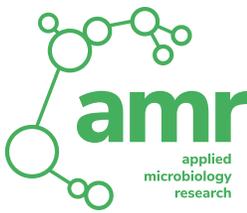




University of  
Zurich<sup>UZH</sup>

Institute of Medical Microbiology

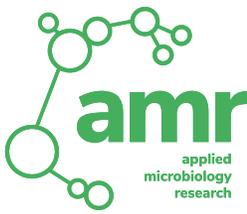


# Typing: Introduction to cgMLST

Helena Seth-Smith PhD

24.03.2023

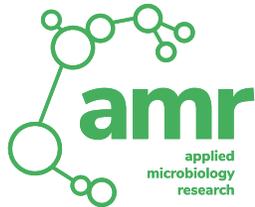
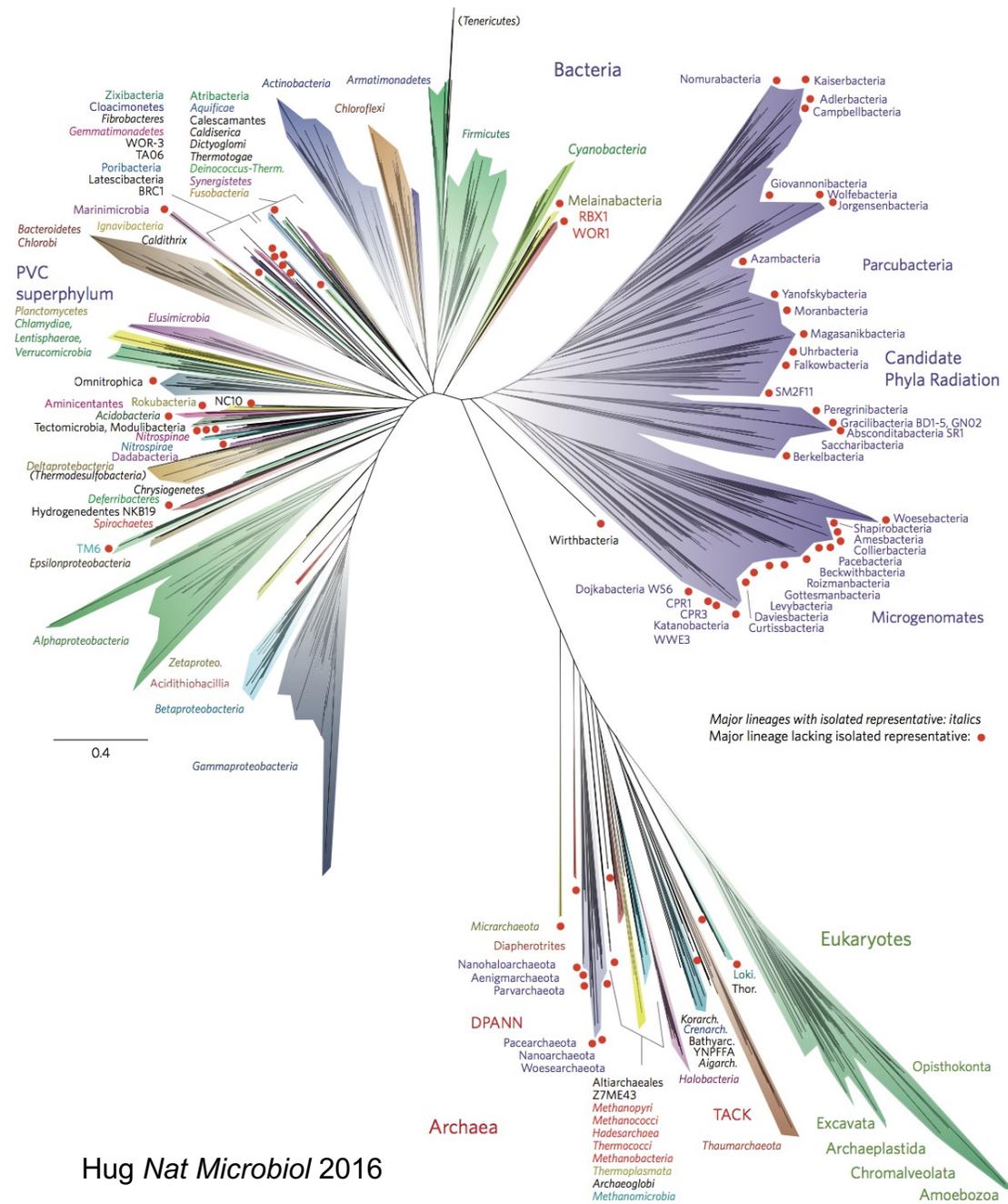
# Table of Contents



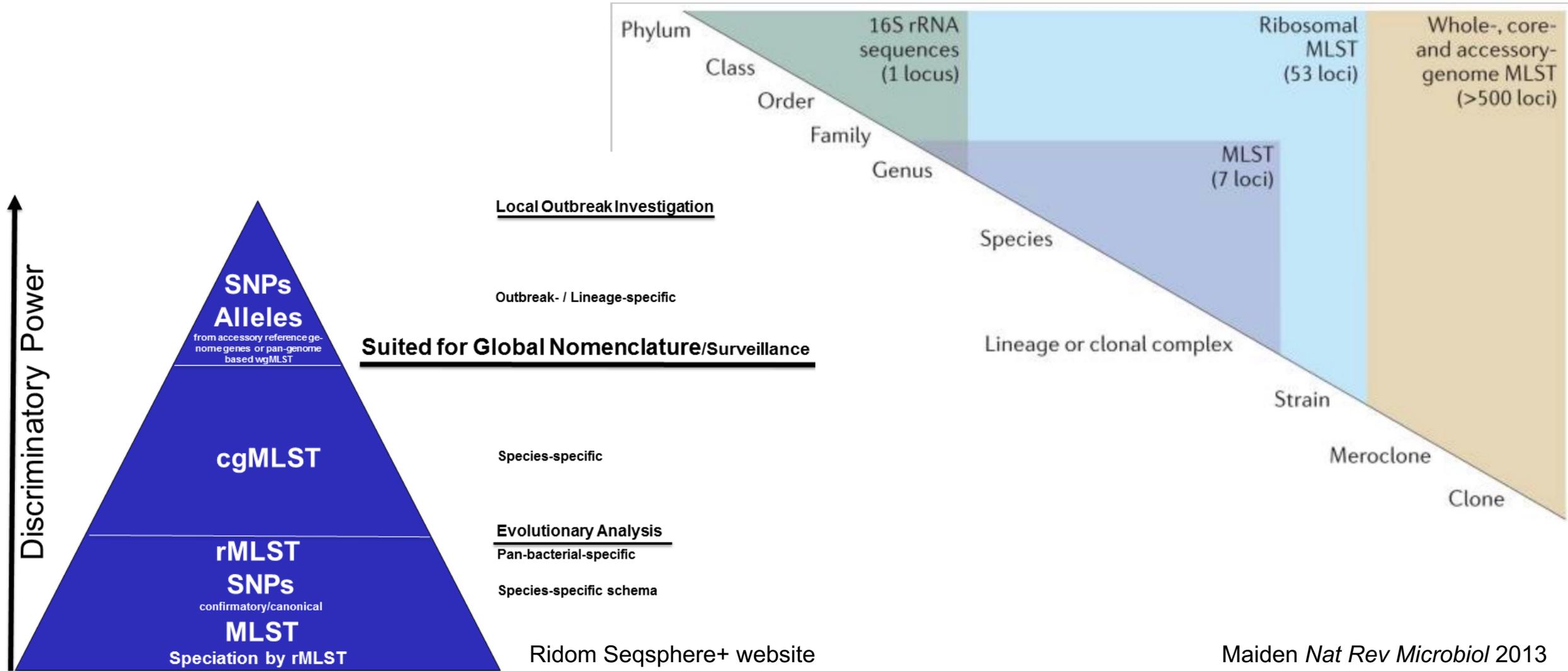
- Recap on resolution
- Ridom Seqsphere+ software and others
- Defined schemes
- Cluster thresholds / cut-offs
- Examples
- Summary

