Genomic Sequencing and Implementation of Genomic Surveillance Strategies

Nuno R. Faria

Co-Lead Pathogen Genomic Epidemiology Unit

MRC Centre for Global Infectious Disease Analysis

WHO Collaborating Centre for Infectious Disease Modelling

School of Public Health, Imperial College London, UK (SPH-ICL)

Institute of Tropical Medicine, University of São Paulo, Brazil (IMT-USP)

Email: nfaria@ic.ac.uk – São Paulo, 10 July 2023

SPSAS

São Paulo School of Advanced Science on Epidemic Preparedness

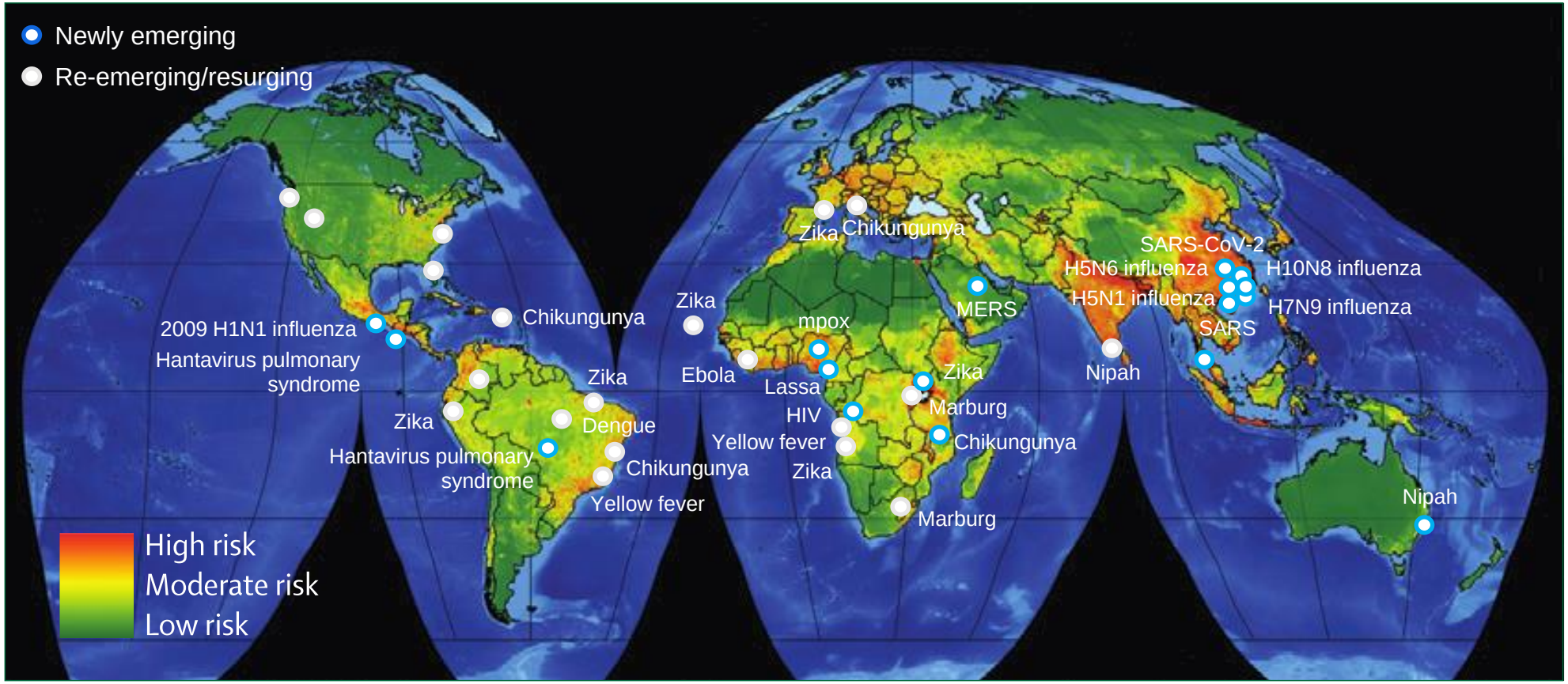
Outline of today's talk

1. Introduction
2. From traditional diagnostics to next generation sequencing
3. Pathogen genome sequencing implementation
4. Genome sequencing strategies
5. Global surveillance networks and resources
6. Disparities in global sequencing capacity
7. Summary and concluding remarks

Context

Global hotspots for emerging infectious diseases that originate in wildlife

- 75% of all emerging pathogens have a zoonotic origin.

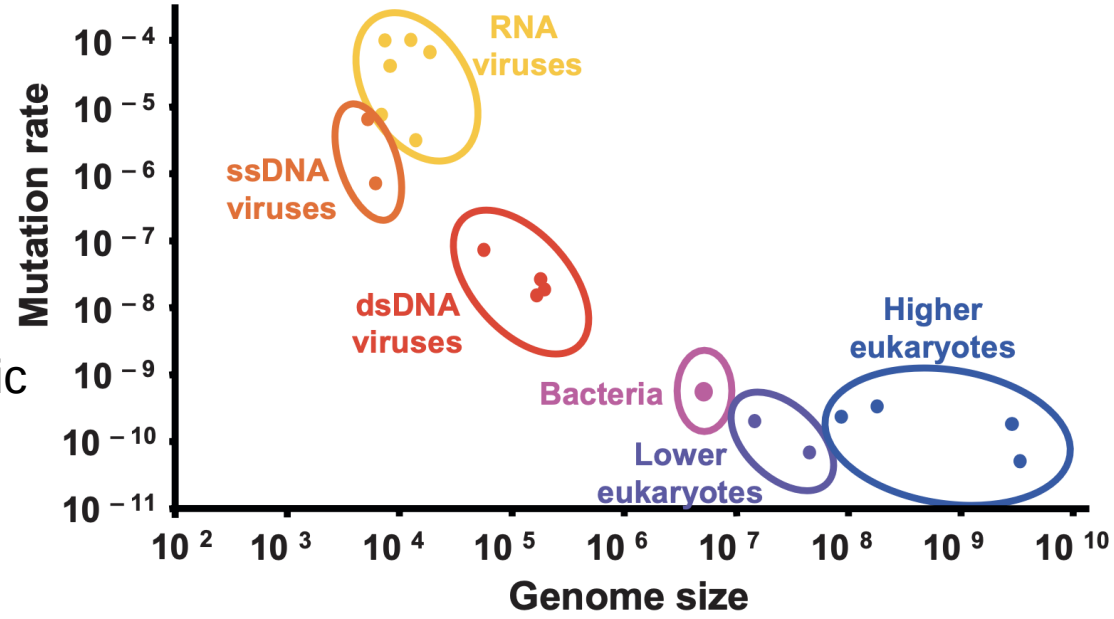


Adapted from Morse et al. *Lancet* 2012; Morens & Fauci *Cell* 2020

Public Health Emergencies of International Concern (PHEIC) have all been caused by viruses

- RNA viruses: rapid mutation rates, short generation times, and large population sizes facilitate the rapid accumulation of genetic diversity over observable timescales (days/months) and host species plasticity.
- Nearly all PHEIC to date have been caused by RNA viruses.

1. H1N1pdm (2009) **RNA** | Zoonotic
2. Polio (2014) **RNA** | Zoonotic?
3. Ebola virus (2014) **RNA** | Zoonotic
4. Zika virus (2016) **RNA** | Zoonotic
5. Kivu Ebola (2018–2020) **RNA** | Zoonotic
6. SARS-CoV-2 (2020) **RNA** | Zoonotic
7. Monkeypox (2022) **dsDNA** | Zoonotic

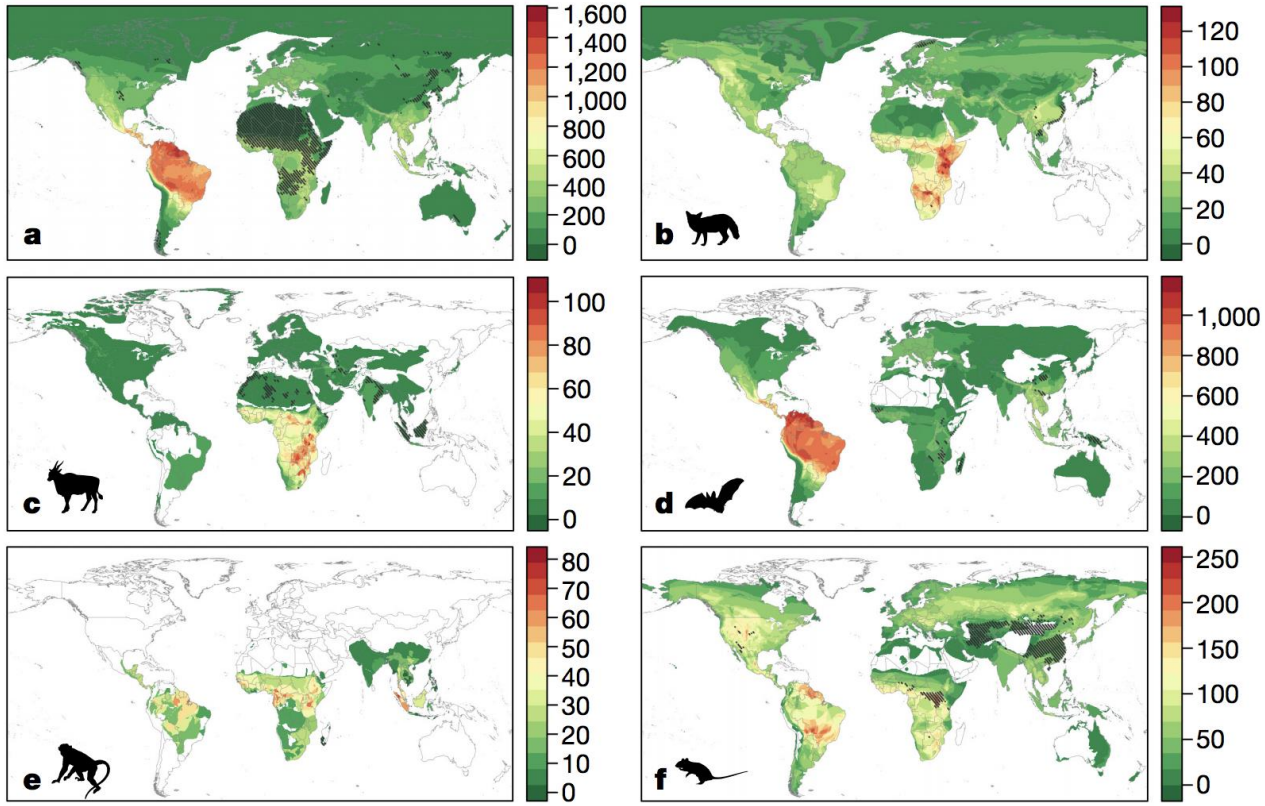


Adapted from Gago et al. *Science* 2009

The emergence of zoonotic viruses with epidemic and pandemic potential is largely unpredictable

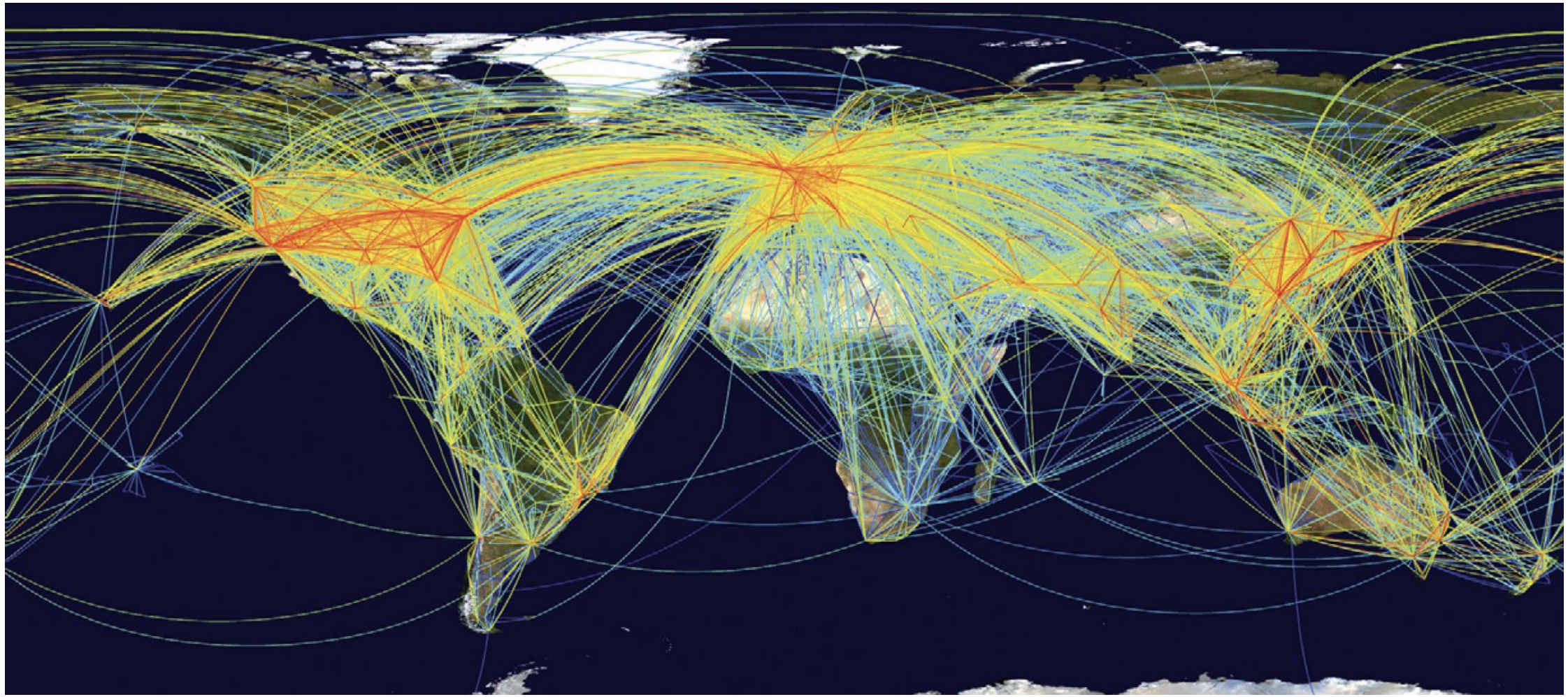
- Estimates of **1.6 million animal viruses**: how many are capable of replicating in humans and transmitting between?
- Around **250 human viruses** have been described, and only a small subset of these have caused major epidemics.
- Unknown viruses are often undetected until they cause disease in humans.
- **Effective surveillance: critical to rapidly identify emerging and re-emerging viruses.**

