

Introduction to genomic epidemiology

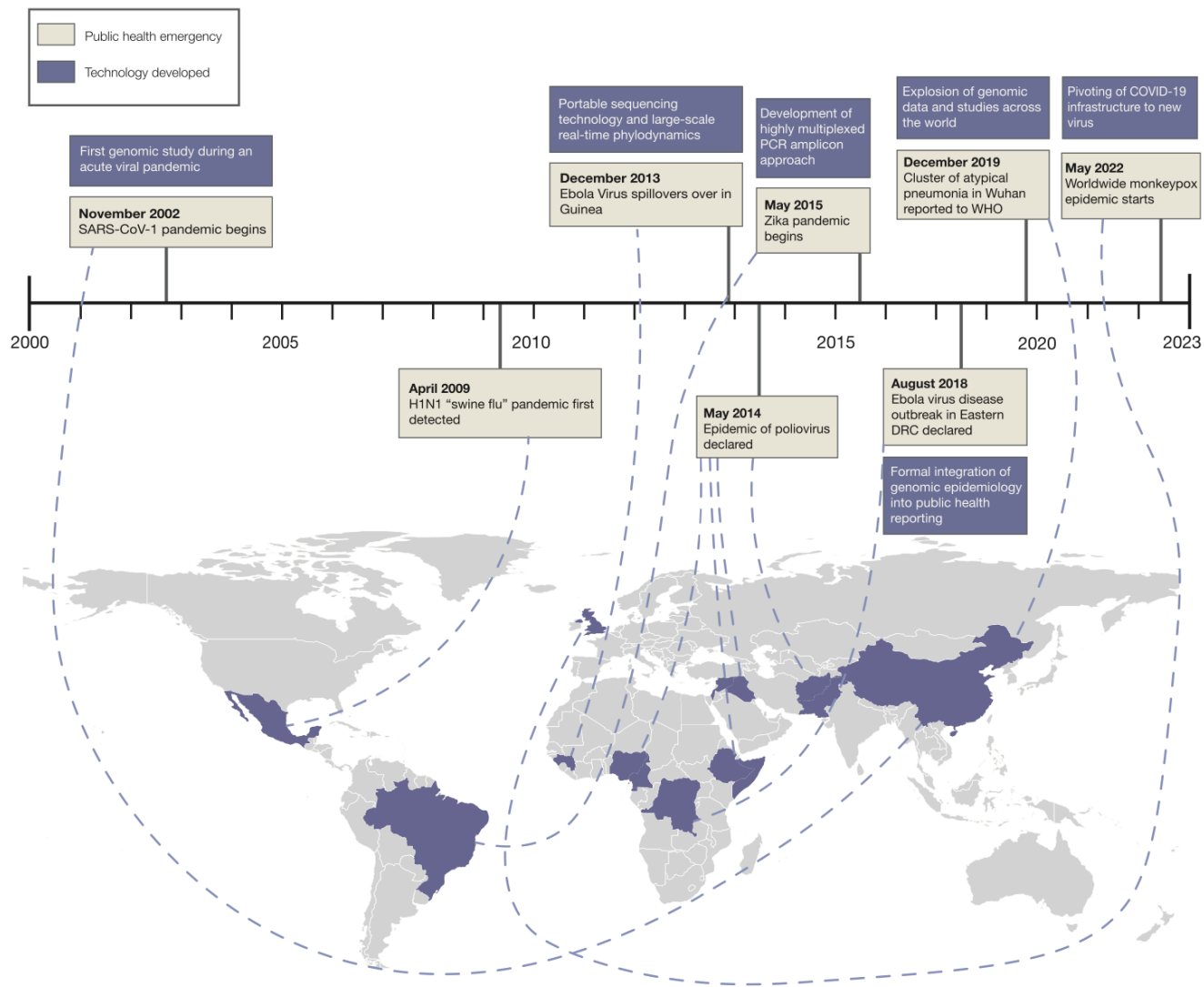
Dr Alicia Arnott

Deputy Head of Epidemiology, VIDRL

Introduction

- Whole genome sequencing (WGS) involves isolating and amplifying the entire genome of a pathogen
- Genomes obtained from different samples containing the same pathogen can be compared to answer questions
 - Are any of these samples novel/unexpected?
 - Were these infections likely acquired from the same source?
- Applying genomics to investigate pathogens is referred to as pathogen genomics
- The use of pathogen genomic data to determine the distribution and spread of an infectious disease in a specified population and the application of this information to control health problems is **genomic epidemiology**

Use of genomics for investigation of viral outbreaks over time



Pathogen genomics increasingly incorporated into public health response

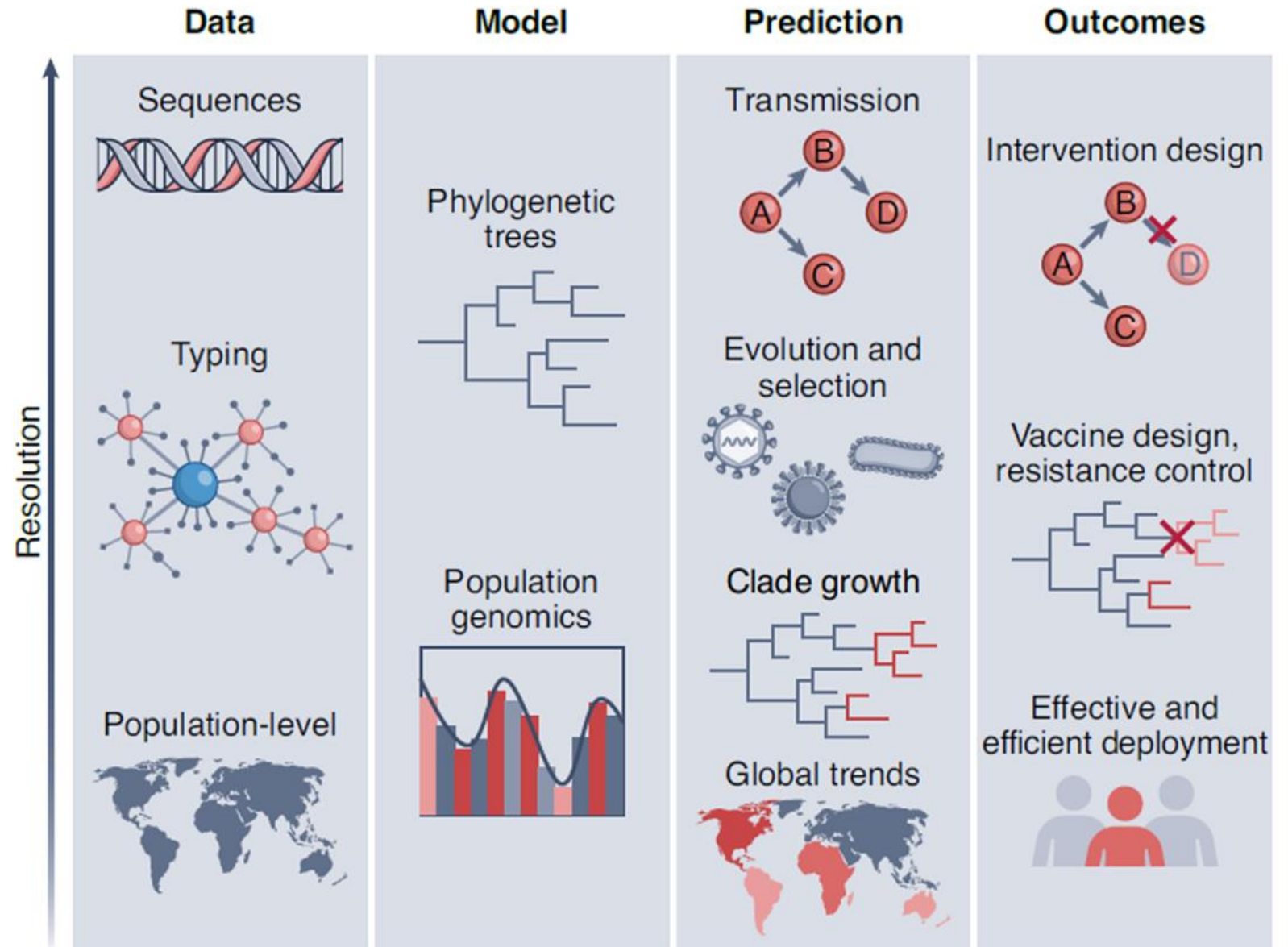


A joint venture between The University of Melbourne and The Royal Melbourne Hospital

The screenshot displays the European Centre for Disease Prevention and Control (ECDC) website. At the top, the ECDC logo and name are visible, along with a search bar labeled "NEW! Improved search". A navigation menu includes links for "Infectious disease topics", "Data", "Analysis and guidance", "Training and tools", and "About ECDC". The breadcrumb trail reads: "Home > About ECDC > Media centre > Training in genomic epidemiology and public health bioinformatics". The main heading is "Training in genomic epidemiology and public health bioinformatics", with a sub-label "E-learning course". Below this, it says "ECDC" with a user icon. A "Manage cookies" button is in the bottom right. On the left, a partial view of the WHO logo and the text "The" are visible. At the bottom, a dark blue banner contains the text "Better data. Better analytics. Better decisions."

