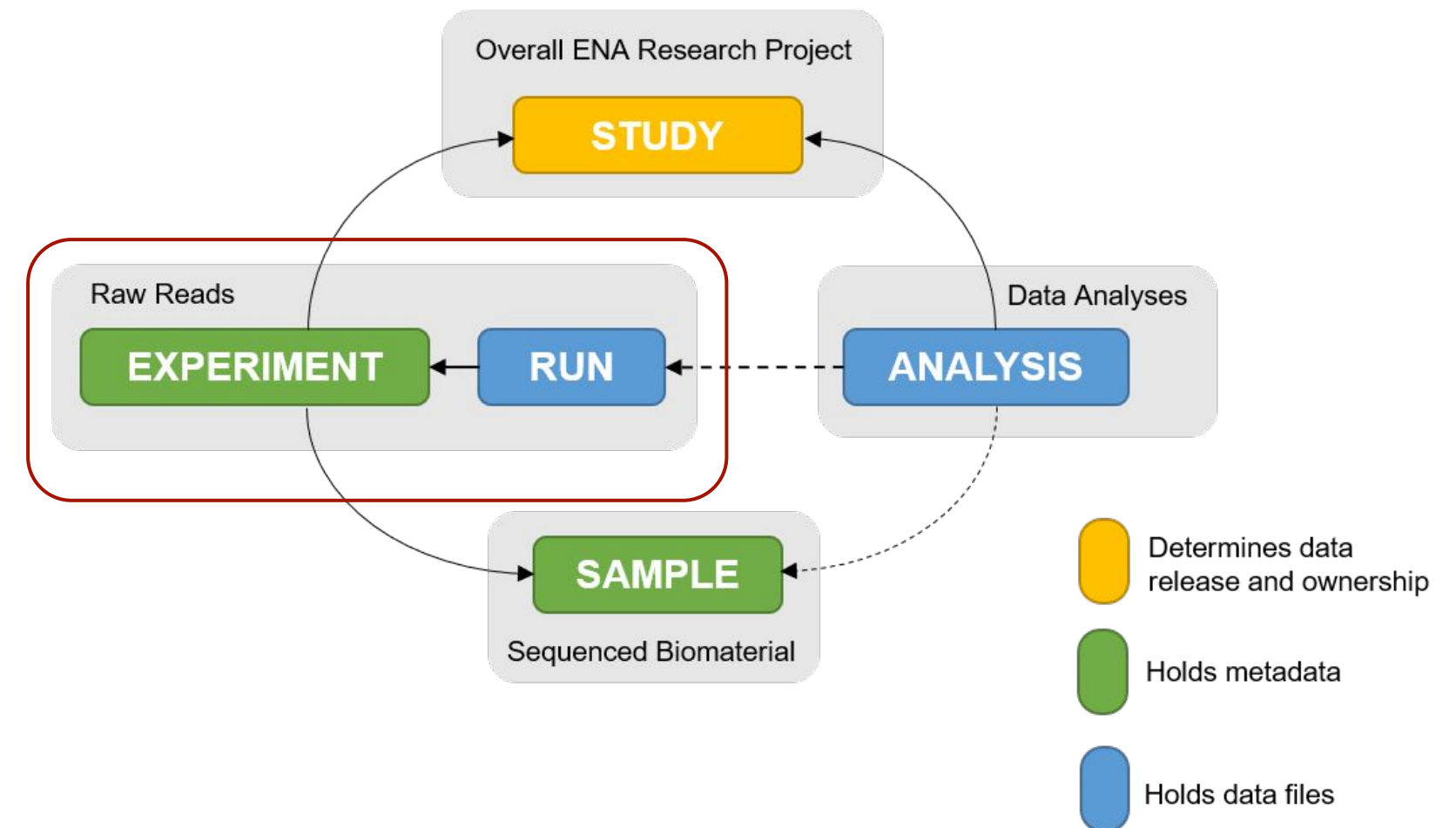
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# ENA Metadata Model: Experiment & Run

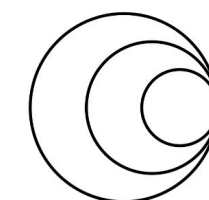
- Experiment
  - Metadata about sequencing methodology
  - Accession like 'ERX\*'
  - Example metadata
    - instrument platform and model
    - library preparation information, e.g. construction protocol, primers
- Run
  - Holds data file, e.g. BAM/CRAM/FASTQ
  - Accession like 'ERR\*'



<https://www.ebi.ac.uk/ena/browser/view/ERX9584325>

<https://www.ebi.ac.uk/ena/browser/view/ERR10044437>

View> XML



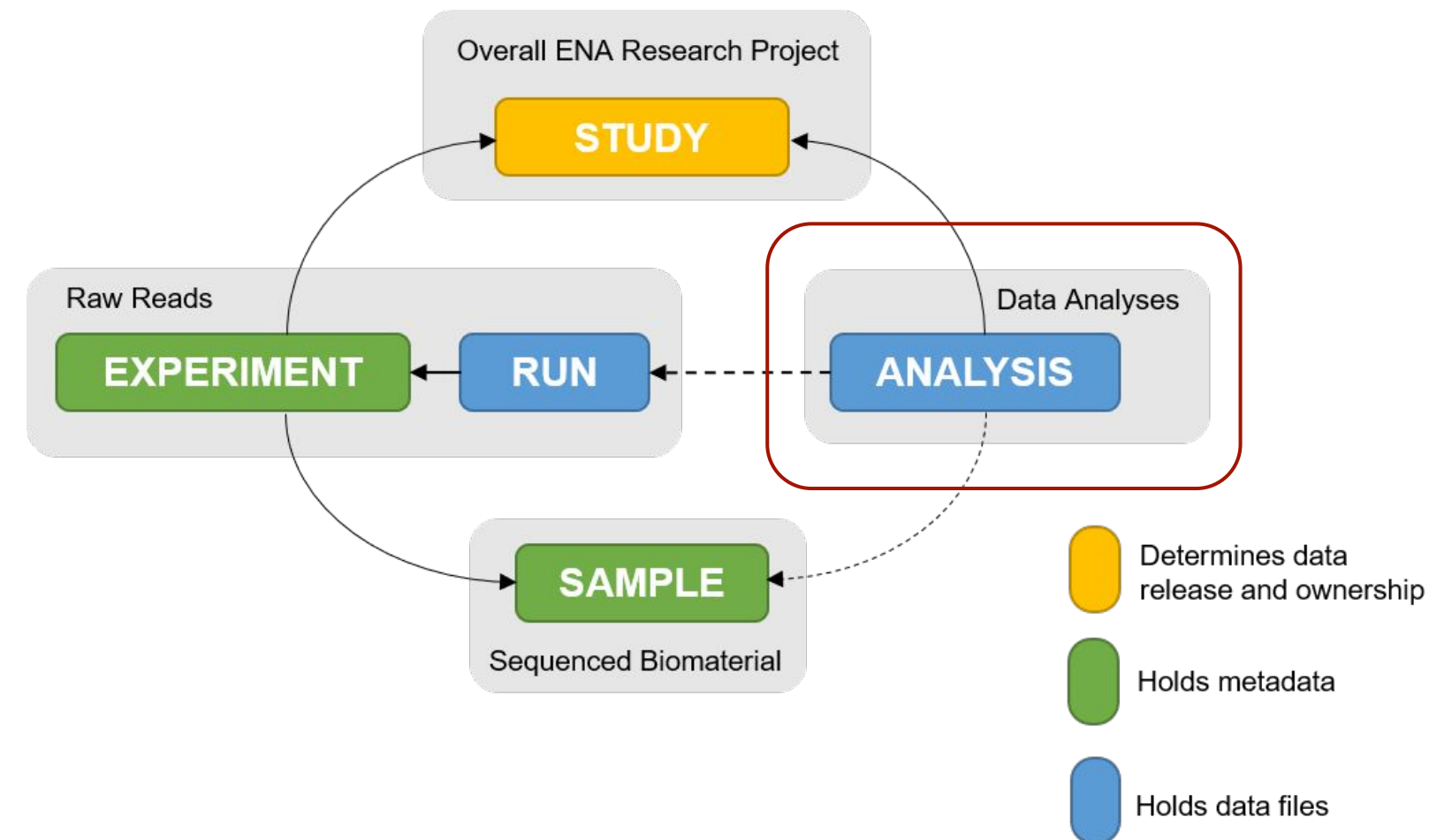
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

