



Phylogenetic Tree Construction and Interpretation

How are SNPs and core-genome trees are used in epidemiology?

August 27, 2025, Niamh Lacy-Roberts, Research Assistant at DTU, nlac@food.dtu.dk

Intended Learning Objectives

Specific objectives of this session:

1. Define basic phylogenetic tree concepts and terminology (clade, branch length, synapomorphy, monophyly).
2. Describe different types of phylogenetic trees derived from WGS data (maximum parsimony, maximum likelihood, distance-based trees).
3. Explain how distance matrices and phylogenetic trees are applied in outbreak investigations.
4. Interpret SNP and allele-based trees to identify clusters and assess clonal relatedness.

Outline

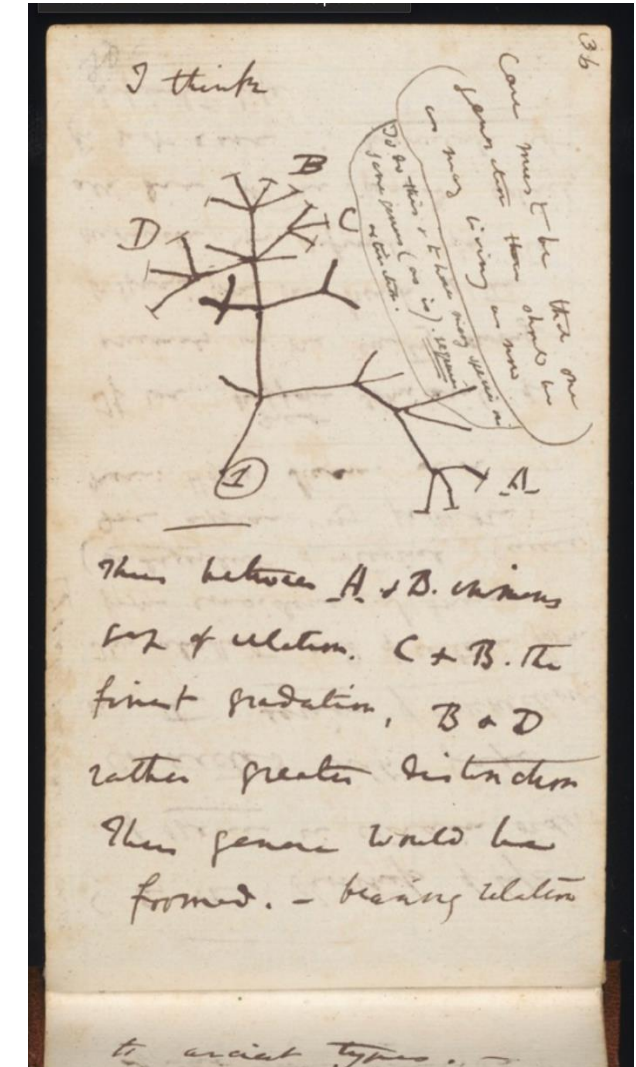
This session consists of the following elements:

1. Introduction to Phylogenetic Trees
2. Reading Phylogenetic trees
3. Building Phylogenetic Trees
4. Interpreting SNP and cgMLST trees

