Joint Training Course of the inter EURLS Working Group on NGS:
Introduction to Bioinformatics for genomic data mining

Introduction to genome comparison: gene-by-gene VS SNPs (Guidance document for cluster analysis of WGS data)

Ásgeir Ástvaldsson

EURL-*Campylobacter*

June 15th 2022

Funded by
the European Union

Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)

Foreword

The WG has been established by the European Commission with the aim to promote the use of NGS across the EURLs' networks, build NGS capacity within the EU and ensure liaison with the work of the EURLs and the work of EFSA and ECDC on the NGS mandate sent by the Commission. The WG includes all the EURLs operating in the field of the microbiological contamination of food and feed and this document represents a deliverable of the WG and is meant to be diffused to all the respective networks of NRLs.

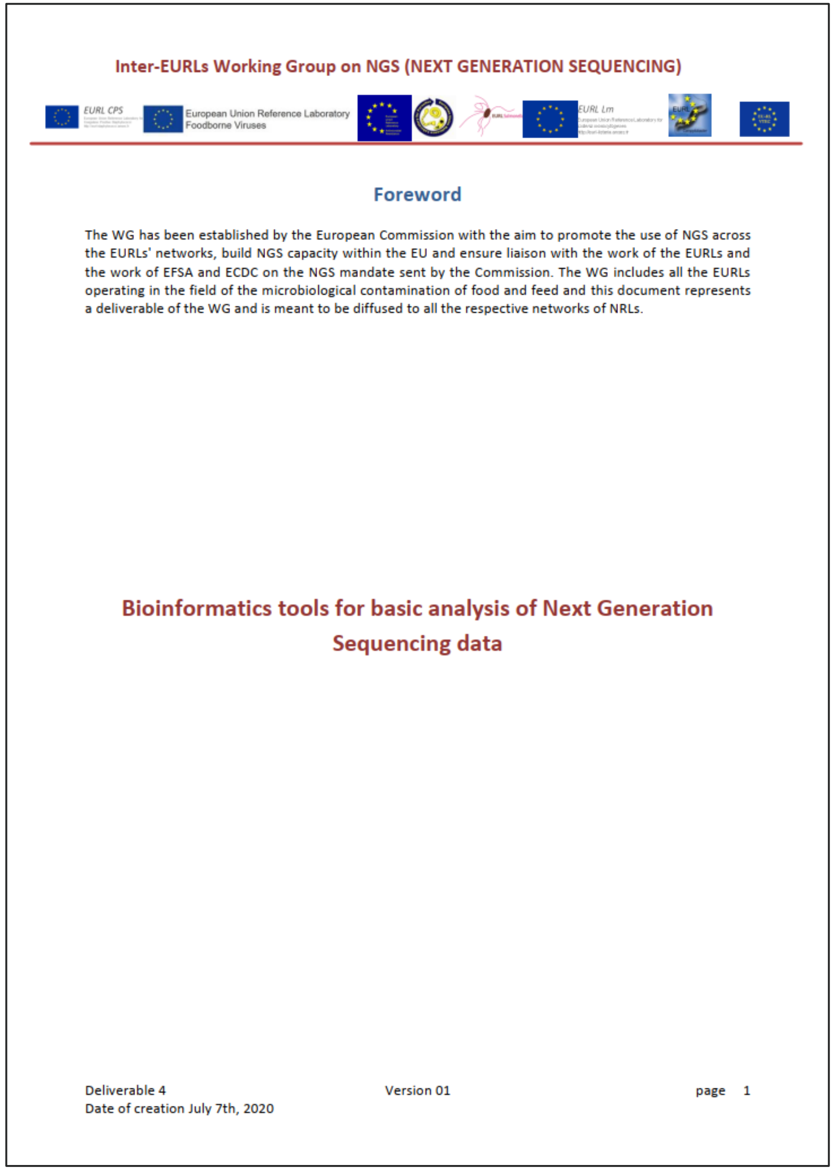
Guidance document for cluster analysis of whole genome sequence data

Version 02

Funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or DG-SANTE. Neither the European Union nor DG-SANTE can be held responsible for them.