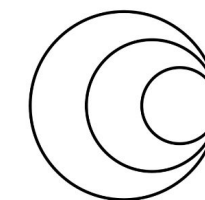
# ENA Metadata Model: Analysis

- Accessions:
  - 'ERZ\*'
  - + additional chromosome level accession, e.g. 'OW296552'
- Example metadata
  - analysis type (COVID-19 OUTBREAK)
  - assembly method and platform
  - depth of coverage
  - molecule type
    - (e.g 'genomic DNA', 'genomic RNA' or 'viral cRNA')
- Holds data file, e.g. FASTA/FLATFILE



<https://www.ebi.ac.uk/ena/browser/view/ERZ1769911> View> XML

<https://www.ebi.ac.uk/ena/browser/view/OA964249> View> EMBL



**wellcome  
connecting  
science**

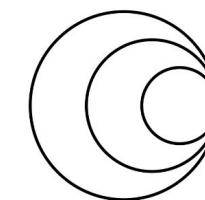


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

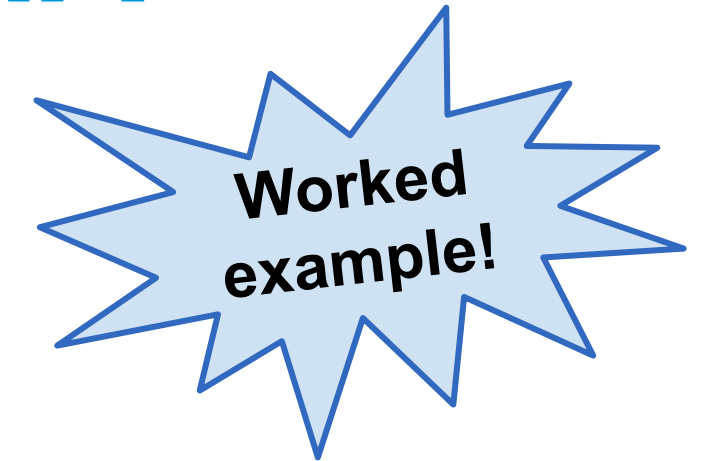
# A note on aliases

- All objects have 'aliases'
- These should be used to link objects together between local system and ENA
- Map your objects to ENA accessions
- Receipt example (programmatic submission):

