Public Health Ontario | Santé publique Ontario

To view an archived recording of this presentation please click the following link:

https://youtu.be/vMTCcQBKRZk

Please scroll down this file to view a copy of the slides from the session.

Disclaimer

This document was created by its author and/or external organization. It has been published on the Public Health Ontario (PHO) website for public use as outlined in our Website Terms of Use. PHO is not the owner of this content. Any application or use of the information in this document is the responsibility of the user. PHO assumes no liability resulting from any such application or use.

Whole genome sequencing: methods, utility, and implementation challenges for Infectious Disease surveillance

Aaron Campigotto, MD, FRCPC

Medical Microbiologist

Hospital for Sick Children

Learning objectives

- Describe the use of pathogen WGS for infectious disease surveillance

- Develop a structured approach to analyze and interpret WGS data for infectious disease surveillance

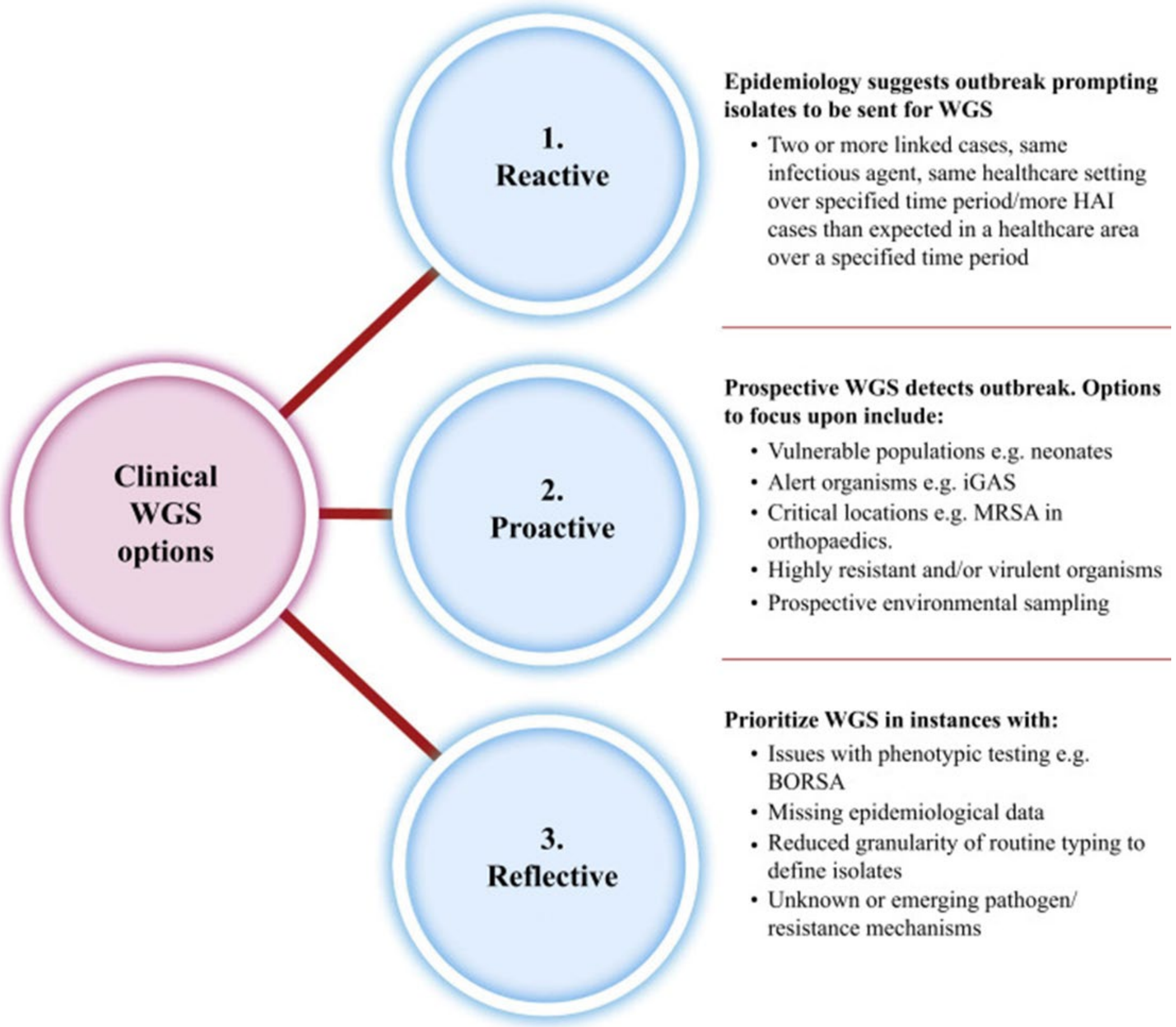
- Appreciate the implementation challenges for the development and interpretation of WGS data

Poll One

What is your experience with analyzing WGS data to understand infectious disease surveillance?

1. Very experienced
2. Experienced
3. Some experience
4. Limited experience
5. No experience

The role of WGS in surveillance and infection prevention and control



Epidemiology suggests outbreak prompting isolates to be sent for WGS

- Two or more linked cases, same infectious agent, same healthcare setting over specified time period/more HAI cases than expected in a healthcare area over a specified time period

Prospective WGS detects outbreak. Options to focus upon include:

- Vulnerable populations e.g. neonates
- Alert organisms e.g. iGAS
- Critical locations e.g. MRSA in orthopaedics.
- Highly resistant and/or virulent organisms
- Prospective environmental sampling

Prioritize WGS in instances with:

- Issues with phenotypic testing e.g. BORSA
- Missing epidemiological data
- Reduced granularity of routine typing to define isolates
- Unknown or emerging pathogen/ resistance mechanisms

- WGS data may be used to:

    1. Provide organism level information that assists in outbreak investigations

    2. Act as a mechanism for surveillance

- Timely data access and sharing

Parcell *et al.,* J Hosp Infect, 2021

Goals of WGS in surveillance and IPAC investigations

- Goal of WGS analysis is to determine if cases are linked (if there is transmission occurring between cases) or monitor for certain traits
  - If strains are different – transmission can be ruled out
  - If strains are identical or similar – transmission not definitively proven based on genomic sequence alone
    - Conserved genomes among organisms in outbreak (low diversity)
    - Incomplete sampling

- *Epidemiological and other supporting evidence is key*

From phenotype to genotype: Laboratory methods used for pathogen surveillance

- Pathogen typing methods to provide information on potential relatedness between cases
  - Surveillance
  - Outbreak investigation

- Methods may include:
  - Phenotypic-based methods (e.g. serotyping, AST)
  - Restriction-based methods (e.g. PFGE)
  - Molecular-based methods (e.g. RAPD, MLVA)
  - Whole genome sequencing

- Key performance characteristics to consider for each method:
  - Resolution
  - Relatedness

Comparison of select pathogen typing methods

	AST profiling	PFGE	WGS
Discriminatory power	Poor	Excellent*	Excellent
Universal applicability	Low	Moderate	High
Complexity of data	Low	Complex	Very complex
Ease of use	Low	Moderately labour-intensive	Labour-intensive
Cost	Low	Moderate	High

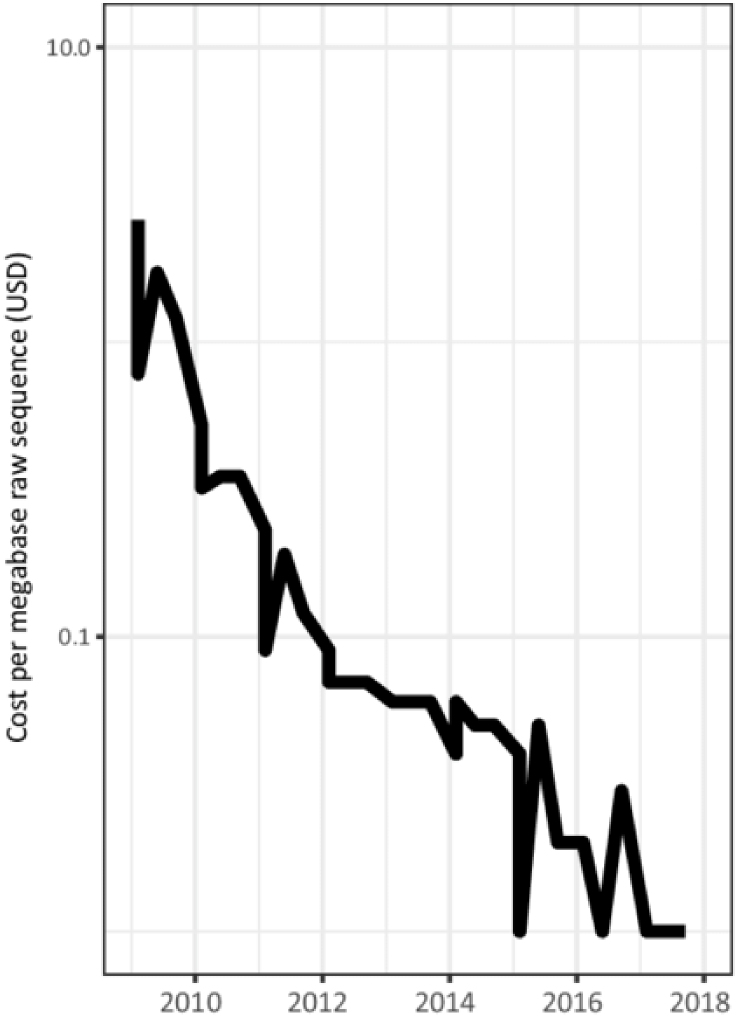
Modified from Manual of Clinical Microbiology, 4th ed