Introduction to Phylogenetic Trees

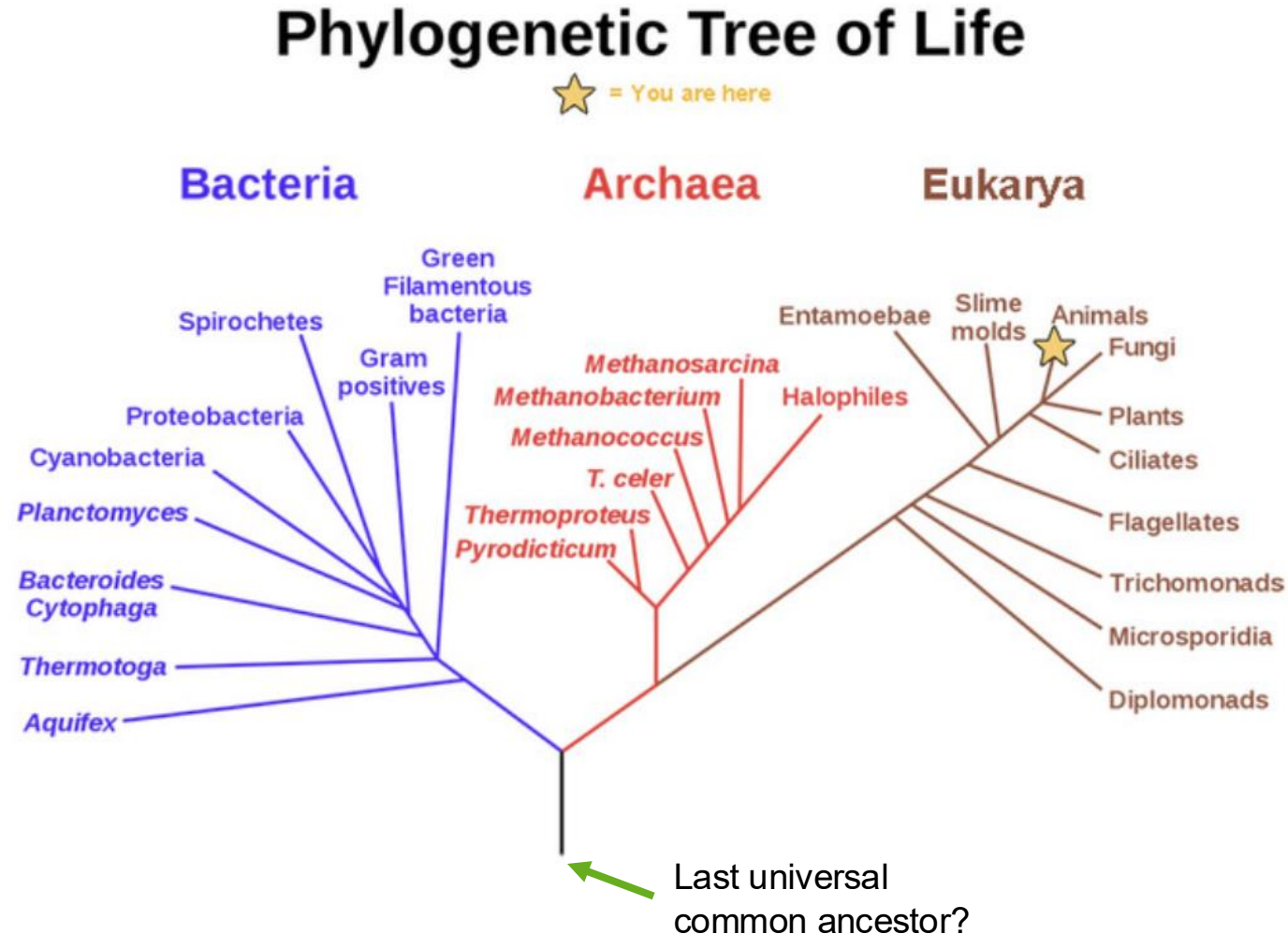
Introduction to phylogenetic trees

- 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!



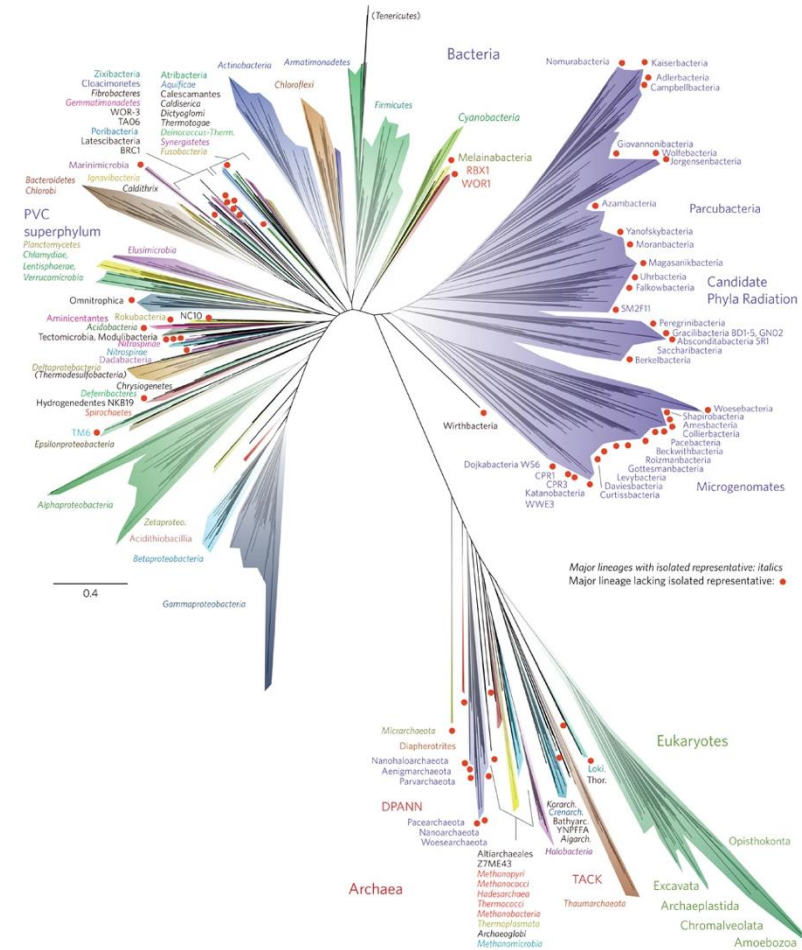
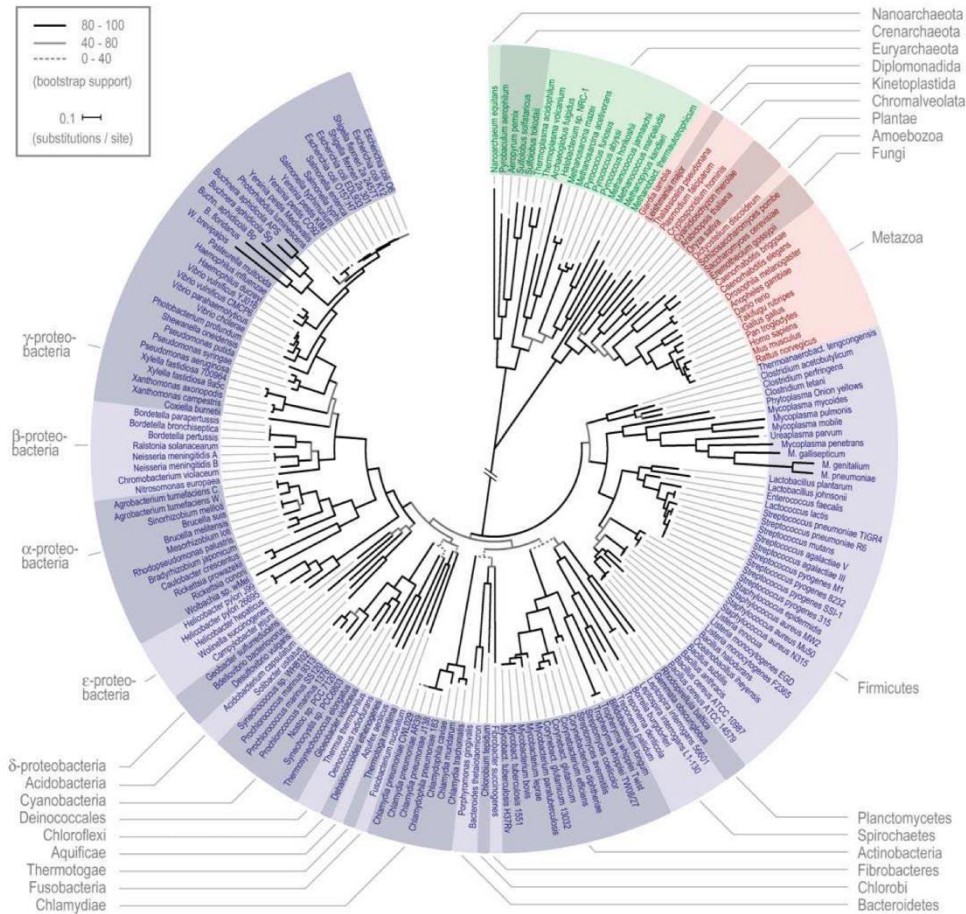
Introduction to phylogenetic trees

- 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).



