Bacterial strain Taxonomy for Genomic Surveillance

Phylogenetic taxonomies based on whole-genome SNPs

22nd October 2025

Intended Learning Objectives

Specific objectives of this session:

1. Explain the principles of SNP-based taxonomies

2. Explore SNP-based nomenclature systems to classify bacterial isolates

3. Consider the strengths and limitations of SNP-based approaches in the context of public health microbiology

Outline

This session consists of the following elements:

1. Overview of SNP-based approaches
2. Taxonomy and variation between different pathogens
3. Inclusion of novel genomes into established nomenclature
4. Considering backwards compatibility with previous schemes
5. Approaches to define and name novel lineages of concern

Have you had much experience with SNP-based approaches?

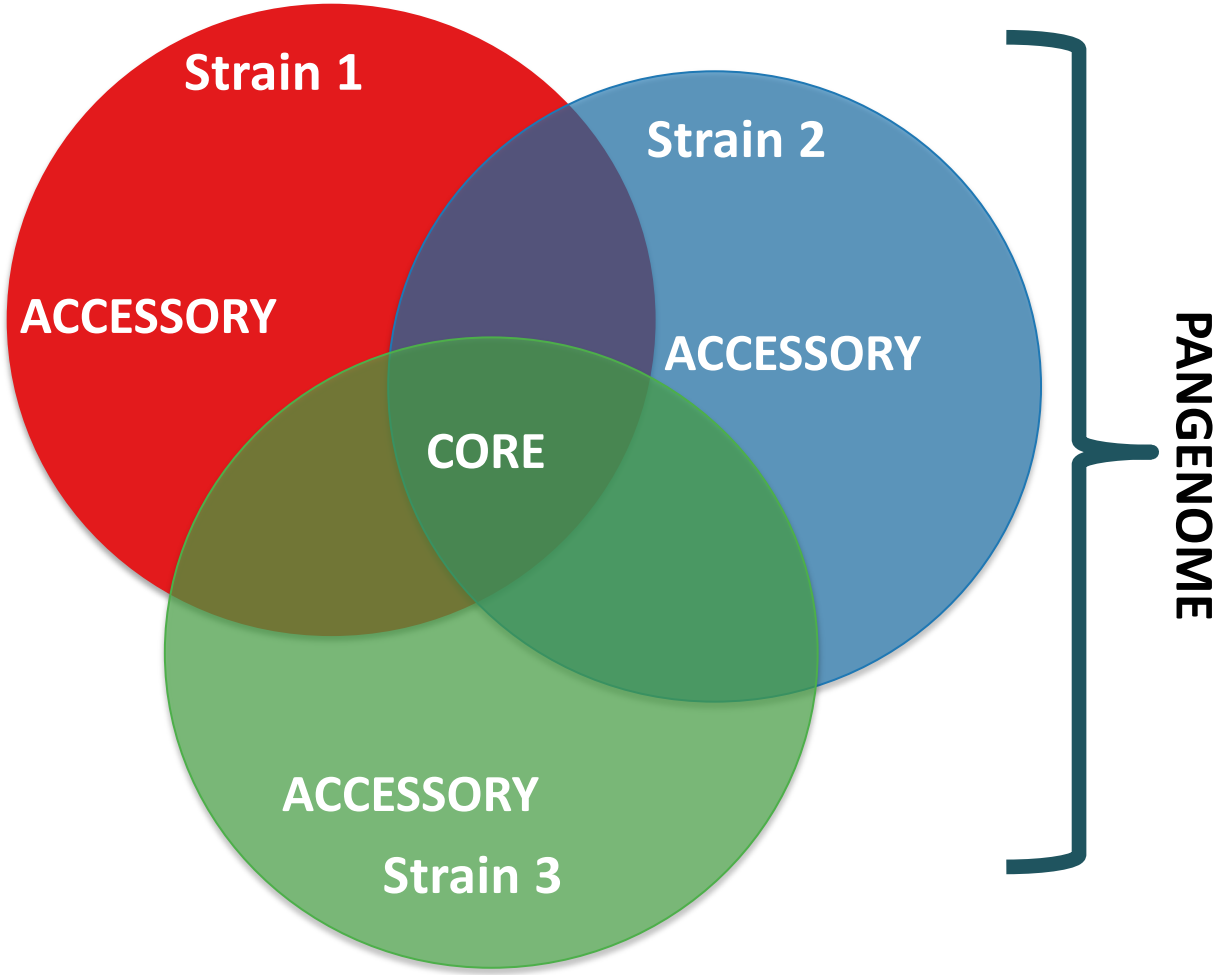
Insights from bacterial genomes

Core genome

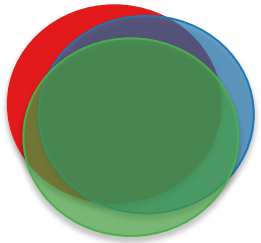
- Present in all isolates
- Infer how related isolates are to each other
- Measurably evolving populations – molecular clock of the bacteria
- Vertical evolution

Accessory genome

- Variable genome content
- Enable adaptation to different ecological niches e.g. AMR genes
- Horizontal evolution

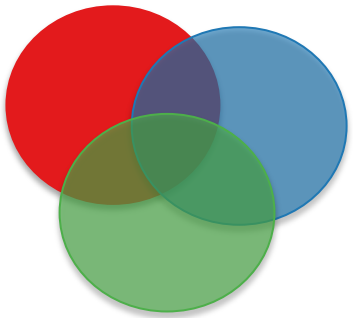


Genome size differs by bacterial pathogen



Specialist

- Often have 'closed pangenomes'
- E.g *Coxiella burnetii* genomes ~2,100 genes, obligate intracellular pathogen



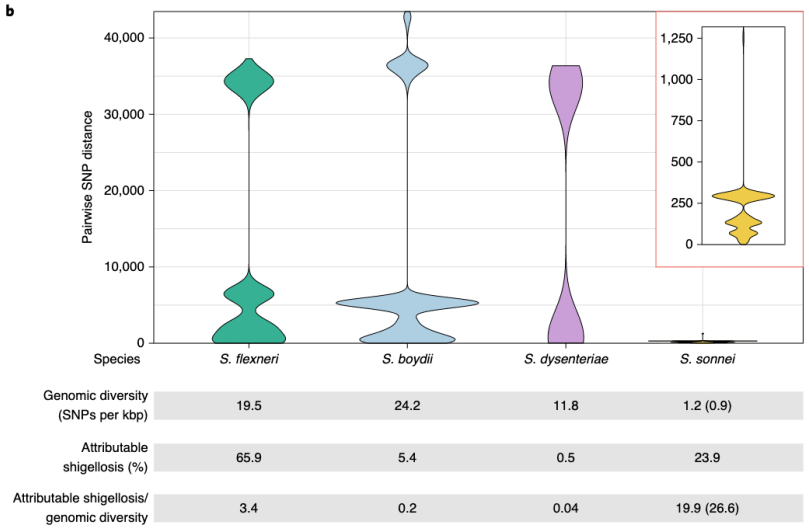
Generalists

- Often have 'open pangenomes'
- E.g. *Klebsiella pneumoniae* genomes ~5000-6000 genes, multiple ecological niches

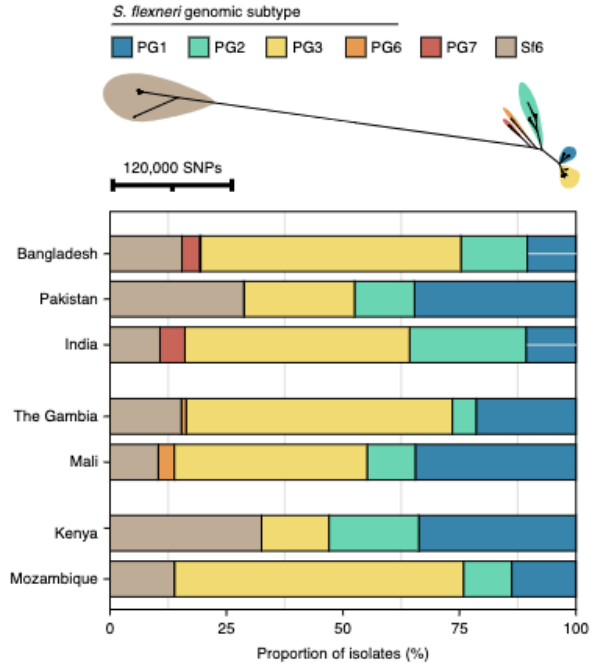
Core genome may vary in size depending on bacteria **and** dataset

Taxonomy differs by bacterial pathogens

- Important to know the pathogen-specific features of interest
- *Shigella*
  - 4 species – defined by gain of virulence plasmid
  - All sit within *Escherichia coli*
  - Difference in nomenclature within each species
- *Klebsiella*
  - Big changes to taxonomy over the past decade
  - Multiple new species
  - Importance of clonal groups
  - Use of LIN (Life Identification Number) codes for diverse population from cgMLST
- *Salmonella*
  - 2 species (discussion if should be more)
  - Multiple subspecies within *S. enterica*
  - 2,500 serovars defined by surface antigens
  - Serovars – Typhoidal, non-typhoidal (NTS) and invasive NTS



