Phylogenetic Analysis

Classical vs genomic approaches, SNP-based phylogeny and. how SNP trees are used in epidemiology.

January 19, 2026, Niamh Lacy-Roberts, Research Assistant at DTU, nlac@food.dtu.dk

Intended Learning Objectives

Specific objectives of this session:

- Understand what phylogenetic analysis is and why it’s used
- Distinguish classical vs genomic (SNP-based) approaches
- Explain what a SNP-based phylogenetic tree represents
- Interpret trees in an epidemiological context

Outline

This session consists of the following elements:

1. Overview of phylogenetic analyses
2. Brief comparison of classical and genomic approaches
3. Explanation of SNP-based phylogeny
4. How SNP trees are used in epidemiology
5. Overview of tree construction and interpretation.

Phylogenetic analysis

Phylogenetic analysis overview

Phylogenetic analysis is the study of evolutionary relationships between organisms or isolates based on shared characteristics.

In genomic epidemiology, it is used to:

- Compare genomes to assess genetic relatedness
- Reconstruct evolutionary history using sequence data
- Visualize relationships as a phylogenetic tree

Phylogenetics in outbreak investigations

Phylogenetics for outbreaks ≠ evolution

In outbreak investigations, we are usually not interested in:
- Long-term evolutionary history
- Ancient divergence
- Species-level evolution

In outbreak investigations, phylogenetic analysis is used to:
- Assess how closely related isolates are
- Identify genetically clustered cases
- Distinguish outbreak-related from unrelated isolates

The focus is usually on:
- Short-term genetic differences (e.g. SNPs)
- Recent shared ancestry
- Measuring genetic distance, not long-term evolution

What is a SNP?

Single Nucleotide Polymorphisms (SNPs)

A SNP is:

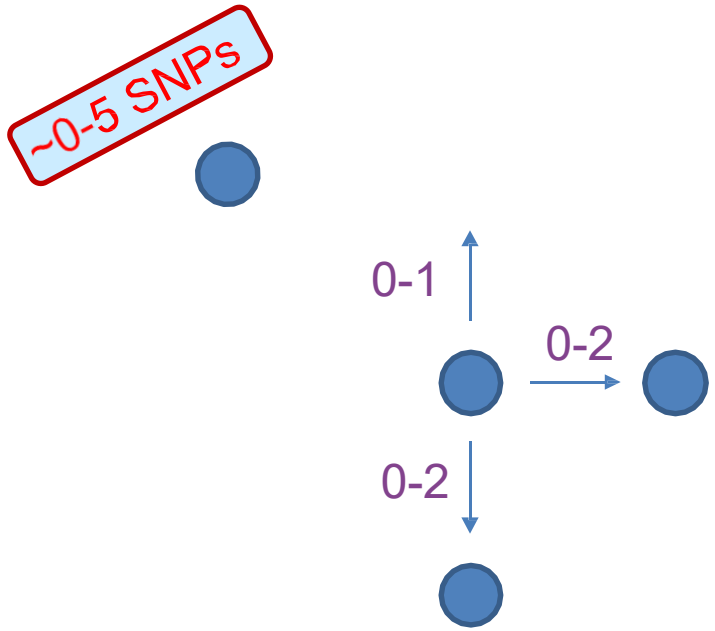
- A difference at a single DNA base (A, T, C, or G)
- Found when comparing genomes to a reference sequence
- One of the smallest possible genetic differences

Example:

Reference	A	T	C	**G**	A
Isolate	A	T	C	**A**	A

→ 1 SNP difference

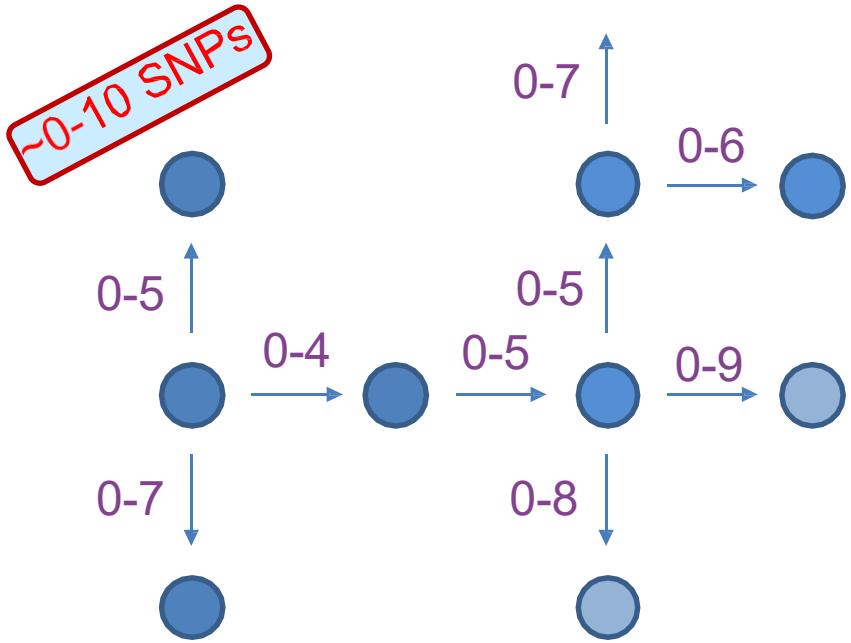
Single source outbreaks



Single source Short time span

"Contaminated dish"

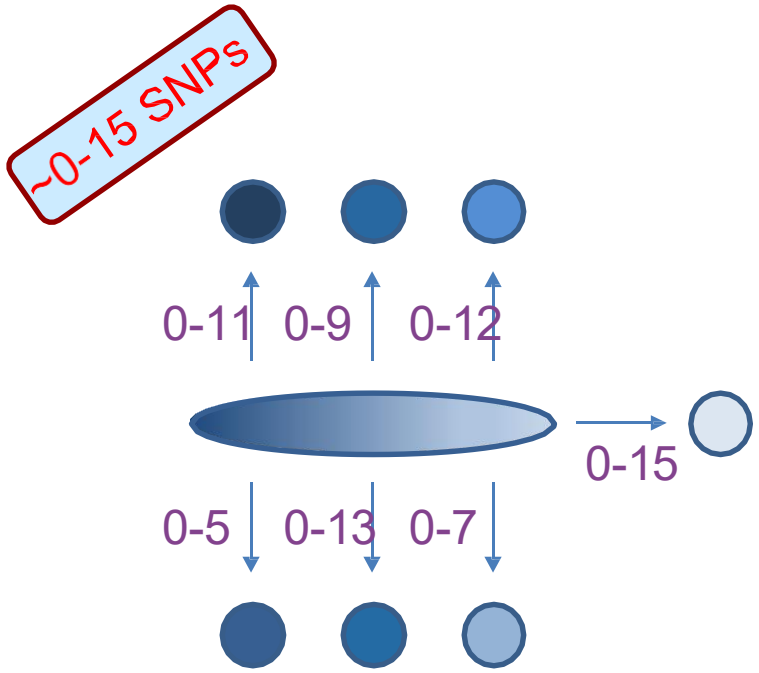
"Single infected patient"



Single source – local spread Long time span

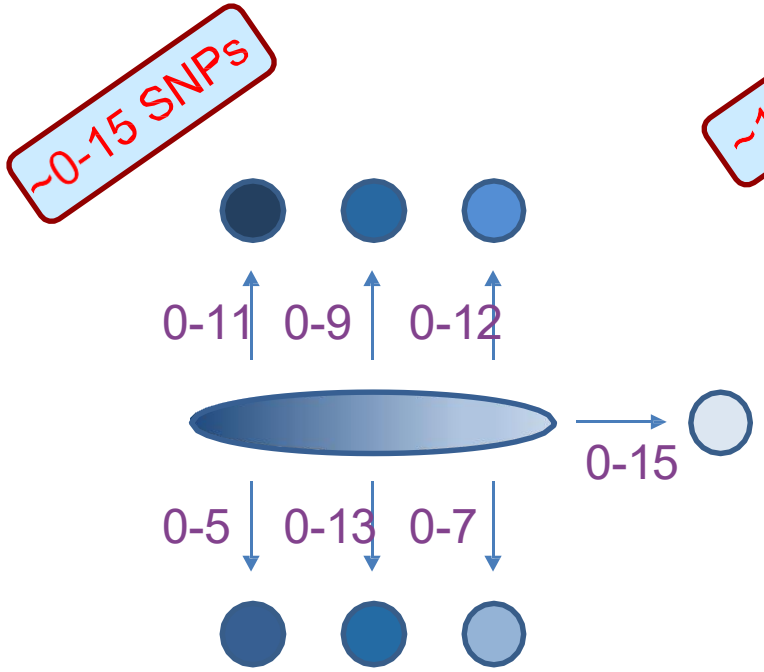
"Hospital or regional outbreak"

Complicated outbreaks

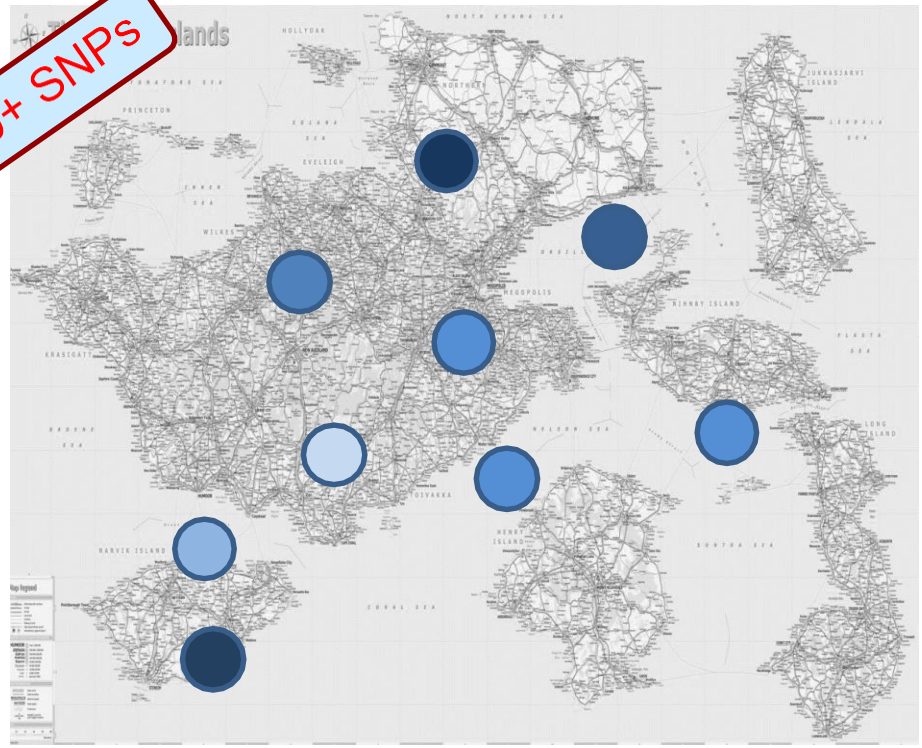


Single source
Long time span
"Contaminated processing plant / industry"
"Long-term colonized patient / healthcare worker"

Complicated outbreaks

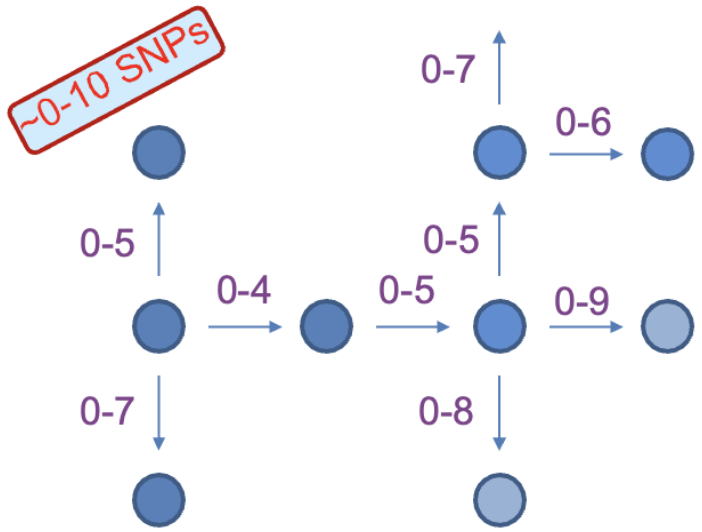


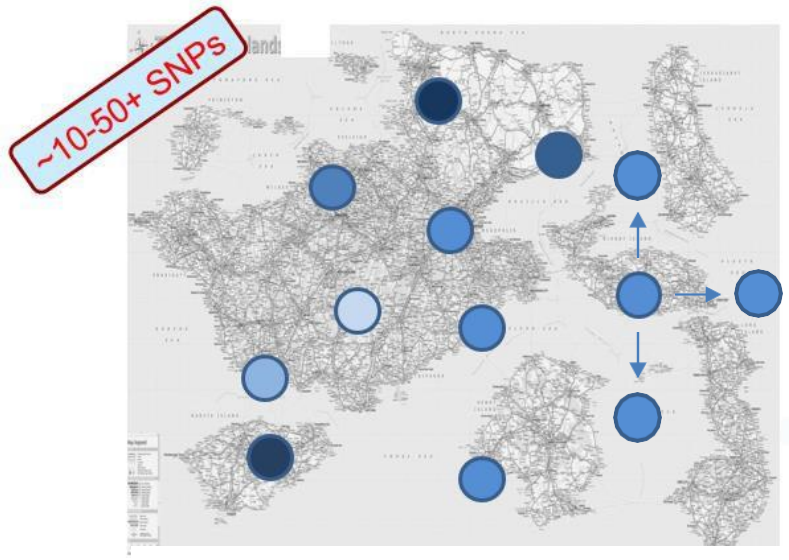
Single source
Long time span
"Contaminated processing plant / industry"
"Long-term colonized patient / healthcare worker"



International s
Long tim
"Importe source"
"Trav ted outbreak"

PO = Possible outbreaks





Tentative definition of possible outbreak (PO)

If two isolates have a SNP distance ≤ 10 (termed PO_{10}), they are considered to be so genetically related that they may be part of the same outbreak.