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="receipt.xsl"?>
<RECEIPT receiptDate="2021-09-29T16:58:08.634+01:00" submissionFile="submission.xml" success="1"
  <PROJECT accession="PRJEB123456" alias="example_project_alias" status="PRIVATE" />
  <SUBMISSION accession="ERA123456" alias="example_submission_alias" />
  <MESSAGES>
    <INFO>This submission is a TEST submission and will be discarded within 24 hours</INFO>
  </MESSAGES>
  <ACTIONS>ADD</ACTIONS>
</RECEIPT>
```



# Submitting SARS-CoV-2 data to the ENA



Please ensure you have first registered for a Webin submission account [here](#)

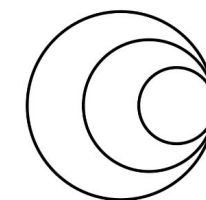
Several methods to submit ENA objects, depending on your needs and technical proficiency:

- interactive (browser-based)
- programmatic (XML-based)
- Webin-CLI (command line tool)

Today you will test the submission of:  
An **ENA Project** and **Samples interactively**

&

SARS-CoV-2 **genomes** using the **Webin-CLI program**



**wellcome  
connecting  
science**

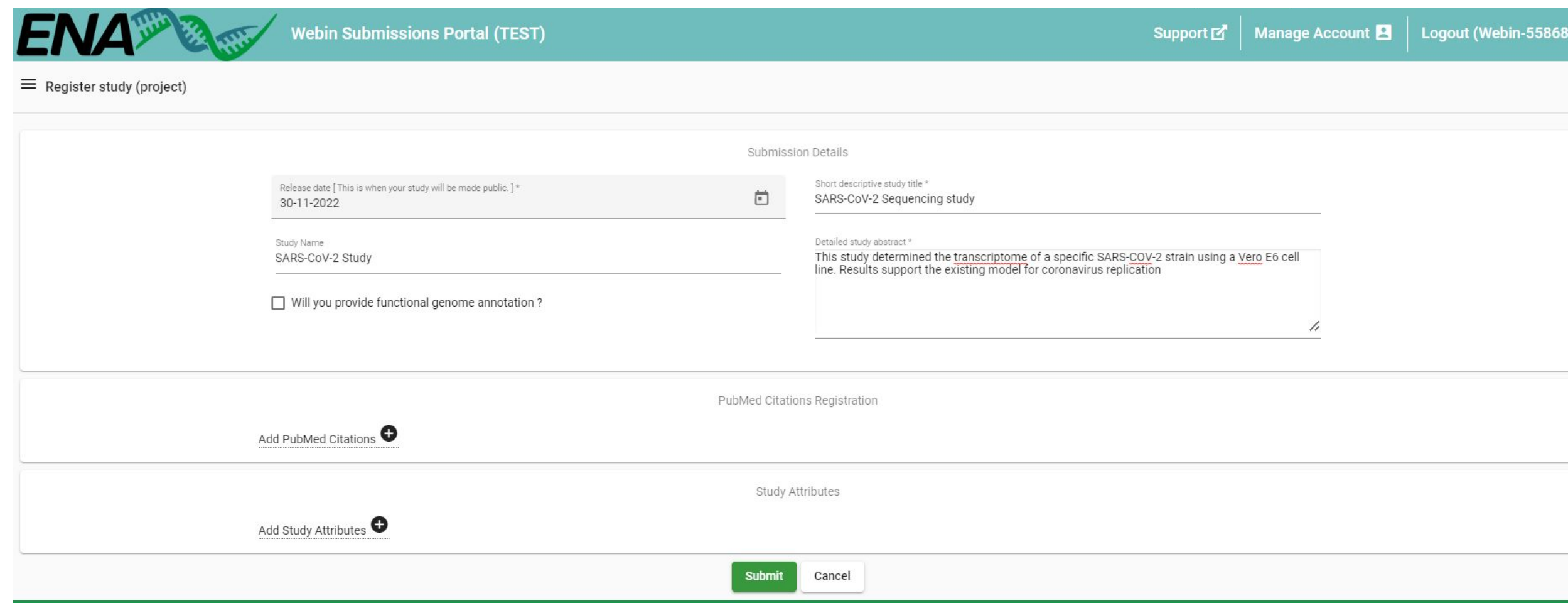


**COVID-19  
GENOMICS  
GLOBAL TRAINING**

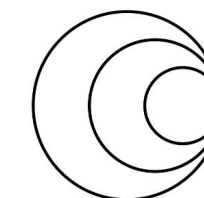


# Creating a COVID Project at the ENA

1. Log in to the Test Webin Submissions Portal:  
<https://wwwdev.ebi.ac.uk/ena/submit/webin/login>
2. Create a **Project** for your submission:



The screenshot shows the ENA Webin Submissions Portal (TEST) interface. The header includes the ENA logo, the text 'Webin Submissions Portal (TEST)', and links for 'Support', 'Manage Account', and 'Logout (Webin-55868)'. The main content area is titled 'Register study (project)' and contains three sections: 'Submission Details', 'PubMed Citations Registration', and 'Study Attributes'. The 'Submission Details' section includes a 'Release date' field (30-11-2022), a 'Short descriptive study title' field (SARS-CoV-2 Sequencing study), a 'Study Name' field (SARS-CoV-2 Study), and a 'Detailed study abstract' field (This study determined the transcriptome of a specific SARS-COV-2 strain using a Vero E6 cell line. Results support the existing model for coronavirus replication). There is also a checkbox for 'Will you provide functional genome annotation?'. The 'PubMed Citations Registration' section has an 'Add PubMed Citations' button. The 'Study Attributes' section has an 'Add Study Attributes' button. At the bottom, there are 'Submit' and 'Cancel' buttons.



wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

# Submitting COVID samples to the ENA - 1

1. Download and unzip the `Module4_data_zip` folder - link [here](#)
2. Using the pre-filled ENA sample spreadsheet: `sample_spreadsheet_COG_Train.tsv...`

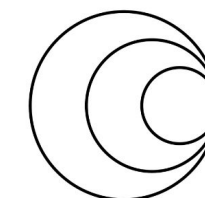
1	Checklist	ERC00003: ENA virus pathogen reporting standard checklist															
2	tax_id	scientific_name	sample_accession	sample_title	sample_description	collection_date	geographic_location	geographic_coordinates	sample_capture_status	host	comorbidity	host_subject	host_health	host_sex	host_science	collector	submitting_institution
3	#units																
4	2697049	Severe acute respiratory syndrome coronavirus 2	Case A	Case A	SARS-CoV-2 sample from passenger A	2020-10-02	New Zealand Auckland		active surveillance	human		A	diseased	not collected	Homo sapiens	Una Ren	Institute of Microbiology and Immunology
5	2697049	Severe acute respiratory syndrome coronavirus 2	Case B	Case B	SARS-CoV-2 sample from passenger B	2020-10-02	New Zealand Auckland		active surveillance	human		B	diseased	not collected	Homo sapiens	Marcela E. Universida	Universidade Federal do Rio de Janeiro
6	2697049	Severe acute respiratory syndrome coronavirus 2	Case C	Case C	SARS-CoV-2 sample from passenger C	2020-10-02	New Zealand Auckland		active surveillance	human		C	diseased	not collected	Homo sapiens	Zahra Wal	EMBL-EBI
7																	

- All mandatory (and some recommended) fields of [ERC000033](#) present within tsv file
- [INSDC missing terms](#) can be used for any mandatory fields where information cannot be provided
- **‘Active surveillance in response to outbreak’** strongly recommended field value
- **‘GISAID Accession ID’** custom attribute

host sex
not collected
not collected
not collected

sample capture status
active surveillance in response to outbreak
active surveillance in response to outbreak
active surveillance in response to outbreak

GISAID Accession ID
EPI_ISL_582019
EPI_ISL_582020
EPI_ISL_582021



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# ERC000033 Sample Checklist

## Checklist: ERC000033



### ENA virus pathogen reporting standard checklist

Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

## Checklist Fields



Filter fields...

### Filter by type:

Collection event information

host description

General collection event information

Intraspecies information

Field Name	Field Format	(Field Restriction)	Requirement Mandatory	(Units)
geographic location (country and/or sea)	text choice	options	mandatory	
host common name	free text		mandatory	
host subject id	free text		mandatory	
host health state	text choice	options	mandatory	
host sex	text choice	options	mandatory	
host scientific name	free text		mandatory	
collector name	free text		mandatory	
collecting institution	free text		mandatory	
isolate	free text		mandatory	

# Submitting COVID samples to the ENA - 2

...upload this directly to the Webin Submissions Portal:

Ensure collection date format is: YYYY-MM-DD

ENA Webin Submissions Portal (TEST)

- Dashboard
- Studies (Projects)
  - Register Study (Project)
  - Submit XML (advanced)
  - Study Report
- Samples**
  - Register Samples**
  - Register Novel Taxonomy
  - Submit XML (advanced)
  - Samples Report
- Raw Reads (Experiments and Runs)
  - Submit Reads
  - Submit XML (advanced)
  - Runs Report



ENA Webin Submissions Portal (TEST) Support Manage Account Log

Register Samples using spreadsheet template

Download spreadsheet to register samples

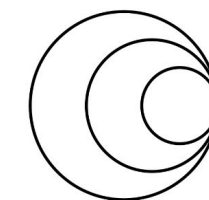
**Upload filled spreadsheet to register samples**

If you have downloaded and filled a template spreadsheet please upload it using the **Submit Completed Spreadsheet** button.

Please note that only spreadsheets in tab-delimited text format are supported (with either .tsv or .tab extensions). If you edited the spreadsheet in Microsoft Excel (or equivalent) please save the spreadsheet as Text (Tab delimited). To do this please see [these instructions](#).

Choose File No file chosen

Submit Completed Spreadsheet



wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

# Submitting COVID samples to the ENA - 3

- Accessions will be provided immediately, and can be viewed in the 'Samples Report' section of the Webin Submission Portal:

## Submission result

✓ The submission was successful.

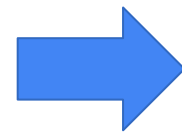
Show receipt XML

[Download accessions](#) [Download receipt XML](#)

Type	Accession	Unique name (alias)
Sample	ERS13666078	COVID Case A
Sample	ERS13666079	COVID Case B
Sample	ERS13666080	COVID Case C
Submission	ERA18575973	ena-SUBMISSION-TAB-06-11-2022-20:58:10:123-25

Items per page: 10 0 of 0

Close



### Samples Report

Shows submitted samples and their release statuses. Search by accession or unique name, or simply click search to show most recent submissions. The results will show the most recently submitted samples in your submission account.

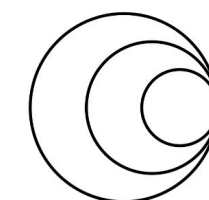
Please click search to see the results.

Search Samples

Accession or Name Release status Maximum rows 100  Show unique name [Search](#) [Reset](#)

[Download all results](#)

Accession	BioSample	Title	Organism	Tax id	Submission date	Status	Action
ERS13666080	SAMEA130171410	Case C	Severe acute respiratory syndrome coronavirus 2	2697049	6th Nov 2022	Private	<a href="#">🔗</a>
ERS13666079	SAMEA130171409	Case B	Severe acute respiratory syndrome coronavirus 2	2697049	6th Nov 2022	Private	<a href="#">🔗</a>
ERS13666078	SAMEA130171408	Case A	Severe acute respiratory syndrome coronavirus 2	2697049	6th Nov 2022	Private	<a href="#">🔗</a>



wellcome  
connecting  
science



COVID-19  
GENOMICS  
GLOBAL TRAINING

# Submitting COVID samples to the ENA - 3

- PHA4GE (Public Health Alliance for Genomic Epidemiology) recommended metadata for COVID data sharing: <https://tinyurl.com/358rhuf4>
- Contains **mapping of PHA4GE fields to ENA ERC000033 checklist** - any extra fields to be added as custom sample fields

# Submitting COVID genomes to the ENA - 1

- 3 files required for a SARS-CoV-2 assembly submission with Webin-CLI:
  - Fasta (gzipped)
  - Manifest file (specifying Project and Sample accessions, and assembly metadata)
  - [Chromosome list file](#) (gzipped)

*Assemblies can be linked to originating run data, via run accession*

