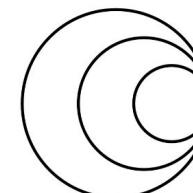


Module 4: Data Sharing and Interpretation

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



wellcome
connecting
science

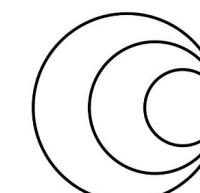


COVID-19
GENOMICS
GLOBAL TRAINING

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
 - Phylogenetics
 - Visualisation (microreact, nextstrain)
 - Data sharing, introduction to GISAID and ENA



wellcome
connecting
science



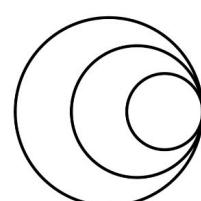
COVID-19
GENOMICS
GLOBAL TRAINING

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**

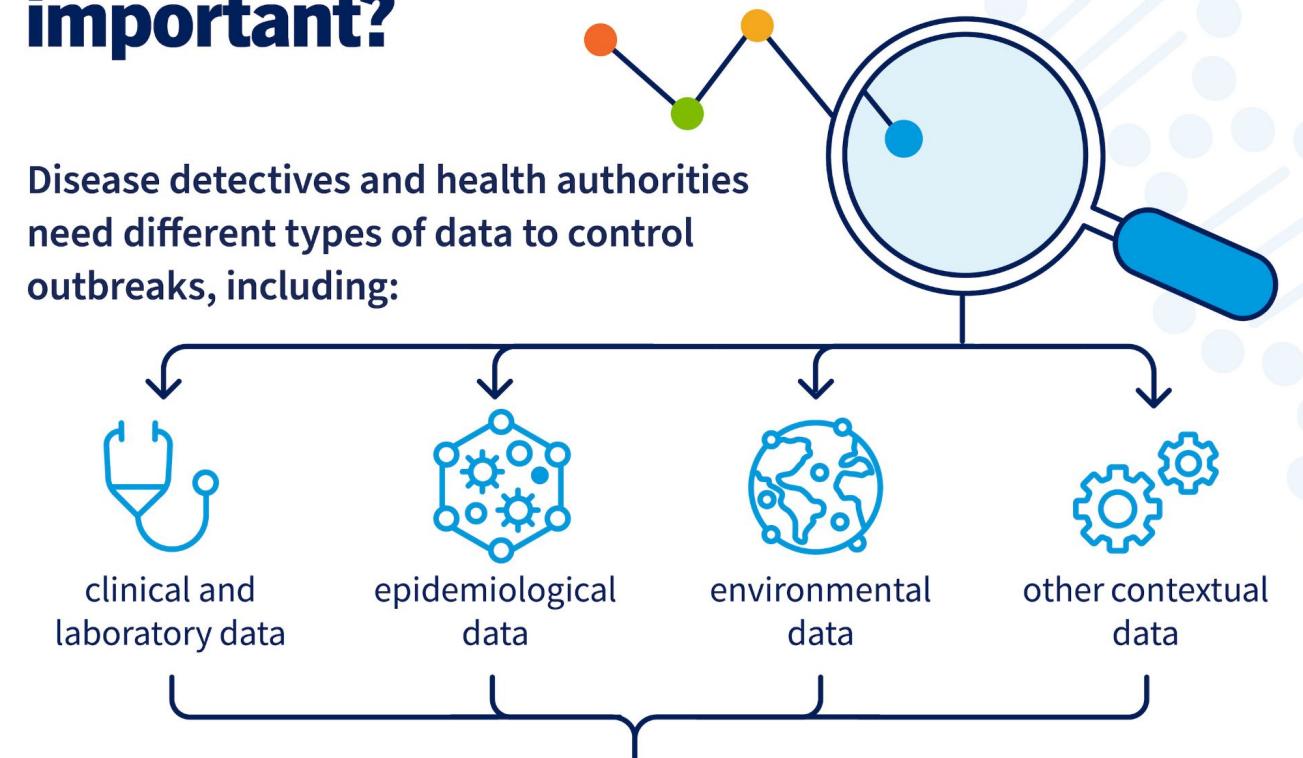
Open tools and protocols drove global surveillance

Enabled:

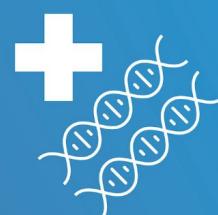
- Identification of variants of concern
- Understand transmission and immune evasion
- Vaccine and diagnostic development

Why is GENOMIC SURVEILLANCE important?

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

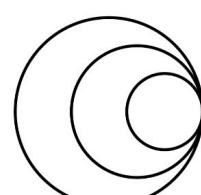


By adding genomic data, they can more quickly understand how a pathogen behaves and how to control it.



This is a powerful tool in public health surveillance.

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



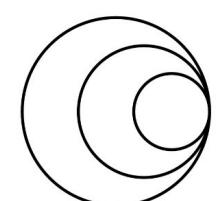
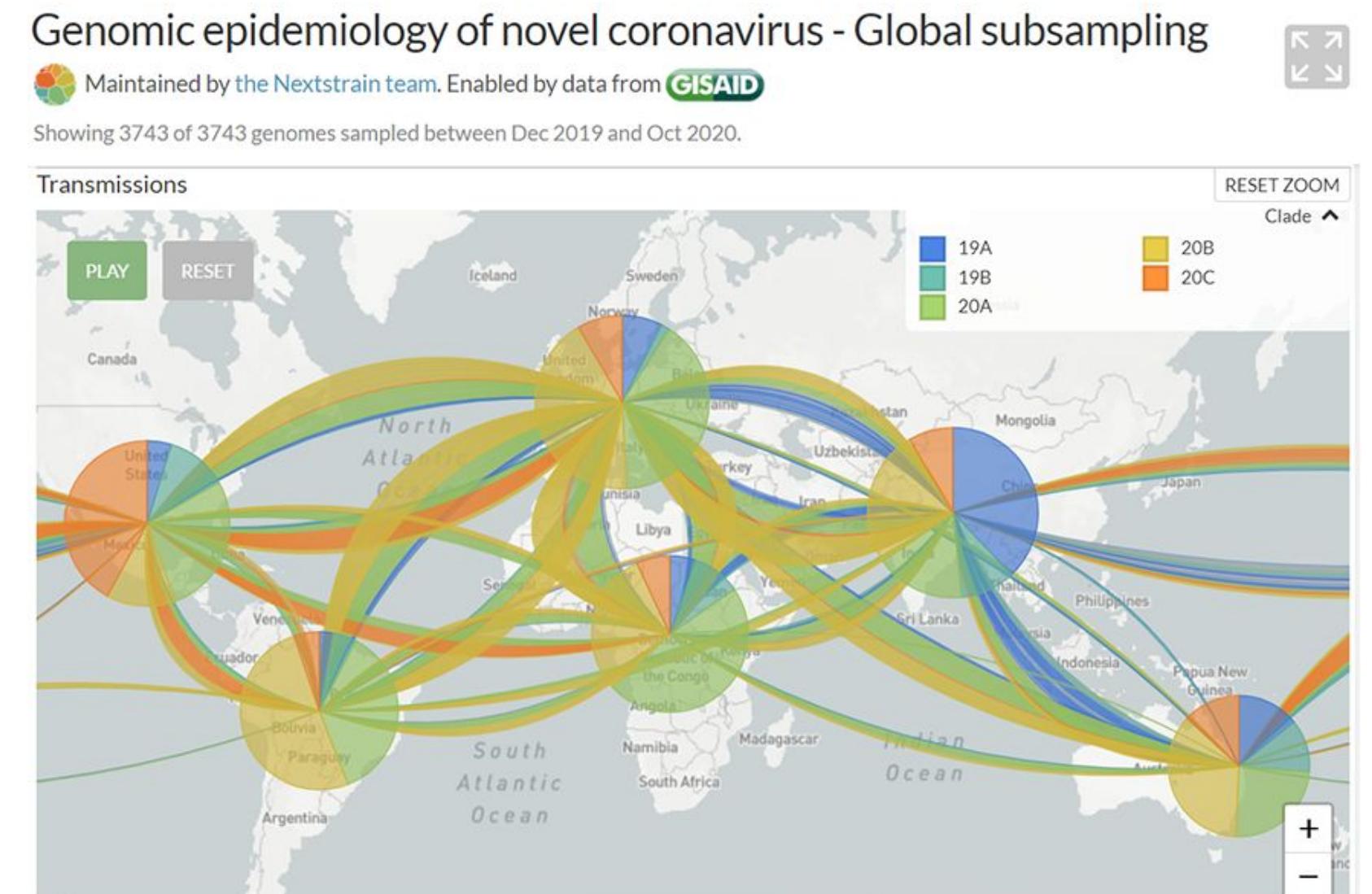
wellcome
connecting
science



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

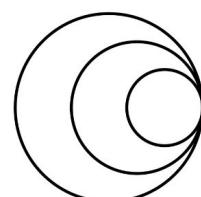


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Section 1: Intro to phylogenetics



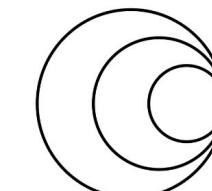
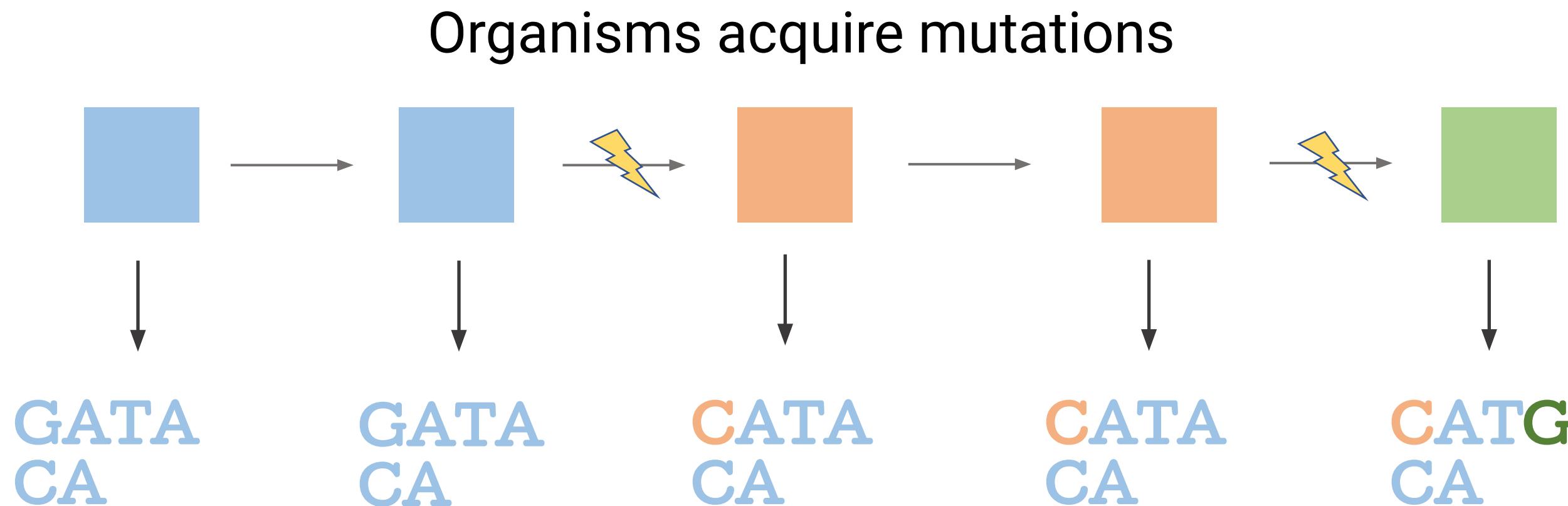
wellcome
connecting
science



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.

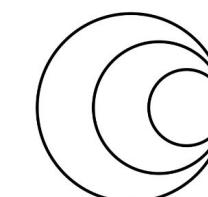
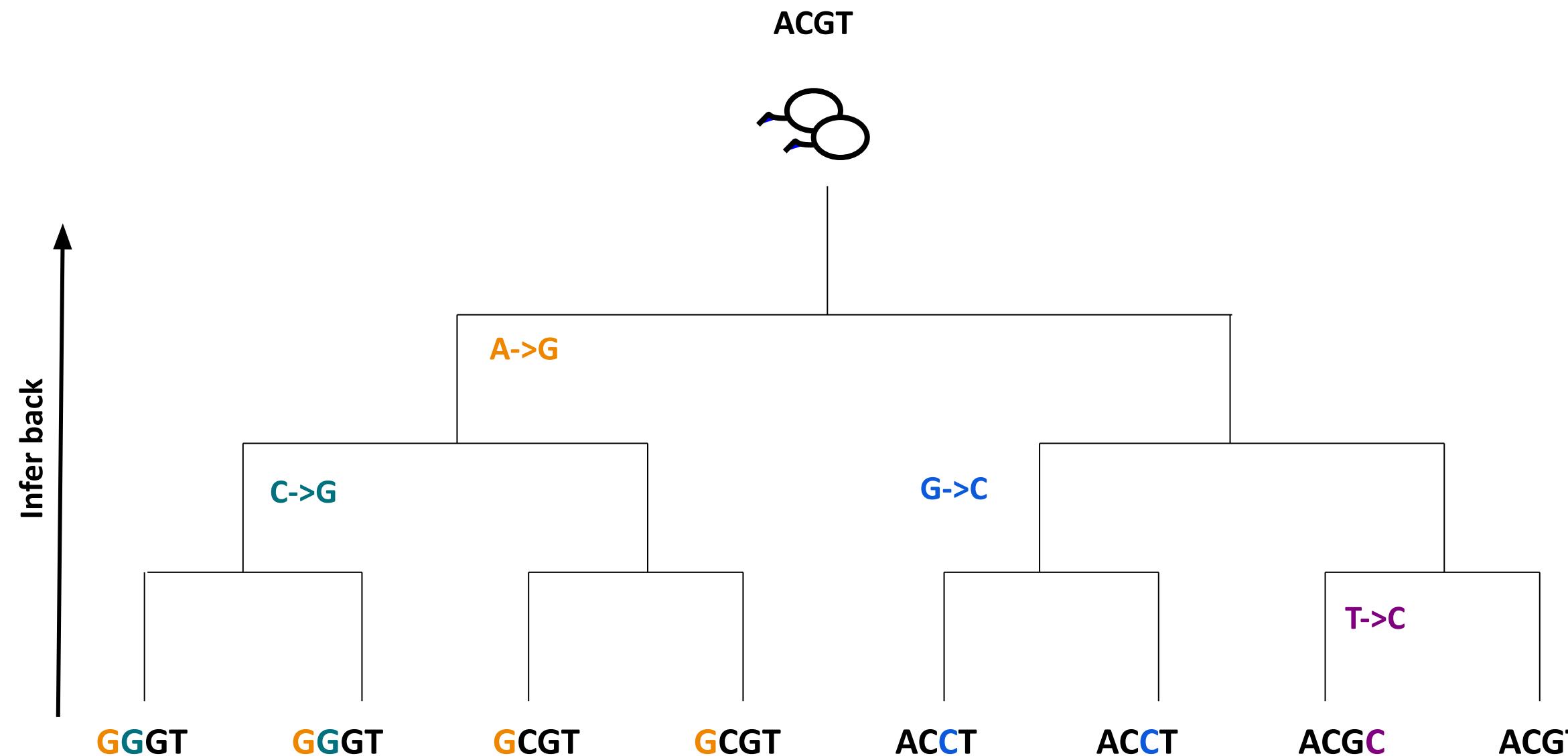


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Mutations tell us about relationships



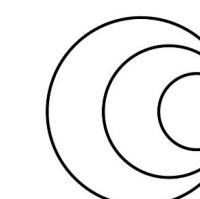
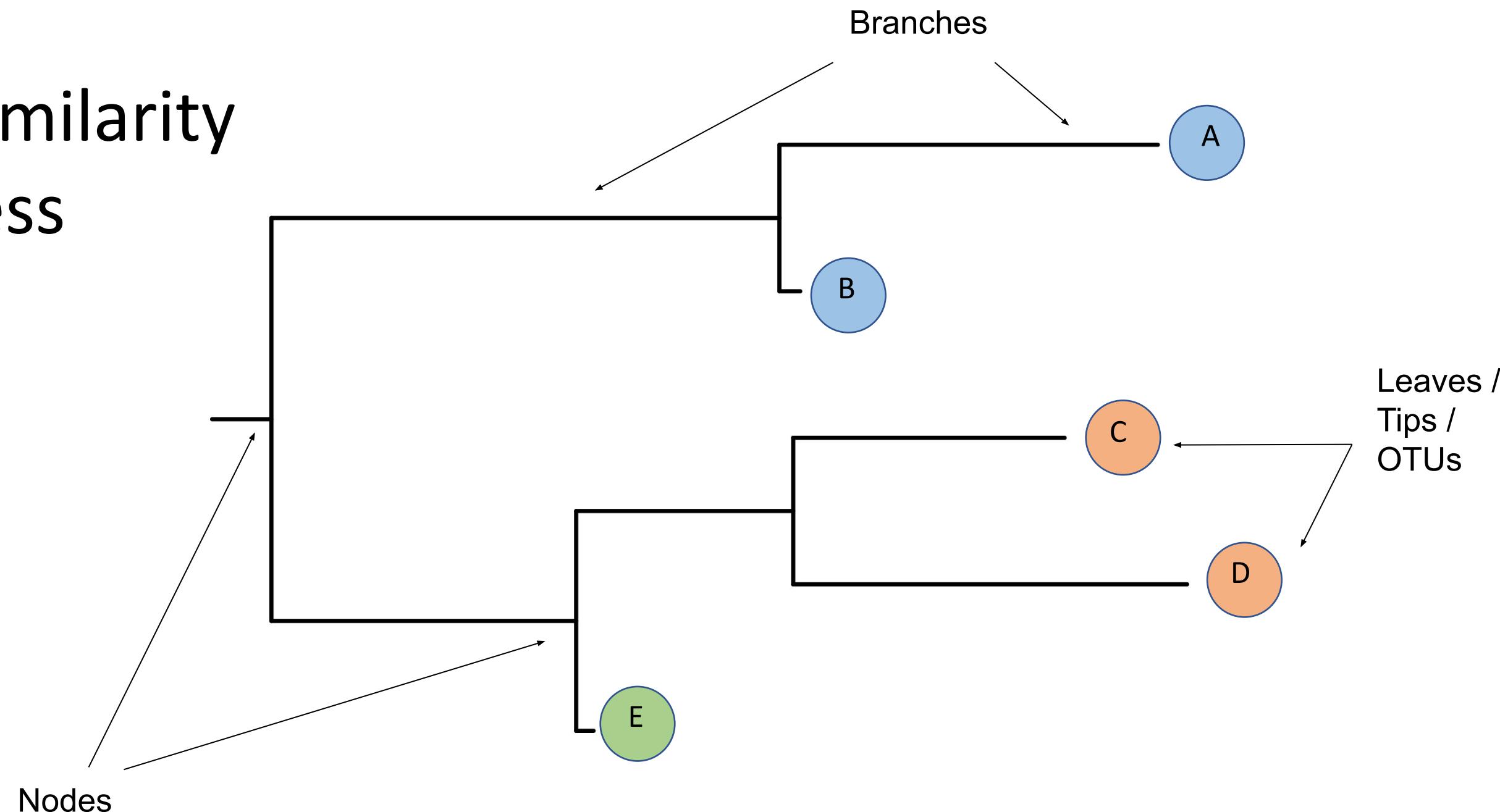
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Phylogenetic trees reveal relationships

Genetic similarity
Relatedness

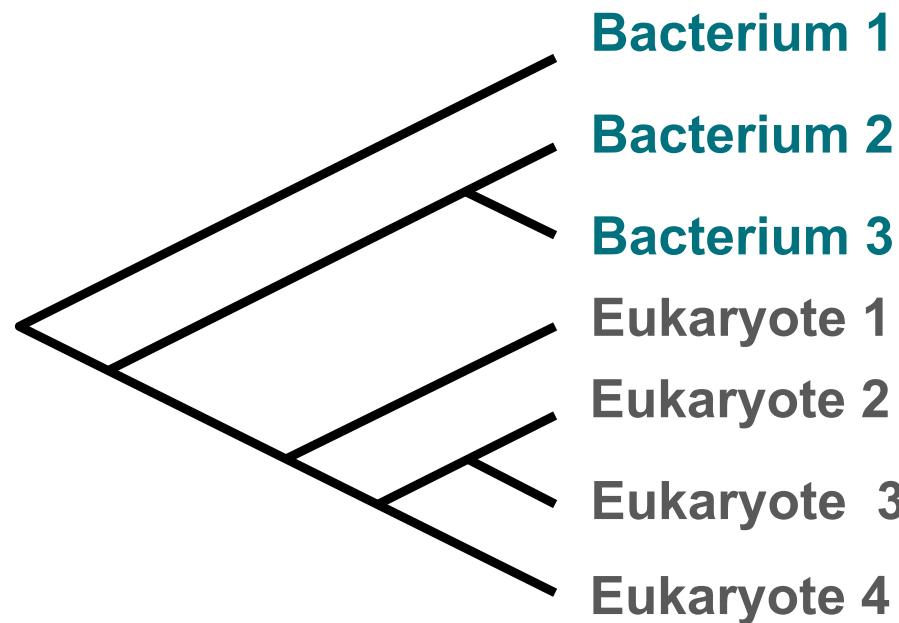


wellcome
connecting
science

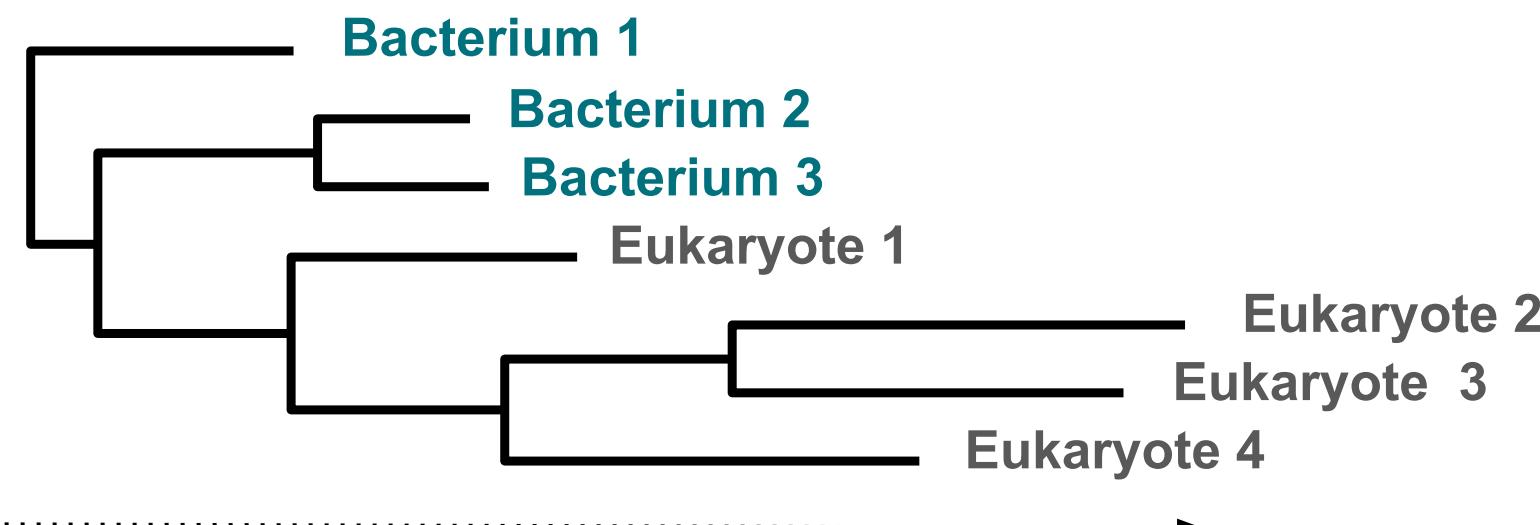


COVID-19
GENOMICS
GLOBAL TRAINING

Cladograms vs Phylogenograms

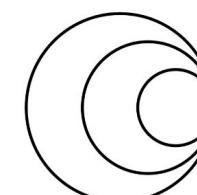


Cladograms show
branch order (topology)
only - branch lengths are
meaningless



Phylogenograms show
branch order and
branch lengths
with scale

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

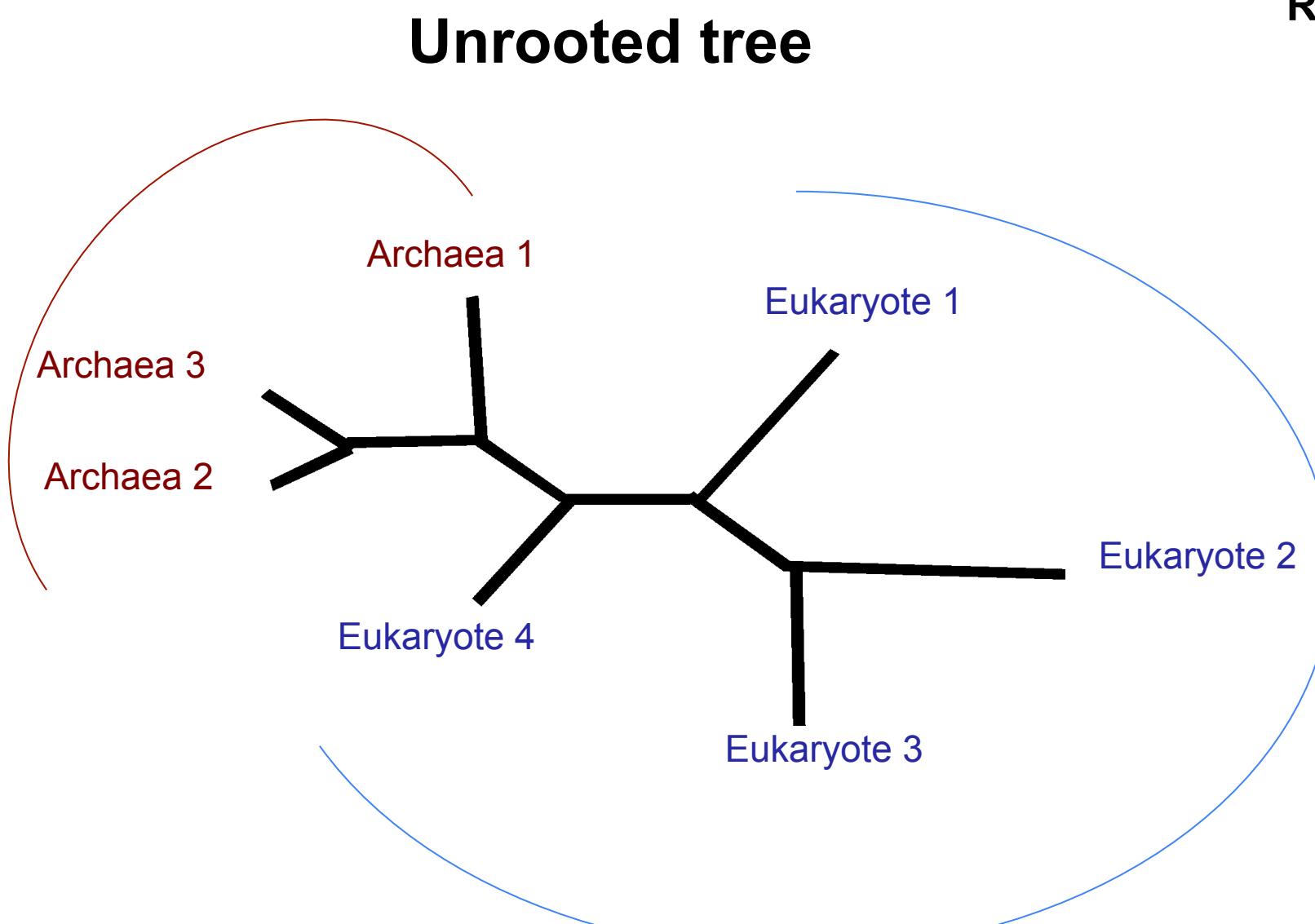


wellcome
connecting
science

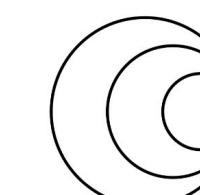
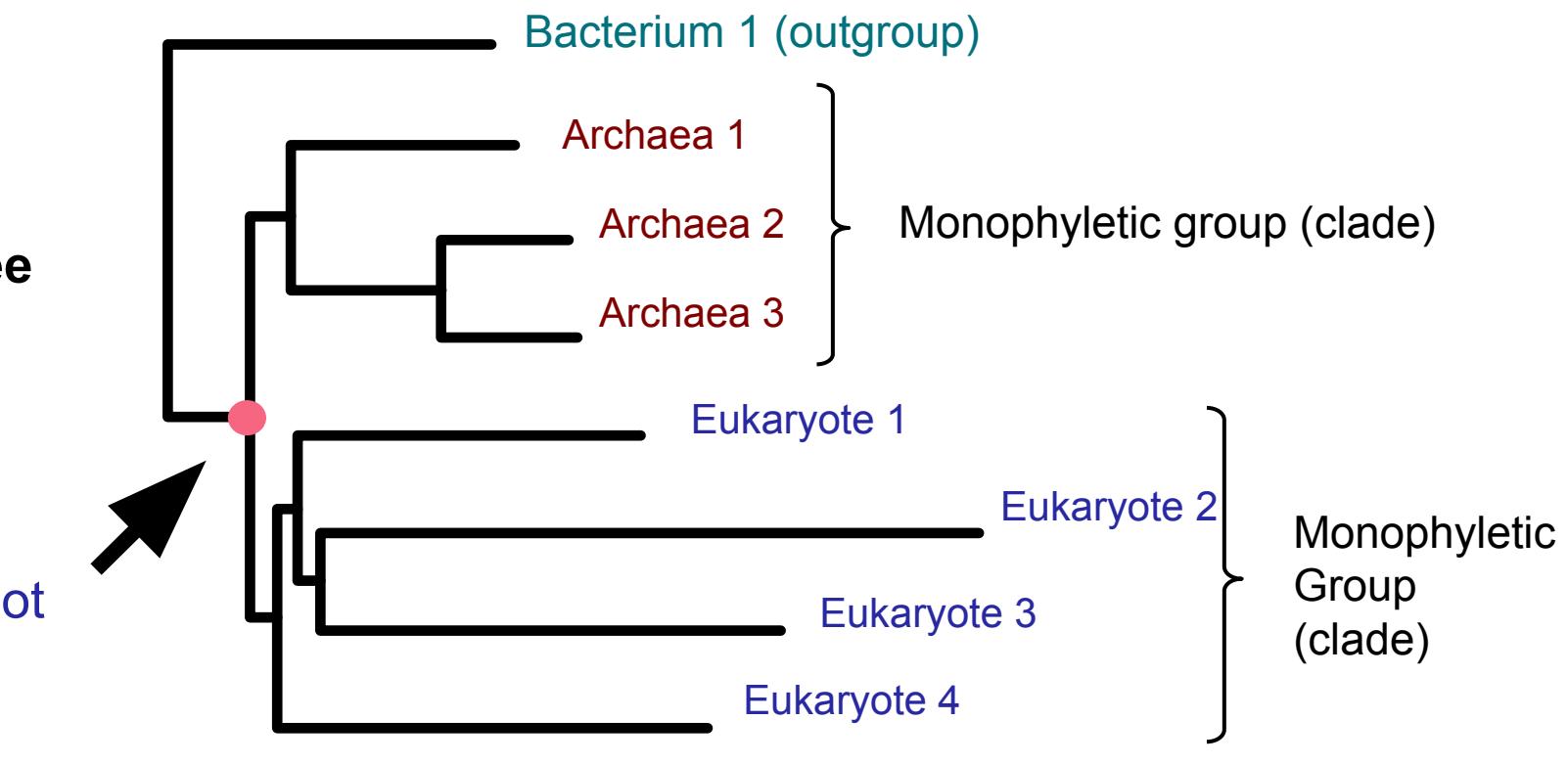