Global distribution of predicted number of missing zoonoses by order



Olival et al. *Nature* 2017; Weiss et al. *Nature* 2018; Holmes et al. *Nature* 2020

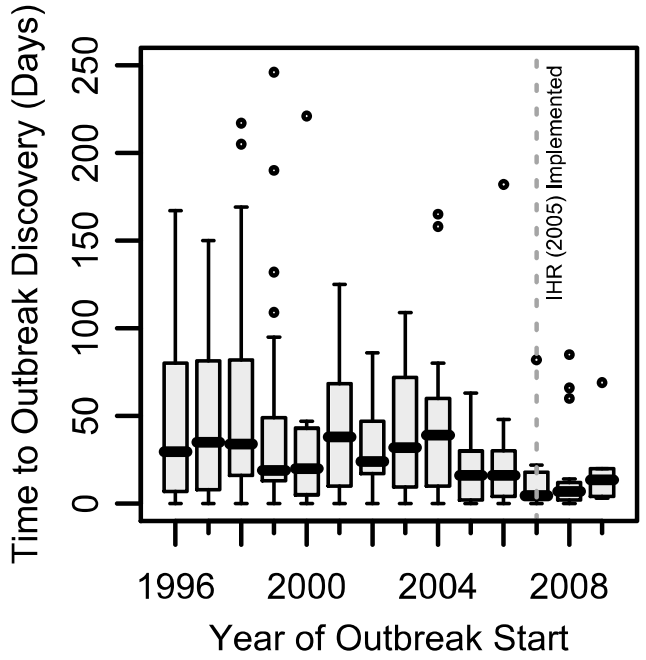
Less than 10% of the worlds population is more than 48h away from a large well-connected city



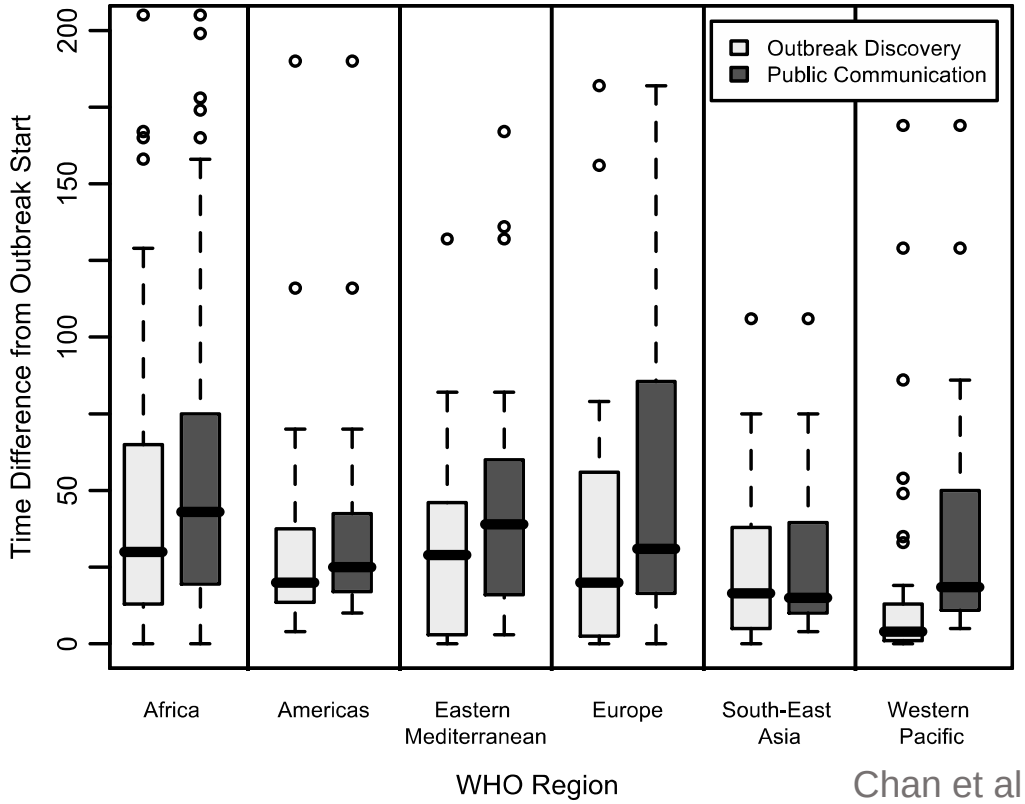
Kilpatrick and Randolph *Lancet* 2012

Global capacity for emerging infectious disease detection from 1996 to 2009

- Time from outbreak start to outbreak discovery decreased from 29.5 d (1996) to 13.5 d (2009)
- Post IHR (2007) implementation, time from outbreak start to outbreak discovery was 7 d
- Lags varied by geographic region: Africa (30 d) to Western Pacific (4 d)



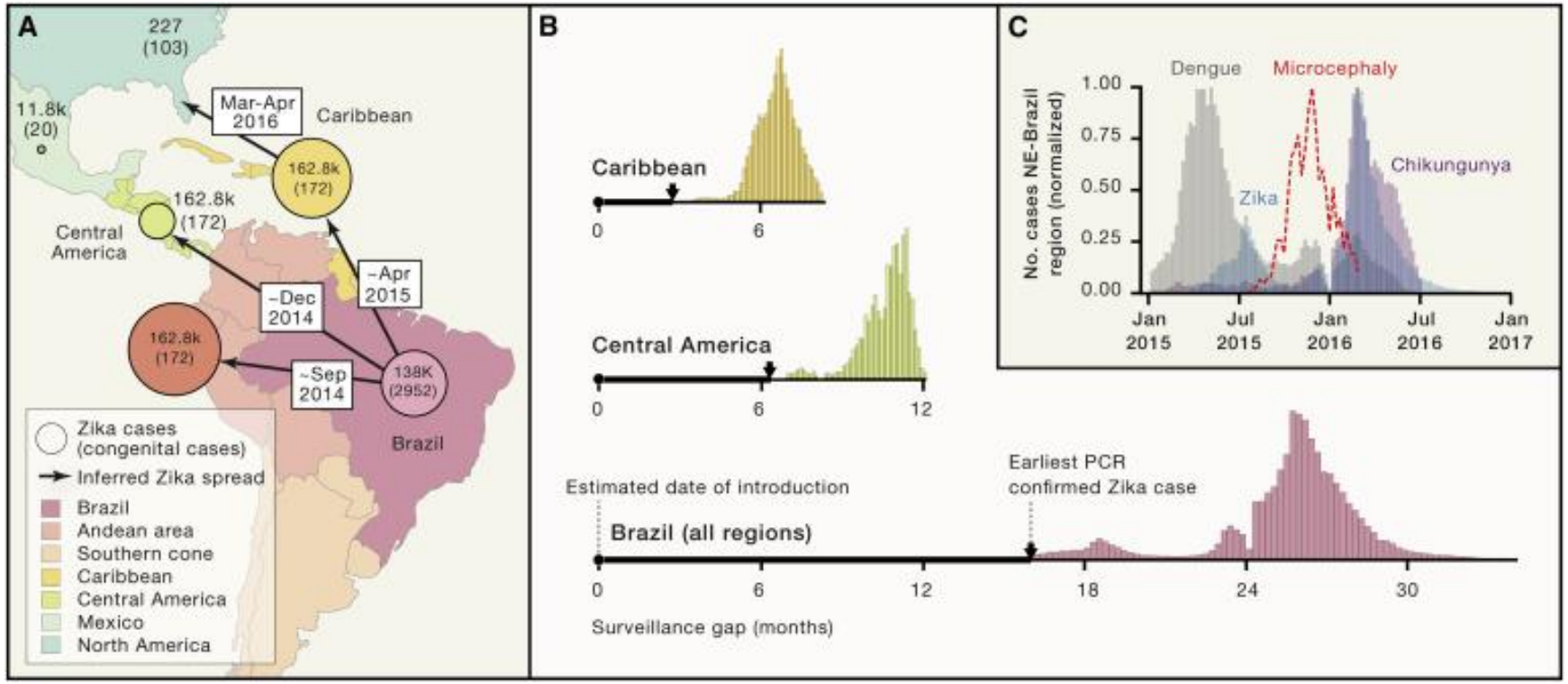
IHR (2005) requires state parties to notify the WHO of any disease event that may constitute a "public health emergency of international concern" (PHEIC).



Chan et al. *PNAS* 2010

Undetected circulation of Zika virus in the Americas highlights important surveillance gaps

- Zika virus undetected transmission highlights the important distinction between date of detection of first cases (outbreak discovery – May 2015), date of first genome sequence shared from outbreak (outbreak sequence – Jan 2016), and estimated date of emergence of the outbreak (outbreak origins – Jan 2014).



Grubaugh, Faria et al. *Cell* 2018

From traditional diagnostics to next generation sequencing

From virus culture to agnostic sequencing for real-time detection of viral pathogens

Viral culture

- 3- to 5-day from sample collection to diagnosis
- Requires robust storage and high costs
- Limitations of culturing techniques for virus detection (majority viruses cannot be cultured)

ELISA (enzyme-linked immunosorbent assays)

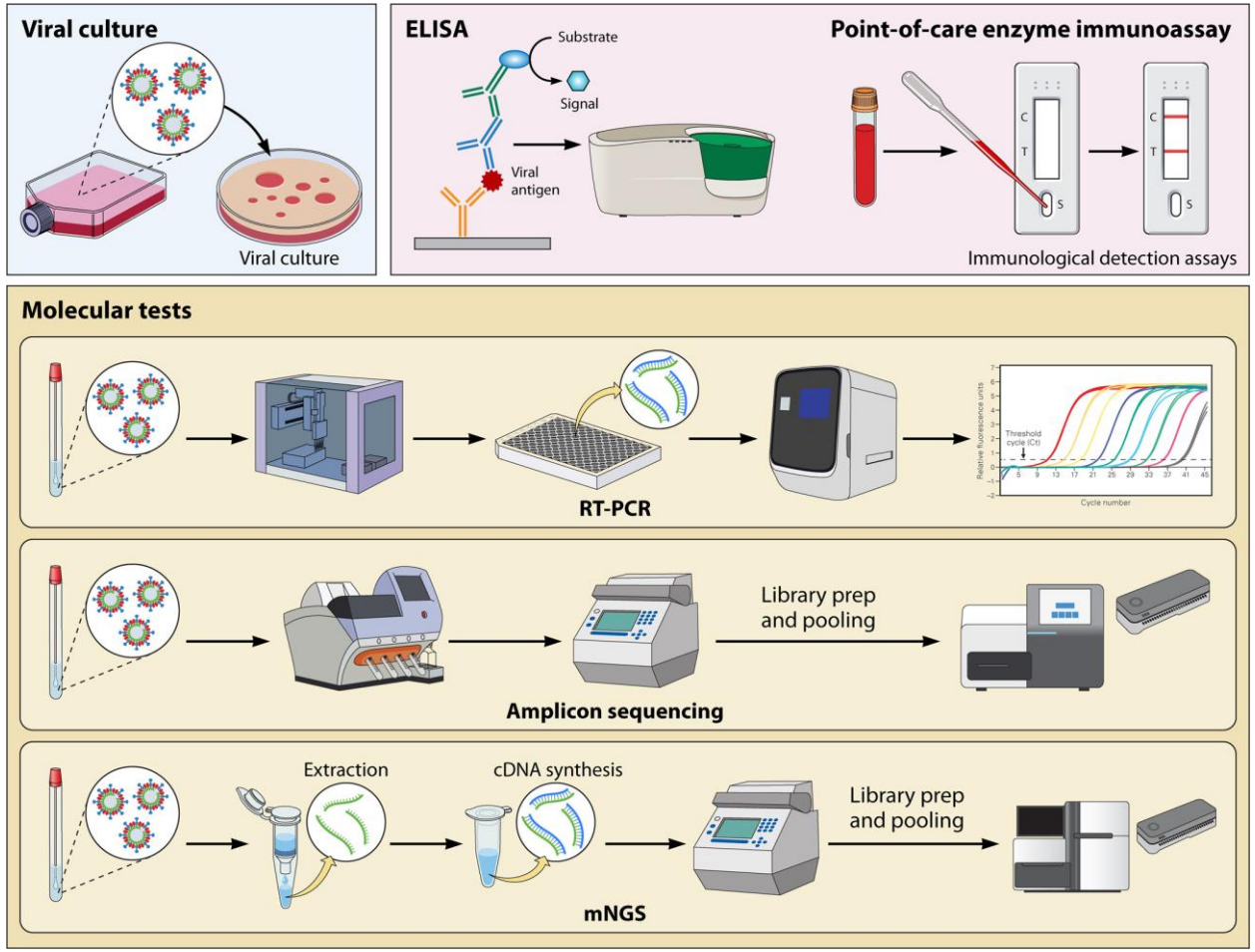
- Measures antibody or protein signals in sample
- Variable sensitivity depending on the test
- sensitivity can decline as a virus' genome mutates

Nucleic acid amplification tests (NAATs)

- Low cost and reliable, extensively validated
- Expanded capacity for routine testing
- Effective response during public health threats.