```
CaseA_manifest.txt
1 STUDY PRJEB####
2 SAMPLE ERS#####
3 RUN_REF ERR#####
4 ASSEMBLYNAME SARS-CoV-2 assembly Case A
5 ASSEMBLY_TYPE COVID-19 outbreak
6 COVERAGE 100
7 PROGRAM ARTIC fieldbioinformatics (minimap2/nanopolish) 1.1.3 (nanopolish 0.13.2)
8 PLATFORM ILLUMINA
9 MINGAPLENGTH 2
10 MOLECULETYPE genomic RNA
11 DESCRIPTION example sequence #1 for workshop
12 FASTA 20CV0408.fasta.gz
13 CHROMOSOME_LIST CaseA_chromosome_list.txt.gz
```

**Manifest file**

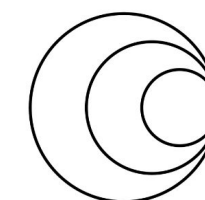
```
CaseA_chromosome_list.txt
1 hCoV-19/NewZealand/20CV0408/2020 1 Monopartite
```

*Tab separated text file containing a single row only:*

- *Fasta header, chromosome number ('1'), chromosome type (**Monopartite** for coronaviruses)*

**Chromosome list file**

*'ASSEMBLY\_TYPE' should be set to 'COVID-19 OUTBREAK'*



# Submitting COVID genomes to the ENA - 2

1. Download the latest release of our Webin-CLI program [here](#)
2. Copy and paste the `webin-cli-*` jar file into the unzipped `Module4_data` folder, so all is one place. Here you'll also find all `fasta.gz`, `manifest` and `chromosome list` files
3. **Edit the manifest files to include your newly created Project (PRJEB###) and Sample (ERS###) accessions**
4. Run the command below in your favourite terminal (e.g. Ubuntu, or Git Bash, etc.), specifying your Webin credentials:

```
java -jar webin-cli-5.2.0.jar -context genome -userName  
'Webin-####' -password '#####' -manifest CaseA_manifest.txt  
-submit -test
```

*Specifies type of submission*

*Validates + submits files defined in manifest file*



# Submitting COVID genomes to the ENA - 2

1. Download and unzip the `Module4_data_zip` folder, where you'll find all required data files, as well as the latest release of our Webin-CLI program (you can also find this here)
2. Copy and paste the `webin-cli-*` jar file into the unzipped `Module4_data` folder, so all is one place

1. **Ensure you edit the manifest files to include your newly created Project (PRJEB###) and Sample (ERS###) accersions**

2. Run the command below in your favourite terminal (e.g. Ubuntu, or Git Bash, etc.), specifying your Webin credentials:

```
java -jar webin-cli-5.2.0.jar -context genome -userName  
'Webin-####' -password '#####' -manifest CaseA_manifest.txt  
-submit -test
```

*Specifies type of submission*

*Validates + submits files defined in manifest file*

# Submitting COVID genomes to the ENA - 3

- Successful output:

```
INFO : Your application version is 5.2.0
INFO :
A dedicated submission API for COVID-19 genomes is available here: <a href="https://www.ebi.ac.uk/ena/submit/webin-cli">https://www.ebi.ac.uk/ena/submit/webin-cli </a> <br>
INFO : Submission has not been validated previously.
INFO : Creating report file: C:\Users\zahra\Documents\COG-Train\.\webin-cli.report
INFO : The submission has been validated successfully.
INFO : Uploading file: C:\Users\zahra\Documents\COG-Train\20CV0408.fasta.gz

INFO : Uploading file: C:\Users\zahra\Documents\COG-Train\CaseA_chromosome_list.txt.gz

INFO : Files have been uploaded to webin2.ebi.ac.uk.
INFO : The TEST submission has been completed successfully. This was a TEST submission and no data was submitted. The following analysis accession was assigned to the submission: ERZ14235939
```

- Test analysis objects can be viewed under 'Analysis Report' of Webin Submissions Portal
- **Repeat step 4** (on previous slide) **specifying a different manifest and chromosome list file each time**, to submit SARS-CoV-2 genomes from Cases B and C

# Bulk Webin-CLI Tool

- To bulk submit assemblies and runs using Webin-CLI

code style black

## ENA Webin-CLI Bulk Submission Tool

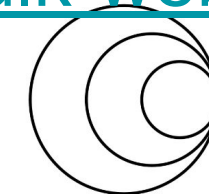
### Introduction

This tool is a wrapper to bulk submit read, un-annotated genome, targeted sequence or taxonomic reference data to the ENA using Webin-CLI.

The tool requires an appropriate metadata spreadsheet which it uses to generate manifest files for the user and validate or submit their submission. The tool does not handle study and sample registration, therefore visit [ENA Submissions Documentation](#) for more information on this. The documentation also provides information on manifest file fields for your type of submission (which correlate to the headers in the spreadsheet file).

An example template spreadsheet has been provided (example\_template\_input.txt). This file is a tab-delimited text file, however the script also consumes spreadsheets in native MS Excel formats (e.g. .xlsx) or comma-separated (.csv).

<https://github.com/enasequence/ena-bulk-webincli>



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Other methods to submit SARS-CoV-2 data to the ENA

## Programmatic

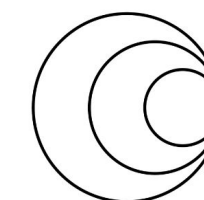
- For high-volume and/or frequent submissions (e.g. brokered data)
- Create and submit XMLs for Projects, Samples, Runs/Experiments ( **X** analysis)
- Submit via cURL

```
<SAMPLE_SET>
<SAMPLE alias="Test SARS-CoV-2 sample 1" center_name="EBI">
  <TITLE>Test SARS-CoV-2 Sample 1 Title</TITLE>
  <SAMPLE_NAME>
    <TAXON_ID>2697049</TAXON_ID>
    <SCIENTIFIC_NAME>Severe acute respiratory syndrome coronavirus 2</SCIENTIFIC_NAME>
    <COMMON_NAME>SARS-CoV-2</COMMON_NAME>
  </SAMPLE_NAME>
  <SAMPLE_ATTRIBUTES>
    <SAMPLE_ATTRIBUTE>
      <TAG>geographic location (country and/or sea)</TAG>
      <VALUE>United Kingdom</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
      <TAG>collection date</TAG>
      <VALUE>2020-04-26</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
      <TAG>host common name</TAG>
      <VALUE>human</VALUE>
    </SAMPLE_ATTRIBUTE>
  </SAMPLE_ATTRIBUTES>
</SAMPLE>
</SAMPLE_SET>
```

E.g. Sample XML