COVID-19
GENOMICS
GLOBAL TRAINING

Rooted and Unrooted trees



Rooted tree
(by using
outgroup)



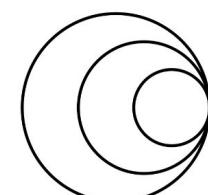
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

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

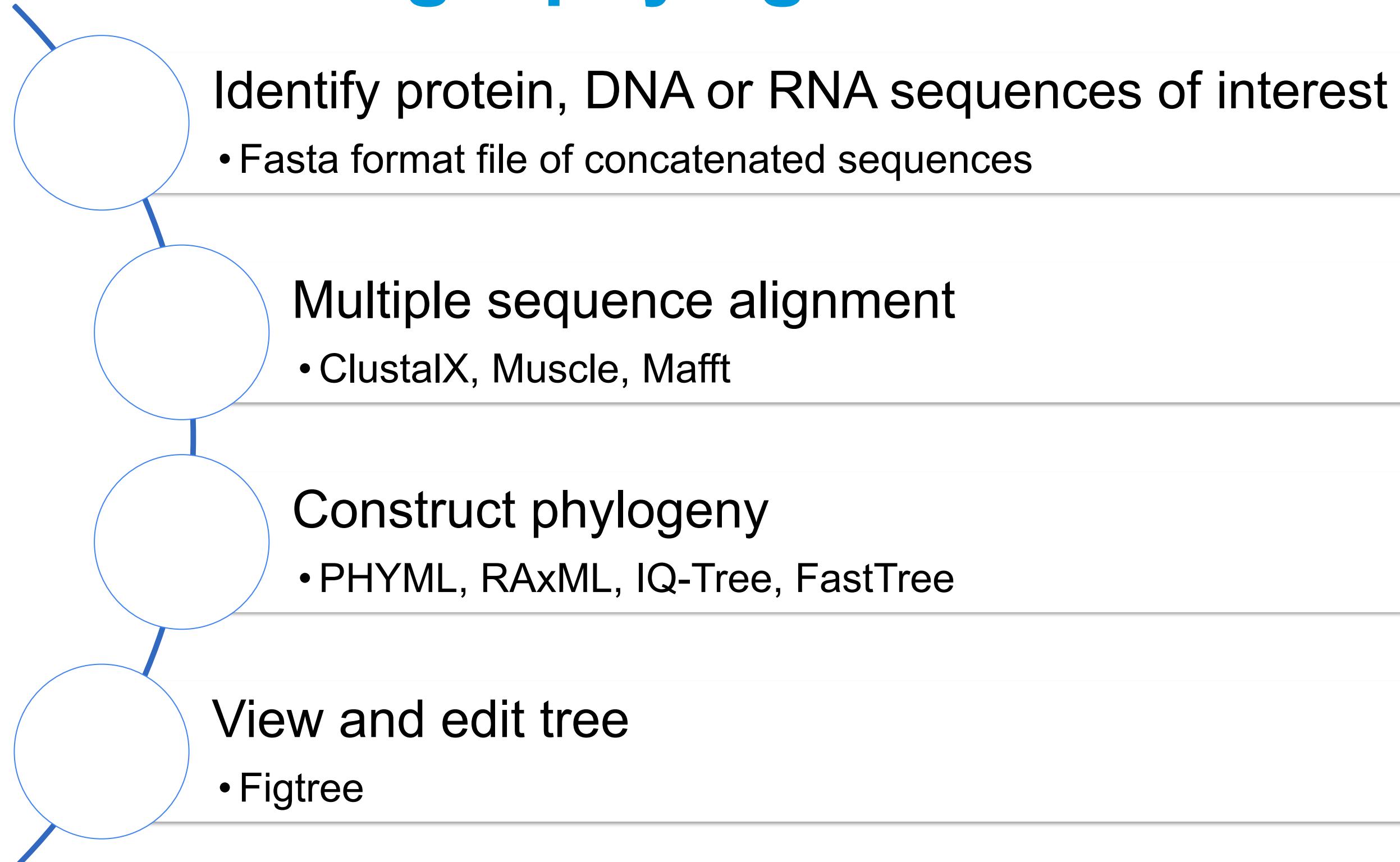


wellcome
connecting
science



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

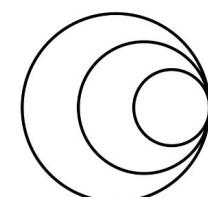
- ClustalX, Muscle, Mafft

Construct phylogeny

- PHYML, RAxML, IQ-Tree, FastTree

View and edit tree

- Figtree



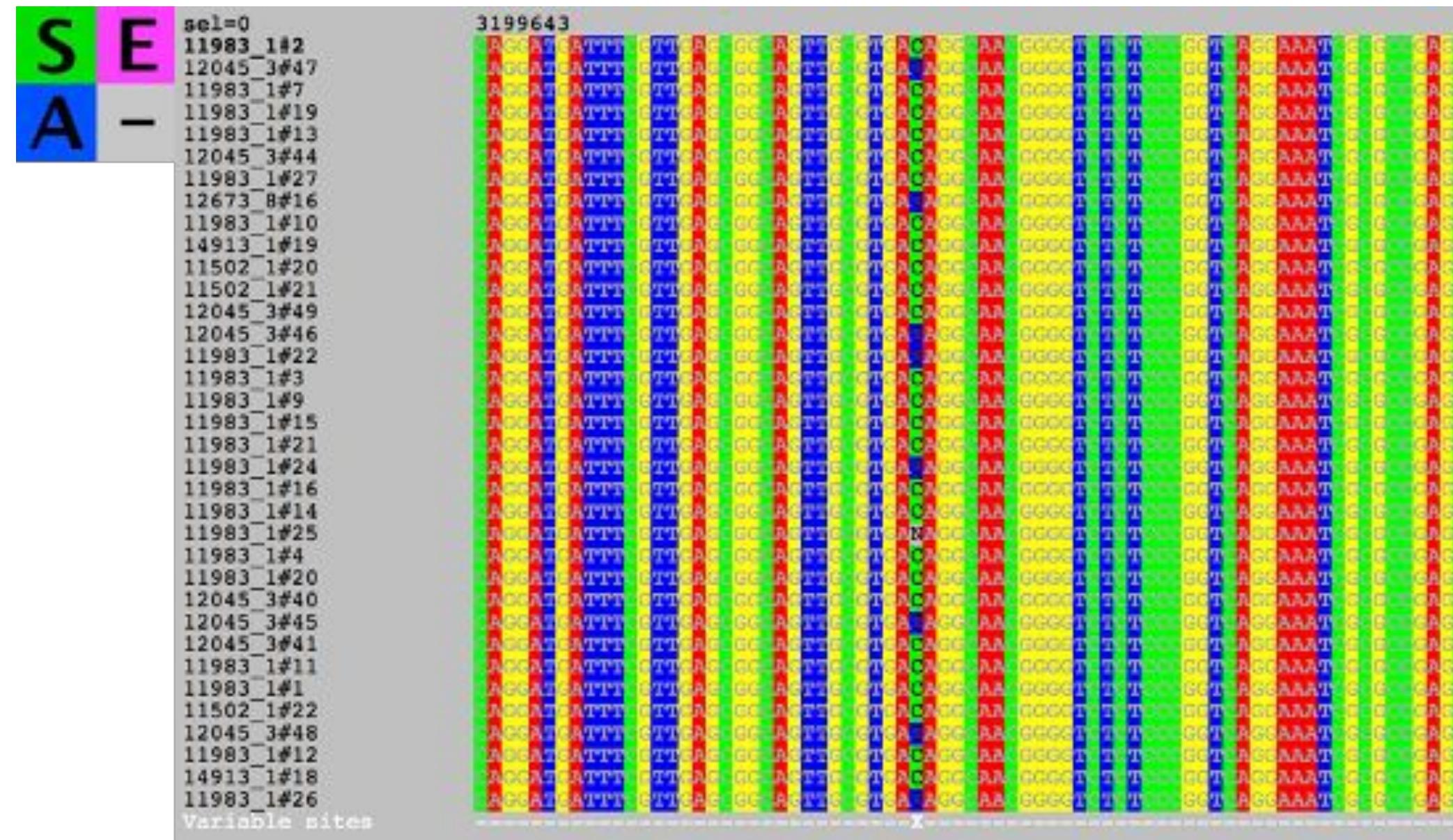
wellcome
connecting
science



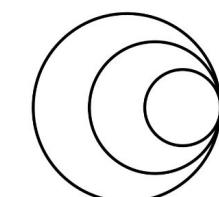
COVID-19
GENOMICS
GLOBAL TRAINING

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

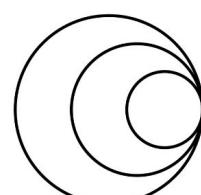


COVID-19
GENOMICS
GLOBAL TRAINING

Constructing a phylogenetic tree

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

* attempt to find the BEST tree



wellcome
connecting
science

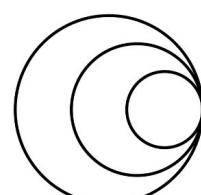


COVID-19
GENOMICS
GLOBAL TRAINING

Constructing a phylogenetic tree

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

* attempt to find the BEST tree



wellcome
connecting
science



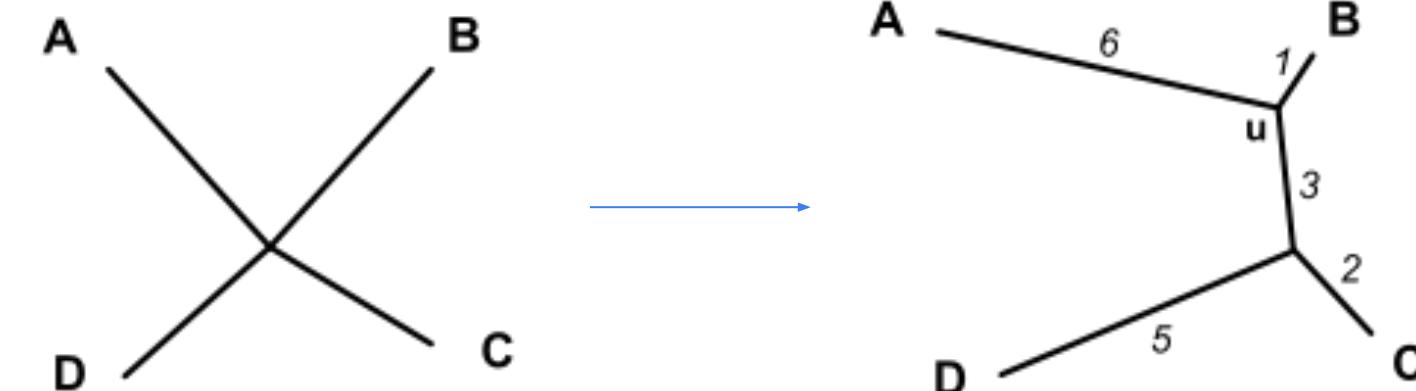
COVID-19
GENOMICS
GLOBAL TRAINING

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



wellcome
connecting
science

e.g. UPGMA, Neighbour joining, minimum evolution, BIONJ

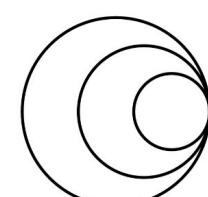


COVID-19
GENOMICS
GLOBAL TRAINING

Constructing a phylogenetic tree

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

* attempt to find the BEST tree

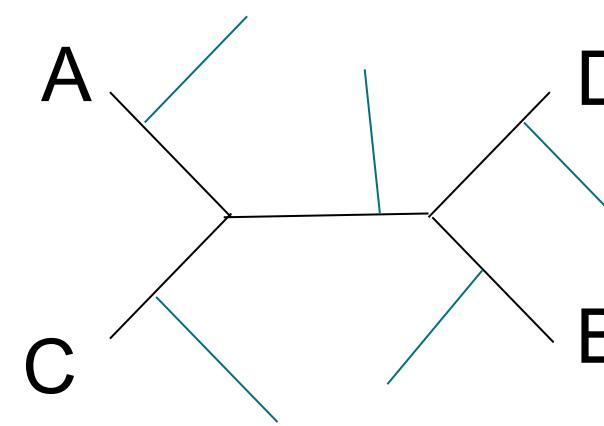
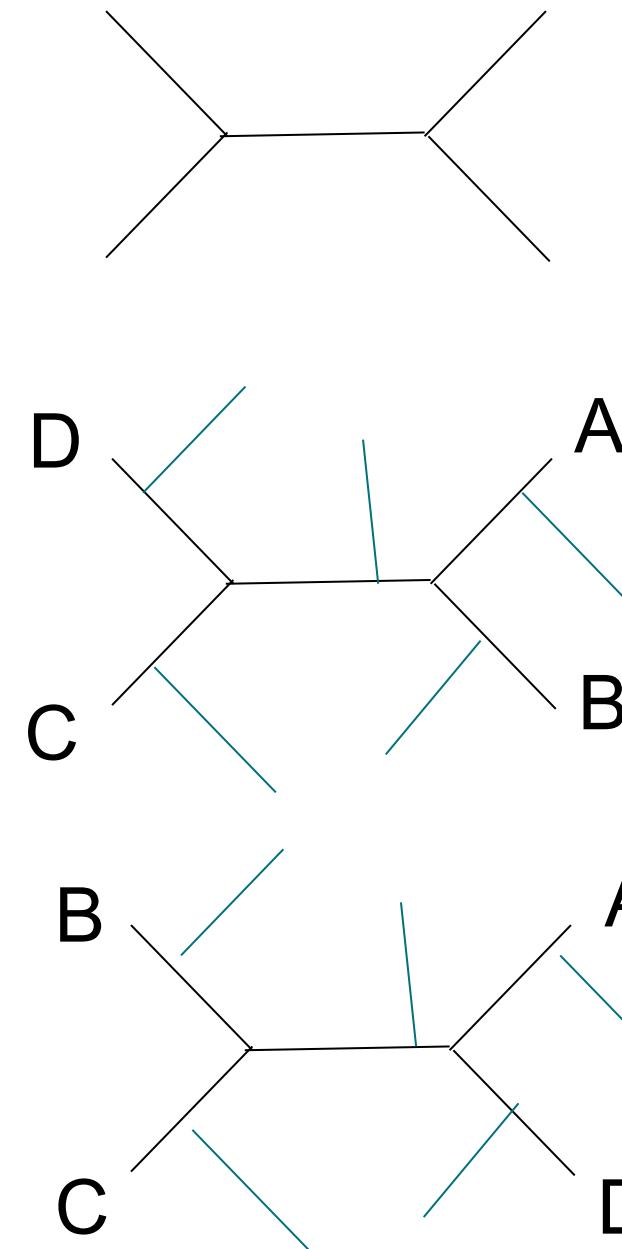


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Tree searching algorithms

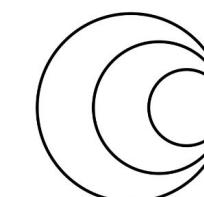


E?

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	2.18×10^{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

Simple



JC69:

all substitutions equally likely,
all bases equally frequent.

JC69+I+ Γ :

as for JC69, but with additional parameters
for invariant sites and gamma distribution.

K2P:

specific probabilities for transitions and transversions,
all bases equally frequent.

HKY85:

specific probabilities for transitions and transversions,
specific base frequencies.

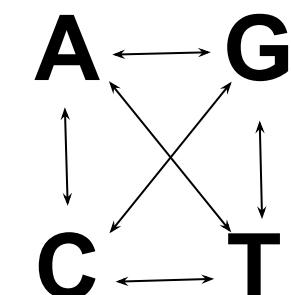
GTR:

each substitution has a specific probability,
moderated by specific base frequencies.

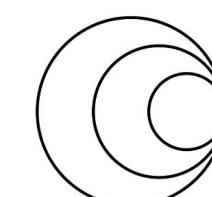
GTR+I+ Γ :

as for GTR, but with additional parameters
for invariant sites and gamma distribution.

Complex



4 equilibrium base
frequency parameters and
6 substitution rate
parameters



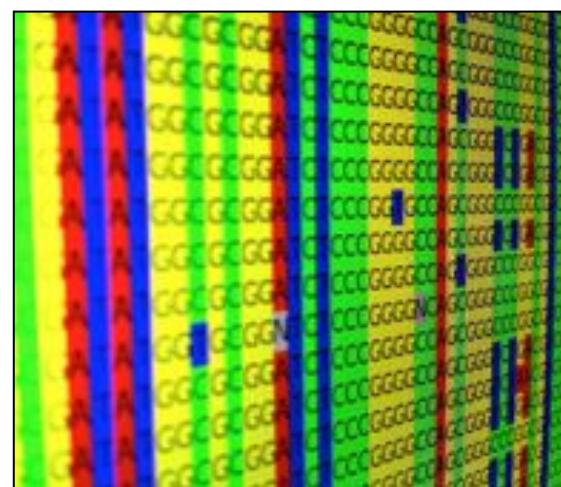
wellcome
connecting
science



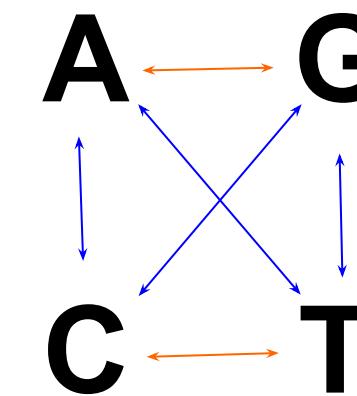
COVID-19
GENOMICS
GLOBAL TRAINING

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

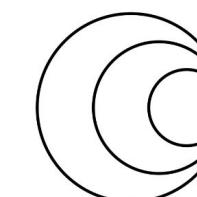
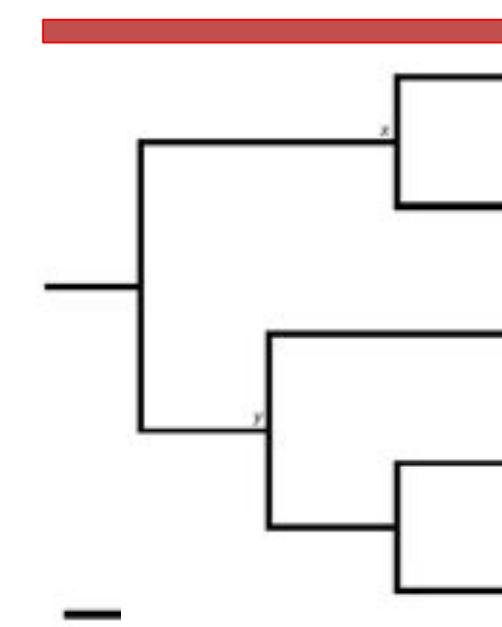
these data...



... if this happens...



... over this tree



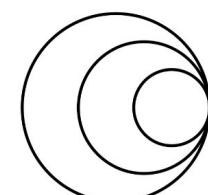
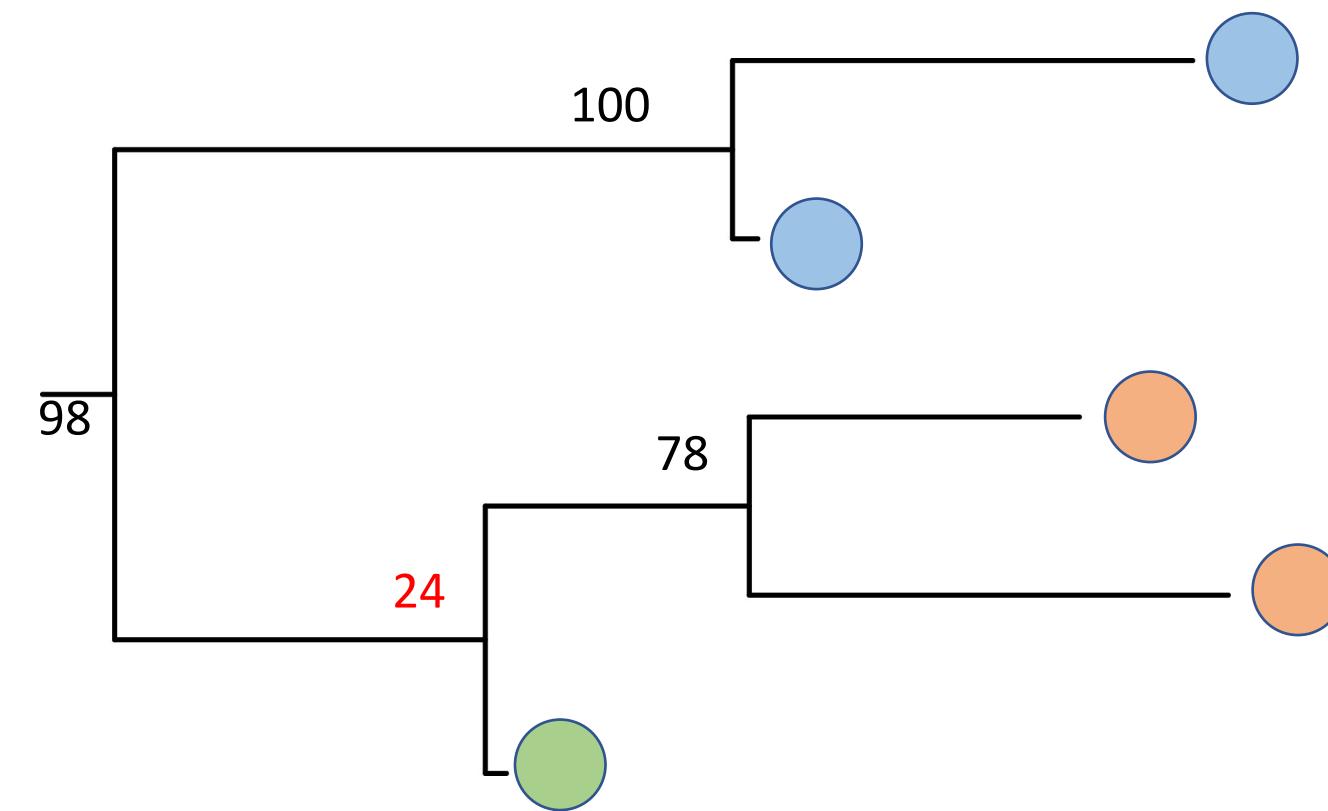
wellcome
connecting
science



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

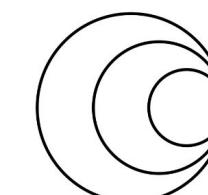
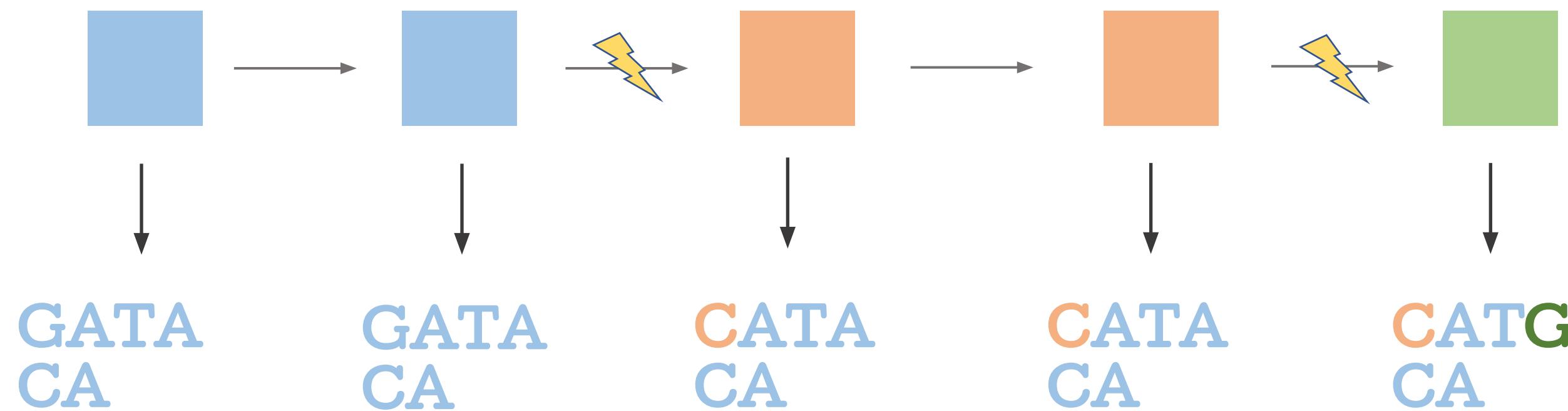


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Pathogens mutate as they transmit

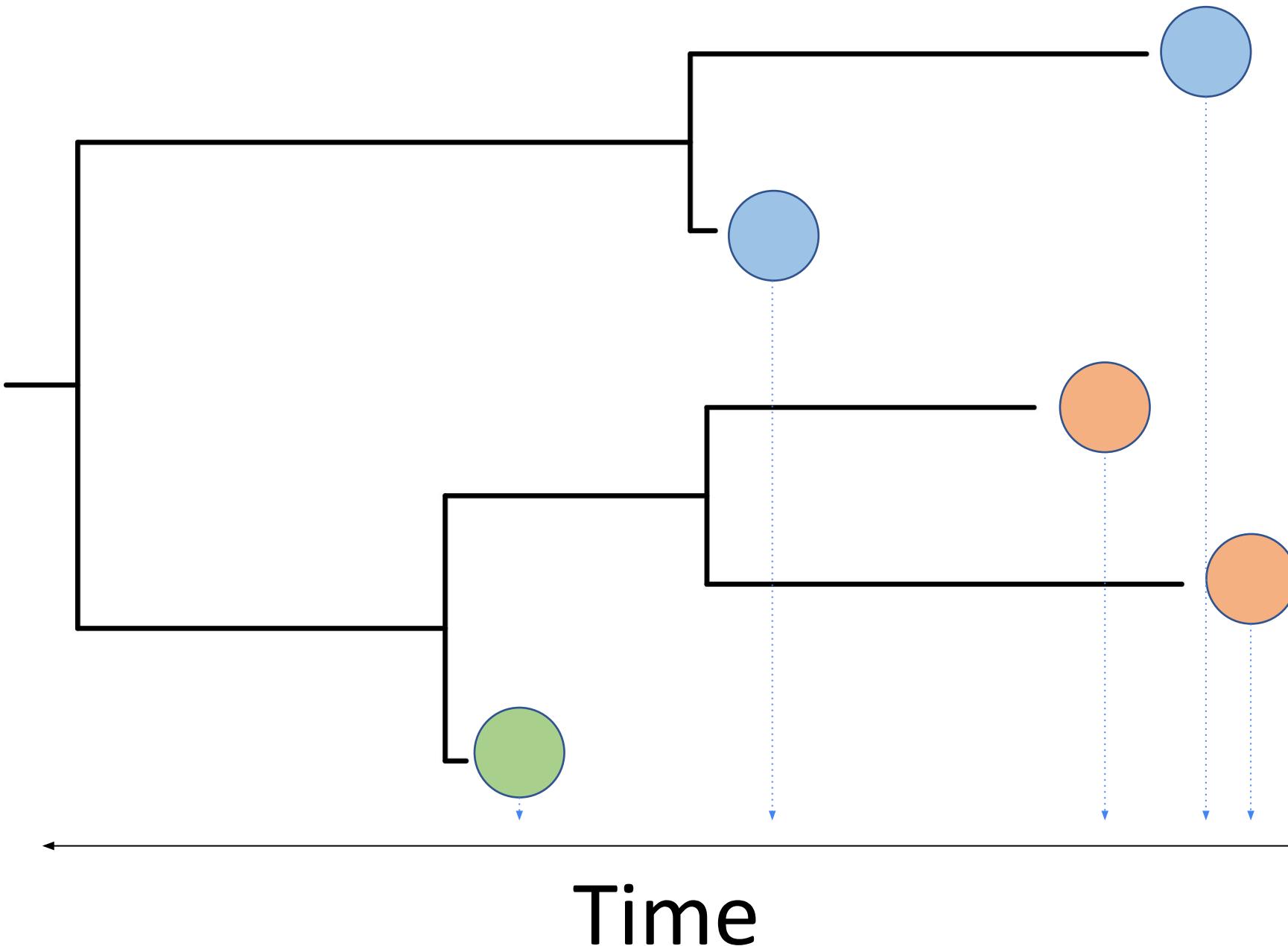


wellcome
connecting
science

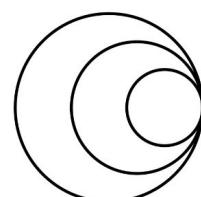


COVID-19
GENOMICS
GLOBAL TRAINING

Trees reveal timing



Typically use BEAST to generate



wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Section 2: Interpreting phylogenetic analysis

Some Resources:

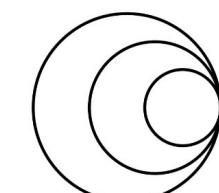
Understanding Evolutionary Trees, <https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-008-0035-x>