# Typing: why genomes?

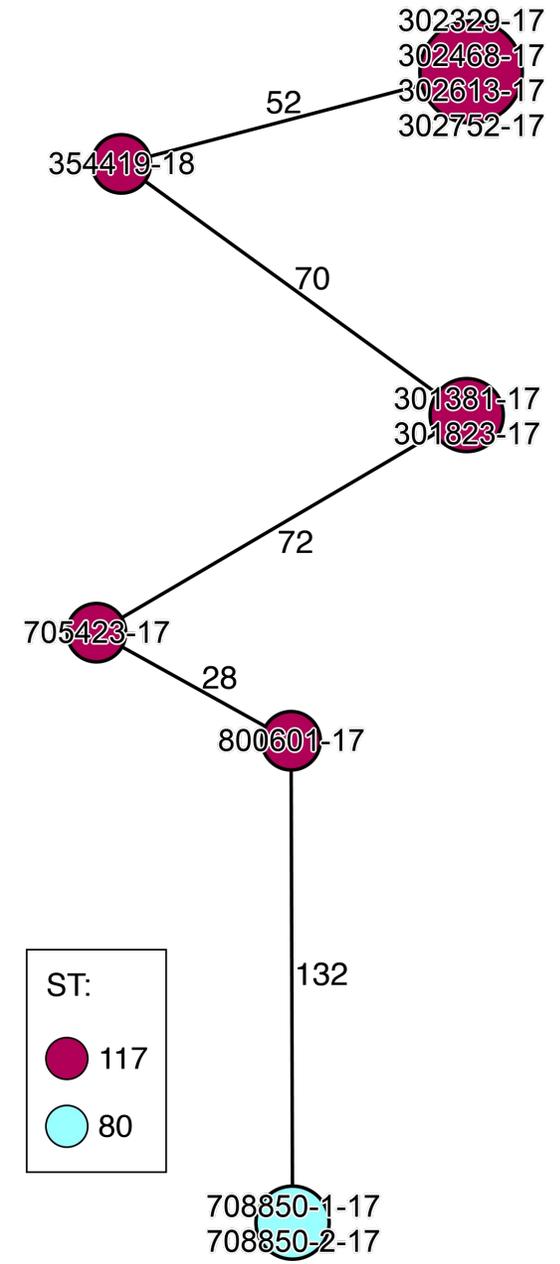
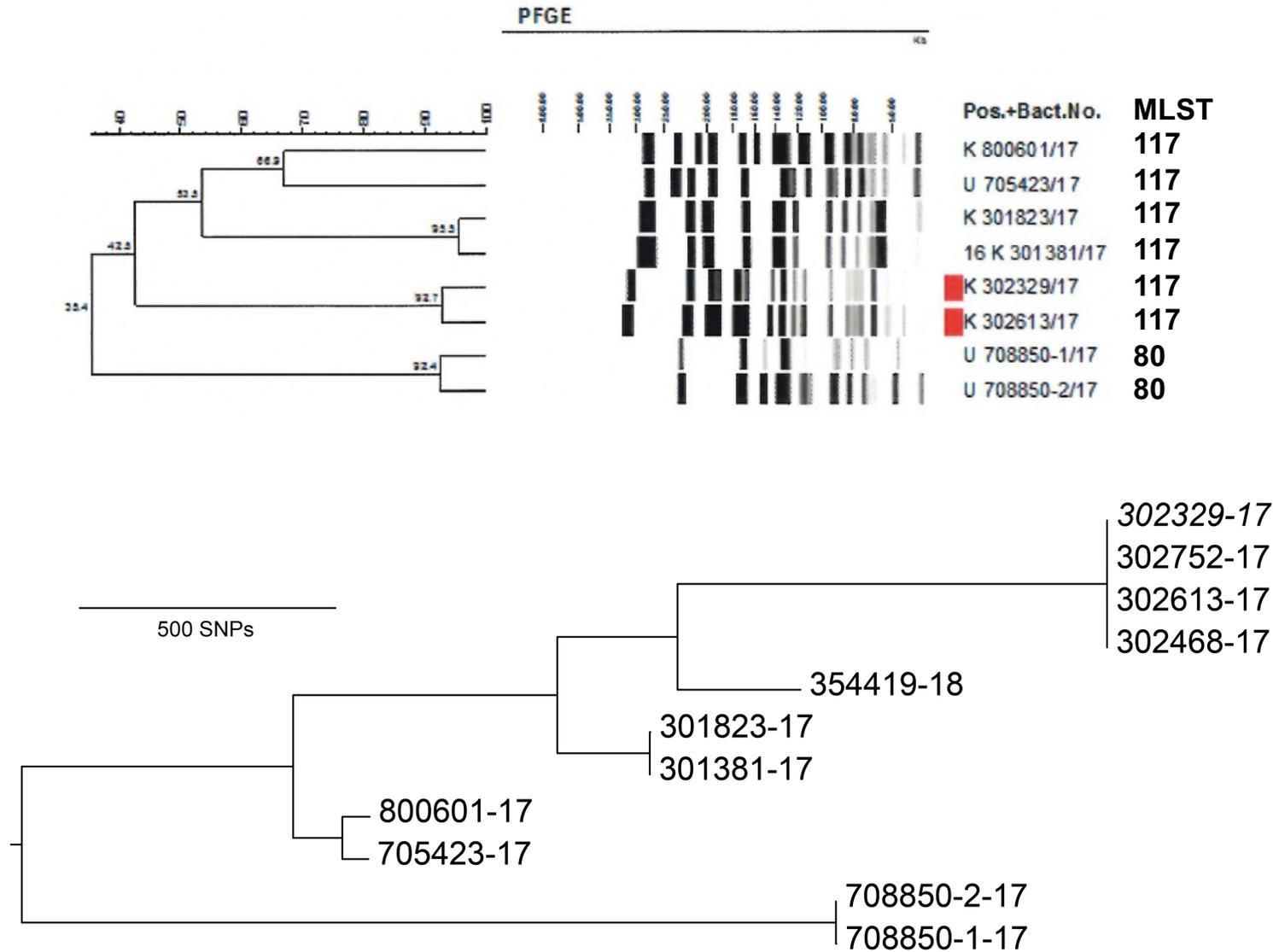
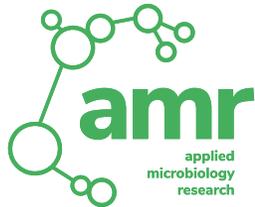
- Genomes **define** the ancestry and relatedness of isolates
- Accessing more information provides higher resolution



# WGS typing methods: resolution

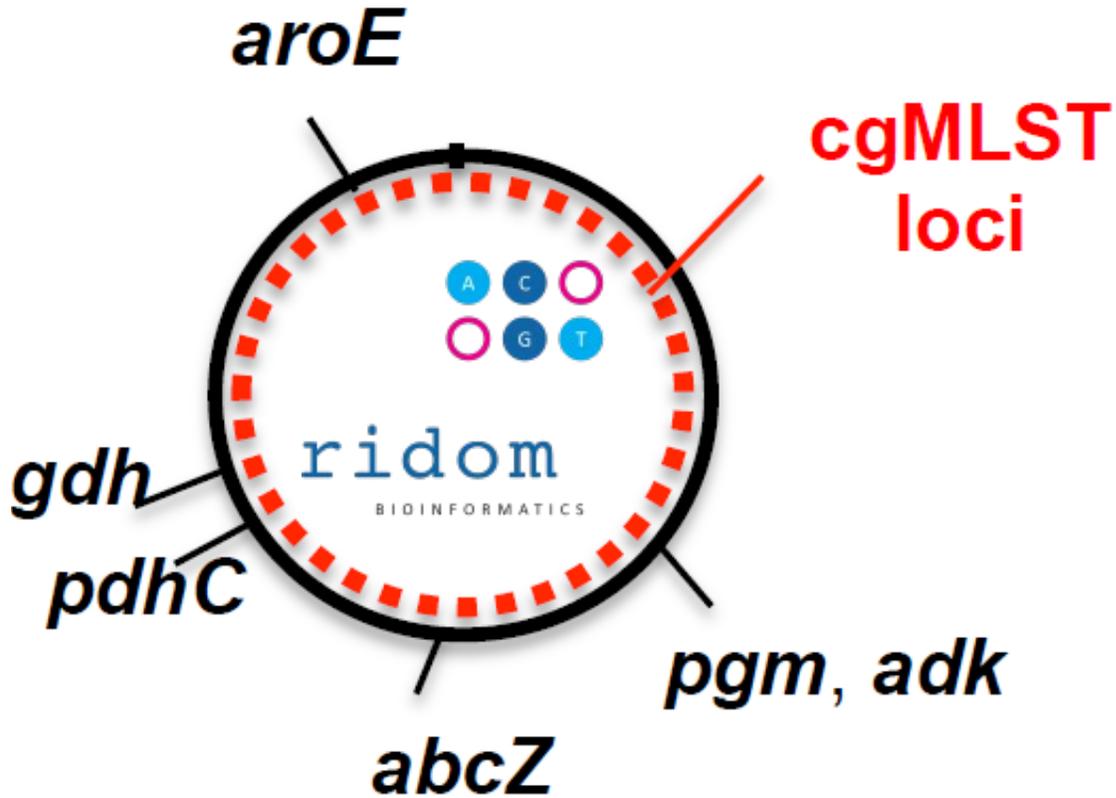


# Typing: increasing resolution

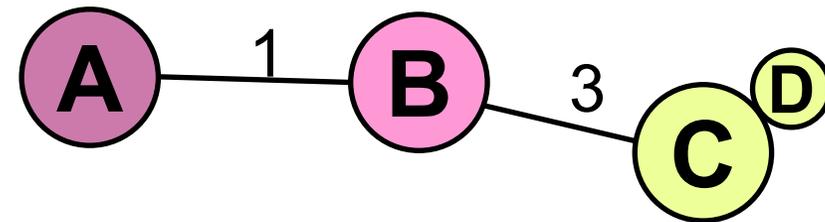


# Core Genome MLST: more data used, higher resolution

- Stable, defined schemes



	Gene 1	Gene 2	Gene 3	Gene 4	<del>Gene 5</del>
A	TCGAT	CGATG	TCGAAT	TGTCGA	<del>AAGCGA</del>
B	TCGAT	CGTTG	TCGAAT	TGTCGA	<del>AAGCGA</del>
C	GCGAT	CGTTG	TCTTAT	TGCGA	
D	GCGAT	CGTTG	TCTTAT	TGCGA	<del>AAGCGA</del>



# Ridom Seqsphere+: Commercial licensed software

Defined, reproducible, results

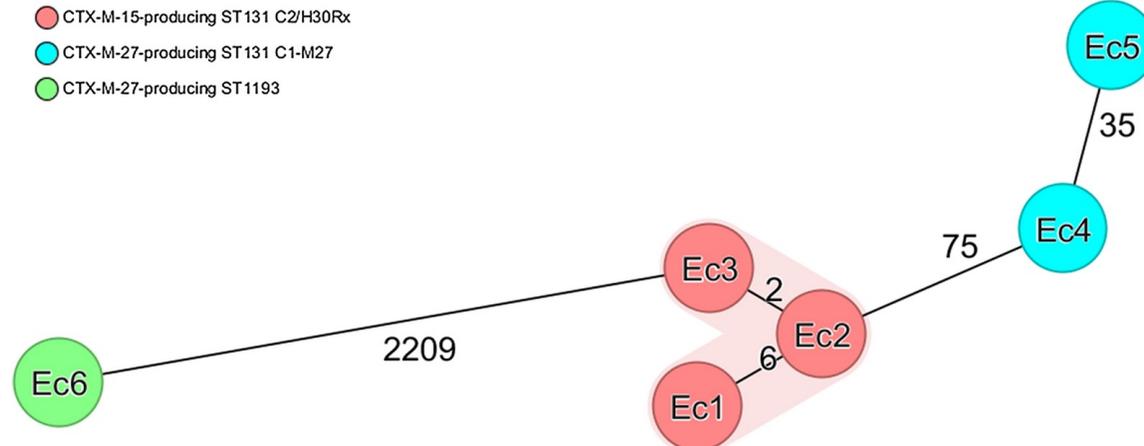
Based on published, justified schemes

Following in-house validation compared to previous "Gold Standard"

Required for ISO accreditation



ridom  
BIOINFORMATICS



ISO 17025

Technically competent to  
produce accurate and  
reliable data

## Other softwares for core genome analysis

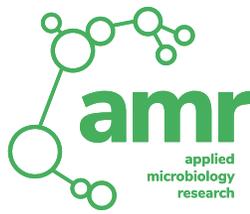
For stable, species-wide, defined schemes, and creation of new schemes:

- chewBBACA (open source)
- MentaLiST (open source)

For *ad hoc*, dataset specific schemes including pan genome analysis:

- Roary/Scory (open source)
- Panaroo (open source)
- Can also use accessory genes to correlate with known phenotypes  
(bacterial Genome Wide Association Study, bGWAS)

# Defined schemes



Many published schemes, where core genomes have been defined

To develop a scheme:

- Annotated seed genome
  - Many query genomes
  - Searches for genes in all query genomes, >90% identity, 100% length
  - Remove outlier genomes
  - (Remove plasmid genes)
  - Generate the core genome scheme
  - (Curate by removing repetitive / phase variable genes)
- 
- Can also make own *ad hoc* schemes for species not in the list using above guidelines

The screenshot shows the website <https://www.cgmlst.org/nsc>. The page title is "cgMLST.org Nomenclature Server". Below the title, there is a paragraph explaining the server's purpose: "This server controls the allelic nomenclature of core genome MLST (cgMLST) bacterial gene schemes. Currently submission of new alleles and optional metadata is only possible by use of the SeqSphere+ software. A cgMLST scheme is a fixed and agreed upon number of genes for each species or group of closely related species that is ideally suited to standardize whole genome sequencing (WGS) based bacterial genotyping. By cgMLST very closely related genomes are 'lumped' together in a **Complex Type** (CT). In addition, this server controls the allelic nomenclature of the **accessory genes** of the species seed genomes." Below this is a link to a privacy policy. The main content is a table with three columns: "Scheme", "Target Count", and "Strain Count".