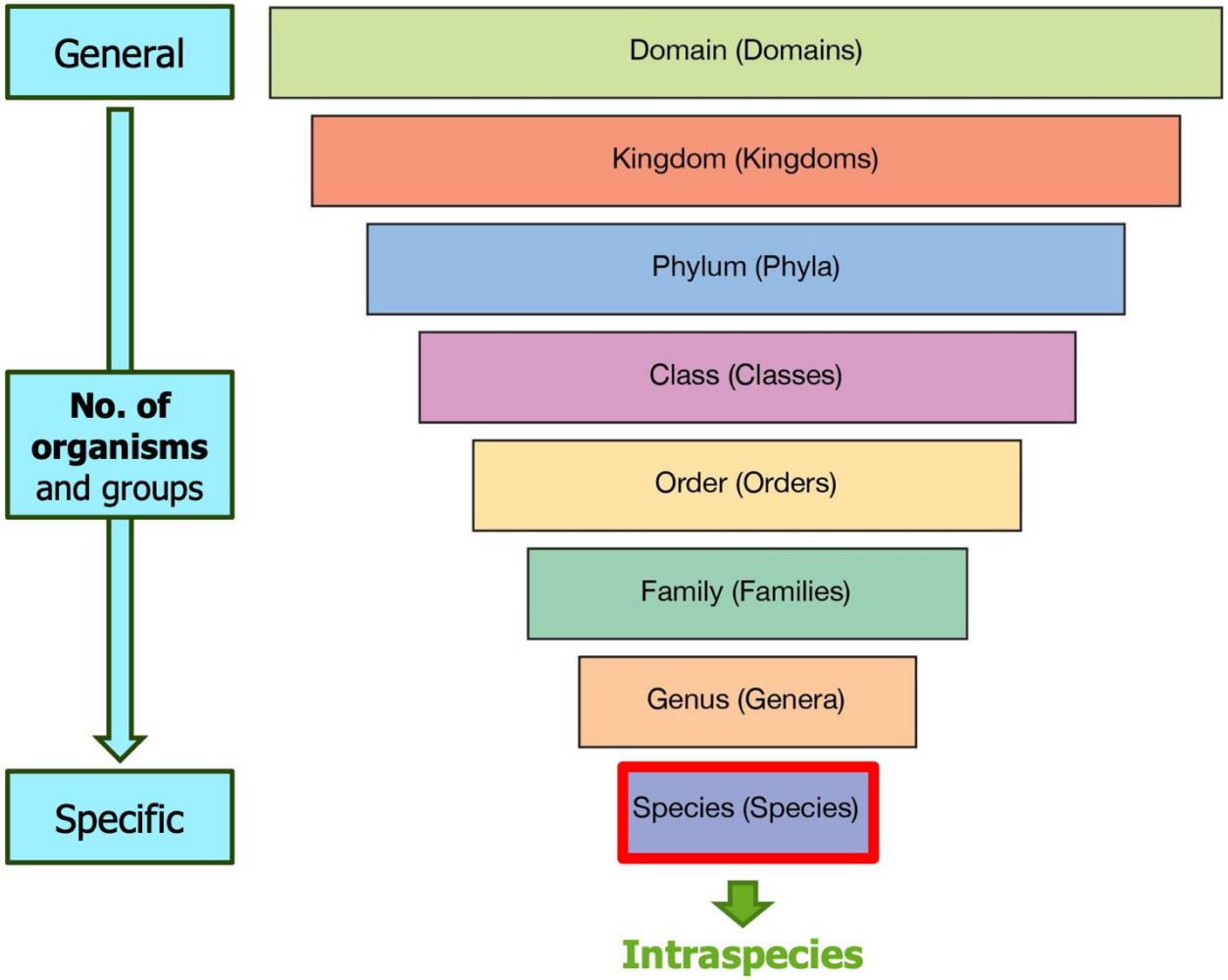
SNP thresholds

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

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

Classical and genomic approaches

Going Beyond species level



Variability:

Phenotypic features (e.g., capsule production)

Antimicrobial resistance patterns

Virulence properties

Gene content (e.g., plasmids/phages)

Ferreira Wanda F. Canas 340; Microbiologia. ISBN: 978-972-757-515-2

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Same Species, different Strains, different Risk

The bacterial risk to humans in most cases resides at the **strain level**, not the species

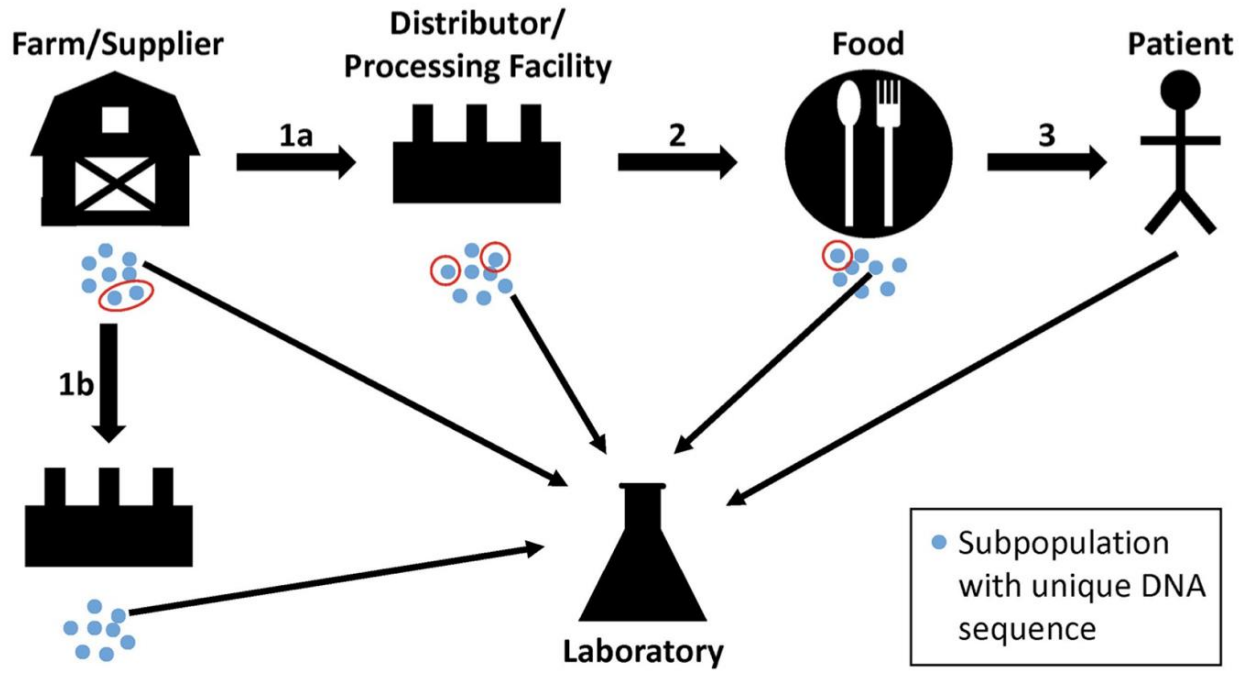
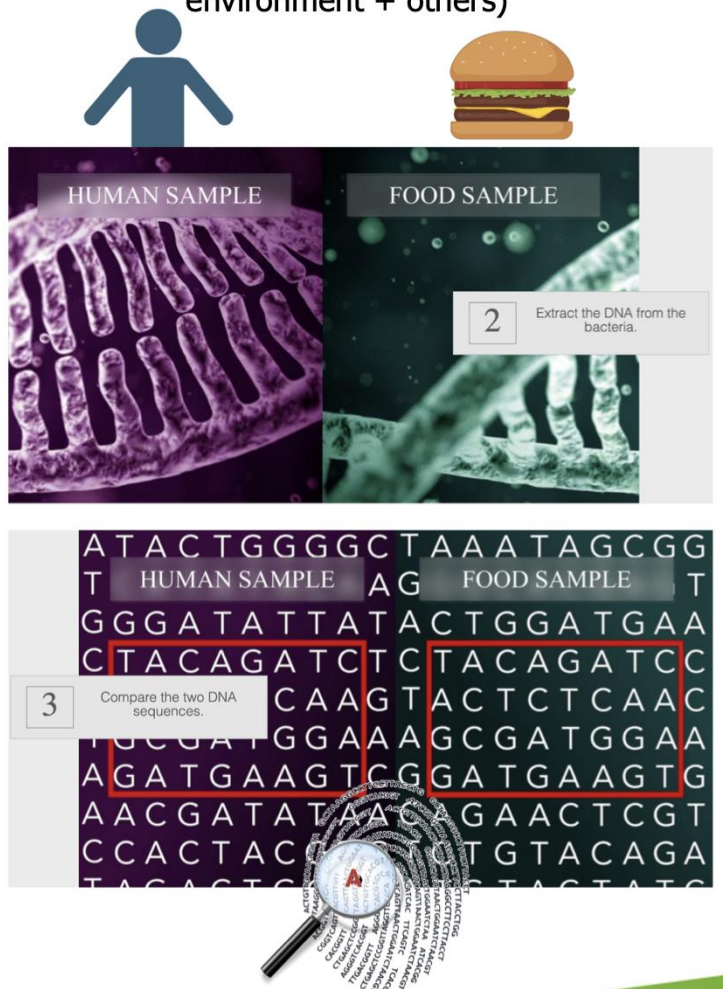


FIGURE 1 | Hypothetical pathway of bacterial pathogens along the farm-to-fork continuum. Genetically unique subpopulations of bacterial pathogens (blue dots can grow in farms or other suppliers from reservoir populations. Representatives of one or more subpopulations can be transmitted to distributors and processing facilities, food, and to patients (red circles; steps 1a, 2, and 3). At each step, the founding bacteria may grow into multiple subpopulations. Product, environmental or clinical samples may yield representative bacteria from closely related subpopulations that are analyzed by the food safety laboratory (thin arrows). Farms and suppliers may have more than one customer, resulting in the dissemination of genetically identical bacteria to multiple distributors or food processing facilities (step 1b).