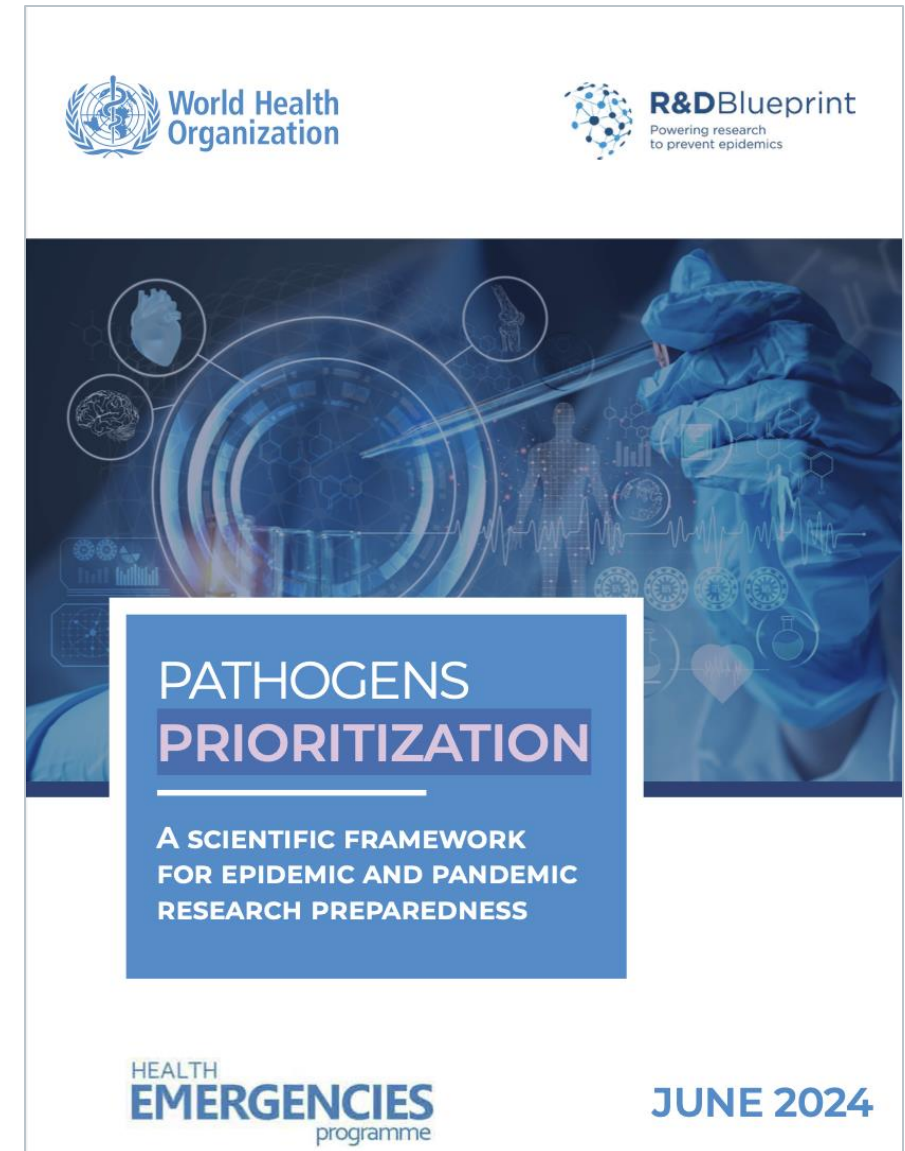
Introduction

- Genomics was being utilised for the public health management of infectious diseases prior to the pandemic
- Pandemic highlighted the public health utility of pathogen genomics, increasing global implementation efforts

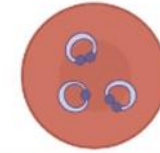


Post-pandemic pathogen prioritization

- The number of pathogens of public health concern is very large, while the resources for disease research and development is limited
- In 2022, the WHO employed a global panel of experts to identify which pathogen families/diseases pose the greatest public health risk due to their epidemic potential and/or whether there is no or insufficient countermeasures



Viruses pose greatest global pandemic threat



	RNA Viruses	DNA Viruses	Bacteria	Fungi	Protozoa ^a (Parasites)
Pandemic Potential	High	Moderate	Moderate, historically high	Low	Low
Features	Continual threat due to RNA genomes prone to mutation; fast viral reproduction and transmission; propensity to cause respiratory infections with airborne potential	Less prone to mutations than RNA viruses and transmission less likely to be airborne	Slower reproduction than viruses, less mutation prone, but able to transfer fitness genes; Threat now reduced due to fast detection and well-developed treatments	Nontraditional pandemic threat due to slow speed of spread and effective antifungals, though multidrug resistant organisms are increasing	Pandemic threat is low due to spread through vector, however malaria is estimated to contribute to half of all human deaths throughout history
Speed of Transmission	Fast	Fast-Medium	Medium	Slow	Slow-Medium
Mode of Transmission	Airborne Respiratory, Droplet, contact with infected fluids	Droplet, contact with infected fluids or materials (fabric)	Droplet (respiratory); Contact with infected humans or animals	Airborne spores; Contact with infected humans or animals; Environmental	Generally through an intermediate host (vector)
Detection Methods	PCR-based (hours) Select Antigens (hours-days)	PCR-based (hours) Antibodies (hours-days)	Bacterial culture (days), PCR-based (hours)	Fungal culture (days) Select antigen tests	Microscopy (hours) PCR-based (hours)
Countermeasures	Masks for respiratory transmission, antivirals, vaccines if available	Vaccines, standard precautions ^b	Vaccines, antibiotics, improved sanitation	Antifungals, control of environmental fungal threats	Barriers between hosts and humans, elimination of reservoirs, anti-parasitic medications
Examples of Pandemic Pathogens	SARS-CoV1 (SARS) SARS-CoV2 (COVID-19) Ebola (Hemorrhagic Fever) HIV (AIDS) Influenza (Spanish, Swine, Avian flus)	<i>Variola major</i> (Smallpox)	<i>Yersinia pestis</i> (Bubonic Plague) <i>Vibrio cholera</i> (cholera) <i>Mycobacterium tuberculosis</i> (TB)	No human pandemics caused by fungi; widespread fungal disease affecting bats, frogs; fungi posited to contribute to dinosaur die-off	<i>Plasmodium spp</i> (malaria) <i>Trypanosoma lewisi</i> (widespread death of the Christmas Island Rat)

Viruses pose greatest pandemic threat

- In 2022, the priority diseases were:

- COVID-19
- Crimean-Congo haemorrhagic fever
- Ebola virus disease and Marburg virus disease
- Lassa fever
- Middle East respiratory syndrome coronavirus (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS)
- Nipah and henipaviral diseases
- Rift Valley fever
- Zika
- “Disease X”*



	2017	2018	2024		
Family	Priority Pathogens	Priority Pathogens	PHEIC risk	Priority Pathogens	Prototype Pathogens
Adenoviridae			Low-Medium		Recombinant Mastadenovirus
Adenoviridae			Low-Medium		Mastadenovirus blackbeard serotype 14
Anelloviridae			Low		
Arenaviridae	Arenaviral hemorrhagic fevers including Lassa fever	Lassa fever virus	High	Mammarenavirus lassaense	Mammarenavirus lassaense
Arenaviridae			High		Mammarenavirus juninense
Arenaviridae			High		Mammarenavirus lujoense
Astroviridae			Low		Mamastrovirus virginiaense
Bacteria			High	Vibrio cholerae serogroup 0139	
Bacteria			High	Yersinia Pestis	
Bacteria			High	Shigella dysenteriae serotype 1	
Bacteria			High	Salmonella enterica non typhoidal serovars	
Bacteria			High	Klebsiella pneumoniae	
Bornaviridae			Low		Orthobornavirus bornaense
Coronaviridae	Middle East Respiratory Syndrome Coronavirus	Middle East Respiratory Syndrome Coronavirus	High	Subgenus Merbecovirus	Subgenus Merbecovirus
Coronaviridae	Other highly pathogenic coronaviral diseases such as Severe Acute Respiratory Syndrome	Severe Acute Respiratory Syndrome	High	Subgenus Sarbecovirus	Subgenus Sarbecovirus
Filoviridae	Filoviral diseases Ebola	Ebola virus disease	High	Orthoebolavirus zairense	Orthoebolavirus zairense
Filoviridae	Filoviral diseases Marburg	Marburg virus disease	High	Orthomarburgvirus marburgense	
Filoviridae			High	Orthoebolavirus sudanense	
Flaviviridae	Zika virus	Zika virus	High	Orthoflavivirus zikaense	Orthoflavivirus zikaense
Flaviviridae			High	Orthoflavivirus denguei	Orthoflavivirus denguei
Flaviviridae			High	Orthoflavivirus flavi	
Flaviviridae			High		Orthoflavivirus encephalitidis
Flaviviridae			High		Orthoflavivirus nilense
Hantaviridae			High	Orthohantavirus sinnombreense	Orthohantavirus sinnombreense
Hantaviridae			High	Orthohantavirus hantanense	
Hepadnaviridae			Low		Orthohepadnavirus hominoid genotype C

	2017	2018	2024		
Family	Priority Pathogens	Priority Pathogens	PHEIC risk	Priority Pathogens	Prototype Pathogens
Hepeviridae			Low		Paslahepevirus balayani genotype 3
Herpesviridae			Low		
Nairoviridae	Crimean Congo Haemorrhagic Fever	Crimean Congo Haemorrhagic Fever	High	Orthonaïrovirus haemorrhagiae	Orthonaïrovirus haemorrhagiae
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H1	Alphainfluenzavirus Influenzae H1
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H2	
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H3	
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H5	Alphainfluenzavirus Influenzae H5
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H6	
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H7	
Orthomyxoviridae			High	Alphainfluenzavirus Influenzae H10	
Papillomaviridae			Low		
Paramyxoviridae	Nipah and related henipaviral diseases	Nipah and related henipaviral diseases	High	Henipavirus nipahense	Henipavirus nipahense
Parvoviridae			Low		Protoparvovirus carnavoran
Peribunyaviridae			Low		Orthobunyavirus oropoucheense
Phenuiviridae	Severe Fever with Thrombocytopenia Syndrome		High	Bandavirus dabiense	Bandavirus dabiense
Phenuiviridae	Rift Valley Fever	Rift Valley Fever	High		Phlebovirus riftense
Picobinaviridae			Low		Orthopicobinavirus hominis
Picornaviridae			Medium	Enterovirus coxsackiepol	
Picornaviridae			Medium		Enterovirus alphacoxsackie 71
Picornaviridae			Medium		Enterovirus deconjecti 68
Pneumoviridae			Low-Medium		Metapneumovirus hominis
Polyomaviridae			Low		
Poxviridae			High	Orthopoxvirus variola	Orthopoxvirus variola
Poxviridae			High		Orthopoxvirus yacina
Poxviridae			High	Orthopoxvirus monkeypox	Orthopoxvirus monkeypox
Retroviridae			Medium	Lentivirus humimdefl	Lentivirus humimdefl
Rhabdoviridae			Low		Genus Vesiculovirus
Sedoreoviridae			Low		Genus Rotavirus
Spinareoviridae			Low		Orthoreovirus mammalis
Togaviridae			High	Alphavirus chikungunya	Alphavirus chikungunya
Togaviridae			High	Alphavirus venezuelan	Alphavirus venezuelan
Pathogen X	Pathogen X	Pathogen X		Pathogen X	