Scheme	Target Count	Strain Count
<a href="#">Acinetobacter baumannii cgMLST</a>	2,390	5,366
<a href="#">Brucella melitensis cgMLST</a>	2,704	88
<a href="#">Clostridioides difficile cgMLST</a>	2,270	4,972
<a href="#">Enterococcus faecalis cgMLST</a>	1,972	1,605
<a href="#">Enterococcus faecium cgMLST</a>	1,423	11,156
<a href="#">Escherichia coli cgMLST</a>	3,152	1,765
<a href="#">Francisella tularensis cgMLST</a>	1,147	240
<a href="#">Klebsiella pneumoniae/variicola/quasipneumoniae cgMLST</a>	2,358	7,374
<a href="#">Legionella pneumophila cgMLST</a>	1,521	900
<a href="#">Listeria monocytogenes cgMLST</a>	1,701	20,654
<a href="#">Mycobacterium tuberculosis/bovis/africanum/canettii cgMLST</a>	2,891	36,484
<a href="#">Mycoplasma gallisepticum cgMLST</a>	425	78
<a href="#">Staphylococcus aureus cgMLST</a>	1,861	25,305

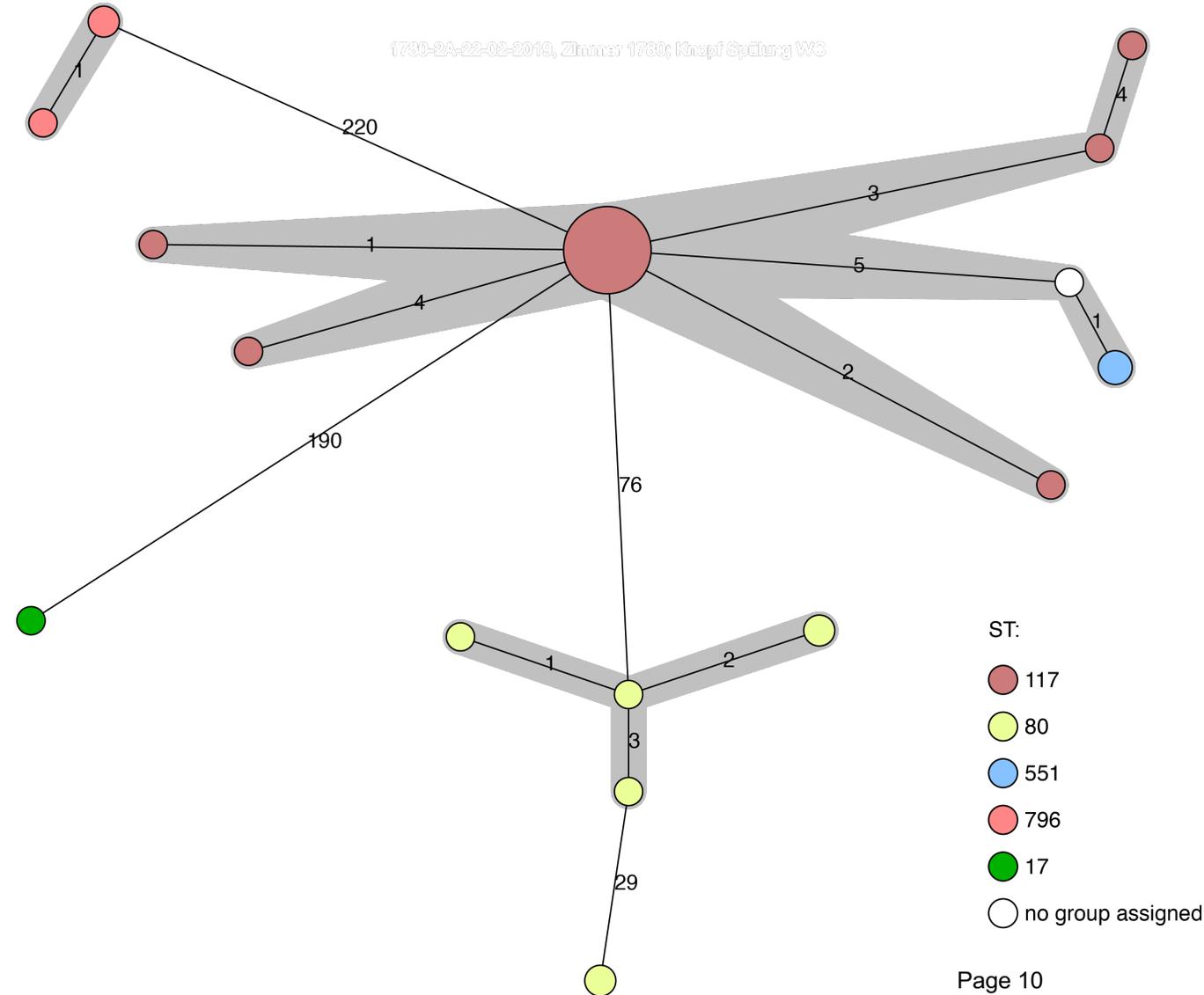
# Cluster thresholds / cut-offs / limits

Cluster thresholds are a constant issue when investigating potential transmissions and outbreaks

Our answer is always that you need to take into account many factors:

- Mutation rate, ie species dependent
- Time between samples
- Genome dynamics (recombination; presence / absence of other mobile elements)

Even when defined, all findings need to be considered in conjunction with epidemiological / metadata



# Species specific clustering distance thresholds

Many published schemes, where core genomes have been defined

To develop a scheme:

- Annotated seed genome
- Many query genomes
- Searches for genes in all query genomes, >90% identity, 100% length
- Remove outlier genomes
- (Remove plasmid genes)
- Generate the core genome scheme
- (Curate by removing repetitive / phase variable genes)
- Run on own set of genomes to define cutoffs
- (Check reproducibility using technical replicates)
- (Publish)
- Data on cgMLST.org

Published Scheme	Distance threshold
<i>Acinetobacter baumannii</i>	9
<i>Bacillus anthracis</i>	5
<i>Brucella melitensis</i>	6
<i>Burkholderia pseudomallei</i>	5
<i>Campylobacter jejuni/coli</i>	13
<i>Clostridioides difficile</i>	6
<i>Clostridium perfringens</i>	60
<i>Enterococcus faecalis</i>	7
<i>Enterococcus faecium</i>	20
<i>Escherichia coli</i>	10
<i>Francisella tularensis</i>	1
<i>Klebsiella pneumoniae/variicola/quasipneumoniae</i>	15
<i>Legionella pneumophila</i>	4
<i>Listeria monocytogenes</i>	10
<i>Mycobacterium tuberculosis/bovis/africanum/canettii</i>	12
<i>Mycoplasma gallisepticum</i>	10
<i>Paenibacillus larvae</i>	10
<i>Pseudomonas aeruginosa</i>	12
<i>Salmonella enterica</i>	7
<i>Staphylococcus aureus</i>	24

# How do we interpret the analysis?

Different species have different genome dynamics: different lifestyles, replication rates, recombination rates, mutations rates....

Also technical issues related to assemblies over repeat genes, coverage etc...

How reliable are the cutoffs?

Publication and validation quality varies  
- No real consensus on optimum

Cutoff values based on local outbreaks and added epi knowledge  
- Sometimes small number of genomes  
- In-host variation (over time) rarely addressed

**Table 1**  
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

Organism	Relatedness threshold <sup>a</sup>		References
	wg/cgMLST (allele)	SNPs	
<i>Acinetobacter baumannii</i>	≤8	≤3	[25,26]
<i>Brucella</i> spp.	Epidemiologic validation in progress <sup>b</sup>		<a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a>
<i>Campylobacter coli</i> , <i>C. jejuni</i>	≤14	≤15	[27,28]
<i>Cronobacter</i> spp.	Epidemiologic validation in progress <sup>b</sup>		<a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a>
<i>Clostridium difficile</i>	Epidemiologic validation in progress <sup>b</sup>	≤4	[29], <a href="http://www.cgmlst.org/ncs">http://www.cgmlst.org/ncs</a> , <a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a>
<i>Enterococcus faecium</i>	≤20	≤16	[30]
<i>Enterococcus raffinosus</i>	Epidemiologic validation in progress <sup>b</sup>		<a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a>
<i>Escherichia coli</i>	≤10	≤10	[31,32], <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a>
<i>Francisella tularensis</i>	≤1	≤2	[33,34]
<i>Klebsiella oxytoca</i>	Epidemiologic validation in progress <sup>b</sup>		<a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a>
<i>Klebsiella pneumonia</i>	≤10	≤18	[35,36]
<i>Legionella pneumophila</i>	≤4	≤15	[37]
<i>Listeria monocytogenes</i>	≤10	≤3	[38,39]
<i>Mycobacterium abscessus</i>		≤30	[40]
<i>Mycobacterium tuberculosis</i>	≤12	≤12	[41]
<i>Neisseria gonorrhoeae</i>	Epidemiologic validation in progress <sup>b</sup>	≤14	[42], <a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a>
<i>Neisseria meningitidis</i>	Epidemiologic validation in progress <sup>b</sup>		<a href="http://www.cgmlst.org/ncs">http://www.cgmlst.org/ncs</a>
<i>Pseudomonas aeruginosa</i>	≤14	≤37	[31,43]
<i>Salmonella dublin</i>	Epidemiologic validation in progress <sup>b</sup>	≤13	[44], <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a>
<i>Salmonella enterica</i>	Epidemiologic validation in progress <sup>b</sup>	≤4	[45], <a href="http://www.cgmlst.org/ncs">http://www.cgmlst.org/ncs</a> , <a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a> , <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a>
<i>Salmonella typhimurium</i>	Epidemiologic validation in progress <sup>b</sup>	≤2	[46], <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a>
<i>Staphylococcus aureus</i>	≤24	≤15	[47,48]
<i>Streptococcus suis</i>		≤21	[49]
<i>Vibrio parahaemolyticus</i>	≤10		[50]
<i>Yersinia</i> spp.	0		[51]

cg, core genome; MLST, multilocus sequence typing; SNP, single nucleotide polymorphism; wg, whole genome.