Guinane et al. 2019. Genomics of Foodborne Microorganisms. Chapter 35, Pightling et al. 2018. Front. Microbiol. 9

Compare microorganisms of the **same species** from **different samples** (clinical + food + environment + others)



Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Typing Methods

Phenotypic

01

- **Serotyping**
- Phagotyping
- Biotyping
- Antimicrobial Susceptibility (AST)

Genotypic/ Molecular

02

- **PFGE**
- **MLST**
- MLVA
- RFLP
- FT-IR (spectrometric)

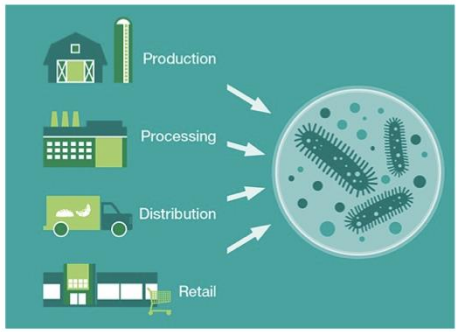
Whole Genome Sequencing

03

WGS

New "Golden Age" of Microbiology Metagenomic

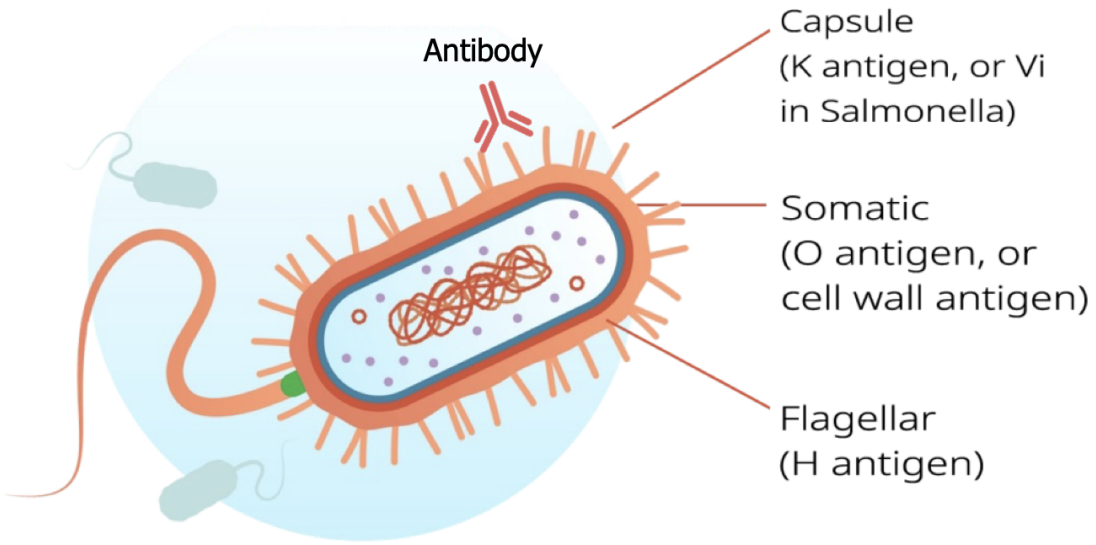
04



Novais et al. 2019. Eur J Clin Microbiol Infect Dis. 38(3):427-448

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Serotyping



Mughini-Gras et al. 2020. Front Microbiol. 10:2578

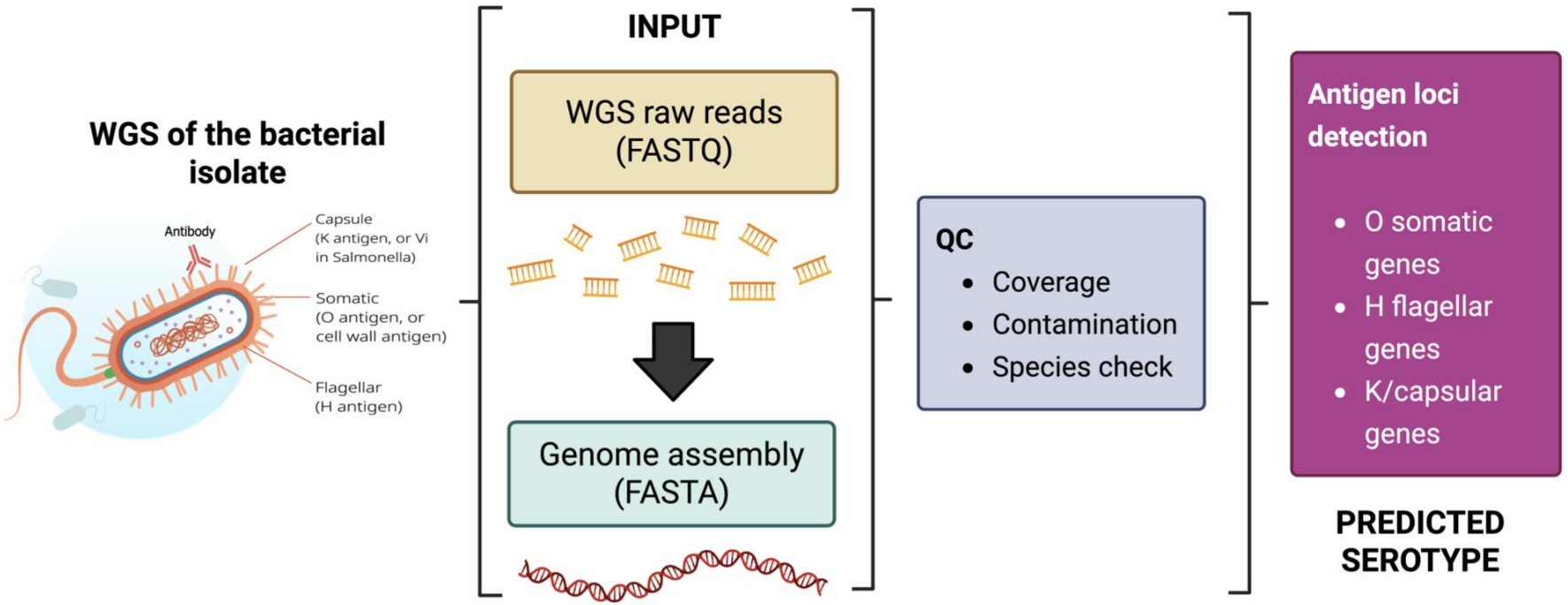
Based on **antigen-antibody** agglutination reactions ➔ serotype represented by an **antigenic formula** (O:H:K)

- Labor- and time-consuming
- Expensive (several sera needed)
- Restricted application

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

In silico serotyping

Addresses wet-lab serotyping limits by using **WGS data** to **analyze antigen loci** (O-somatic, H-flagellar, K/capsular) **directly** → portable, reproducible



Scalable: works with genome raw reads or assemblies; analysis of thousands of genomes at the same time.

Uelze et al. 2020. One Health Outlook. 2(3); Mughini-Gras et al. 2020. Front Microbiol. 10:2578; Image created with BioRender.com

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Molecular Typing Methods

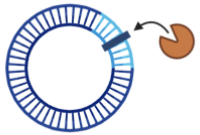
Strain discrimination from genetic content — DNA is unique to each microorganism

Food or other type of matrix

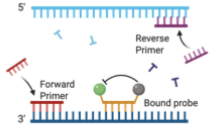
Culture and isolation

Identification at the genus and/or species level

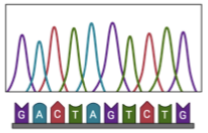
Extraction of total/genomic DNA - isolates to type



Restriction profile–based techniques



Amplification-based techniques



Sequencing-based techniques

Without amplification

- ***Pulsed-Field Gel Electrophoresis (PFGE)***

With amplification

- *Polymerase Chain Reaction–Restriction Fragment Length Polymorphism **(PCR–RFLP)***

- ***Multiple-Locus Variable-Number Tandem Repeat Analysis (MLVA)***
- *Random Amplified Polymorphic DNA **(RAPD)***
- *Amplified Fragment Length Polymorphism **(AFLP)***

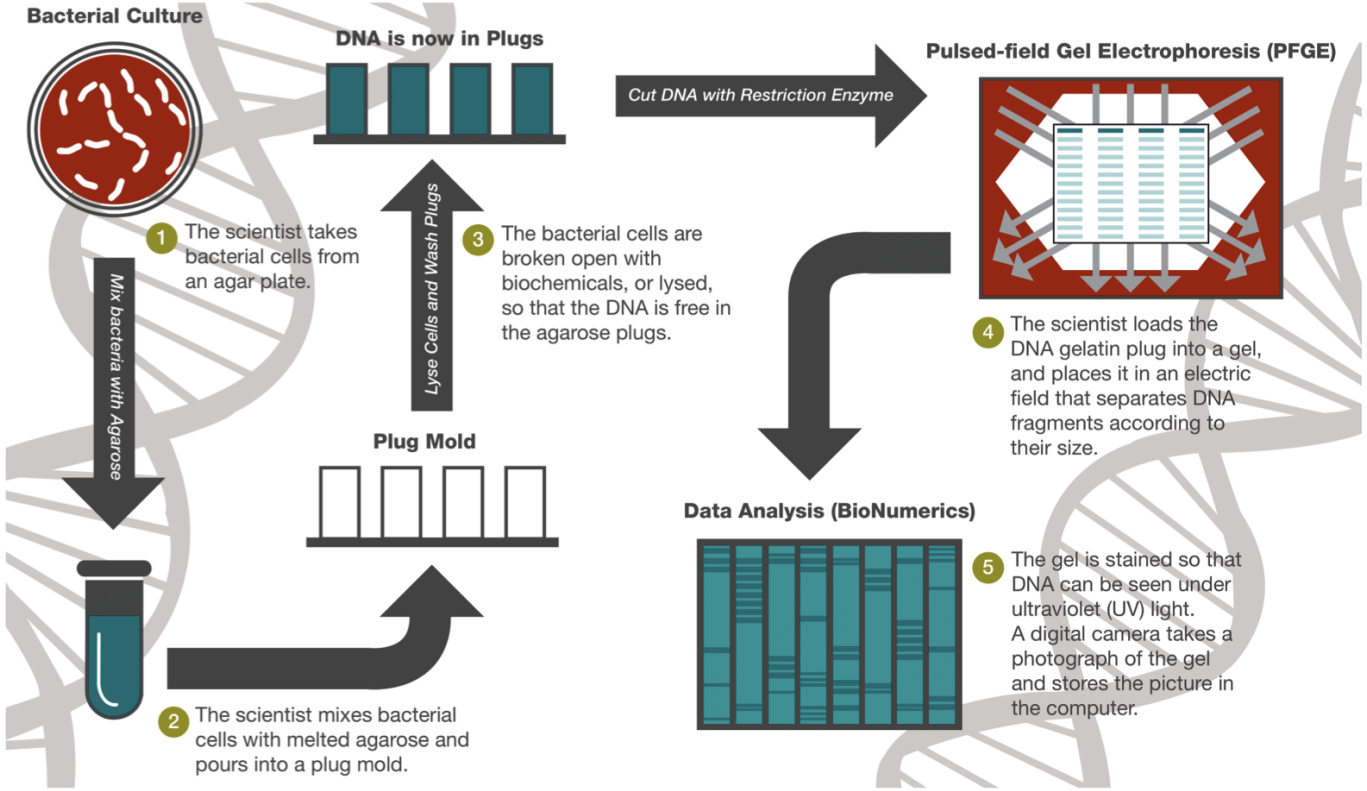
- ***Multilocus Sequence Typing (MLST)***
- ***Single-Nucleotide Polymorphism–based analysis (SNPs) (via WGS — whole-genome sequencing)***

Sabat et al. 2013. Euro Surveill. 18(4):):pii=20380; Image created with BioRender.com

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Pulsed-Field Gel Electrophoresis - PFGE

Divides a bacterial population into pulsotypes: macro-restriction of specific DNA sites with restriction enzymes, followed by separation of the fragments by pulsed-field gel electrophoresis.