The guidelines aim to inform and support NRLs in the choices of methods to be used for the so-called cluster analysis, in which comparisons of genomes are performed followed by visualisations of the results to allow an interpretation of how closely the genomes are related to each other.



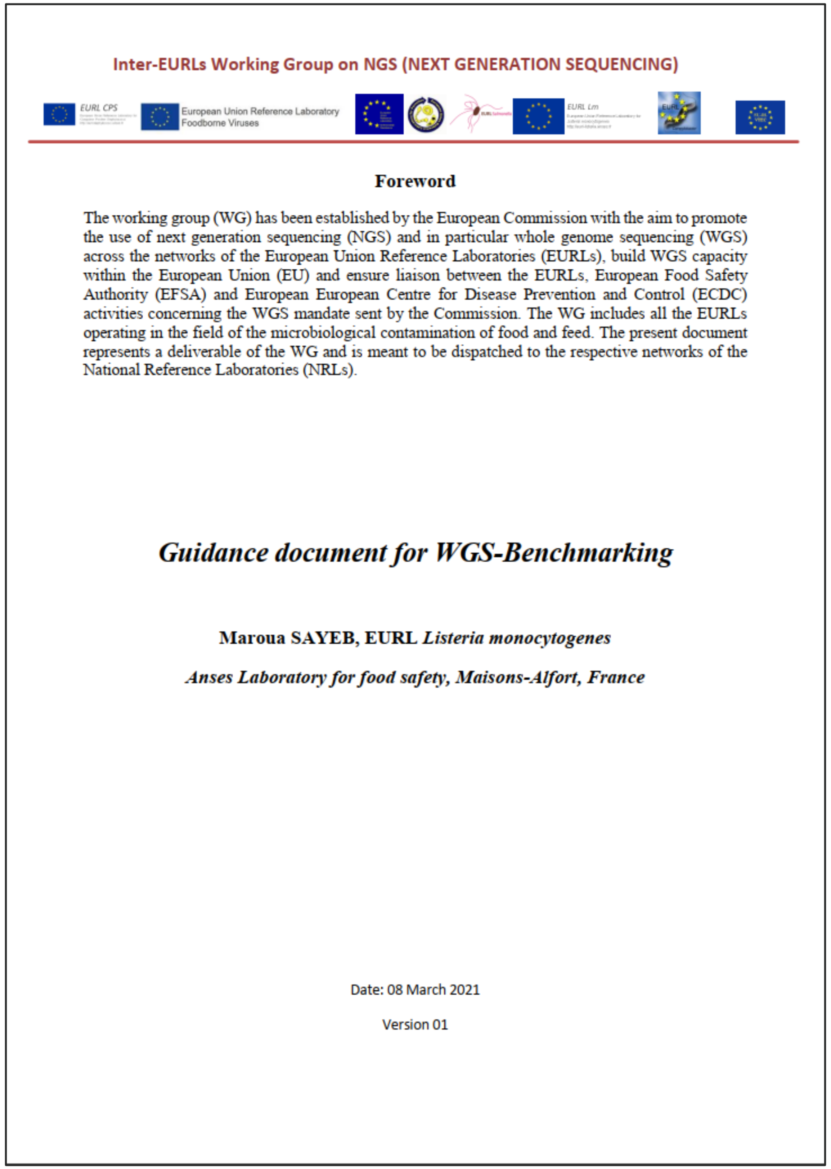
Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)

Foreword

The WG has been established by the European Commission with the aim to promote the use of NGS across the EURLs' networks, build NGS capacity within the EU and ensure liaison with the work of the EURLs and the work of EFSA and ECDC on the NGS mandate sent by the Commission. The WG includes all the EURLs operating in the field of the microbiological contamination of food and feed and this document represents a deliverable of the WG and is meant to be diffused to all the respective networks of NRLs.