<sup>a</sup> Data often represent single studies that can be used to begin formulation of species-specific interpretation criteria. Thus, these data should be coupled with newly published similar studies to ensure that resulting values are not atypical and can be generally applied.

<sup>b</sup> Proposed wg/cgMLST schemes are available online (<http://www.cgmlst.org/ncs>, <http://www.applied-maths.com/applications/wgmlst>, <https://enterobase.warwick.ac.uk/>) but as yet have not been epidemiologically validated.

# Checking schemes in an ideal world

Large datasets including outbreaks and sporadic cases

In-host variation over **time**

Outbreaks (also over **time**)

- bearing in mind foodborne or environmental source single point case clusters / outbreaks

Technical controls:

- repeat sequencing of same strain, also on different platforms
- repeat assembly of same data

Other ideas???

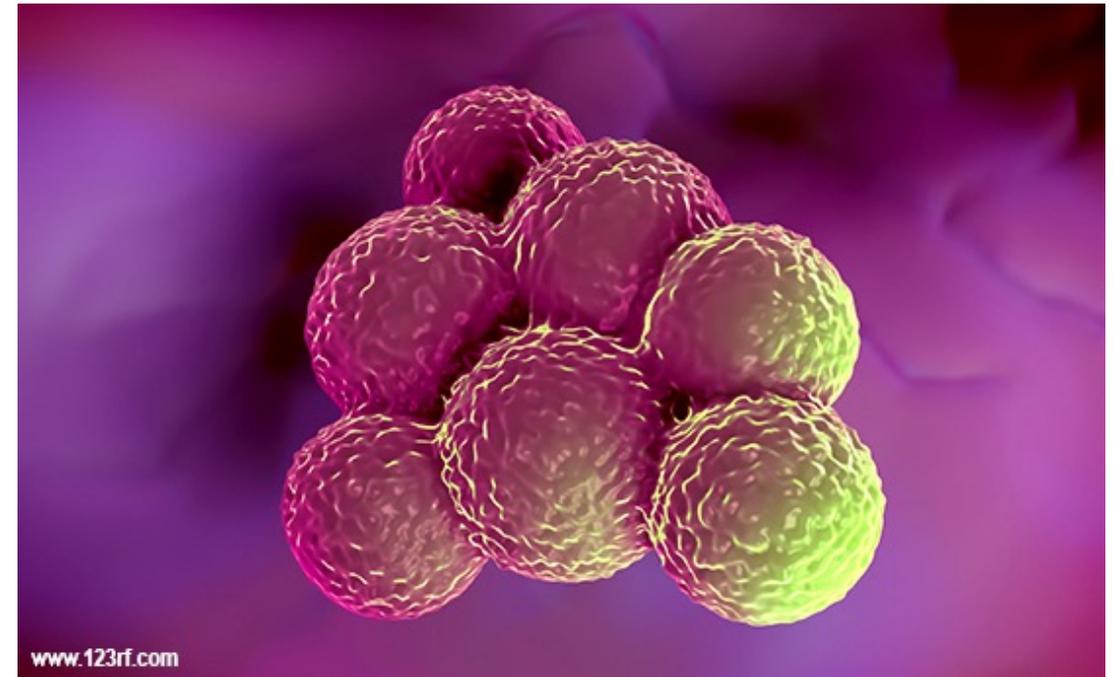
# *Staphylococcus aureus* – cutoff 24

Good scheme, stable and used as NGS control

In-host diversity quite high, but often sampled long-term

Most cases seen are independent

Direct transmissions are typically far below cut-off



# *Pseudomonas aeruginosa* – cutoff 12

Large, complicated genome

Valid scheme recently published

Technical replicates not identical, but within cutoff

Only a handful of putative transmissions identified

- followed by epidemiological investigation

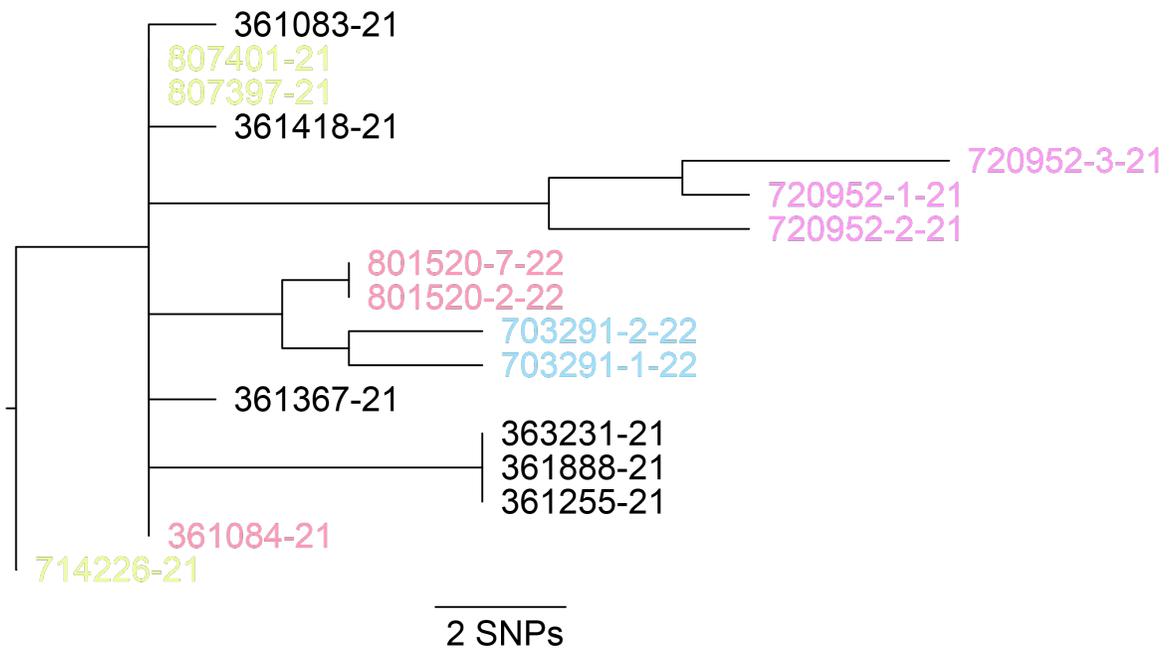
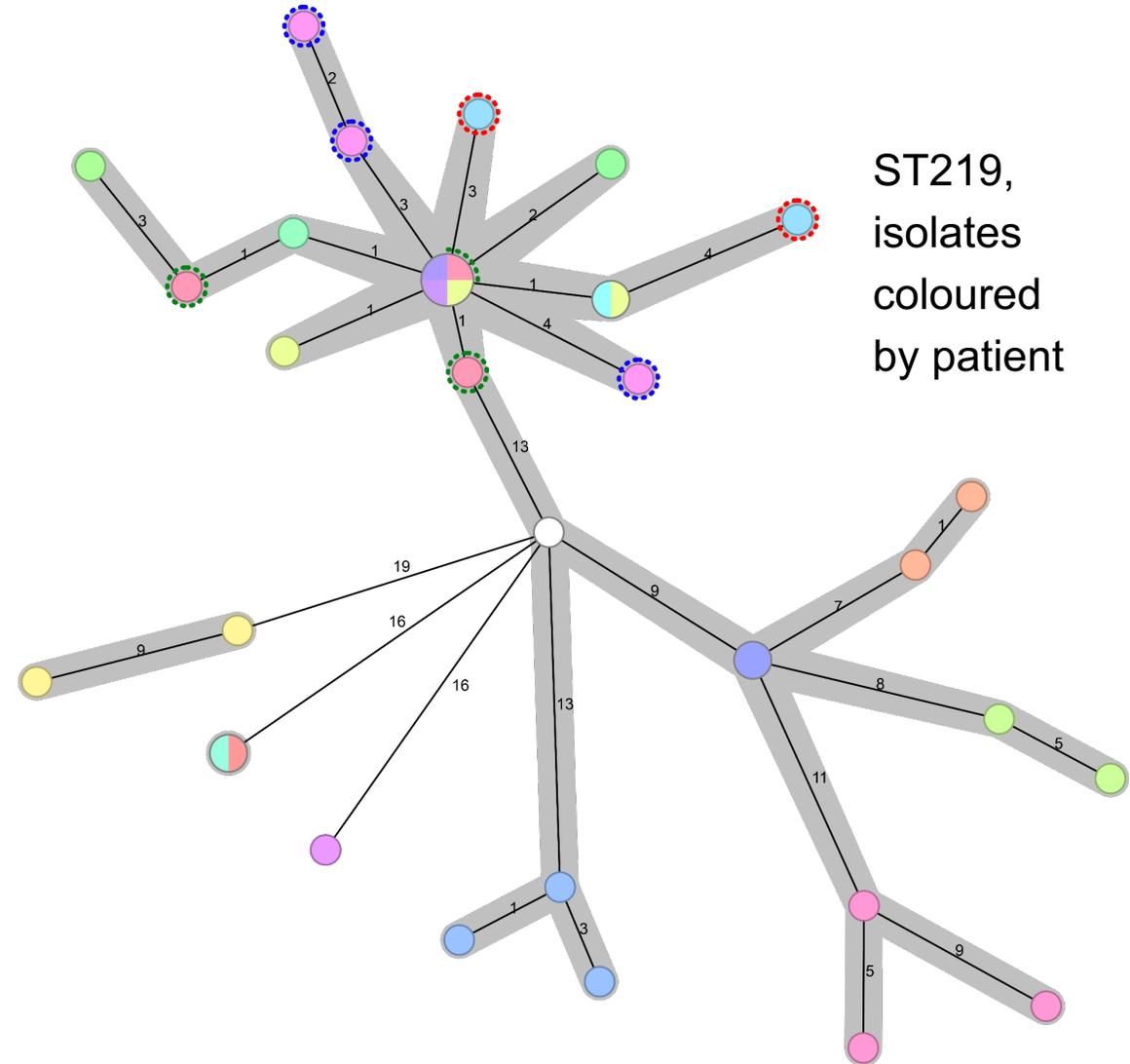