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.

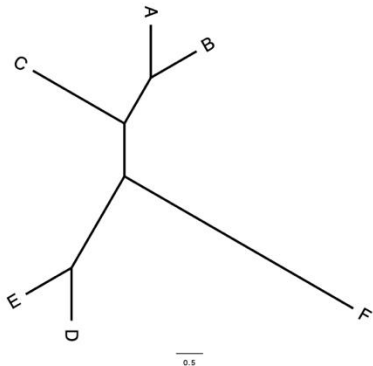
Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... & Banfield, J. F. (2016). A new view of the tree of life. *Nature microbiology*, 1(5), 1-6.

Reading Phylogenetic trees

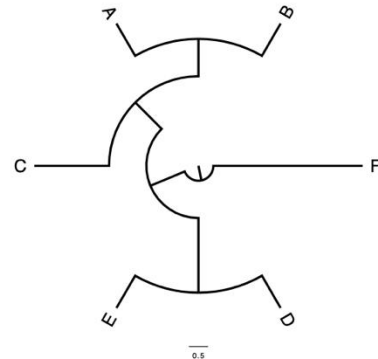
Reading phylogenetic trees

Different visualisations of the same tree!

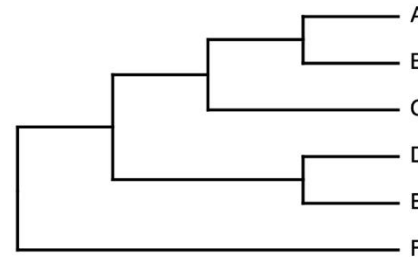
unrooted / polar



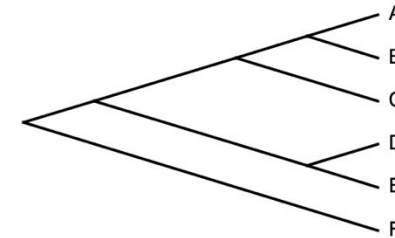
circular / radial



rectangular



(straight)

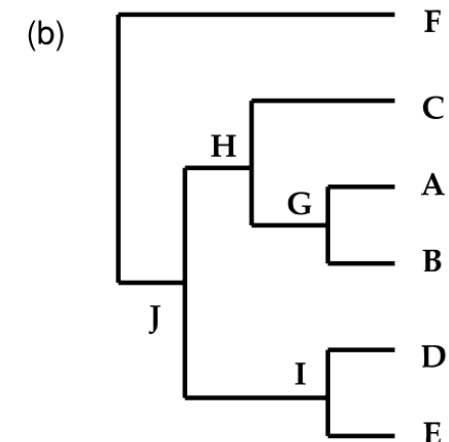
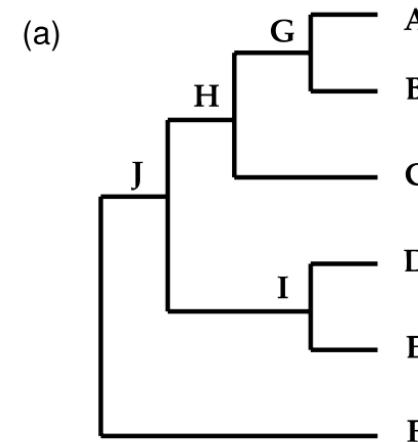
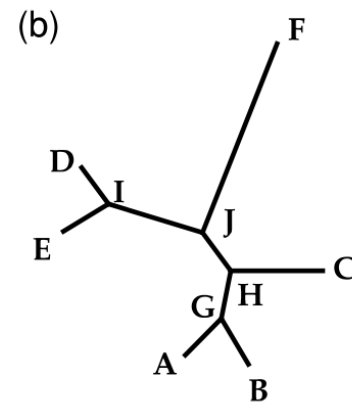
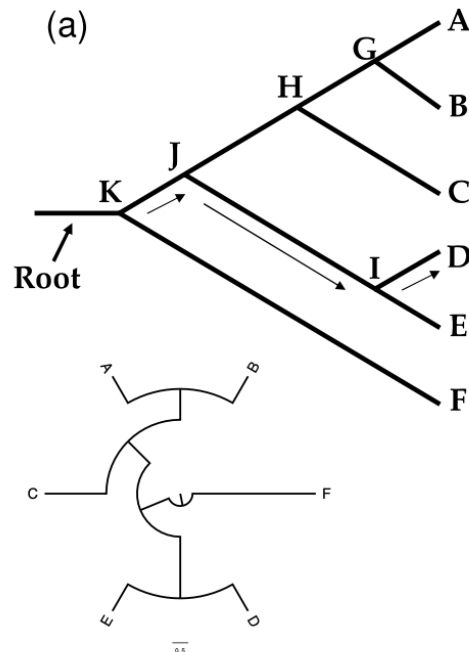


(curved)