Genomics is an essential tool for public health management of infectious diseases

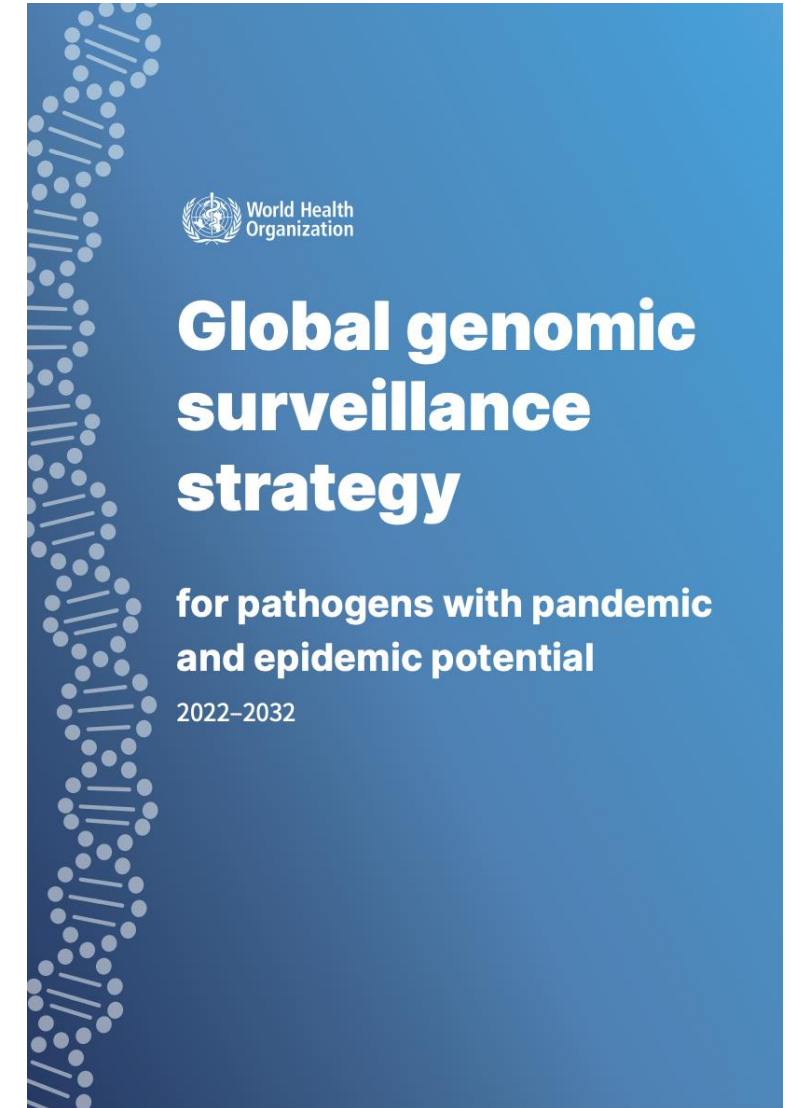
- WGS can overcome important limitations of conventional epidemiological and laboratory methods:

Epidemiological

- Poor case recall
- Case lost to follow-up
- Deliberately misleading information
- Missing cases
- Confirm/refute tenuous epidemiological links

Laboratory

- Multiple tests required to answer different questions about the pathogen
- Often cannot definitively determine whether pathogens are identical



Added resolution provided by WGS

PCR

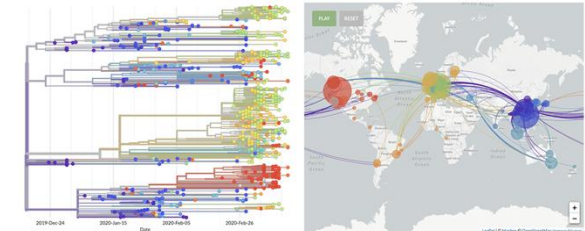
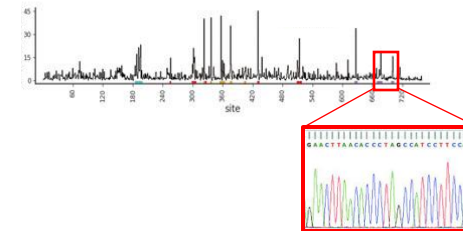
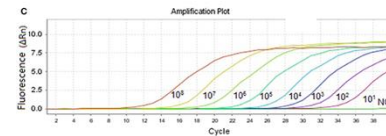
Real time
PCR

Sanger
sequencing

WGS



The fragments are separated by size.



Output:

Presence/
absence

Presence/absence, viral
load, species identification

Species identification,
identification of limited #
mutations

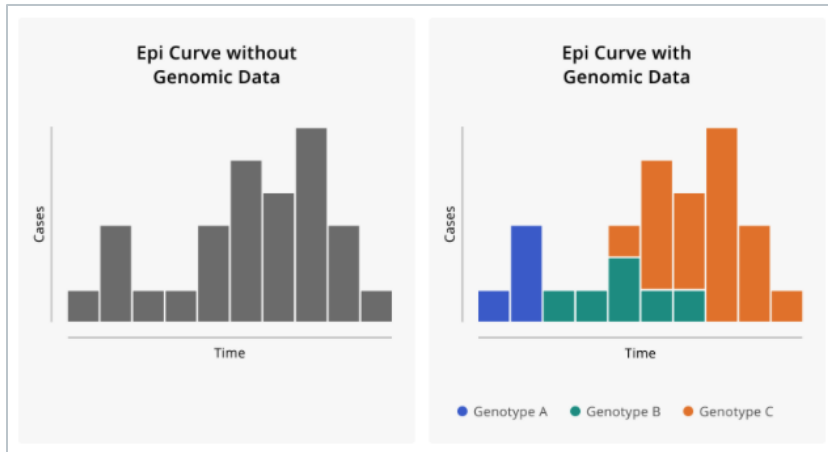
Species/cluster
identification, identification
of all mutations

Low

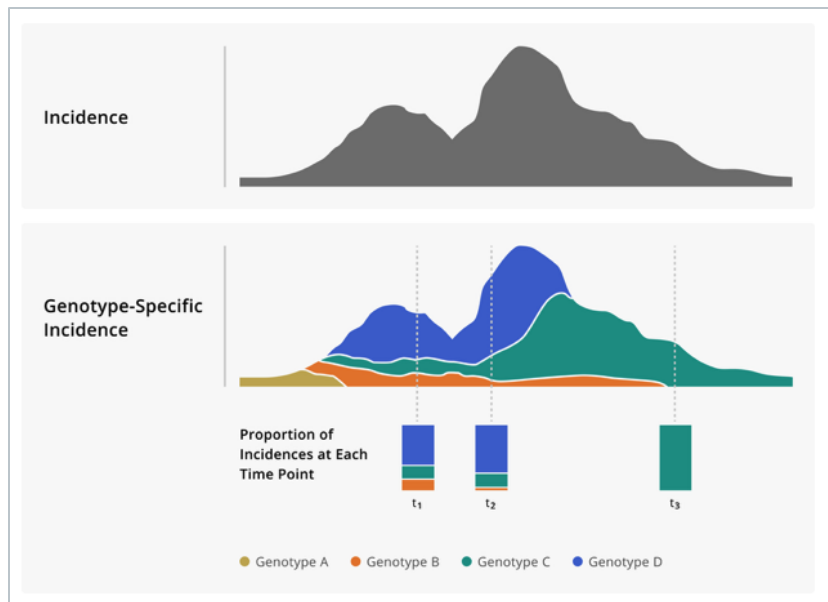
Resolution

High

Genomic data can enable accurate interpretation of epidemiological data

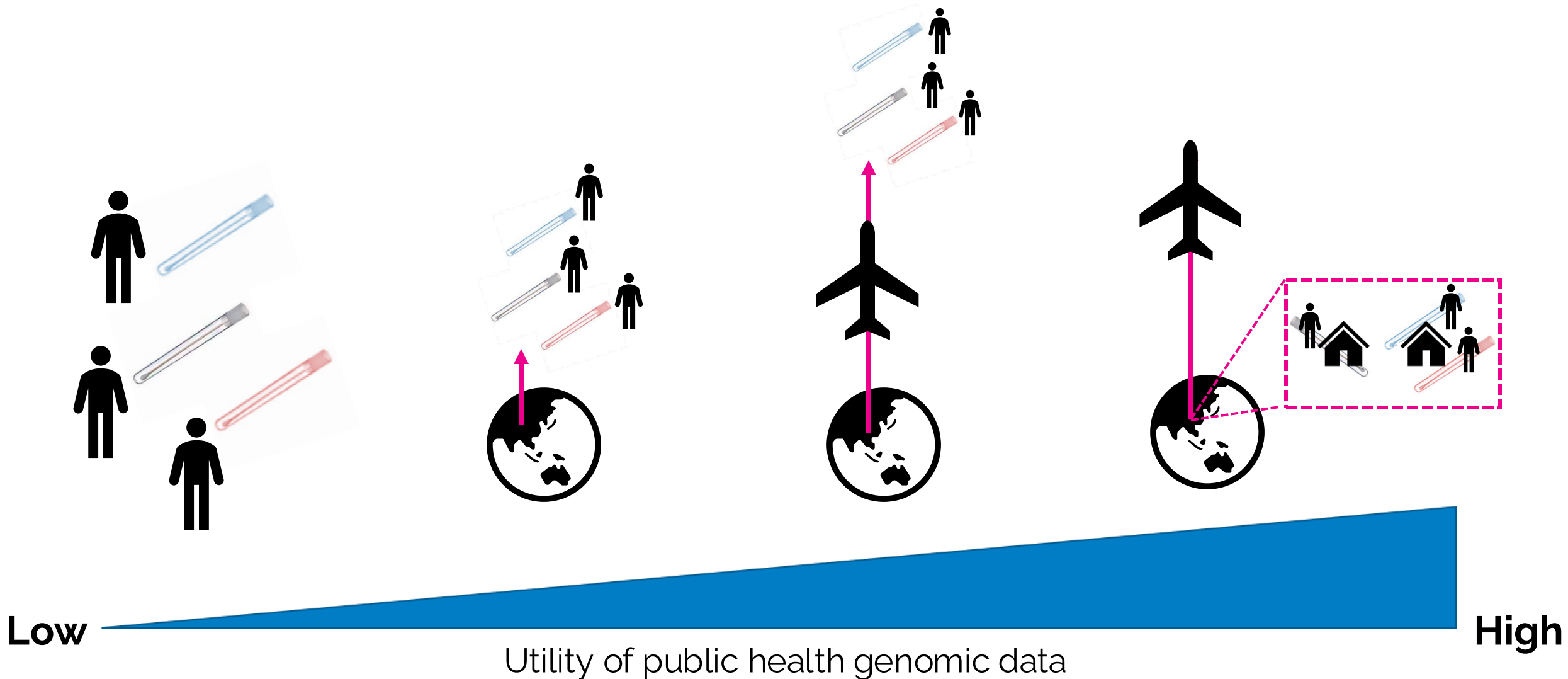


- **Better understand epidemiology at the local level**
 - Investigate clusters
 - Provide evidence confirming or refuting suspected transmission
 - Identify undetected clusters