# *Klebsiella pneumoniae* / *variicola* – cutoff 15

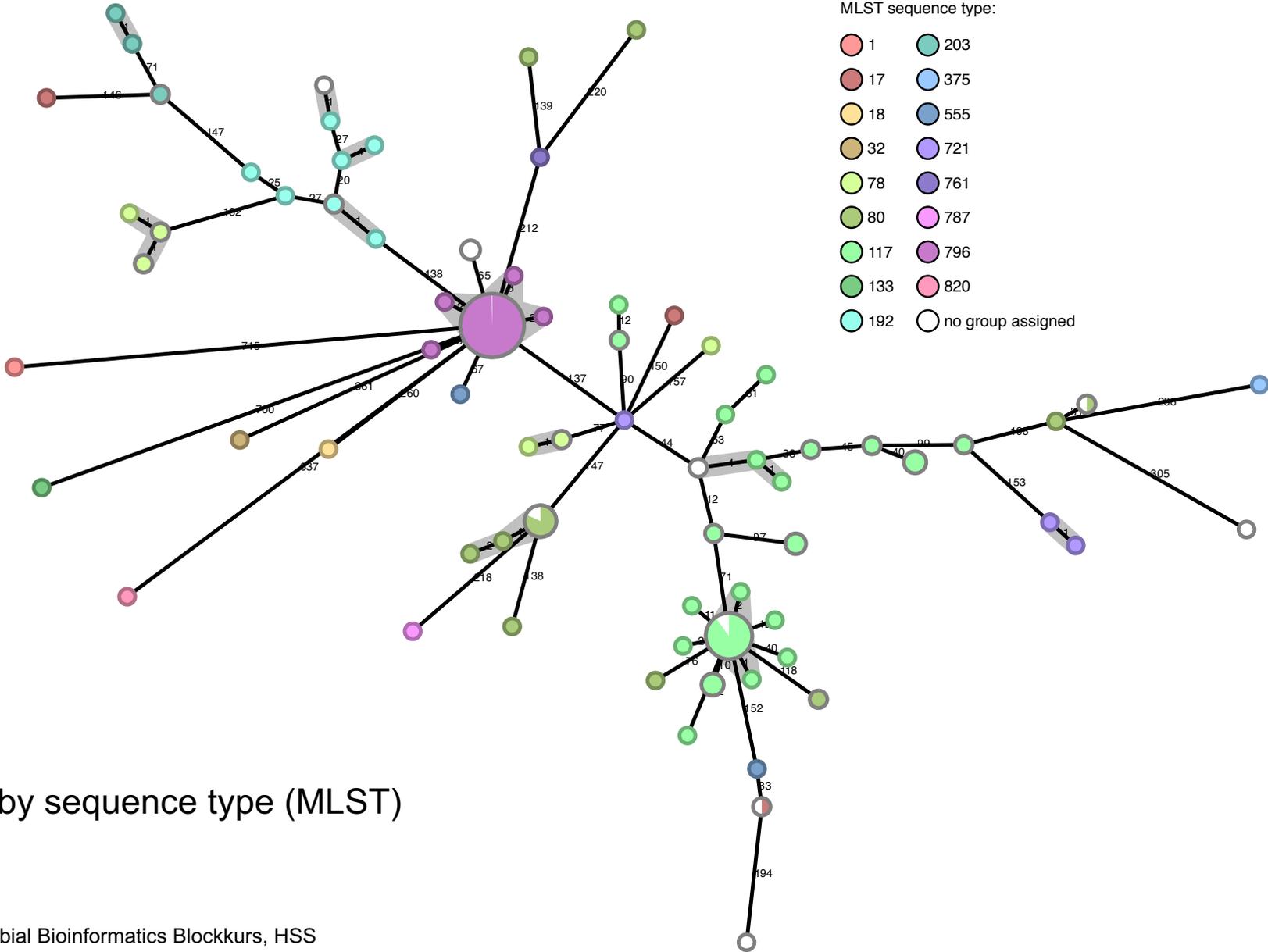
Replicate sequences not identical, but within cutoff

In-host diversity can be high (but within cutoff)

Within clusters, things can look messy:  
isolates from same patients do not cluster  
but in a SNP tree they do

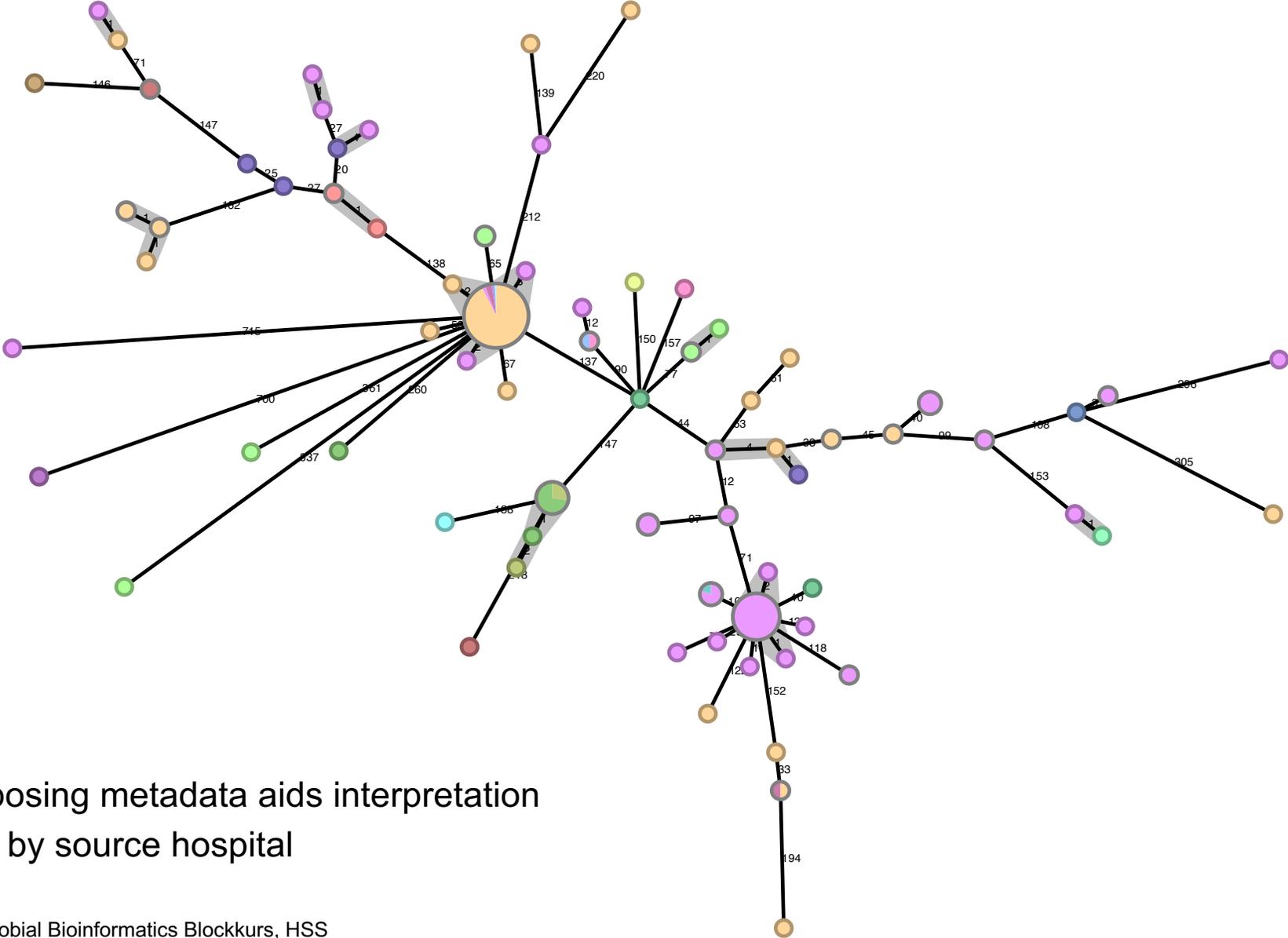


# Visualisation example: VRE: minimum spanning tree



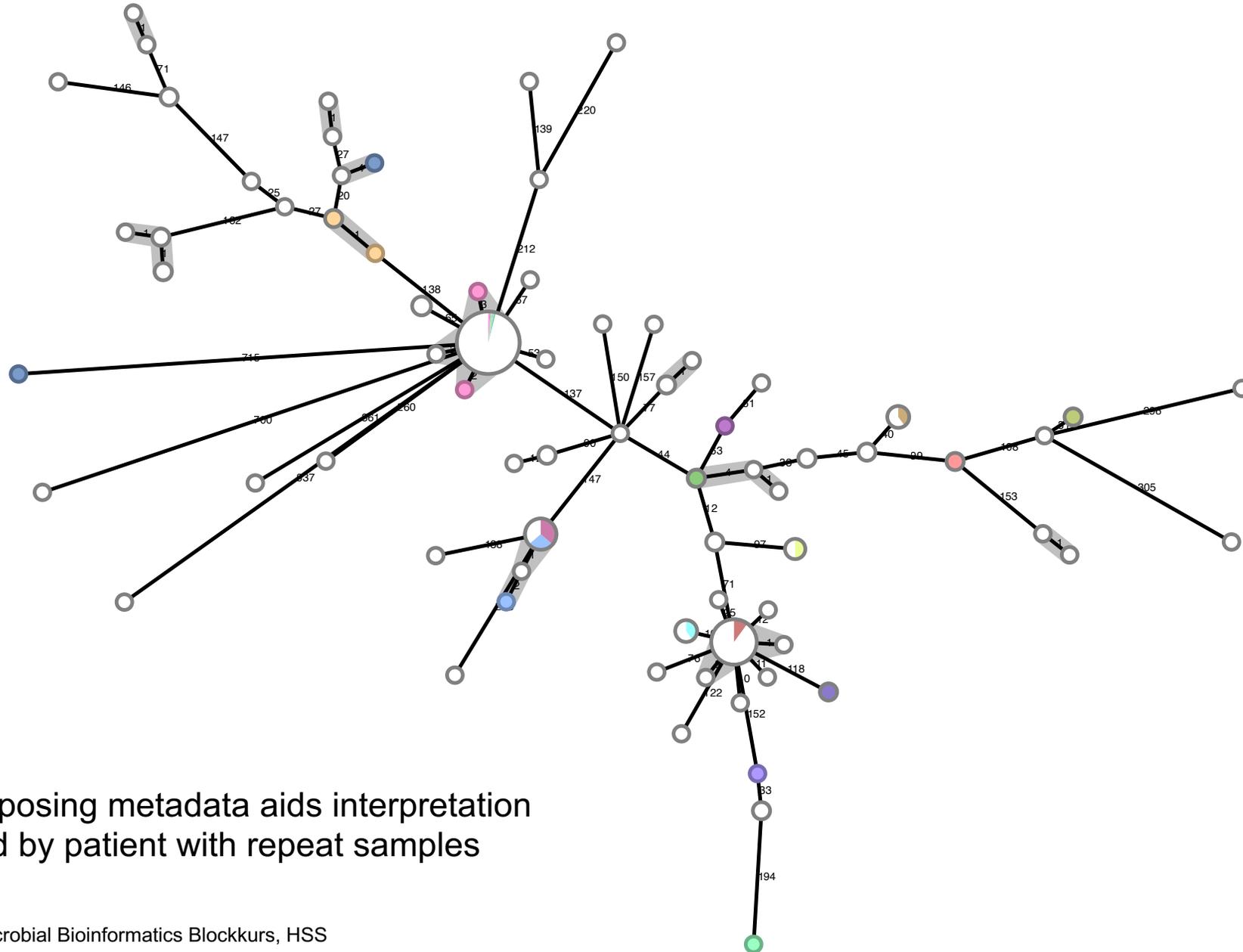
Coloured by sequence type (MLST)