Next-Generation Sequencing (NGS) Based Diagnostic and Surveillance Strategies

- "Targeted" NGS (culture-independent)
- Unbiased mNGS (sequence- & culture-independent)



Adapted from Gauthier et al. *Clinical Microb Rev* 2023

Laboratory workflows for targeted and untargeted detection of viral pathogens

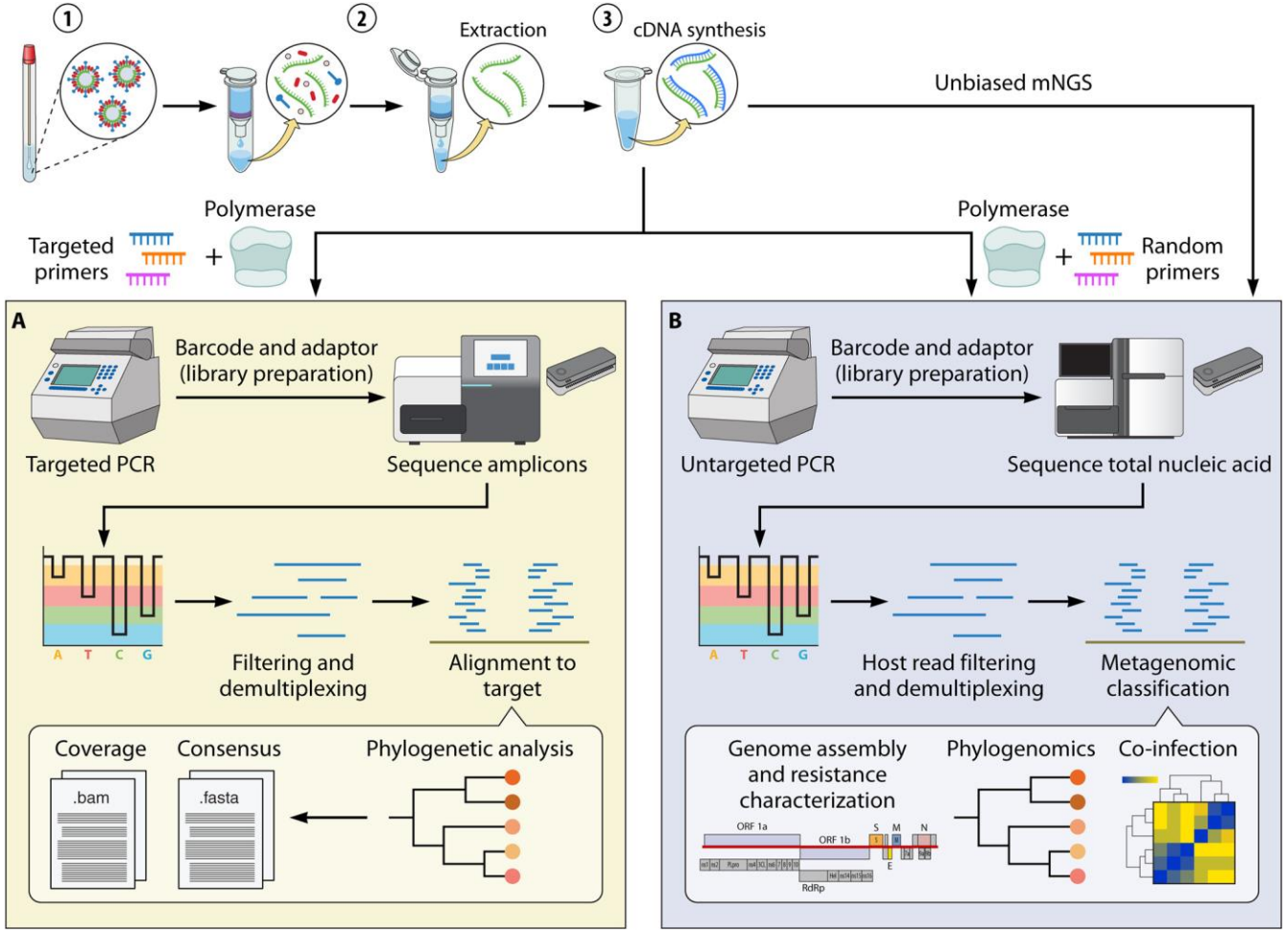
Sample collection, processing (1), and nucleic acid extraction (2) are similar for the two approaches.

Untargeted (unbiased) mNGS

- Don’t require knowledge of pathogen’s genome
- Can detect multiple pathogens within a sample
- Useful for emerging pathogens, rare pathogens, clinical specimens from infections of unknown etiologic origins, or coinfections.
- May require more sample volume, additional steps to reduce host nucleic acid and/or increase ratio viral nucleic acid to host background.

Targeted NGS

- Require knowledge of pathogen’s genome
- Can detect single pathogen within a sample
- Allow clinical laboratories to perform massively parallel sequencing for outbreak investigation and public health surveillance for a single virus.

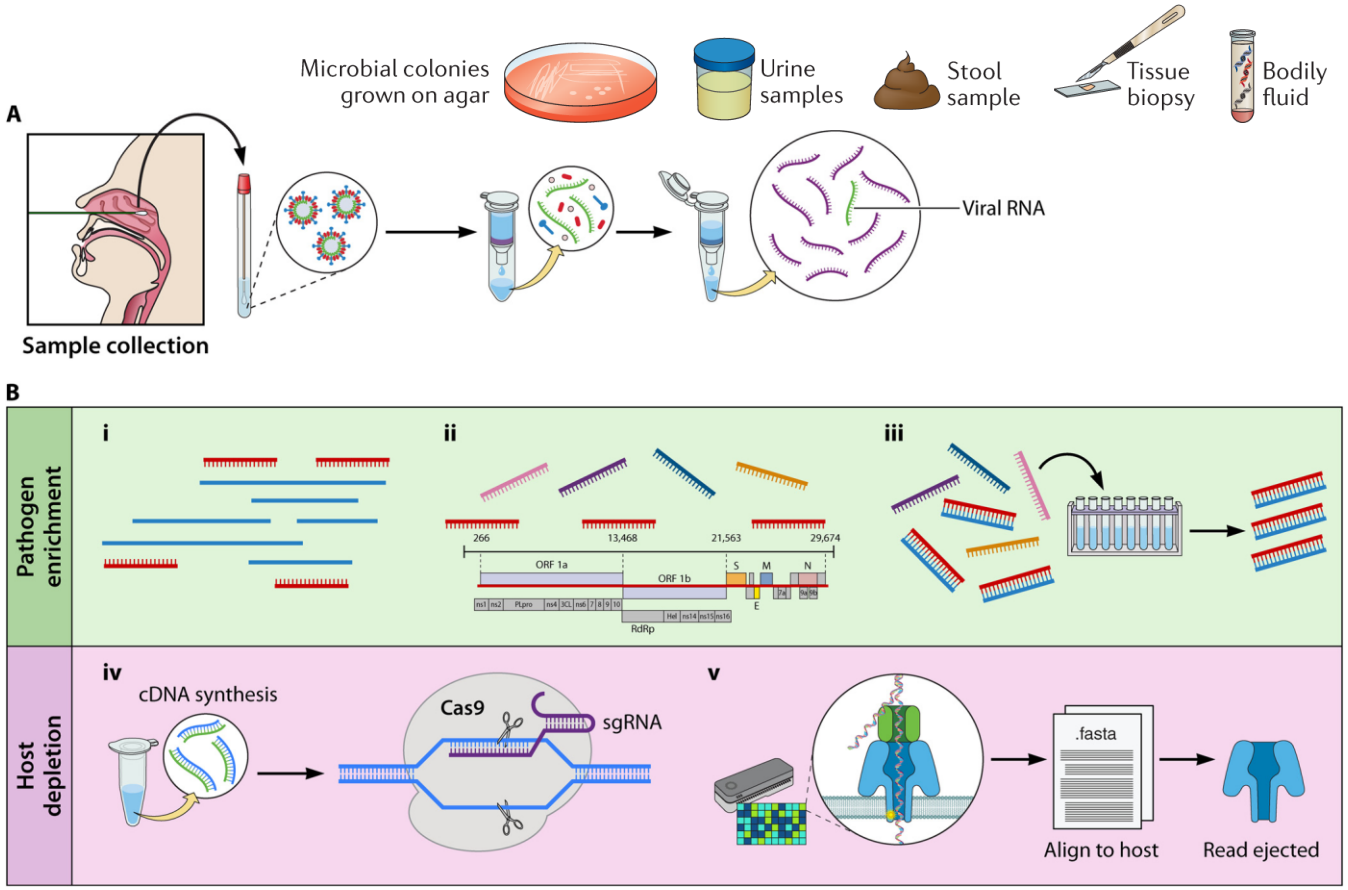


Adapted from Gauthier et al. *Clinical Microb Rev* 2023

Advances in sample processing and library preparation for real-time sequencing at outbreak epicenter

- **Host depletion with hexamers** that do not bind to human rRNA – useful for samples with high abundance of host nucleic acid but can miss certain viral targets.
- **Metagenomic sequencing with spiked primer enrichment (MSSPE):** uses primers targeting a specific virus or panel of viruses in addition to random primers to detect untargeted viruses – useful for samples with low abundance of viral nucleic acid.
- **SMART-9N:** performs random priming and one-step cDNA synthesis to reduce workflow time (up to 6h from sample to sequence) (by Ingra Claro).

Other examples of types of samples for sequencing (general):

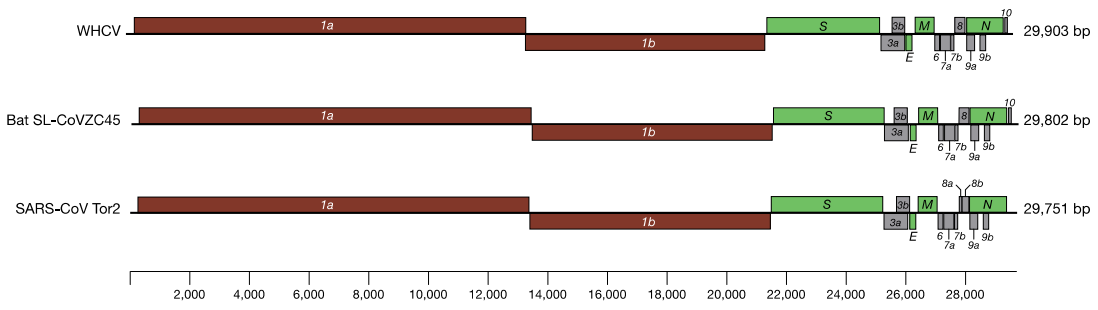


Adapted from Gauthier et al. *Clinical Microb Rev* 2023

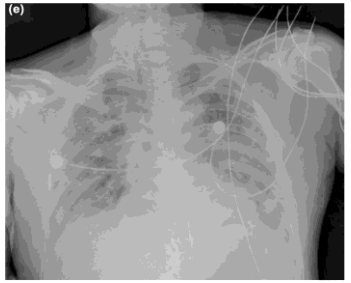
Untargeted sequencing to identify SARS-CoV-2 in Wuhan, China

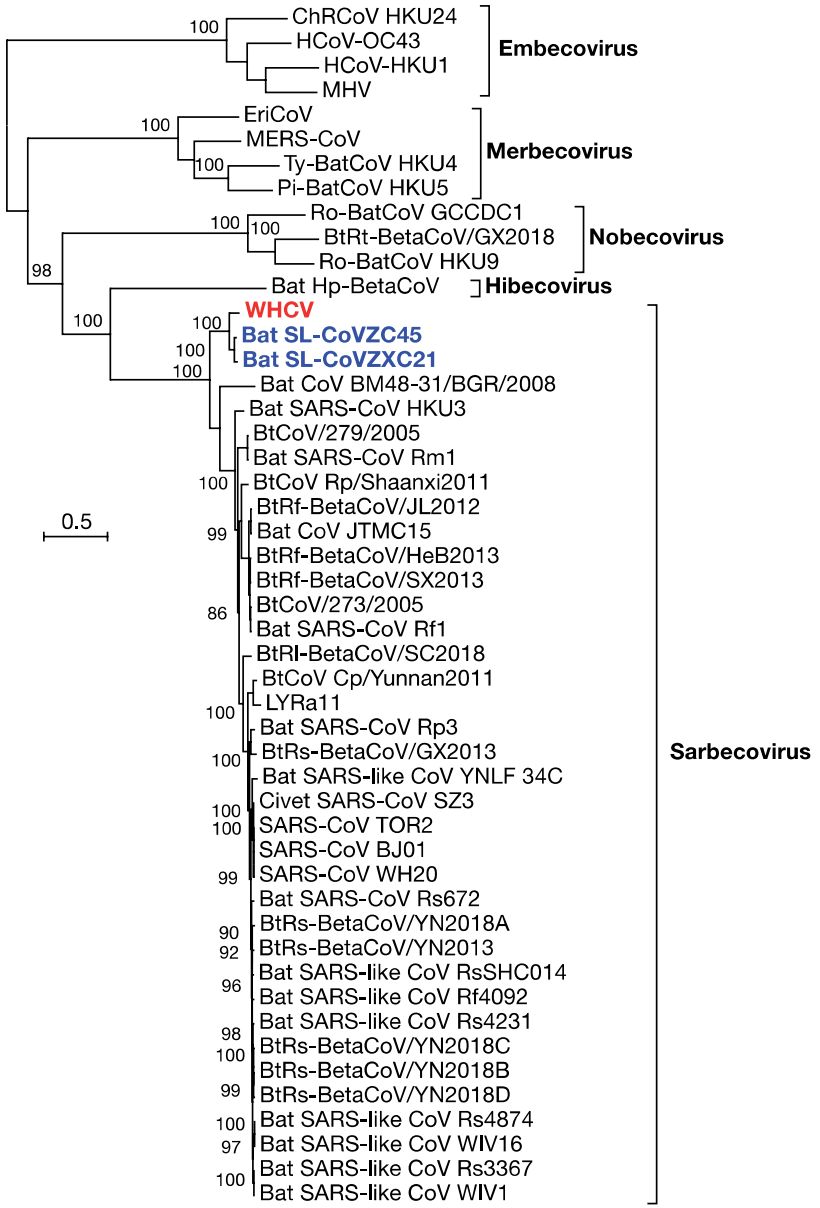
- Worker at the Wuhan market, hospital admission: 26 Dec 2019
- Clinical: severe respiratory syndrome (fever, dizziness, cough)
- Untargeted RNA sequencing of bronchoalveolar lavage fluid
- **Discovery of a novel pandemic disease 2019-nCoV (later COVID-19) caused by a virus WH-Human 1 (later SARS-CoV-2)**
- Genome shared online 10 Jan 2020 => NAAT tests 2 days later
- Phylogenetic analysis: novel virus most closely related (89.1% nucleotide similarity) to a group of SARS-like viruses previously been found in bats in China.