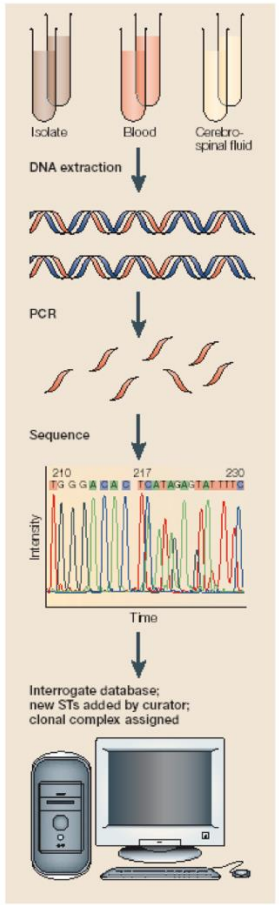
- Time-consuming assay (≈3–5 days)
- Limited inter-laboratory reproducibility
- Low throughput (limited number of strains per run)
- Difficult to interpret without software
- Profiles depend on enzyme and run conditions

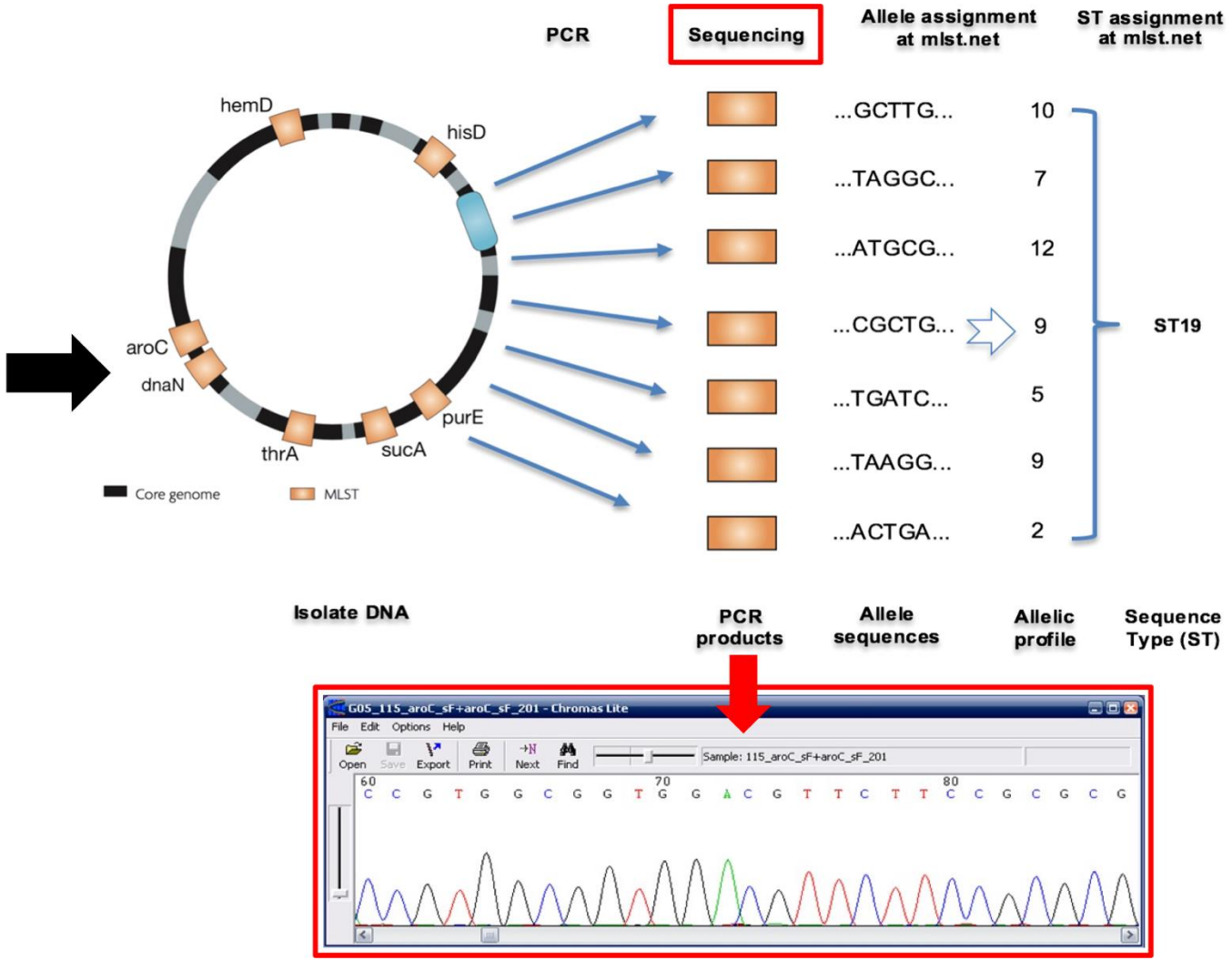
Largely replaced by WGS for surveillance

Li et al. 2009. FEMS Microbiol Rev. 33(5):892-916, EFSA Journal. 2019. 17(12):5898, https://www.cdc.gov/pulsenet/pathogens/pfge.html

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Multilocus Sequence Typing - MLST





- Detects sequence variation in ~7 housekeeping genes — conserved, but with some diversity
- Each unique allelic profile is assigned a **Sequence Type (ST)**

✅ Widely used, standardized nomenclature

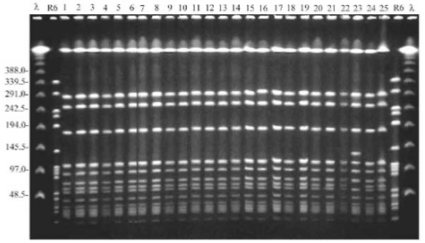
❌ Limited resolution, labor-intensive, costly

Online Salmonella MLST minimum spanning tree (MST): https://lfz.corefacility.ca/sistr-app/ (select Phylogeny and MST)

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Whole Genome Sequencing-based Typing

Traditional typing methods
(e.g., serotyping, PFGE, MLVA, MLST)



- Difficult to automate
- Limited digital integration
- Low/moderate power of discrimination
- Specific for each microorganism
- Time-consuming

Methods based on **(total) DNA sequencing**
(e.g., SNPs, cgMLST, wgMLST)

- **Compare "the whole" available genome**
- **High accuracy** and **discrimination** power
- Applicable **across organisms**
- **Data** can be **shared** between laboratories
- Cost and time – already competitive
- Enables global real-time surveillance

Guinane et al. 2019. Genomics of Foodborne Microorganisms. Chapter 35, Pightling et al. 2018. Front. Microbiol. 9

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

Whole Genome Sequencing-based Typing

PulseNet International
Connects the Dots for FOOD SAFETY

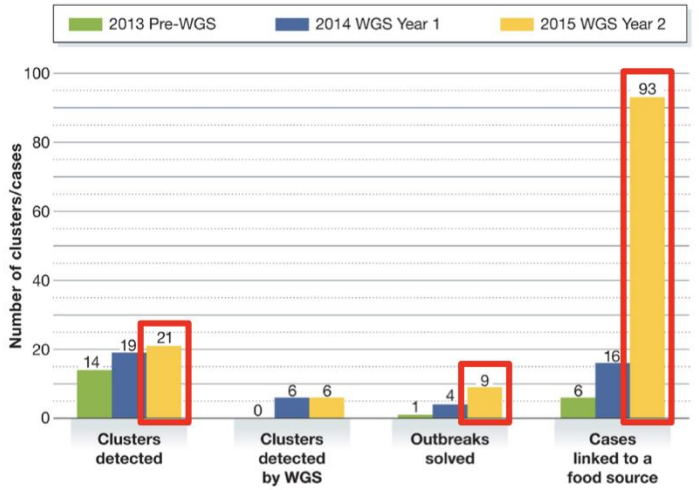
WGS is **replacing traditional methods** to type foodborne pathogens, as recommended by multiple organizations for outbreak detection, surveillance, and source tracking

REVIEW ARTICLES

PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance

PulseNet International is a global network dedicated to laboratory-based surveillance for food-borne diseases. The network comprises the national and regional laboratory networks of Africa, Asia Pacific, Canada, Europe, Latin America and the Caribbean, the Middle East, and the United States. The PulseNet International vision is the standardised use of whole genome sequencing (WGS) to identify and subtype food-borne bacterial pathogens worldwide, replacing traditional methods to strengthen preparedness and response, reduce global social and economic disease burden, and save lives. To meet the needs of real-time surveillance, the PulseNet International network will standardise subtyping via WGS using whole genome multilocus sequence typing (wgMLST), which delivers sufficiently high resolution and epidemiological concordance, plus unambiguous nomenclature for the purposes of surveillance. Standardised protocols, validation studies, quality control programmes, database and nomenclature development, and training should support the implementation and decentralisation of WGS. Ideally, WGS data collected for surveillance purposes should be publicly available, in real time where possible, respecting data protection policies. WGS data are suitable for surveillance and outbreak purposes and for answering scientific questions pertaining to source attribution, antimicrobial resistance, transmission patterns, and virulence, which will further enable the protection and improvement of public health with respect to food-borne disease.

https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202500179



Listeria monocytogenes **surveillance pre and post WGS implementation**

It offers higher **accuracy** and **discrimination** than PFGE (old gold standard) — detecting more outbreaks, tracing contamination sources, and identifying sporadic cases

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

WGS Mandated for EU Outbreak Investigations

Official Journal
of the European Union

EN
L series

2025/179

3.2.2025

COMMISSION IMPLEMENTING REGULATION (EU) 2025/179

of 31 January 2025

on the collection and transmission of molecular analytical data within the frame of epidemiological investigations of food-borne outbreaks in accordance with Directive 2003/99/EC of the European Parliament and of the Council

(3) Whole genome sequencing (WGS) is a modern molecular analytical technique for microbiological studies which facilitates greatly the swift identification of clusters of microorganisms, supporting the epidemiological investigations. It enables to establish links between isolates of food-borne pathogens recovered from humans, food, animals, feed and the related environment during the outbreak investigation.

(4) To substantially facilitate food-borne outbreak investigations and the timely detection of the sources of those outbreaks, Member States should be required to collect *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli*, *Campylobacter jejuni* and *Campylobacter coli* isolates derived from food, animal, feed and related environmental samples from food and feed business operators and during official controls, where those isolates are associated or suspected to be associated with a food-borne outbreak. Member States should also be required to carry out WGS on those isolates.

2. The competent authority shall carry out WGS on at least one isolate of each serovar, biotype or molecular type of the collected isolates referred to in paragraph 1 of this Article in an official laboratory which is referred to in Article 37 of Regulation (EU) 2017/625 of the European Parliament and of the Council (5) and which is accredited for this method according to ISO 17025.

WGS now mandatory in EU

WGS-based typing has become **mandatory** in the **EU** for foodborne **outbreak investigations** of major foodborne pathogens

https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202500179

Adapted from Joana Mourão, Technical University of Denmark (DTU), slides: Typing Tools: From PFGE to WGS, August 2025.

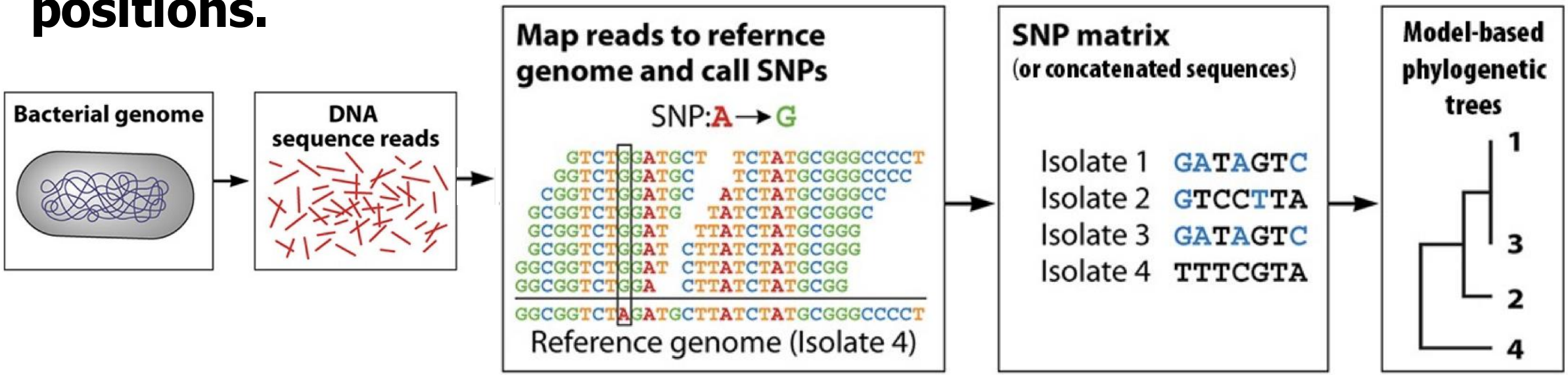
SNP based phylogeny

What is SNP-based phylogeny?