Bioinformatics tools for basic analysis of Next Generation Sequencing data

Deliverable 4 Version 01 page 1
Date of creation July 7th, 2020



Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)

Foreword

The working group (WG) has been established by the European Commission with the aim to promote the use of next generation sequencing (NGS) and in particular whole genome sequencing (WGS) across the networks of the European Union Reference Laboratories (EURLs), build WGS capacity within the European Union (EU) and ensure liaison between the EURLs, European Food Safety Authority (EFSA) and European European Centre for Disease Prevention and Control (ECDC) activities concerning the WGS mandate sent by the Commission. The WG includes all the EURLs operating in the field of the microbiological contamination of food and feed. The present document represents a deliverable of the WG and is meant to be dispatched to the respective networks of the National Reference Laboratories (NRLs).

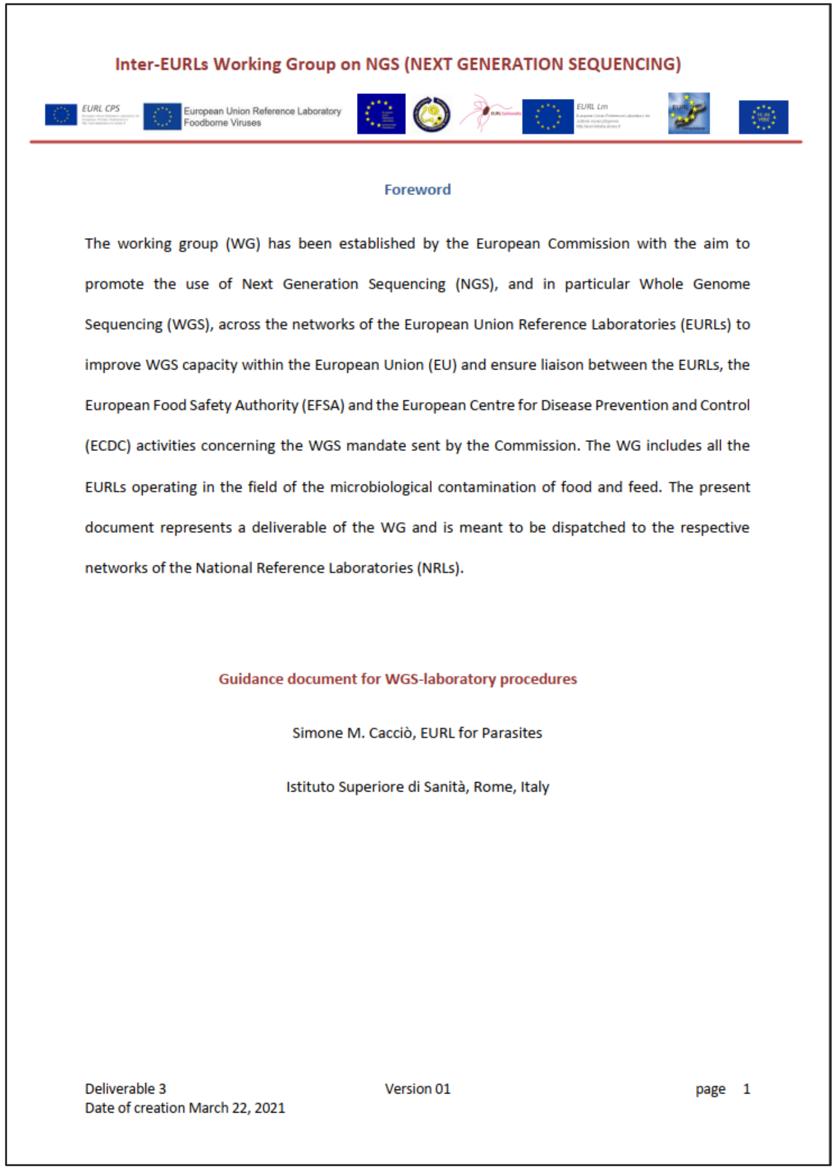
Guidance document for WGS-Benchmarking

Maroua SAYEB, EURL *Listeria monocytogenes*

Anses Laboratory for food safety, Maisons-Alfort, France

Date: 08 March 2021

Version 01



Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)