Reading phylogenetic trees

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



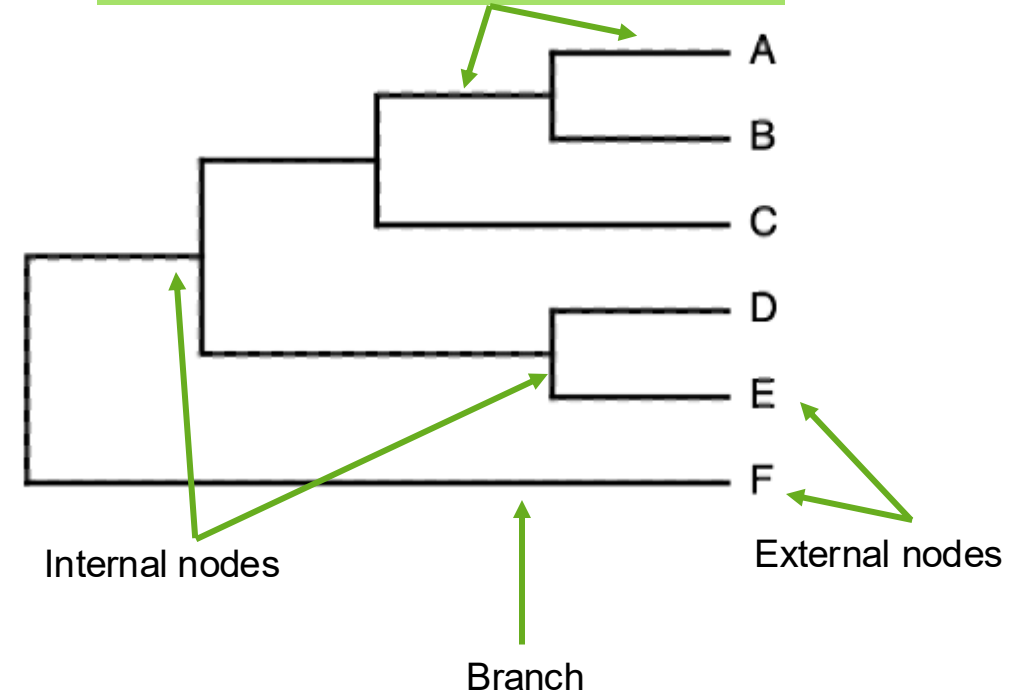
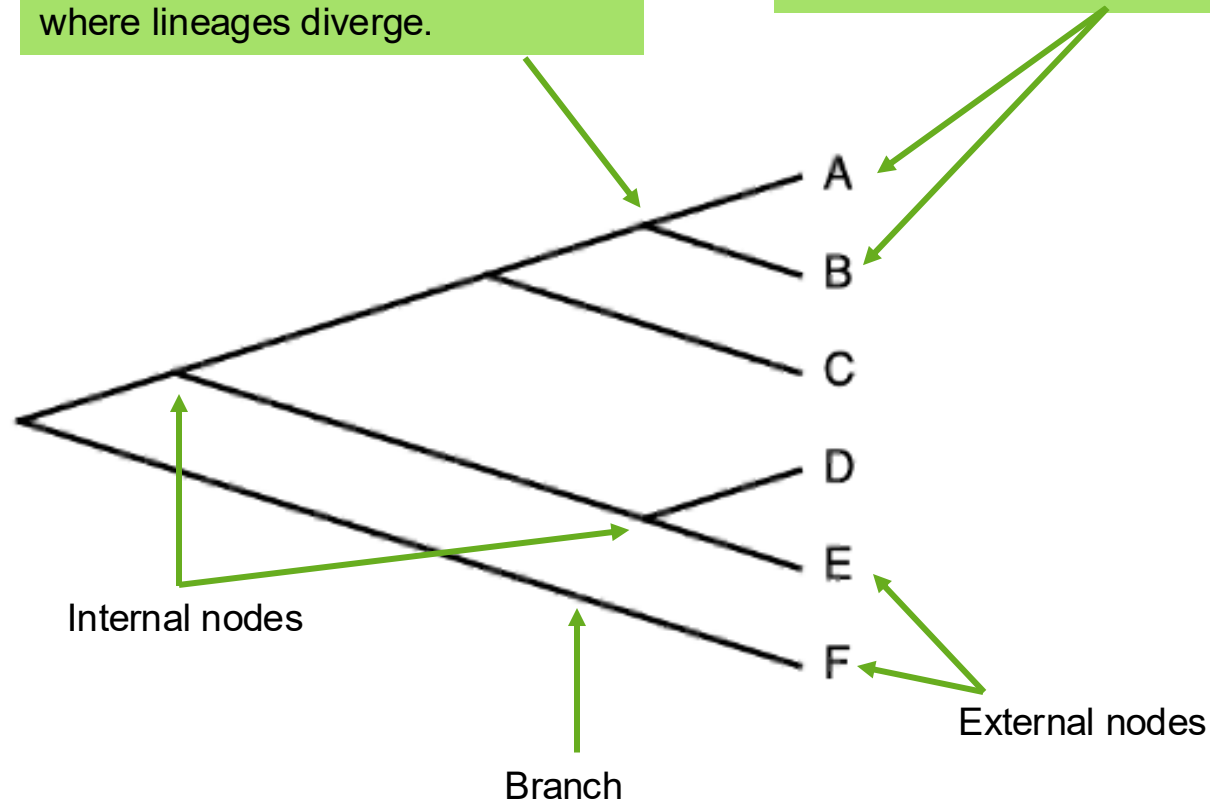
Reading phylogenetic trees - terminology