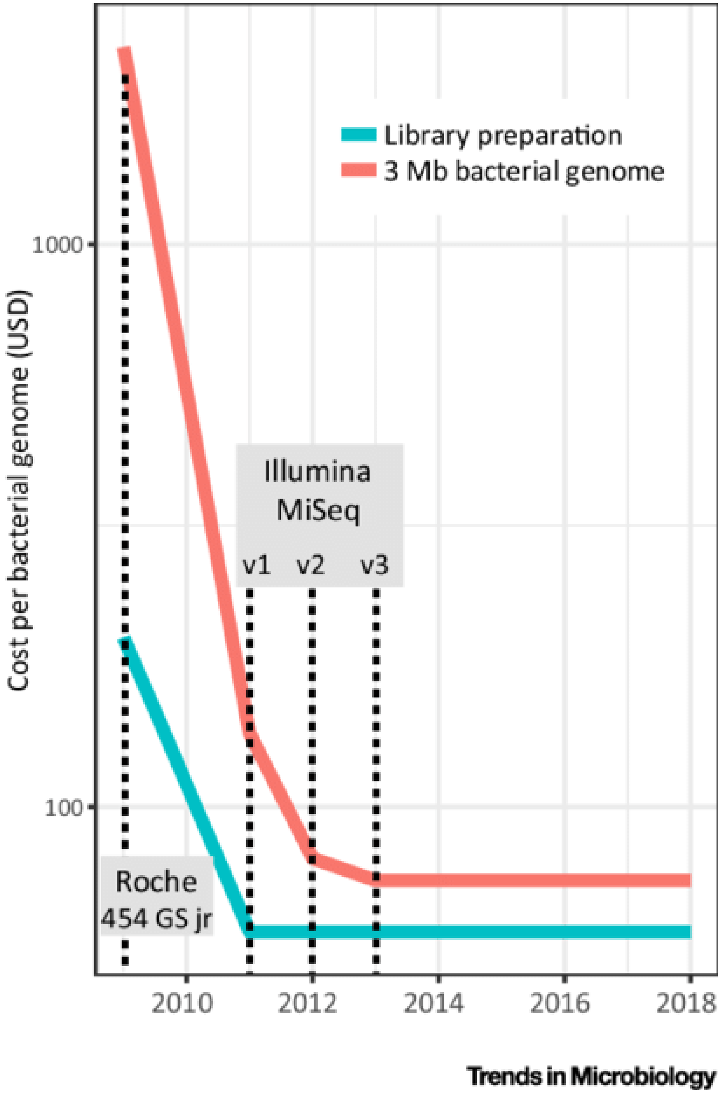
Evolution of sequencing technologies

- Significant changes in the way we sequence and associated costs over the past 2 decades

- From sanger sequencing to NGS

- Comparing sequencing methods
  - Monomicrobial vs polymicrobial
    - Bacterial, fungal, viral
  - Performance characteristics
    - Diagnostic vs surveillance/outbreak investigation





Balloux et al, Trends in Microbiology 2018

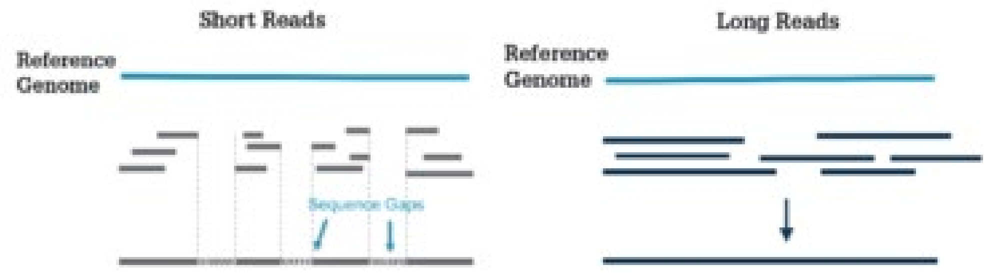
Next generation sequencing

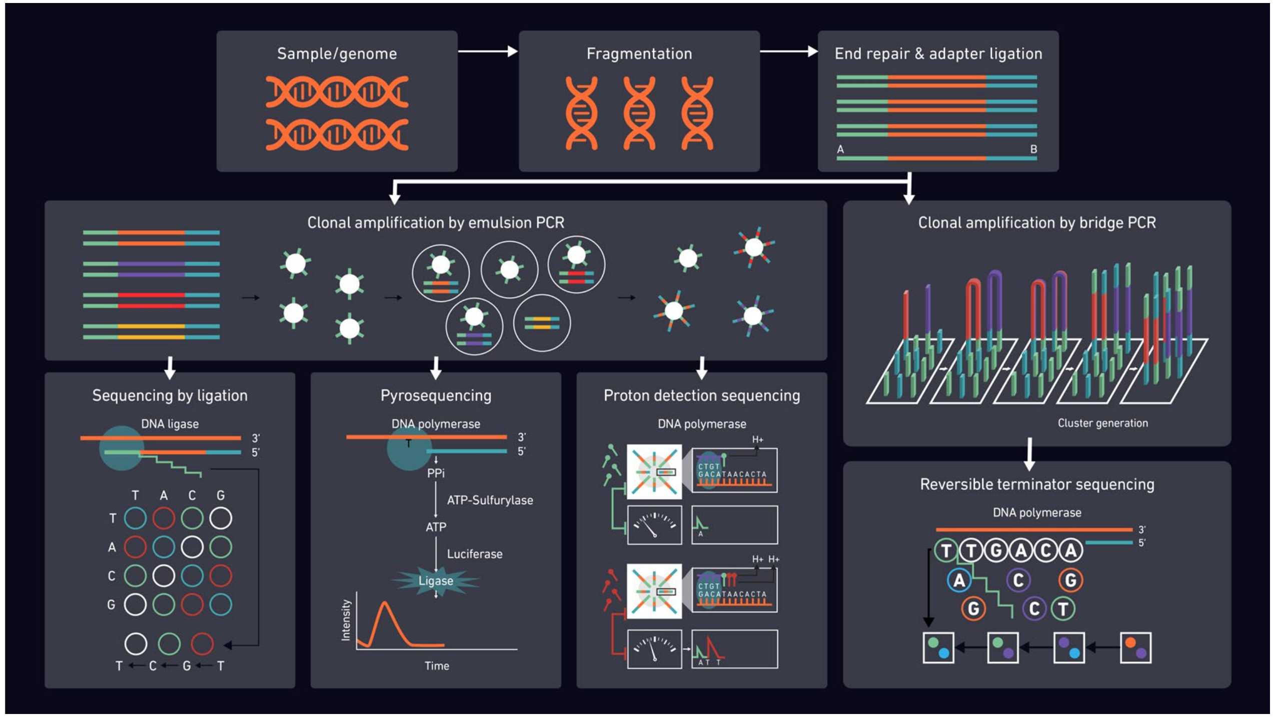
- Allows for rapid and accurate generation of a full pathogen genome (WGS)

- Different techniques each with their own advantages and disadvantages
  - Long vs short read sequencing

- May sequence from isolate (e.g. bacteria culture) or directly from primary specimen

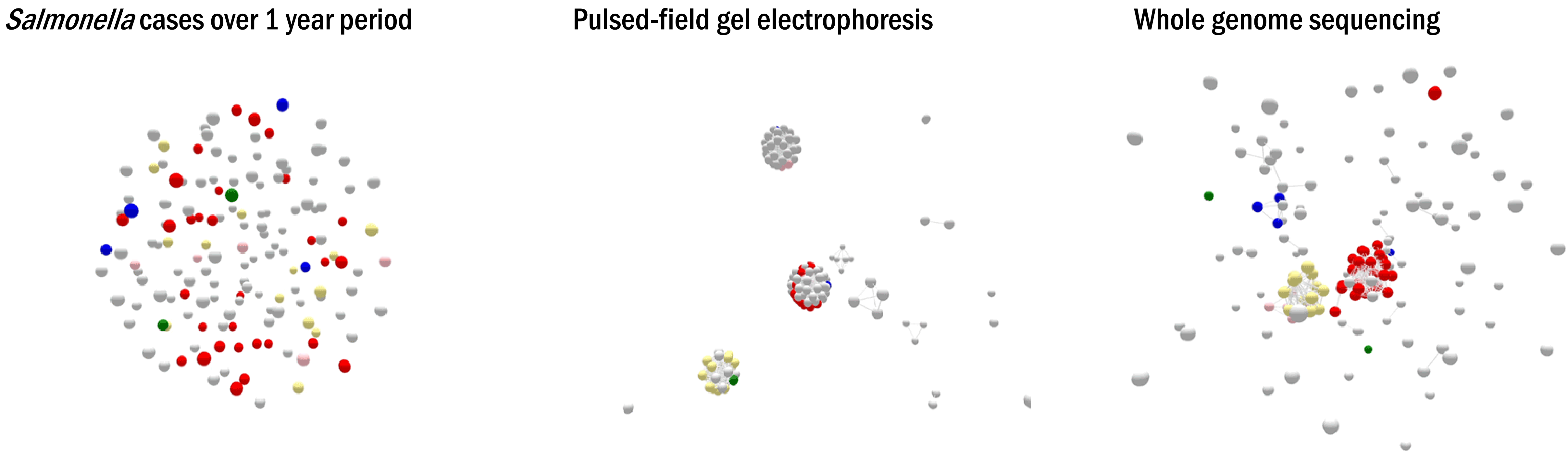
- Complex data analysis (informatics!)