Foreword

The working group (WG) has been established by the European Commission with the aim to promote the use of Next Generation Sequencing (NGS), and in particular Whole Genome Sequencing (WGS), across the networks of the European Union Reference Laboratories (EURLs) to improve WGS capacity within the European Union (EU) and ensure liaison between the EURLs, the European Food Safety Authority (EFSA) and the European Centre for Disease Prevention and Control (ECDC) activities concerning the WGS mandate sent by the Commission. The WG includes all the EURLs operating in the field of the microbiological contamination of food and feed. The present document represents a deliverable of the WG and is meant to be dispatched to the respective networks of the National Reference Laboratories (NRLs).

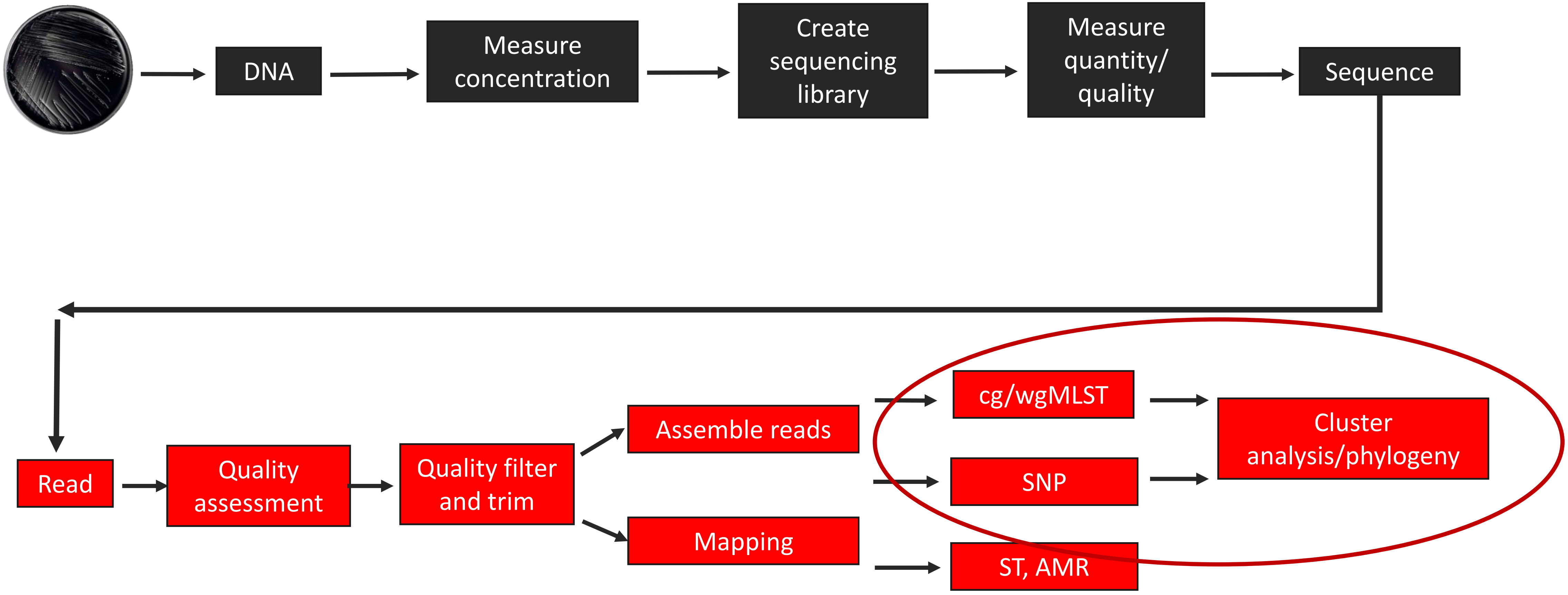
Guidance document for WGS-laboratory procedures

Simone M. Cacciò, EURL for Parasites

Istituto Superiore di Sanità, Rome, Italy

Deliverable 3 Version 01 page 1
Date of creation March 22, 2021

A typical WGS workflow



DNA

Measure concentration

Create sequencing library

Measure quantity/ quality

Sequence

Read

Quality assessment

Quality filter and trim

Assemble reads

Mapping

cg/wgMLST

SNP

ST, AMR

Cluster analysis/phylogeny

Why perform WGS cluster analysis?