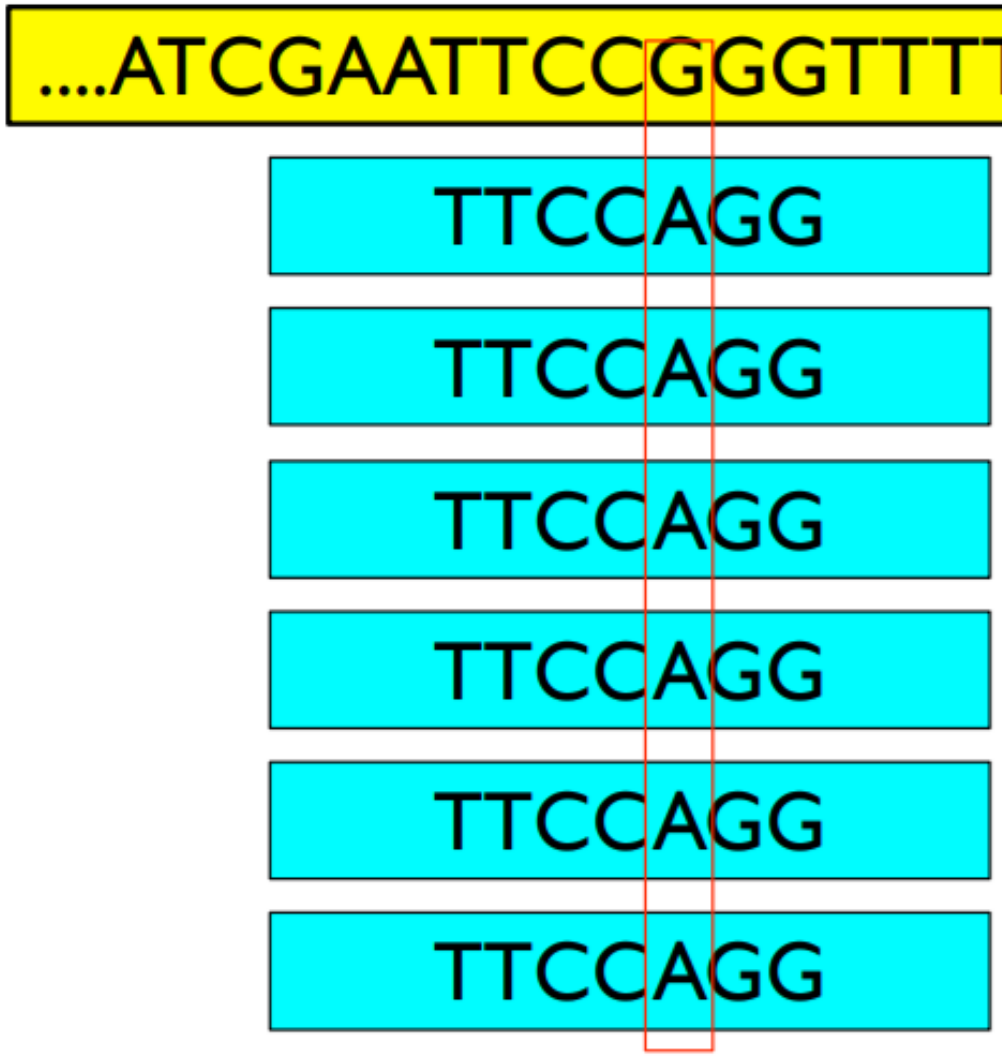
A SNP-based phylogeny is a way to compare bacterial genomes by:

- Identifying single nucleotide differences (SNPs)
- Comparing these differences across multiple isolates
- Inferring genetic relatedness between isolates

Closely related isolates differ at only a small number of SNP positions.



Reference genomes in SNP analysis



SNP-based phylogenies are reference-based:

- Sequencing reads are mapped to a reference genome
- SNPs are identified as differences from that reference

Choosing a good reference is critical:

- Should be closely related to the isolates

A distant reference can:

- Reduce mapping quality
- Reduce the number of comparable positions
- Bias SNP distances

Data quality

- SNPs are called only at:
  - High-quality positions
  - Positions covered in all isolates

- Low coverage or poor assemblies:
  - Increase missing data
  - Shrink the core genome
  - Can inflate SNP distances

More divergence from the reference → fewer shared positions for comparison.

Filtering and masking

To avoid misleading signal, SNP pipelines often:

- Remove low-quality positions
- Mask repetitive regions
- Mask recombination events

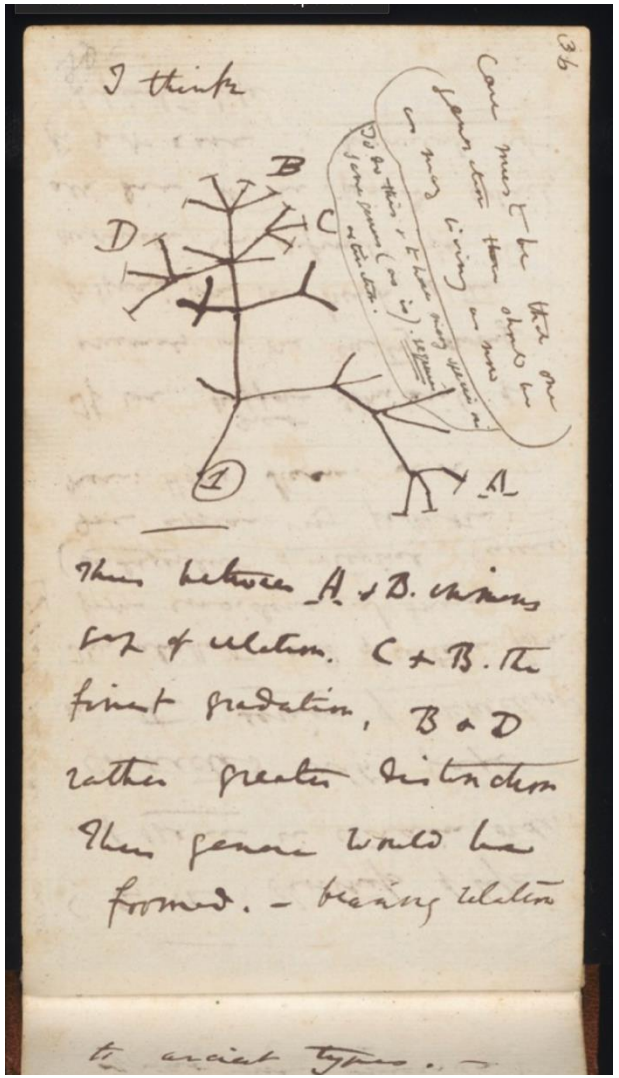
Why mask recombination?

- Recombination introduces clusters of SNPs
- These SNPs do not reflect clonal descent
- Masking improves outbreak resolution

Introduction to Phylogenetic Trees

Introduction to phylogenetic trees

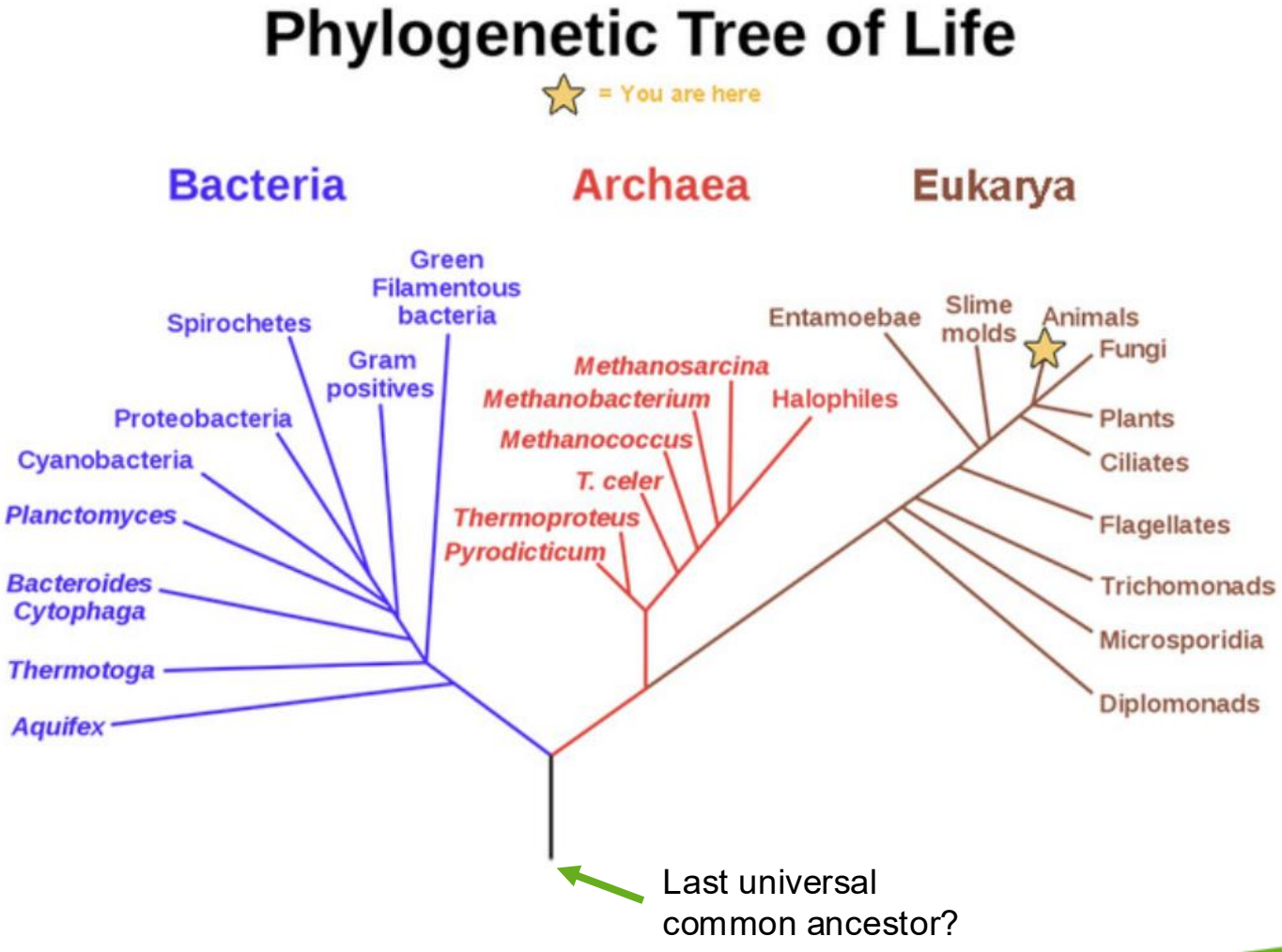
- In 1837 Charles Darwin sketched his “Tree of Life”
- First known visual representation of common ancestry
- Modern phylogenetics builds on this principle
- Origin on Species wasn’t published until 1859!



Darwin's Tree of Life sketch from his "B" notebook on Transmutation of Species, CUL DAR 121, page 36, Cambridge University Library - https://www.darwinproject.ac.uk/commentary/evolution/darwin-s-species-notebooks-i-think

Introduction to phylogenetic trees

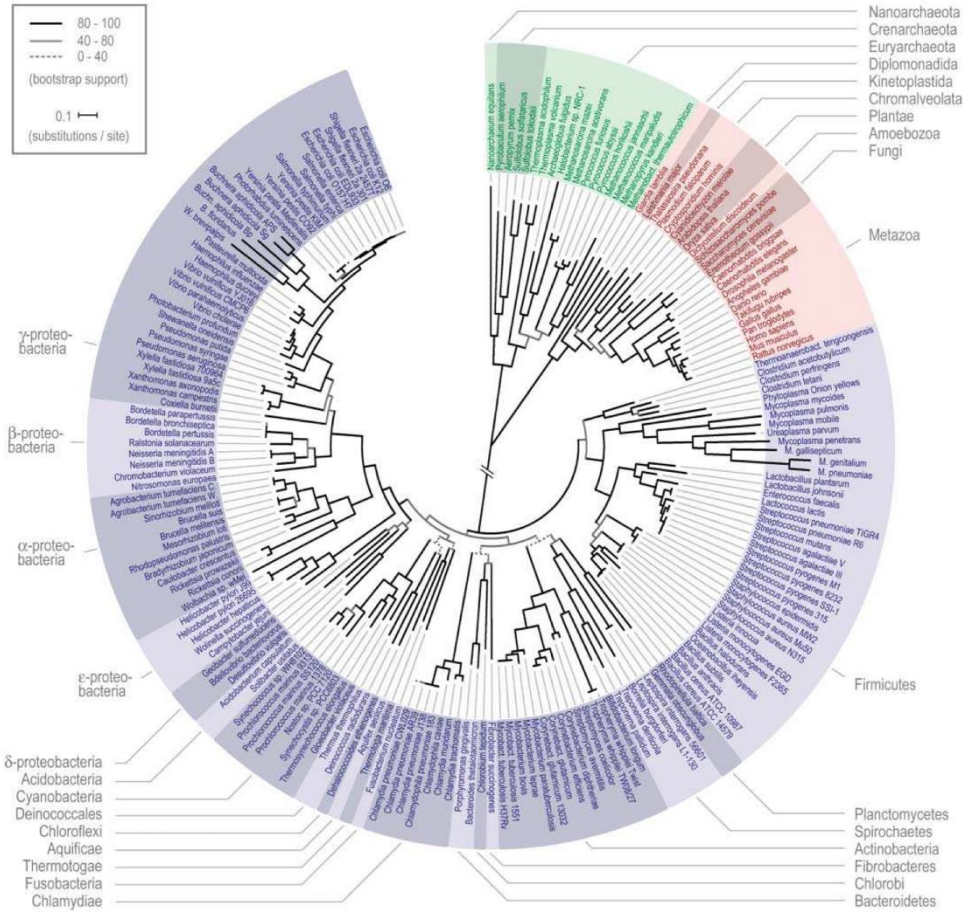
- Phylogenetic trees aim to describe the relationships between taxa (e.g. species, genera).
- They can be based on:
  - Phenotypic traits (e.g. hair colour, presence of wings).
  - Genotypic markers (e.g. restriction patterns, genes, SNPs, nucleotide/amino acid sequences).
- This tree was constructed by Carl Woese using ribosomal RNA gene sequences in 1977- a breakthrough that revealed the three domains of life (Bacteria, Archaea, Eukarya).