https://thepathologist.com/diagnostics/smrt-long-read-sequencing-solves-genetic-mysteries;
Gkazi, Technology Networks 2021

Resolution of pathogen typing methods



Salmonella cases over 1 year period

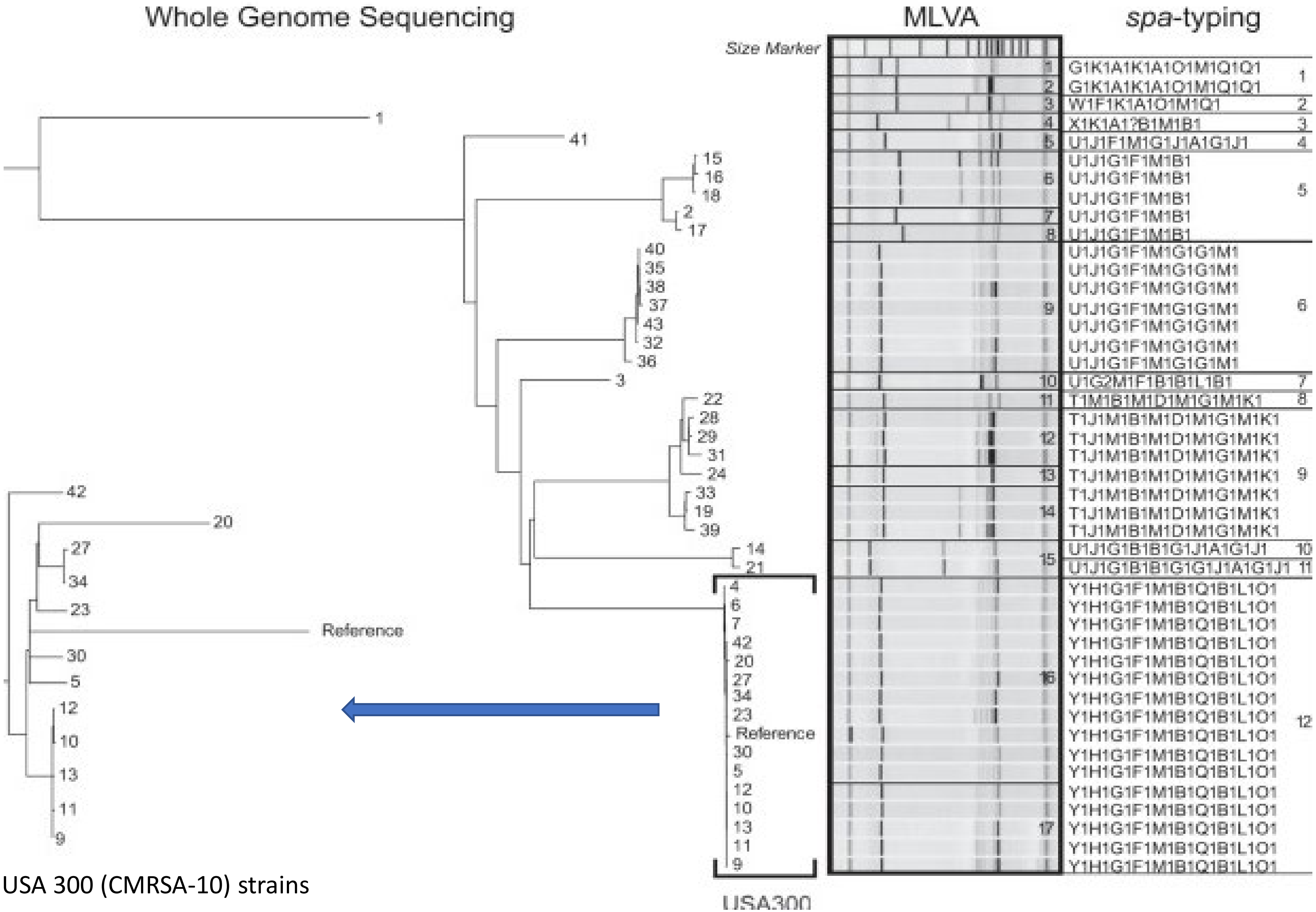
Pulsed-field gel electrophoresis

Whole genome sequencing

- The resolution or relatedness of organisms by each subtyping method may assist in ***ruling-out*** cases in a cluster
- Consider use in ruling-in cases or defining transmission patterns

https://www.cdc.gov/amd/how-it-works/detecting-outbreaks-wgs.html

Comparison of resolution by different methods



USA 300 (CMRSA-10) strains

- MRSA typing methods demonstrating differences in resolution:

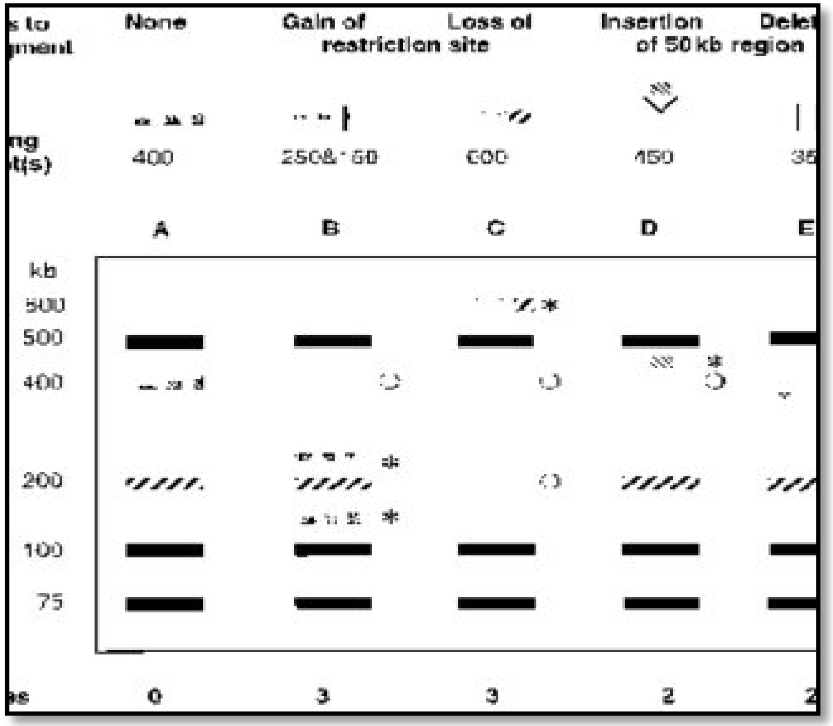
  WGS>MLVA>spa-typing

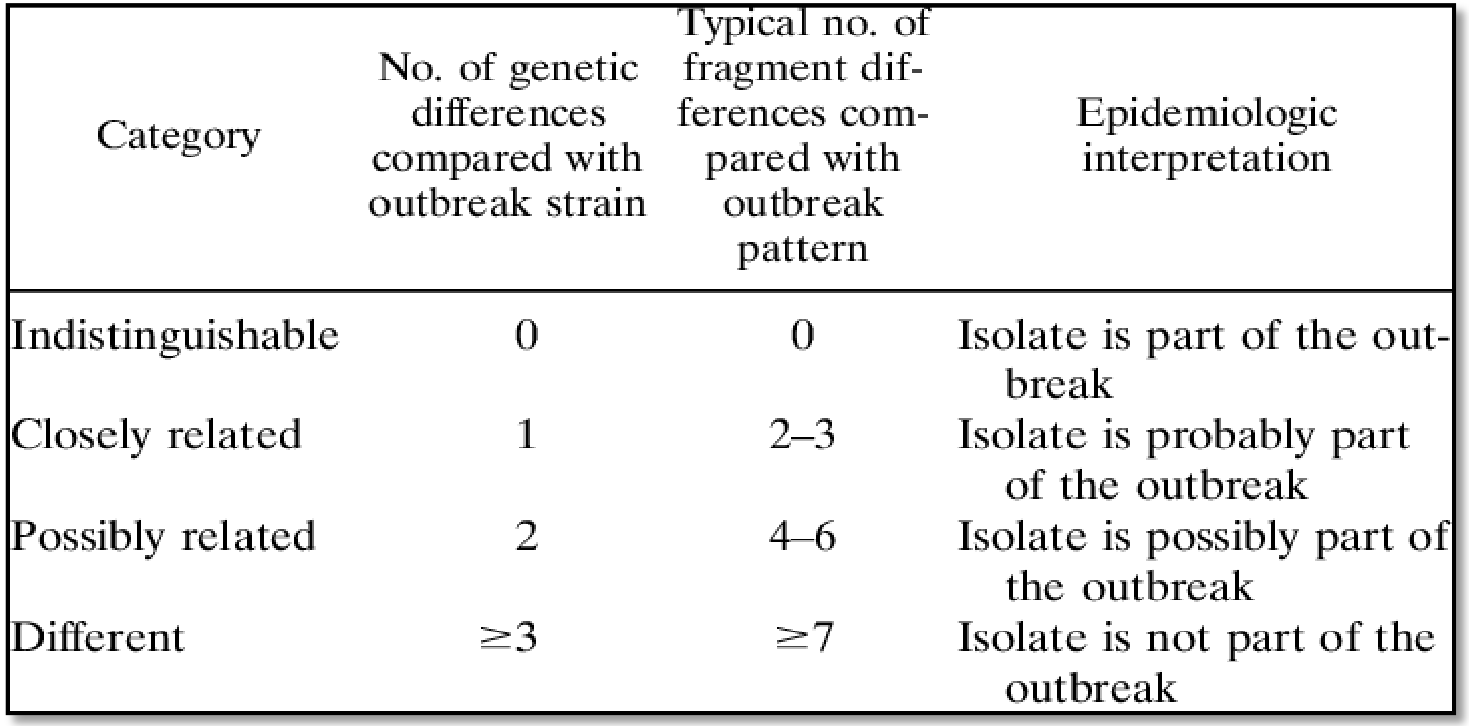
- WGS most sensitive
  - Lower false identification of outbreaks (e.g. rule-out cases)

SenGupta *et al.*, JCM, 2014

How can we define **relatedness**?

- **Interpretation criteria!**
  - E.g. PFGE - “gold-standard” for bacterial comparison
  - Advantages
    - Many epidemiologic studies showing concordance



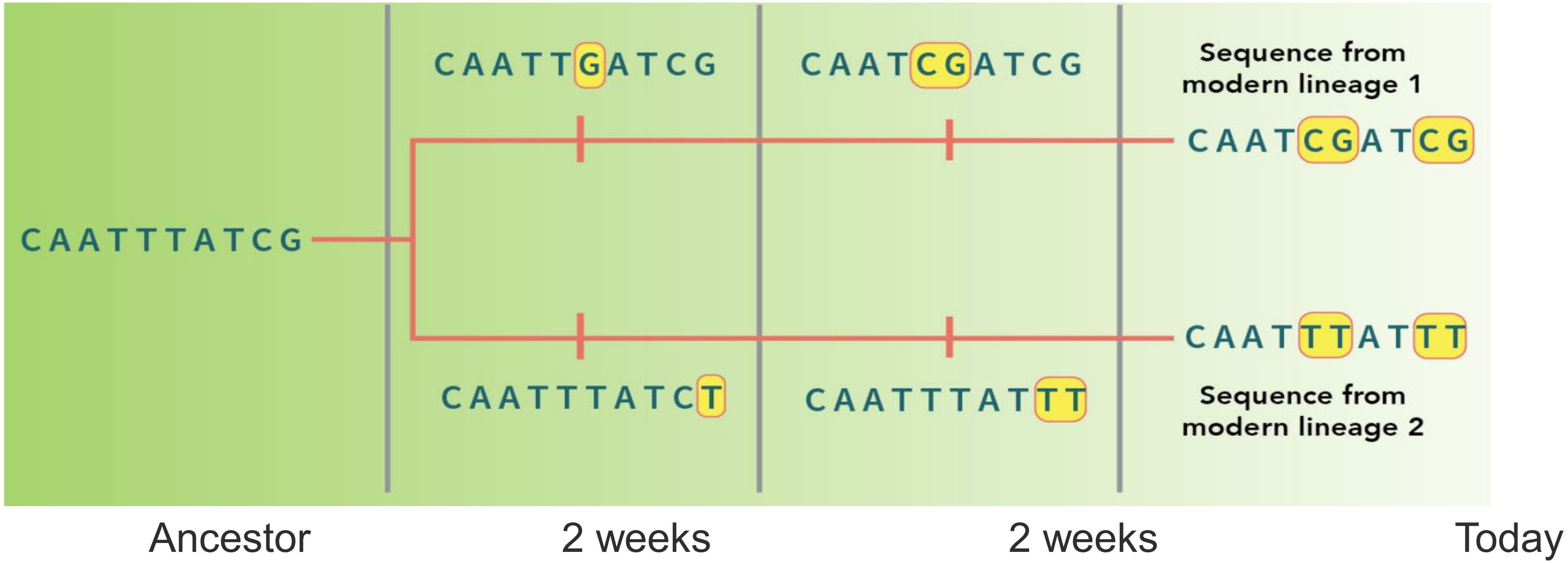


Category	No. of genetic differences compared with outbreak strain	Typical no. of fragment differences compared with outbreak pattern	Epidemiologic interpretation
Indistinguishable	0	0	Isolate is part of the outbreak
Closely related	1	2–3	Isolate is probably part of the outbreak
Possibly related	2	4–6	Isolate is possibly part of the outbreak
Different	≥3	≥7	Isolate is not part of the outbreak

Tenover *et al.*, JCM, 1995

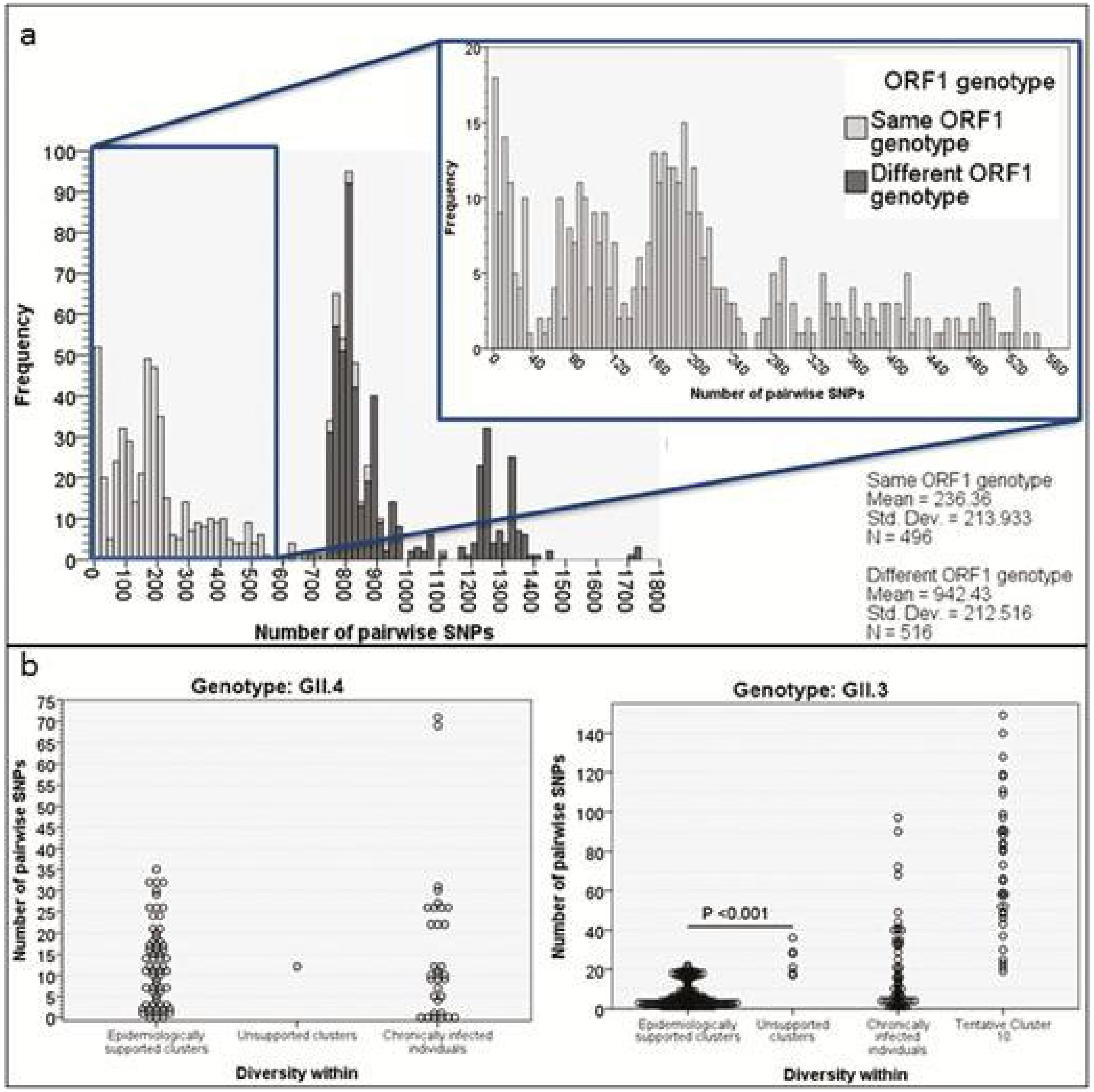
Pathogen mutation rate

- Key to define and understand relatedness between pathogens
- Mutation rate (molecular clock) of an organism assists in understanding the number of mutations that would be expected overtime to consider an organism as different or unrelated
- Can vary significantly based on organism
  - Errors in sequencing technology and amplification process should be considered



https://evolution.berkeley.edu/molecular-clocks/

Defining pathogen relatedness – Norovirus



- Define the number of SNPs (different base pairs at each positive) between various groups
  - Genotypes
  - Clusters of cases/outbreaks
  - Same patient over time