S. sonnei less genetic diversity

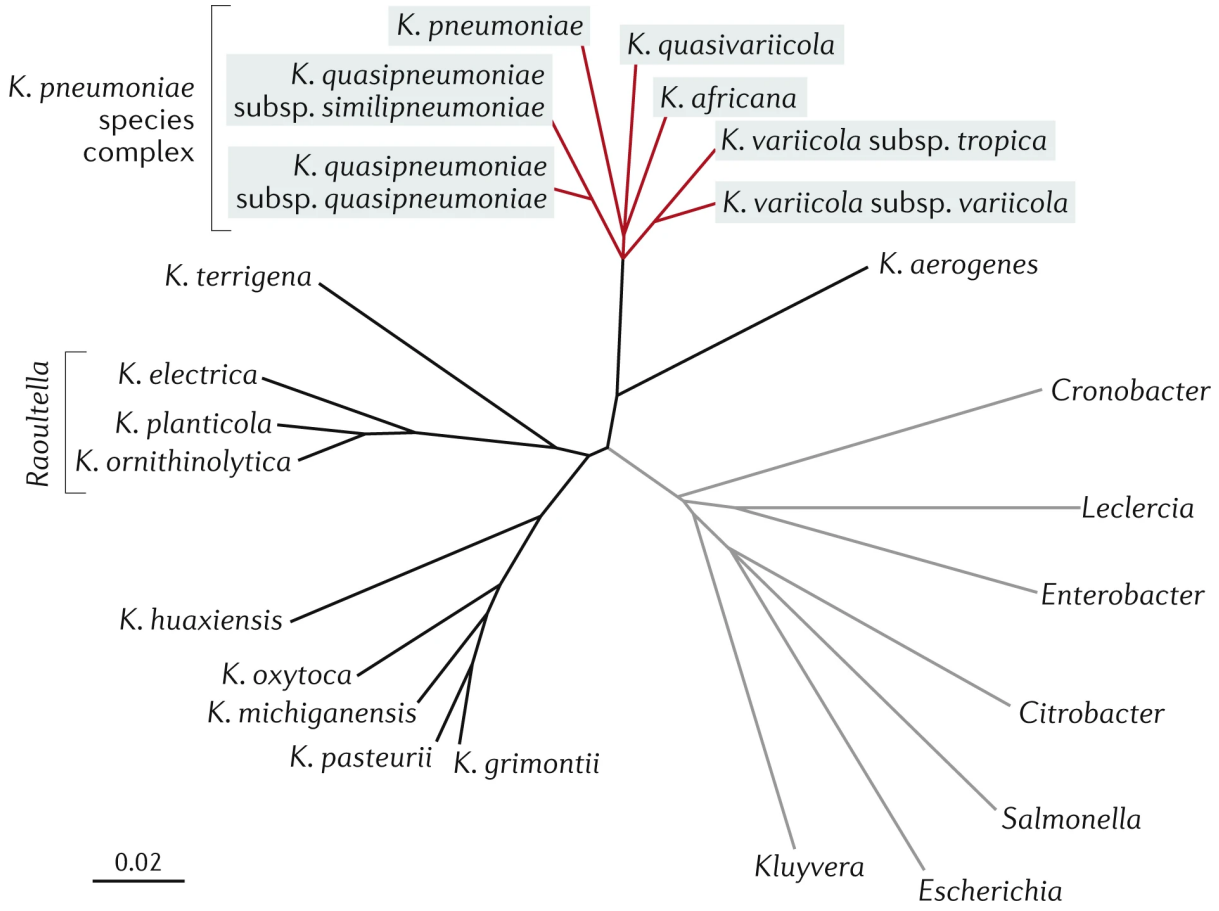


Phylogroups within *S. flexneri*

Bengtsson et al 2022 *Nature Microbiology*

Taxonomy differs by bacterial pathogens

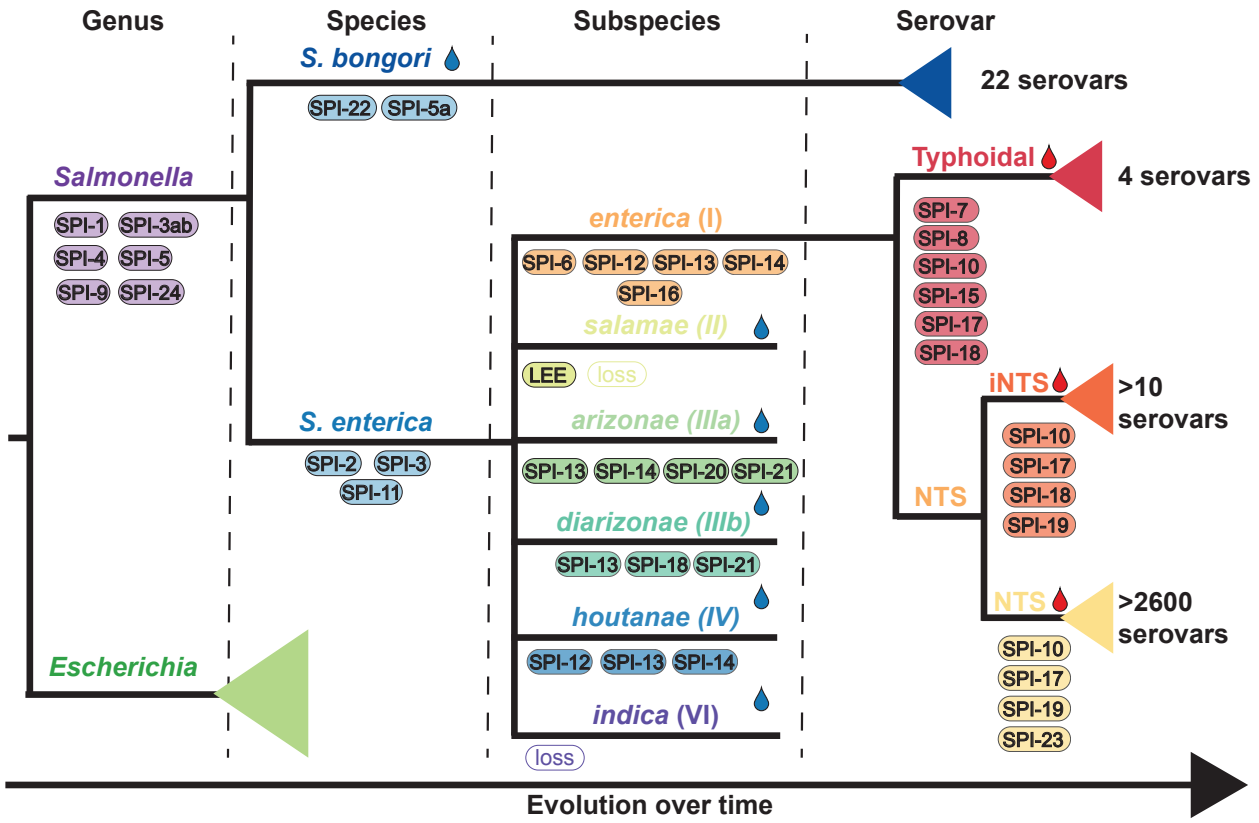
- Important to know the pathogen-specific features of interest
- *Shigella*
  - 4 species – defined by gain of virulence plasmid
  - All sit within *Escherichia coli*
  - Difference in nomenclature within each species
- *Klebsiella*
  - Big changes to taxonomy over the past decade
  - Multiple new species
  - Importance of clonal groups
  - Use of LIN (Life Identification Number) codes for diverse population from cgMLST
- *Salmonella*
  - 2 species (discussion if should be more)
  - Multiple subspecies within *S. enterica*
  - 2,500 serovars defined by surface antigens
  - Serovars – Typhoidal, non-typhoidal (NTS) and invasive NTS



Wyres et al 2020 *Nature Reviews Microbiology*, Tian et al 2020 *Nucleic Acids Research*, Hennart et al 2022 *Molecular Biology and Evolution*

Taxonomy differs by bacterial pathogens

- Important to know the pathogen-specific features of interest
- *Shigella*
  - 4 species – defined by gain of virulence plasmid
  - All sit within *Escherichia coli*
  - Difference in nomenclature within each species
- *Klebsiella*
  - Big changes to taxonomy over the past decade
  - Multiple new species
  - Importance of clonal groups
  - Use of LIN (Life Identification Number) codes for diverse population from cgMLST
- *Salmonella*
  - 2 species (discussion if should be more)
  - Multiple subspecies within *S. enterica*
  - 2,500 serovars defined by surface antigens
  - Serovars – Typhoidal, non-typhoidal (NTS) and invasive NTS