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

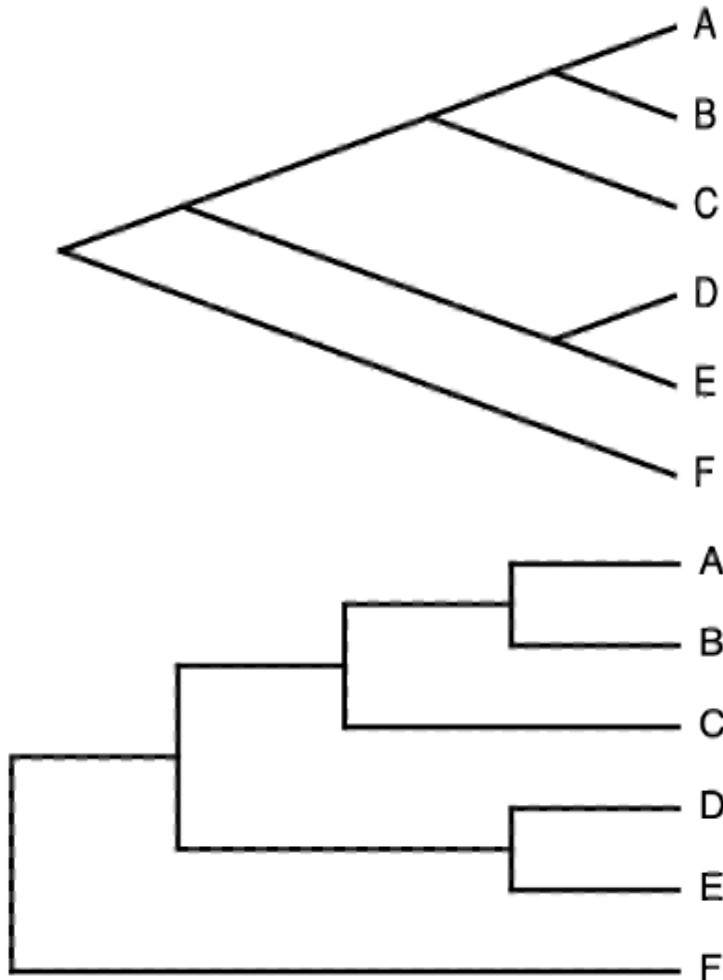
External nodes: the leaves or tips of the tree. These are the things you sequenced: bacterial isolates, species, or sequences.

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



Reading phylogenetic trees – rooted and unrooted

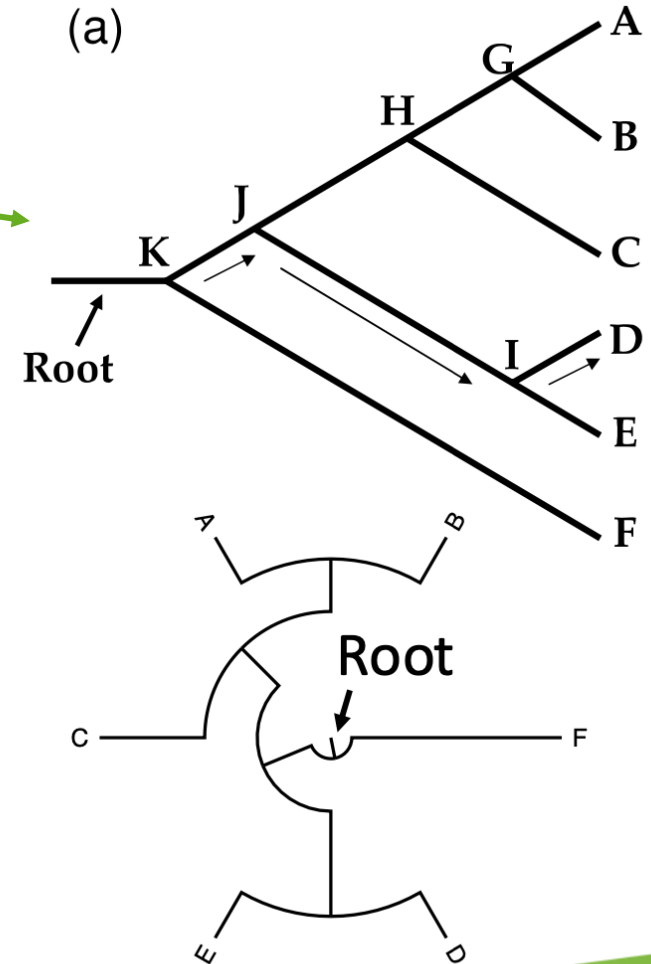


Rooted tree:

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

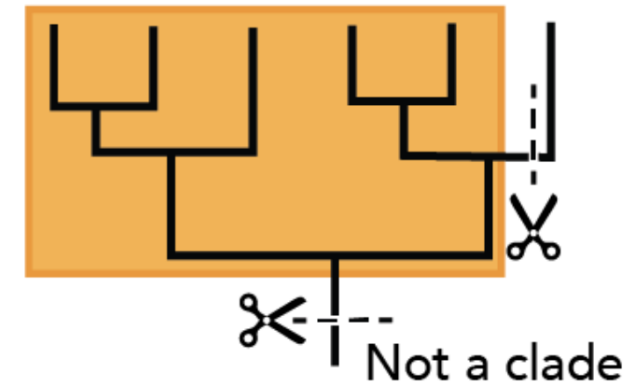
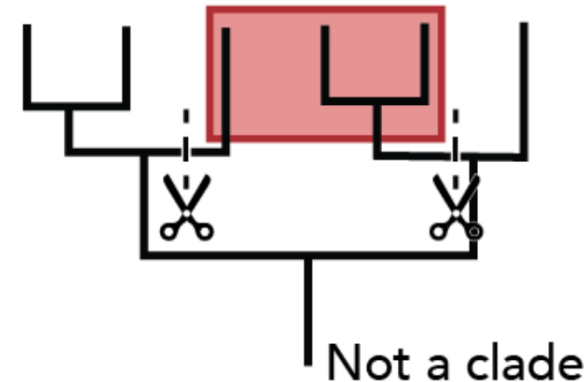
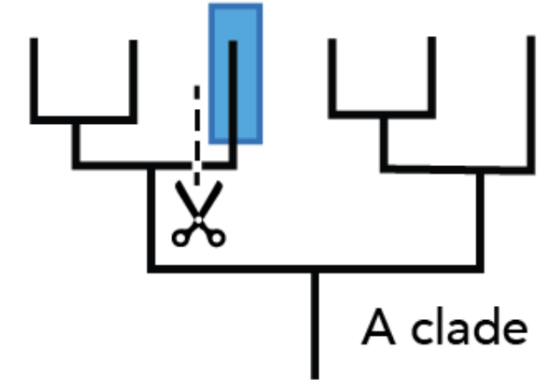
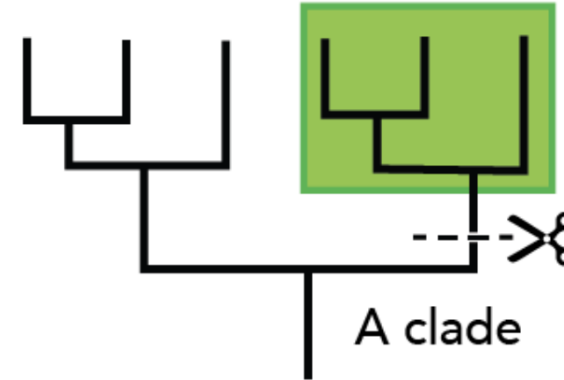
Unrooted tree:

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



Reading phylogenetic trees - Clades (Monophyletic Groups)

- A clade = an ancestor + all its descendants
- Represents an unbroken line of descent
- One cut rule: if a single cut on the tree separates the group, it's a clade
- More than one cut → not a clade (paraphyletic / polyphyletic):
- Polyphyletic: a group of organisms that does not include the most recent common ancestor of those organisms.
- Paraphyletic: a group of organisms that includes the most recent common ancestor of all of its members but does not include all of the descendants of that most recent common ancestor.



Building 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



Building a Phylogenetic Tree – different methods



Tree method	Maximum Parsimony (MP)	Maximum Likelihood (ML)	Distance-based (Neighbor-Joining, UPGMA)
Overview	- Parsimony = simplicity. - Finds the tree that requires the fewest genetic changes.	- Uses a statistical model of sequence evolution (substitution model). - Evaluates many possible trees and chooses the one with the highest probability of producing the observed data.	- Starts with a distance matrix (pairwise SNP or allele differences). - Groups taxa iteratively to minimize overall branch length.
Strengths	- Intuitive, easy to understand. - Works well for small datasets.	- More accurate and robust than parsimony or distance methods. - Can incorporate complex substitution models. - Gold standard for outbreak trees in bacterial genomics.	- Very fast, can handle large datasets. - Easy to compute and visualize.
Weaknesses	- Can oversimplify if there's recombination or parallel evolution. - Doesn't account for different probabilities of changes (e.g., transitions vs transversions in DNA).	Computationally intensive for large datasets.	- Less accurate than ML (can be misled if distances don't reflect true ancestry). - Doesn't explicitly use a model of sequence evolution.
Outbreak context	Useful for teaching or quick approximations but not widely used in modern outbreak pipelines because bacteria recombine, and MP doesn't handle that well.	Widely used in tools like RAxML, IQ-TREE, CSI Phylogeny. Often the method behind "final" outbreak trees because it's statistically robust.	- Good for exploratory analysis or quick clustering (e.g. GrapeTree, cgMLST analyses). - Often used as a first-pass to visualize relatedness before more rigorous ML analysis.

Interpreting SNP and cgMLST trees

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/>

Input data

Upload reference genome (fasta format)
Note: Reference genome must not be compressed.

Ingen fil valgt.

☐ Include reference in final phylogeny.

Select min. depth at SNP positions
10x

Select min. relative depth at SNP positions
10 %

Select minimum distance between SNPs (prune)
10 bp

Select min. SNP quality
30

Select min. read mapping quality
25

Select min. Z-score
1.96

☐ Ignore heterozygous SNPs

Comment (to yourself)
This comment will appear unaltered on your output page. It has no effect on the analysis.

☒ Use altered FastTree (more accurate)
Note: Read more [here](#)

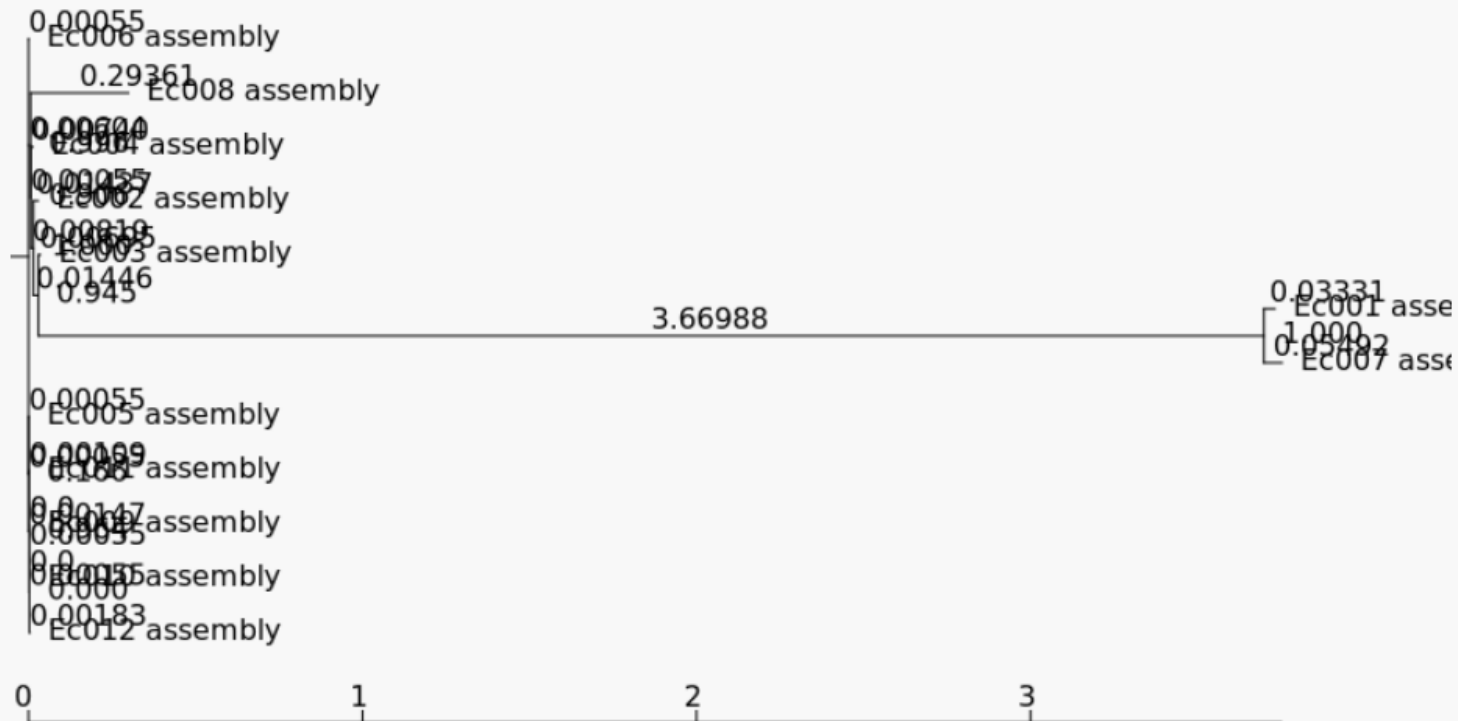
Upload read files and/or assembled genomes (fasta or fastq format)
Note: Read files must be compressed with gzip (compressed files often ends with .gz).
If you get an "Access forbidden. Error 403": Make sure the start of the web adress is https and not just http. Fix it by clicking [here](#).