Genome organization of SARS-CoV-2 (29,903 nucleotides)



Chest radiograph at day 5

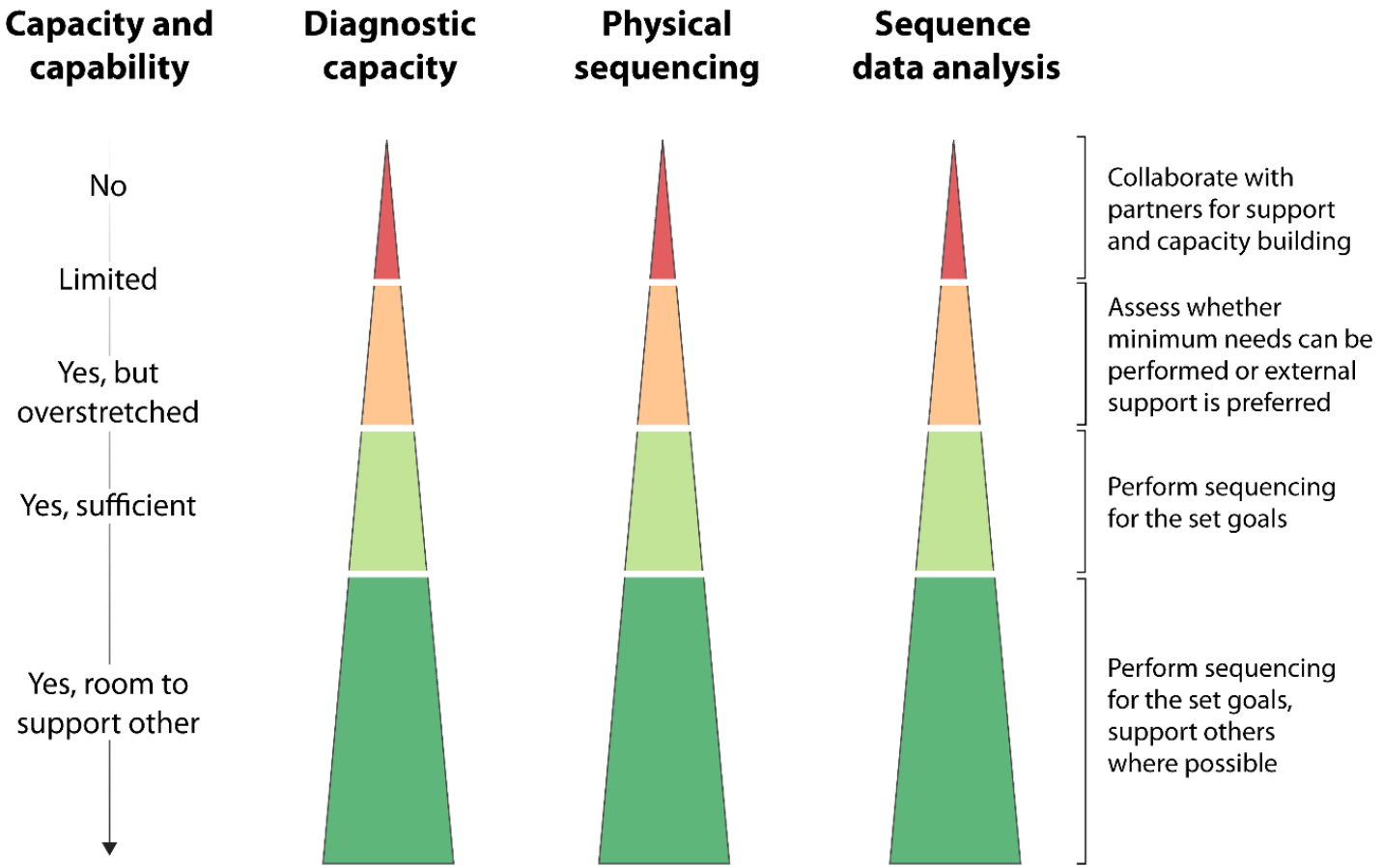


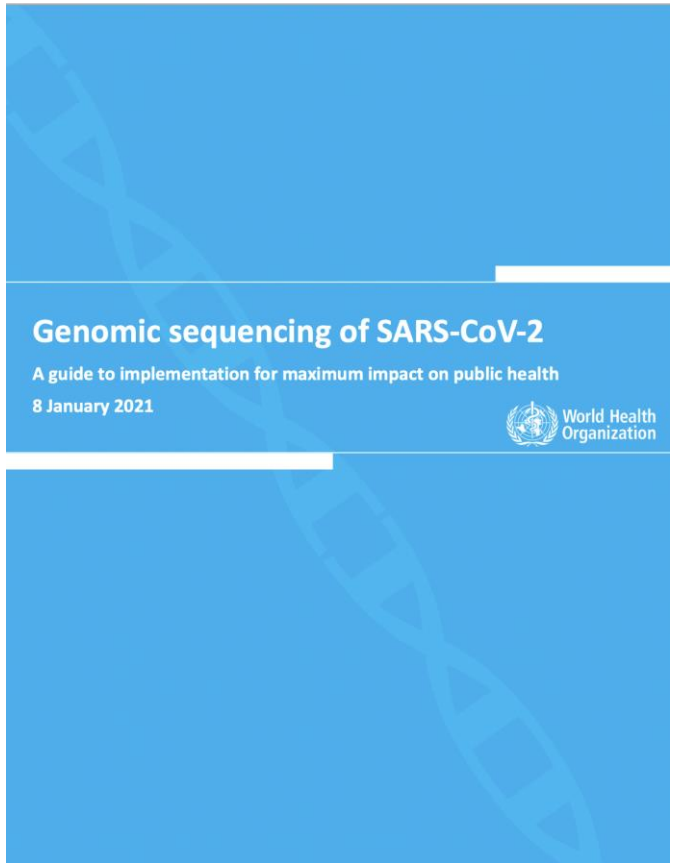


Wu et al. *Nature* 03 Feb 2020

Pathogen genome sequencing implementation

Establishing capacity in the three pillars required for pathogen genome sequencing





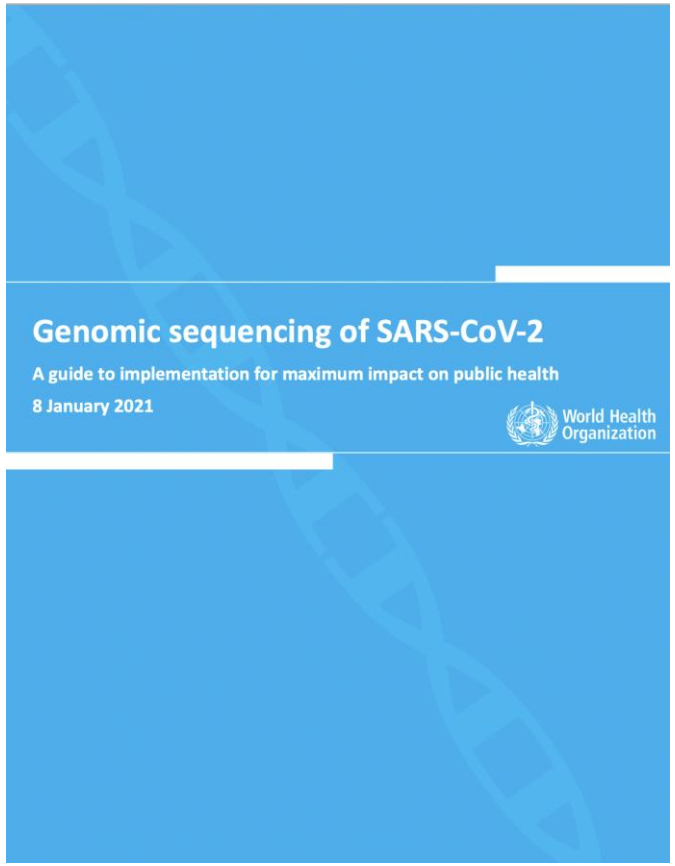
Genomic sequencing of SARS-CoV-2, *Geneva WHO* 2021

Genomic sequencing for maximizing public health impact

Key considerations before starting a sequencing programme:

- **Clear understanding of the objectives and expectations of sequencing**
- **Clear strategy for data analysis, sharing, storage and data integration**
- **Plan for how findings will be used to inform public health responses**

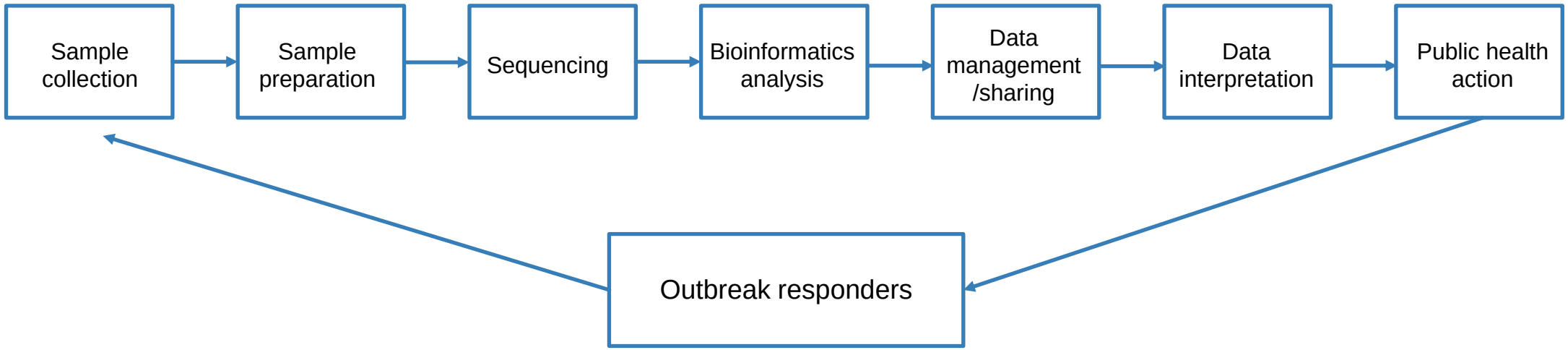
- Decide sequencing goals within multidisciplinary framework that includes representatives of all stakeholders
- Identify **funding sources to ensure sustainable** support
    - Costs of specialist personnel
    - Sequencing devices and consumables
    - Computational architecture to process and store data.
- Evaluate carefully **ethical aspects** of the project
- Conduct laboratory **biosafety** and **biosecurity risk** assessments for every step in chosen protocol.
- Focus on stepwise capacity-building programmes to build up competencies.
- **Communicate results** in a timely and clear manner to stakeholders who can use the information directly for public health benefit.



Genomic sequencing of SARS-CoV-2, *Geneva WHO* 2021

Pathogen genome surveillance workflow

Successfully implementing of genomic surveillance workstreams involve **bilateral communication between experts** involved at different stages; for example, those conducting data interpretation should discuss which samples will be chosen for sequencing with those involved in sample selection and preparation.



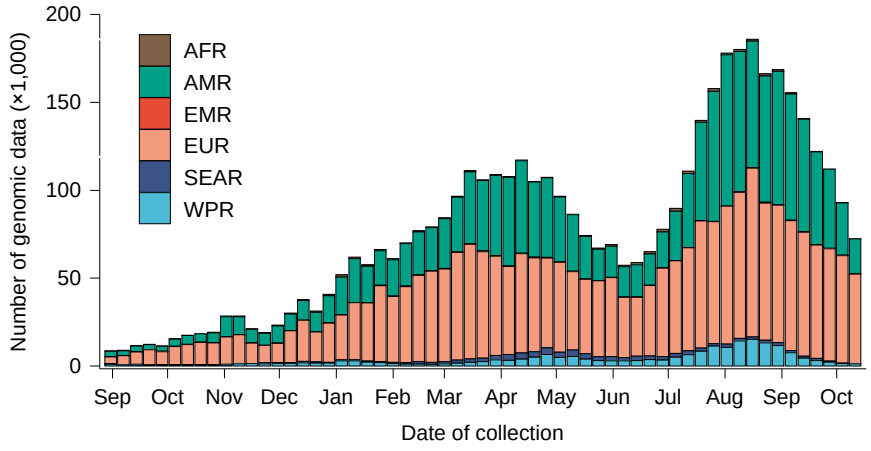
COVID-19 framework can be adapted for any other virus.