# Visualisation example: VRE: minimum spanning tree



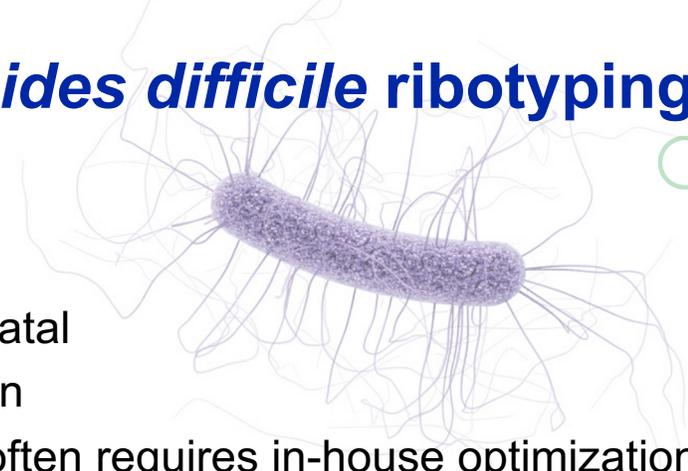
Superimposing metadata aids interpretation  
Coloured by source hospital

# Visualisation example: VRE: minimum spanning tree

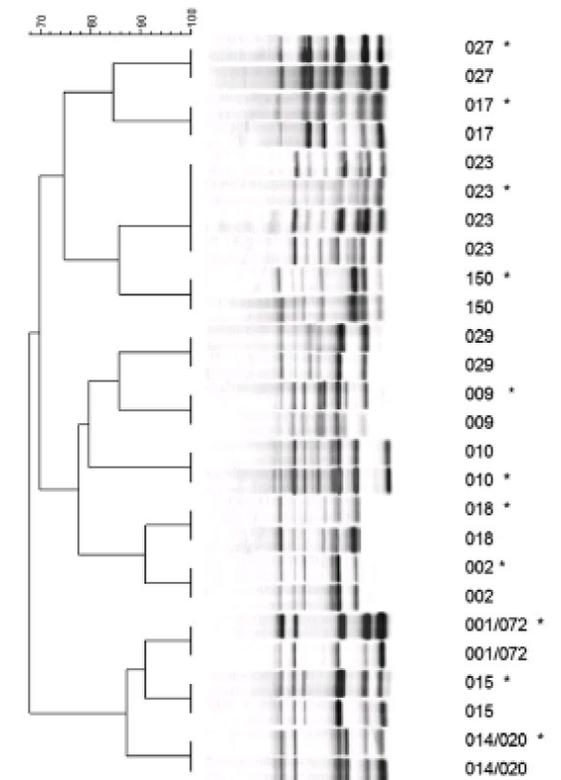
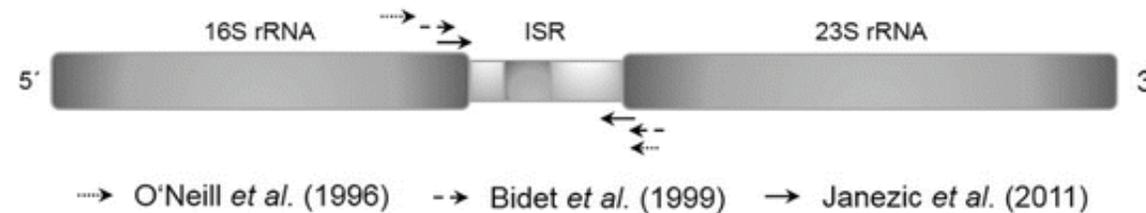
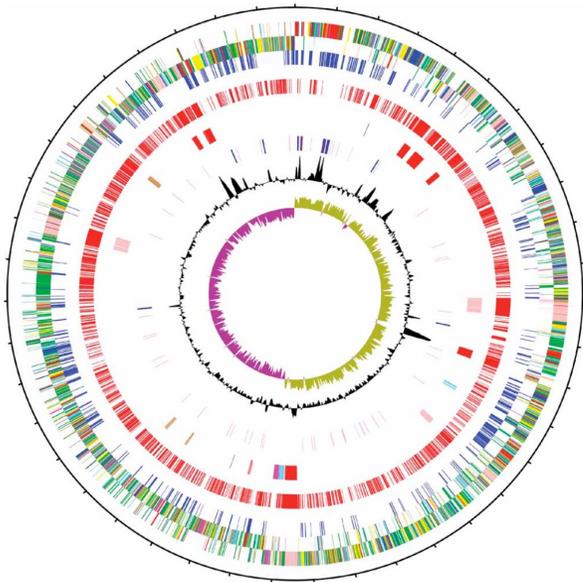


Superimposing metadata aids interpretation  
Coloured by patient with repeat samples

# Correlation with conventional typing: *Clostridioides difficile* ribotyping



- Important nosocomial pathogen
- Asymptomatic -> *C. difficile* infection (CDI) -> severe colitis, sepsis, fatal
- PCR-ribotyping commonly used typing tool: good strain discrimination
- BUT not fully portable between laboratories, labour intensive, slow, often requires in-house optimization
- Hypervirulent lineages have been defined by ribotype: RT027 and RT078
- What is the connection between ribotype, virulence and genome?

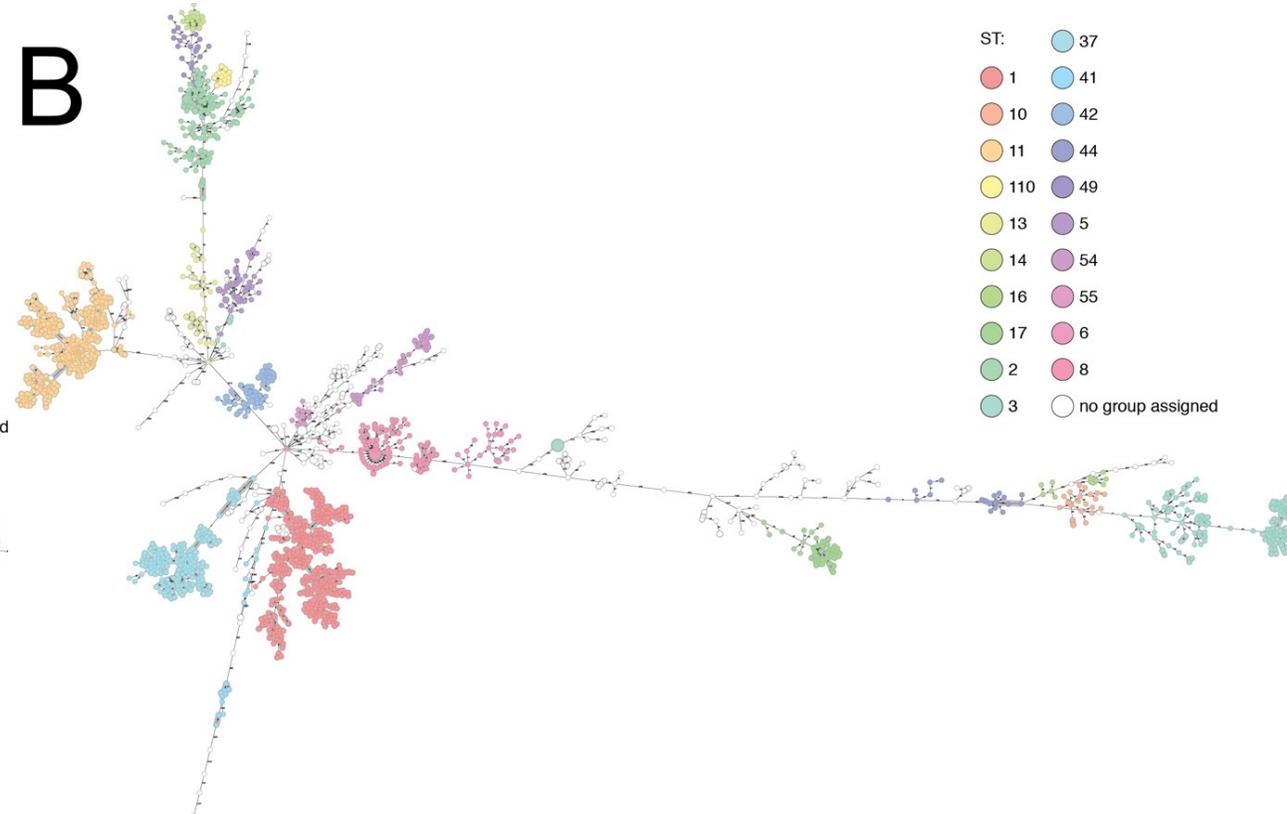
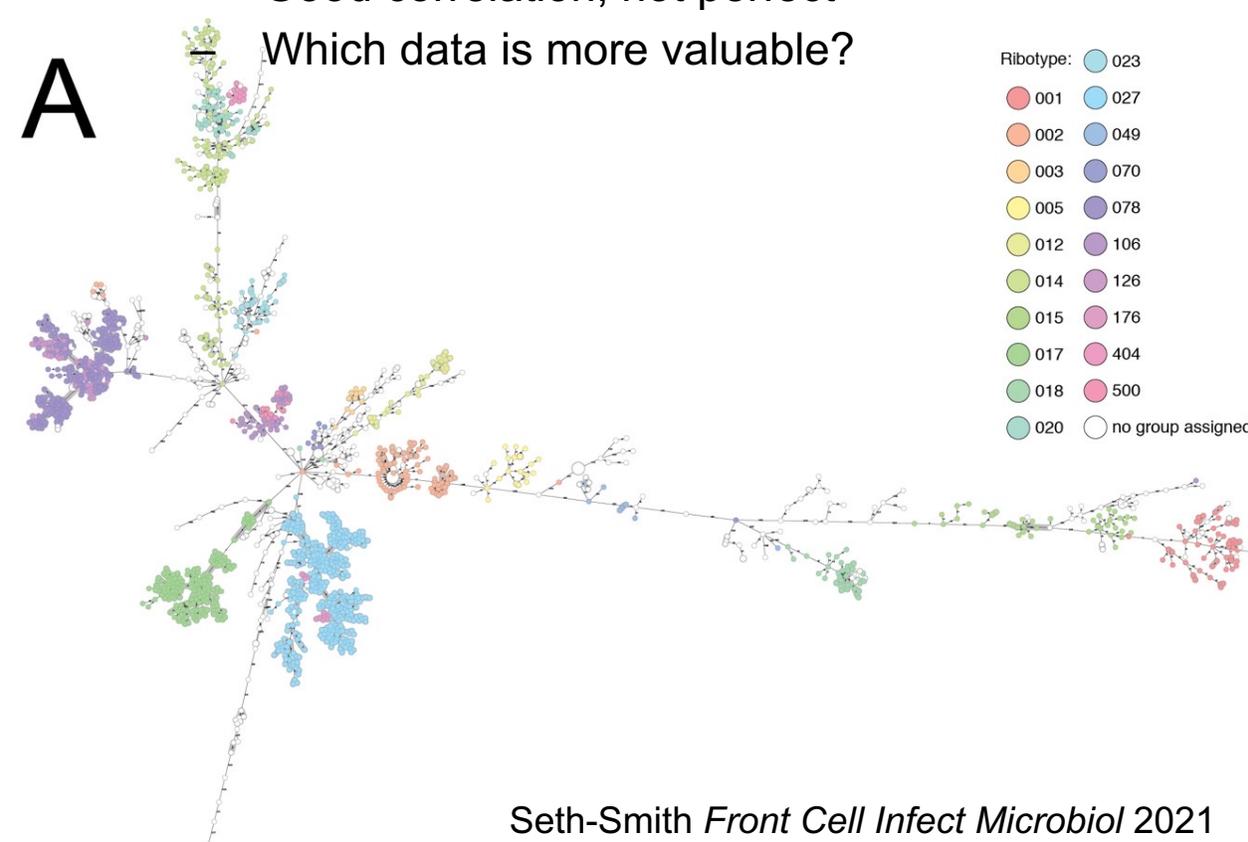