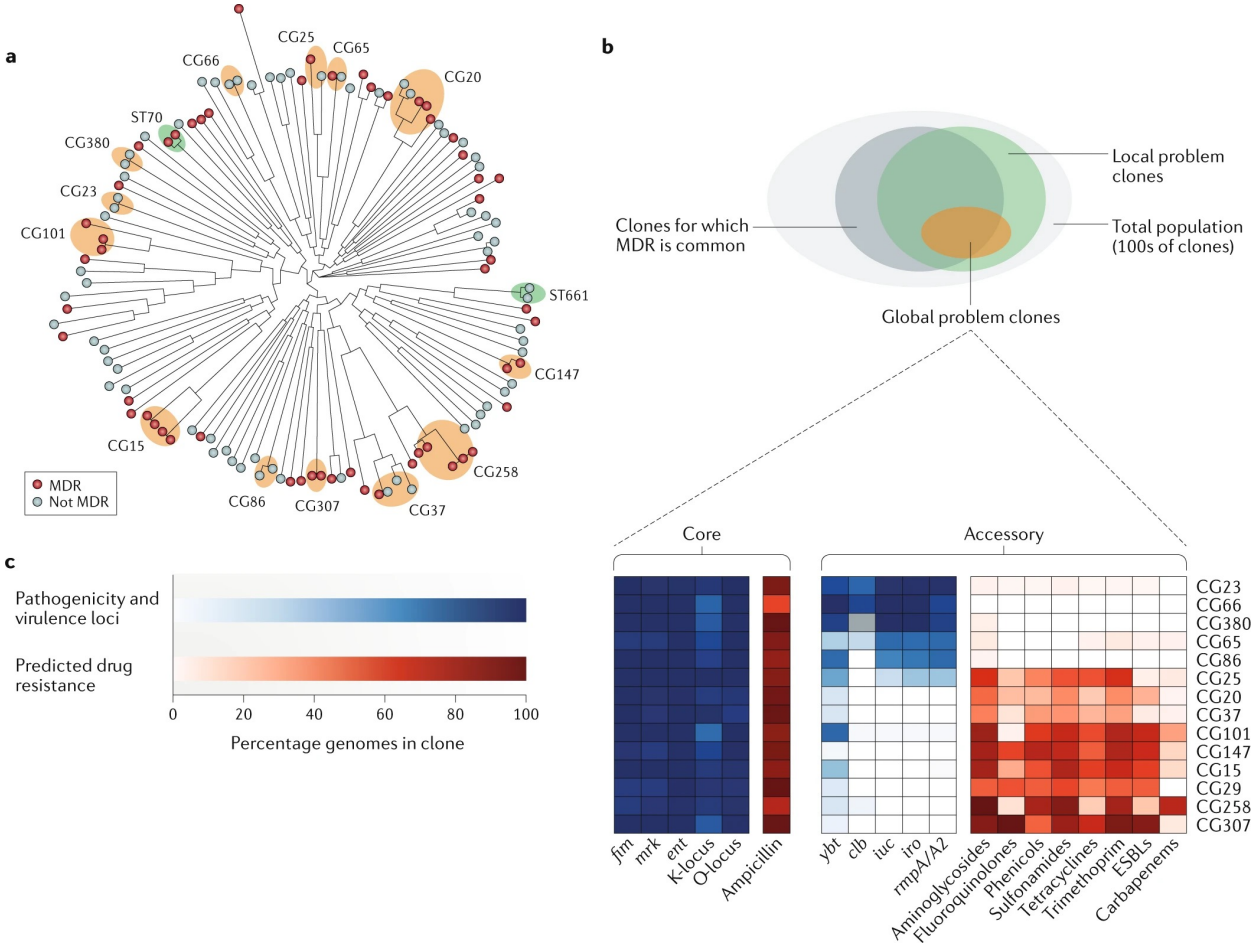
Chattaway et al 2023 *Pathogens*, Pierce et al 2021 *Genomics*,
Fookes et al 2011 *PLos Pathogens*, Sia et al 2025 *Trends in Microbiology*

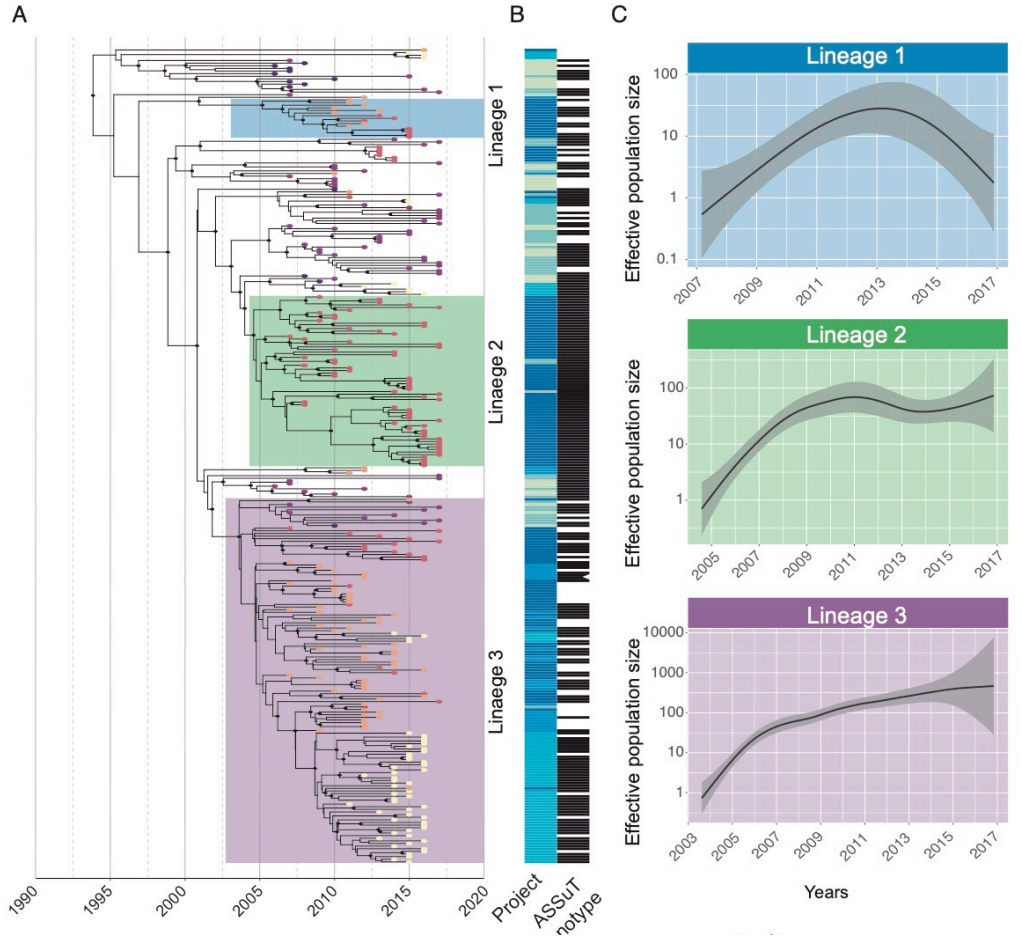
Considerations for SNP-based approaches

- What is important to characterise from the genome?
- How has evolution shaped the bacterial genome (e.g. open / closed)?
- When would SNP-based approaches be useful? Or are there other tools that would suit pathogen better?
- What data would help public health surveillance efforts?
- How should SNP-based approaches be maintained?

What are some applications for SNP-based phylogenies?

Different applications of SNP-based phylogenies





Species level - Population framework

Klebsiella pneumoniae – Clonal Groups (CG)

Wyres et al 2020 *Nature Reviews Microbiology*

Lineage/Serovar Level - Timed tree

Monophasic *Salmonella* – evolutionary dynamics

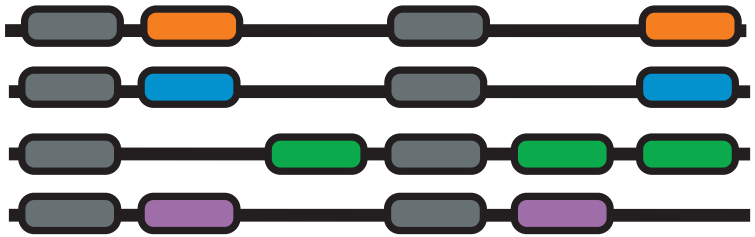
Ingle et al 2021 *Nature Communications*

SNP-based taxonomies need SNP alignments

Considerations for SNP alignments from core genome

- Alignment of reads to reference genome
  - Selection of reference
  - Different tools e.g. Snippy
  - Masking of regions (phage, recombination)
  - Thresholds for filtering of SNPs
- Assembly and pangenome analysis
  - Core genome e.g. of genes in >95% isolates with Panaroo
  - SNPs from core genes