Genomic sequencing of SARS-CoV-2, *Geneva WHO* 2021

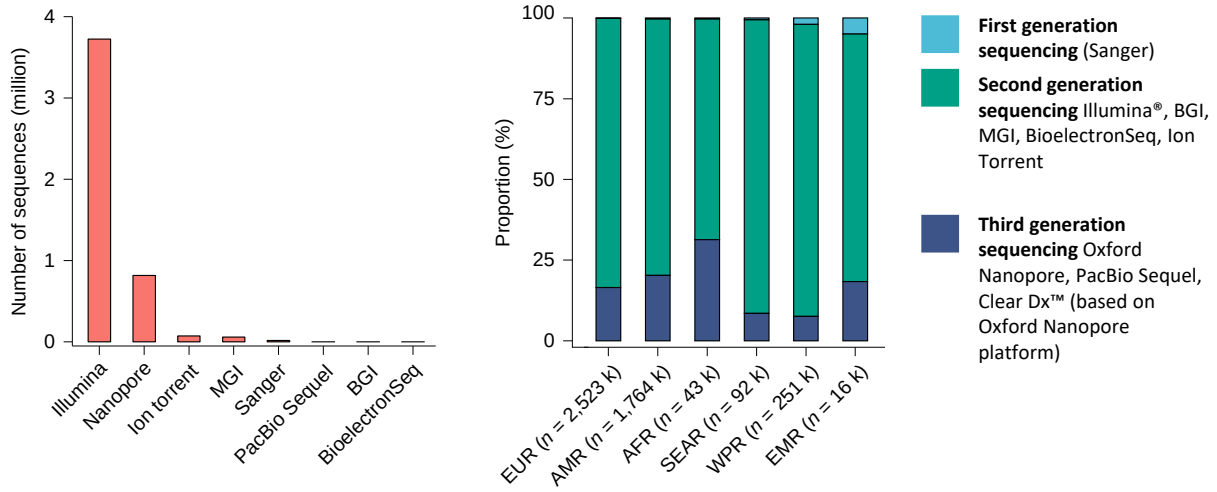
Sequencing technologies during the COVID-19 pandemic

Over 15.5 million SARS-CoV-2 sequences have been deposited in GISAID from 216 countries – can we leverage existing genomic sequencing expertise for arbovirus genomic surveillance?

Weekly numbers of publicly deposited SARS-CoV-2 genomic data by region



Sequencing counts per sequencing platform used globally, by income group and WHO region (n = 4.69 million)



Heterogeneity of sequencing percentage, sequencing technologies, turnaround time and completeness of released metadata across regions and income groups – representativity & standardized protocols needed.

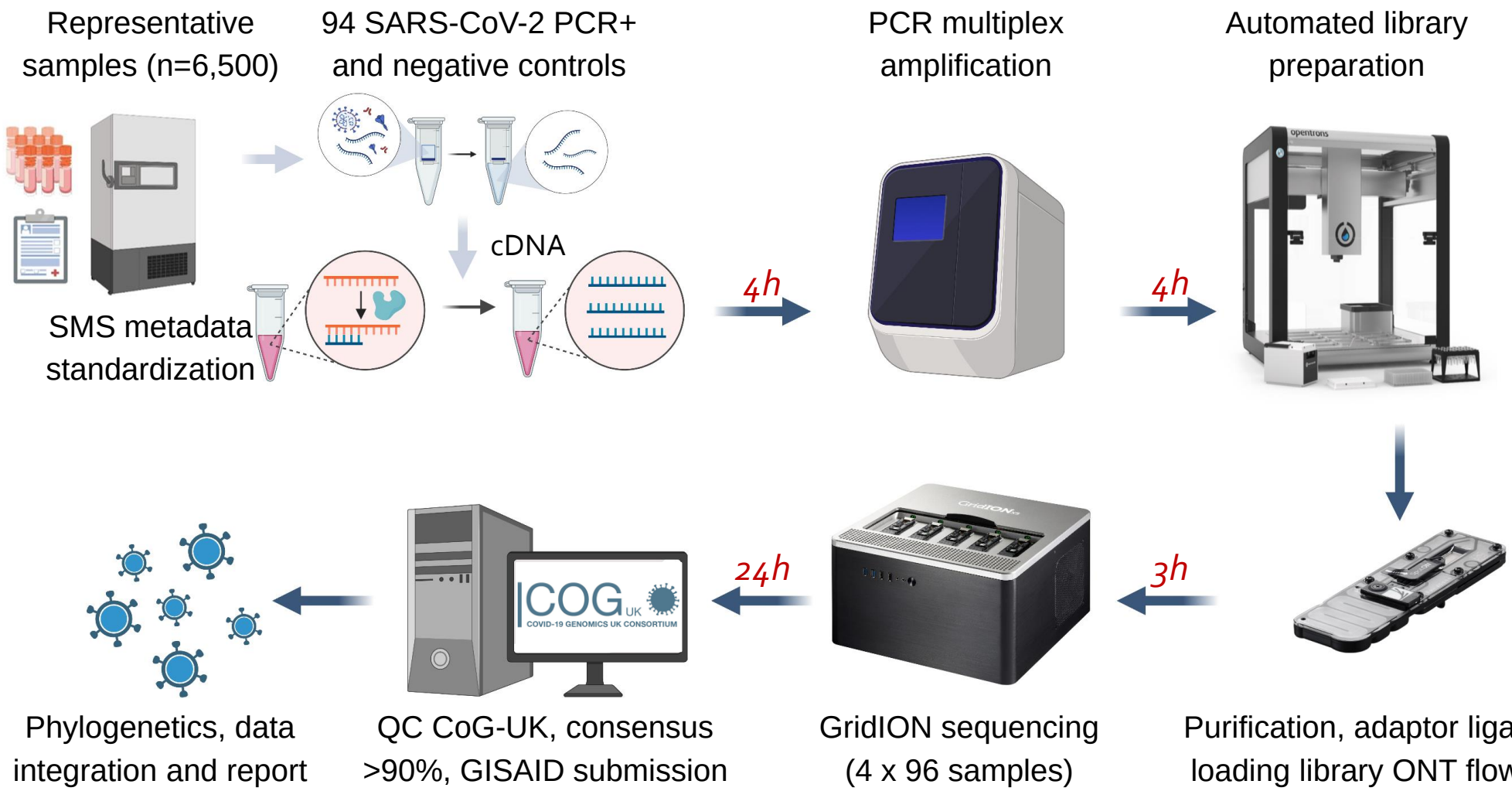
Chen et al. *Nature Genetics* 2021

Pathogen genome sequencing platforms

Instrument	Advantages	Limitations	Instrument run-time	Sequencing throughput	Relative availability and cost
Sanger sequencing	Widely accessible. Easy to use. Cost-effective sequencing if few targets required.	Very low throughput. Amplicons (often no more than 1000 bp). must be individually amplified and sequenced. Expensive for full genomes. Inappropriate for metagenomics.	Typically a few hours.	100 kB-2 Mb per single run.	Widely available. Relatively low cost for a few targets.
Illumina (e.g. iSeq, MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq)	Very high sequencing yields possible; very high accuracy. iSeq is portable; methods for handling data are well established.	With the exception of Illumina iSeq, expensive to purchase and maintain compared with some other platforms. Maximum read length 2 x 300 bp.	10–55 h, depending on the instrument.	1.2–6000 Gb, depending on instrument.	High maintenance and start-up costs. Moderate running costs
Ion Torrent	Fast turnaround once sequencing starts.	Challenges with homopolymers. Expensive to purchase. Maximum typical read lengths ~400 bp.	2 h–1 day, depending on chip and device.	30 Mb–50Gb depending on device and chips.	Moderate costs.
Oxford Nanopore Technologies (Flongle, MinION, GridION, PromethION)	Portable, direct sequencing real-time data; low start-up and maintenance costs; can stop sequencing as soon as sufficient data are achieved; very long read lengths achievable (exceeding the full length of the SARS- CoV-2 genome).	Challenges with homopolymers. Error rate per read is ~5% (R9.4 flowcells) so use of appropriate pipelines is critical to obtain high- accuracy consensus sequences. Latest R10 flowcells have now lower error rates. Currently unsuitable for determining intra-host variation unless replicate sequencing is used.	Reads available immediately. Can be monitored and run for up to several days as required.	Ranging from < 2 Gb for Flongle flow cell to 220 Gb for PromethION flow cell Up to 48 flow cells can be used on PromethION	No maintenance and low startup costs. Moderate running costs.

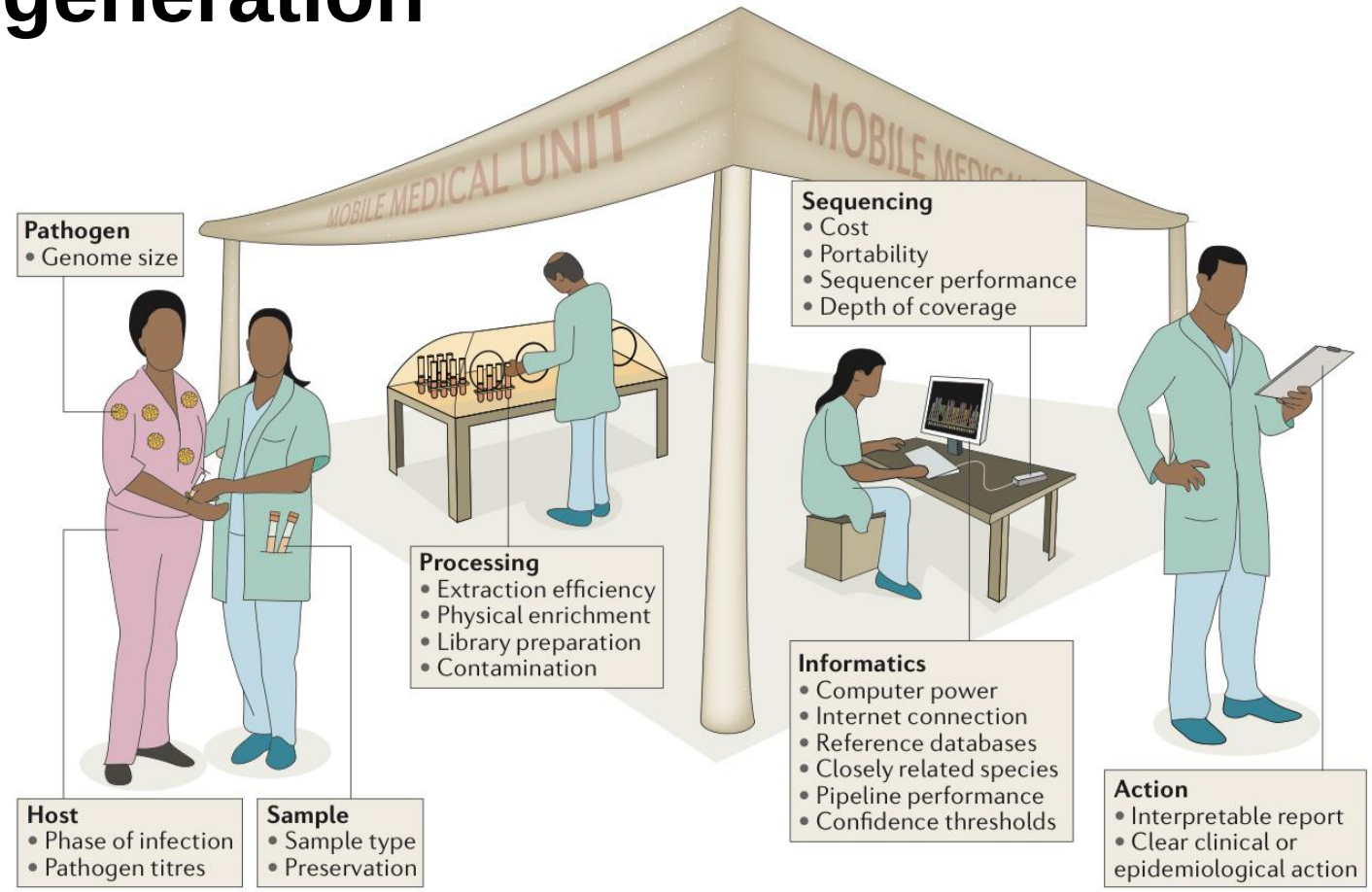
Genomic sequencing of SARS-CoV-2, *Geneva WHO* 2021

Sample processing and sequencing – ONT example

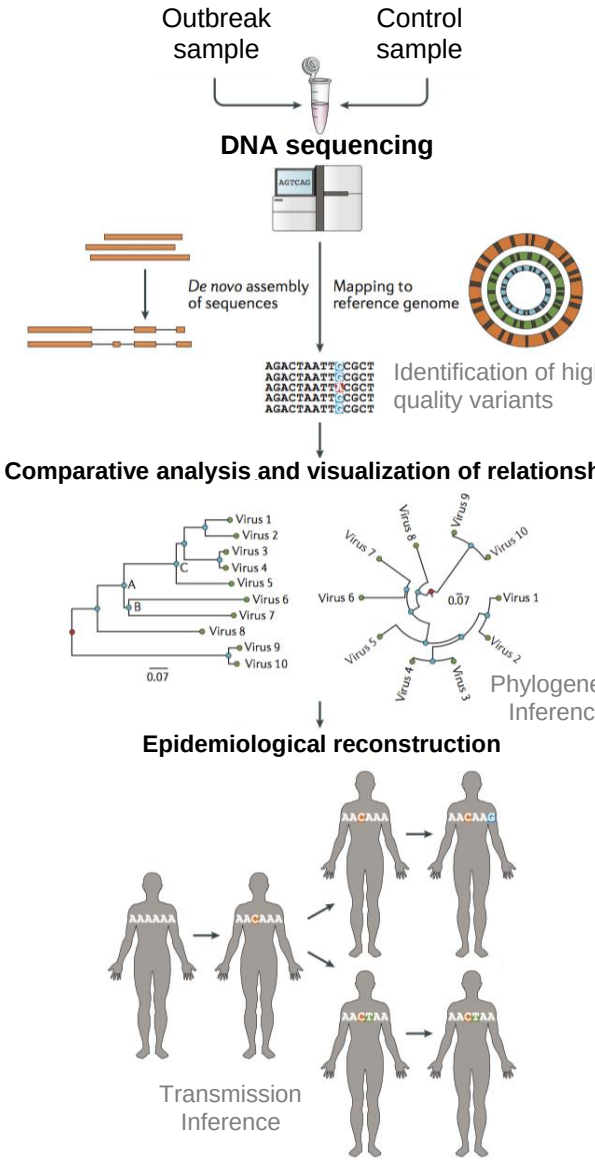


BMGF COSERGE Project (Nuno Faria, Ester Sabino)

Challenges for on-site pathogen whole genome sequence data generation



Short-lived viremia, RNA preservation (cold chain), collection *vs* onset symptoms; metadata availability, patient clinical outcome.