- **Monitor trends at the population/national level**
 - Monitor emergence of new strains
 - Monitor trends after interventions (i.e. vaccines)

Epidemiological context is crucial to accurately interpret genomic data



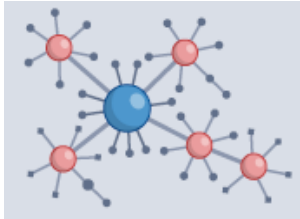
Examples of how genomic data can be used to address specific public health objectives

Resolution of analysis performed depends on public health objective

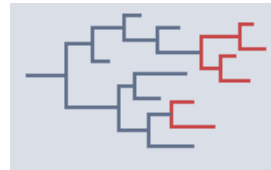
Genome
analysis



Lineage/clade
designation



Phylogenomics/
clustering



Trends: local



Trends: global



High

Low




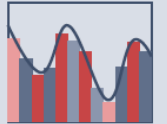
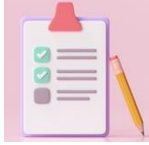



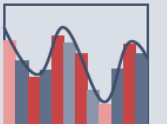






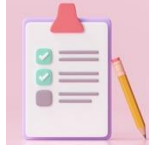



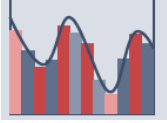

- Contextualising metadata
- Local context
- Translation

- Population overview
- International context

Resolution of public health genomic data

Typical public health (PH) objectives for genomic epidemiological analysis

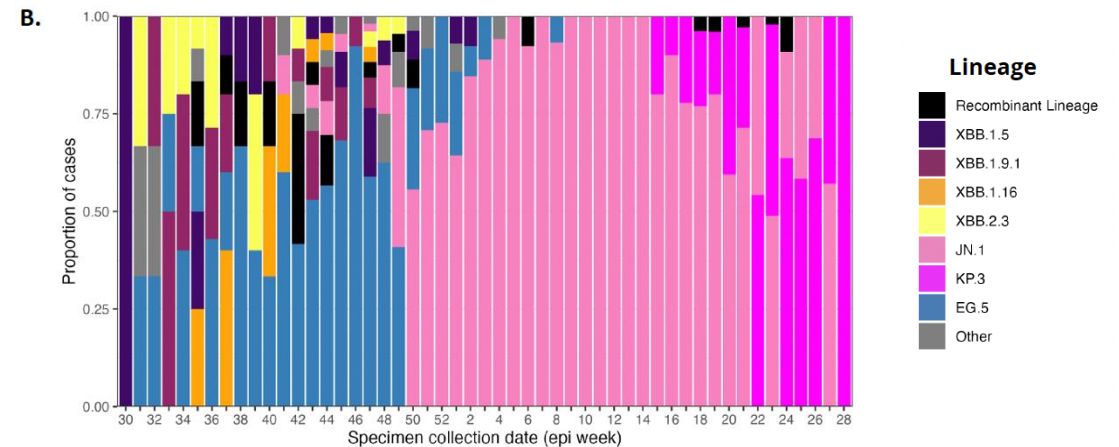
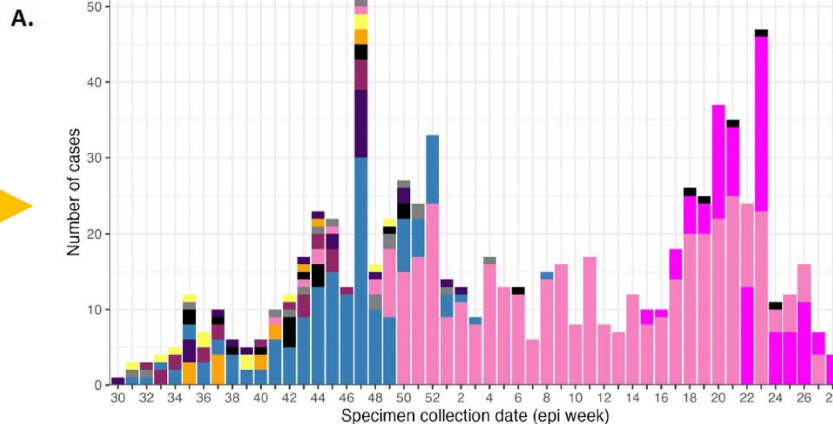
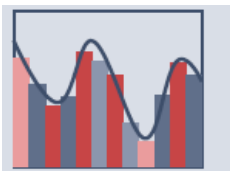
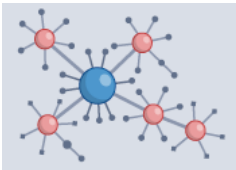
- Are cases linked/exposed to the same source?
- Is the viral lineage/sub-clade novel?
- Will the vaccines/diagnostic tests/antivirals still work?
- Is the lineage/sub-clade associated with severe disease?

Epi	Genomic
	  
	    
	    
	    

How genomic data can be used to address specific public health objectives: Example 1

Objectives:

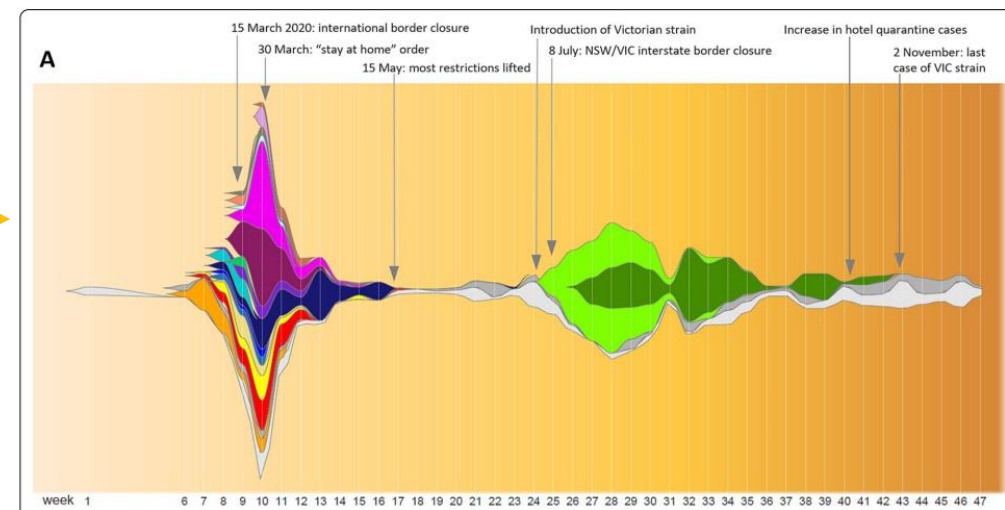
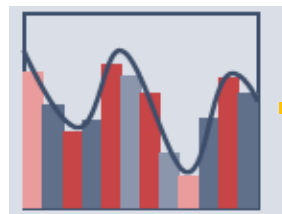
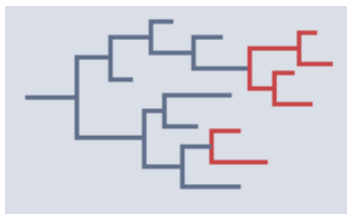
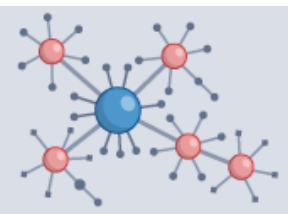
- Are the circulating variants changing?
- Are specific variants associated with severe disease?



How genomic data can be used to address specific public health objectives: Example 2

Objectives:

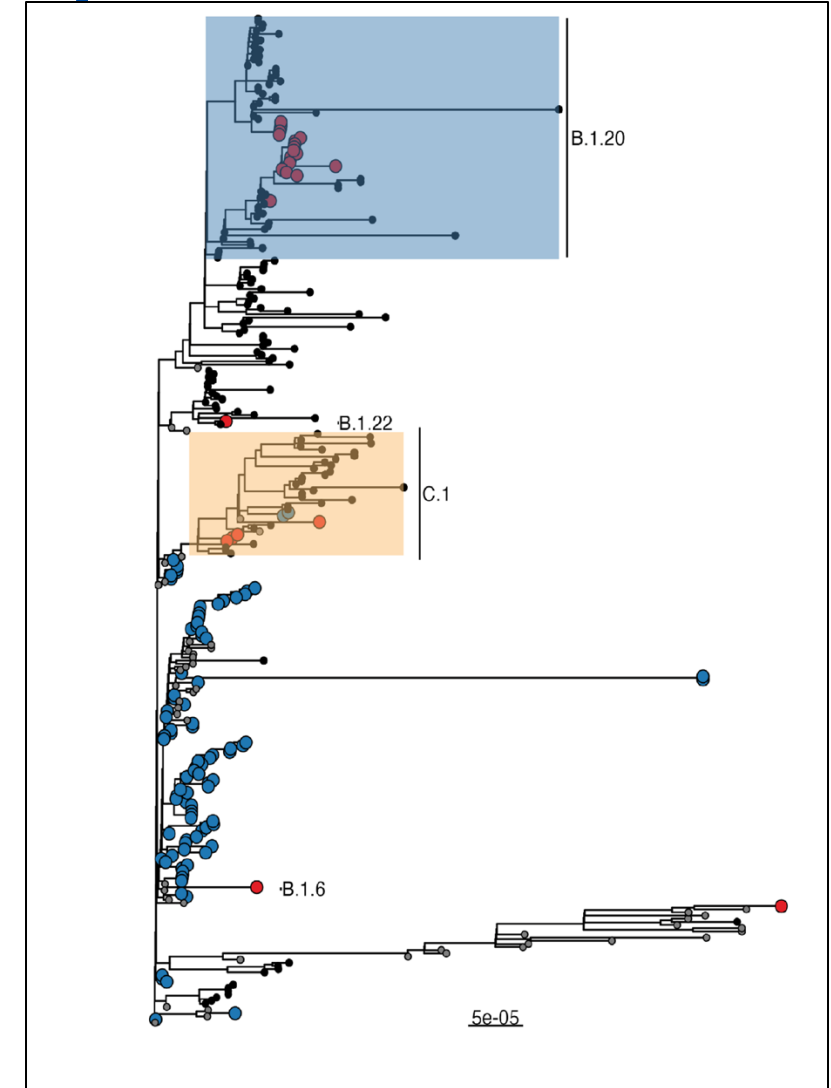
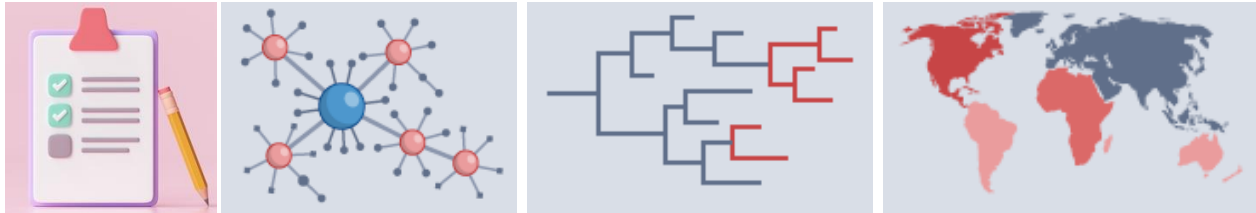
- Are the public health interventions working?