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

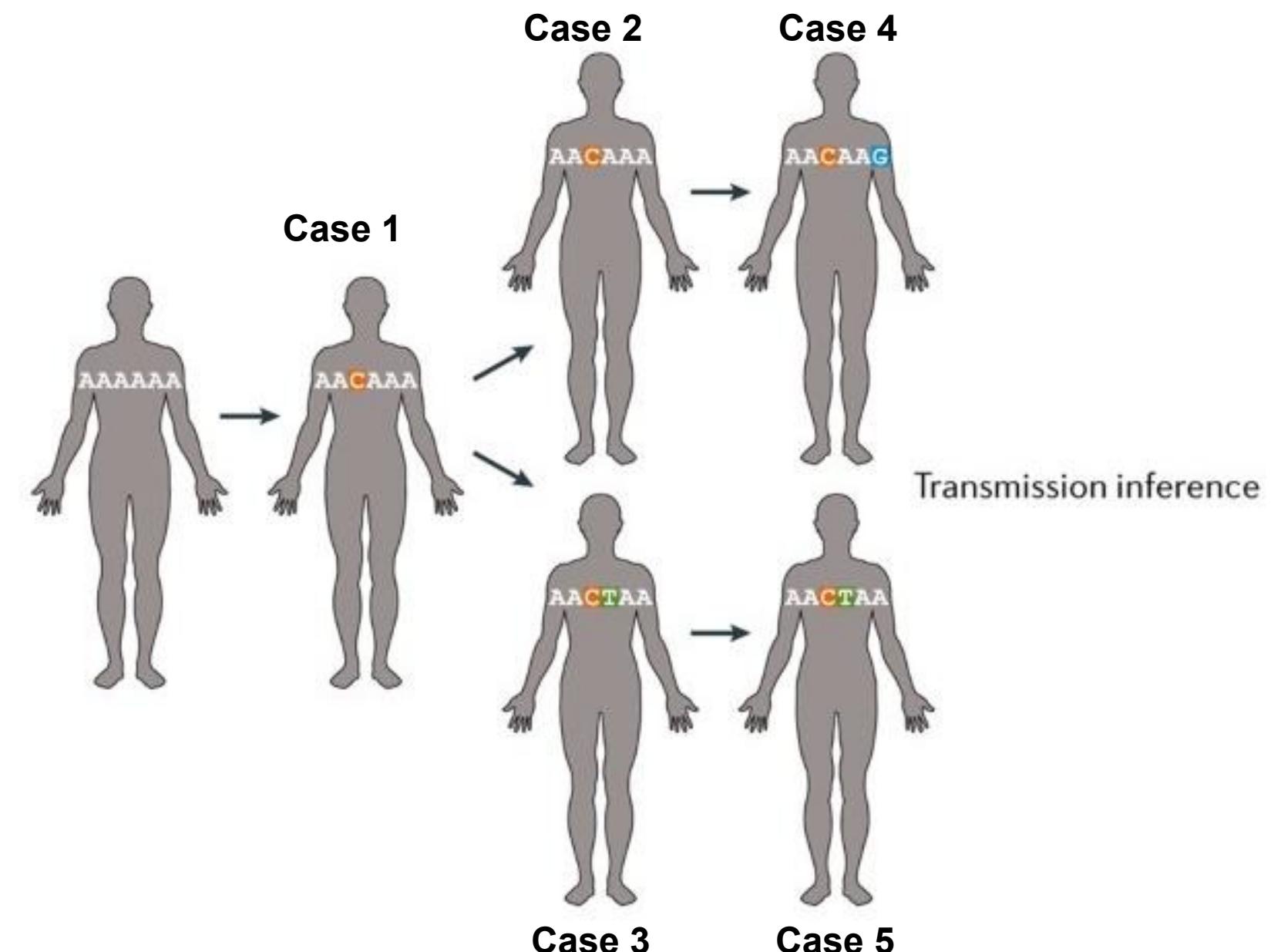


**COVID-19
GENOMICS
GLOBAL TRAINING**

Sequence relatedness can be used to infer transmission

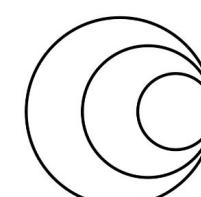
Virus replicate, random mutations occur

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



Nature Reviews | Genetics

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

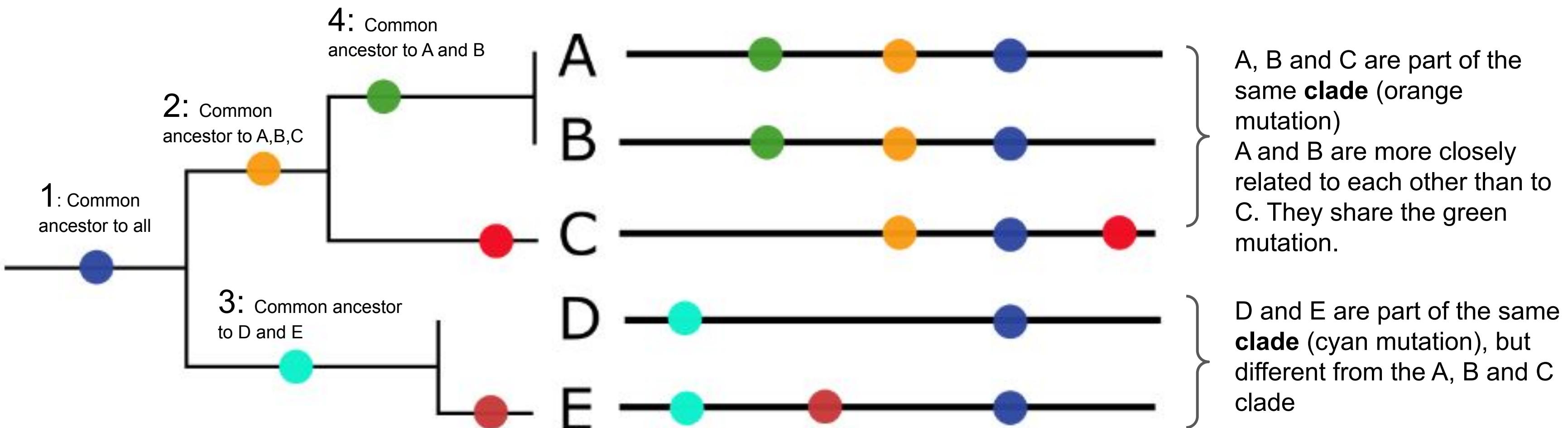


wellcome
connecting
science

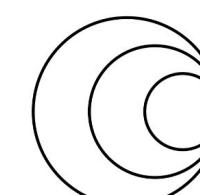


COVID-19
GENOMICS
GLOBAL TRAINING

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

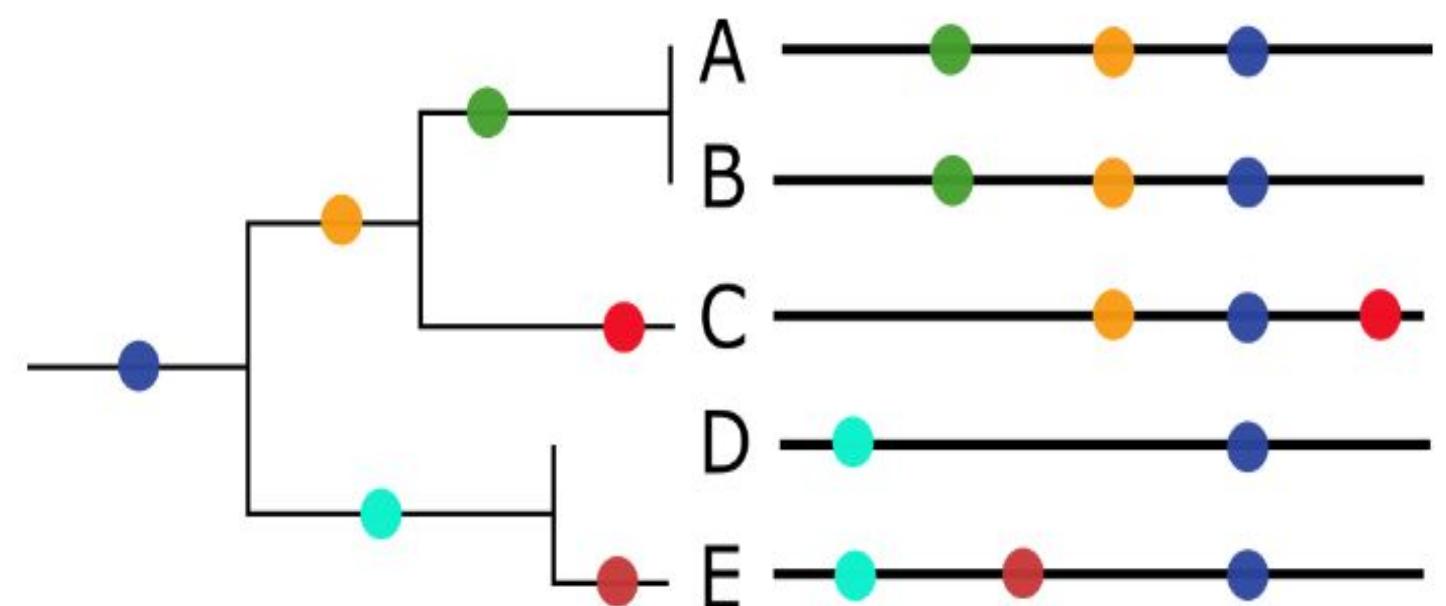


COVID-19
GENOMICS
GLOBAL TRAINING

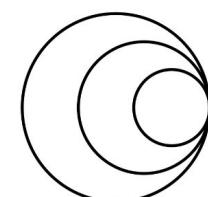
Phylogeny can be used to generate hypothesis about transmission

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

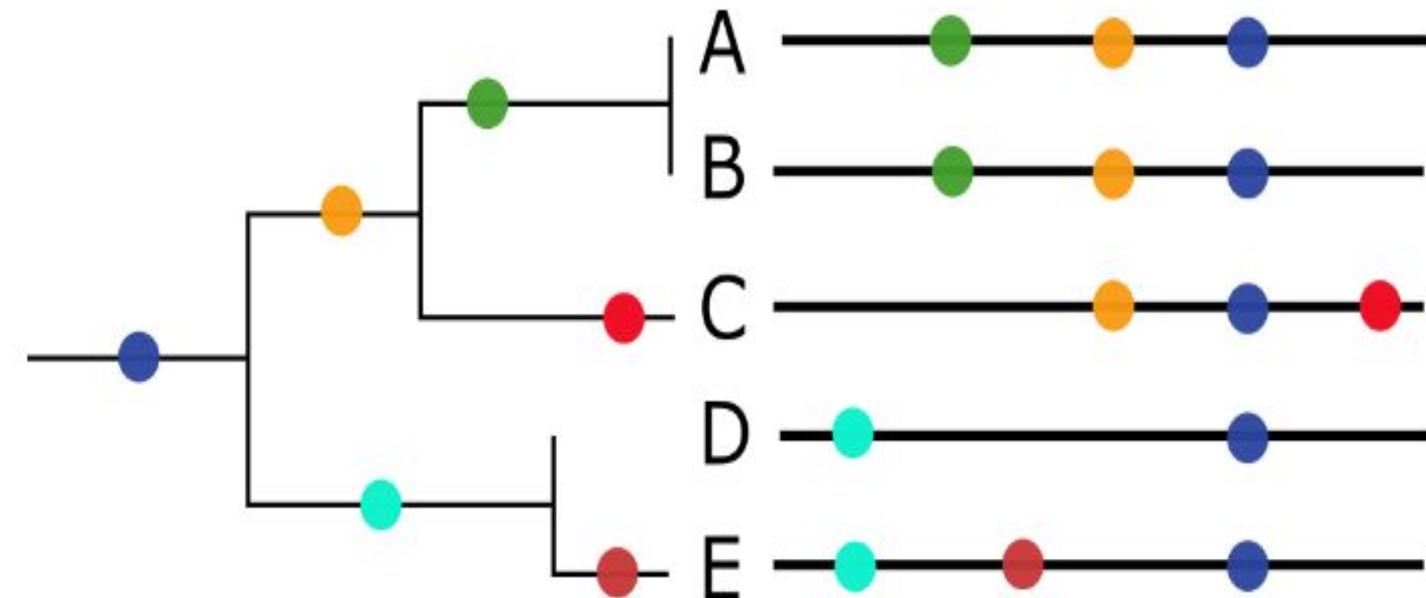


COVID-19
GENOMICS
GLOBAL TRAINING

Phylogenetics can assist epidemiological investigations related to outbreaks

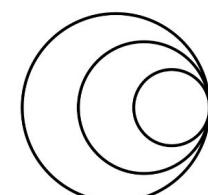
By refining outbreak by ruling in or out individuals

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/>

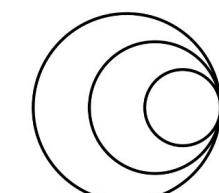
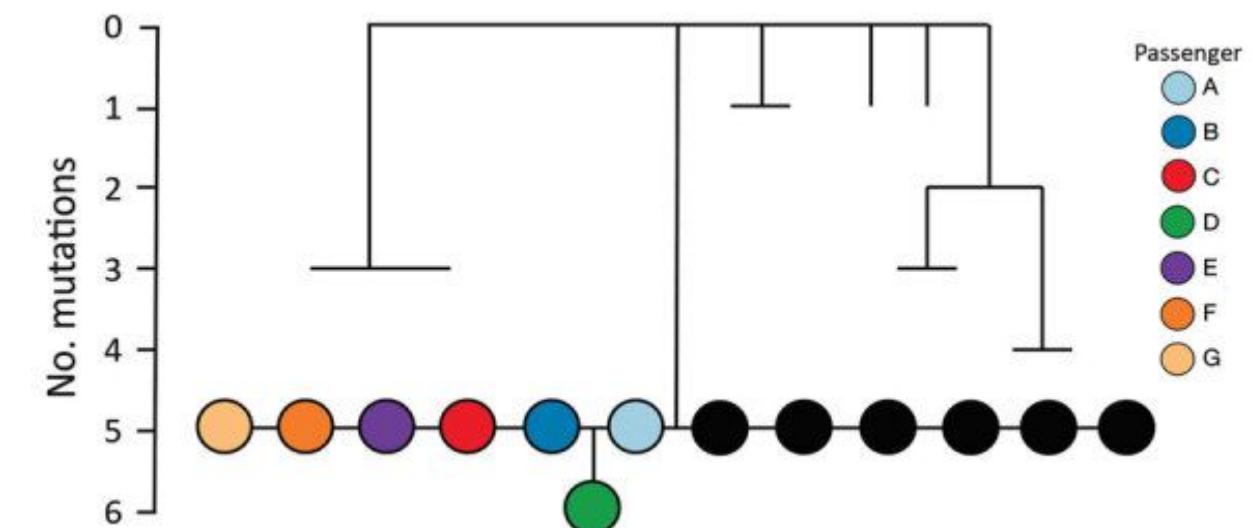
Dataset_1_flight

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

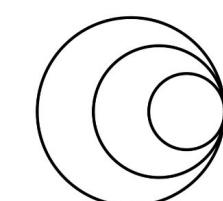
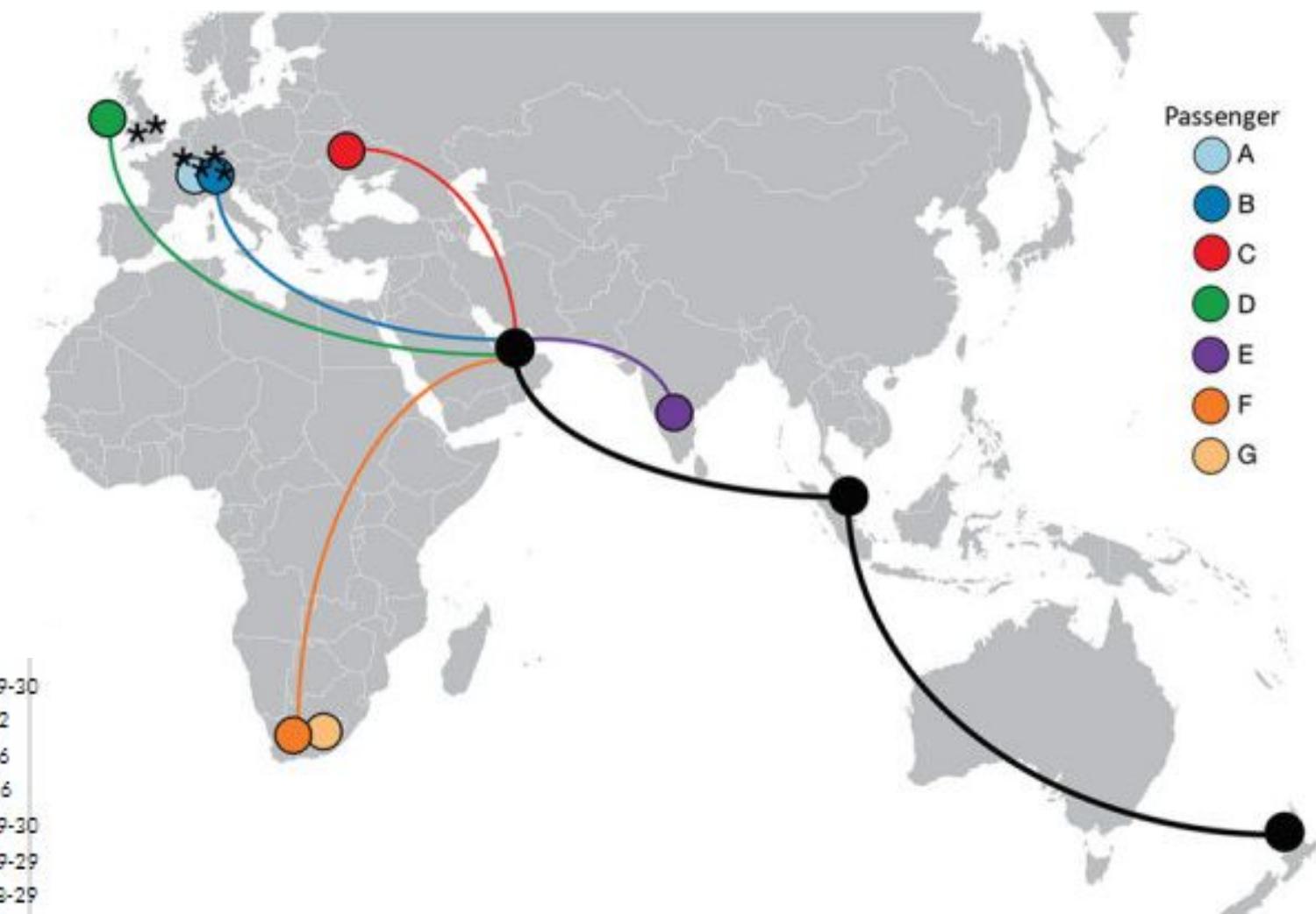
Example 1: inflight transmission?

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.

Switzerland/VD-ETHZ-300250/2020 EPI_ISL_603282 2020-09-30
Switzerland/SO-ETHZ-500136/2020 OU000186.1 2020-11-12
Switzerland/AG-ETHZ-340116/2020 OD997195.1 2020-10-26
Switzerland/VD-ETHZ-340113/2020 OU007787.1 2020-10-26
Switzerland/VD-ETHZ-300425/2020 EPI_ISL_603382 2020-09-30
Switzerland/VD-ETHZ-300423/2020 EPI_ISL_603381 2020-09-29
Switzerland/ZH-ETHZ-270170/2020 EPI_ISL_541529 2020-08-29
Switzerland/BE-ETHZ-400341/2020 OD998810.1 2020-12-08
Switzerland/ZH-ETHZ-270081/2020 EPI_ISL_541516 2020-09-02
Switzerland/ZH-ETHZ-0000068/2020
Switzerland/VD-ETHZ-270050/2020 EPI_ISL_541470 2020-09-01
Germany/BAM-PL-virotum_LSEIF/2021 EPI_ISL_
Switzerland/VD-CHUV-GEN2429/2020
● hCoV-19/New Zealand/20CV0398/2020
Switzerland/VD-ETHZ-531538/2020 OU193410.1 2020-11-12
Switzerland/VD-ETHZ-310353/2020 EPI_ISL_603427 2020-10-07
Switzerland/VD-ETHZ-300469/2020 EPI_ISL_603416 2020-09-30
Switzerland/VD-ETHZ-300467/2020 EPI_ISL_603414 2020-09-30
Switzerland/SO-ETHZ-430540/2020 OE999357.1 2020-12-29
Switzerland/VD-ETHZ-321364/2020 EPI_ISL_603613 2020-10-13
Switzerland/VD-ETHZ-350197/2020 OU007795.1 2020-10-30
Switzerland/BE-ETHZ-360092/2020 OD998635.1 2020-11-08



wellcome
connecting
science

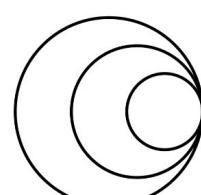
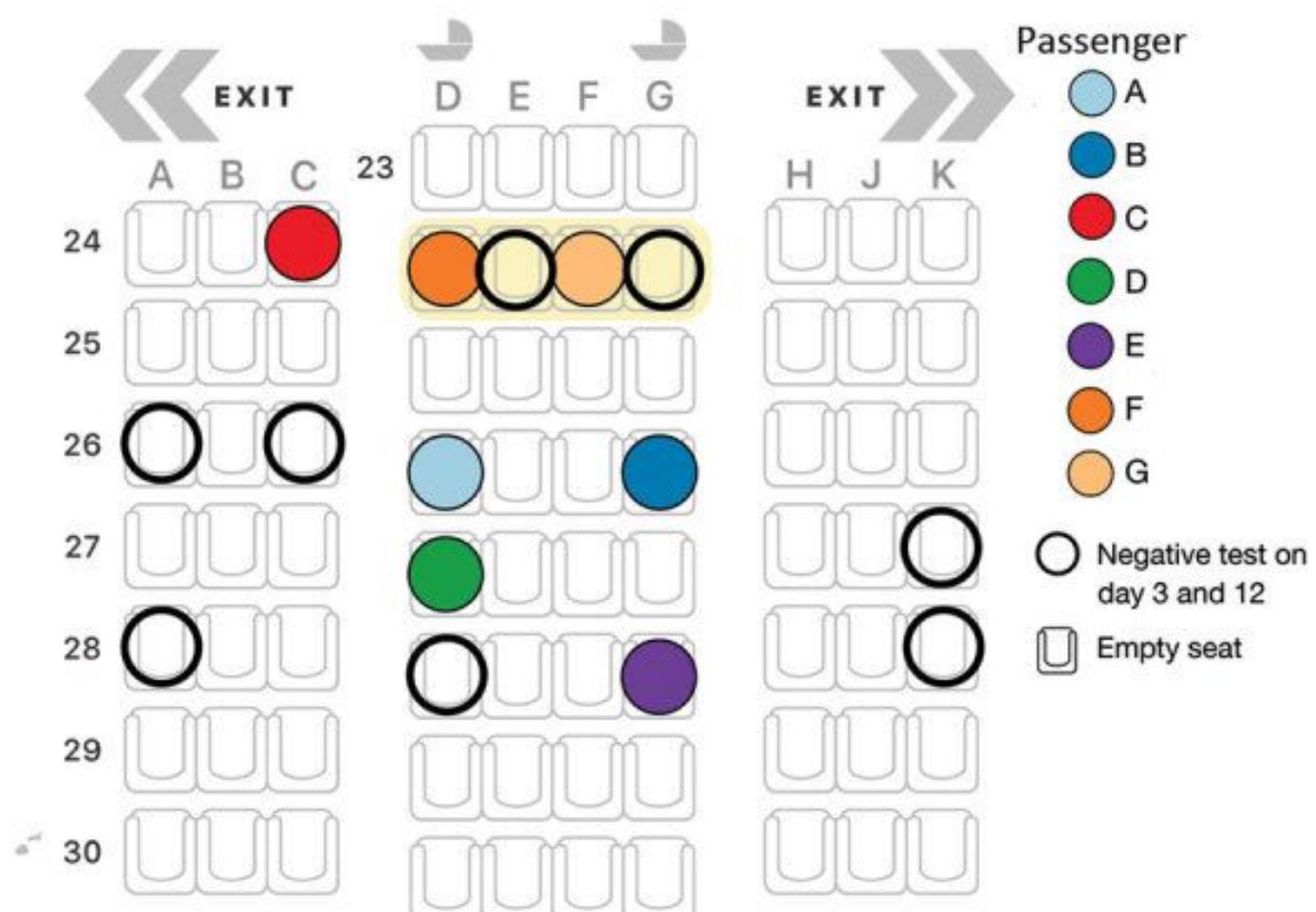


COVID-19
GENOMICS
GLOBAL TRAINING

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.



wellcome
connecting
science



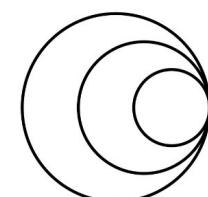
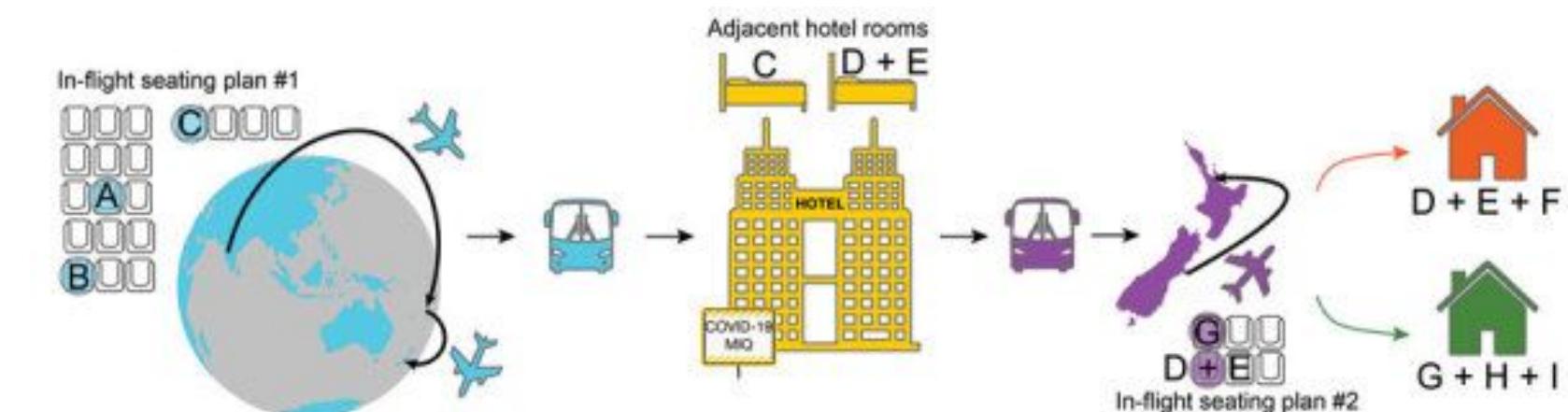
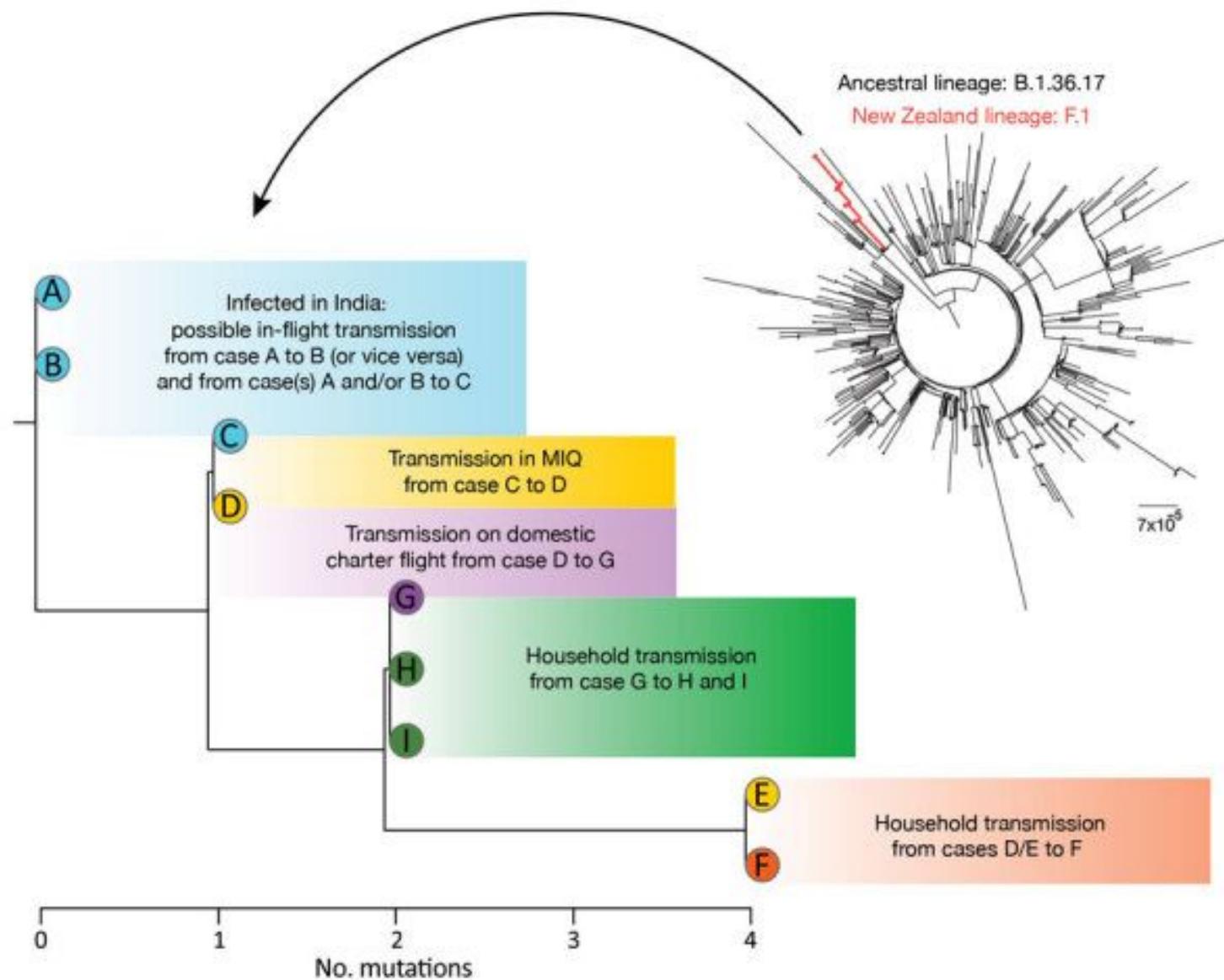
COVID-19
GENOMICS
GLOBAL TRAINING