Unbiased characterization of the pathogen (rapid & low-cost; protocol updates; standardization)

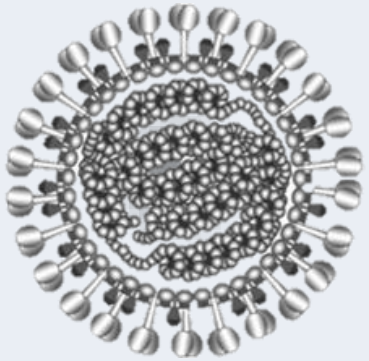
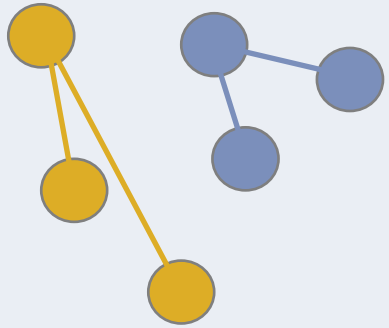
Phylogenetic inference and epidemiological reconstruction (bioinformatic skills)

Transmission inference and Interpretation (capacity building/ strengthening and training)

Adapted from Gardy and Loman, *Nature Rev Genetics*, 2018

Sequencing strategies

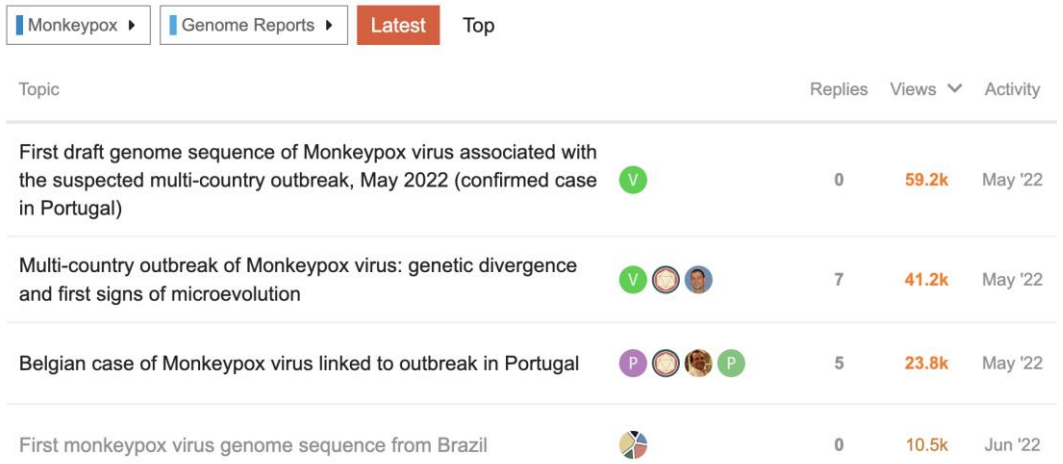
Sequencing strategies for SARS-CoV-2

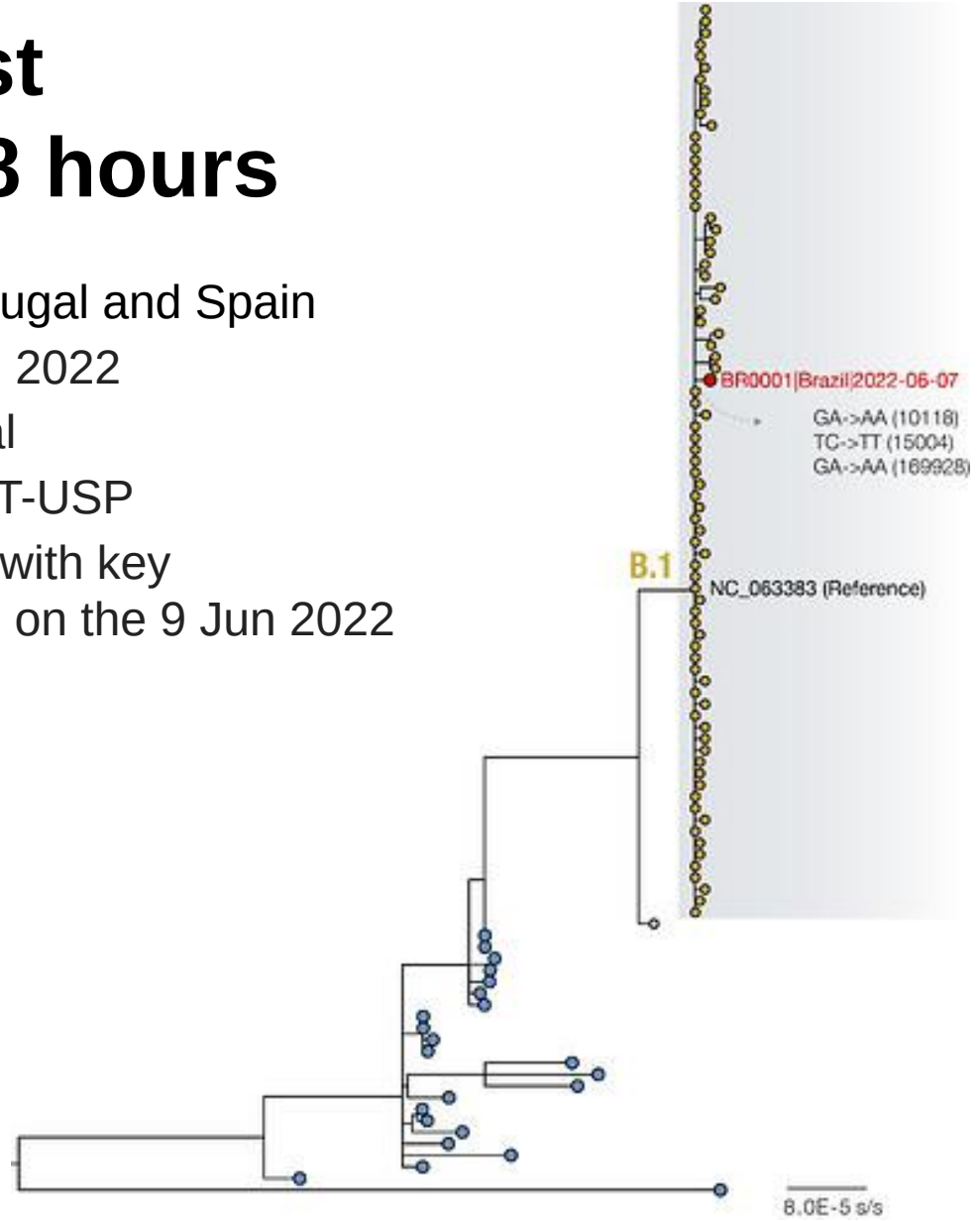
	Untargeted sequencing	Targeted sequencing		Population sequencing
Goal		GACAACCAAGTAGG TGCAACCAAICAGG CGCTACCAAICAGG		
Goal	Pathogen identification, understand origins of outbreak	Design of diagnostics, therapeutics, phenotypic changes (e.g., Alpha, Omicron)	Investigate transmission clusters & complement epi approaches (e.g., outbreaks in hospitals, travelers)	Large-scale patterns, drivers and assess detection (e.g., R, pop size, dispersal patterns)
Requires	Untargeted sequencing (metagenomics), human (H) and non-H samples	Targeted sequencing of S-gene dropouts, extensive experimental work	Non-random, dense cluster & community sequencing	Random representative population sequencing

Genomic sequencing of SARS-CoV-2, *Geneva WHO* 2021

Untargeted sequencing of the first monkeypox case in Brazil took 18 hours

- 41-yo male from São Paulo with recent travel history to Portugal and Spain
- Skin swab of the lesions (vesicle & crust) collected on 7 Jun 2022
- Viral DNA was isolated from 200 µl of the recovered material
- Sequencing and bioinformatics completed in 18 hours at IMT-USP
- Complete genome, raw sequence data shared immediately with key stakeholders; and with scientific community (*Virological.org*) on the 9 Jun 2022
- Results confirmed 24 hours later by reference laboratory

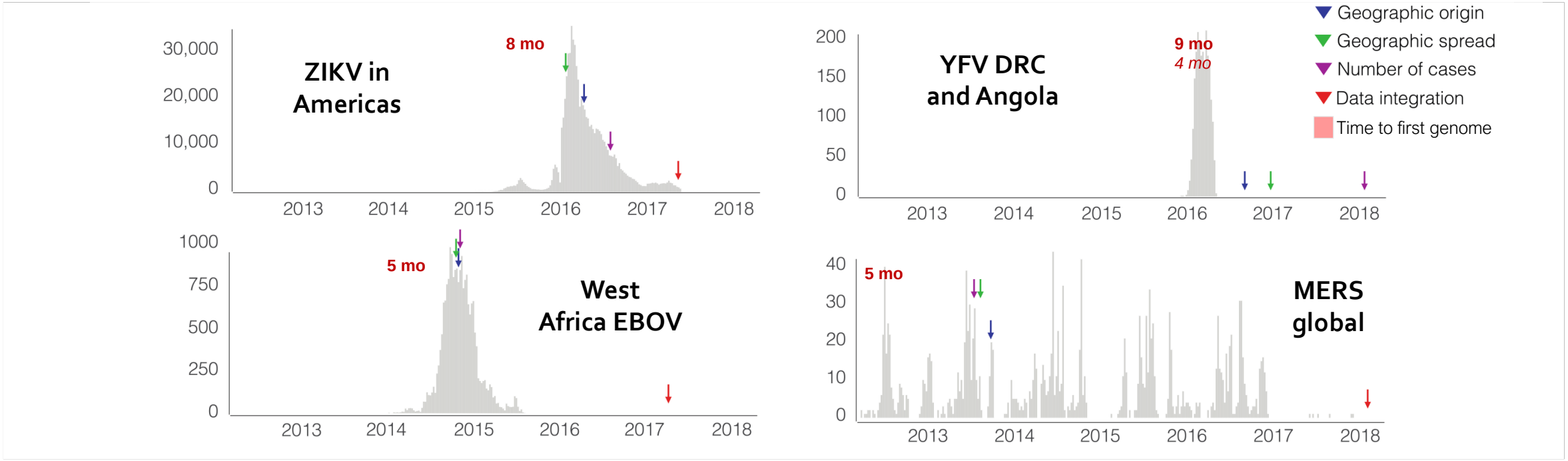




Claro et al. RIMT 2022; https://virological.org/t/first-monkeypox-virus-genome-sequence-from-brazil/850

Untargeted sequencing to reduce *genomic* surveillance gaps for emerging viral pathogens?