Name	Size
------	------

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 strongly recommend that the 'Newick' file is downloaded and processed using software created for this purpose. We suggest ([FigTree](#)).



Download 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

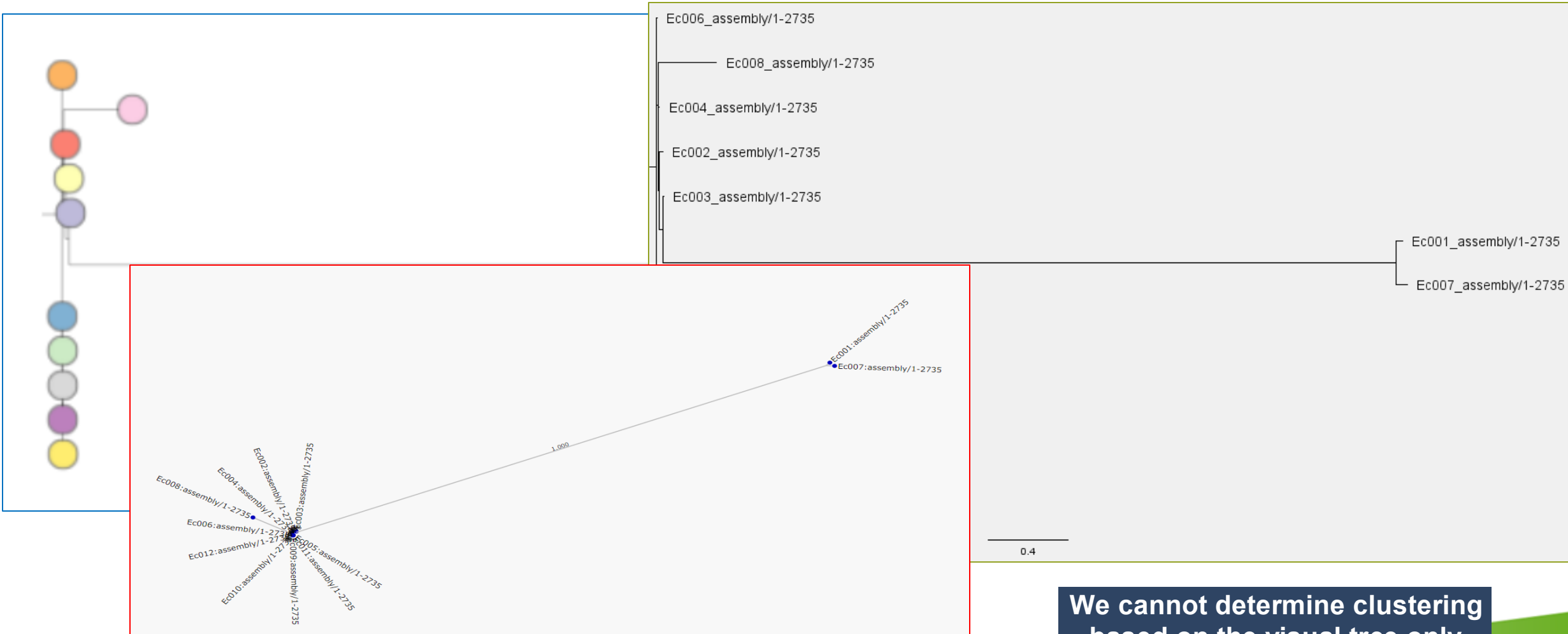
Visualisation of tree (newick file)

Download phylogeny as:

Newick

PDF

SVG



We cannot determine clustering
based on the visual tree only

SNP matrix – pairwise comparison of SNPs

A1 prune 10

	Ec001_asse mbly/1-2735	Ec002_asse mbly/1-2735	Ec003_ass embly/1- 2735	Ec004_ass embly/1- 2735	Ec005_asse bly/1-2735	Ec006_asse bly/1-2735	Ec007_asse bly/1-2735	Ec008_asse bly/1-2735	Ec009_asse bly/1-2735	Ec010_asse bly/1-2735	Ec011_asse bly/1-2735	Ec012_asse bly/1-2735
Ec001_asse mbly/1-2735	0	2176	2122	2180	2176	2176	216	2280	2179	2179	2182	2184
Ec002_asse mbly/1-2735	2176	0	94	80	78	78	2212	644	81	81	84	86
Ec003_asse mbly/1-2735	2122	94	0	98	96	96	2170	662	99	99	102	104
Ec004_asse mbly/1-2735	2180	80	98	0	38	38	2222	604	41	41	44	46
Ec005_asse mbly/1-2735	2176	78	96	38	0	2	2218	598	5	5	8	10
Ec006_asse mbly/1-2735	2176	78	96	38	2	0	2218	598	5	5	8	10
Ec007_asse mbly/1-2735	216	2212	2170	2222	2218	2218	0	2322	2221	2221	2224	2226
Ec008_asse mbly/1-2735	2280	644	662	604	598	598	2322	0	601	601	604	606
Ec009_asse mbly/1-2735	2179	81	99	41	5	5	2221	601	0	0	3	5
Ec010_asse mbly/1-2735	2179	81	99	41	5	5	2221	601	0	0	3	5
Ec011_asse mbly/1-2735	2182	84	102	44	8	8	2224	604	3	3	0	8
Ec012_asse mbly/1-2735	2184	86	104	46	10	10	2226	606	5	5	8	0

min: 0 max: 2322

What is going on with that isolate?