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

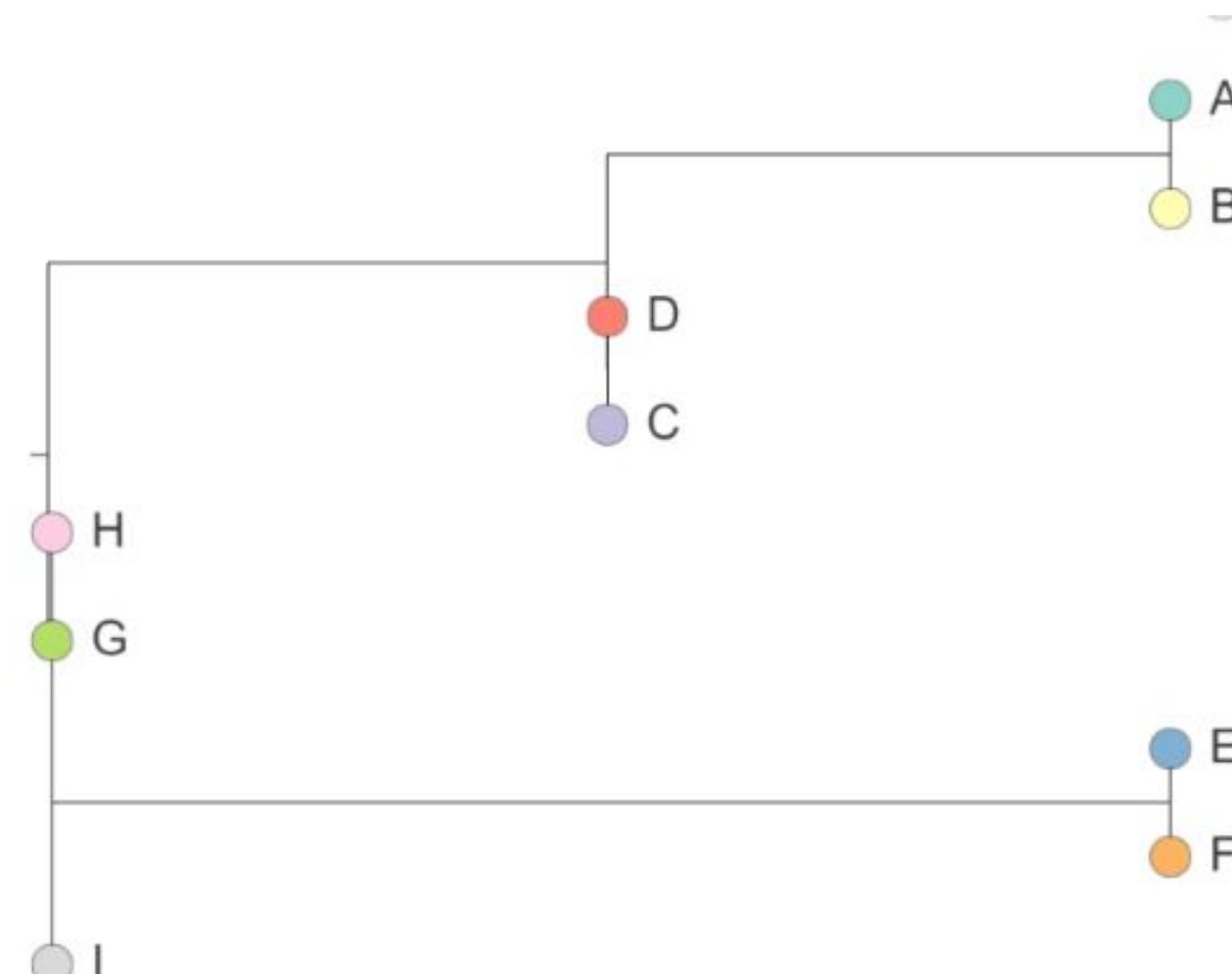


COVID-19
GENOMICS
GLOBAL TRAINING

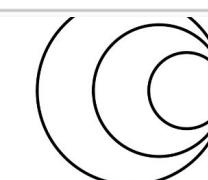
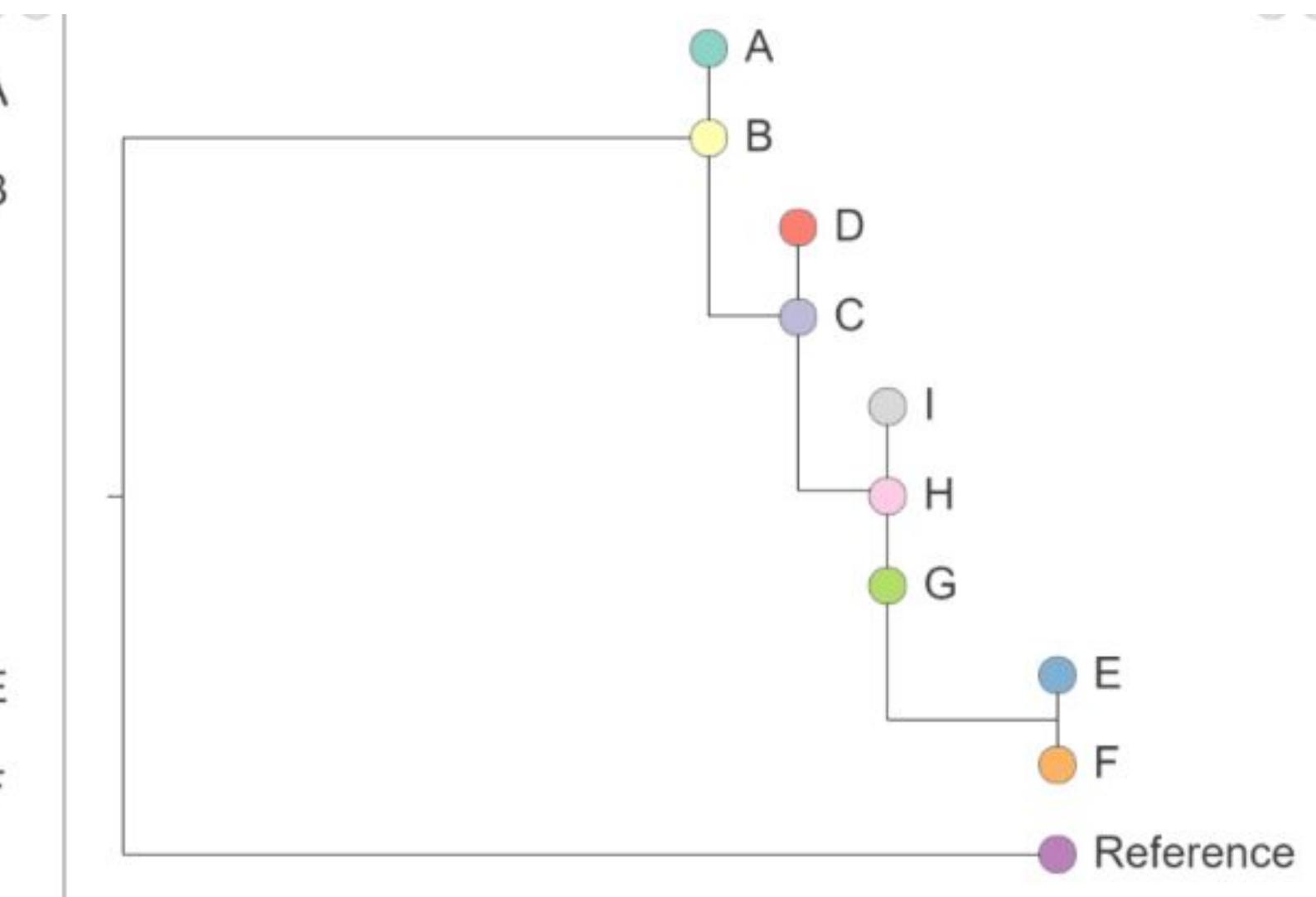
Example 2: following an incursion

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



wellcome
connecting
science



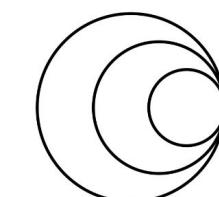
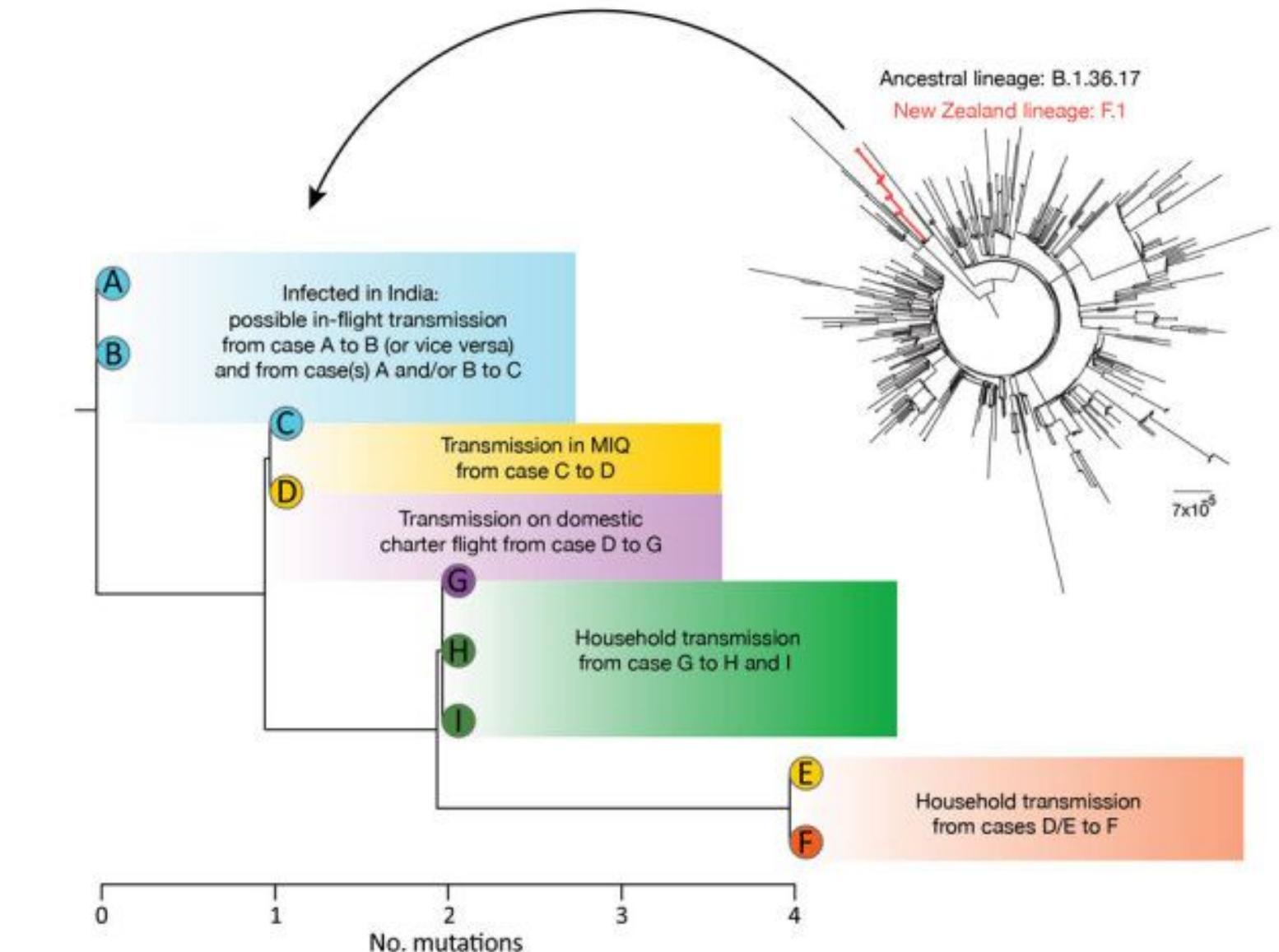
COVID-19
GENOMICS
GLOBAL TRAINING

Example 2: following an incursion

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

Here is an interactive visualisation you can follow:

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



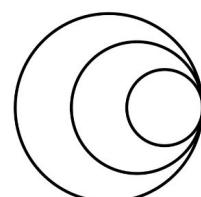
wellcome
connecting
science



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



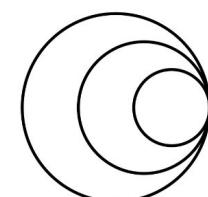
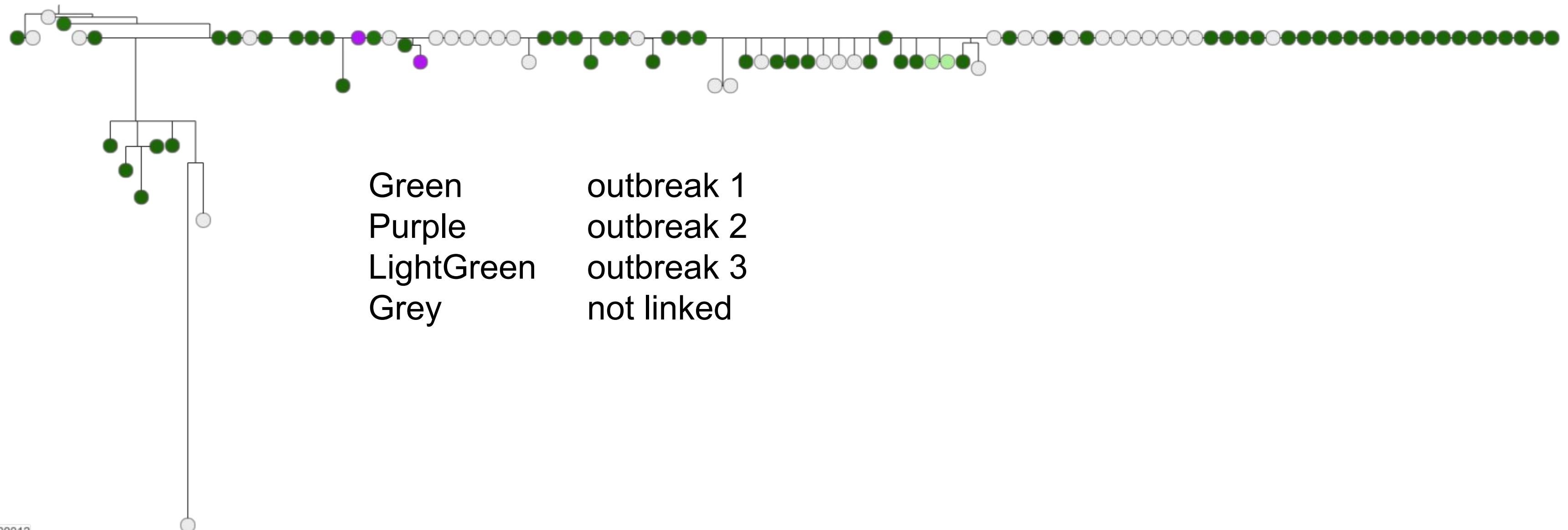
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Considerations:

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



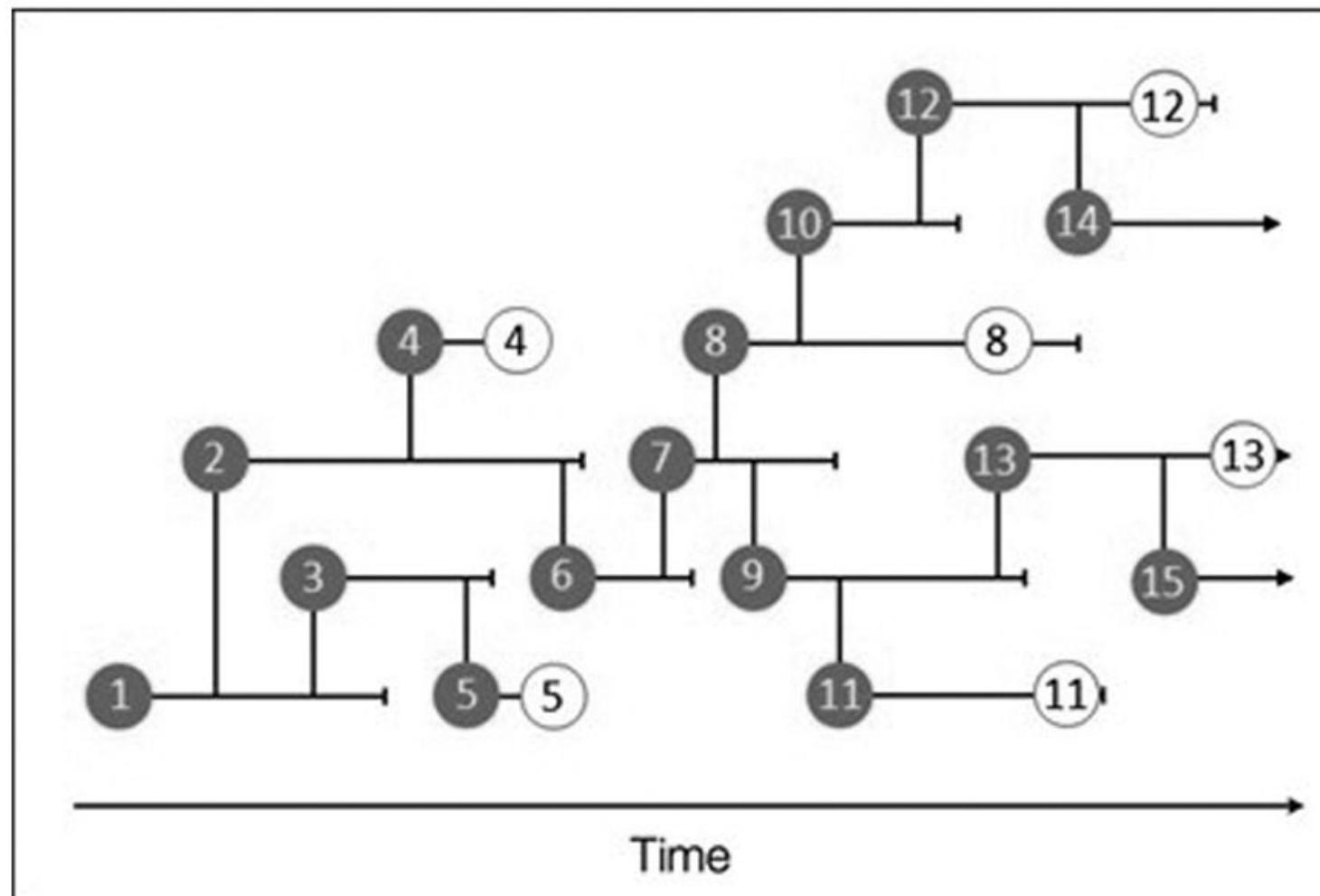
wellcome
connecting
science



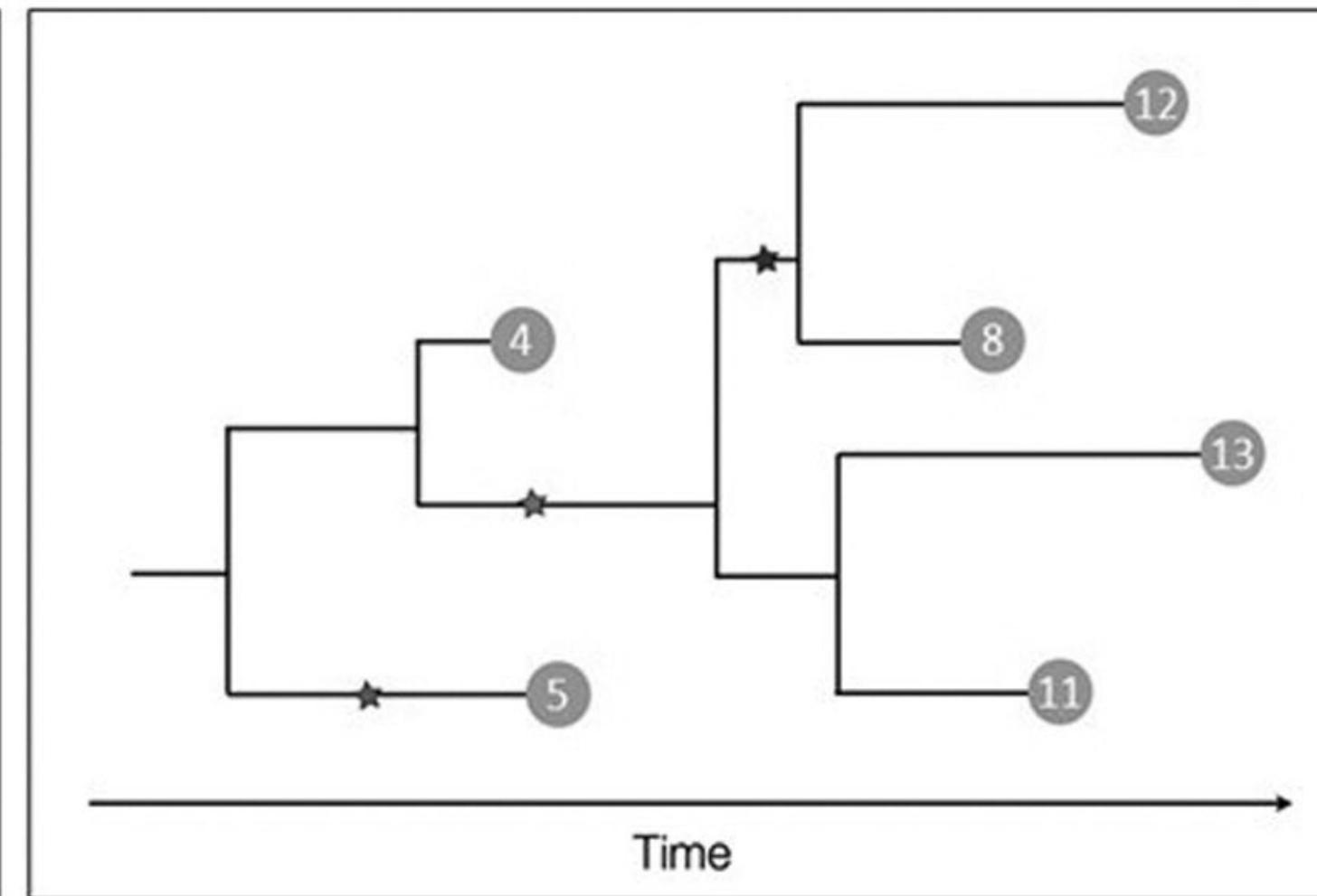
COVID-19
GENOMICS
GLOBAL TRAINING

Considerations:

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



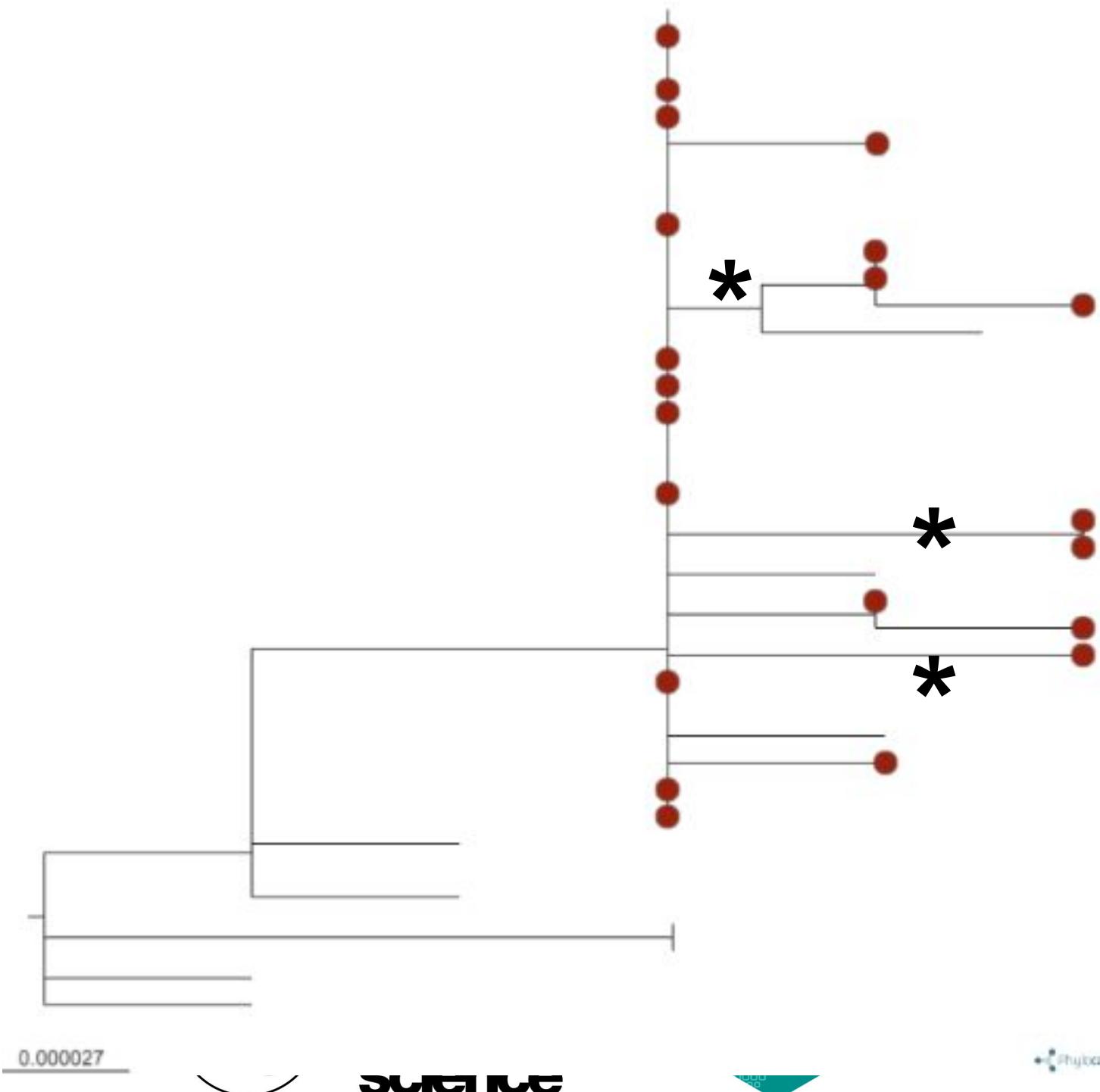
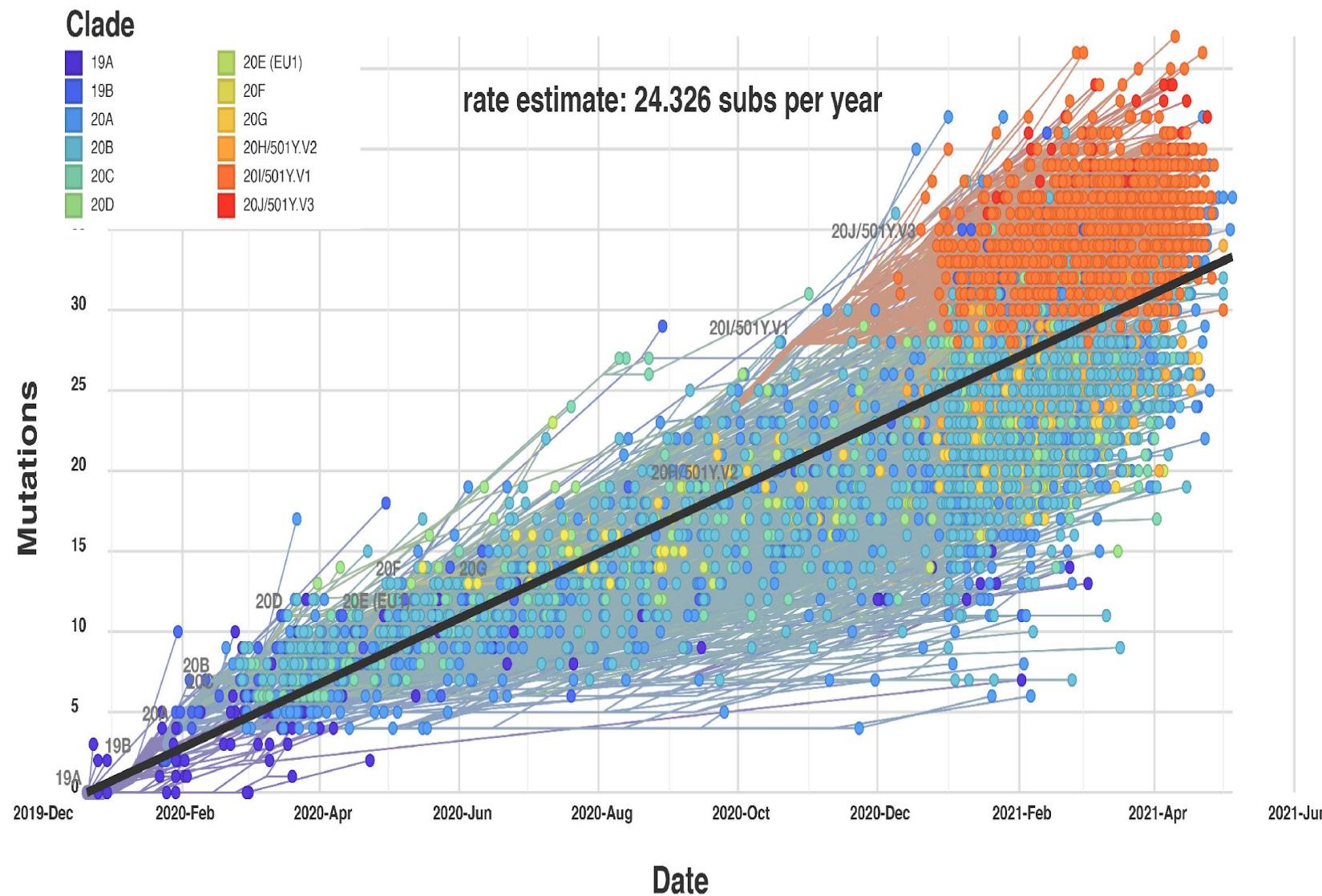
Actual transmission tree
15 infections
6 samples