How genomic data can be used to address specific public health objectives: Example 3

Objectives:

- Are all outbreak cases carrying the same variant?
- Have we had multiple or a single incursion?



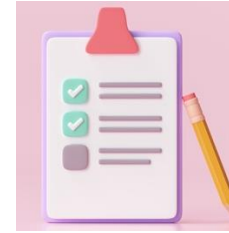
So how do we combine genomic and epidemiological data?

Generating and obtaining data for reporting: epidemiological data

- Necessary for context
- Collected by local/state/national public health departments/stakeholders
- Held in external databases
- Typically demographic details and risk factors specific for the pathogen:
 - Travel history
 - Vaccination status
 - Exposure history: food/contact with known cases



Epidemiological data example: Victoria



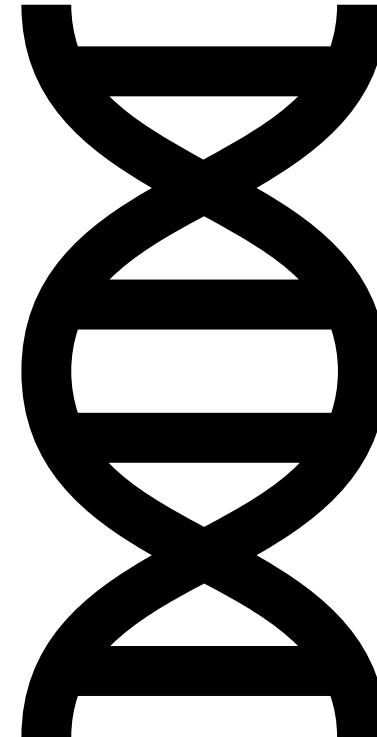
- The availability of detailed epidemiological data depends on the pathogen
- High consequence pathogens:
 - Jurisdictional health departments/PHUs investigate and perform public health follow up
 - Case interviews, sampling, collation of relevant clinical data
- Seasonal/high burden pathogens
 - Follow-up and investigation is not typically performed unless outbreak occurs in specific settings/unusual clinical presentation
- Targeted surveillance is performed for some high burden pathogens, i.e. influenza and FluCAN, resulting in collection of detailed epidemiological data
 - May not be representative of general population



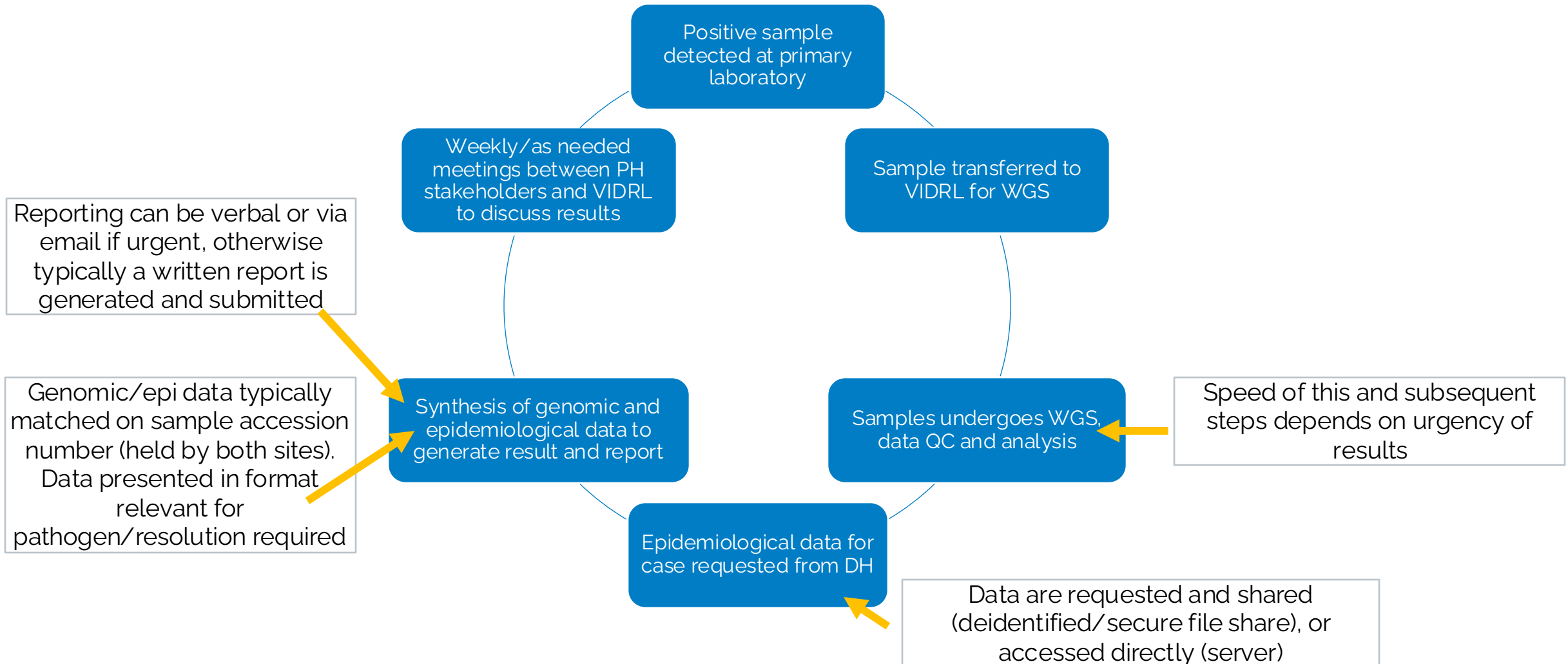
<ul style="list-style-type: none">• Anthrax• Botulism• Candida auris• Cholera• Diphtheria• Food-borne and water-borne illness (two or more related cases)• Haemophilus influenza, type B (meningitis, epiglottitis, other invasive infections)• Hepatitis A• Japanese encephalitis	<ul style="list-style-type: none">• Legionellosis• Listeriosis• Measles• Meningococcal infection (invasive)• Middle East Respiratory Syndrome coronavirus (MERS-CoV)• Murray Valley encephalitis (MVE) virus infection• Novel coronavirus 2019 (2019-nCoV)• Paratyphoid	<ul style="list-style-type: none">• Plague• Poliovirus infection• Rabies• Severe Acute Respiratory Syndrome (SARS)• Smallpox• Tularemia• Typhoid• Viral haemorrhagic fevers• Yellow fever
--	--	---

Generating and obtaining data for reporting: genomic data

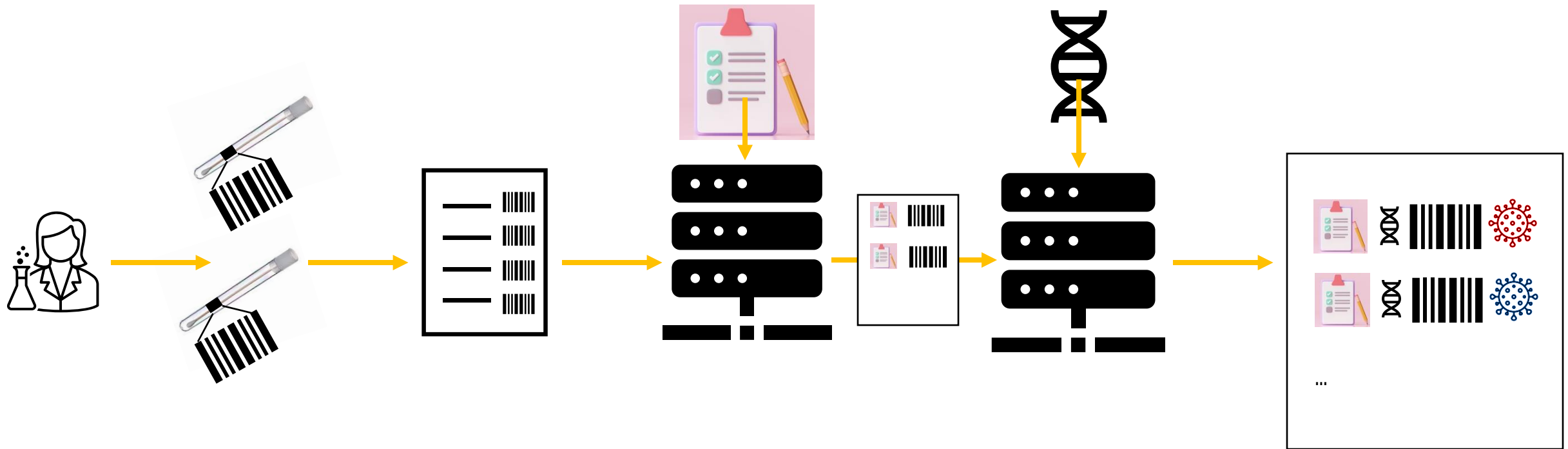
- Generated from case sample by a public health laboratory
- Held in internal databases/servers
- Output is specific for pathogen of interest:
 - Lineage, clade, sub-clade
- Often analyzed in context of international data for genomic context, especially if emergent/non-endemic pathogen