```
▼<EXPERIMENT_SET>
▼<EXPERIMENT accession="ERX9541016" alias="COG-UK/LSPA-3EBF5EC/SANG:220708_A01404_0494_BH3J3TDRX2/2t183" center_name="Wellcome Sanger Institute">
  ▼<IDENTIFIERS>
    <PRIMARY_ID>ERX9541016</PRIMARY_ID>
    <SUBMITTER_ID namespace="Wellcome Sanger Institute">COG-UK/LSPA-3EBF5EC/SANG:220708_A01404_0494_BH3J3TDRX2/2t183</SUBMITTER_ID>
  </IDENTIFIERS>
  <TITLE>Illumina NovaSeq 6000 paired end sequencing; Illumina NovaSeq 6000 paired end sequencing; COG-UK/LSPA-3EBF5EC/SANG:220708_A01404_0494_BH3J3TDRX2/2t183</TITLE>
  ▼<STUDY_REF accession="ERP121228">
    ▼<IDENTIFIERS>
      <PRIMARY_ID>ERP121228</PRIMARY_ID>
      <SECONDARY_ID>PRJEB37886</SECONDARY_ID>
    </IDENTIFIERS>
  </STUDY_REF>
  ▼<DESIGN>
    <DESIGN_DESCRIPTION>Illumina NovaSeq 6000 amplicon sequencing. Samples prepared and sequenced by Donald Fraser, Suki Lee, Rob Howes, The Rosalind Franklin Institute, and Alex Alderton, Roberto Amato, Jeffrey Barrett, Sonia Goncalves, Ewan Harrison, David K. Jackson, Ian Johnston, Dominic Kwiatkowski, Cordelia Langford, and the Wellcome Sanger Institute COVID-19 Surveillance Team</DESIGN_DESCRIPTION>
  </DESIGN>
  ▼<SAMPLE_DESCRIPTOR accession="ERS12524969">
    ▼<IDENTIFIERS>
      <PRIMARY_ID>ERS12524969</PRIMARY_ID>
      <EXTERNAL_ID namespace="BioSample">SAMEA110427043</EXTERNAL_ID>
    </IDENTIFIERS>
  </SAMPLE_DESCRIPTOR>
</EXPERIMENT>
</EXPERIMENT_SET>
```

E.g. Experiment XML



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# SARS-CoV-2 specific tools

## Webin-CLI JSON API

- For high-volume and/or frequent submissions
- Submit SARS-CoV-2 sequence and metadata as a JSON payload (no manifest file nor chromosome list)
- **Genome assembly submissions only**

Covid-19 GenomeAPI Validation and submission of Covid-19 genome sequence

POST /api/v1/genome/covid-19

Submit Covid-19 genome sequence data.

Parameters Try it out

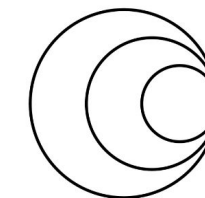
No parameters

Request body <sup>required</sup> application/json

Example Value | Schema