Sebahia *Nat Gen* 2006; Janezic *Clostridium difficile: Methods and Protocols* 2016; *C. difficile* *Methods and Protocols*, Springer

Fig. 2 Comparison of PCR-ribotyping patterns obtained from total stool DNA (marked with \*) and reference strains using primers and protocol described in Janezic et al. [20] and in this chapter

# Correlation with conventional typing: *Clostridioides difficile* ribotyping

- Ridom Seqsphere+ : 2270 targets, official scheme, based on reference strain 630
- n=2094 genomes (294 Basel collection); defined cluster limit 7 cutoff of 6 alleles
- n=141 RTs; n=118 STs
- Good correlation, not perfect
- Which data is more valuable?



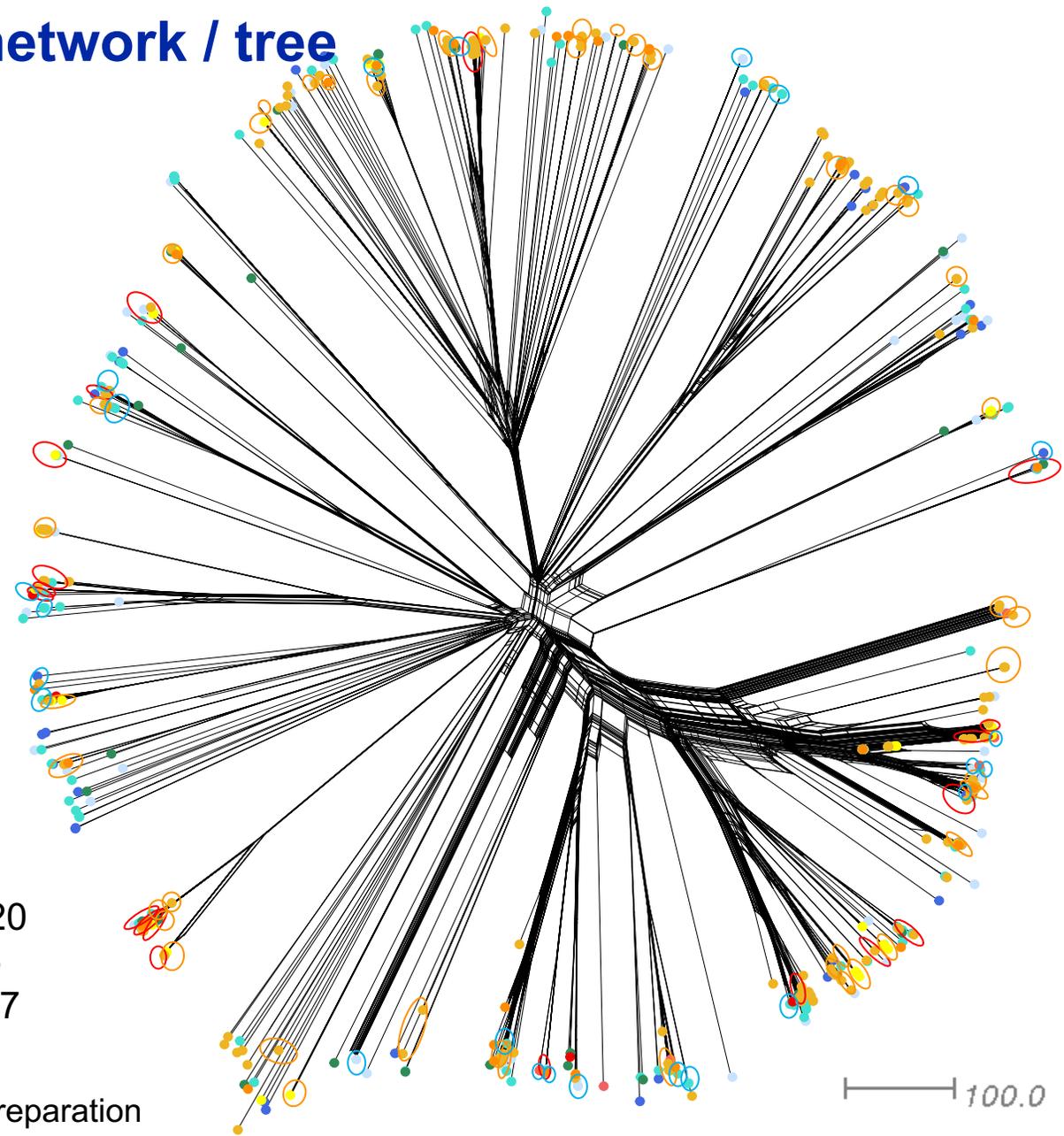
Seth-Smith *Front Cell Infect Microbiol* 2021

# *Campylobacter jejuni/coli* in Basel: network / tree

- Samples from patients and chicken meat
- PubMLST cgMLST scheme of 1343 loci
- Displayed as a Splitstree

- Stool isolates 2015
- Stool isolates 2016
- Stool isolates 2017
- Stool isolates 2018
- Blood isolates 2015
- Blood isolates 2016
- Blood isolates 2017
- Blood isolates 2018
- Chicken meat isolates 2015
- Chicken meat isolates 2016
- Chicken meat isolates 2017
- Chicken meat isolates 2018

- Transmission cluster n= 20
- Patient only cluster n= 26
- Chicken only cluster n= 47