Phylogenetic tree
Based on 6 sequences

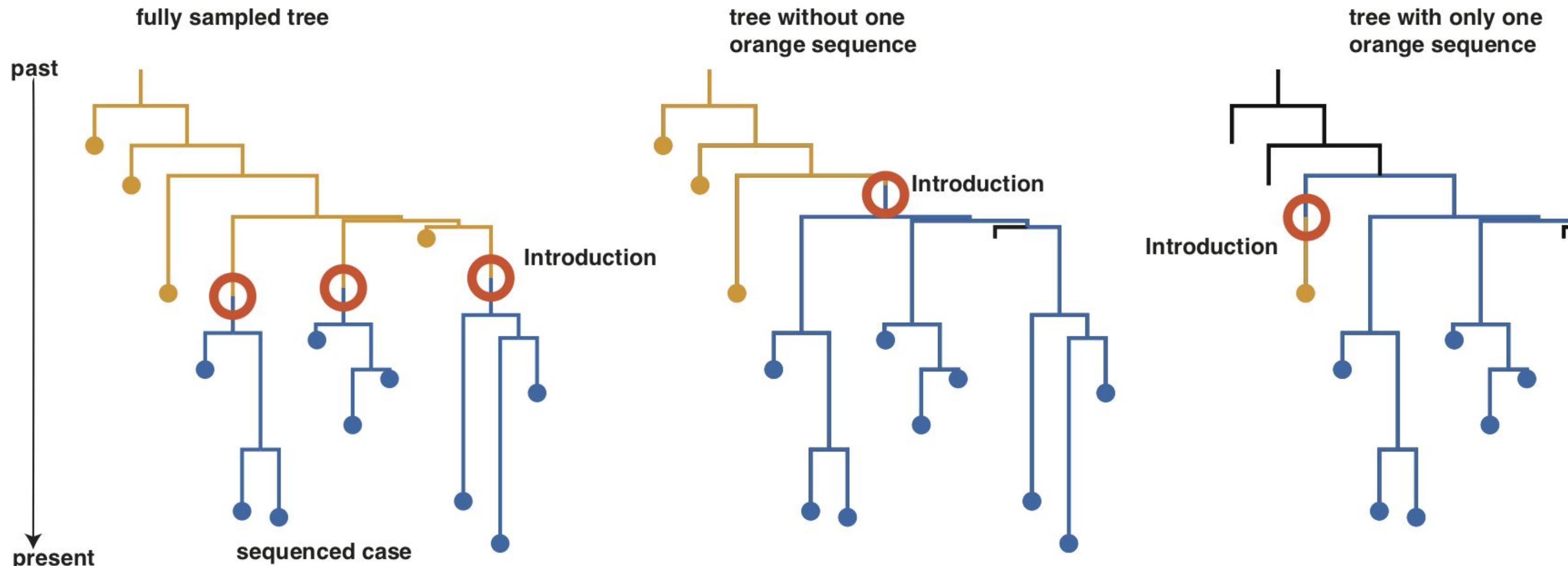
Signature of missing events

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

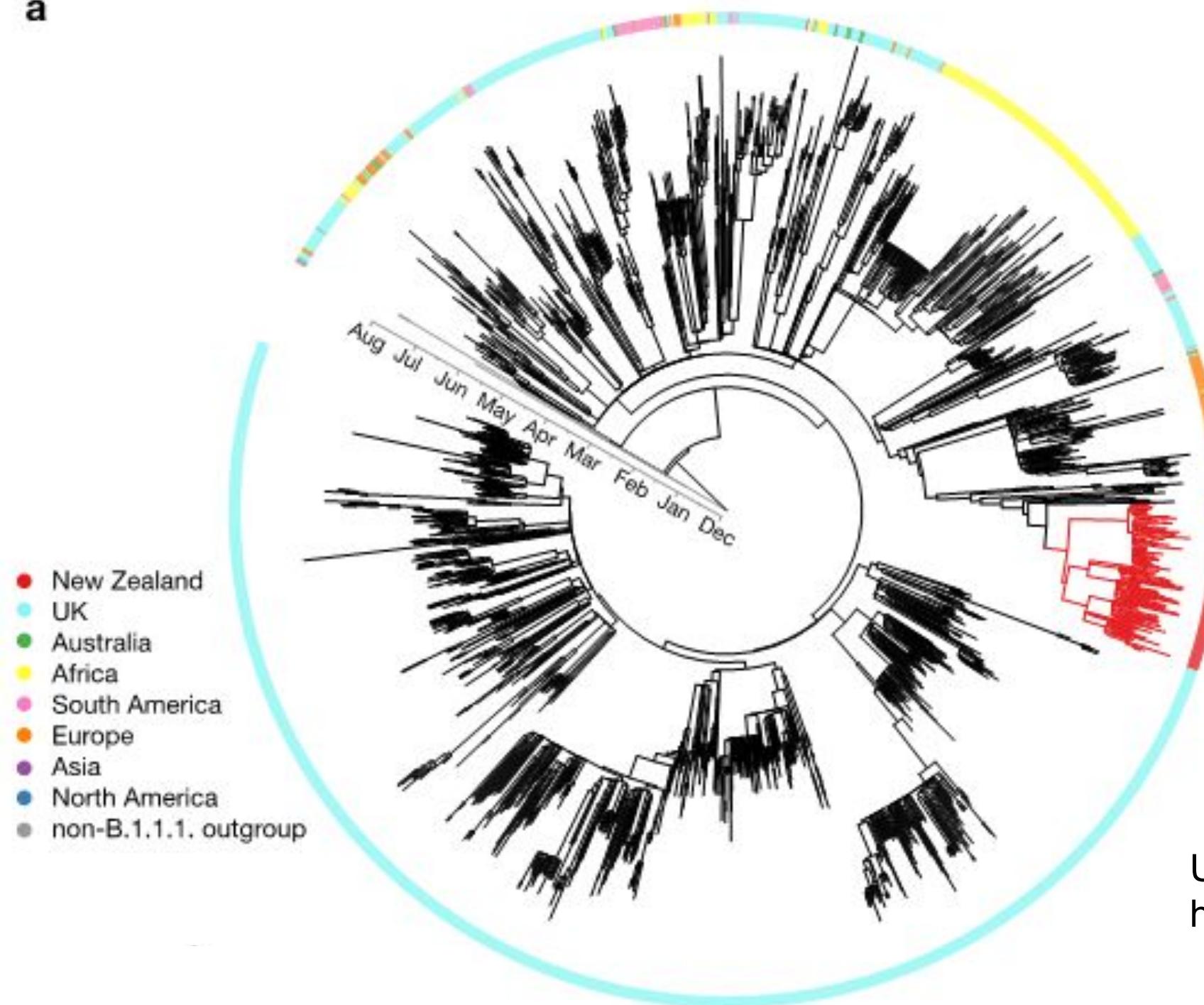


Considerations:

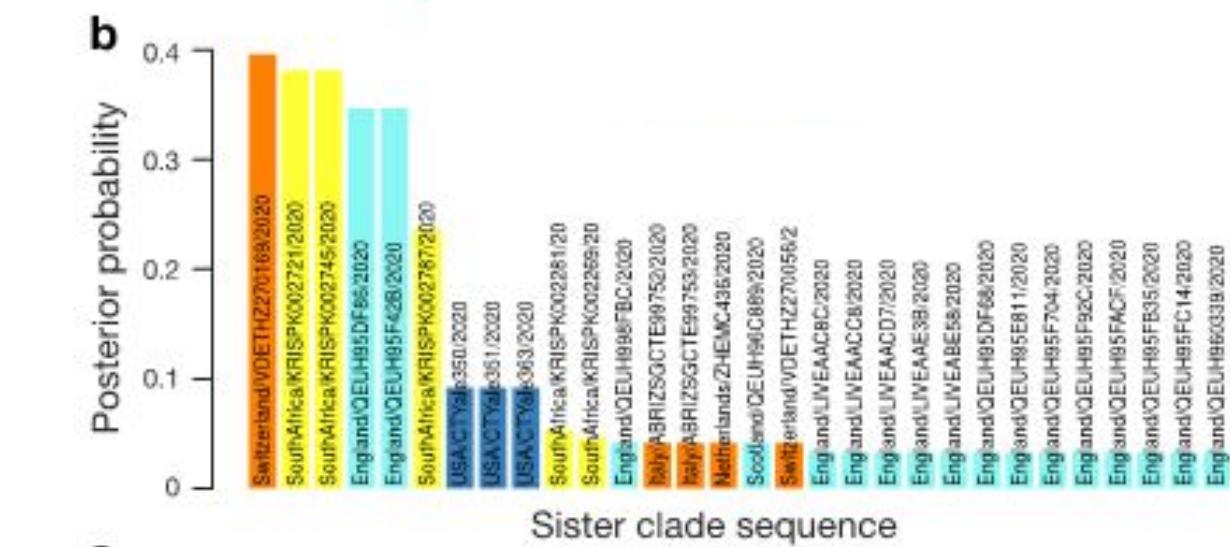
Poor sampling means cautions is needed for interpreting geographical origin

Biases in international data

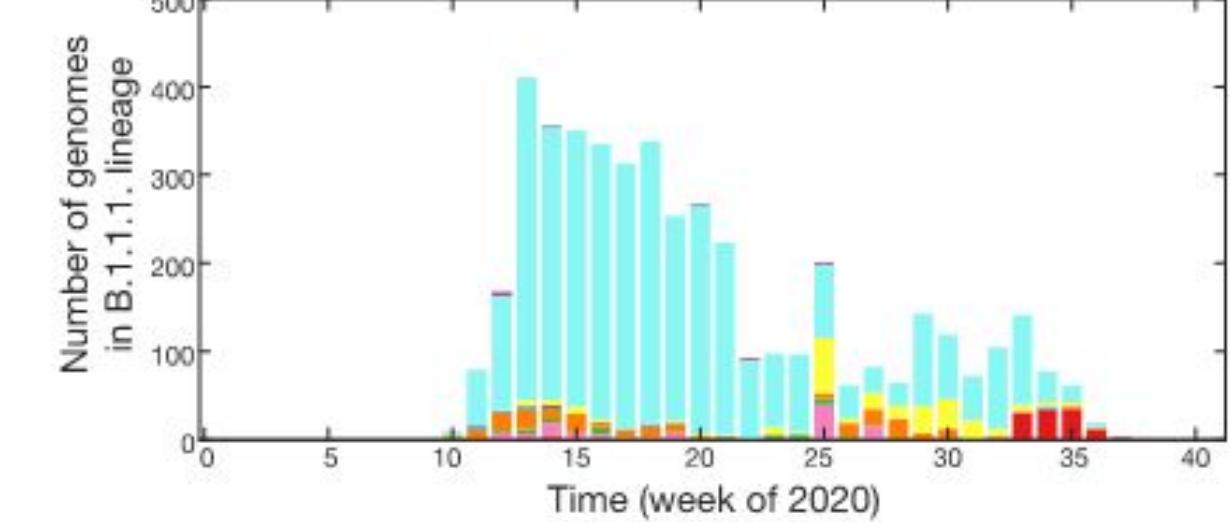
a



b



c



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



GLOBAL TRAINING

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

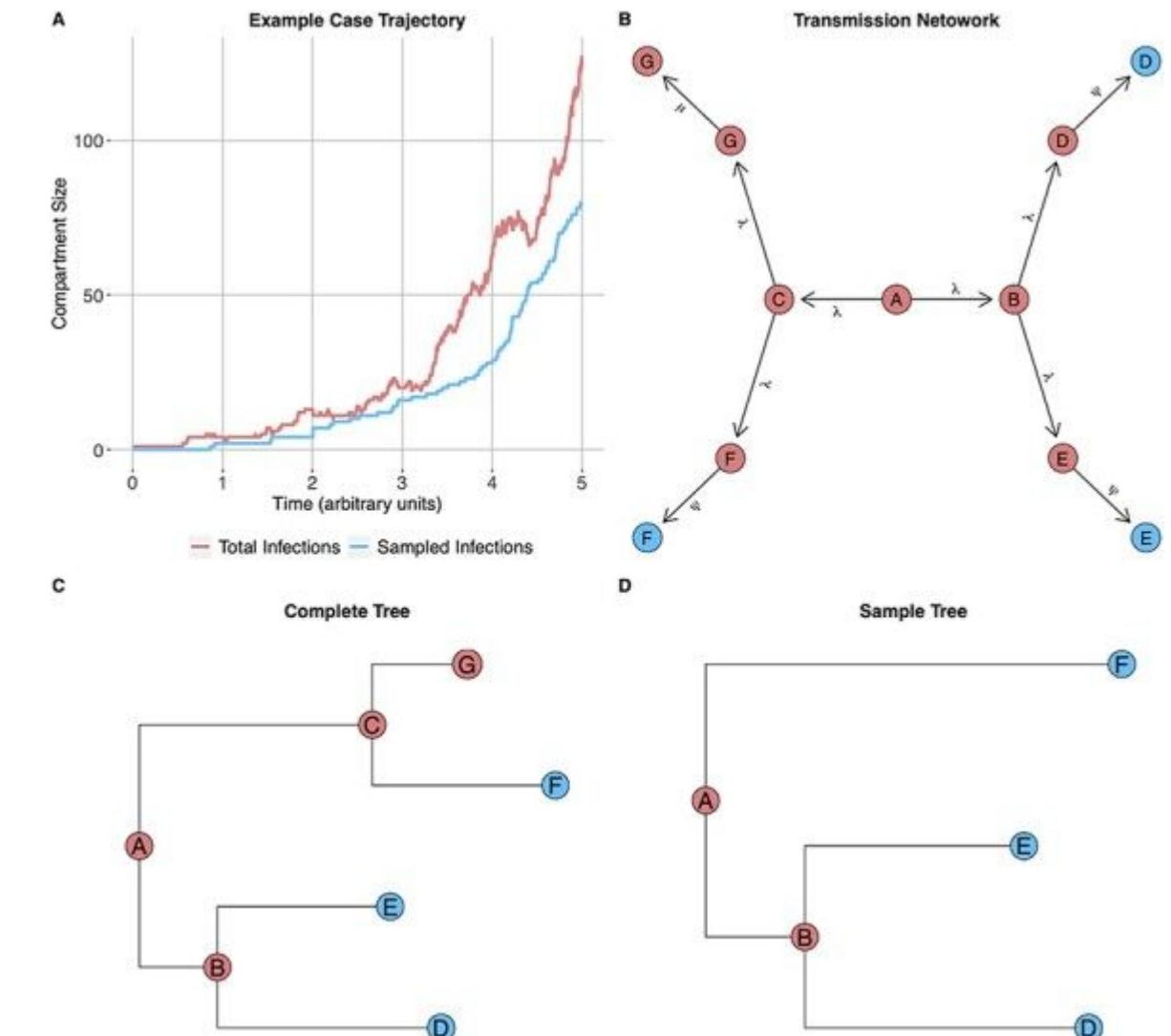
Estimate missing cases and population changes

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/>



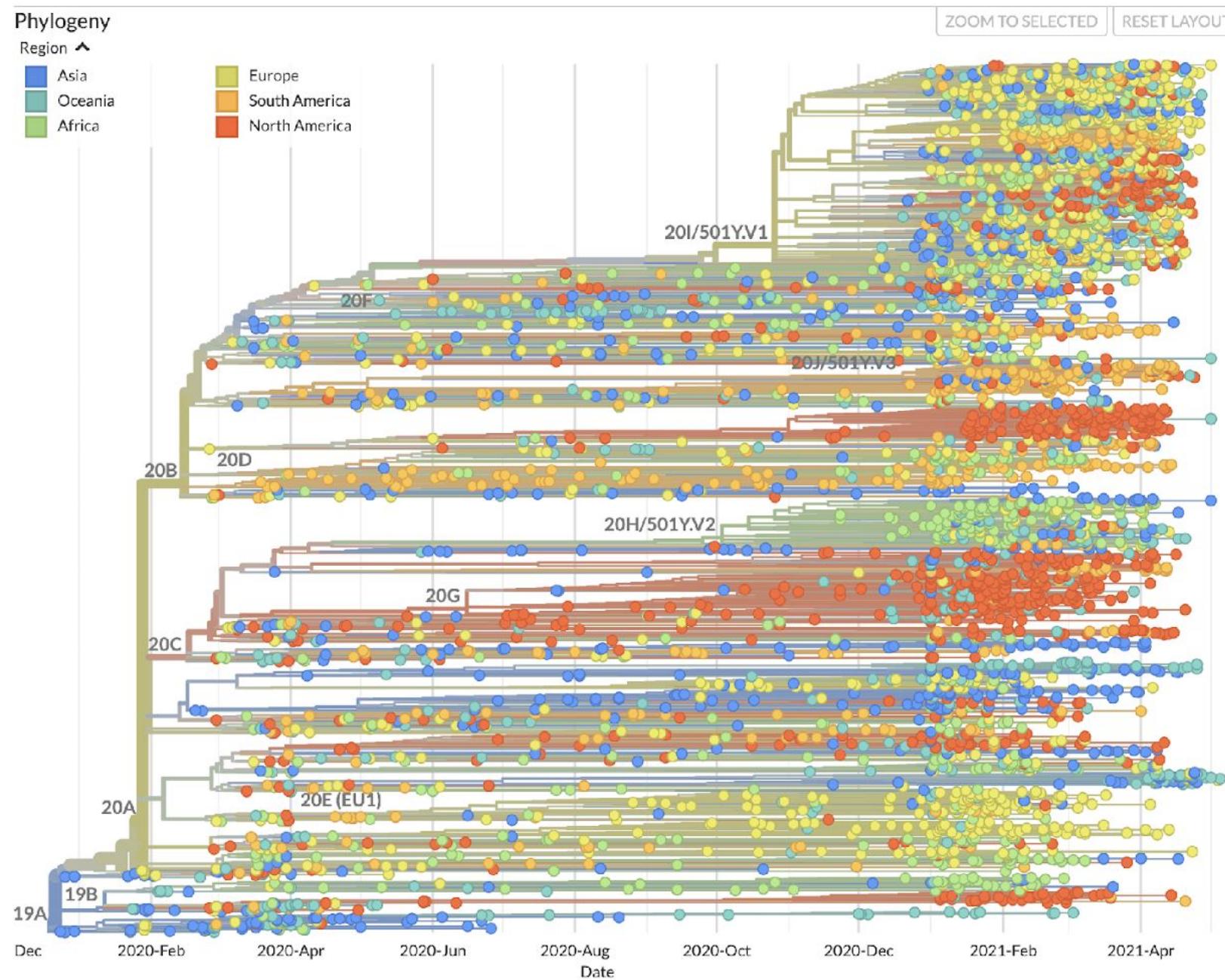
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9241095/>

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/>

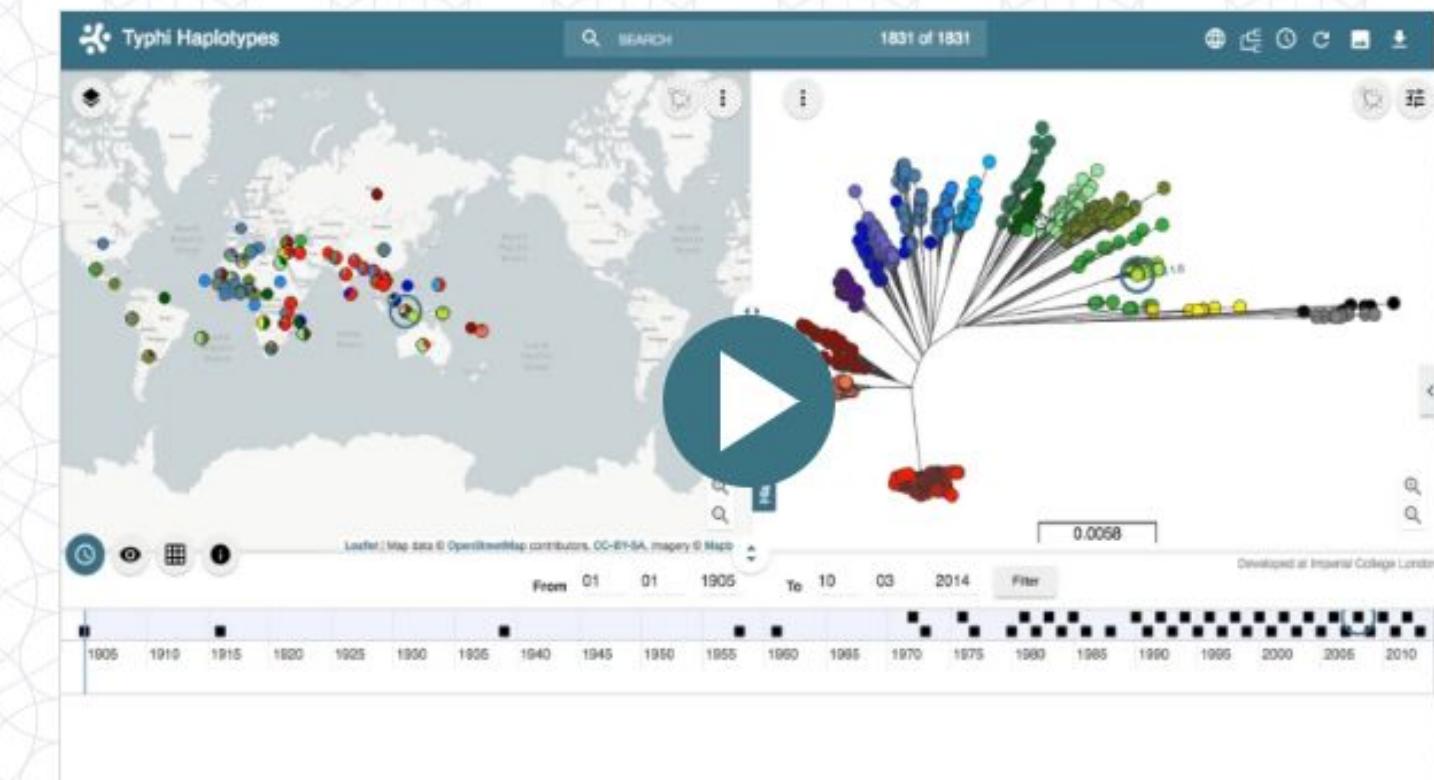
Interactive tree

Annotation

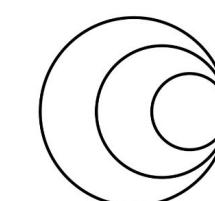
Network diagrams

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

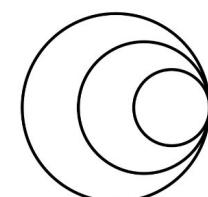
Nextstrain

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.

[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

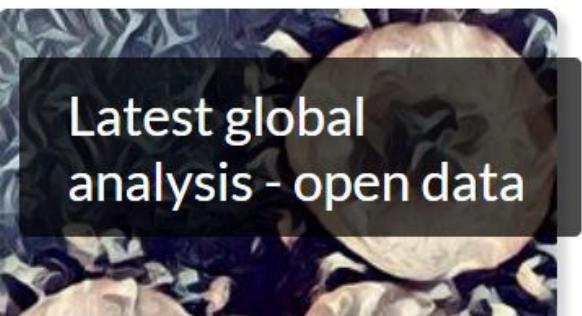
Data sharing is crucial for genomic surveillance and epidemiology

Data sharing enables comparisons between cases

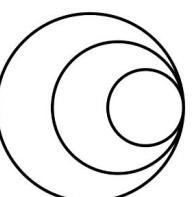
Data sharing informs and drives public health changes

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:

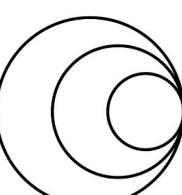
Reproducibility

Adherence to FAIR principles

Collaboration

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

Advancing scientific discovery - e.g. vaccine development



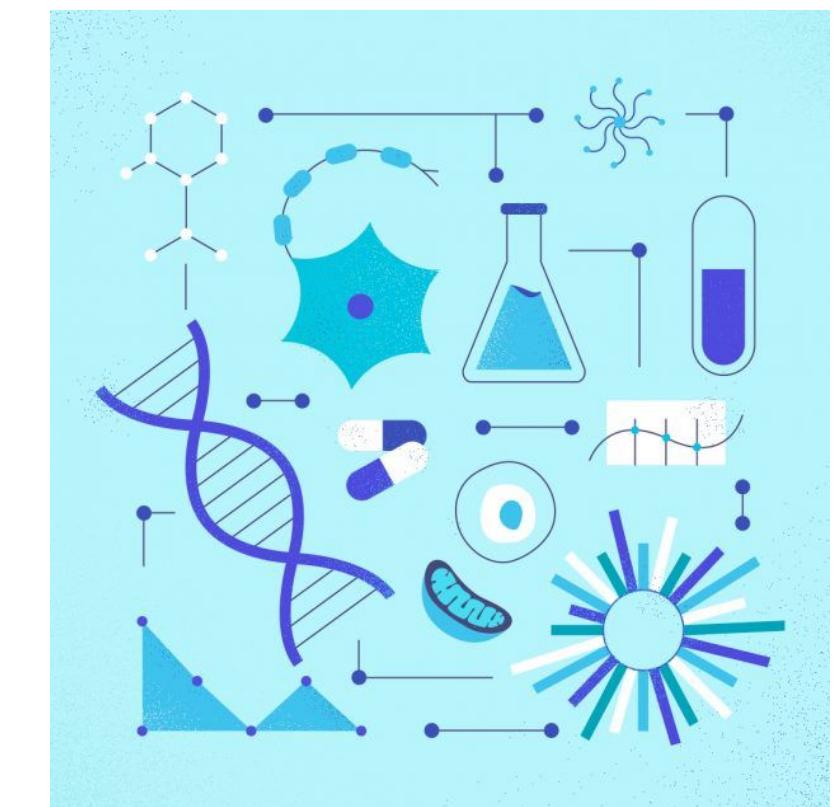
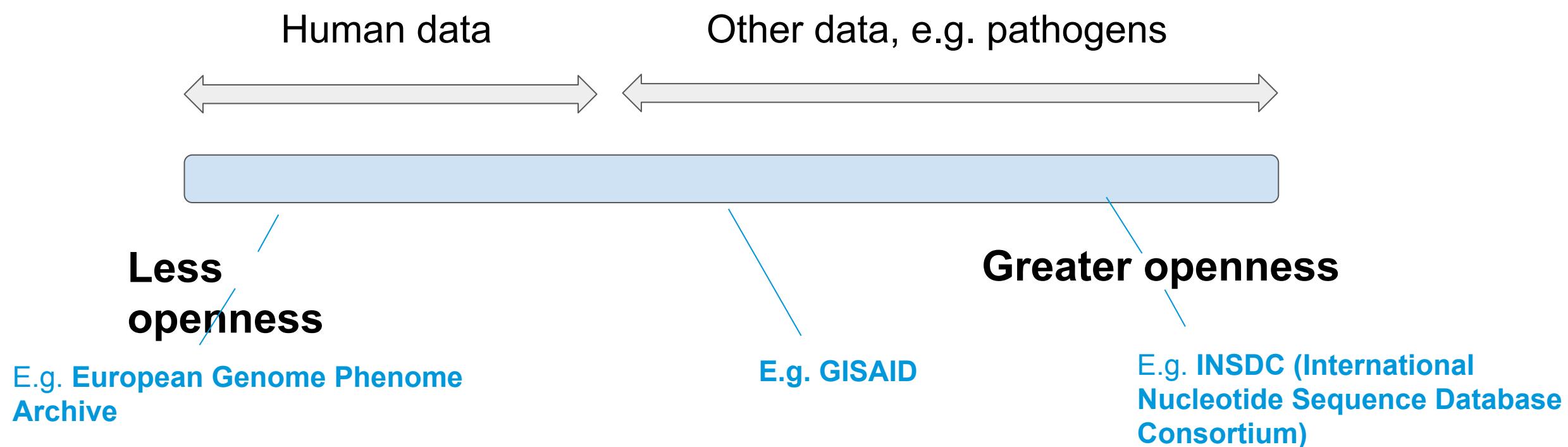
**wellcome
connecting
science**



**COVID-19
GENOMICS
GLOBAL TRAINING**

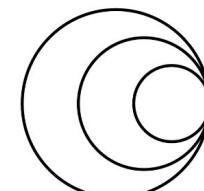
Genomic databases

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

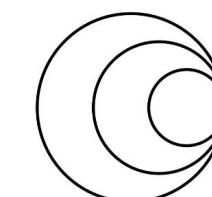
Genomic data repositories

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

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 CISAID	Browse data and metadata standards endorsed by the Genome Standards Consortium
Genetic variation data	dbSNP (human variations less than 50bp) dbVar (human variations greater than 50bp) ClinVar (human genotype & phenotype) European Variation Archive (EVA) (all species)	

SARS-CoV-2 data also shared here

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



wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

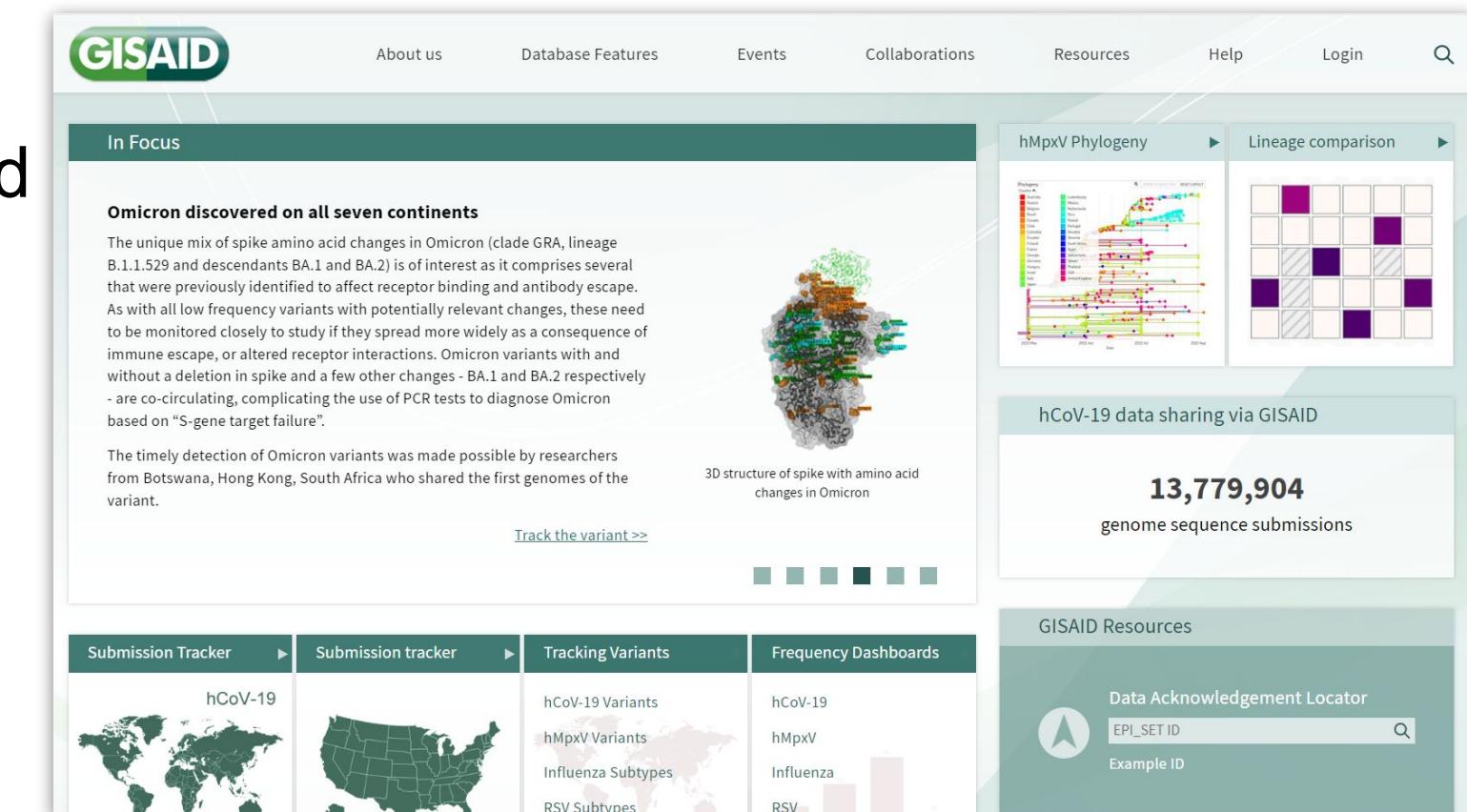
Sharing SARS-CoV-2 data - GISAID

Global Initiative on Sharing Avian Influenza Data

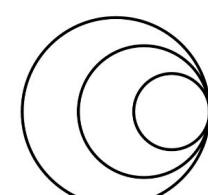
Established in 2008, first SARS-CoV-2 sequence shared in 2020. Now a popular SARS-CoV-2 data sharing platform

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

Assembled sequence submissions only



<https://gisaid.org/>



wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Sharing SARS-CoV-2 data - ENA



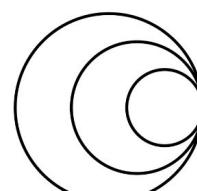
European Nucleotide Archive (European arm of INSDC) -
data is mirrored between all 3 nodes

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

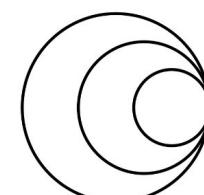
What open access is *not*:

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

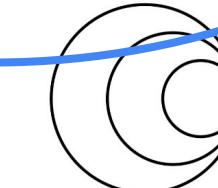
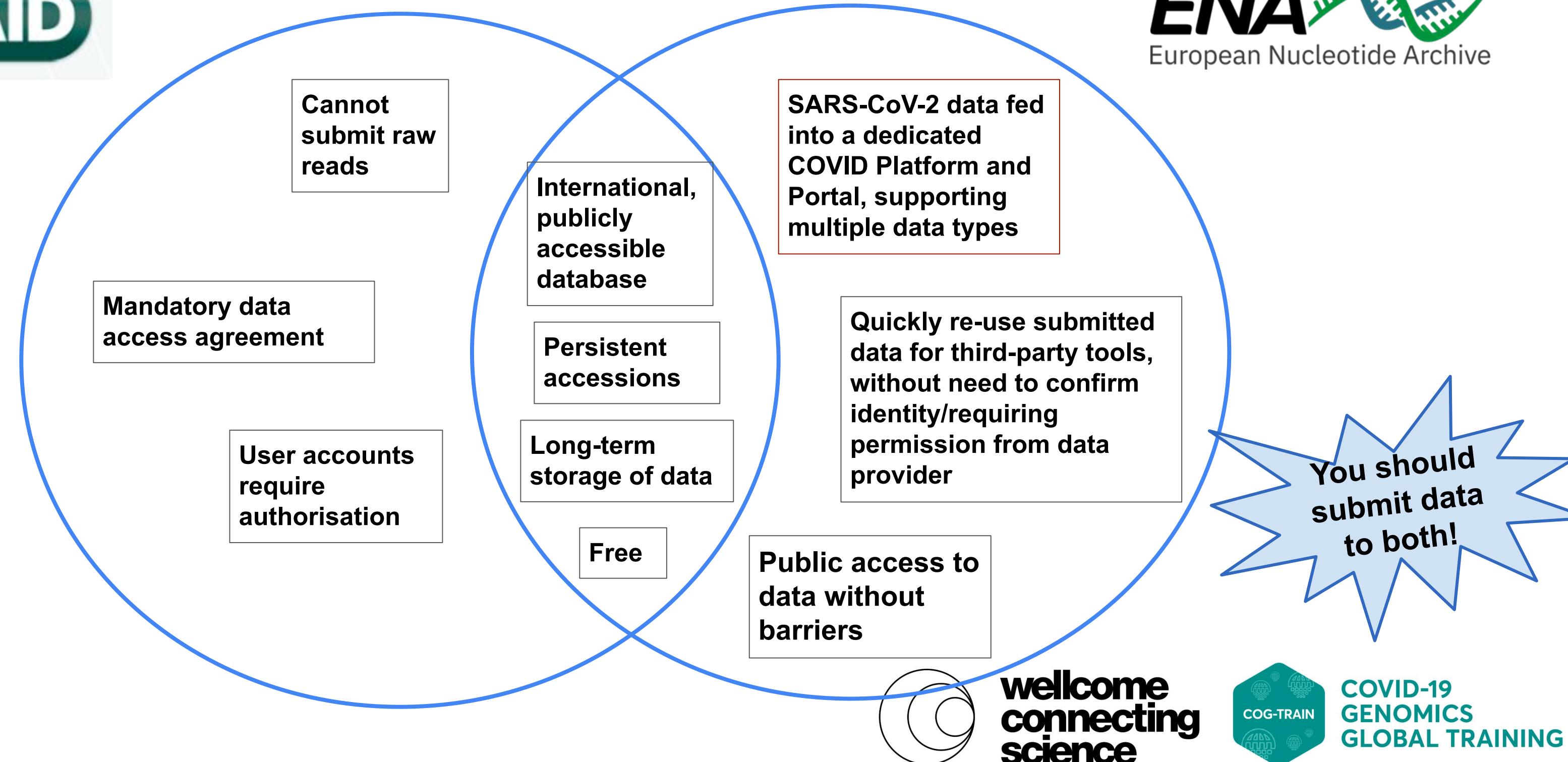


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

GISAID & ENA databases - a comparison



**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

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')

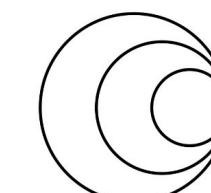
- **GISAID to ENA xml/xls sample converter**

- Can use custom GISAID<->ENA field mapping file if desired
- ‘GISAID Accession ID’ user-defined ENA attribute included by default

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

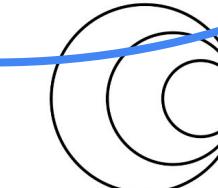
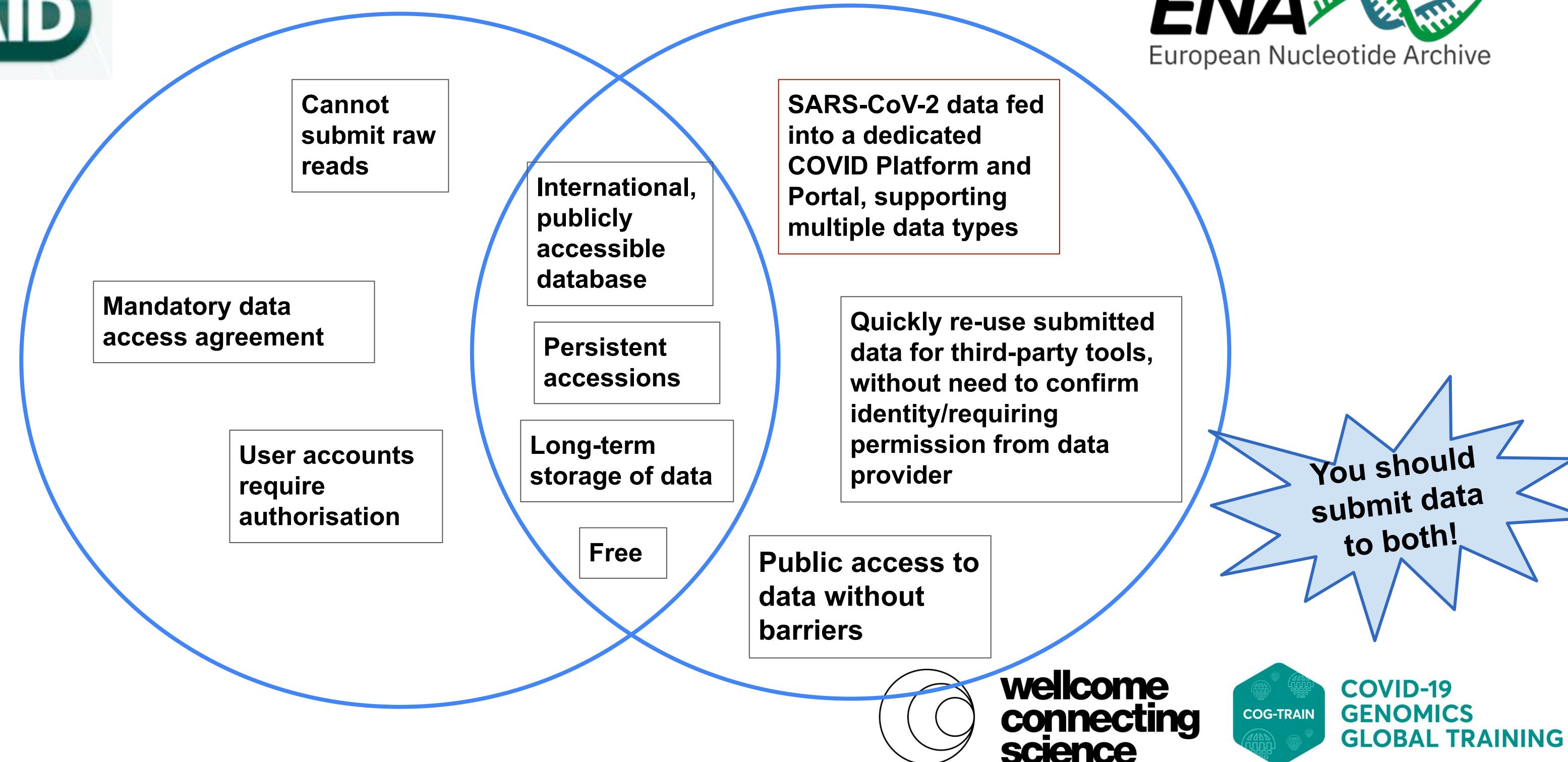


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

GISAID & ENA databases - a comparison

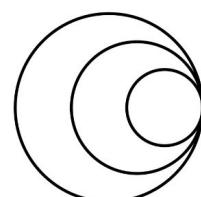


**wellcome
connecting
science**



**COVID-19
GENOMICS
GLOBAL TRAINING**

Section 4: European COVID-19 Data Platform

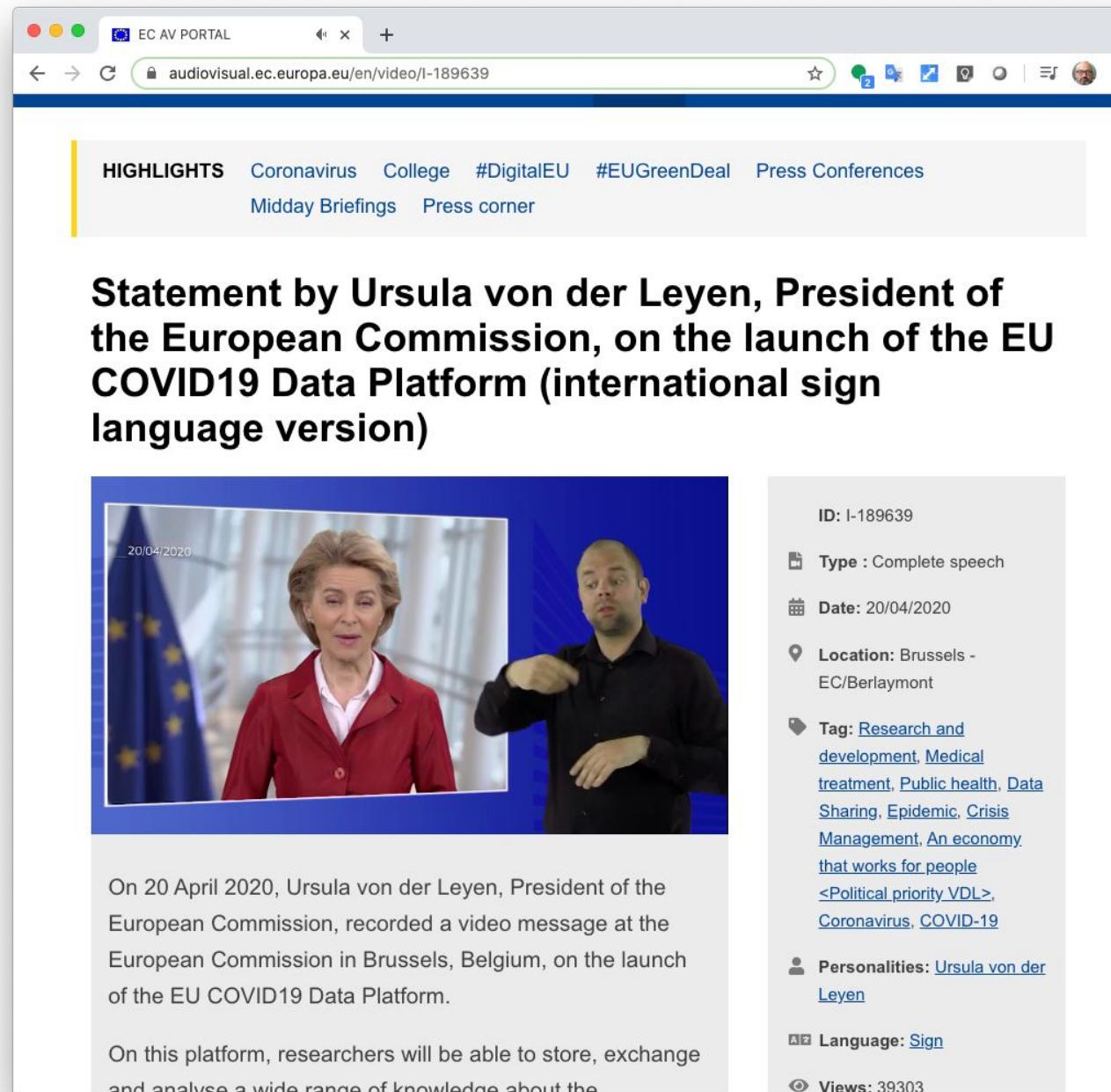


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

The European COVID-19 Data Platform



A screenshot of a web browser displaying a video player. The video shows Ursula von der Leyen, President of the European Commission, speaking at a podium with the European Union flag in the background. A sign language interpreter is visible to her right. The video player interface includes a timestamp (20/04/2020), a play button, and a progress bar.

HIGHLIGHTS Coronavirus College #DigitalEU #EUGreenDeal Press Conferences
Midday Briefings Press corner

Statement by Ursula von der Leyen, President of the European Commission, on the launch of the EU COVID19 Data Platform (international sign language version)

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

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
Personality: Ursula von der Leyen
Language: Sign
Views: 39303

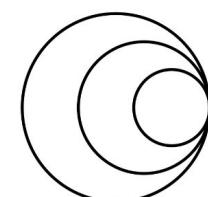
Launched Apr 2020

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



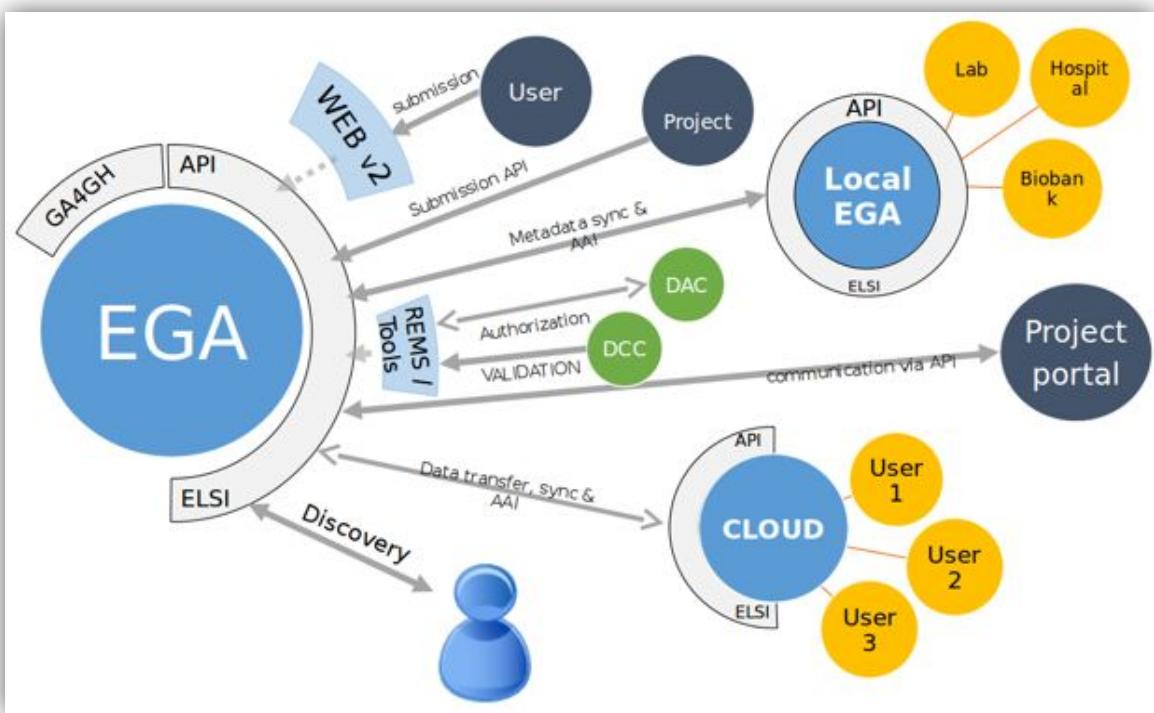
COVID-19
GENOMICS
GLOBAL TRAINING

Components

1. SARS-CoV-2 Data Hubs



2. Federated European Genome-phenome Archive



Support for sensitive human data

Restricted/controlled data access

Federated data model

Workspace enabling controlled access sharing of pre-publication sequence data

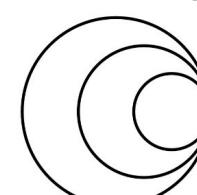
Tools for data analysis and visualisation

3. COVID-19 Data Portal

The screenshot shows the COVID-19 Data Portal homepage. At the top, there's a navigation bar with links for 'About', 'Tools', 'FAQ', 'Related Resources', 'Bulk Downloads', and 'Submit Data'. Below the header, a banner reads 'COVID-19 Data' and 'Accelerating research through data sharing'. It features a search bar and a 'Search' button. To the right is a large image of a COVID-19 virus particle. The main content area is divided into several sections: 'Viral sequences' (14,718,460 records), 'Host sequences' (30,694 records), 'Expression' (183 records), and 'Proteins' (3,633 records). Each section includes a brief description and a link to more details. A 'Latest news' section highlights the '13th VEO report on SARS-CoV-2 mutations and variations now published' (29 Sep 2022).

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

Entry point for data sharing and visualisation tools

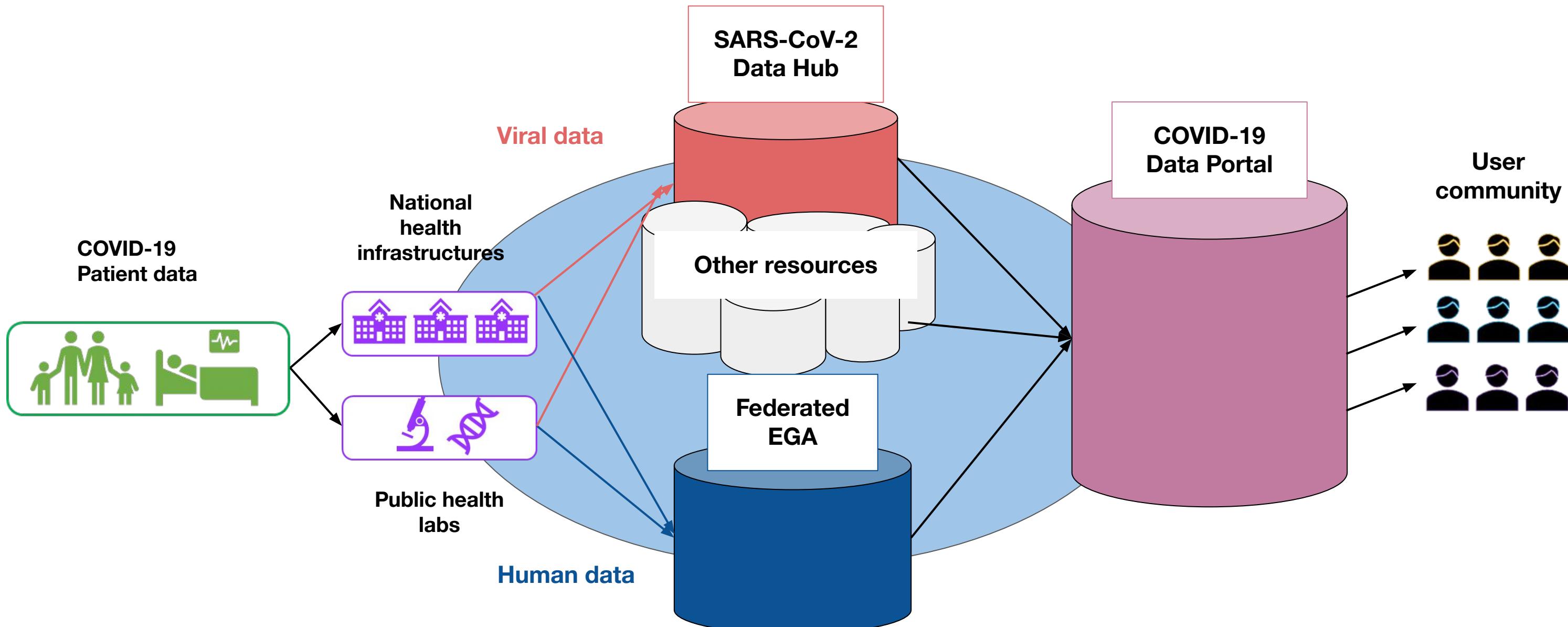


wellcome
connecting
science

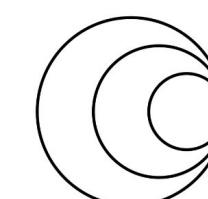


COVID-19 GENOMICS GLOBAL TRAINING

Data flow through the COVID-19 Data Platform



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

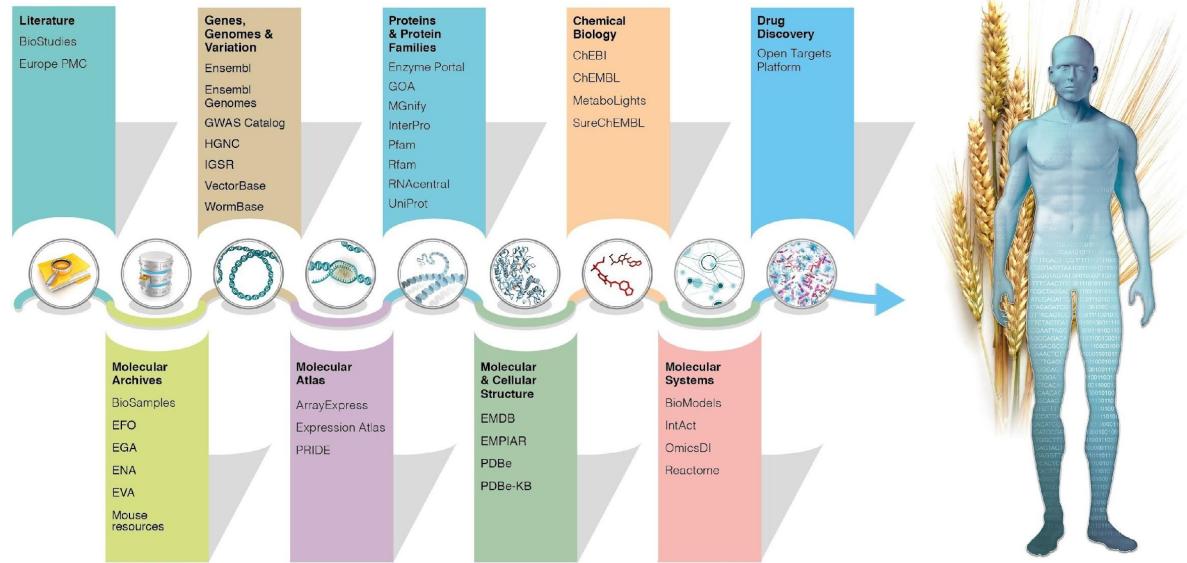


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Foundations



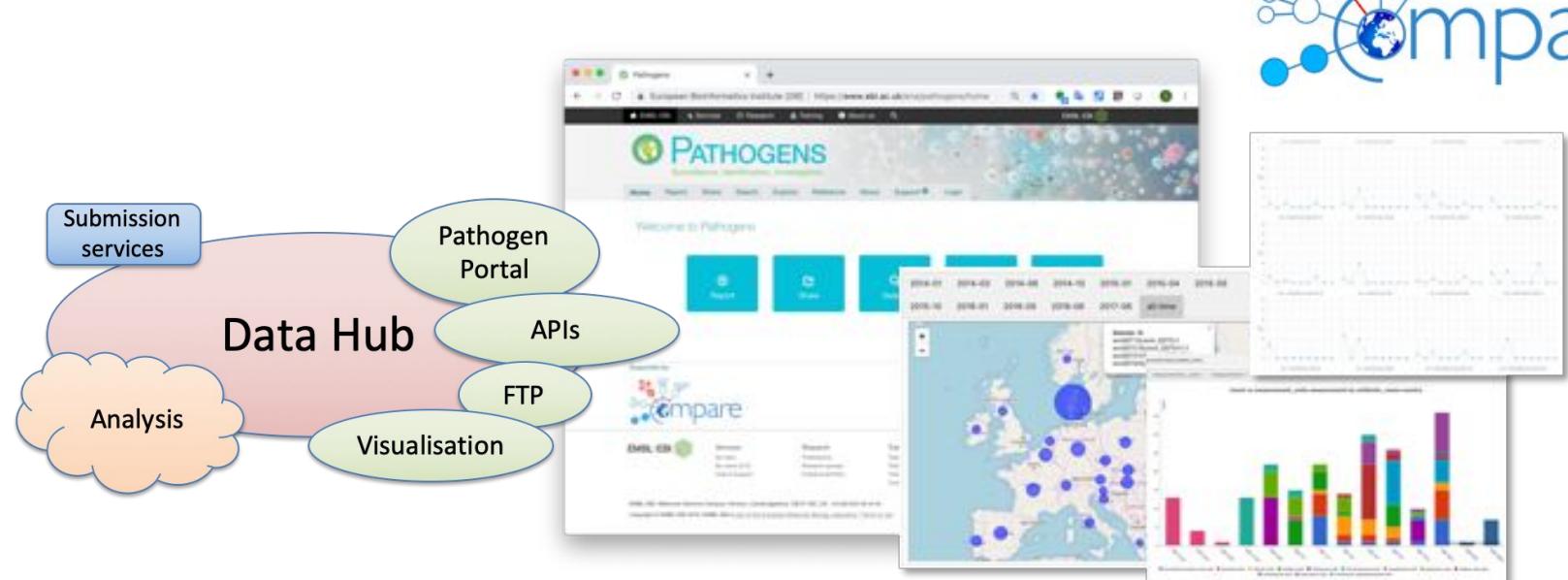
<https://elixir-europe.org/>

The EOSC Portal website features a banner for the COVID-19 Data Portal and sections for accessing services like networking, compute, storage, and sharing/discovery.

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



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



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 Medical Centre, the Netherlands



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



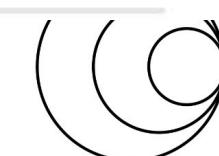
Technical University of Denmark (DTU)



Eötvös Loránd University, Hungary



Universitätsklinikum Heidelberg, Germany



COG-TRAIN
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's a navigation bar with links for About, Tools, FAQ, Related Resources, Bulk Downloads, and Submit Data. Below the navigation is a search bar with a placeholder 'Search' and a 'Search' button. To the right of the search bar is a 'More' dropdown menu. On the left, there's a sidebar with links for Viral Sequences, Host Sequences, Expression, Proteins, Networks, Cohorts, and another 'More' dropdown. The main content area features a large image of a brain. Overlaid on the image are four icons: a magnifying glass for 'Findable', a hand pointing for 'Accessible', three gears for 'Interoperable', and a recycling symbol for 'Reusable'. Below the image is a table titled 'COVID-19 Data' with the subtitle 'Accelerating research through data sharing'. The table has two columns: 'Type of Data' and 'Total'. The data is summarized as follows:

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/>



COVID-19
GENOMICS
GLOBAL TRAINING

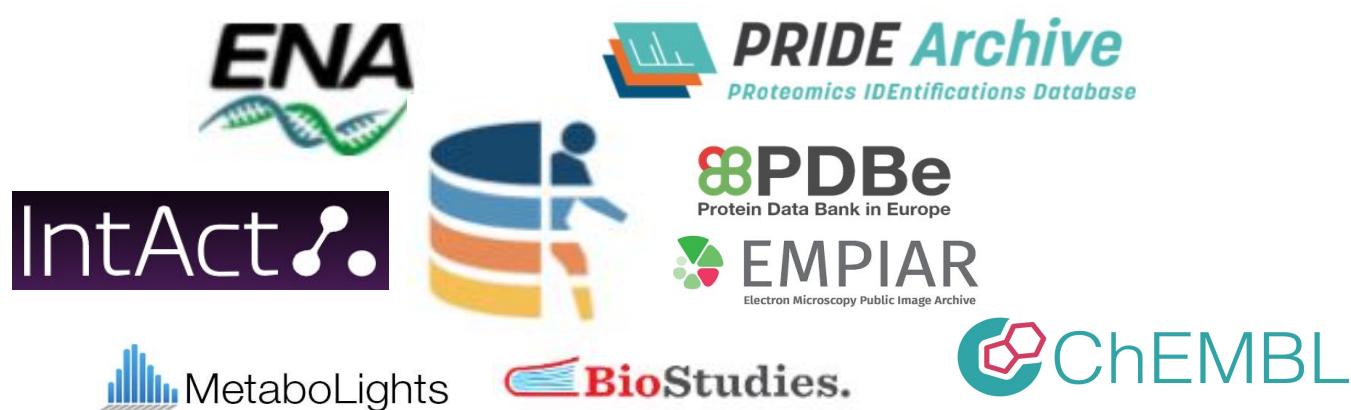
The COVID-19 Data Portal - data submission

Data is not submitted to the Portal itself...

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

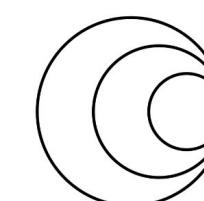
Spans 19 different datatypes

Different resources = different data
submission methods



The screenshot shows the COVID-19 Data Portal's submission interface. At the top, there are navigation links: About, Tools, FAQ, Related Resources, Bulk Downloads, and Submit Data. Below that is a menu bar with links: Viral Sequences, Host Sequences, Expression, Proteins, Networks, Cohorts, and More. The main content area has a header "Submit new data". Step 1 asks "What kind of data would you like to submit?" with options: Non-human data, Human data, Combination of human and non-human data, Multiple datatypes from the same species, and Non-biological data. Below this, there is a note about collaboration and data sharing, followed by a footer with links to 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, EOSC-Life, and the Horizon 2020 programme.

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



wellcome
connecting
science



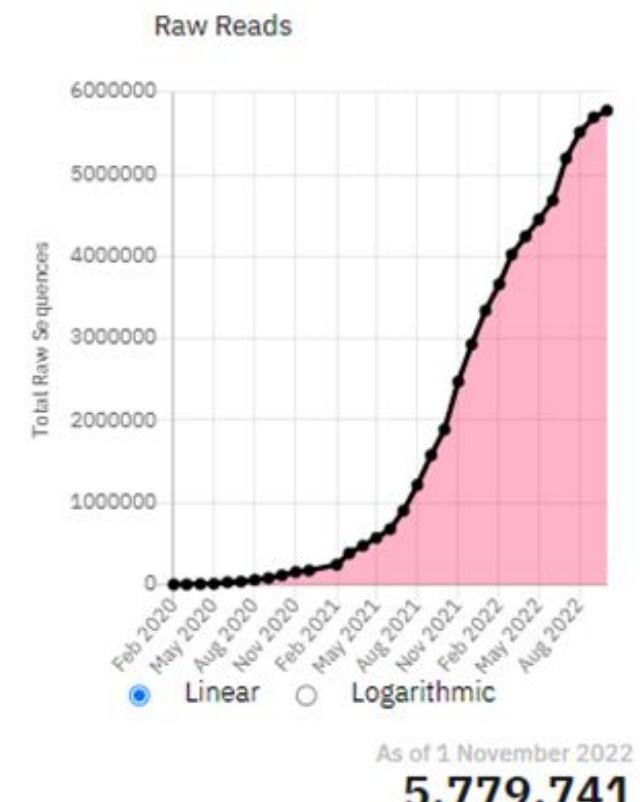
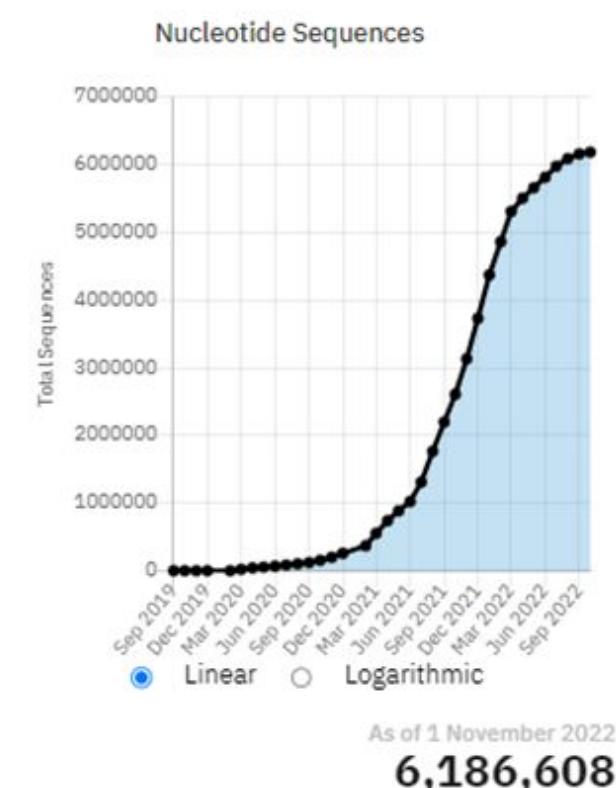
COVID-19
GENOMICS
GLOBAL TRAINING

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



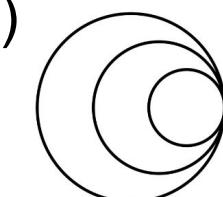
<https://www.covid19dataportal.org/statistics>

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

Raw reads submitted by Country



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



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



The screenshot shows the CABANA project website. The header features the CABANA logo with a globe icon, the tagline "Capacity building for bioinformatics in Latin America", and a navigation menu with links to Home, About, Workshops, Research secondments, Train the trainer, eLearning, Webinars, News, and Contact us. A "Change language" dropdown is also present. The main content area includes a large image of a virus and a toucan, a section titled "What is CABANA?", a video thumbnail, and four cards for Workshops, Research Secondments, Train the Trainer, and eLearning Resources.