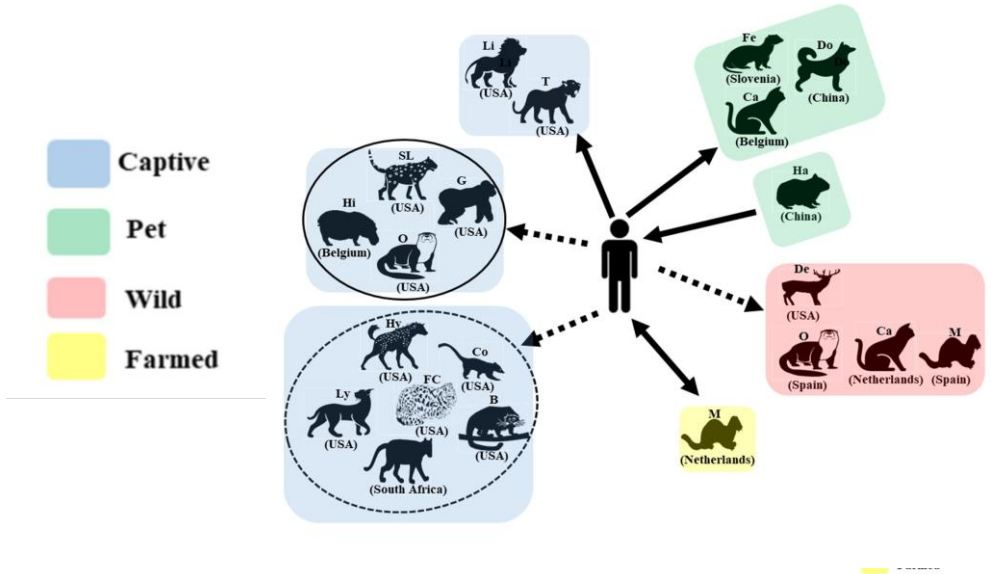
- **2019-nCoV –** 15 days between first cases and GSD: decreasing time-lag between outbreak detection and GSD release should be encouraged both for PHEIC and non-PHEIC
- **mpox in Brazil** – Untargeted sequencing decreased surveillance gap and allowed timely deployment of control measures

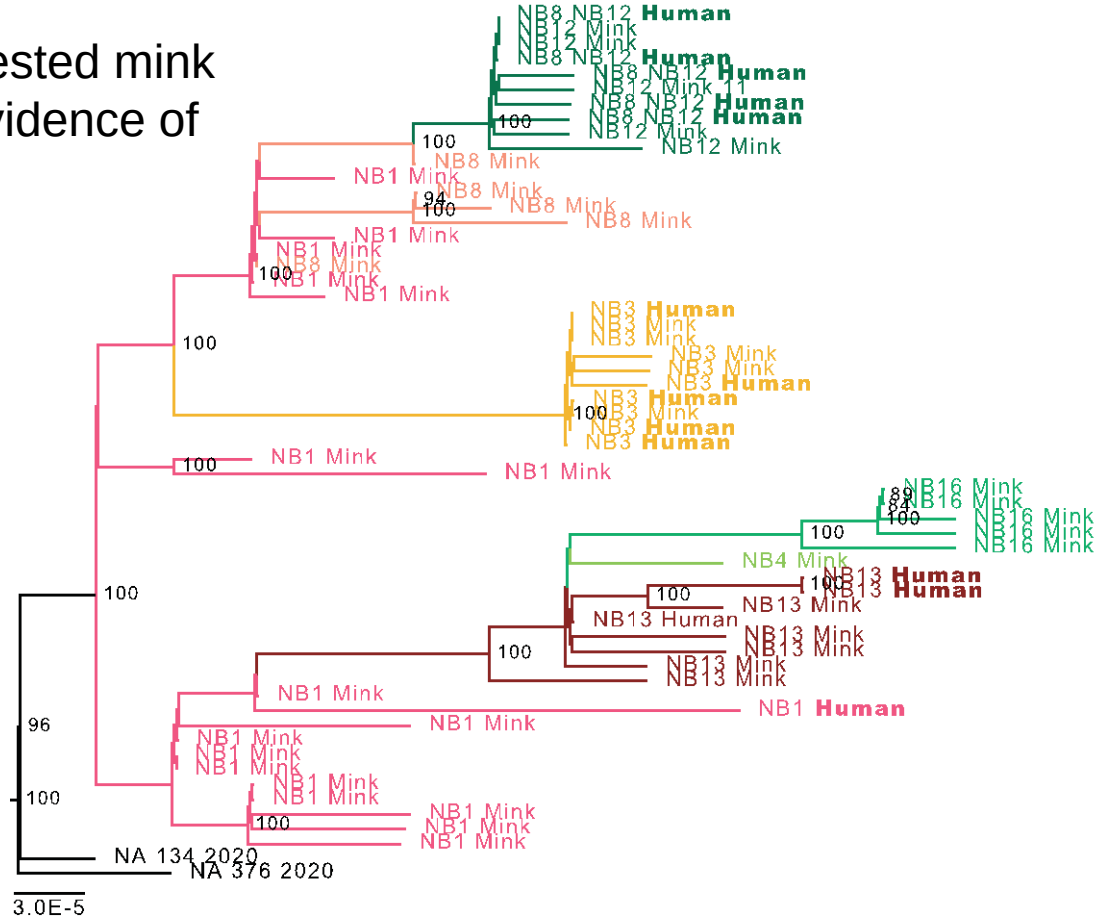


Kraemer et al. *Epid and Infection* 2019

Targeted sequencing of SARS-CoV-2 in animals

- Outbreaks investigation on 16 mink farms & humans living or working on these farms in The Netherlands.
- **Spillover followed by spillback of the virus**: 68% of tested mink farm residents, employees, and/or their contacts, had evidence of SARS-CoV-2 infection with animal sequence signature
- **SARS-CoV-2 has since been identified in 18 animal species (pet, captive, farmed, and wild animals)**





Oude Munnick et al. *Science* 2021; Cui et al. *Viruses*, 2022

Surveillance networks

Global Surveillance Networks from 1994 onwards

• **Program for Monitoring Emerging Diseases (ProMED) (1994)**: "largest publicly-available system conducting global reporting of infectious disease outbreaks" – it is a free, nonprofit, noncommercial, moderated e-mail list that today serves over 37,000 subscribers in more than 150 countries.

• **Global Public Health Intelligence Network (GPHIN) (1997, in collaboration with WHO)**: event-based multilingual early-warning and situational awareness network for potential chemical, biological, radiological and nuclear (CBRN) public health threats worldwide.

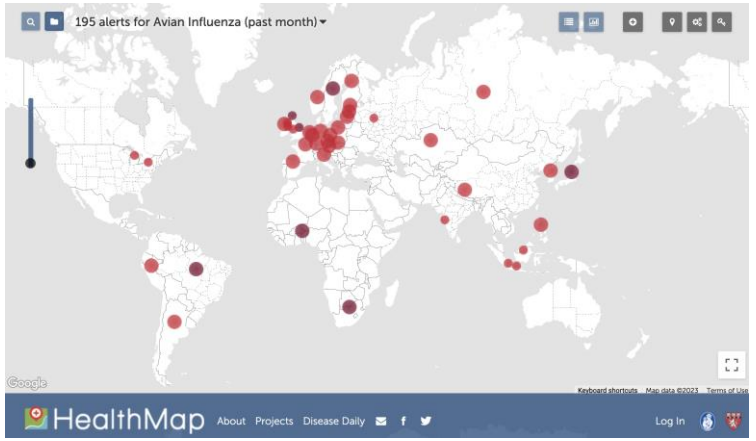
• **GOARN (2000)**: WHO's operational network through which human and technical resources from over 600 existing institutions and networks in global epidemic surveillance are pooled "combating the international spread of outbreaks; ensuring that appropriate technical assistance reaches affected states rapidly and contributing to long-term epidemic preparedness and capacity building." GOARN has responded to over 120 occurrences in 85 countries and has deployed over 2,300 experts into the field.

ISID ProMED INTERNATIONAL SOCIETY FOR INFECTIOUS DISEASES

GPHIN

Global Outbreak Alert and Response Network

Global Surveillance Networks from 2020 onwards (COVID-19)



• **HealthMap (2006)**: freely available internet-based emerging infectious disease intelligence, automated network that collects information from multiple web-based data sources on infectious outbreaks (currently news wires, Really Simple Syndication (RSS) feeds, ProMED, EuroSurveillance and WHO alerts).

• **Our World In Data (OWID) (2011)**: interactive large dataset with a variety of metrics: confirmed cases, deaths, intensive care unit (ICU) admissions, epidemic reproduction rates and vaccinations, among others.

Our World in Data

• **Johns Hopkins University COVID-19 Dashboard (2020):** Updated source of COVID-19 data and expert guidance. Data available on cases, deaths, tests, hospitalizations, and vaccines to help respond to the pandemic.

JOHNS HOPKINS UNIVERSITY & MEDICINE | CORONAVIRUS RESOURCE CENTER

• **Global.Health (2020):** open access standardised line list of >100 million individual-level anonymised cases based on open sources and voluntary submissions from countries. Allows the study of transmissibility, routes of transmission and risk factors for infection, as well as to inform planning of response and containment efforts.

GLOBAL .hEALTH | a Data Science Initiative

Global Pathogen Genome Sequence Data Databases

There are two key pathogen genome sequence data (GSD) global databases:

INSDC (International Nucleotide Sequence Database Collaboration):

Access is free and unrestricted for reproducibility and reuse:

- NCBI GenBank
- ENA-EBI (European Nucleotide Archive)
- DDBJ (DNA Data Bank of Japan)

GISAID (Global Initiative on Sharing All Influenza Data):

Access is free but restricted to users and data agreements:

- SARS-CoV-2, mpox, RSV only
- EpiArbo (to be released?)

Recent controversies with GISAID linked to **transparency** and **sustainability**. Guidelines for virus GSD code conduct for data generators and data users missing.

WHO Principles of pathogen GSD sharing*
1. Capacity development. Preference for local analysis for local decision-making
2. Collaboration and cooperation. Promote collaboration and cooperation between data generators/analysers
3. High-quality, reproducible data. High quality of the data and supporting metadata
4. Global and regional representativeness. Critical to identify the emergence of new infectious threats
5. Timeliness. Timely generation and sharing of pathogen genome data
6. Acknowledgement and intellectual credit
7. Equitable access to health technologies as a benefit
8. As open as possible and as closed as necessary
9. Interoperability and relevance for national, regional and global decision makers
10. Trustworthiness and ease of use
11. Transparency
12. Consistency with applicable law and ethical regulations
13. Compliance and enforcement

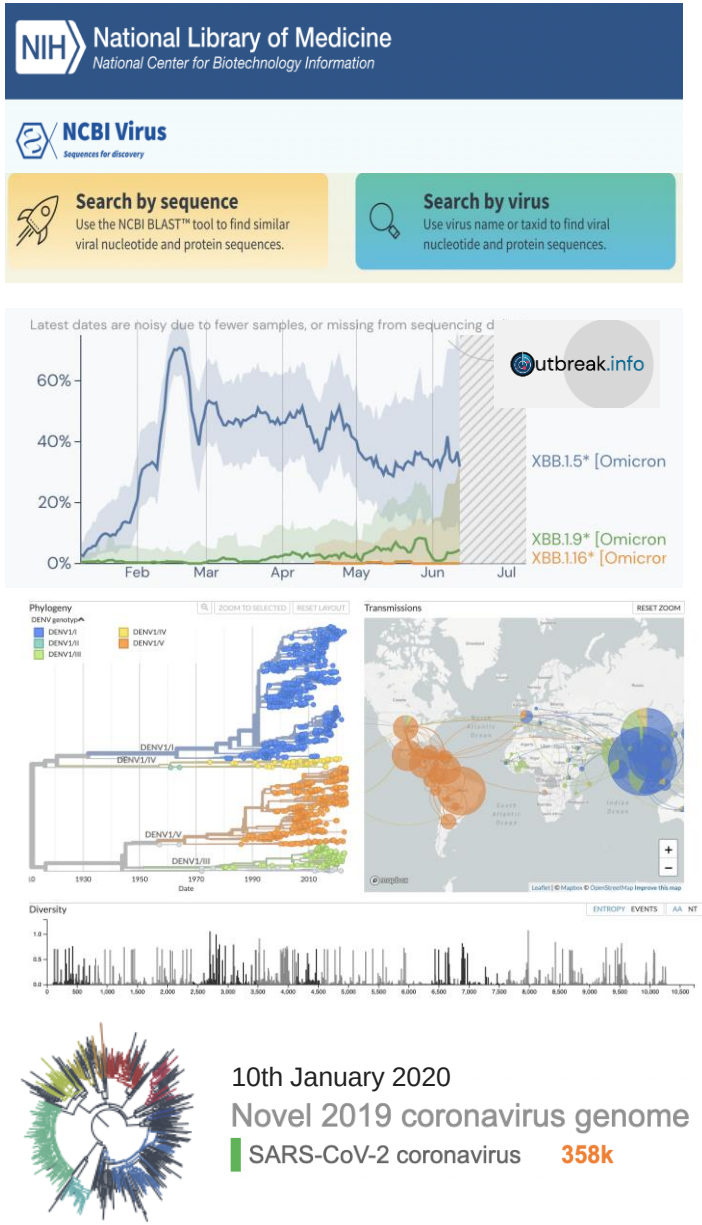
Global Pathogen Genome Sequence Data (GSD) Resources

- **NCBI GenBank Virus Variation (2017):** a community portal to find, retrieve and analyse any viral sequence data from RefSeq, GenBank and other NCBI repositories (12,025,335 nucleotides).

- **Outbreak.Info (2020):** aggregates genomic and epidemiological data across scientific sources to monitor SARS-CoV-2 variants; integrate publications, preprints, clinical trials, datasets, protocols, and other resources into one searchable library of COVID-19 research; track trends in COVID-19 cases and deaths.

- **NextStrain (2015)**: collection of open-source tools for visualising the genetics behind the spread of viral outbreaks. Aim is to support public health measures and surveillance by facilitating understanding of the spread and evolution of pathogens.

- **Virological.org**: Discussion forum for analysis and interpretation of virus molecular evolution and epidemiology. Includes latest reports on outbreak genomics, where first sequences and analyses from outbreaks are often shared (including SARS-CoV-2, Zika, Ebola).