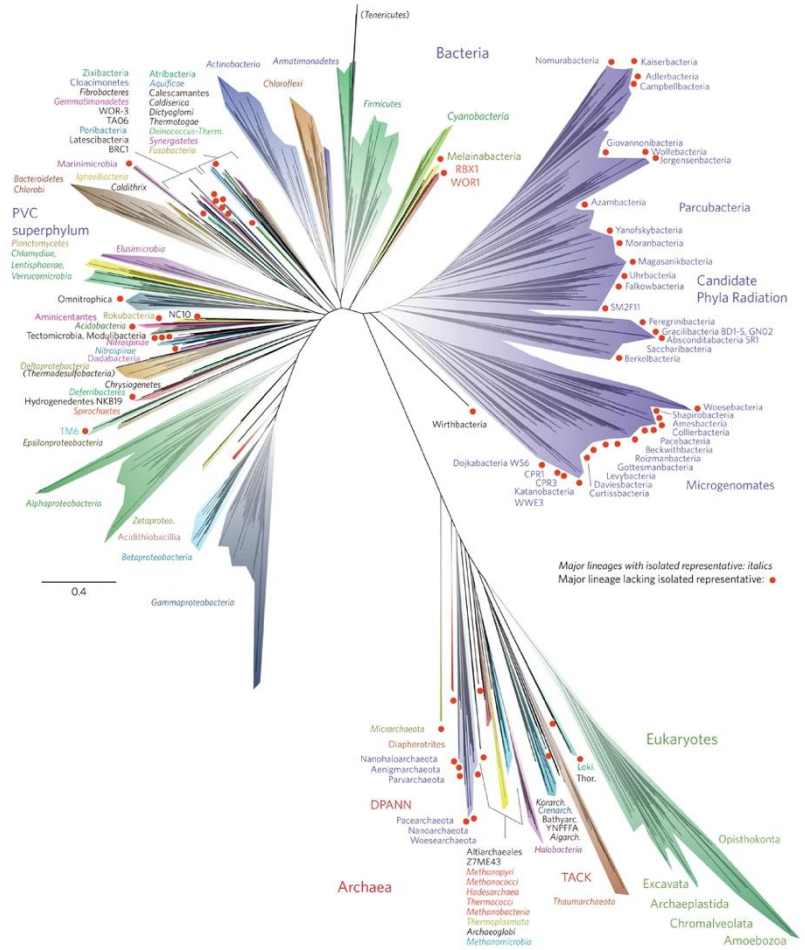


1.4: Carl Woese and the Phylogenetic Tree, LibreTexts Biology, https://bio.libretexts.org/Courses/Prince_Georges_Community_College/PGCC_Microbiology/01%3A_Introduction_to_Microbiology/1.04%3A__Carl_Woese_and_the_Phylogenetic_Tree

Introduction to phylogenetic trees

- Today trees are built on the comparison of whole genomes!





Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. science, 311(5765), 1283-1287.

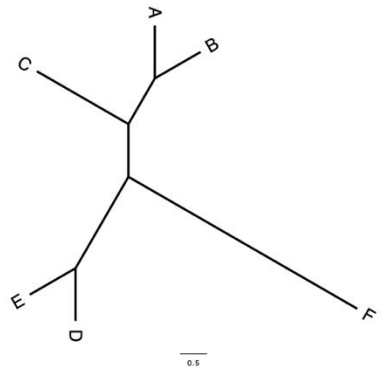
Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... & Banfield, J. F. (2016). A new view of the tree of life. Nature microbiology, 1(5), 1-6.

Reading Phylogenetic trees

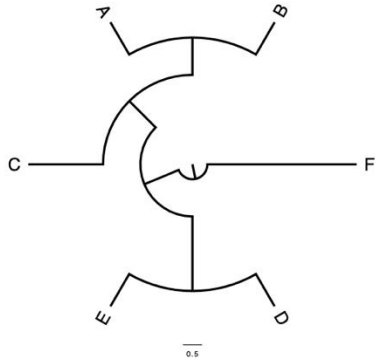
Reading phylogenetic trees

Different visualisations of the same tree!

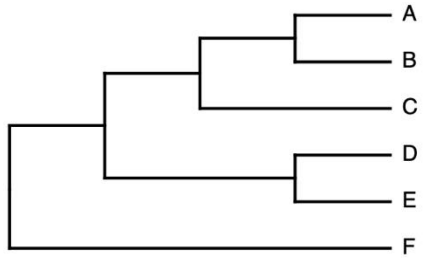
unrooted / polar



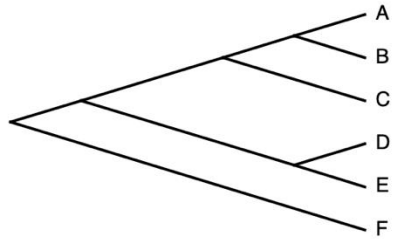
circular / radial



rectangular



(straight)

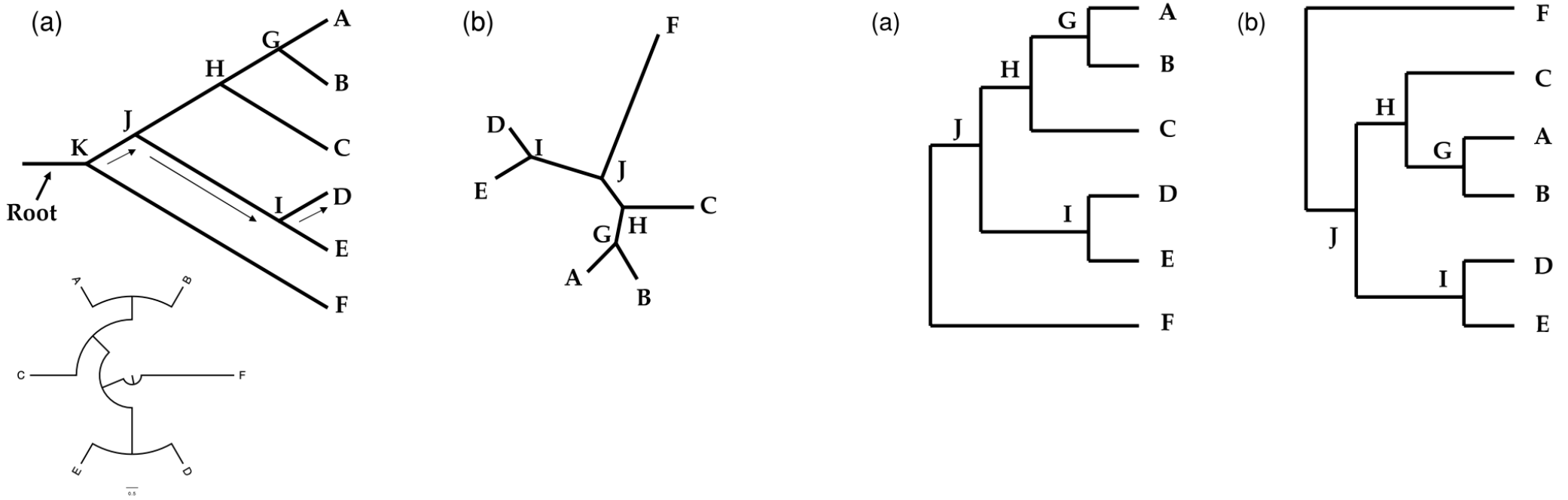


(curved)

Adapted from Raphael N. Sieber, Statens Serum Institut, Denmark, slides: “Phylogenetic analysis” Presentation, Bridging the Gaps, GenEpiBioTrain, March 2024.

Reading phylogenetic trees

- Always read the trees horizontally (left to right), never vertically.
- Rotations are arbitrary and can be confusing
- Rooted vs unrooted?



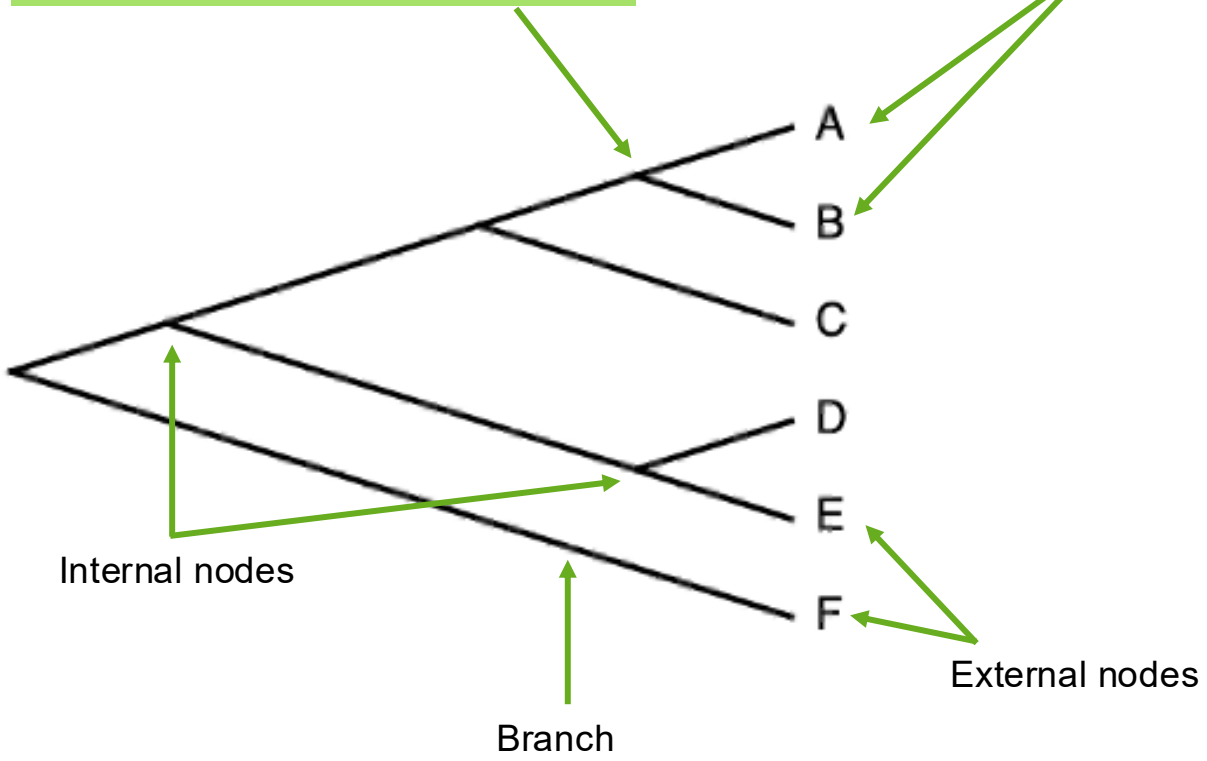
Reading phylogenetic trees - terminology