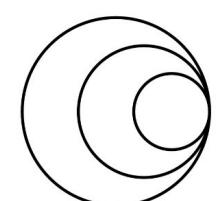
What is CABANA?

- 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

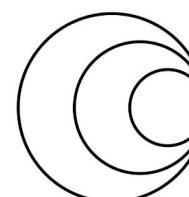


wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

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



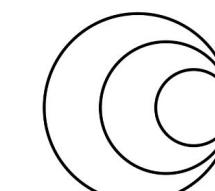
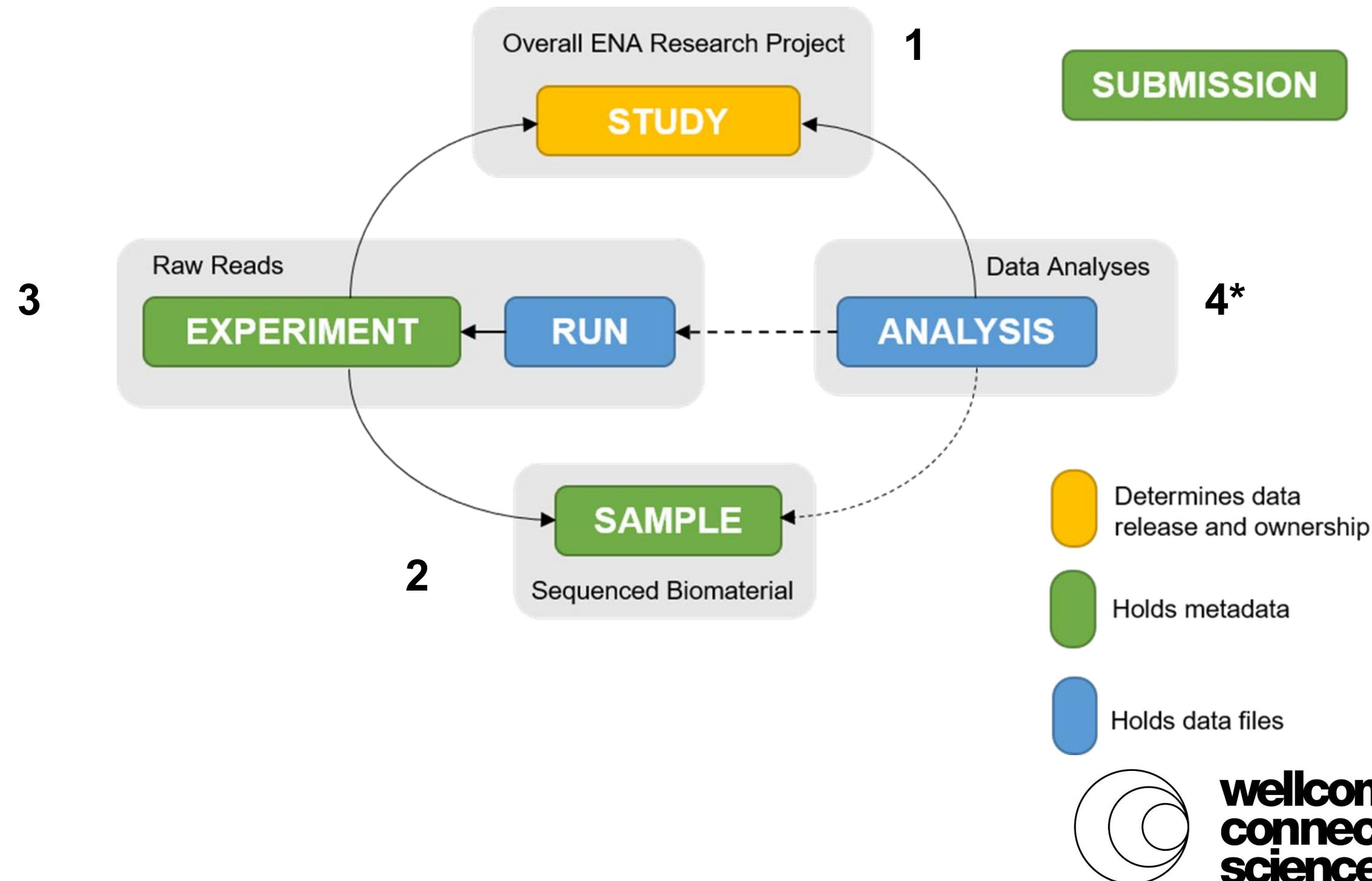
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Submitting SARS-CoV-2 data to the ENA

The ENA Metadata Model



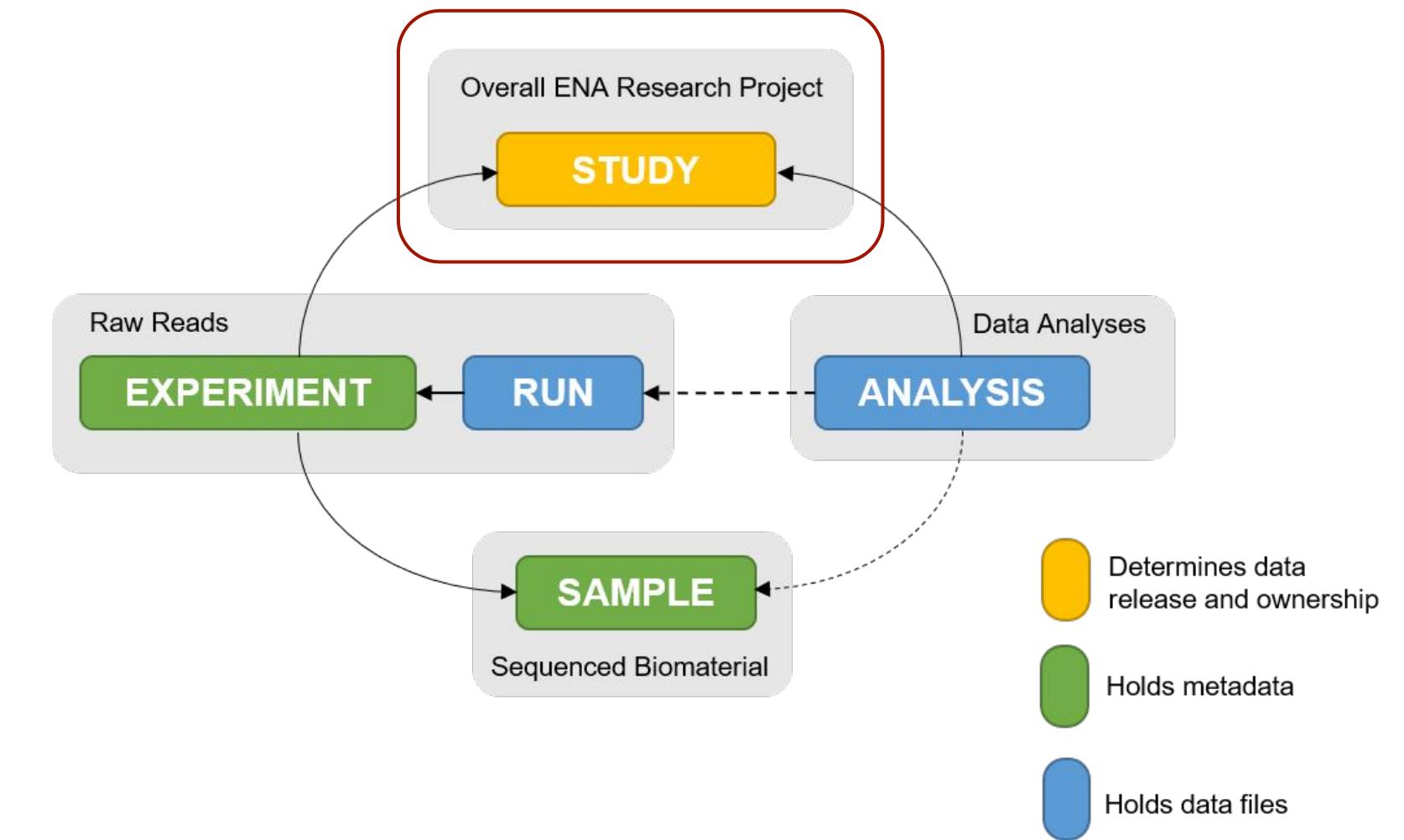
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

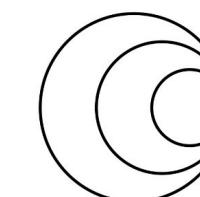
ENA Metadata Model: Study

- 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



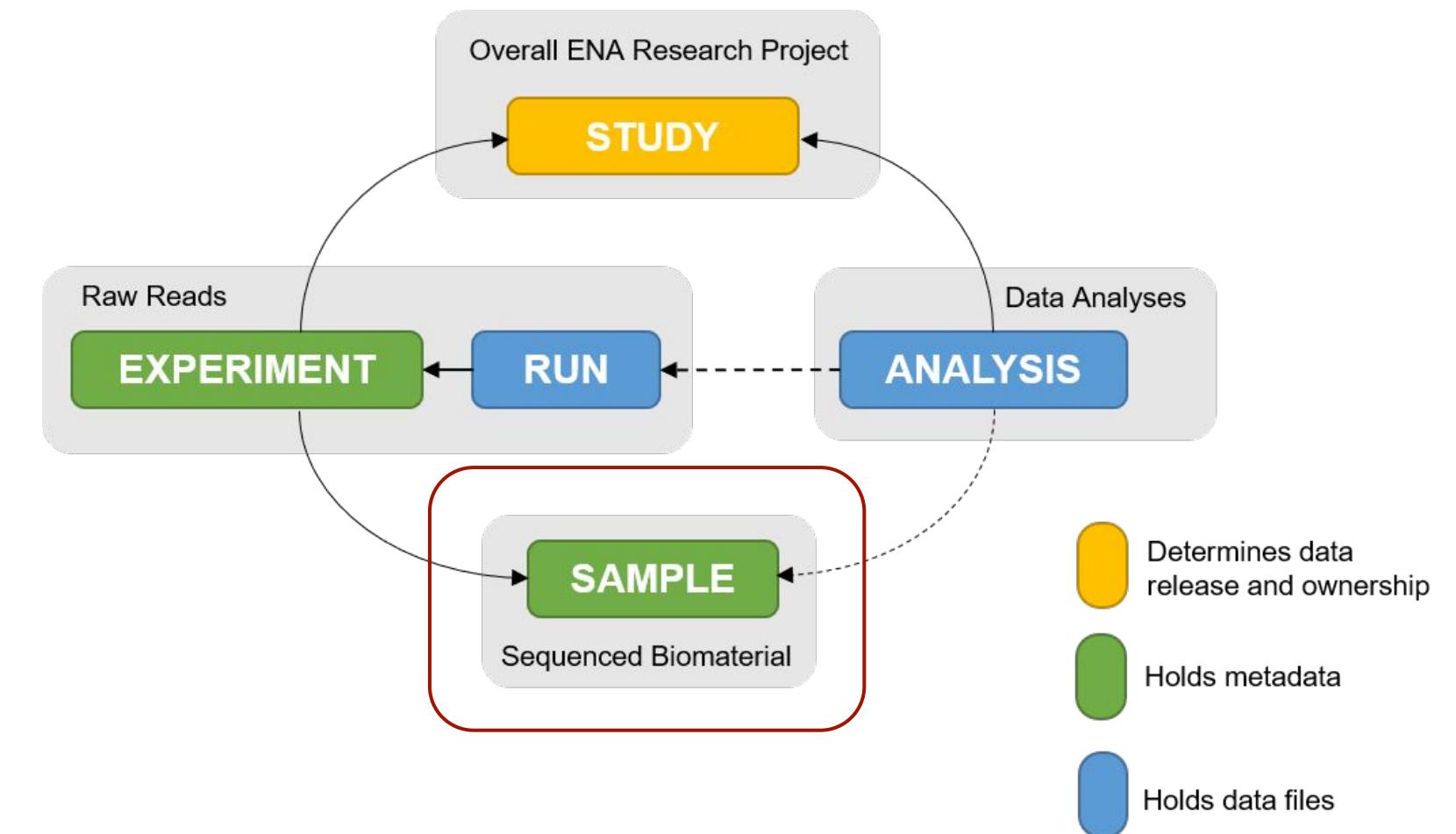
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

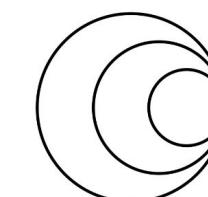
ENA Metadata Model: Sample

- 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



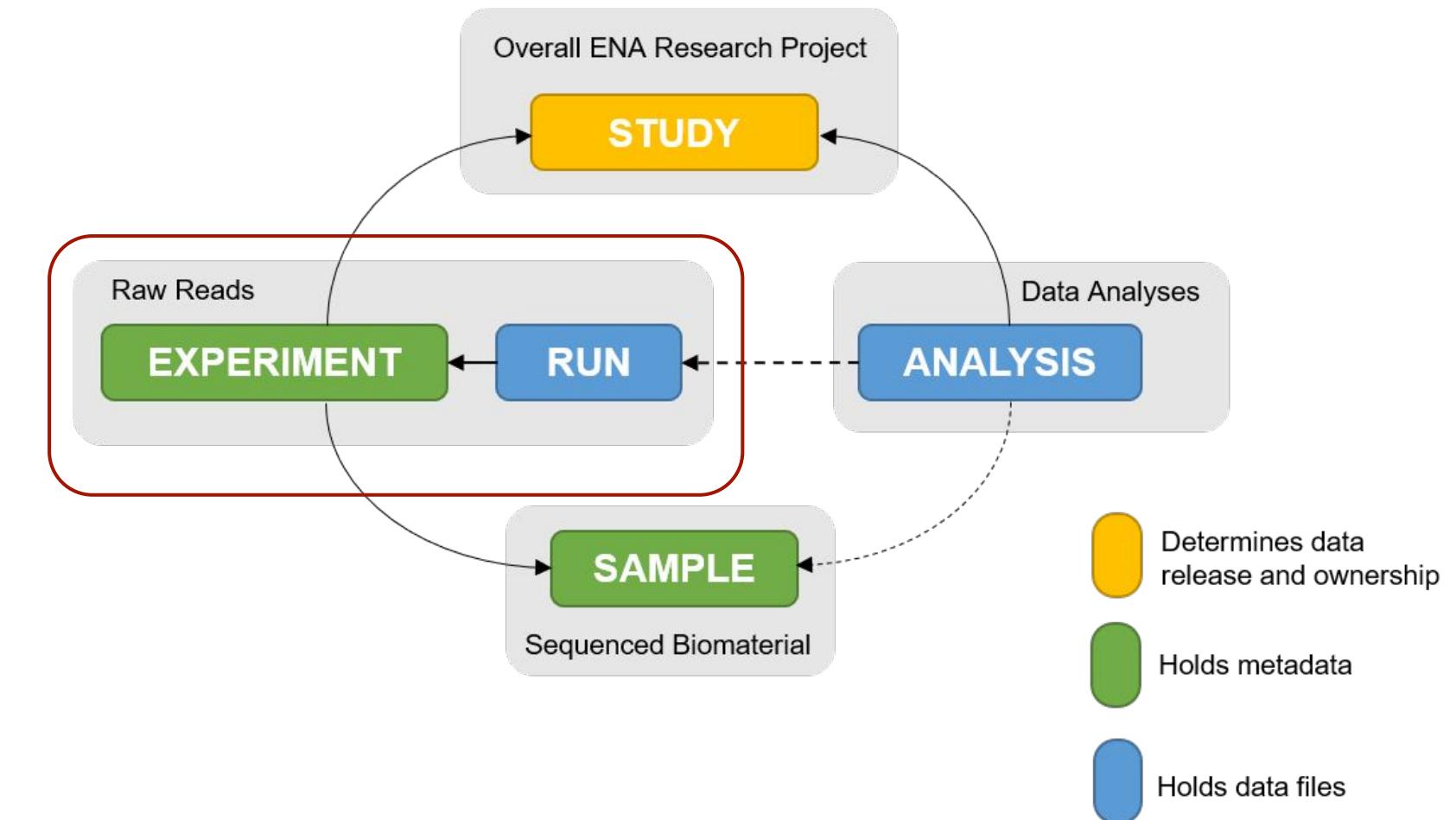
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

ENA Metadata Model: Experiment & Run

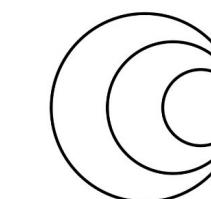
- Experiment
 - Metadata about sequencing methodology
 - Accession like 'ERX*'
 - Example metadata
 - instrument platform and model
 - library preparation information, e.g. construction protocol, primers
- Run
 - Holds data file, e.g. BAM/CRAM/FASTQ
 - Accession like 'ERR*'



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

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

View> XML



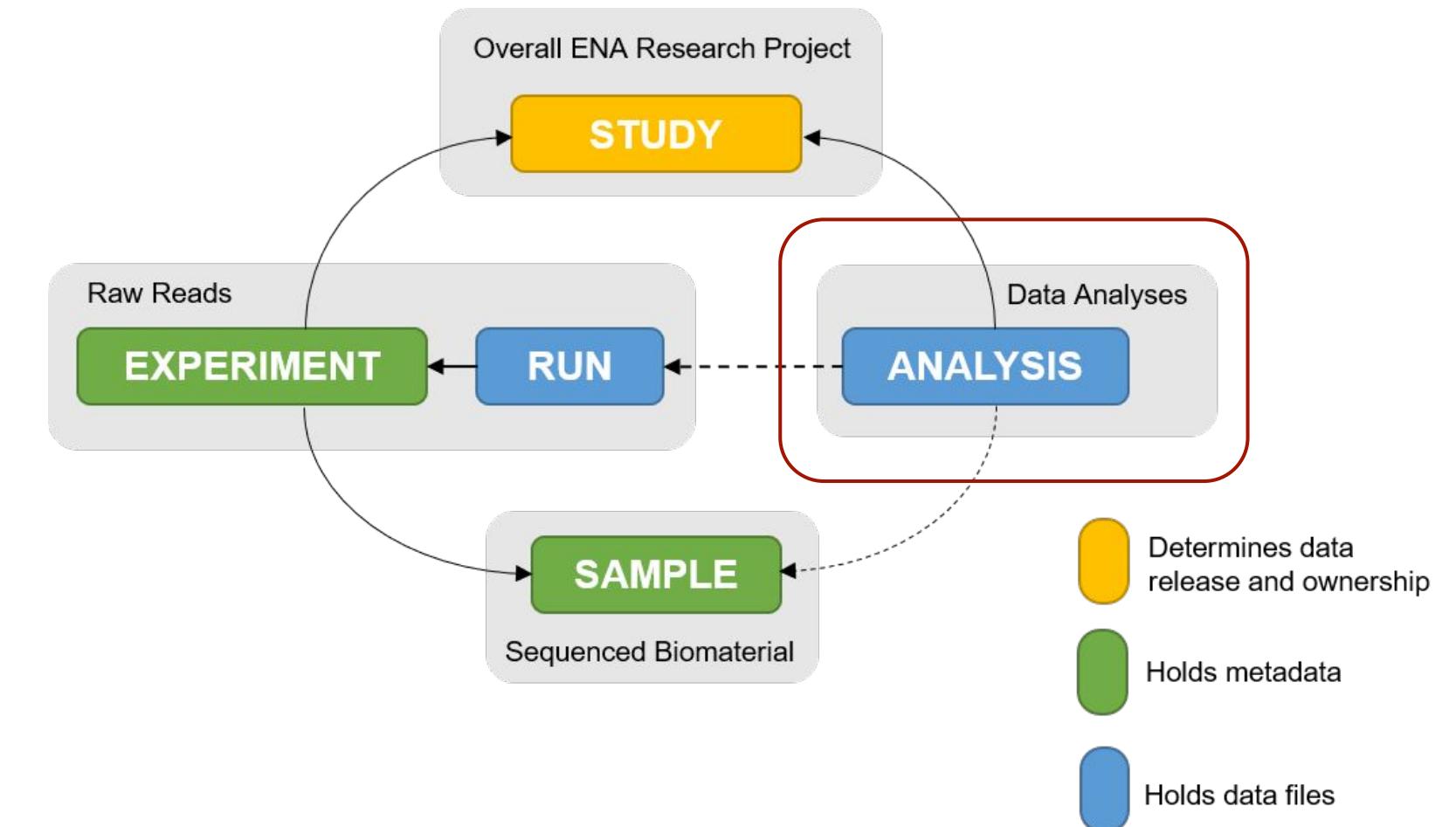
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

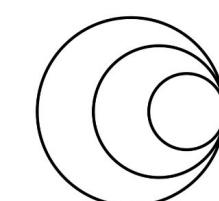
ENA Metadata Model: Analysis

- 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

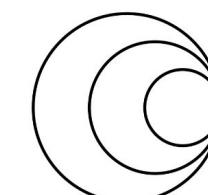


COVID-19
GENOMICS
GLOBAL TRAINING

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>
```



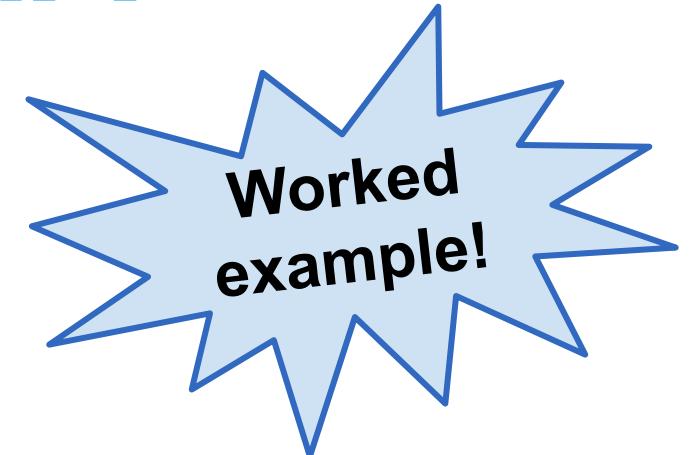
wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

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:

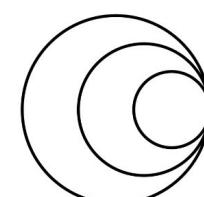
- interactive (browser-based)
- programmatic (XML-based)
- Webin-CLI (command line tool)

Today you will test the submission of:

An **ENA Project** and **Samples interactively**

&

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



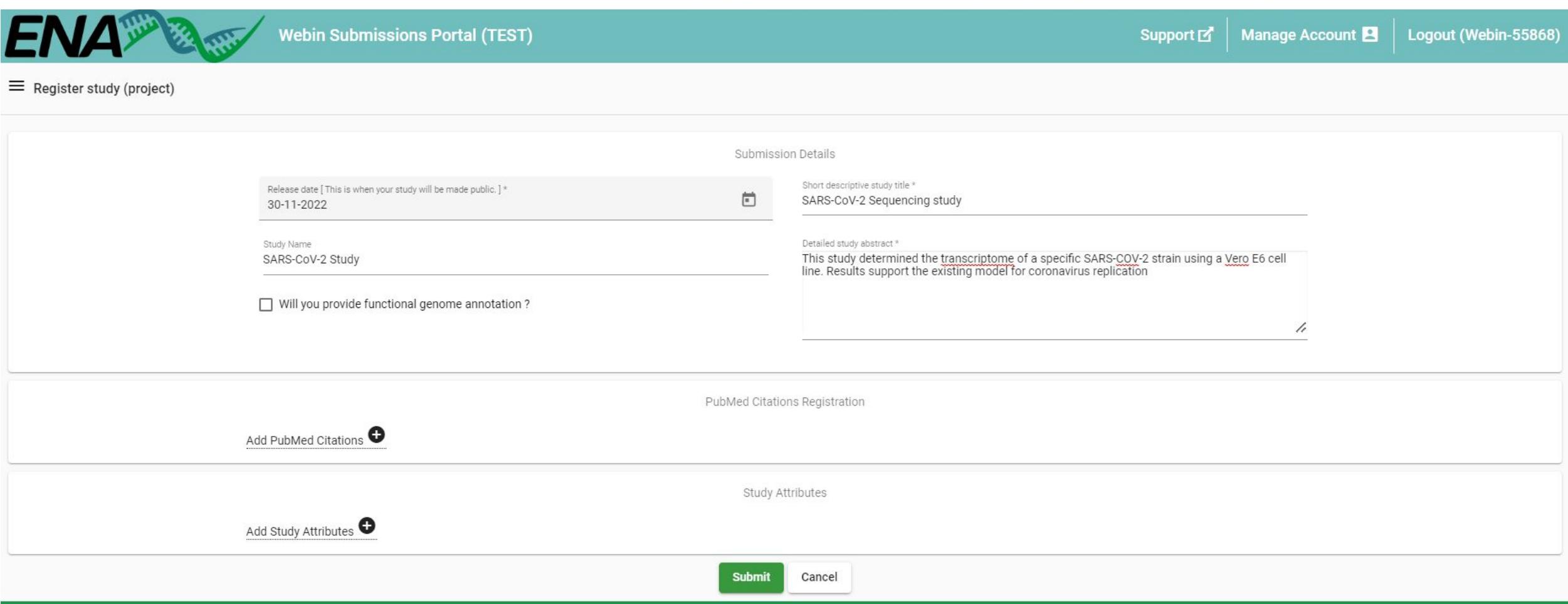
wellcome
connecting
science



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:



ENAS Webin Submissions Portal (TEST)

Support | Manage Account | Logout (Webin-55868)

Register study (project)

Submission Details

Release date [This is when your study will be made public.] *
30-11-2022

Short descriptive study title *
SARS-CoV-2 Sequencing study

Detailed study abstract *
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

Will you provide functional genome annotation ?

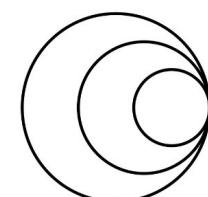
PubMed Citations Registration

Add PubMed Citations +

Study Attributes

Add Study Attributes +

Submit Cancel



wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Submitting COVID samples to the ENA - 1

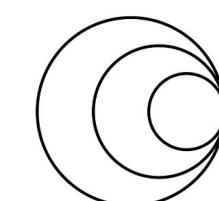
1. Download and unzip the `Module4_data_zip` folder - link [here](#)
2. Using the pre-filled ENA sample spreadsheet: `sample_spreadsheet_COG_Train.tsv...`

1	Checklist	ERC00003: ENA virus pathogen reporting standard checklist	2	tax_id	scientific_sample_id	sample_type	sample_description	collection_date	geographic_sample_type	host_composition	host_subject	host_health	host_sex	host_scientific_name	host_collector_institution
3	#units														
4	2697049	Severe acute 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	University of
5	2697049	Severe acute 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
6	2697049	Severe acute 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 WalEBI
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

host sex	sample capture status
not collected	active surveillance in response to outbreak
not collected	active surveillance in response to outbreak
not collected	active surveillance in response to outbreak
	active surveillance in response to outbreak

- '[GISAID Accession ID](#)' custom attribute



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... 

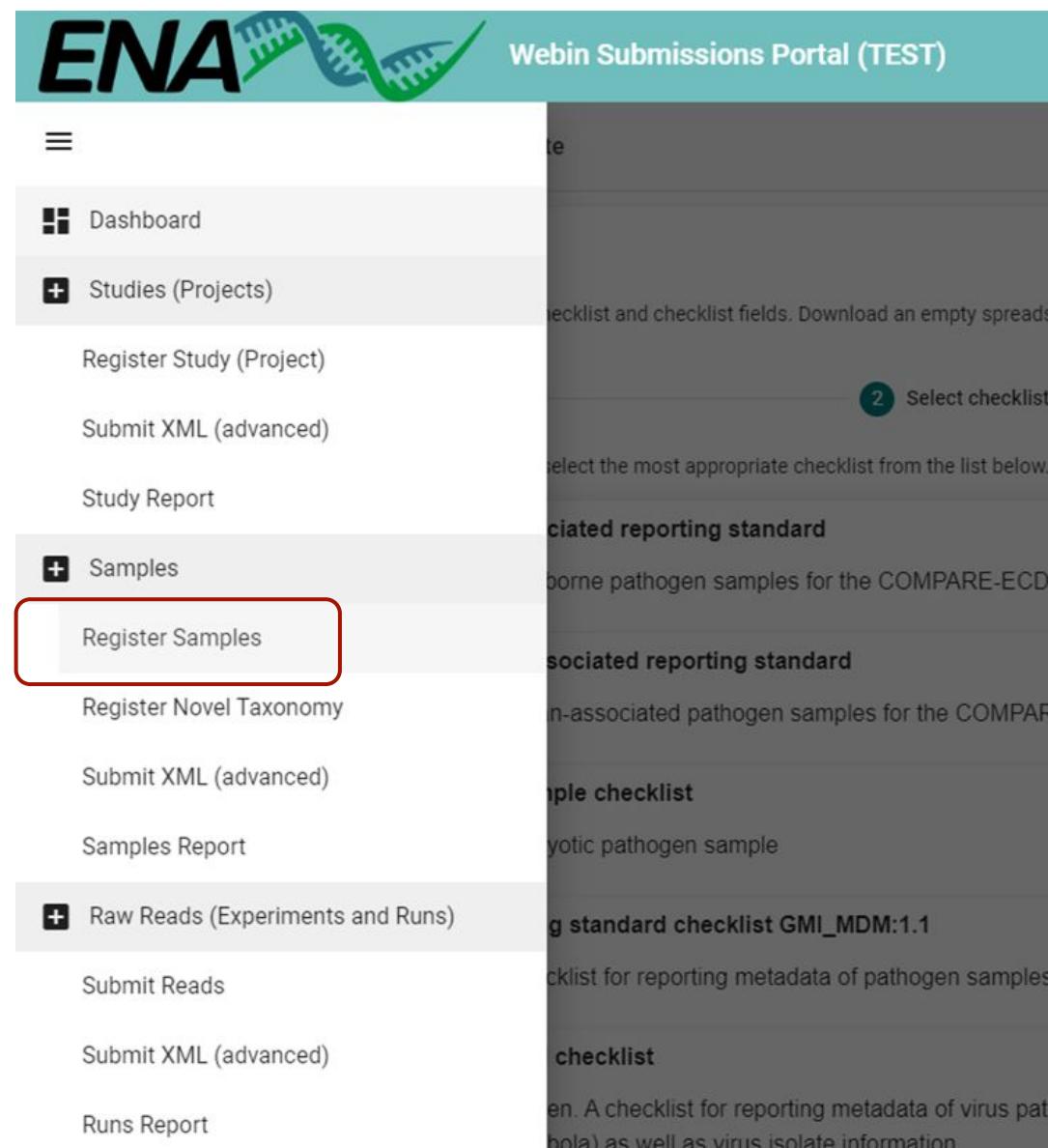
Filter by type:

- Collection event information
- host description
- General collection event information
- Infraspecies information

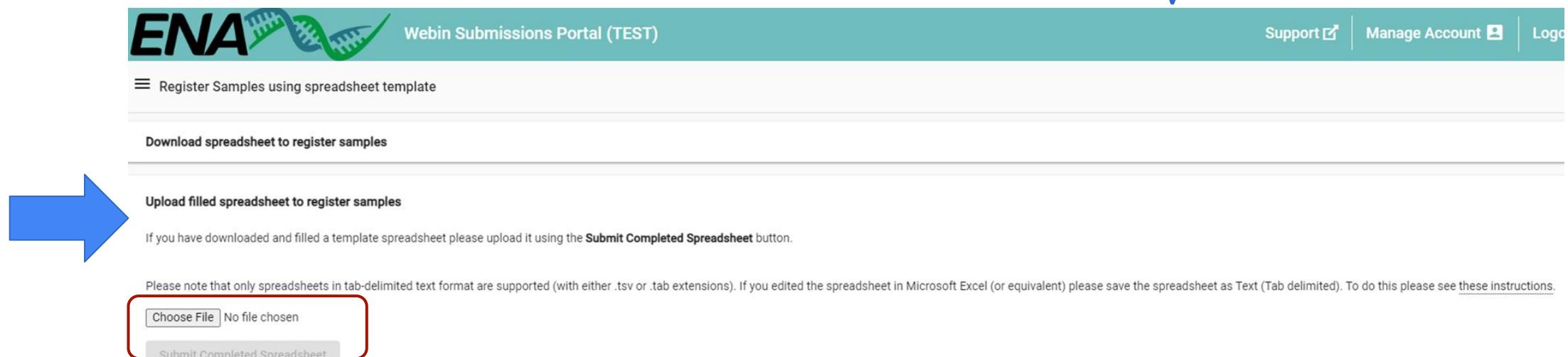
Field Name	Field Format	(Field Restriction)	Requirement	Mandatory	(Units)
geographic location (country and/or sea)	 text choice	options		mandatory	
host common name	 free text			mandatory	
host subject id	 free text			mandatory	
host health state	 text choice	options		mandatory	
host sex	 text choice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

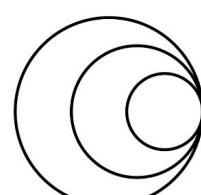
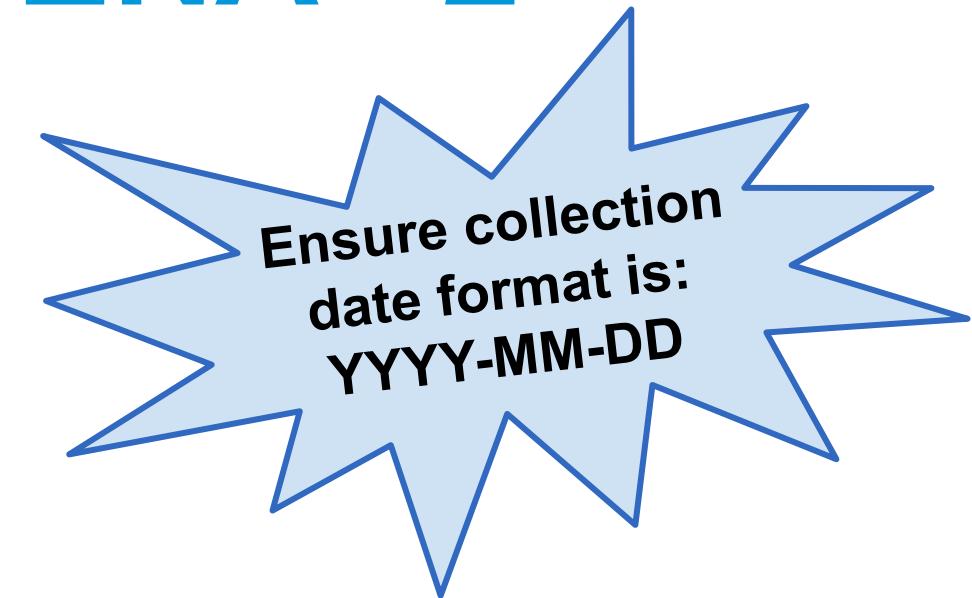
...upload this directly to the Webin Submissions Portal:



The screenshot shows the ENA Webin Submissions Portal (TEST) interface. The left sidebar has a 'Samples' section where 'Register Samples' is highlighted with a red border. A large blue arrow points from this screen to the next one.



The screenshot shows the 'Upload filled spreadsheet to register samples' step. It includes instructions, a 'Choose File' button with 'No file chosen', and a 'Submit Completed Spreadsheet' button. A large blue arrow points from the previous screen to this one.



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.

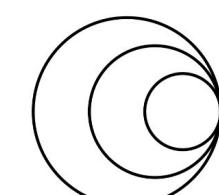
Please click search to see the results.

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	
ERS13666079	SAMEA130171409	Case B	Severe acute respiratory syndrome coronavirus 2	2697049	6th Nov 2022	Private	
ERS13666078	SAMEA130171408	Case A	Severe acute respiratory syndrome coronavirus 2	2697049	6th Nov 2022	Private	



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 x
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

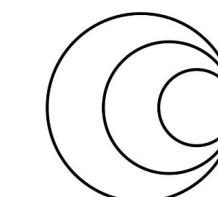
'ASSEMBLY_TYPE' should be set to 'COVID-19 OUTBREAK'