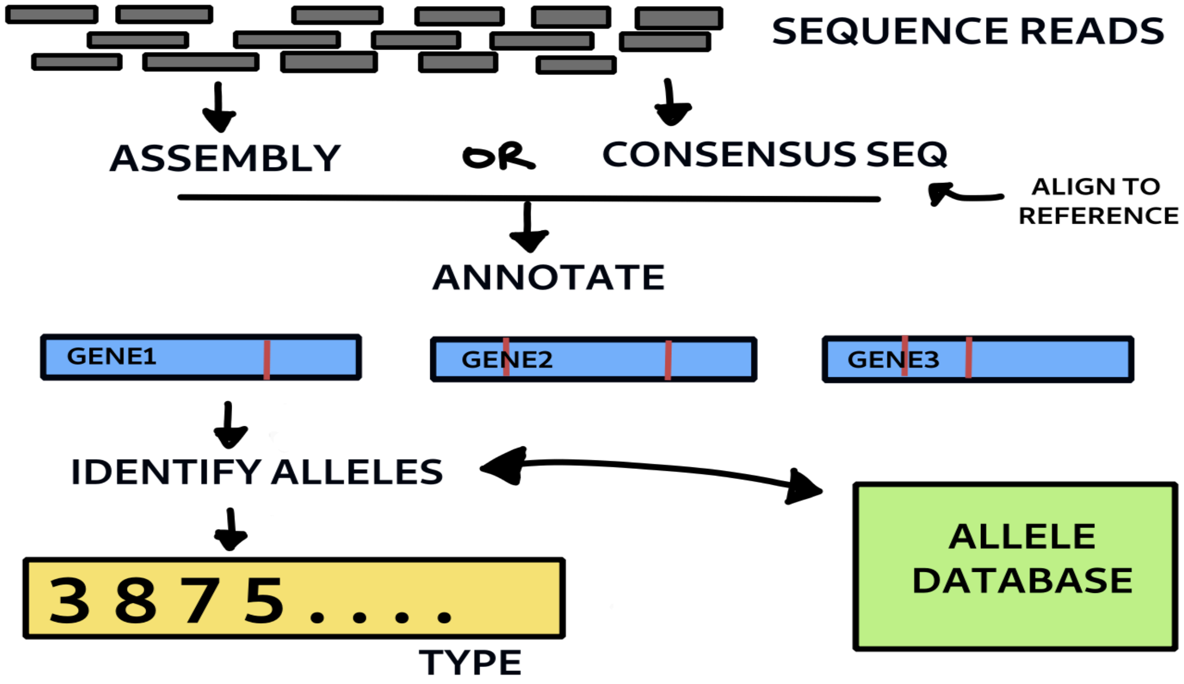
- Clinical epidemiology necessary in development and interpretation

Brown *et al.,* CID 2019

Analysis of WGS data for typing and surveillance

1. Whole genome MLST (wgMLST)/ core genome MLST (cgMLST)

- 1000’s of different genetic regions are analyzed (instead of 5-10 in MSLT)
- Identity between each genes (alleles) is compared and the number of genes with differences is used to define relatedness NOT individual point mutations
- Not standardized
  - E.g. MRSA ≤8 vs 18-24 differences to define related isolates has been proposed
  - Requires ongoing clinical validation