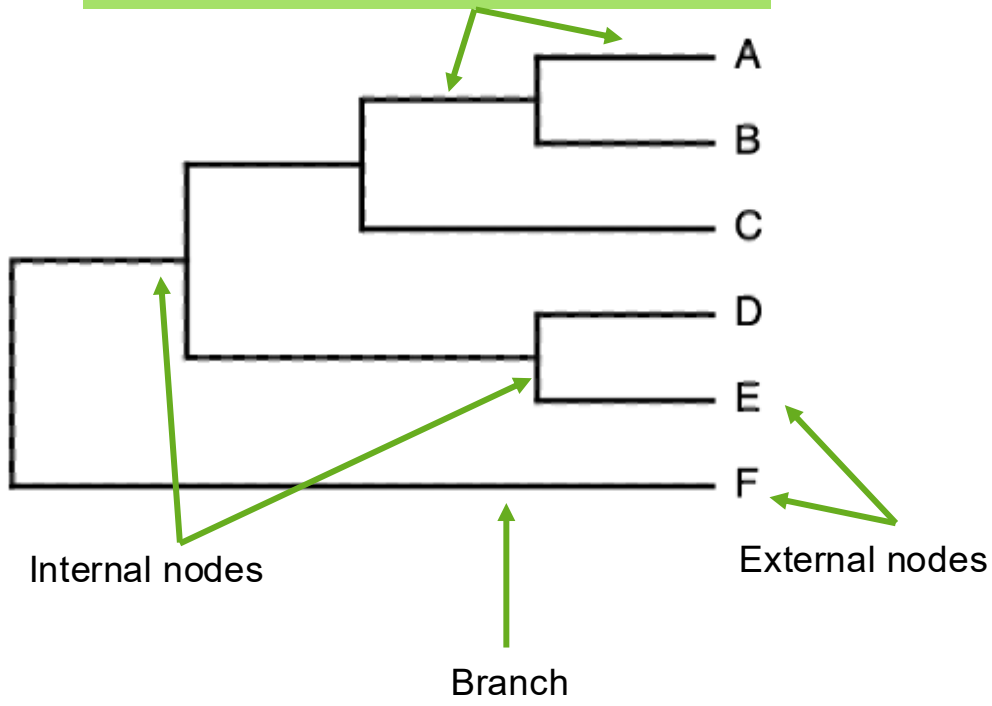
Internal nodes: the branching points. They're not observed directly - they're inferred ancestors where lineages diverge.

External nodes: the leaves or tips of the tree. These are the things you sequenced: bacterial isolates, species, or sequences.

Branches (edges): connect nodes. Their length reflects genetic distance -

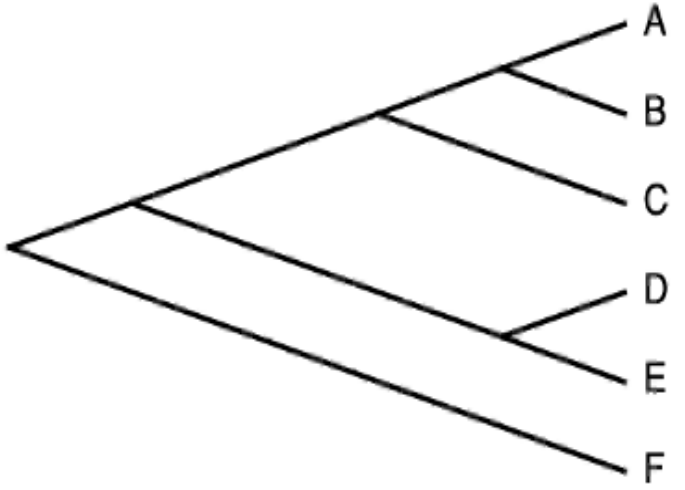
- In outbreak trees: SNPs or allele differences (clonal relatedness)
- In evolutionary trees: substitution rates / evolutionary distance

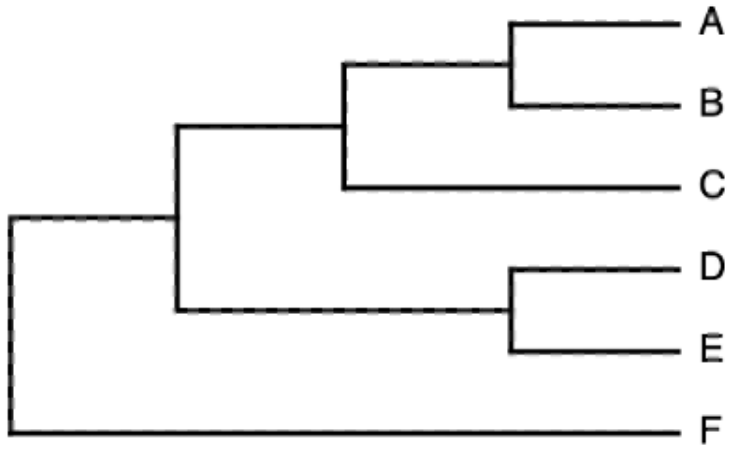




Adapted from Raphael N. Sieber, Statens Serum Institut, Denmark, slides: "Phylogenetic analysis" Presentation, Bridging the Gaps, GenEpiBioTrain, March 2024.

Reading phylogenetic trees – rooted and unrooted



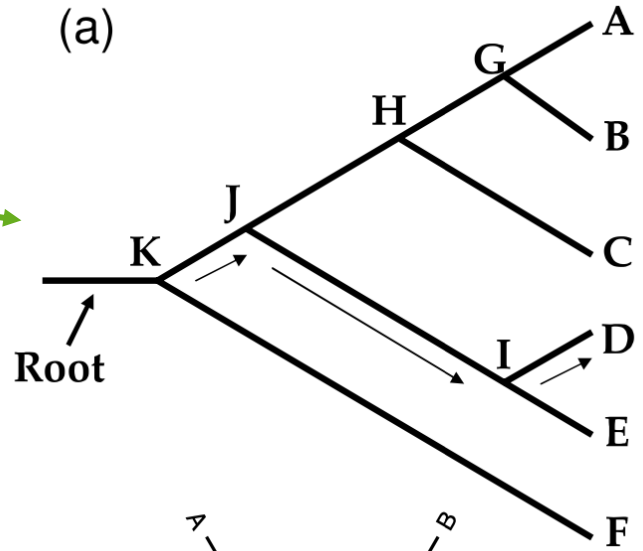


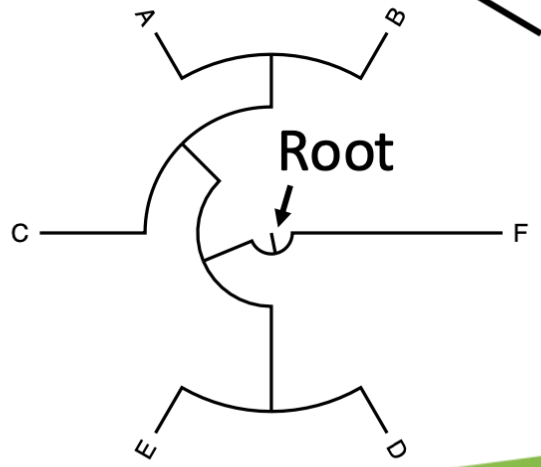
Rooted tree:

- Has a defined most recent common ancestor at the root
- Branch lengths may represent evolutionary distance or time

Unrooted tree:

- Shows relatedness but no direction of ancestry
- Can look rooted depending on display - always check the figure legend!





Adapted from Raphael N. Sieber, Statens Serum Institut, Denmark, slides: "Phylogenetic analysis" Presentation, Bridging the Gaps, GenEpiBioTrai

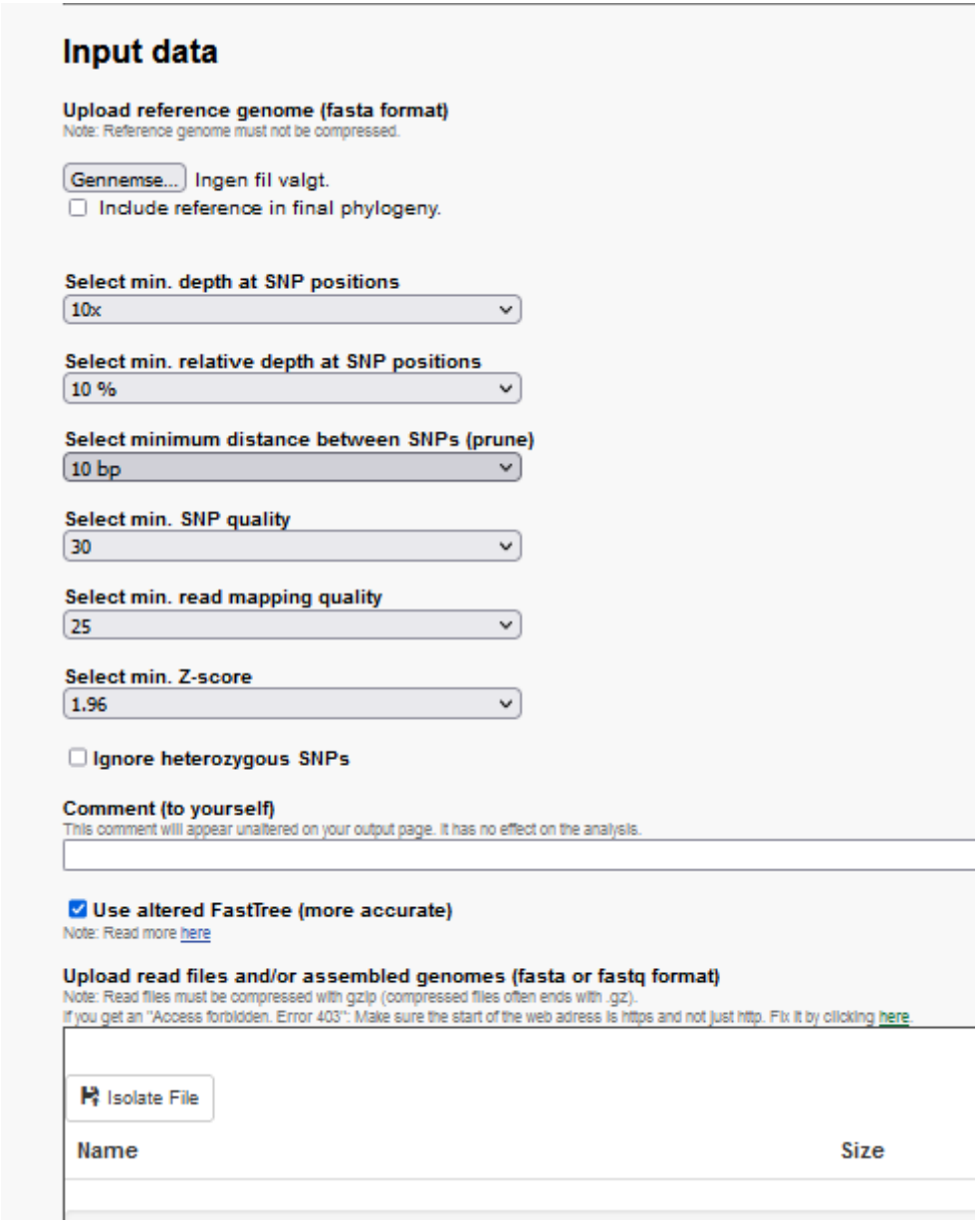
Interpreting SNP trees

Building a Phylogenetic Tree

1. Collect data
   - Whole genome sequences, genes, or proteins
2. Identify variation
   - Outbreak trees: SNP calling or cgMLST allele assignment
   - Evolutionary trees: Multiple sequence alignment (MSA) of sequences
3. Create a matrix
   - SNP/allele difference matrix (outbreaks)
   - Aligned sites from MSA (evolutionary)
4. Choose a tree-building method
   - Distance-based (Neighbor-Joining)
   - Maximum Parsimony
   - Maximum Likelihood
5. Visualize & annotate
   - Add metadata in iTOL, Microreact, GrapeTree