Reference genome



Reference free - pangenome

Tonkin-Hill et al 2020 *Genome Biology*, Taouk et al 2025 *Microbial Genomics*

How to get SNP-based information

Alignment of short read data to reference genome

SNPs in the 'core genome' for alignment

Infer (Maximum Likelihood) tree

Identification of clusters/ groups from tree

Use of nomenclature to identify important groups for surveillance

Development of SNP-based methods

What are some points to consider for SNP-based approaches

How to get SNP-based information

Alignment of short read data to reference genome

← Selection of reference genome, inclusion of contextual genomes

SNPs in the 'core genome' for alignment

← Size of core genome in dataset

Infer (Maximum Likelihood) tree

← Computational resources available

Approach to take to define e.g. BAPs or SNP threshold→

Identification of clusters/ groups from tree

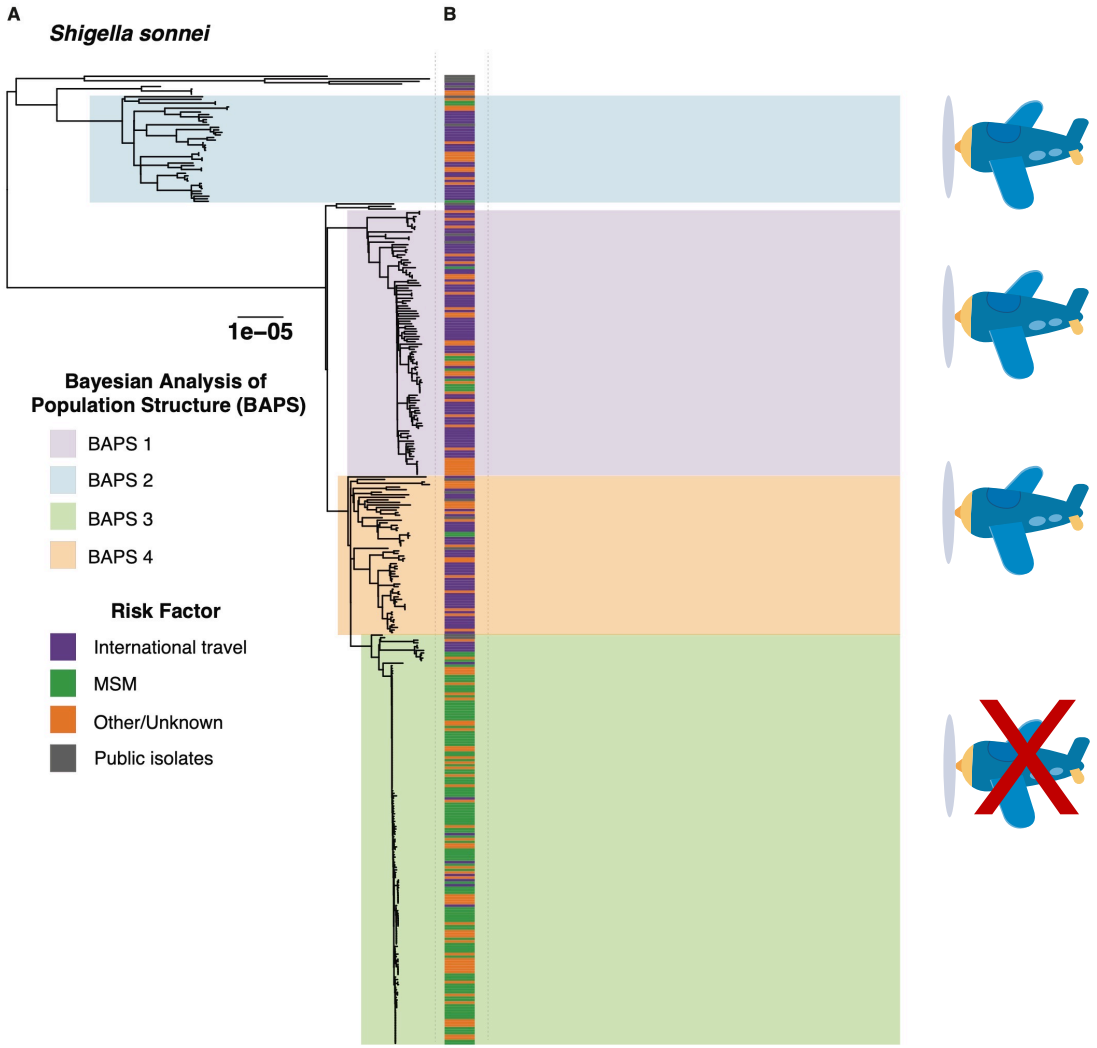
Development of nomenclature for broader use →

Use of nomenclature to identify important groups for surveillance

Development of SNP-based approaches that don't need a tree every time →

Development of SNP-based methods

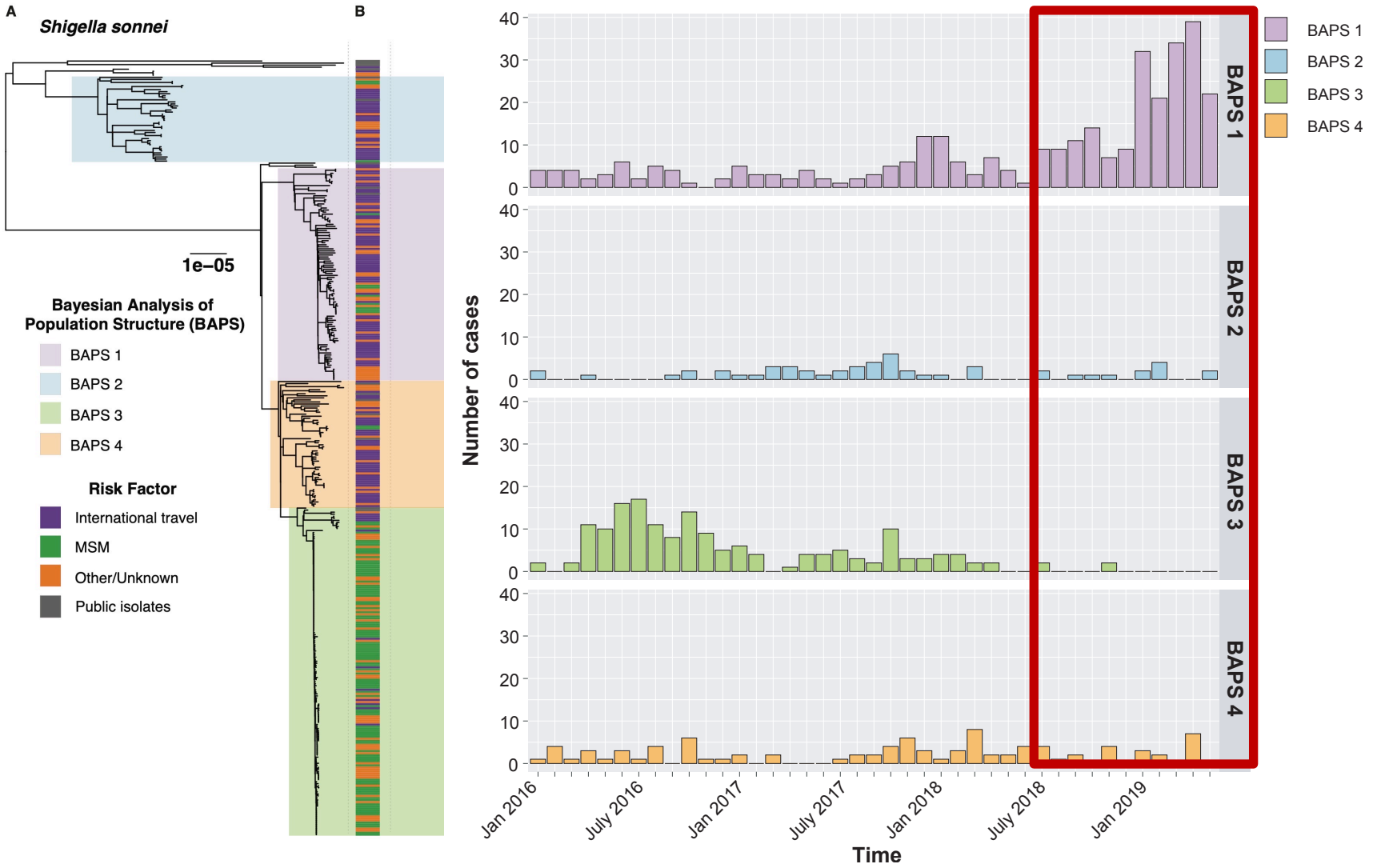
Example of SNP-based approaches: *Shigella sonnei*



- Increase in cases since 2015
- New epidemiological risk factor
- Infer ML tree and BAPS groups
- Lineages associated with different AMR profiles

Ingle et al 2019 *Clinical Infectious Diseases*

Example of SNP-based approaches: *Shigella sonnei*



- 2019: new year, new outbreak lineage
- Change in AMR profile
- New isolates added to original dataset
- New SNP alignment and new tree
- 2020: new outbreak lineage

Ingle et al 2019 *Clinical Infectious Diseases,*
Williamson, Ingle and Howden 2019 *New England Journal of Medicine*
Ingle et al 2020 *Antimicrobial Agents and Chemotherapy*

Example of SNP-based approaches: *Shigella sonnei*

- Increase in drug-resistant *S. sonnei* globally
- Need approaches to track international spread and standard nomenclature

nature communications

Article

https://doi.org/10.1038/s41467-023-37672-w

The evolution and international spread of extensively drug resistant *Shigella sonnei*

Received: 12 September 2022

Accepted: 24 March 2023

Published online: 08 April 2023

Check for updates

Lewis C. E. Mason[1,2], David R. Greig[3], Lauren A. Cowley[4], Sally R. Partridge[5,6,7,8], Elena Martinez[7,9], Grace A. Blackwell[7,9], Charlotte E. Chong[2], P. Malaka De Silva[2], Rebecca J. Bengtsson[2], Jenny L. Draper[7,9], Andrew N. Ginn[7,8,9,10], Indy Sandaradura[6,7,9], Eby M. Sim[5,6,8], Jonathan R. Iredell[5,6,7,8], Vitali Sintchenko[5,6,7,8,9,11], Danielle J. Ingle[12], Benjamin P. Howden[12], Sophie Lefèvre[13], Elisabeth Njamkepo[13], François-Xavier Weill[13], Pieter-Jan Ceyssens[14], Claire Jenkins[3] & Kate S. Baker[1,2]

nature communications

Article

https://doi.org/10.1038/s41467-023-36222-8

Rapid emergence of extensively drug-resistant *Shigella sonnei* in France

Received: 22 November 2022

Accepted: 19 January 2023

Published online: 28 January 2023

Sophie Lefèvre[1], Elisabeth Njamkepo[1], Sarah Feldman[2,3], Corinne Ruckly[1], Isabelle Carle[1], Monique Lejay-Collin[1], Laëtitia Fabre[1], Iman Yassine[1], Lise Frézal[1], Maria Pardos de la Gandara[1], Arnaud Fontanet[2] & François-Xavier Weill[1]

What information do we want from Typhi genome?