```
CaseA_chromosome_list.txt x
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



wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

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###) accsesions**
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###) accesions**
 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

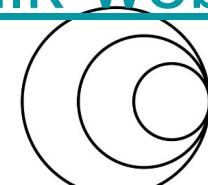
Introduction

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

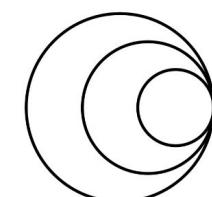
- For high-volume and/or frequent submissions (e.g. brokered data)
- Create and submit XMLs for Projects, Samples, Runs/Experiments (**X** analysis)
- Submit via cURL

```
<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>
```

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闻, on behalf of the Wellcome Sanger Institute COVID-19 Surveillance Team</DESIGN_DESCRIPTION>
      <SAMPLE_DESCRIPTOR accession="ERS12524969">
        <IDENTIFIERS>
          <PRIMARY_ID>ERS12524969</PRIMARY_ID>
          <EXTERNAL_ID namespace="BioSample">SAMEA110427043</EXTERNAL_ID>
        </IDENTIFIERS>
      </SAMPLE_DESCRIPTOR>
    </DESIGN>
  </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

No parameters

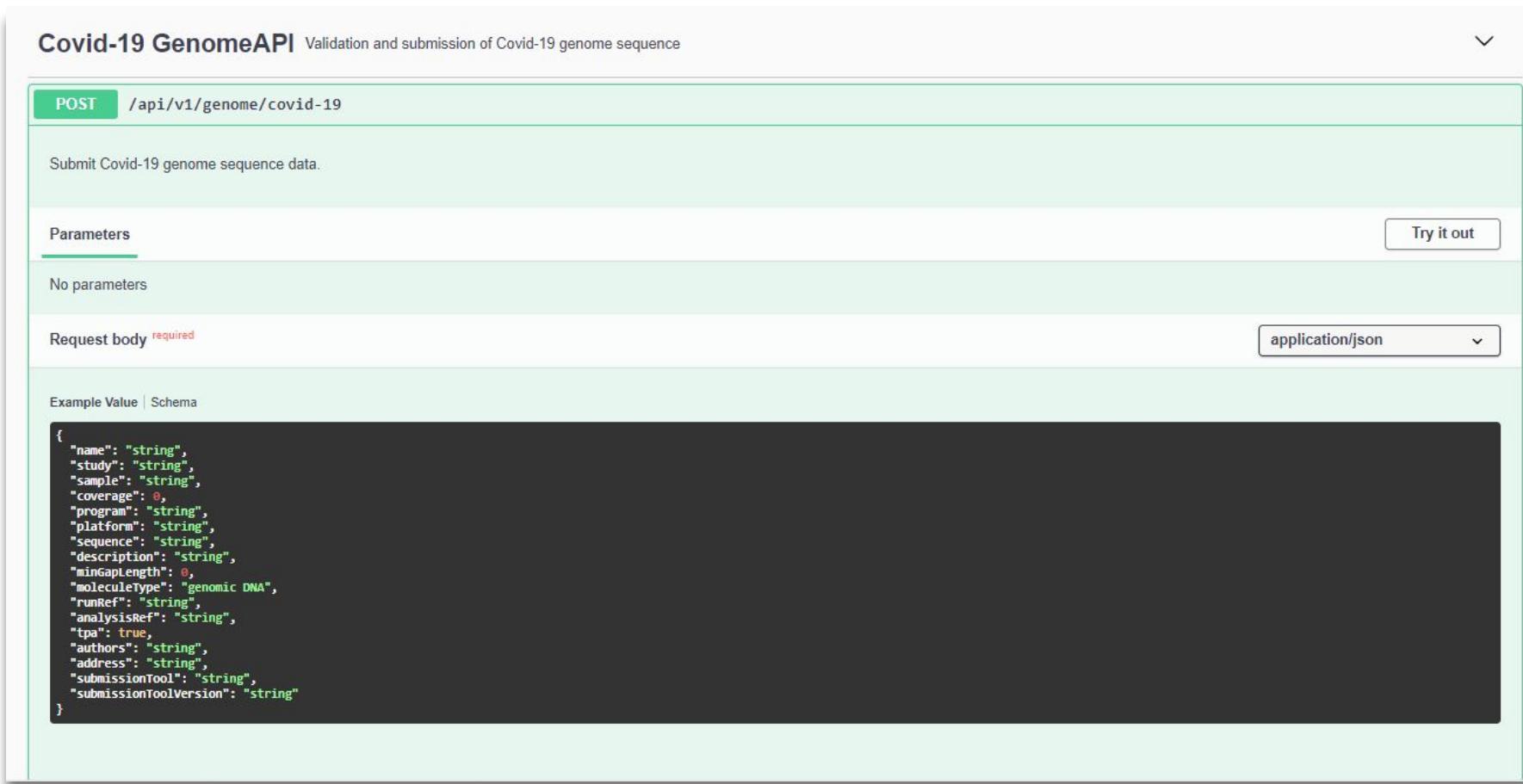
Request body required

application/json

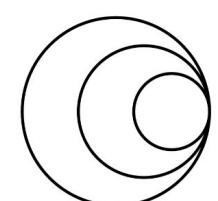
Example Value | Schema

```
{  
  "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"  
}
```

Try it out



<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

SARS-CoV-2 Data Hubs

Data Uploader
The gateway to submit your data

Login

Your secure key
d64d2fbe-3aaa-4783-8a3c-a6adc57e137f

Enter your name
user

Enter your email
user@institution.com

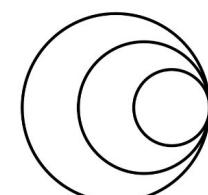
Do you have a Webin Submission Account? Yes No

Enter your Webin Submission Account ID
Webin-####

Reset

Email virus-dataflow@ebi.ac.uk for login details & metadata spreadsheet

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



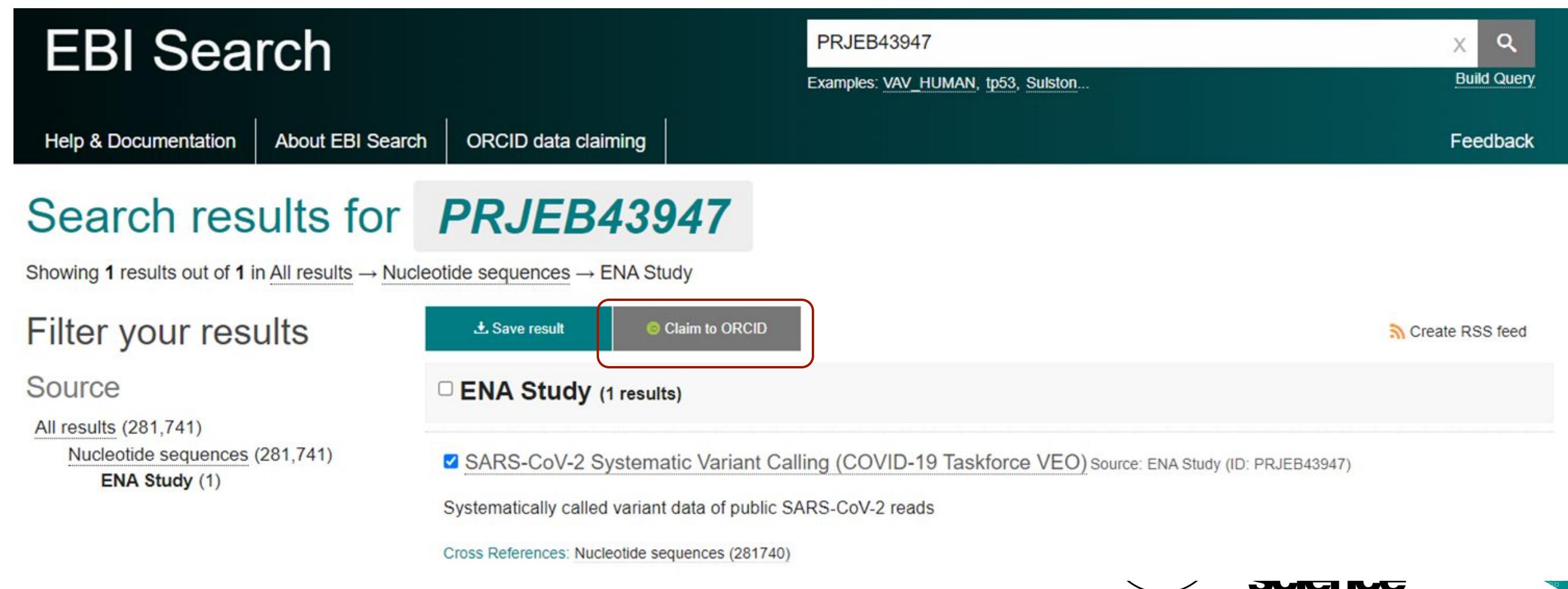
wellcome
connecting
science



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



EBI Search

PRJEB43947

Examples: VAV_HUMAN, tp53, Sulston...

Help & Documentation | About EBI Search | ORCID data claiming | Build Query | Feedback

Search results for **PRJEB43947**

Showing 1 results out of 1 in All results → Nucleotide sequences → ENA Study

Filter your results

Save result | **Claim to ORCID** | Create RSS feed

Source

All results (281,741) | Nucleotide sequences (281,741) | ENA Study (1)

ENA Study (1 results)

SARS-CoV-2 Systematic Variant Calling (COVID-19 Taskforce VEO) Source: ENA Study (ID: PRJEB43947)

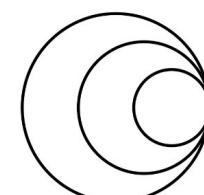
Systematically called variant data of public SARS-CoV-2 reads

Cross References: Nucleotide sequences (281740)

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
- Detailed SARS-CoV-2 workshop:
https://ena-covid19-docs.readthedocs.io/en/latest/submission_workshop/getting_started.html

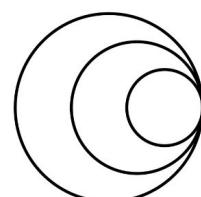


wellcome
connecting
science



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

The screenshot shows the 'Viral sequences' section of the COVID-19 Data Portal. It features a search bar with 'country:United Kingdom' and a 'Search' button. Below the search bar is a text input field with examples like 'lineage:B.1.1.7, whocomicron, Severe acute respiratory syndrome 2...Advanced search'. The main area displays a 3D model of a virus particle. A sidebar on the left lists data types: All, Viral sequences (5,850,770), Sequences (2,180,670), Representative sequences (58), Raw reads (2,609,828), Systematic Analyses (1,060,214), Center name (COVID-19 Genomics UK Consortium (2,178,333), Modernising Medical Microbiology (1,395), Quadram Institute Bioscience (889)).

COVID Portal Advanced Search

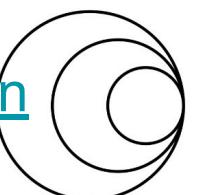
The screenshot shows the 'Advanced search' page of the COVID Portal. It has tabs for 'Build query' and 'Load query'. Under 'Choose category', there are dropdown menus for 'Viral sequences' and 'Sequences'. Below these are radio buttons for 'AND', 'OR', and 'NOT', with 'AND' selected. There are 'Add rule' and 'Add ruleset' buttons. At the bottom are 'Save', 'Reset', and 'Search' buttons. A note at the bottom right says 'on the European COVID-19 platform: ecovid19@ebi.ac.uk.' and 'your data on the COVID-19 Data Portal: virus-dataflow@ebi.ac.uk.'

COVID Portal API

The screenshot shows the 'COVID-19 Data Portal API' documentation. It includes sections for 'Endpoints' and 'Parameters'. The 'Endpoints' section lists various data resources and their corresponding URLs, many of which are powered by ENA (European Nucleotide Archive). The 'Parameters' section details the fields and types for querying the API.

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: <ul style="list-style-type: none">• JSON• XML• TSV• CSV• idlist• acclist• cs_idlist• cs_acclist

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



wellcome
connecting
science



COVID-19
GENOMICS
GLOBAL TRAINING

Retrieval - interactive & programmatic

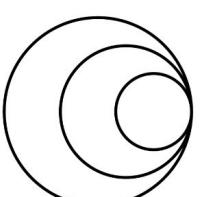
Download button on web interface

The screenshot shows the 'Viral sequences' section of the COVID-19 Data Portal. A search bar at the top has 'lineage:B.1.1.7' entered. Below it is a table of 15 results. A modal dialog box titled 'Download details: Sequences' is open, showing 'Selected results: 4' (radio button selected) and 'All results: 6341404' (radio button unselected). Under 'Data format', 'EMBL' is checked. The 'Download' button is highlighted in blue.

Bulk downloader tool

The left side of the image shows the 'Bulk Downloads' page with sections for 'Bulk Downloads' and 'CDP File Downloader'. The 'Bulk Downloads' section contains links to 'FTP Server' and 'cdp-file-downloader'. The 'CDP File Downloader' section includes links to 'CDP File Downloader', 'Features', 'How to run', 'Download Functionality', 'Support', and 'Privacy Notice'. The right side shows a terminal window titled 'Welcome to the Covid-19 Data Portal's data downloader utility!' with a copyright notice from EMBL 2021. It displays a grid of symbols (#, ., /, //, etc.) and a command prompt for selecting options: 1 for Viral Sequences, 2 for Host Sequences, 3 for Help, 4 for Privacy Notice, and 0 (zero) to exit.

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



wellcome
connecting
science



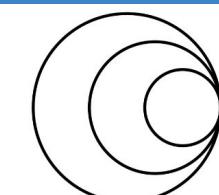
COVID-19
GENOMICS
GLOBAL TRAINING

Bulk downloader tool

- For downloading a range of [ENA](#) COVID-19 data, in a range of formats (XML/FASTA/EMBL/FASTQ)
- Create scripts to easily download data in *Non-interactive mode*
- Can re-attempt downloads if not completed, automatically tries 3 times
- Once run again, only new/updated files are downloaded

The screenshot shows the 'Bulk Downloads' section of the COVID-19 Data Portal. At the top, there's a banner with a microscopic image of cells and the text 'Bulk downloads'. Below it, a sidebar lists data types: Sequences, Reference sequences, Raw reads, Sequenced samples, Studies, Systematic analyses, Human reads, and Other species reads. To the right, under 'Bulk Downloads', there are two sections: 'Bulk Downloads' (with links to Protein data via FTP Server and ENA data via cdp-file-downloader) and 'CDP File Downloader'. The 'CDP File Downloader' section includes a brief description, a note about Java requirements, a 'Download CDP-File-Downloader' button, and a terminal window showing command-line options for the tool.

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



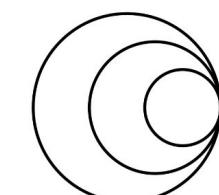
wellcome
connecting
science



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?

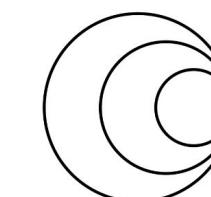


wellcome
connecting
science



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](#)

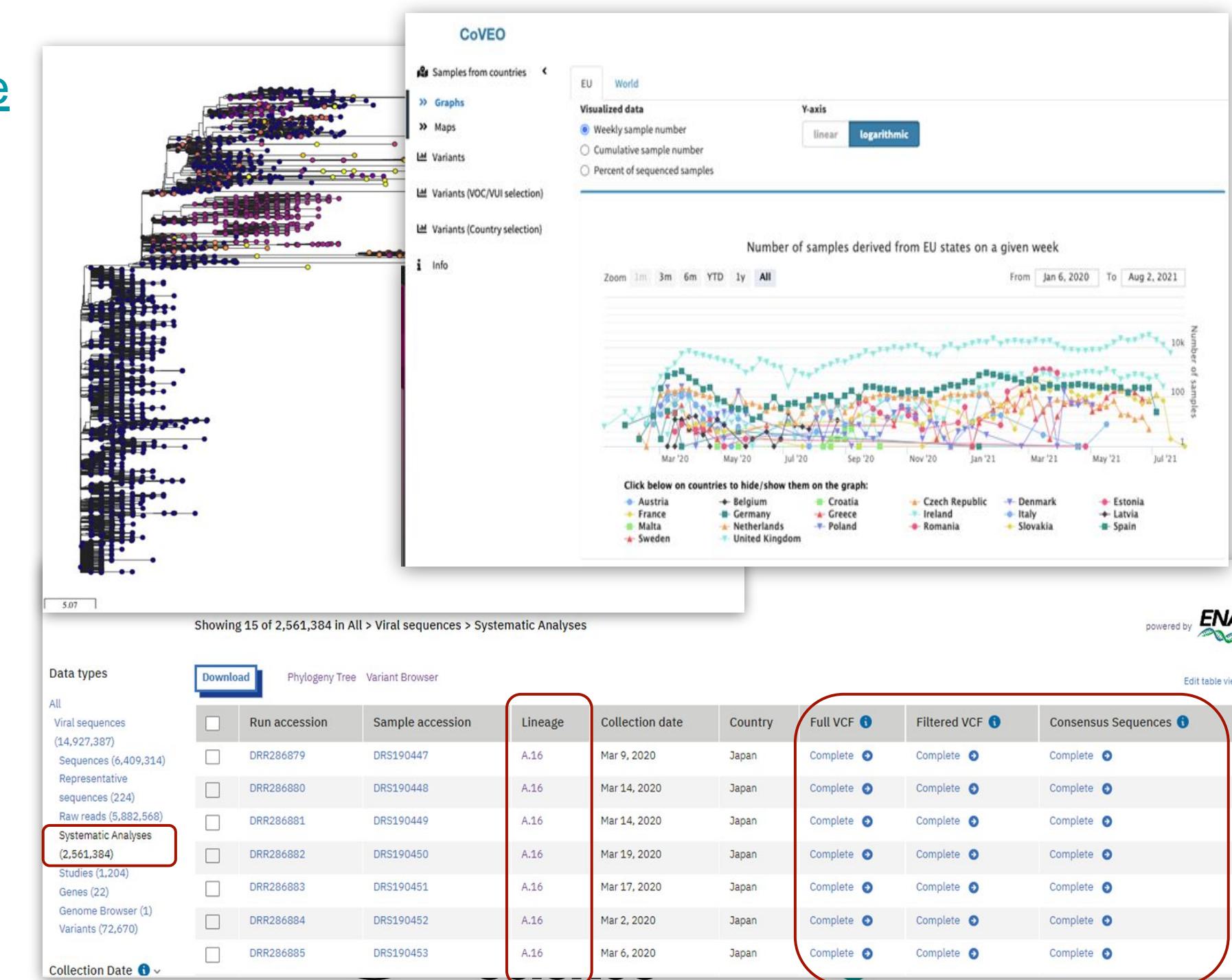
1. Consensus sequences

- Pangolin lineage assignment
- View on phylogeny tree

2. Variant calls

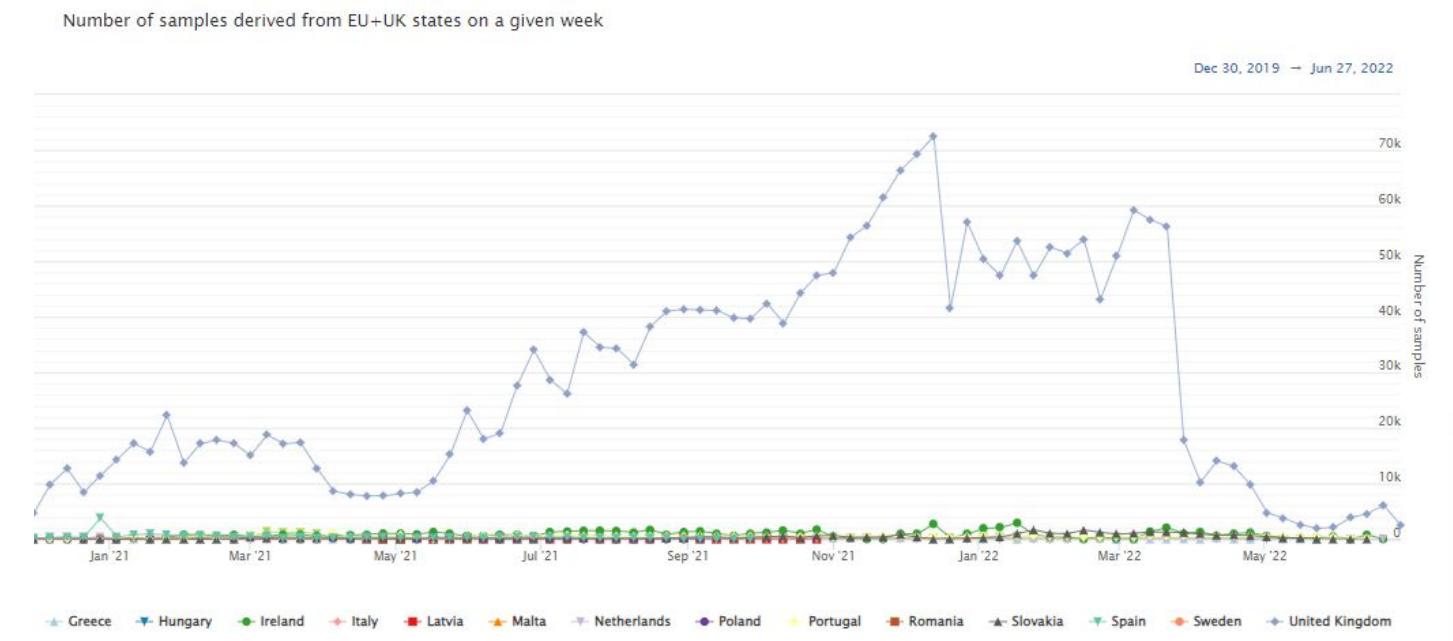
- Fed into CoVEO Variant Browser
- Submitted to European Variation Archive (EVA)

All products visualised on COVID-19 Data Portal



CoVEO Variant Browser

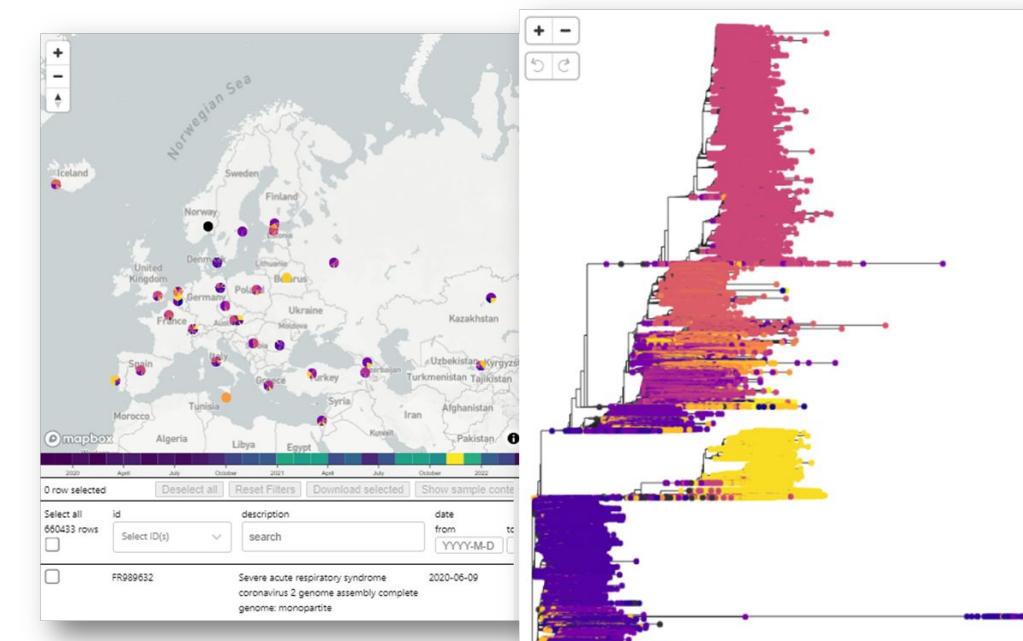
- CoVEO ingests unfiltered variant data to generate a range of plots
- Allows users to track emergence and distribution of SARS-CoV-2 variants across the world



<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/assets/pdf/evergreen_method_notes_2021-10-08.pdf

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



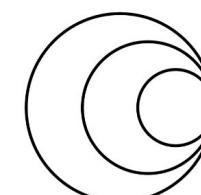
DTU
Technical University of Denmark



COVID-19 GENOMICS GLOBAL TRAINING

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?



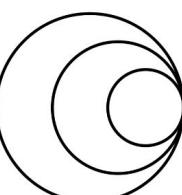
wellcome
connecting
science



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