CSI Phylogeny

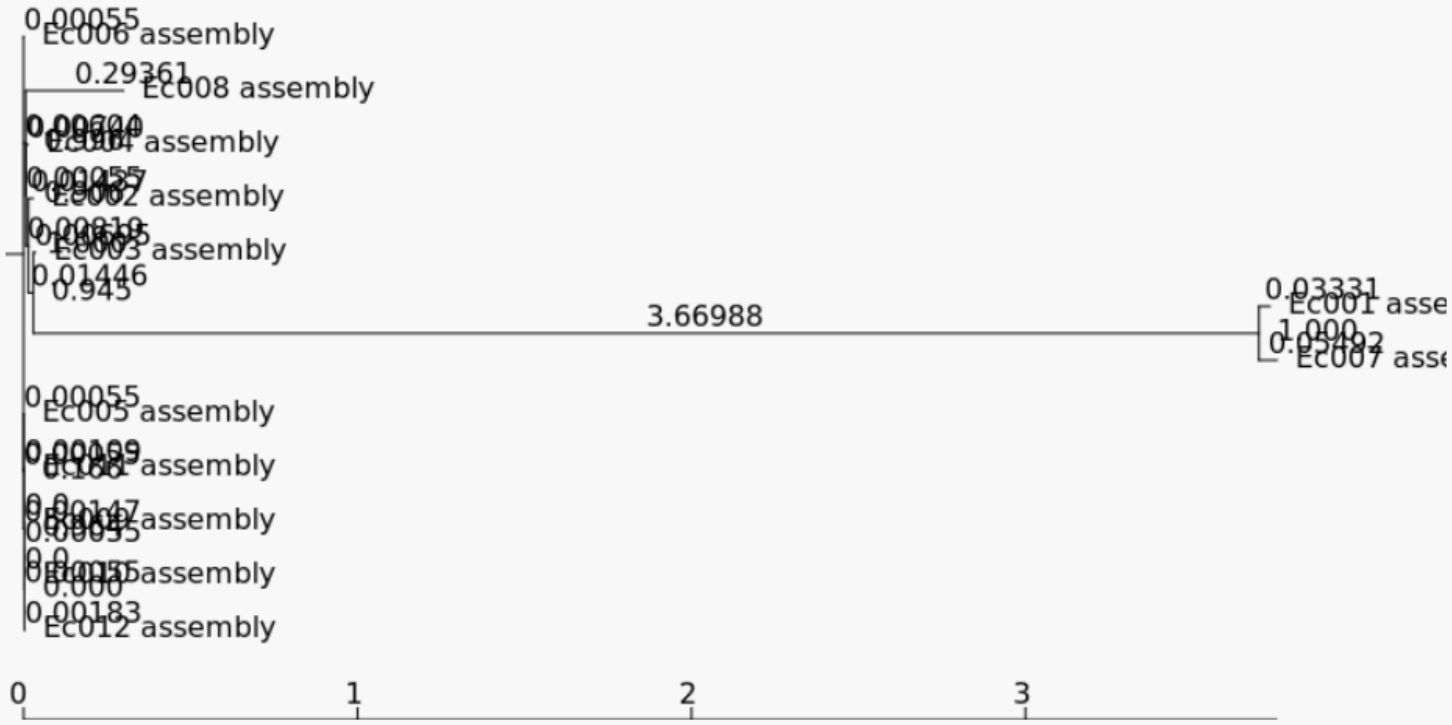
- Webtool on the CGE website for doing SNP calling (designed for Illumina data, use Mintyper for ONT data)
- You can run both FASTA and FASTQ files
- Reference must be in FASTA format
- https://cge.food.dtu.dk/services/CSIPhylogeny/



CSI Phylogeny - Output

CSIPhylogeny Results

The tree presented in the picture below is only meant as a preview. If the tree is meant to be shared or published, we strognly recommend that the 'Newick' file is downloaded and processed using software created for this purpose. We suggest (FigTree).



Dowload phylogeny as: Newick PDF SVG

CSI Phylogeny - Output

Percentage of reference genome covered by all isolates: 90.018466029857
4383433 positions was found in all analyzed genomes.
Size of reference genome: 4869482

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

File	Valid positions	Pct. of reference
Ec002_assembly.ignored_snps	4582041	94.0970928735336
Ec003_assembly.ignored_snps	4593110	94.324406579591
Ec001_assembly.ignored_snps	4612609	94.7248393155576
Ec009_assembly.ignored_snps	4544249	93.3209938962707
Ec005_assembly.ignored_snps	4607895	94.6280323040521
Ec004_assembly.ignored_snps	4607871	94.6275394384865
Ec008_assembly.ignored_snps	4578451	94.0233683993493
Ec007_assembly.ignored_snps	4622929	94.936771508756
Ec011_assembly.ignored_snps	4567232	93.792974283507
Ec012_assembly.ignored_snps	4586595	94.1906141146019
Ec006_assembly.ignored_snps	4610605	94.6836850408319
Ec010_assembly.ignored_snps	4570946	93.8692452297801

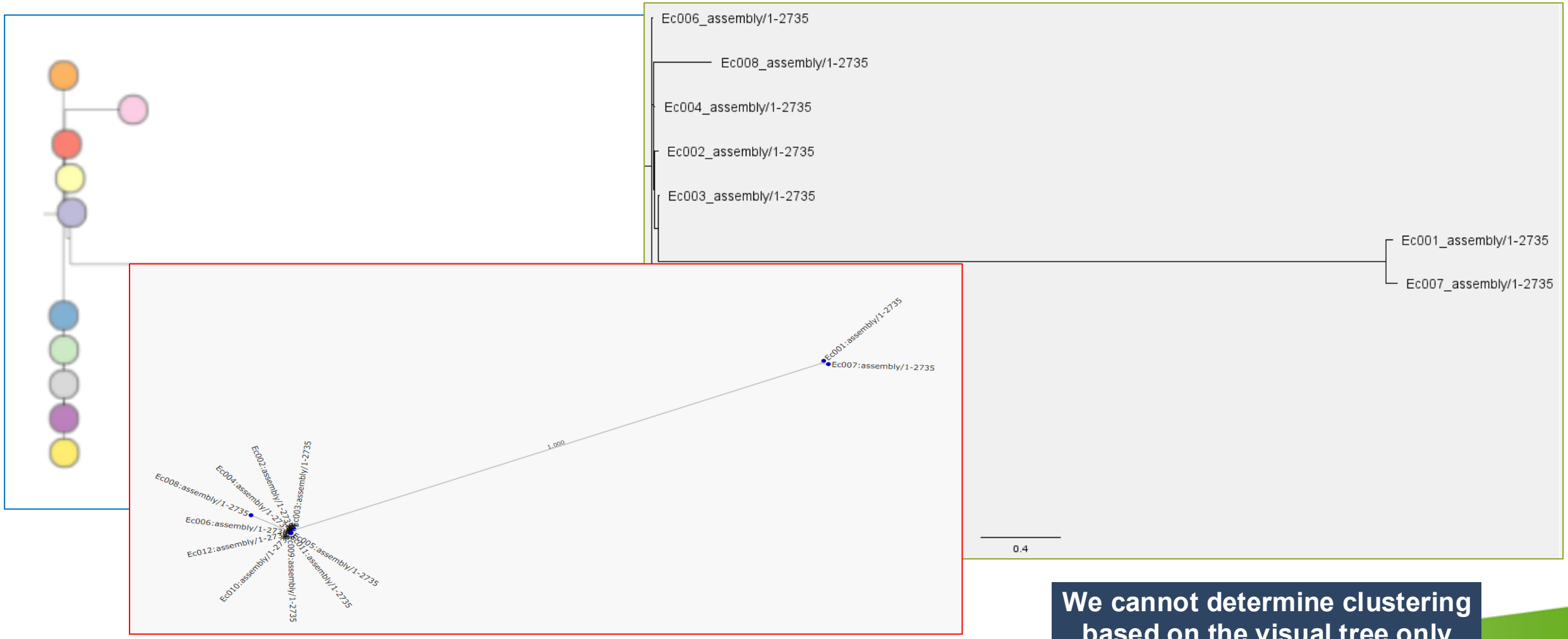
CSI Phylogeny - Output

Below is listed the number of positions, covered by at least one read, in each isolate that did not meet the minimum thresholds.

Index	Isolate	Ignored pos.
1	Ec004_assembly	0
2	Ec006_assembly	0
3	Ec005_assembly	0
4	Ec012_assembly	0
5	Ec009_assembly	0
6	Ec008_assembly	0
7	Ec010_assembly	0
8	Ec002_assembly	0
9	Ec007_assembly	0
10	Ec003_assembly	0
11	Ec001_assembly	0
12	Ec011_assembly	0

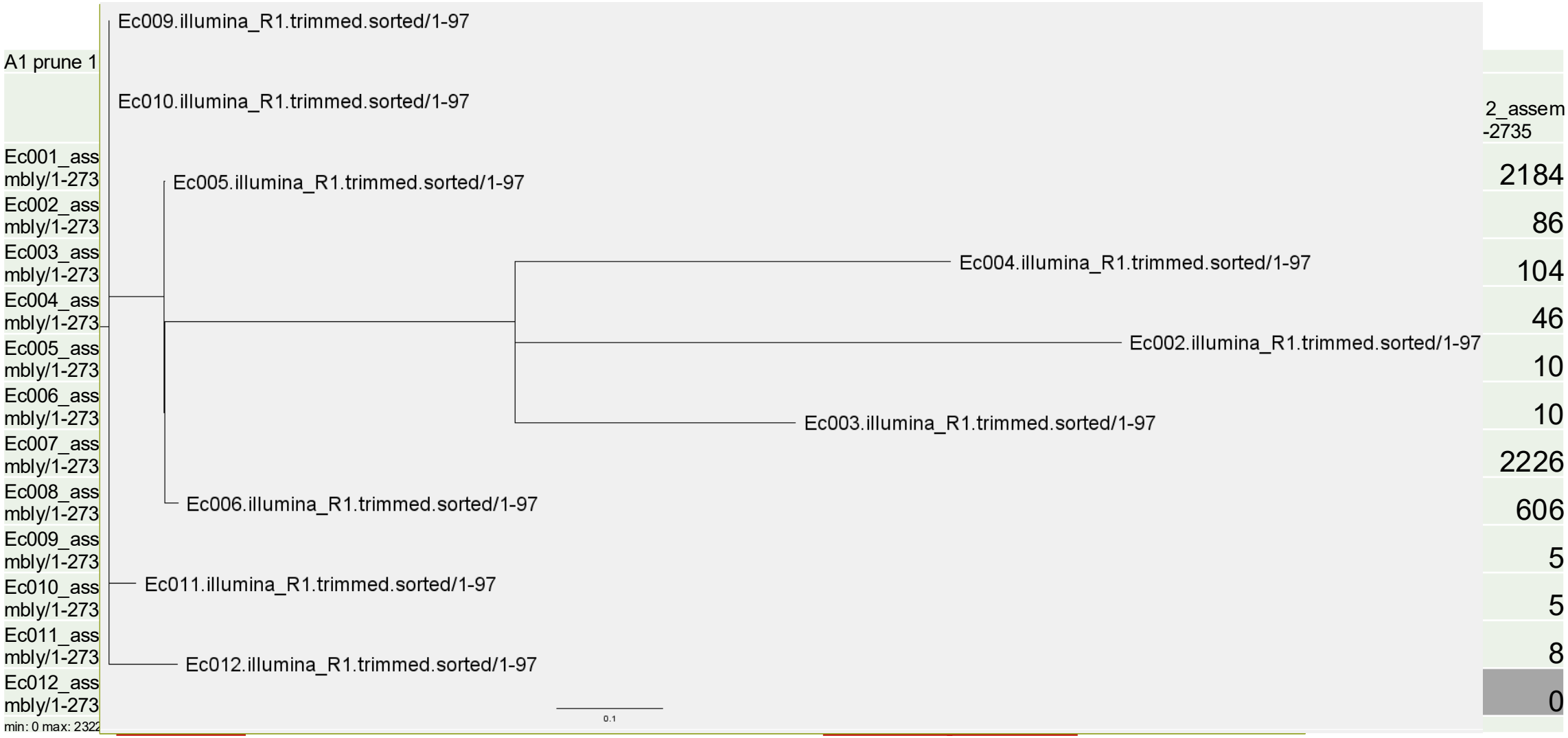
Visualisation of tree (newick file)

Dowload phylogeny as: Newick PDF SVG



We cannot determine clustering based on the visual tree only

SNP matrix – pairwise comparison of SNPs



Which pylogenetic analysis tools have you used before? (e.g. CSI Phylogeny, MinTyper, EnteroBase, cgMLSTFinder, GrapeTree, iTOL, PathogenWatch, snippy)

Thank you for listening! ☺

Acknowledgements

The creation of this training material was commissioned by ECDC to Institut Pasteur with the direct involvement of Niamh Lacy-Roberts, Research Assistant at DTU, nlac@food.dtu.dk