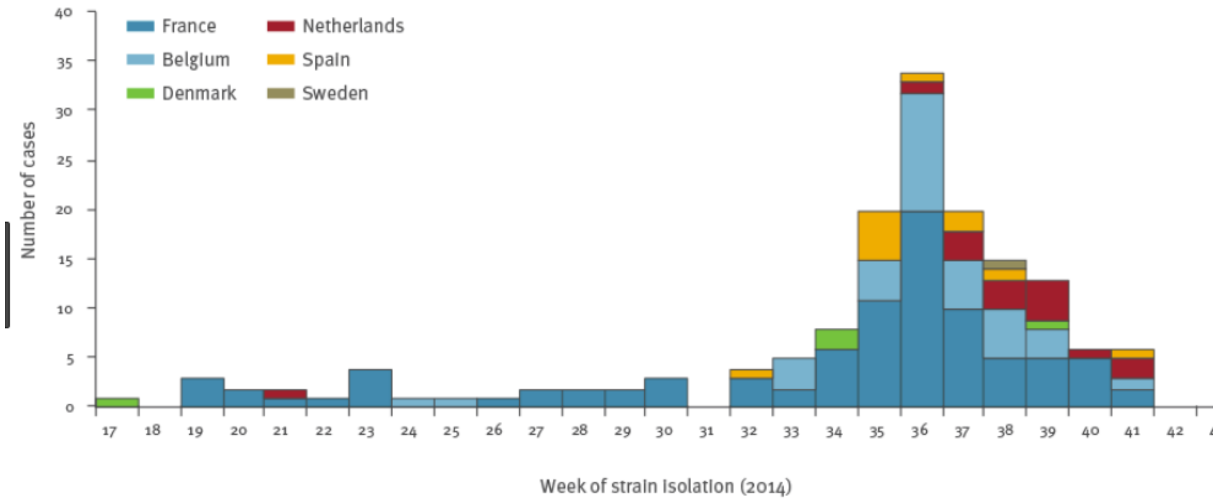
Outbreak investigations

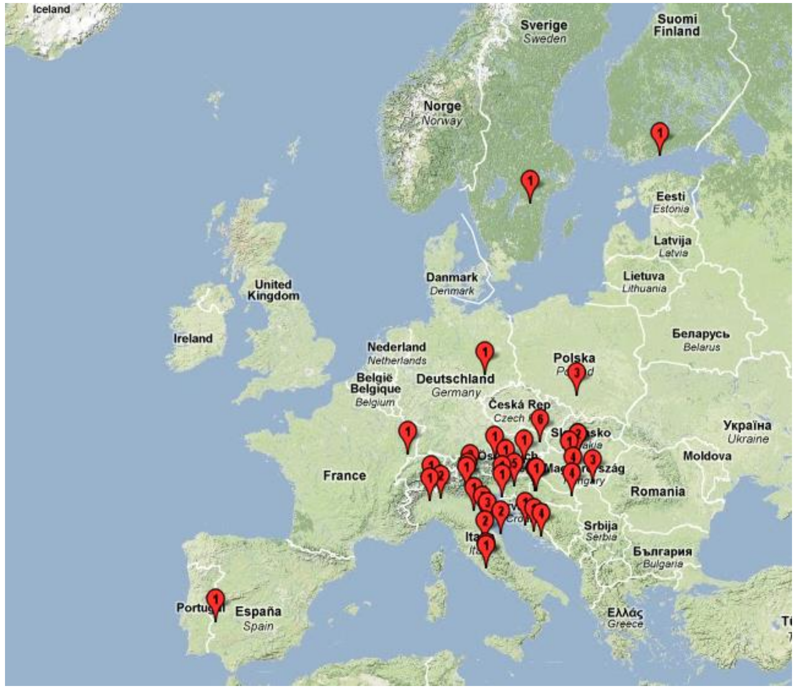
– determine the source of an outbreak,

– determine routes of infection/spread -> interventions

Surveillance

– detecting outbreaks, detect multi country clusters





How to perform WGS cluster analysis

Most common approaches

Single nucleotide polymorphism (SNP) approach

Individual mutations used as separate phylogenetic markers

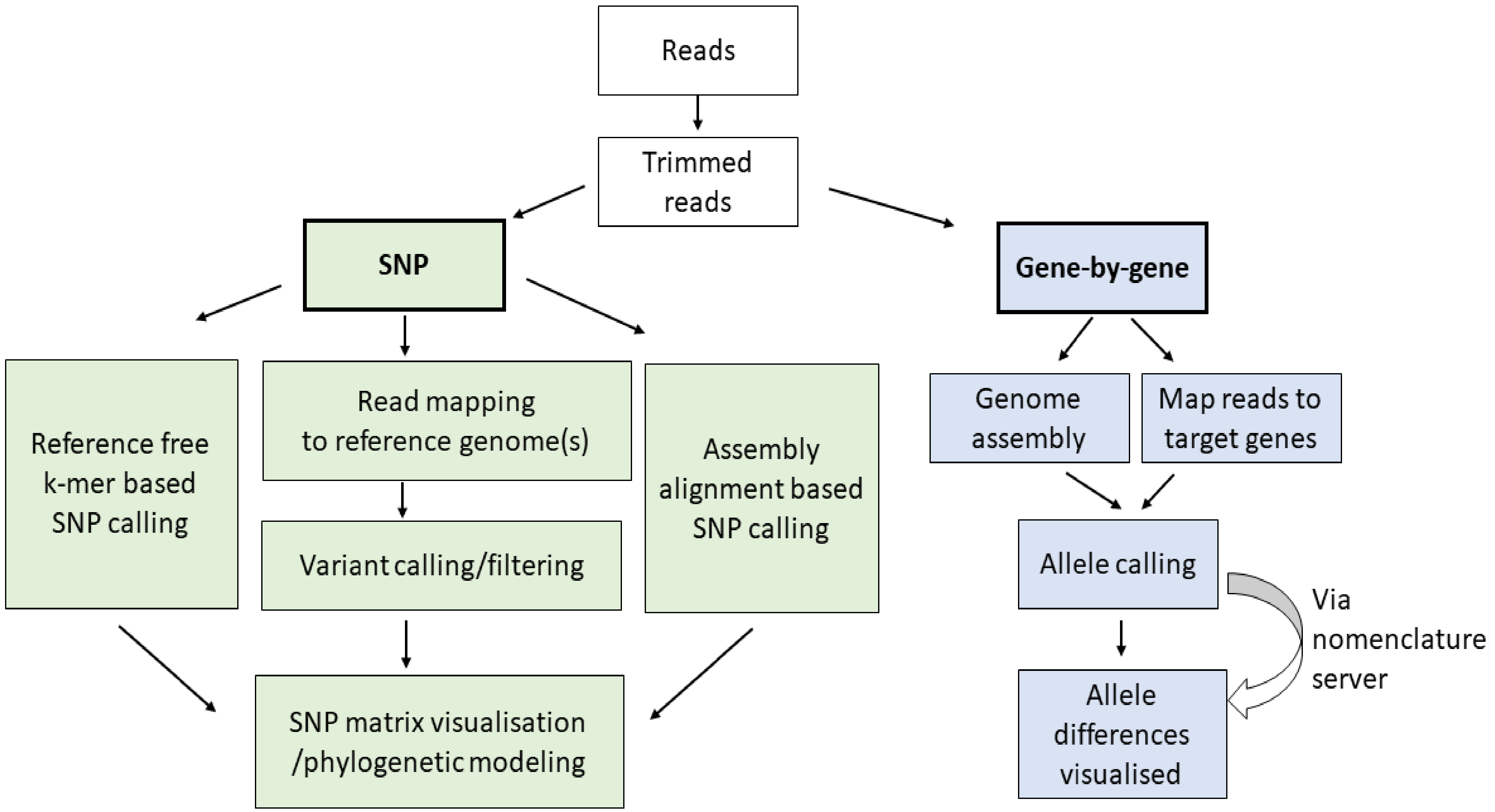
Gene-by-gene approach

Each variant of a gene or part of a gene is considered an allele

- Both approaches involve several steps of analysis, that all can affect the end results
  - e.g., read trimming, assembly, read-mapping, alignment, variant calling, allele calling and dendrogram/tree production