```
{
  "name": "string",
  "study": "string",
  "sample": "string",
  "coverage": 0,
  "program": "string",
  "platform": "string",
  "sequence": "string",
  "description": "string",
  "minGapLength": 0,
  "moleculeType": "genomic DNA",
  "runRef": "string",
  "analysisRef": "string",
  "tpa": true,
  "authors": "string",
  "address": "string",
  "submissionTool": "string",
  "submissionToolVersion": "string"
}
```

<https://tinyurl.com/4d6nymzs>



wellcome  
connecting  
science

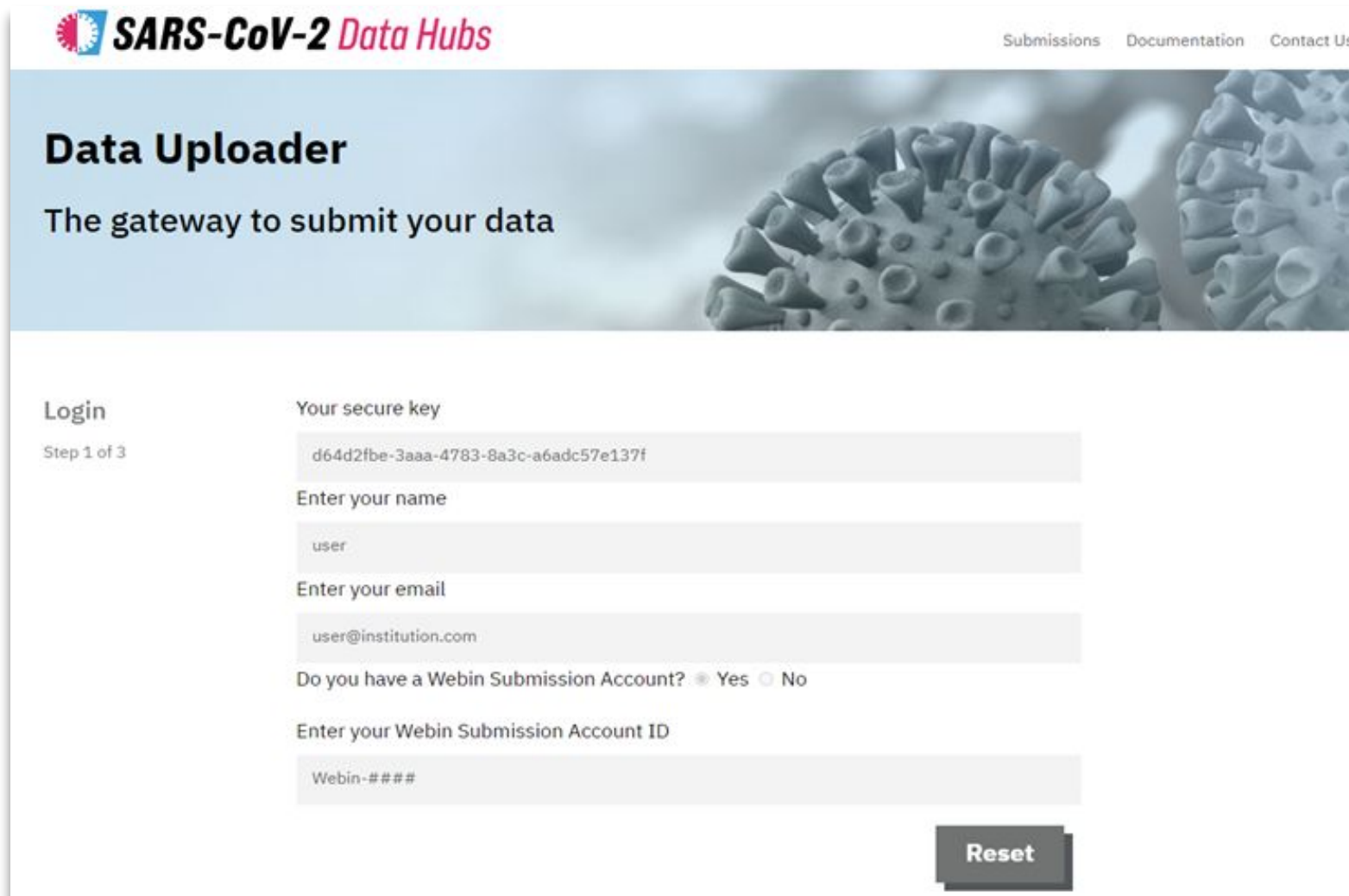


COVID-19  
GENOMICS  
GLOBAL TRAINING

# SARS-CoV-2 specific tools

## Drag and Drop Uploader Tool

- For small-scale/one-off submissions
- Submit any SARS-CoV-2 datatype
- Easy to use, simply drag and drop data files + metadata spreadsheet

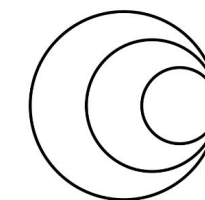


The screenshot shows the 'Data Uploader' page on the SARS-CoV-2 Data Hubs website. The page has a light blue header with the logo and navigation links for 'Submissions', 'Documentation', and 'Contact Us'. Below the header is a large image of a virus particle. The main content area is white and contains a 'Login' section with the following fields and options:

- Login** (Step 1 of 3)
- Your secure key**: A text input field containing the alphanumeric string 'd64d2fbe-3aaa-4783-8a3c-a6adc57e137f'.
- Enter your name**: A text input field containing the text 'user'.
- Enter your email**: A text input field containing the email address 'user@institution.com'.
- Do you have a Webin Submission Account?**: A radio button selection with 'Yes' selected and 'No' unselected.
- Enter your Webin Submission Account ID**: A text input field containing the text 'Webin-####'.
- Reset**: A dark grey button located at the bottom right of the form.

Email [virus-dataflow@ebi.ac.uk](mailto:virus-dataflow@ebi.ac.uk) for login details & metadata spreadsheet

<https://ebi-ait.github.io/sars-cov2-data-upload/>



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

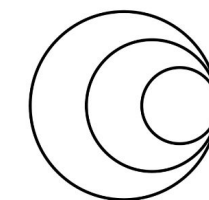
# ORCID Data Claiming

- You can also claim ENA Projects to your ORCID ID
- Search for your project in the 'ENA Study' search box:  
<https://www.ebi.ac.uk/ebisearch/orcidclaimdocumentation.ebi>
- Select 'Claim to ORCID' to login to your ORCID account and claim the ENA Study

The screenshot shows the EBI Search interface. At the top, the search bar contains 'PRJEB43947'. Below the search bar, there are navigation links: 'Help & Documentation', 'About EBI Search', 'ORCID data claiming', and 'Feedback'. The main content area displays 'Search results for PRJEB43947'. Below this, it says 'Showing 1 results out of 1 in All results → Nucleotide sequences → ENA Study'. There are three buttons: 'Save result', 'Claim to ORCID' (highlighted with a red box), and 'Create RSS feed'. The search results list shows 'ENA Study (1 results)' with a checkbox. The first result is 'SARS-CoV-2 Systematic Variant Calling (COVID-19 Taskforce VEO) Source: ENA Study (ID: PRJEB43947)'. Below the result, it says 'Systematically called variant data of public SARS-CoV-2 reads'. At the bottom, there are logos for 'JGI' and 'RAIN', and the text 'COVID-19 GENOMICS GLOBAL TRAINING'.

# ENA submission documentation

- SARS-CoV-2 specific ENA submission guide:  
[https://ena-browser-docs.readthedocs.io/en/latest/help\\_and\\_guides/sars-cov-2-submissions.html](https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/sars-cov-2-submissions.html)
- Detailed SARS-CoV-2 workshop:  
[https://ena-covid19-docs.readthedocs.io/en/latest/submission\\_workshop/getting\\_started.html](https://ena-covid19-docs.readthedocs.io/en/latest/submission_workshop/getting_started.html)



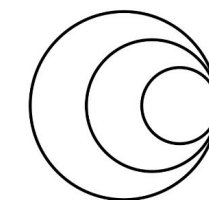
**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**



# Section 6: COVID-19 Data Portal - Search & Retrieval



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Search- interactive & programmatic

Facets on web interface

COVID Portal Advanced Search

**COVID-19 Data Portal**

Viral Sequences Host Sequences Expression Proteins Networks Cohorts More

**Viral sequences**  
Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses

country:United Kingdom **Search**

Examples: lineage:B.1.1.7 , whoomicron , Severe acute respiratory syndrome 2 ...Advanced search

**COVID-19 Data Portal**

Viral Sequences Host Sequences Expression Proteins Networks Cohorts More