Targets of Distance Columns (CPO C. freundii ST18)

Right-click on the allele type columns to jump to the according contig position in the Sample

Target	Begin	End	GenBank gene	GenBank product	GenBank note	GenBank protein_id	200117_A19...	AMA003417	AMA003565	CPO20190159	AMA00338
.322_RS02285	465,622	467,868		phosphoenolpyruvate--protein phosphotransferase PtsP	member of a ...	WP_003033984.1	? (failed)	1	1	1	1
.322_RS03425	716,674	721,020		autotransporter domain-containing protein	Derived by a...	WP_071684359.1	? (failed)	1	1	1	? (not found)
.322_RS03765	802,371	803,735		PTS sugar transporter subunit IIC	Derived by a...	WP_054528657.1	? (failed)	1	? (not found)	1	1
.322_RS04195	912,678	914,693		tRNA(Met) cytidine acetyltransferase TmcA	cetylates the...	WP_054528641.1	? (failed)	1	1	1	1
.322_RS05765	1,252,228	1,254,714		fimbrial assembly protein	Derived by a...	WP_054528576.1	? (failed)	1	1	1	1
.322_RS06635	1,433,991	1,435,370		cobyrinic acid a,c-diamide synthase	Derived by a...	WP_044701540.1	? (failed)	1	1	1	1
.322_RS06975	1,490,987	1,491,643		DNA-binding response regulator	Derived by a...	WP_003030486.1	? (failed)	1	1	1	1
.322_RS09275	1,966,895	1,967,473		TetR family transcriptional regulator	Derived by a...	WP_046670695.1	? (failed)	1	1	1	1
.322_RS09820	2,083,766	2,084,500		DNA-binding response regulator	Derived by a...	WP_003836390.1	? (failed)	1	1	1	1
.322_RS11920	2,514,178	2,514,801		DSBA oxidoreductase	Derived by a...	WP_003035975.1	? (failed)	1	1	1	1
.322_RS12760	2,702,258	2,703,679		2-oxoglutarate/malate translocator	Derived by a...	WP_003837022.1	? (failed)	1	1	1	1
.322_RS13805	2,920,181	2,921,314		LPS O-antigen length regulator	Derived by a...	WP_054528176.1	? (failed)	1	1	1	1
.322_RS14935	3,176,896	3,177,909		4-hydroxy-2-oxovalerate aldolase	Derived by a...	WP_003021379.1	? (failed)	1	1	1	1
.322_RS15475	3,301,194	3,301,901		flagellar basal body L-ring protein	Derived by a...	WP_042270212.1	? (failed)	1	1	1	1
.322_RS15765	3,357,582	3,359,564		type IV secretion protein Rhs	Derived by a...	WP_072143931.1	? (failed)	1	1	1	1
.322_RS19670	4,225,537	4,226,988		potassium transporter	Derived by a...	WP_003017848.1	? (failed)	1	1	1	1
.322_RS21515	4,624,296	4,625,513		MFS transporter	Derived by a...	WP_054528867.1	? (failed)	1	1	1	1
.322_RS07405	1,575,702	1,576,793		enterohemolysin	Derived by a...	WP_054528497.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found)
.322_RS07605	1,603,957	1,604,262		hypothetical protein	Derived by a...	WP_057101149.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found)
.322_RS08700	1,830,199	1,830,762		hypothetical protein	Derived by a...	WP_003843940.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found)
.322_RS17560	3,773,522	3,773,764		transcriptional regulator	Qin prophag...	WP_003839576.1	? (not found)	1	? (not found)	? (not found)	1
.322_RS22180	4,766,146	4,767,330		elongation factor Tu	Derived by a...	WP_003031109.1	? (not found)	? (not found)	? (not found)	? (not found)	? (not found)
.322_RS06380	1,382,478	1,383,593		amino acid oxidase	Derived by a...	WP_054528547.1	? (not found)	1	? (not found)	? (not found)	? (not found)
.322_RS17175	3,686,035	3,686,850		AraC family transcriptional regulator	Derived by a...	WP_054528023.1	? (not found)	1	1	1	1
.322_RS17930	3,844,410	3,846,275		DNA mismatch repair protein MutL	Derived by a...	WP_054527983.1	? (not found)	1	1	1	1
.322_RS20890	4,482,716	4,483,960		O-antigen polymerase	Derived by a...	WP_046671022.1	? (not found)	1	1	? (not found)	? (not found)
.322_RS22035	4,737,170	4,739,713		nitrite reductase large subunit	Derived by a...	WP_003023592.1	? (not found)	1	1	1	1
.322_RS23080	211,671	211,868		hypothetical protein					1	1	1
.322_RS23130	545,890	546,069		hypothetical protein						? (not found)	1
.322_RS02845	588,933	589,343		formate hydrogenlyase maturation protein Hyd						1	2
.322_RS03030	623,789	624,124		L-valine transporter subunit YgaH					(not found)	? (not found)	1
.322_RS03050	627,414	628,478		proline/betaine ABC transporter permease ProW	Derived by a...	WP_003846040.1	1	1	? (not found)	? (not found)	1
.322_RS03070	633,115	633,525	nrdI	ribonucleotide reductase assembly protein NrdI	in Salmonella...	WP_003037273.1	1	1	? (not found)	? (not found)	1
.322_RS03235	667,222	668,508		capsular polysaccharide biosynthesis protein	Derived by a...	WP_003839728.1	1	1	1	1	1

Hypermutator!!



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