- Freely available and commercial software can perform all these steps
- Important for users to have a solid knowledge of the software and methodology in order to produce correct and comparable results
- Different steps of the analysis should be evaluated for each pathogen, sequencing machine and software

- Validation of all steps of the end-to-end WGS workflow has been described in the document ‘**Guidance document for WGS benchmarking**’ also produced by the Inter-EURLs WG on NGS

Fundamental steps in cluster analysis



Reads

Trimmed reads

SNP

Reference free k-mer based SNP calling

Read mapping to reference genome(s)

Variant calling/filtering

Assembly alignment based SNP calling

SNP matrix visualisation /phylogenetic modeling

Gene-by-gene

Genome assembly

Map reads to target genes

Allele calling

Allele differences visualised

Via nomenclature server

The SNP approach

The approach with highest resolution for relatedness studies

Read-mapping

Most common SNP approach

Steps:

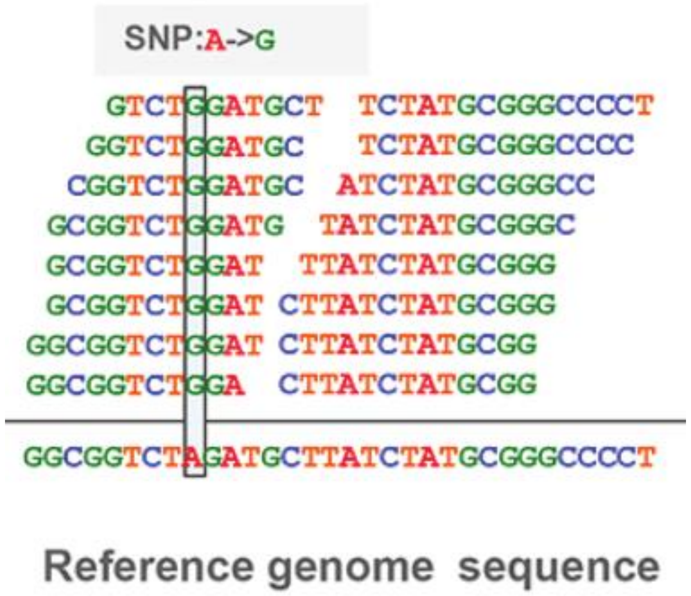
- Mapping reads to a reference
- Variant calling
- Variant filtering
- SNP matrix visualisation

Drawbacks

- Difficult to standardize
- Can be computationally intensive

Other approaches

- Reference free k-mer based SNP calling
- Assembly alignment based SNP calling

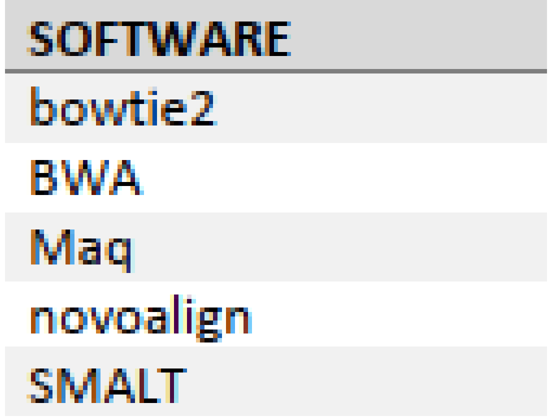

SNP:A->G
GTCTGGATGCT TCTATGCGGGCCCCT
GGTCTGGATGC TCTATGCGGGCCCC
CGGTCTGGATGC ATCTATGCGGGCC
GCGGTCTGGATG TATCTATGCGGGC
GCGGTCTGGAT TTATCTATGCGGG
GCGGTCTGGAT CTTATCTATGCGGG
GGCGGTCTGGAT CTTATCTATGCGG
GGCGGTCTGGA CTTATCTATGCGG
GGCGGTCTAGATGCTTATCTATGCGGGCCCCT
Reference genome sequence