Lineage – genotype

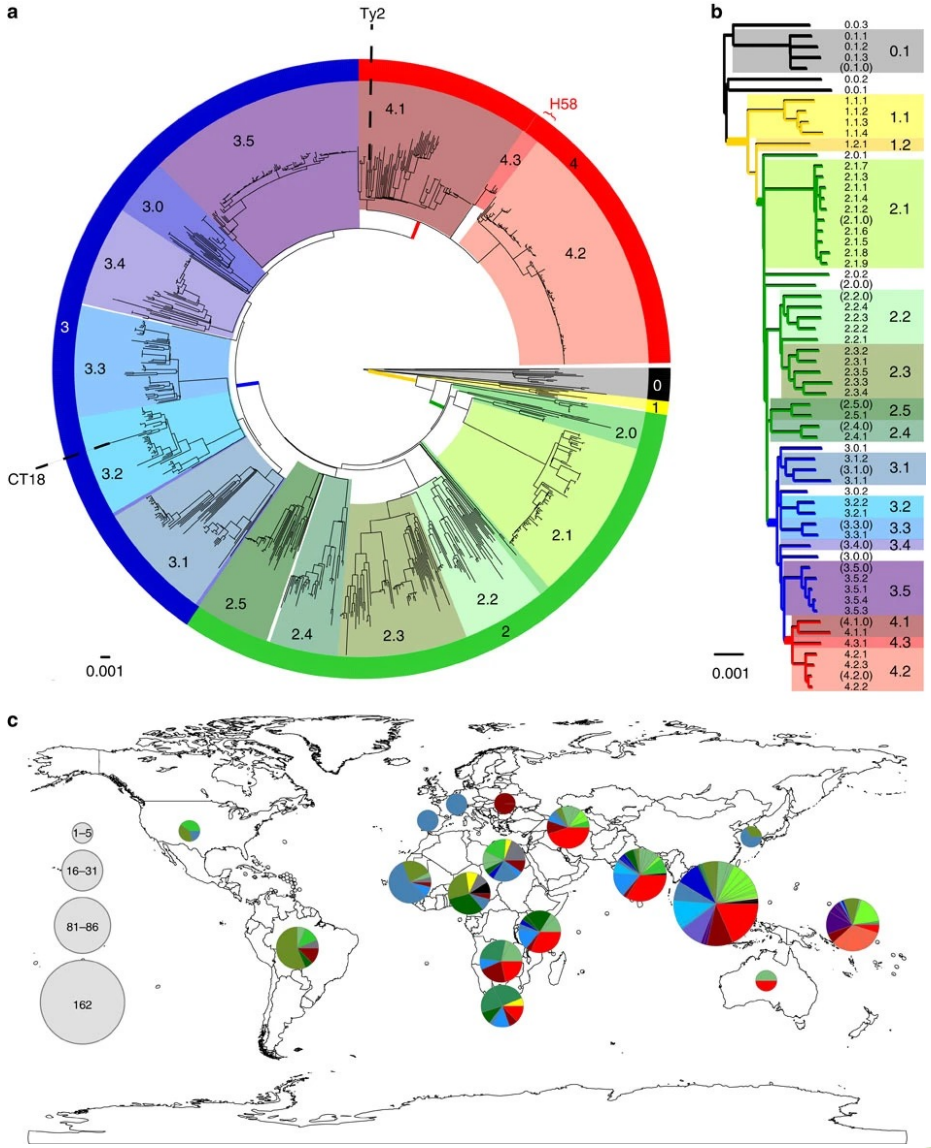
- GenoTyphi scheme
- https://github.com/typhoidgenomics/genotyphi
- 4 major lineages, > 80 genotypes
- Widely adopted in research and public health settings

AMR profile

- SNVs
- Genes

Plasmid profile

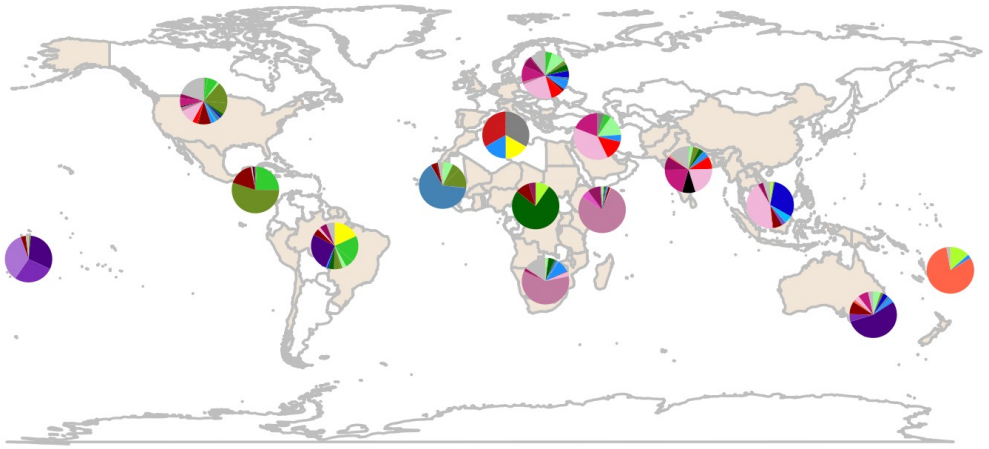
- Epidemiologically important plasmids



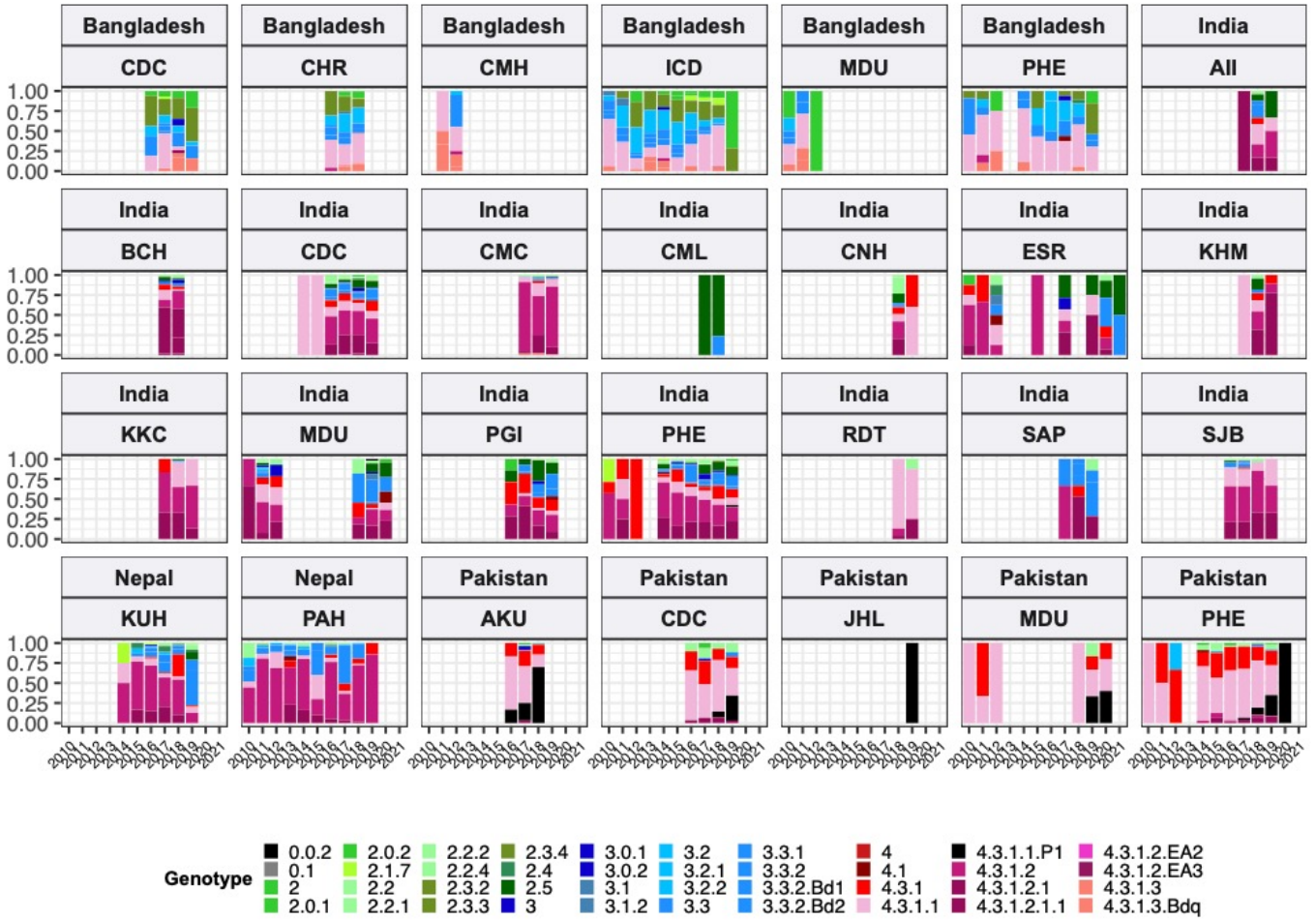
Wong et al 2016 *Nature Communications,* Dyson and Holt 2021 *Journal of Infectious Diseases*