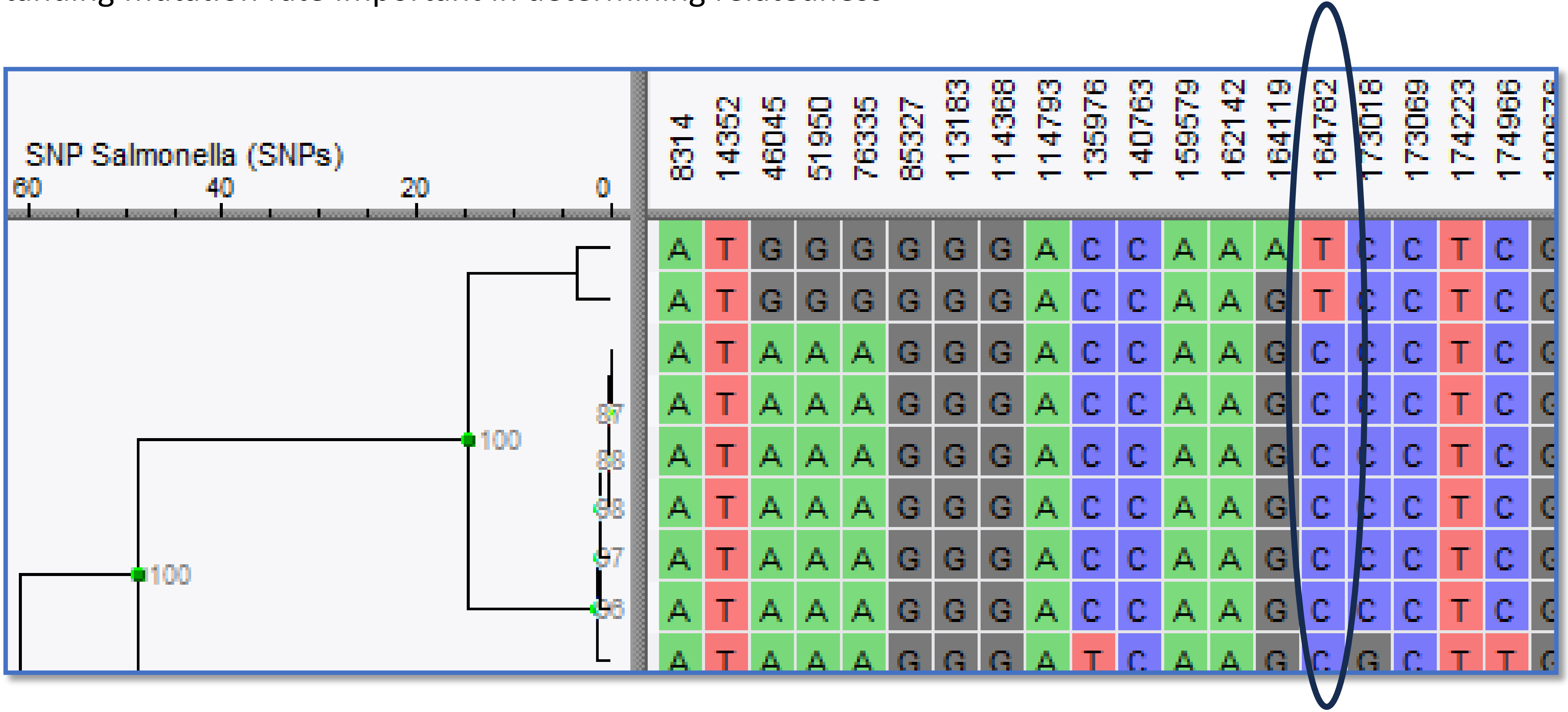


https://dmnfarrell.github.io/bioinformatics/wgmlst-mbovis

Analysis of WGS data for typing and surveillance

2. Whole genome single nucleotide polymorphisms

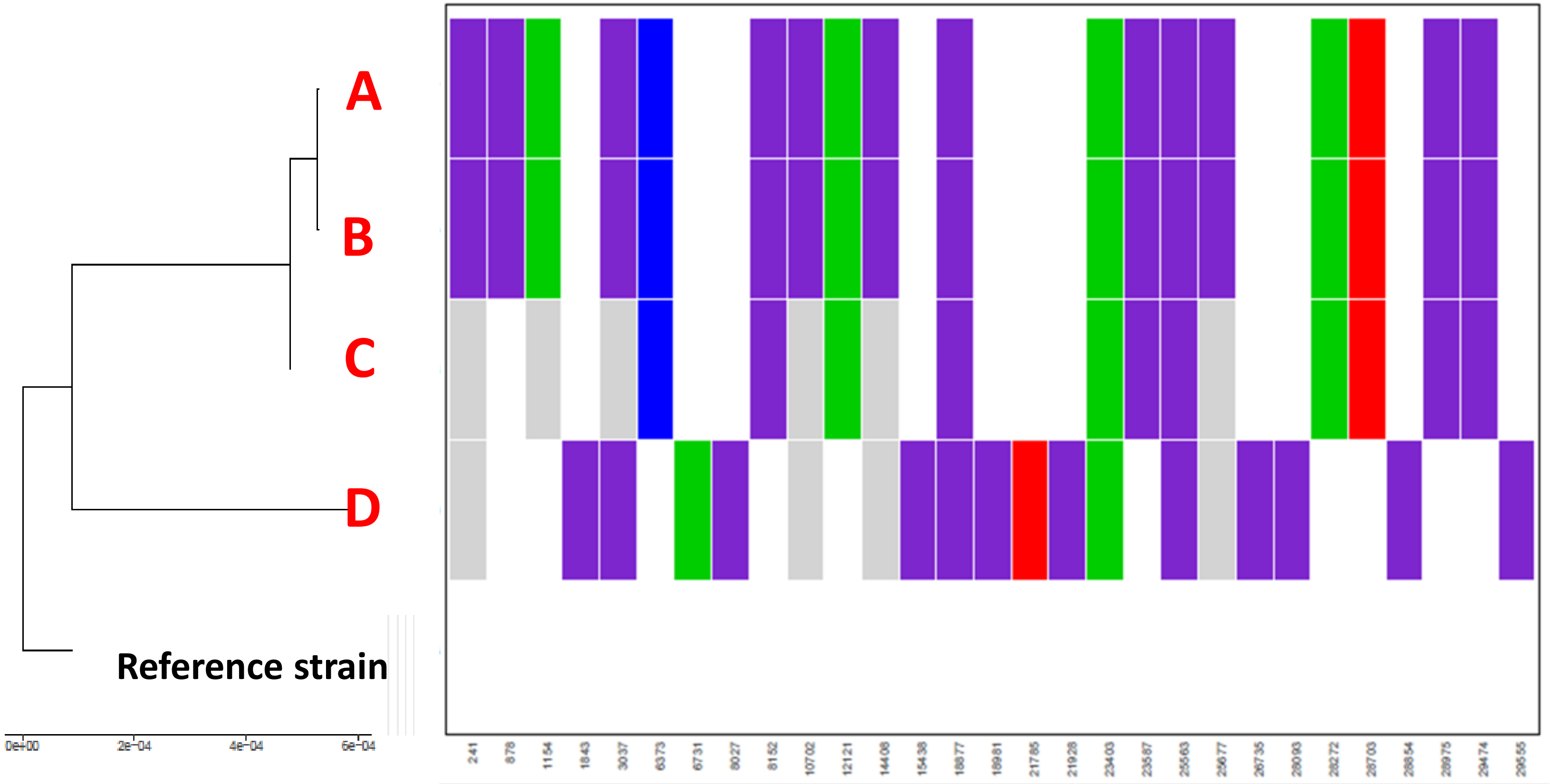
- Each nucleotide difference is captured throughout the organism genome
- Understanding mutation rate important in determining relatedness



8314	14352	46045	51950	76335	85327	113183	114368	114793	135976	140763	159579	162142	164119	164782	173018	173069	174223	174966	[illegible]
A	T	G	G	G	G	G	G	A	C	C	A	A	A	T	C	C	T	C	G
A	T	G	G	G	G	G	G	A	C	C	A	A	G	T	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	T	C	A	A	G	C	G	C	T	T	G

https://www.applied-maths.com/applications/whole-genome-snp-

Interpreting the data



- A and B are **identical**
  1. Related (based on epi) *or*
  2. Not related

- C and A/B are **nearly identical**
  1. Related (based on epi) *or*
  2. Not related

- D is **different** from A/B/C
  1. Transmission *did not* occur between cases
     - Not related

- Example relatedness classification:
  - 0 mutations are **Identical**
  - 1 - 2 mutations are **Nearly Identical**
  - 3 mutations **Similar**
  - >3 mutations **Different**
    - Note: samples must be collected within a similar time period

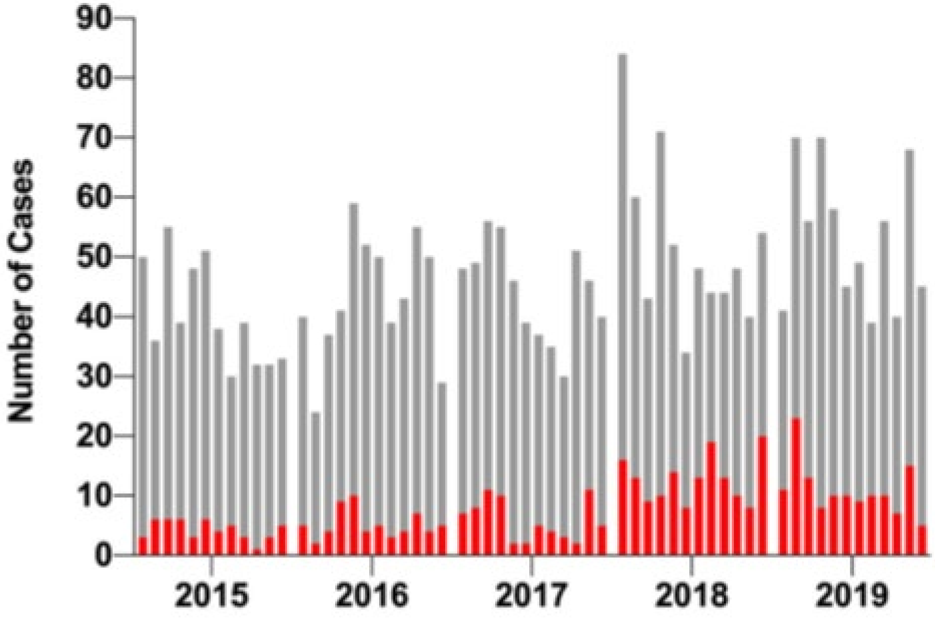
Poll Two

What is your preferred method to receive WGS data to understand transmission within cluster?

1. Discrete sequence data
2. Phylogenetic tree
3. An assigned cluster number
4. Other or not sure

Communicating and reporting results

- Background
  - Case example Human adenovirus surveillance within a pediatric center (GOSH)
  - Human adenovirus (HAdV) infections among paediatric HSCT population may be a cause of significant clinical disease
  - Nosocomial spread may occur with genomics facilitating a more robust understanding
  - Genomics data may be used to monitor transmission to evaluate and modify IPAC policy
  - Assigning and characterization of genetic clusters is undefined but core to understanding viral transmission

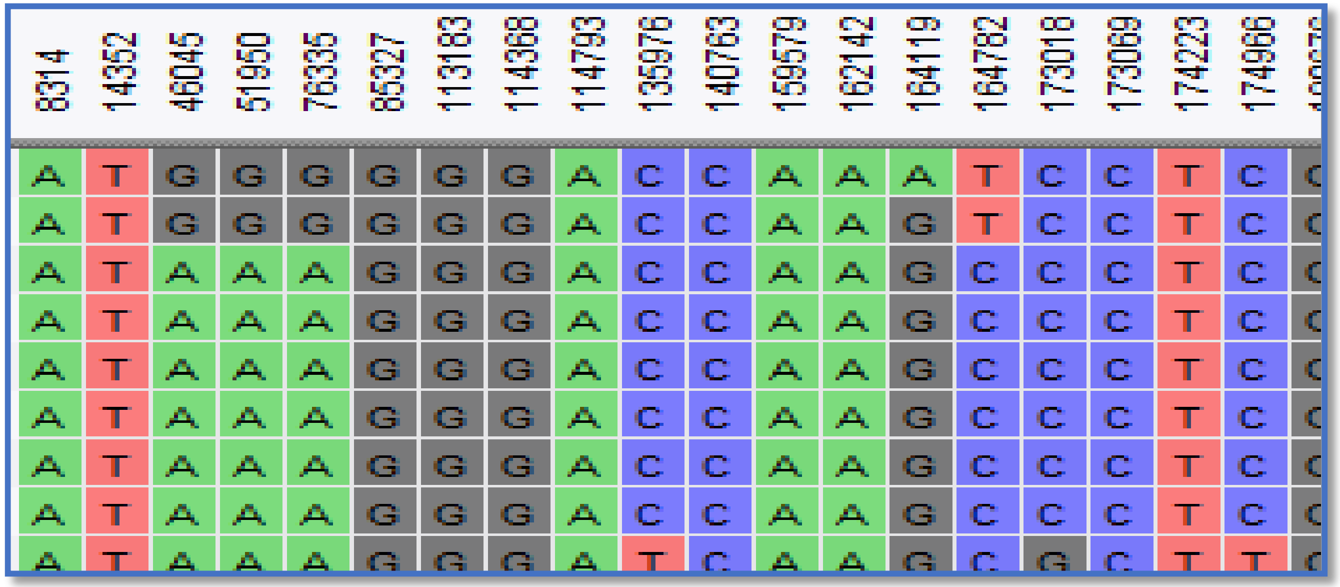


Hospital acquired infections increasing (diagnosis ≥48 hours after admission)