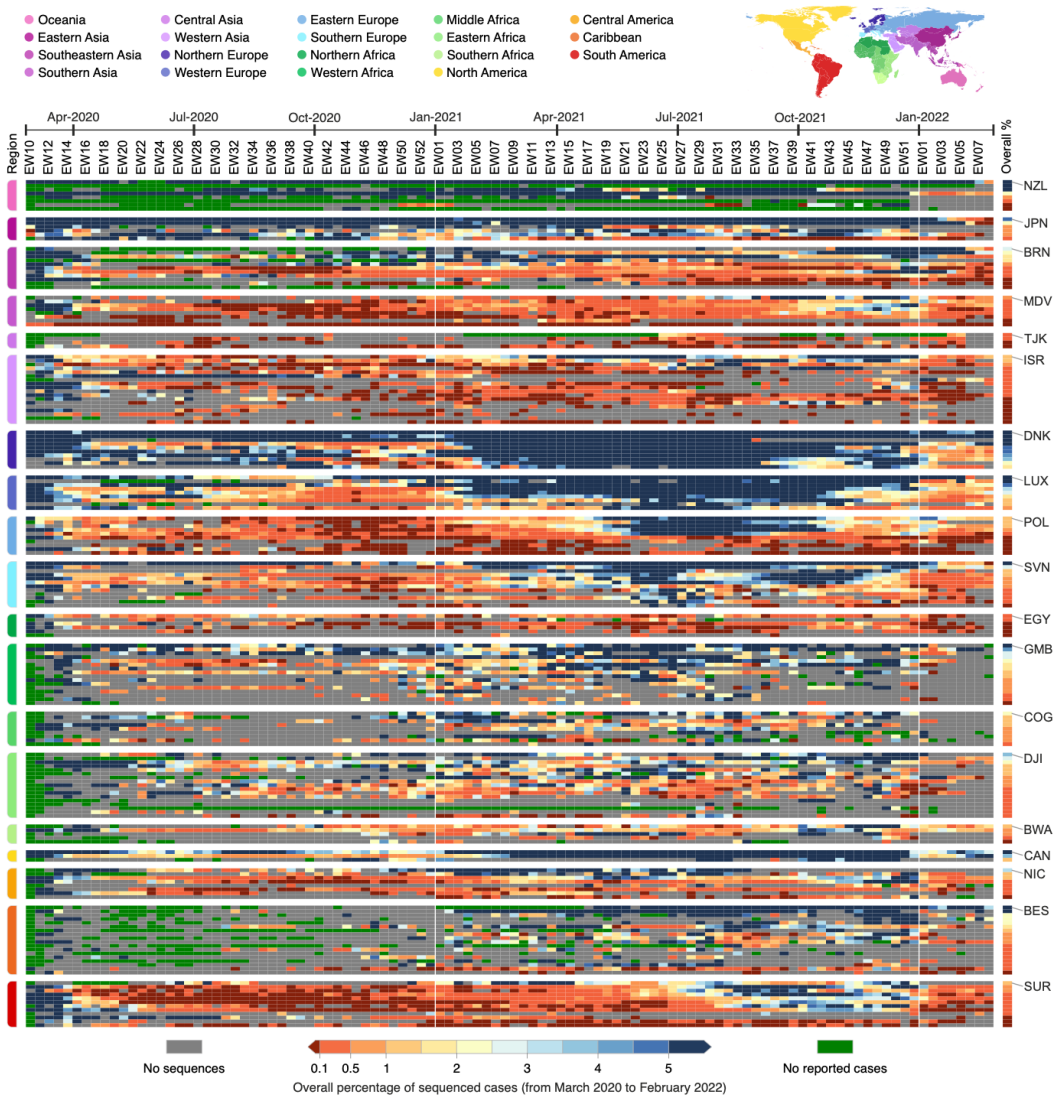


Disparities in global sequencing capacity

Global disparities in population sequencing of SARS-CoV-2: implications for Disease X

Global disparities can hamper our understanding of the epidemiological properties of novel viral variants and delay detection of "Disease X".

- During the first two years of the SARS-CoV-2 pandemic (until March 2022):
  - **Sequenced >0.5% of their COVID-19 cases:**
    - 78% of high-income countries
    - 42% of low- and middle-income countries
  - **Submitted genome data within 21 days:**
    - 25% of high-income countries
    - 5% of low- and middle-income countries
- Genome sequence dataset representativity is hampered by low **sequencing volume** and heterogeneous sampling **strategies** (e.g., random vs. targeted to travellers).

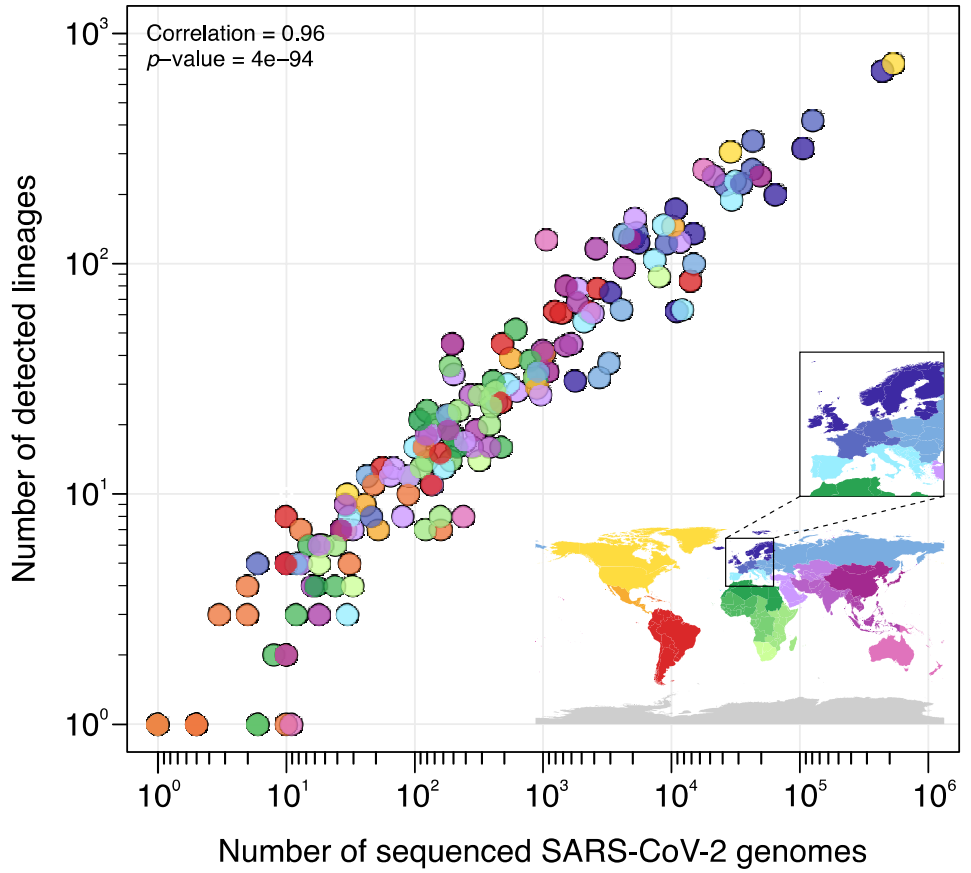


Brito (…) Faria, *Nature Comms* 2022

Impact of SARS-C0V-2 sequencing intensity in pandemic preparedness

Limited genome sequencing capacity affects identification and response to VOCs/VOIs and other viral agents.

- Number of globally observed lineages correlates with number SARS-CoV-2 genomes available per country and overall proportion of sequenced cases.
- Low **sequencing proportion** and **delays:**
  - Disparities in testing
  - Disparities in sequencing capacity
  - Lack of clear sampling strategy
  - Lack of standardization in wet lab and bioinformatic pipelines
  - Lack of virus genome integration with appropriate metadata.

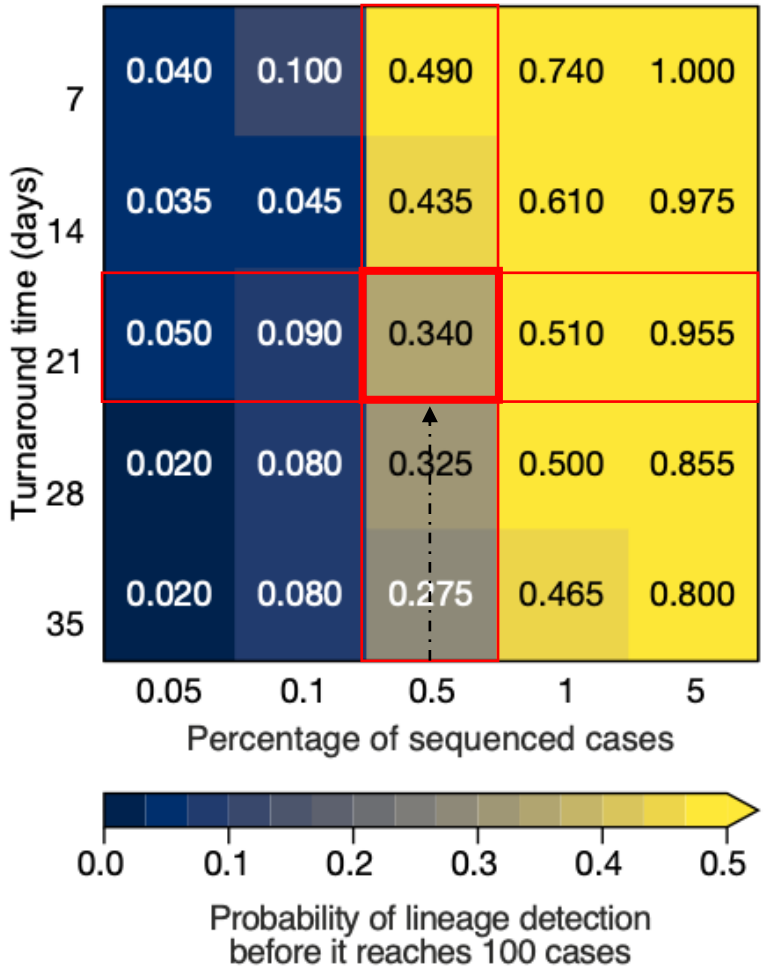


Brito (…) Faria, *Nature Comms* 2022

Detection of viral lineages under different genomic surveillance scenarios

Effective response to emerging variants requires timely and accurate characterisation of epidemiological properties.

- **Simulation study:** Turnaround time (TAT)=21 days and sequencing 0.5% of cases => **34% probability that a lineage is detected before it reaches 100 cases.**
  - **Manaus city** (2.2m inhabitants): sequencing 0.5% cases = 11 randomly selected genomes / week
  - **São Paulo city** (12.4m inhabitants): sequencing 0.5% cases = 62 randomly selected genomes / week
  - **Brazil** (212.6m inhabitants): sequencing 0.5% cases. = 1,063 randomly selected genomes / week



Brito (…) Faria, *Nature Comms* 2022

Summary and Conclusion

Summary of genomic surveillance applications

1. Identification and characterization of lineages and development of countermeasures

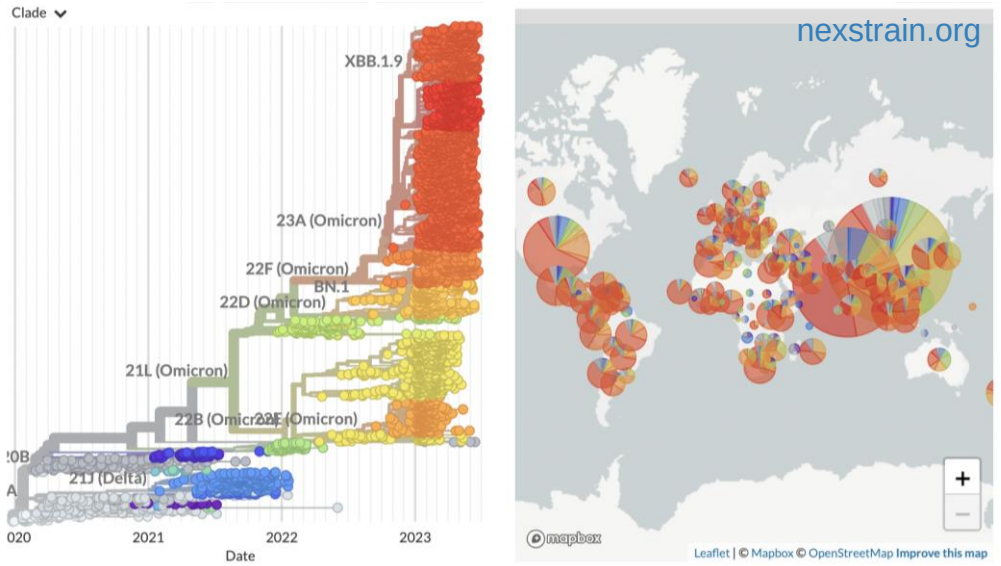
- Identifying the causative agent of an outbreak
- Origins and transmission of novel lineages
- Rapid development of diagnostics and vaccines
- Identification of the initial animal source and/or potential intermediate hosts

2. Monitoring transmission and geographic spread

- Identifying drivers of transmission may help to shape new strategies for preventing spread
- Establish the contribution of local transmission compared to imported cases to help make policy decisions
- Support cluster and outbreak investigations
- Estimate the upper and lower limits of time of circulation
- Improved diagnostic surveillance programmes
- Support assessment of relative changes in outbreak size
- Environmental surveillance as an “early warning” system?
- Investigate reinfections and vaccine reversions

3. Monitoring SARS-CoV-2 evolution

- Clinical genomic studies to investigate transmissibility or virulence of specific mutations
- Monitoring genomic changes that might reduce vaccine efficacy, control strategies, drug resistance
- Investigate spillover and spillback in animals



Genomic sequencing of SARS-CoV-2, *Geneva WHO* 2021

Concluding remarks

- Implementation of genomic sequencing for surveillance needs to consider **how findings will be used to inform public health responses** and requires multidisciplinary effort: diagnostic capacity, physical capacity for sequencing, and capacity for data analysis and interpretation.
- Maximum usefulness of genome sequence data (GSD) requires high quality linked patient/host **associated metadata**: exact location and date of sampling, clinical symptoms, age, sex, occupation, contact/travel history, and host species (when applicable).
- Pathogen GSD sharing: WHO recommendation is to share all data for pathogens during PHEIC. **Acknowledgement and collaboration with data generators is crucial.**
- Modernised **human & non-human genomic surveillance data for "Disease X" and One Health** strategies could build upon the infrastructure and expertise gained during COVID-19.
- **Future steps for genomic surveillance: development and implementation of multi-host, multi-level pathogen agnostic sequencing strategies to improve detection of zoonotic Disease X.**

Thank you!

Questions?