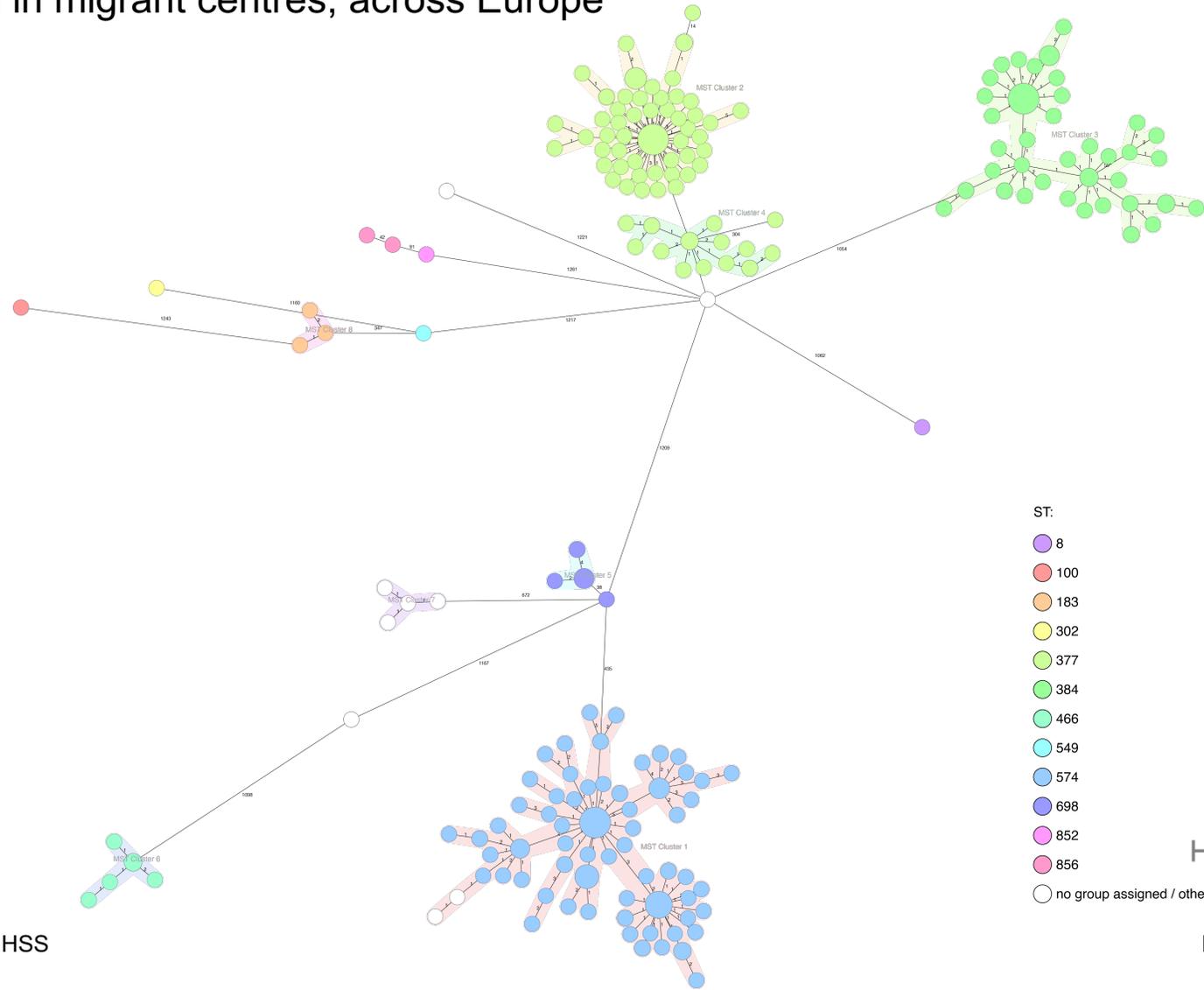
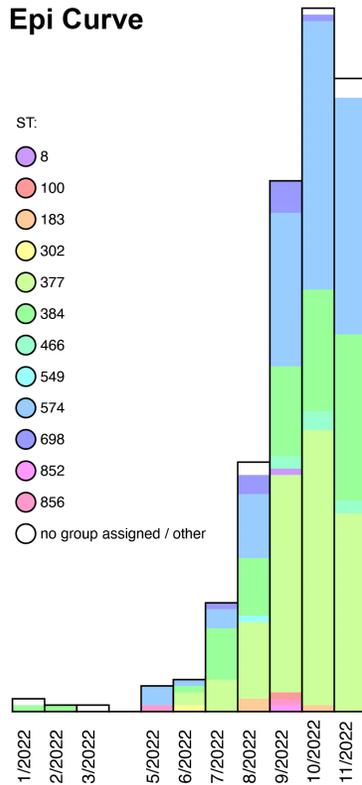


Seth-Smith *et al* in preparation

# Corynebacterium diphtheriae and cgMLST resolution

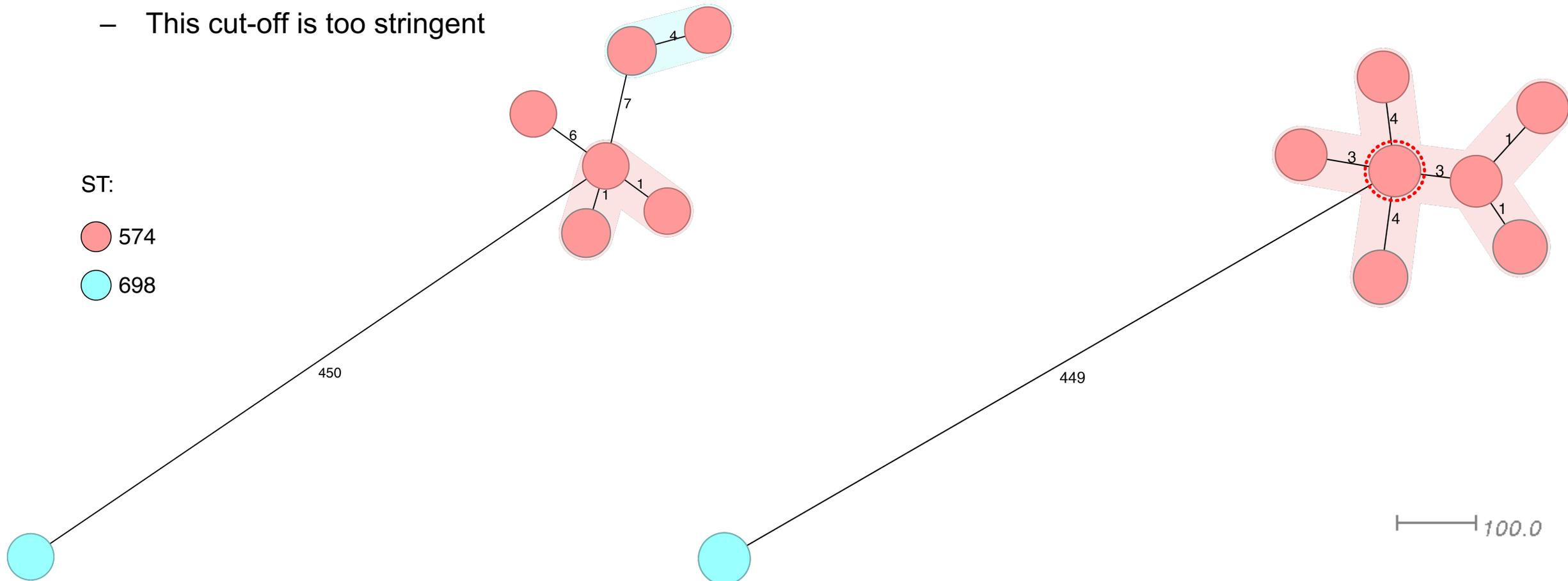
- Recent outbreak of diphtheria in migrant centres, across Europe
- Three dominant STs
- But FOUR dominant clusters

Epi Curve



# *Corynebacterium diphtheriae* and the cut-off dilemma

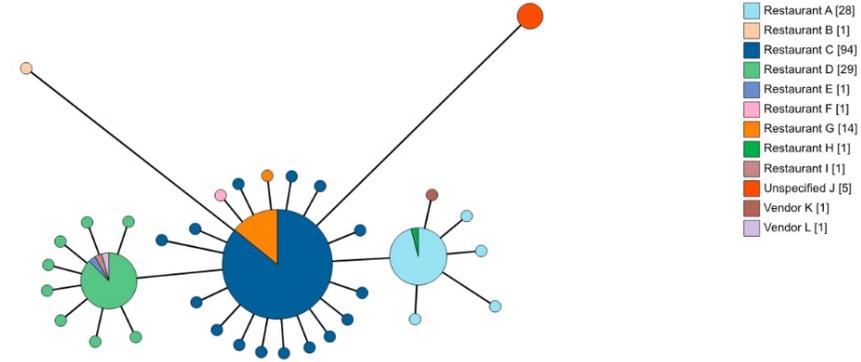
- Publication by Dangel gives cut-off of 5 alleles
- What in this instance?
- Sequence more...
- This cut-off is too stringent



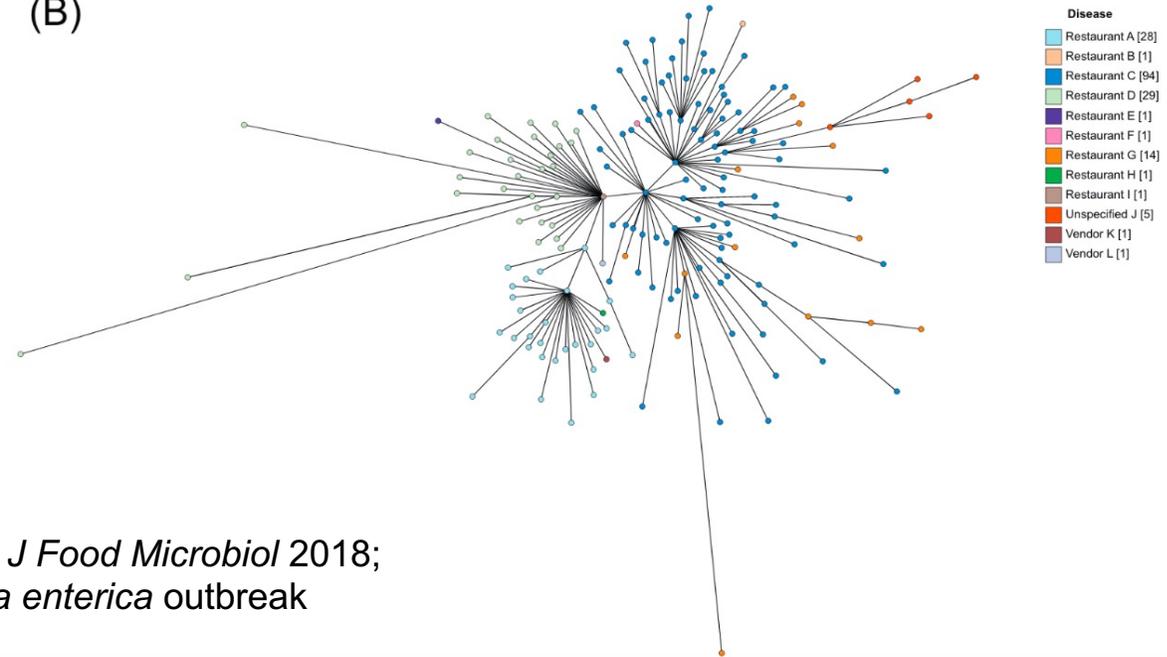
# Whole genome MLST (wgMLST)

- Recalculated with every new isolate = not stable scheme
- Allele based
- Uses *\*all\** alleles found, ie plus accessory genes = higher resolution

(A)



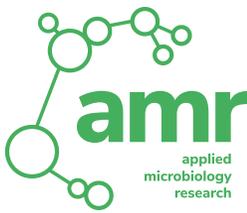
(B)



Pearce *Int J Food Microbiol* 2018;  
*Salmonella enterica* outbreak

Fig. 3. Core genome minimum spanning tree (A) and whole genome minimum spanning tree (B).

# Summary



- Whole genomes carry the information
- What do you want to know?
- With what certainty?
- cgMLST is a fantastic, transportable, fast, clear way to cluster genomes
- Careful analysis and wise interpretation is required
- Not all cgMLST schemes are equally reliable / interpretable
- Different species have different genome dynamics
- There can be technical reasons for allele differences: changes in protocol, assemblies
- Cutoffs are a guide rather than a hard rule
- Time plays a role
- Within clusters, depending on the question, it can be best to look at the SNP level
- Although SNP thresholds are also rarely defined
  
- Stay tuned for SNP analysis next week!

**Many thanks for your  
attention**

**Questions??**