Insights from 13,000 Typhi genomes

a) Genotype prevalence by world region, 2010–2020



a) Annual genotype prevalence, by lab

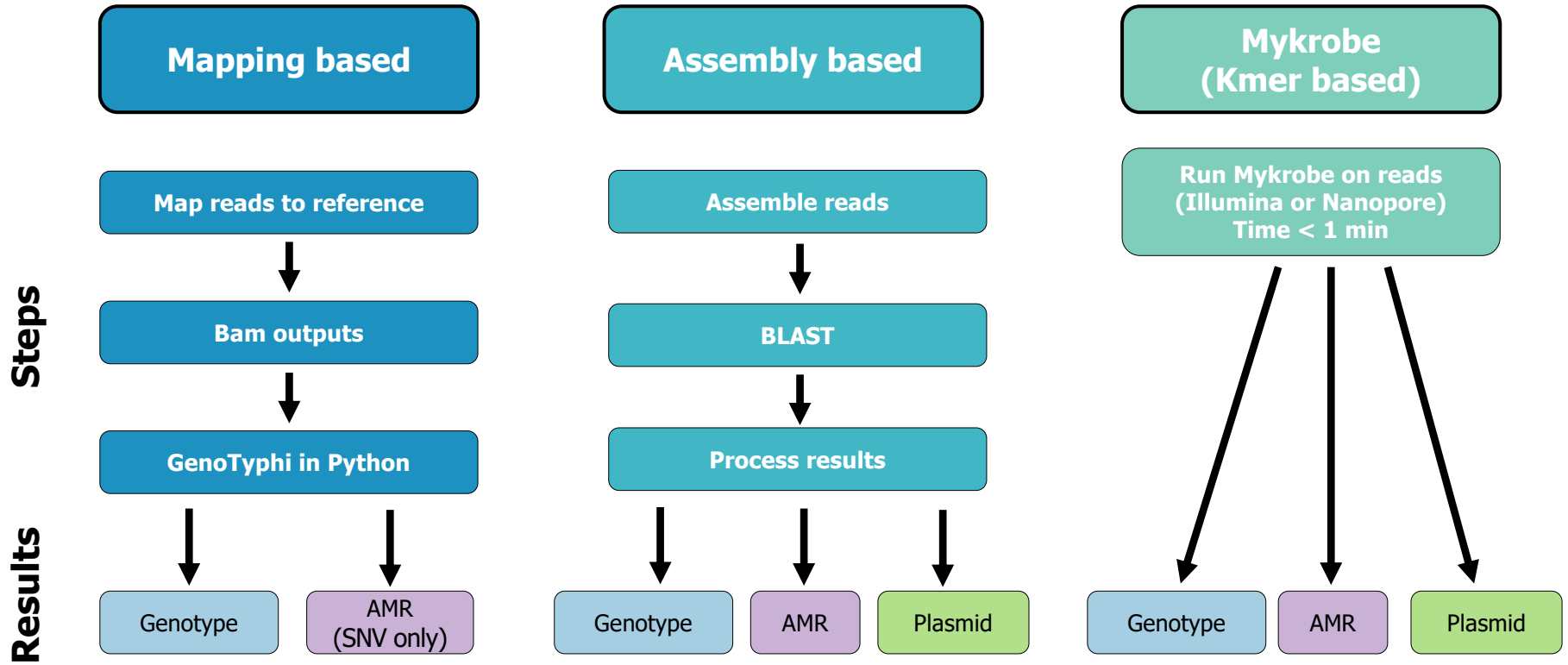


Collaborative efforts >200 researchers/ public health workers

Enhanced genomic surveillance

Carey et al. 2023 *eLife*
https://www.typhoidgenomics.org/
https://github.com/typhoidgenomics/TyphoidGenomicsConsortiumWG1

How to get SNP-based taxonomy information



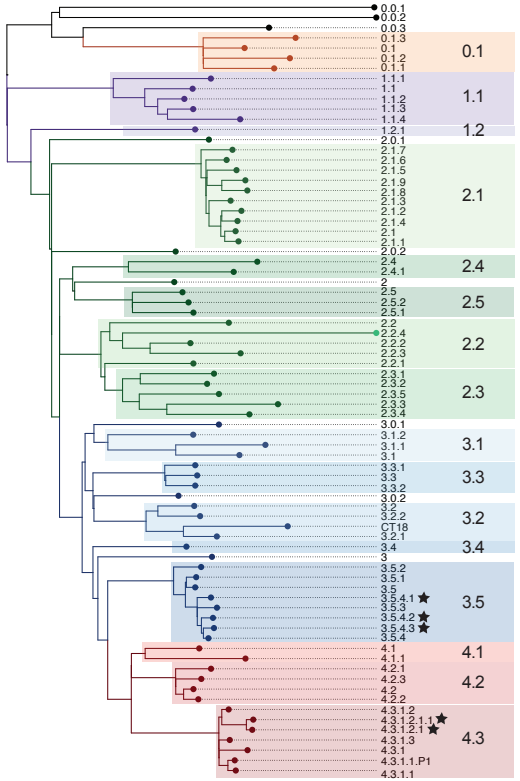
Genotyping schemes can be implemented on different platforms

Bradley et al 2015 *Nature Communications*
https://github.com/Mykrobe-tools/mykrobe

Argimón et al 2021 *Nature Communications*
https://next.pathogen.watch/

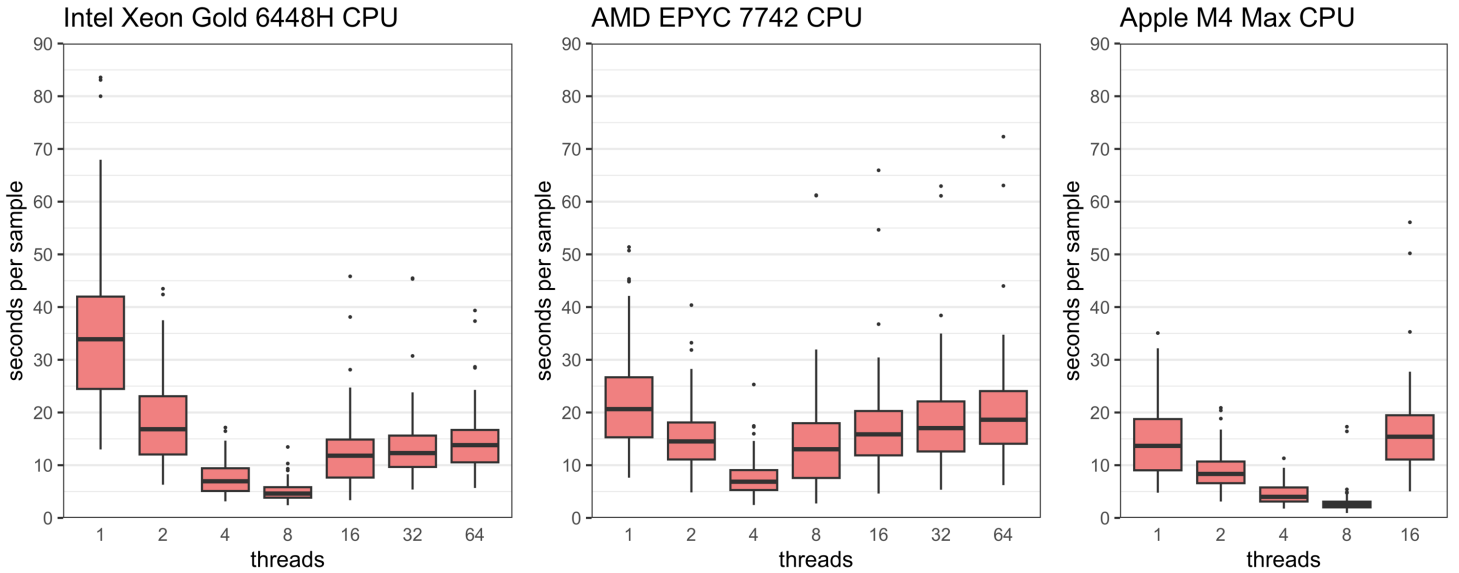
Typhi Mykrobe

- Implementation of GenoTyphi scheme in Mykrobe
- SNP-based approach – use of established poplation framework
- Fast. Typically <1 min genome
- Enables inference of tree position without the inferring a tree each time
- Provides standard nomenclature for lineages and sub- lineages



GenoTyphi framework

- 87 genotypes
- 4 primary lineages
- 16 clades
- 49 subclades



Ingle et al 2024 BioRxiv
https://github.com/typhoidgenomics/TyphoidGenomicsConsortiumMykrobe

Running Typhi Mykrobe

Installing mykrobe

First, install Mykrobe (v0.10.0+) as per the instructions on the Mykrobe github.