Combining genomic and epidemiological data at VIDRL



Combining genomic and epidemiological data at a jurisdictional PHL



Combining genomic and epidemiological data: example SARS-CoV-2

Lab ID	DH ID	Accession #	Date of collection	Date of birth	Name	Lineage	Collapsed lineage
VIDRL-XXX	DH-XXX	12345	01/01/2024	30/06/1972	Mary Smith	JN.2.3	JN.1
VIDRL-XXX	DH-XXX	12346	07/01/2024	13/04/1999	John Green	JN.3.4.5	JN.1
VIDRL-XXX	DH-XXX	12347	14/01/2024	20/12/2001	Nancy Blue	KP.2.7	KP.2

- Additional epidemiological data often included:
 - Case address/postcode, relevant exposure/travel information
- Additional genomic data often included:
 - WGS QC results (i.e. pass/fail, coverage), date of sequencing, genomic cluster membership (if relevant)

Reporting formats



Lab ID	DH ID	Accession #	Date of collection	Date of birth	Collapsed lineage
VIDRL-XXX	DH-XXX	12345	01/01/2024	30/06/1972	JN.1
VIDRL-XXX	DH-XXX	12346	07/01/2024	13/04/1999	JN.1
VIDRL-XXX	DH-XXX	12347	14/01/2024	20/12/2001	KP.2

- Urgent reports typically communicated by supervising pathologist
- Written reports must be approved by supervising pathologist

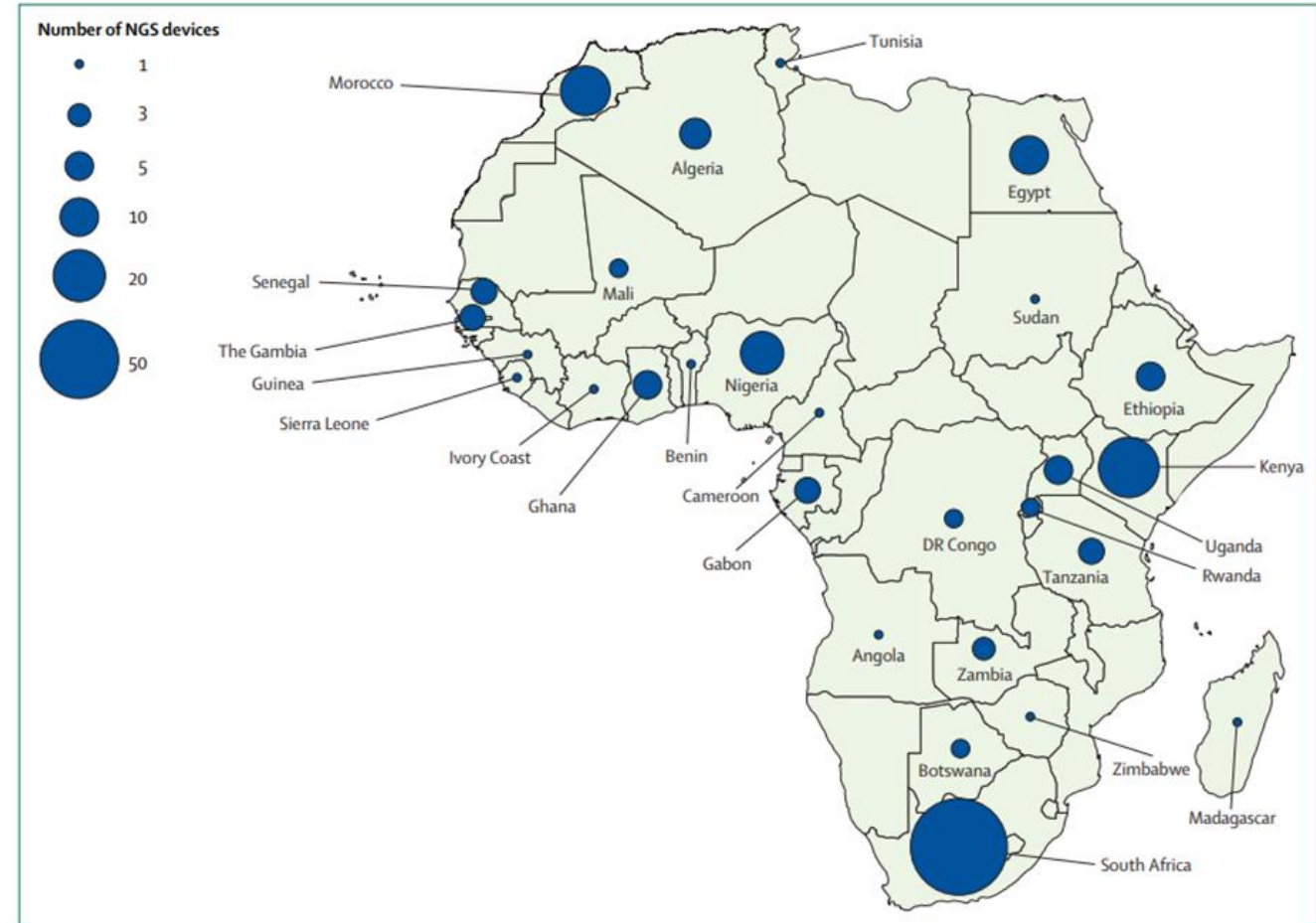
The type of analyses and who receives the report are context dependent

- Resolution of analyses included in report are determined by the public health objectives of the analysis
- Who receives the report is often determined by the situational context:
 - Local DH/LPHU
 - Central DH
 - Outbreak investigation requested by PH/clinical stakeholder
 - Another jurisdictional PHL
 - National surveillance network (i.e. OzFoodNet, FluCAN)
 - Commonwealth DH
 - International health agency (i.e. WHO)

Limitations/considerations

Limitations of WGS: local and global

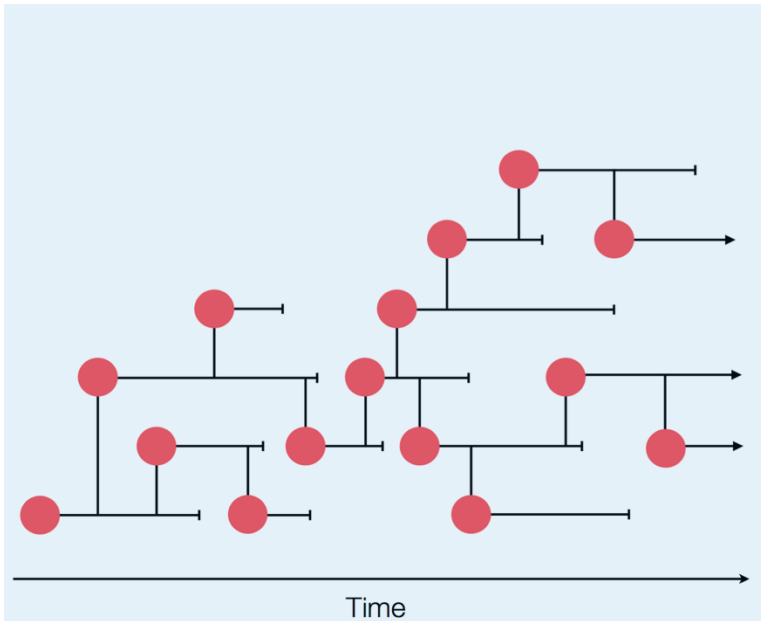
- Directionality of transmission
- Sampling bias
 - Locally
 - Globally
- Global 'blindspots'
- Sample quality/availability
- Speed: timeliness essential
- Reduced utility in the absence of epidemiological data



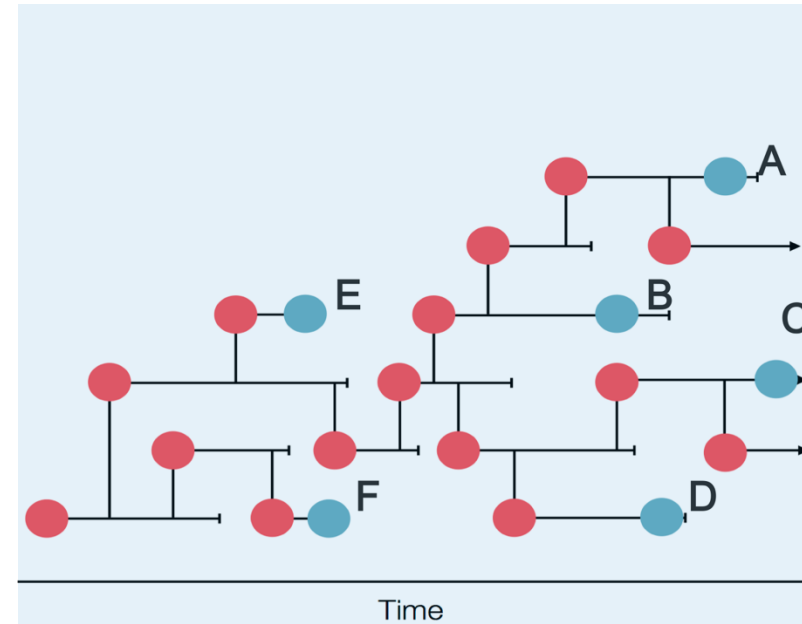
Sampling bias and the importance of metadata

- Sampling bias occurs when the samples collected for sequencing are not representative of the population under investigation:
 - Globally/locally:
 - Geographic areas or populations not represented in sequencing databases
 - Unclear how samples were collected (i.e. outbreak investigation or routine surveillance)
 - Targeted sample collection occurred (not generalizable)
 - Representative sampling occurred (cannot answer specific questions about infection, such as disease severity)
- Sampling bias can affect phylogenetic analysis and interpretation