Myers *et al.,* Frontiers in Microbiol, 2021

Communicating and reporting results

1. Comparison of consensus genetic sequence for all isolates



8314	14352	46045	51950	76335	85327	113183	114368	114793	135976	140763	159579	162142	164119	164782	173018	173069	174223	174966	[illegible]
A	T	G	G	G	G	G	G	A	C	C	A	A	A	T	C	C	T	C	G
A	T	G	G	G	G	G	G	A	C	C	A	A	G	T	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	C	C	A	A	G	C	C	C	T	C	G
A	T	A	A	A	G	G	G	A	T	C	A	A	G	C	G	C	T	T	G

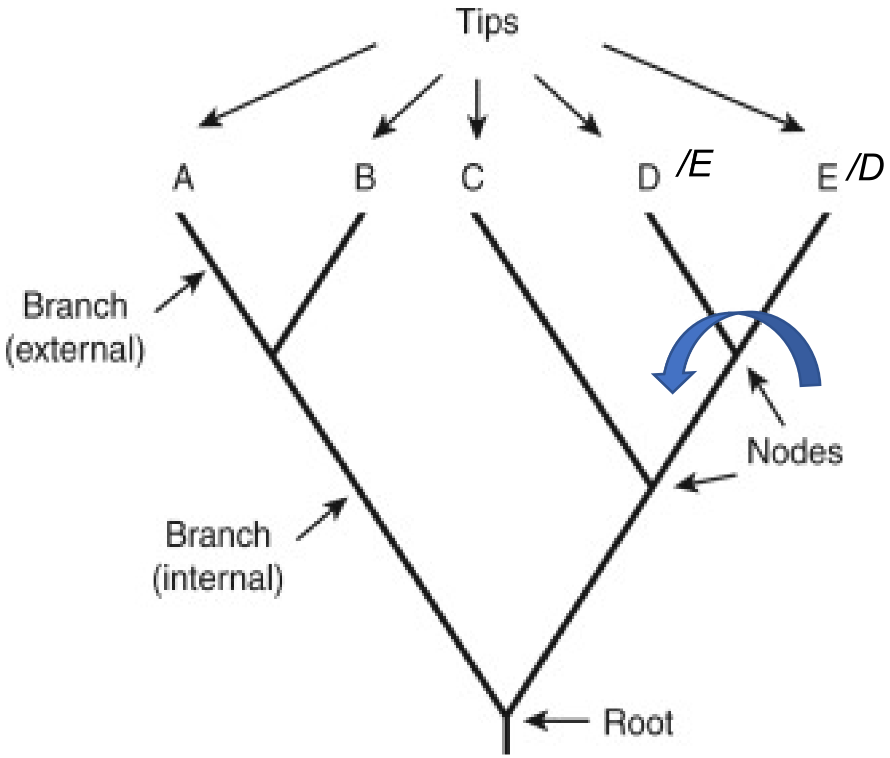
- Advantages
  - Lots of data (raw data)

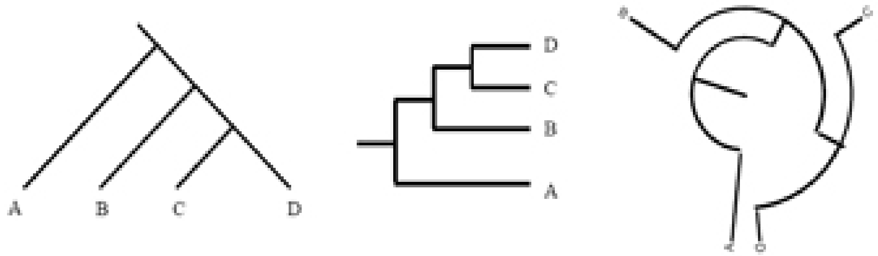
- Challenges
  - May be difficult to review as more differences accumulate or within large datasets
  - Interpretation

Communicating and reporting results

2. Phylogentic trees

Order of tips of trees has no meaning





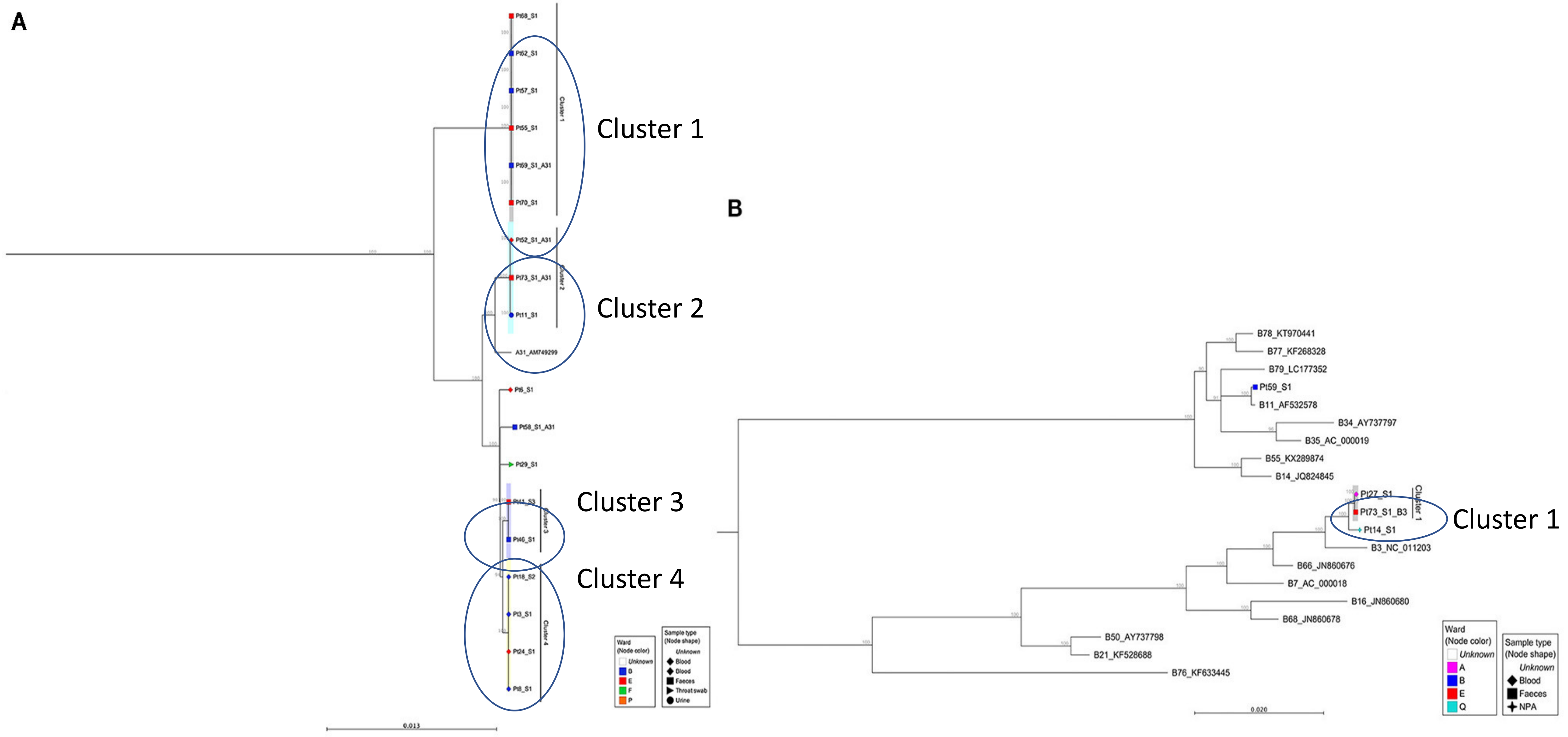
Multiple formats:
Equivalent relationship

- Advantages
  - Able to visualize relationship
  - Easier with larger datasets

- Challenges
  - Biases in creation and interpretation

https://www.applied-maths.com/applications/whole-genome-snp-analysis;
Scott and Baum, Encyclopedia of Evolutionary Biology, 2016

Communicating and reporting results– Phylogenetic tree



Myers *et al.,* Frontiers in Microbiol, 2021

Communicating and reporting results

3. Line list with assigned numbers

Case	Sample ID	Collection Date YYYY-MM-DD	Genome data	Genotype (if applicable)	Genomic Cluster Details	Clinical Cohort Details
1					A.1	1
2					A.2	1
3					B.1	1
4					A.1	2