Once Mykrobe is installed, you can run the following two commands to ensure you have the most up-to-date panels for genotyping, including the Typhi panel (latest version, v20240407):

```
mykrobe panels update_metadata
mykrobe panels update_species all
```

You can check what version of the scheme is currently loaded in your Mykrobe installation via:

```
mykrobe panels describe
```

Run Mykrobe on fastq file/s for a given genome:

```
mykrobe predict --sample aSample \
  --species typhi \
  --format json \
  --out aSample.json \
  --seq aSample_1.fastq.gz aSample_2.fastq.gz
```

Tabulate Mykrobe results for one or more genomes:

(requires Python3 + pandas library)

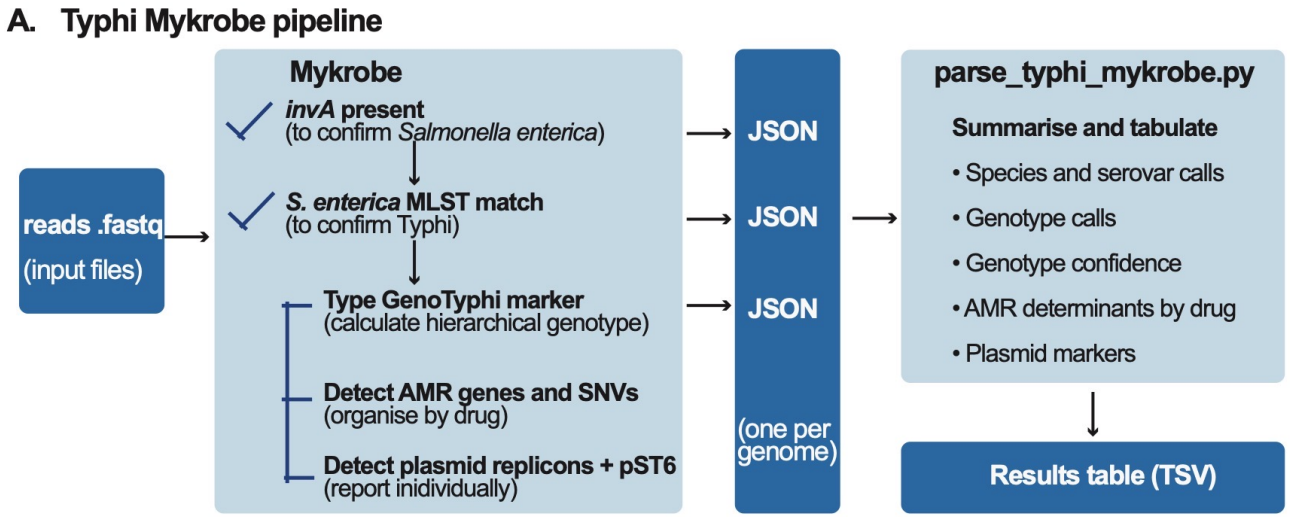
(python script `parse_typhi_mykrobe.py` is in this repository in the `/typhimykrobe` directory in this repository)

```
python parse_typhi_mykrobe.py --jsons *.json --prefix mykrobe_out
```

https://github.com/typhoidgenomics/genotyphi?tab=readme-ov-file#typhi-mykrobe

Typhi Mykrobe

- Direct from reads (Illumina or ONT)
- Output – tabular format

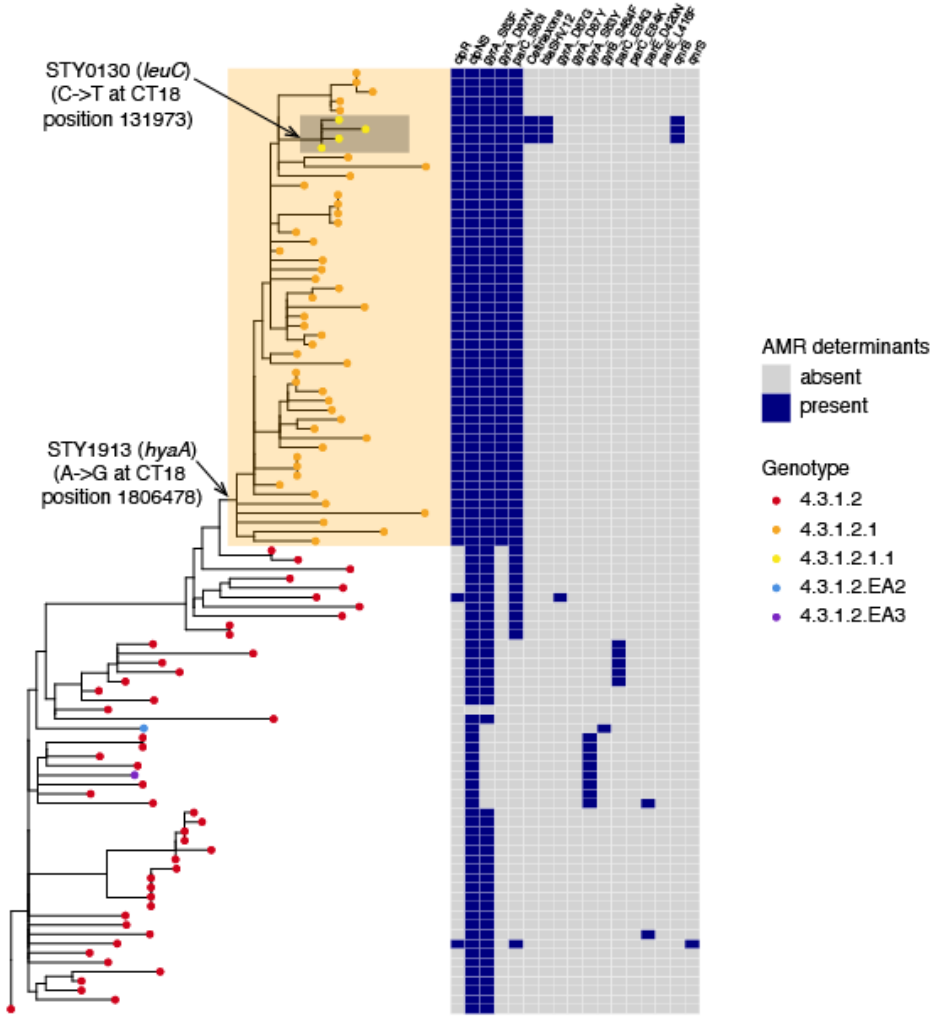


species	spp_percent	genotype	confidence	lowest support for genotype marker	supported markers	for additional	additional markers	node support
typhi	91.518	4.3.1.1	strong	-	-	-	-	1 (1; 0/159); 2 (1; 0/180); 3 (1; 0/155); 4 (1; 133/0); 4.3.1 (1; 138/0); 4.3.
typhi	91.565	4.3.1.2	strong	-	-	-	-	1 (1; 0/167); 2 (1; 0/176); 3 (1; 0/146); 4 (1; 155/0); 4.3.1 (1; 153/0); 4.3.
typhi	91.723	2.2	strong	-	-	-	-	1 (1; 0/133); 2 (1; 1/138); 2.2 (1; 129/0)
typhi	91.504	4.3.1.2	strong	-	-	-	-	1 (1; 0/133); 2 (1; 0/147); 3 (1; 0/160); 4 (1; 139/0); 4.3.1 (1; 121/0); 4.3.
typhi	91.48	4.1	strong	-	-	-	-	1 (1; 0/148); 2 (1; 0/137); 3 (1; 0/152); 4 (1; 164/0); 4.1 (1; 120/0)
typhi	91.892	3	strong	-	-	-	-	1 (1; 0/61); 2 (1; 0/63); 3 (1; 0/75)
typhi	91.478	3	strong	-	-	-	-	1 (1; 0/57); 2 (1; 0/62); 3 (1; 0/66)
typhi	91.494	4.1	strong	-	-	-	-	1 (1; 0/86); 2 (1; 1/58); 3 (1; 0/88); 4 (1; 48/1); 4.1 (1; 56/0)
typhi	91.746	3.2.1	strong	-	-	-	-	1 (1; 1/63); 2 (1; 0/63); 3 (1; 0/63); 3.2 (1; 0/54); 3.2.1 (1; 0/67)
typhi	91.715	3.2.1	strong	-	-	-	-	1 (1; 0/71); 2 (1; 0/58); 3 (1; 0/80); 3.2 (1; 1/57); 3.2.1 (1; 0/62)

Ingle et al 2024 BioRxiv
https://github.com/typhoidgenomics/TyphoidGenomicsConsortiumMykrobe