Sampling bias: impact on interpretation of genomic data

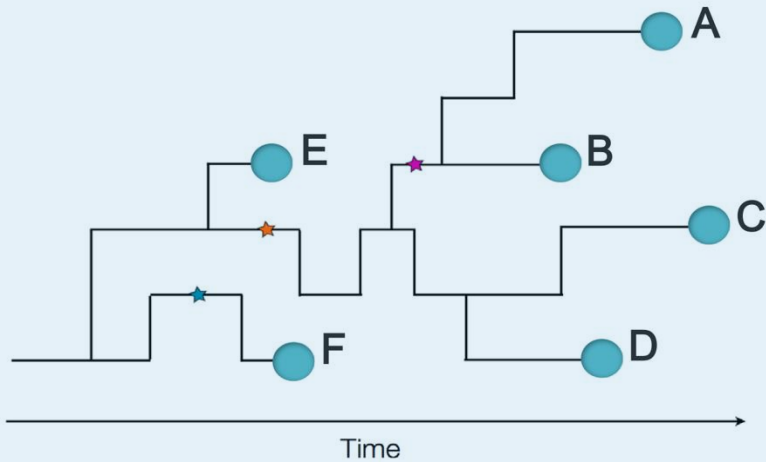


All cases in a population over time

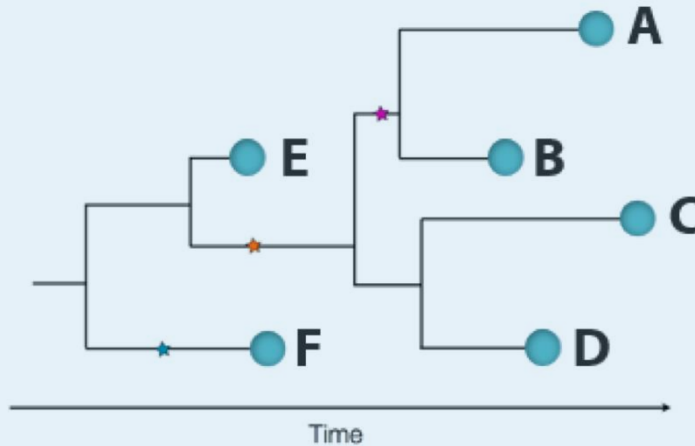


Genomes only available from some cases

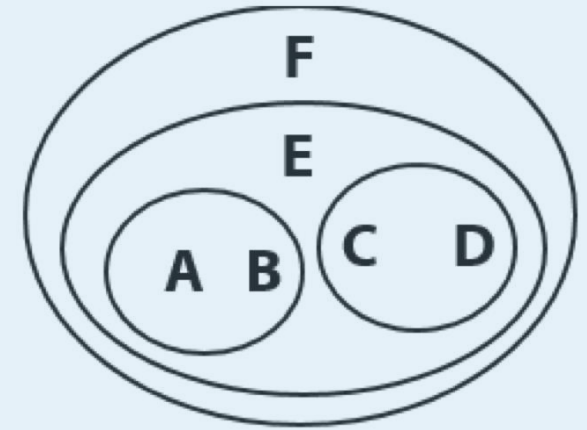
Sampling bias: impact on interpretation of genomic data



Similarity of genomes can be compared

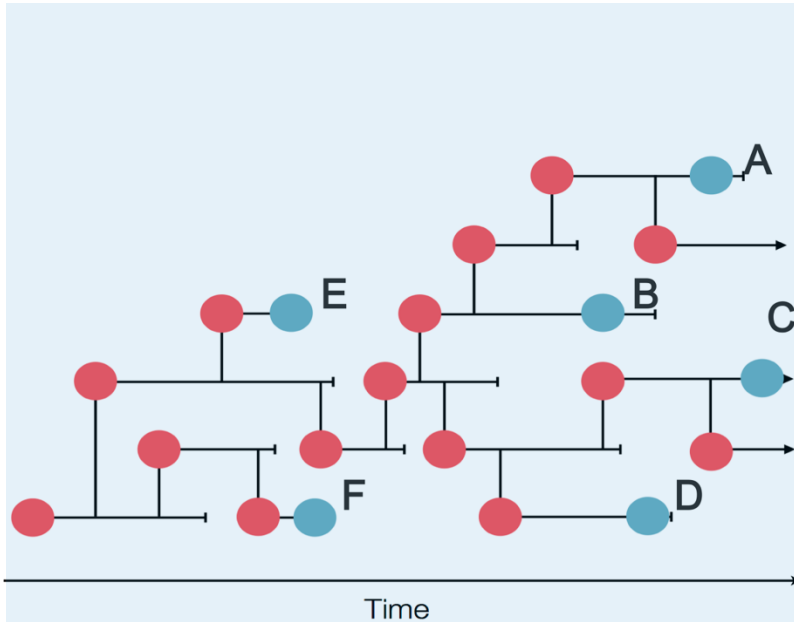


Genomes are represented in a phylogenomic tree



Relatedness between cases inferred

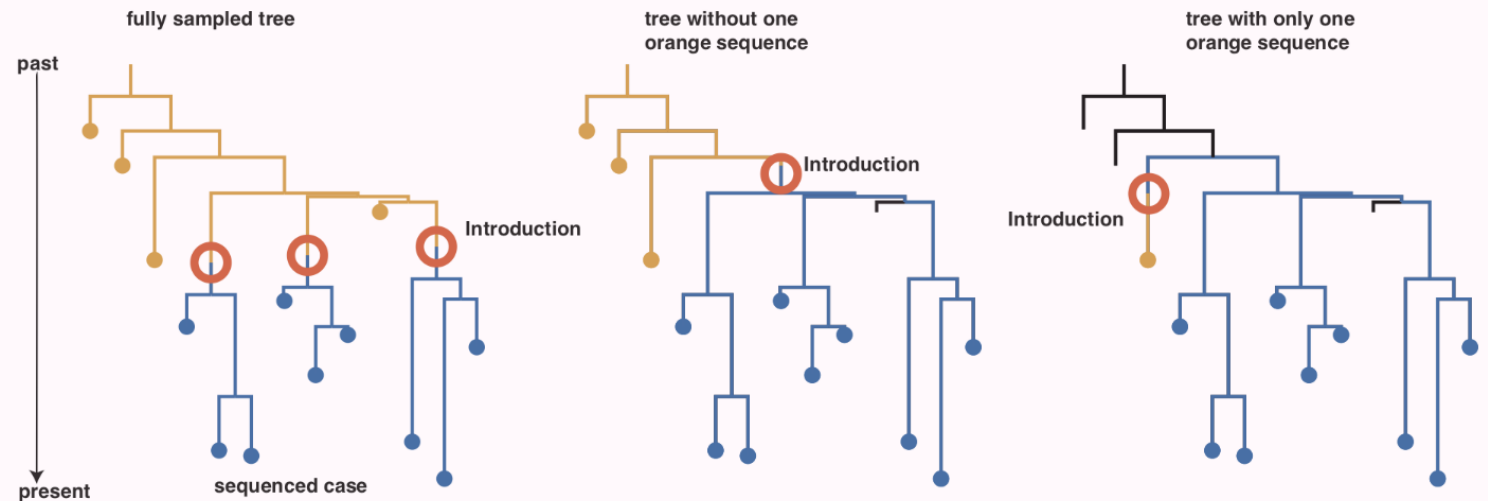
Sampling bias: impact on interpretation of genomic data



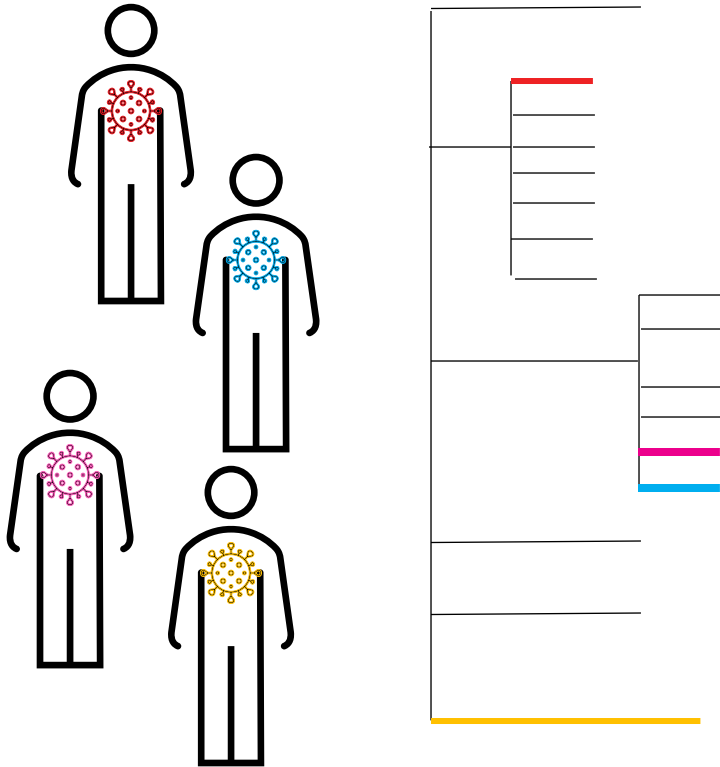
Genomes only available from some cases



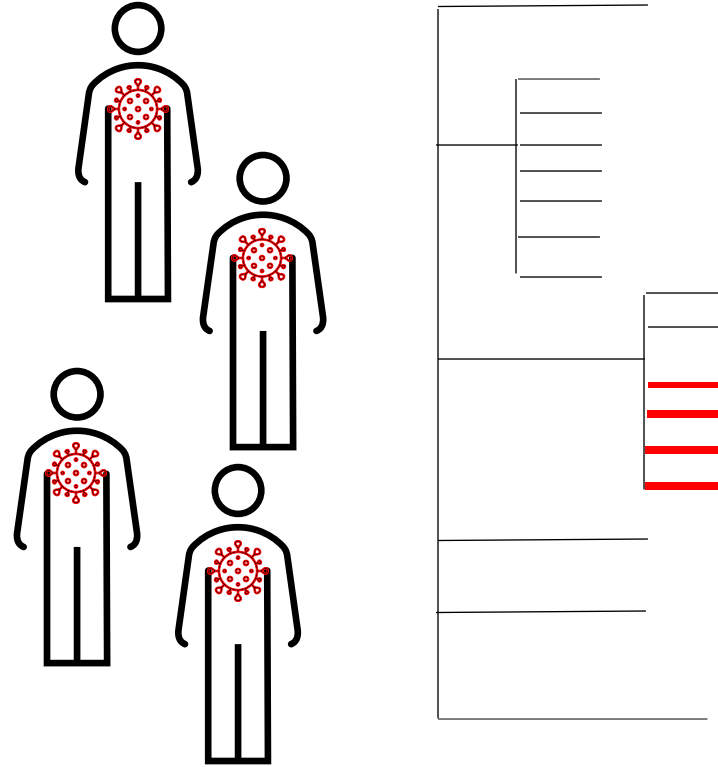
Representativeness of sequenced samples can impact phylogenetic tree structure, impacting interpretation.