**Advanced search**  
Build complex queries for more detailed results.

Build query Load query

Choose category: Viral sequences

Sequences

AND OR NOT

Add rule Add ruleset

Save Reset Search

on the European COVID-19 platform: [ecovid19@ebi.ac.uk](mailto:ecovid19@ebi.ac.uk).  
your data on the COVID-19 Data Portal: [virus-dataflow@ebi.ac.uk](mailto:virus-dataflow@ebi.ac.uk).

COVID Portal API

Search results for: **country:United Kingdom** [X]

Showing 15 of 2,180,670 in All > Viral sequences > Sequences

Data types: Download Phylogeny Tree Variant Browser

Accession	Lineage	Cross-references	Collection date
OA982812	B.1.1.253	BioSamples (2)	Sep 11, 2020
OA982828	B.1.1.37	BioSamples (2)	Sep 10, 2020
OA982890	B.1.258.3	BioSamples (2)	Sep 10, 2020
OA982900	B.1.1.37	BioSamples (2)	Sep 10, 2020
OA982901	B.1.1.303	BioSamples (2)	Sep 10, 2020
OA982937	AD.2	BioSamples (2)	Sep 10, 2020
OA982940	B.1.1.253	BioSamples (2)	Sep 10, 2020
OA982945	B.1.1.37	BioSamples (2)	Sep 10, 2020
OA982956	B.1.1.12	BioSamples (2)	Sep 10, 2020

Center name: COVID-19 Genomics UK Consortium (2,178,333), Modernising Medical Microbiology (1,385), Quadram Institute Bioscience (869)

**COVID-19 Data Portal API**  
Powering the COVID-19 Data Portal through EBI Search

The COVID-19 Data Portal is powered by the EBI Search API. With its flexible use cases.

We also offer a proxy to EBI Search, which provides endpoints with domain. Base URL: <https://www.ebi.ac.uk/ebisearch/ws/rest>

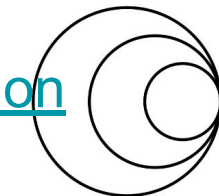
Proxy base URL: <https://www.covid19dataportal.org/api/backend/>

**Parameters**

Name	Description	Type
query	The EBI search term	String
fields	Comma separated values of field identifiers to retrieve	String
start	The index of the first entry in the results	Integer
size	The number of entries to retrieve (response page size)	Integer
format	The response format	A string of either: • JSON • XML • TSV • CSV • idlist • acclist • cs_idlist • cs_acclist

Category	Data resource	Request	Proxy request	Powered by
Viral sequences	Sequences Assembled/consensus sequences for SARS-CoV-2 and other coronaviruses	<a href="#">/embl-covid19?...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	ENA
	Raw reads Raw sequencing read experiments for SARS-CoV-2, other coronaviruses and coronavirus-related metagenomes	<a href="#">/sra-experimen...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	ENA
	Systematic Analyses Analysis of publicly released raw SARS-CoV-2 sequence data	<a href="#">/sra-analysis-c...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	ENA
	Studies ENA studies containing data from SARS-CoV-2, other coronaviruses and coronavirus-related metagenomes	<a href="#">/project-covid1...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	ENA
	Genes Information about genes found in SARS-CoV-2	<a href="#">/ensemblGeno...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	e!
	Genome Browser SARS-CoV-2 whole genome browser	<a href="#">/ensemblGeno...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	e!
	Variants Information about variants found in SARS-CoV-2 and, where available, associated disease and phenotype information	<a href="#">/eva-variants-c...</a> Copy Copy as oURL	<a href="#">/viral-sequence...</a> Copy Copy as oURL	

<https://www.covid19dataportal.org/api-documentation>



**wellcome connecting science**



**COVID-19 GENOMICS GLOBAL TRAINING**

# Retrieval - interactive & programmatic

Download button on web interface

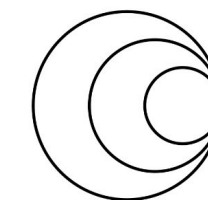
The screenshot shows the ENA 'Viral sequences' page. A table lists sequences with columns for Accession, Lineage, Cross-references, Center name, Host, and Taxonomy. A 'Download' button is highlighted on the table. A modal window titled 'Download details: Sequences' is open, showing options for 'Data format' (EMBL, FASTA) and 'Metadata' (List of IDs, TSV). The modal also includes a 'Download' button and a 'Cancel' button.

Bulk downloader tool

The screenshot shows the 'Bulk downloads' page on the ENA website. It includes sections for 'Bulk Downloads', 'Sections', 'CDP File Downloader', 'Features', 'How to run', 'Download Functionality', 'Support', and 'Privacy Notice'. A 'Download CDP-File-Downloader' button is visible. Below the page content is a terminal window showing the output of the CDP File Downloader utility, including a welcome message and instructions for use.