Ongoing efforts for genotyping Typhi

- Typhi Mykrobe led by a working group within Global Typhoid Genomics Consortium
- Another working group exploring how to add new genotypes
  - GenoTyphi Genotyping Scheme Development and Curation Working Group
- Considerations for incorporating new genomes into the taxonomy and genotypes
  - Need to new tree to identify SNPs specific to lineages
- Considerations for how to name lineages going forward



https://www.typhoidgenomics.org/
https://zenodo.org/records/7407985

SNP-based schemes useful for clonal pathogens

Mykrobe schemes already exist for:

- *Mycobacterium tuberculosis*
- *Staphylococcus aureus*
- *Shigella sonnei*
- *Salmonella enterica* serovar Paratyphi B
- *Salmonella enterica* serovar Typhi

Mykrobe schemes in development for:

- *Shigella flexneri*

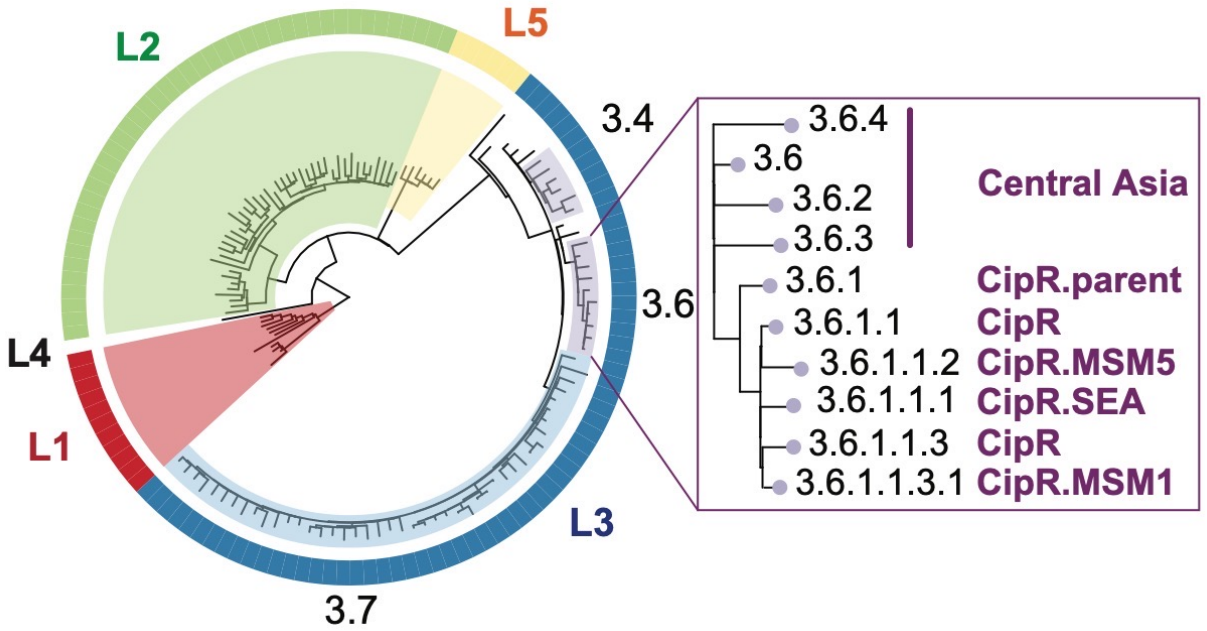
Considerations for developing schemes

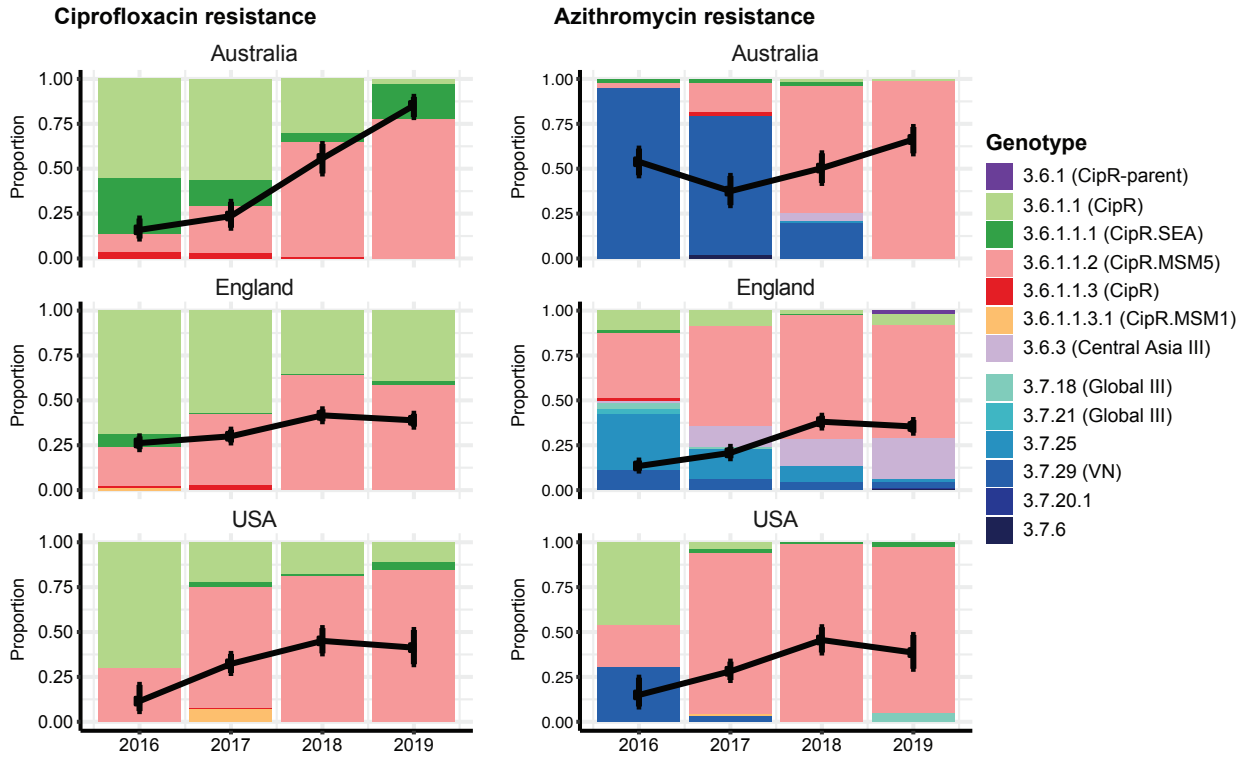
- Need discovery tree to define genotypes and identify SNPs
- Use of previously identified lineages/ clusters to inform design
- Naming of lineages
- When to define new lineages

https://github.com/Mykrobe-tools/mykrobe/wiki

Development of genotype schemes implemented in Mykrobe

- Genotyping scheme developed for *Shigella sonnei*
- Suited as clonal pathogen

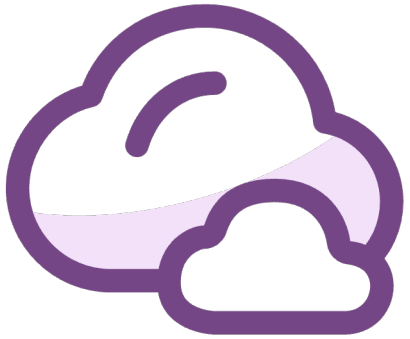




Five lineages with 128 genotypes

Applied to >4,000 genomes from 3 public health labs

Hawkey et al 2021 *Nature Communications*



What should be considered when naming a novel lineage?

In summary

List of learning points in this session:

- SNP-based approaches can be useful in tracking priority bacterial pathogens
  - Trees remain important for defining SNP-based genotyping schemes
  - Genotyping schemes implemented in Mykrobe can provide rapid and robust lineage identification for public health surveillance

- Taxonomy and nomenclature varies between bacterial pathogens
  - This is underpinned by evolution of the pathogen
  - Important to consider the dataset and research/ public health questions when using SNP-based approaches

Further reading

- https://github.com/rrwick/Core-SNP-filter
- https://github.com/gtonkinhill/rhierbaps
- https://github.com/gtonkinhill/panaroo
- https://github.com/Mykrobe-tools/mykrobe
- https://github.com/typhoidgenomics/genotyphi
- https://github.com/typhoidgenomics/TyphoidGenomicsConsortiumMykrobe
- https://github.com/katholt/sonneityping

References

- Wong et al. 2016. An extended genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. *Nature Communications.* https://doi.org/10.1038/ncomms12827

- Ingle et al. Typhi Mykrobe: fast and accurate lineage identification and antimicrobial resistance genotyping directly from sequence reads for the typhoid fever agent *Salmonella* Typhi. BioRxiv: https://doi.org/10.1101/2024.09.30.613582

- Ingle et al. 2019. Co-circulation of Multidrug-resistant Shigella Among Men Who Have Sex With Men in Australia. *Clinical Infectious Diseases.* https://doi.org/10.1093/cid/ciz005

- Hawkey et al. 2021. Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, Shigella sonnei. *Nature Communications.* https://doi.org/10.1038/s41467-021-22700-4

Acknowledgements

The creation of this training material was commissioned by ECDC to The University of Melbourne and the Peter Doherty Institute for Infection and Immunity with the direct involvement of Dr. Danielle Ingle