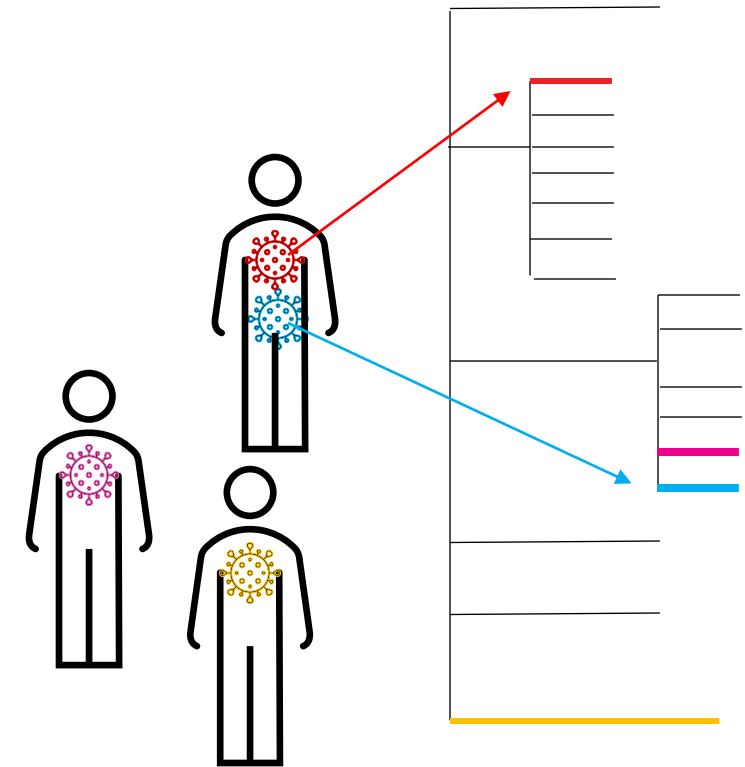
Limitations of genomic epidemiology: not all viral pathogen are the same



One case = one distinct
consensus

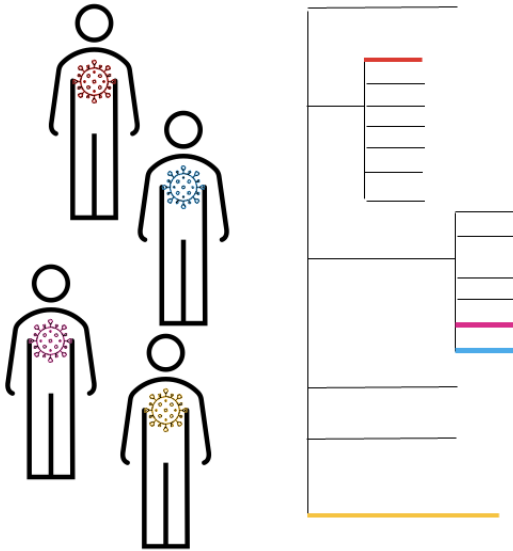


Multiple cases = one identical
consensus (i.e. beginning explosive
transmission)



One case > one distinct
consensus (i.e. Mpox)

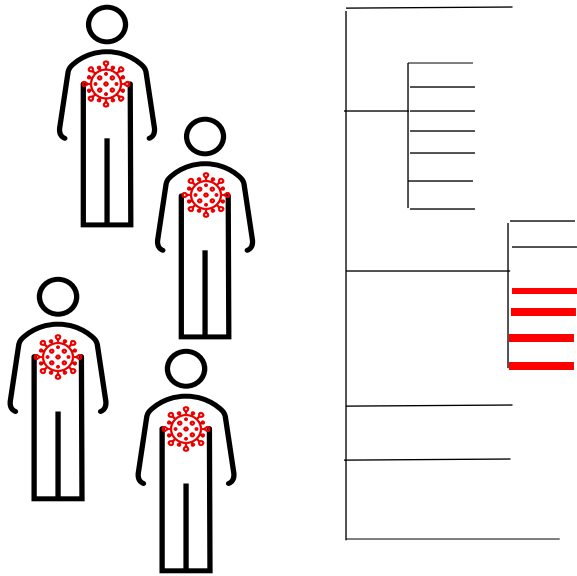
Key considerations in interpretation of genomic epidemiological data: resp. virus



One case = one distinct
consensus

- Directly linking cases must be interpreted in epidemiological context
- Comprehensive global genomic surveillance and high geographical representation permits geographical source attribution and understanding of currently circulating lineages
- Genomic relatedness between cases can be attributed to direct link(s) in the context of robust epidemiological evidence
- A genomic threshold for relatedness between cases has been determined, thus genomic clustering can be performed

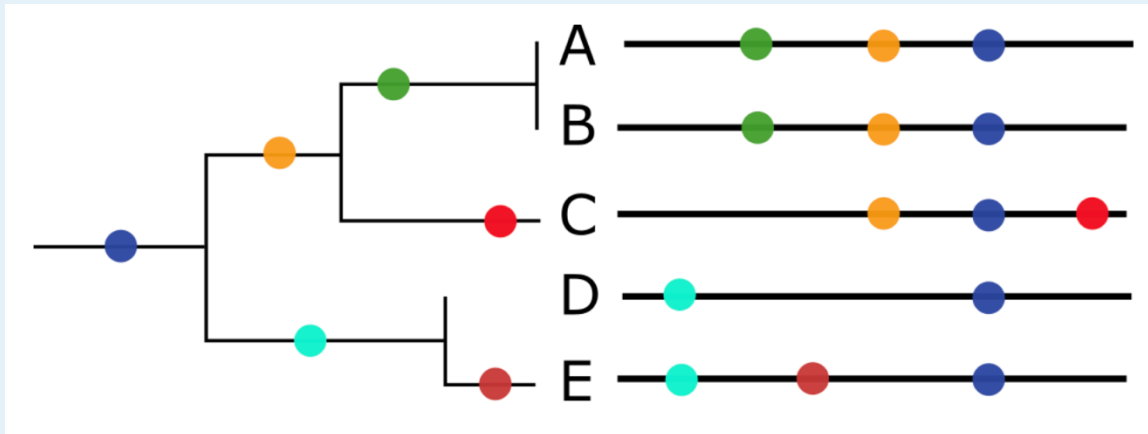
Key considerations in interpretation of genomic epidemiological data: emergent virus



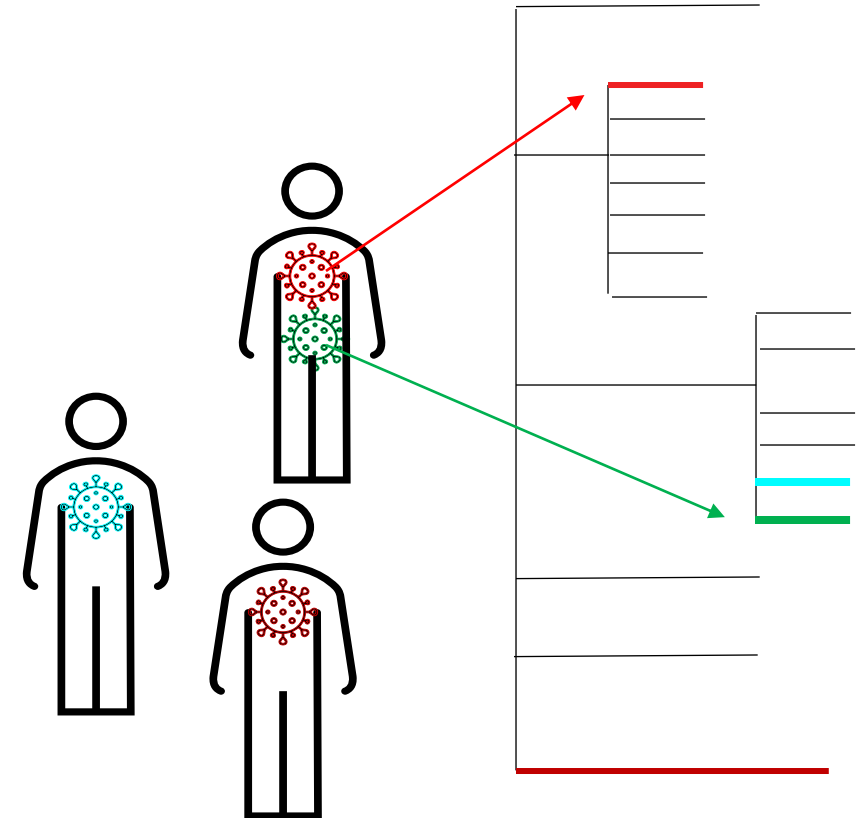
Multiple cases = one identical consensus (i.e. beginning explosive transmission)

- Directly linking cases cannot be conducted using genomics alone and genomic data must be interpreted in epidemiological context
- Lack of global genomic surveillance and bias in geographical representation limits geographical source attribution and understanding of currently circulating lineages
- Genomic relatedness between cases should only be attributed to a common exposure(s) rather than direct link(s) in the absence of robust epidemiological evidence
- Within-host diversity: longitudinal sampling required to better understand
- A genomic threshold for relatedness between cases has not been determined. Genomic clustering may not be able to be performed

Key considerations in interpretation of complex genomic epidemiological data: Mpox



Within-host diversity: sequence diversity between different bodily sites within the same host has been detected, potentially related to complexity of transmission



One case > one distinct consensus (i.e. Mpox)

Conclusions

- Epidemiological (public health and clinical data) are crucial for the accurate interpretation of viral genomic data
- Caveats of genomic and epidemiological data must be well understood to enable accurate interpretation of data
- Reporting must often be tailored to address specific public health objectives
- Genomic epidemiology is not pathogen agnostic

THANK YOU

alicia.arnott@unimelb.edu.au