- Advantages
  - Easy to convey relationship between isolates

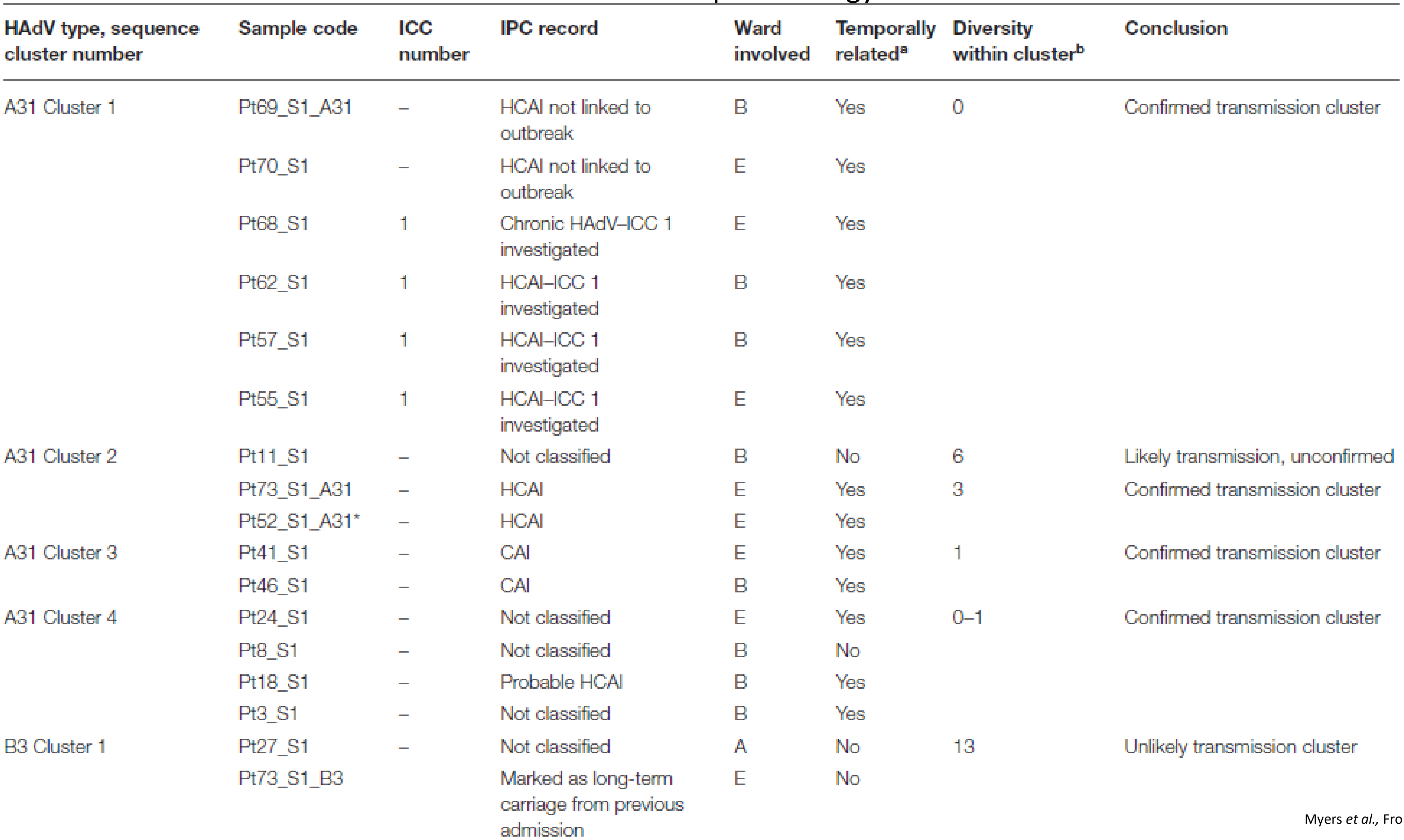
- Challenges
  - Relatedness defined when data created

Communicating and reporting results– Line list

Genetic data

Clinical epidemiology

Genetic data



HAdV type, sequence cluster number	Sample code	ICC number	IPC record	Ward involved	Temporally related[a]	Diversity within cluster[b]	Conclusion
A31 Cluster 1	Pt69_S1_A31	–	HCAI not linked to outbreak	B	Yes	0	Confirmed transmission cluster
	Pt70_S1	–	HCAI not linked to outbreak	E	Yes		
	Pt68_S1	1	Chronic HAdV–ICC 1 investigated	E	Yes		
	Pt62_S1	1	HCAI–ICC 1 investigated	B	Yes		
	Pt57_S1	1	HCAI–ICC 1 investigated	B	Yes		
	Pt55_S1	1	HCAI–ICC 1 investigated	E	Yes		
A31 Cluster 2	Pt11_S1	–	Not classified	B	No	6	Likely transmission, unconfirmed
	Pt73_S1_A31	–	HCAI	E	Yes	3	Confirmed transmission cluster
	Pt52_S1_A31*	–	HCAI	E	Yes		
A31 Cluster 3	Pt41_S1	–	CAI	E	Yes	1	Confirmed transmission cluster
	Pt46_S1	–	CAI	B	Yes		
A31 Cluster 4	Pt24_S1	–	Not classified	E	Yes	0–1	Confirmed transmission cluster
	Pt8_S1	–	Not classified	B	No		
	Pt18_S1	–	Probable HCAI	B	Yes		
	Pt3_S1	–	Not classified	B	Yes		
B3 Cluster 1	Pt27_S1	–	Not classified	A	No	13	Unlikely transmission cluster
	Pt73_S1_B3		Marked as long-term carriage from previous admission	E	No		

Myers *et al.,* Frontiers in Microbiol, 2021

Ongoing challenges for the use of WGS

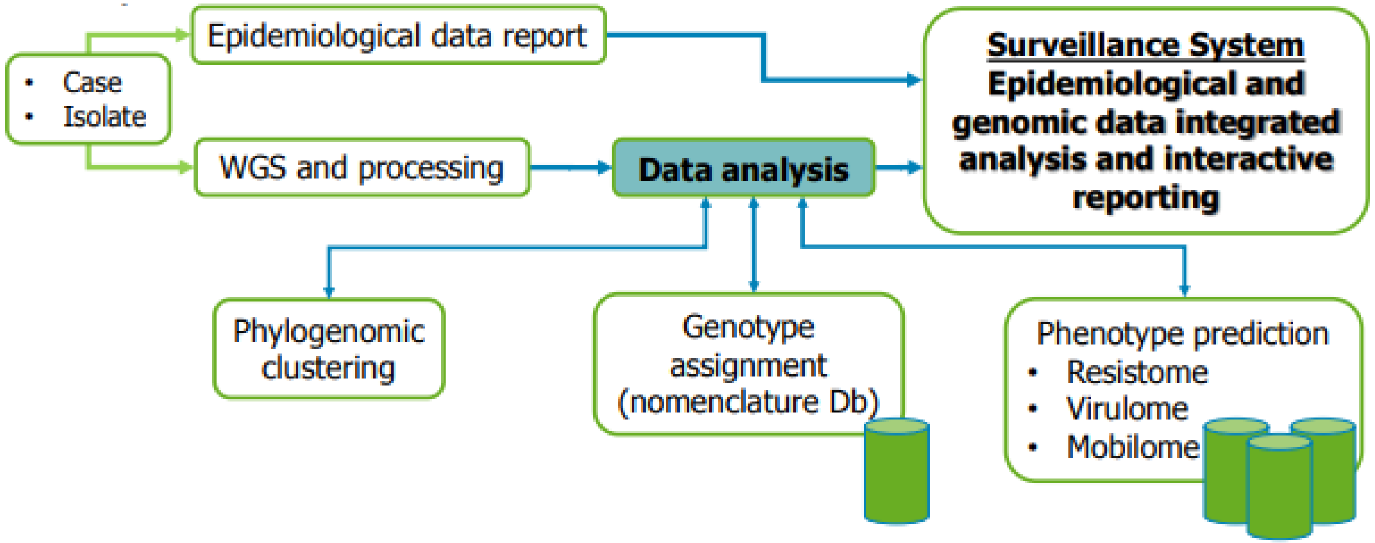
- Mixed populations/genotypes
    - Culture may select for strains that grow best
    - Within host diversity
        - Sequence depth may be insufficient to define or identify multiple strains

- **Standardization**
    - Sequencing depth and genome coverage
    - Quality of sequence data
    - Mutation rate/definition of relatedness

Ongoing challenges for the use of WGS

- Lab experience (wet lab/dry lab) and infrastructure
  - Technical experience may be applicable to the use of NGS for diagnosis

- Available databases/reference databases and **data sharing**



Epidemiological data report
• Case
• Isolate
WGS and processing
Data analysis
Surveillance System
Epidemiological and genomic data integrated analysis and interactive reporting
Phylogenomic clustering
Genotype assignment (nomenclature Db)
Phenotype prediction
• Resistome
• Virulome
• Mobilome

European Centre for Disease Prevention and Control. Expert opinion on whole genome sequencing for public health surveillance, 2016

Take home points

Whole genome sequencing is an important and powerful tool for infectious disease surveillance

Consider data resolution and result interpretation (relatedness)

The interpretation of genomic data is organism specific – clinical validation is essential

It is important to understand the best methods to communicate results with stakeholder group

Clinical epidemiological investigations and surveillance continues to be a fundamental component to infectious disease monitoring