```
(Copyright © EMBL 2021)
Welcome to the Covid-19 Data Portal's data downloader utility!
Select from the options below:
_____
For Viral Sequences enter 1
For Host Sequences enter 2
For Help enter 3
For Privacy Notice enter 4
To exit enter 0 (zero)
```

<https://www.covid19dataportal.org/bulk-downloads>



wellcome  
connecting  
science

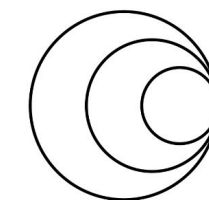


COVID-19  
GENOMICS  
GLOBAL TRAINING



# COVID19 Data Portal search and retrieval - exercise

1. Navigate to COVID-19 Data Portal: <https://www.covid19dataportal.org/>
2. Search for all sequences from a country of your choice  
*Filter by Severe acute respiratory syndrome coronavirus 2*
3. Note the different submitting centers/institutions
4. Which submitting center has contributed the most SARS-CoV-2 data for this country?
5. Repeat all steps for Raw Reads. What is the predominant type of sequencing here?

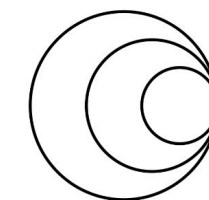


**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# Section 7: COVID19 Data Portal analysis & visualisation tools



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**

# ENA's large scale, systematic analysis of COVID reads

All public SARS-CoV-2 raw read data submitted to INSDC analysed according to [Illumina](#) or [Nanopore workflows](#)

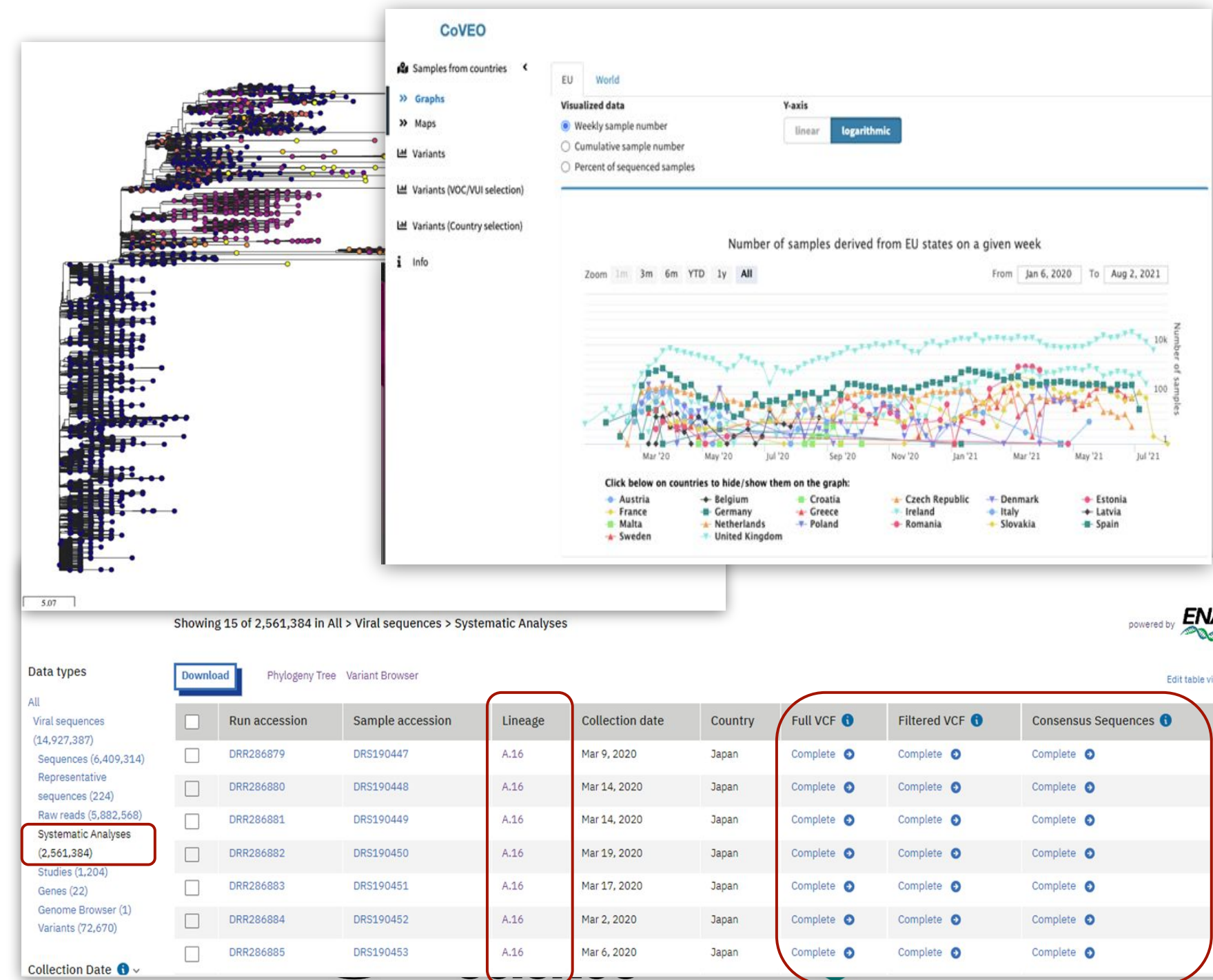
## 1. Consensus sequences

- Pangolin lineage assignment
- View on phylogeny tree

## 2. Variant calls

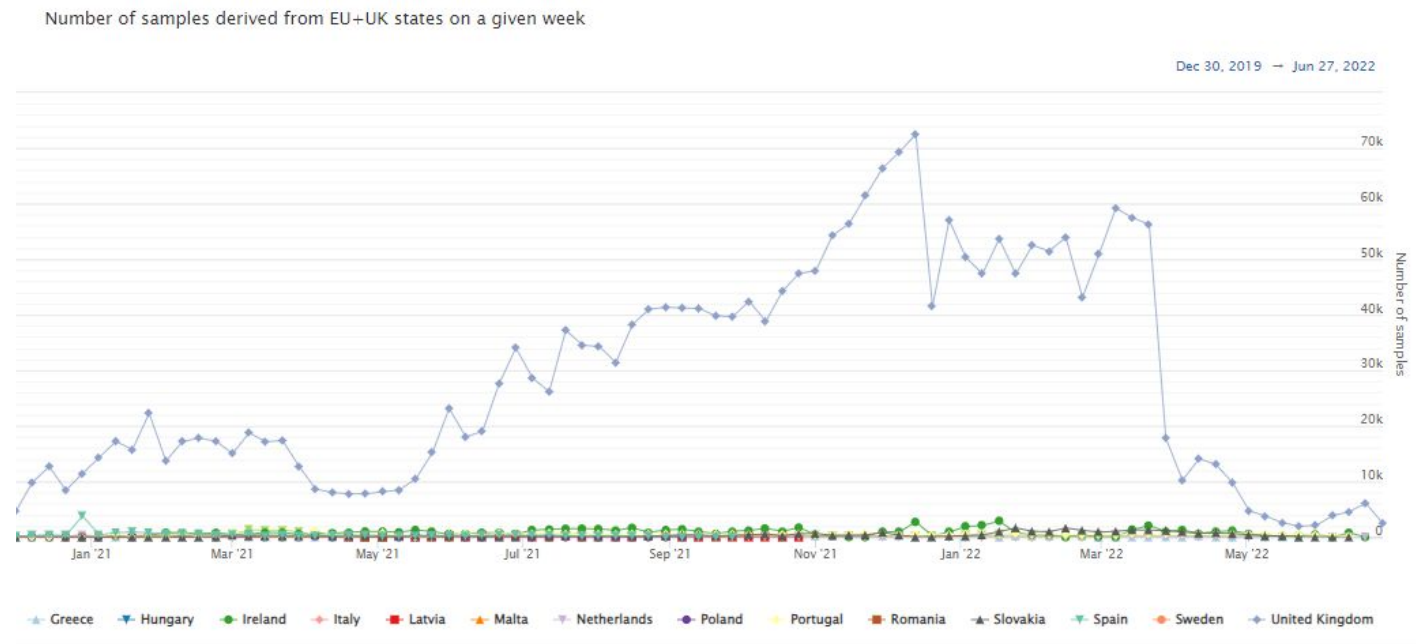
- Fed into CoVEO Variant Browser
- Submitted to European Variation Archive (EVA)

All products visualised on COVID-19 Data Portal



# CoVEO Variant Browser

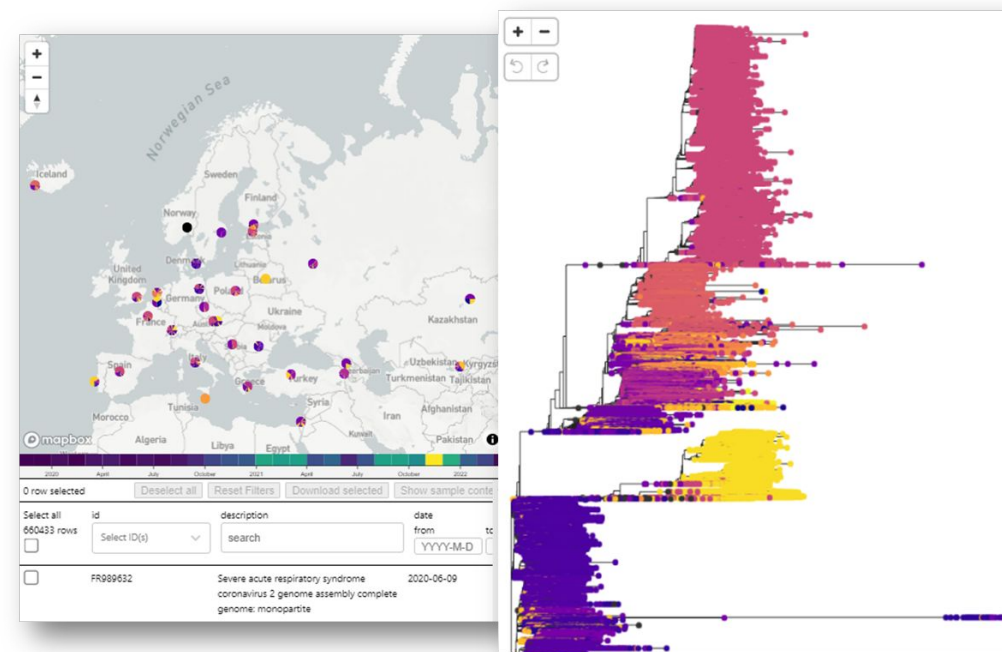
- CoVEO ingests unfiltered variant data to generate a range of plots
- Allows users to track emergence and distribution of SARS-CoV-2 variants across the world



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

# Phylogeny Tree

- Interactive phylogenetic tree built from public consensus sequences
- Features world map and metadata table, including filters on country and lineage



<https://www.covid19dataportal.org/phylogeny-tree>

[https://www.covid19dataportal.org/assets/pdf/evergreen\\_method\\_notes\\_2021-10-08.pdf](https://www.covid19dataportal.org/assets/pdf/evergreen_method_notes_2021-10-08.pdf)



COVID-19 GENOMICS GLOBAL TRAINING

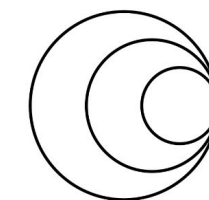


Technical University of Denmark



# CoVEO Variant Browser exercise:

1. Navigate to the CoVEO Explorer on the Covid-19 Data Portal:  
<https://www.covid19dataportal.org/coveo>
2. Under the generic 'Variants' facet on the left, select a country of your choice
3. What is the predominant variant/s in this country, across the full timeline?
4. Which 2 other countries have reported the highest prevalence of this variant overall?



**wellcome**  
**connecting**  
**science**

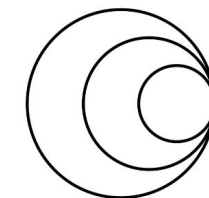


**COVID-19**  
**GENOMICS**  
**GLOBAL TRAINING**

**Thank you very much!**

**We hope you enjoyed  
the COG-Train  
sessions**

**:)**



**wellcome  
connecting  
science**



**COVID-19  
GENOMICS  
GLOBAL TRAINING**