The SNP approach

Mapping reads to a reference

- Reads mapped to a reference genome using a read mapping software
- Normaly only one reference genome is used, but some methods use several
- Choose reference genome representative of the pathogen to maximise resolution

List of common read mappers:



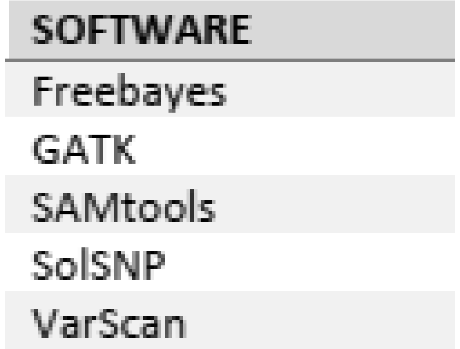
SOFTWARE
bowtie2
BWA
Maq
novoalign
SMALT

The SNP approach

Variant calling

- The process of identifying in which position bases differ from the reference sequence
- Done by using the read mapping results and a variant calling software

List of variant calling software:



SOFTWARE
Freebayes
GATK
SAMtools
SolSNP
VarScan

The SNP approach

Variant filtering

Incorrect SNPs/variants may be called for a number of reasons, including quality issues and repetitive sequence regions. The variant calling procedure often includes, or is combined with, a number of filtering steps to reduce errors and make the analysis more robust. These filtering steps may include:

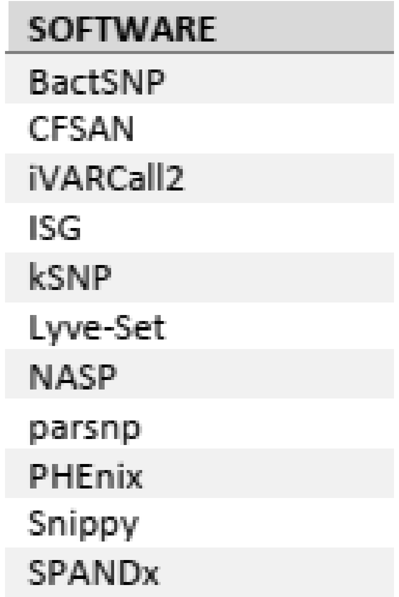
- Genomic regions with low coverage.
- Genomic regions with coverage much larger than the average coverage (possibly repetitive).
- Threshold for how large fraction of reads that must support the allele.
- Minimum quality values for the base calling of the reads at the SNP position.
- Minimum quality value of the read mapping (is the read uniquely mapped).
- Mapping positions close to the reference sequence contig ends may be excluded.
- Regions where many SNPs are found in close proximity to each other may be excluded (possible recombination).
- Duplicate reads in the alignment may be removed (may be PCR duplicates, not true unique sequenced fragments).

The SNP approach

SNP „pipelines“

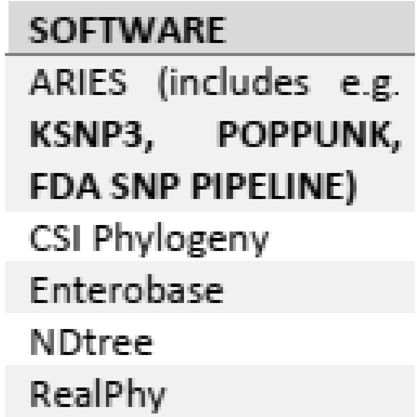
- Several „pipelines“ publicly available for SNP analyses
- Combine the required steps for SNP analysis
- Some pipelines also available as online services

Common SNP pipelines

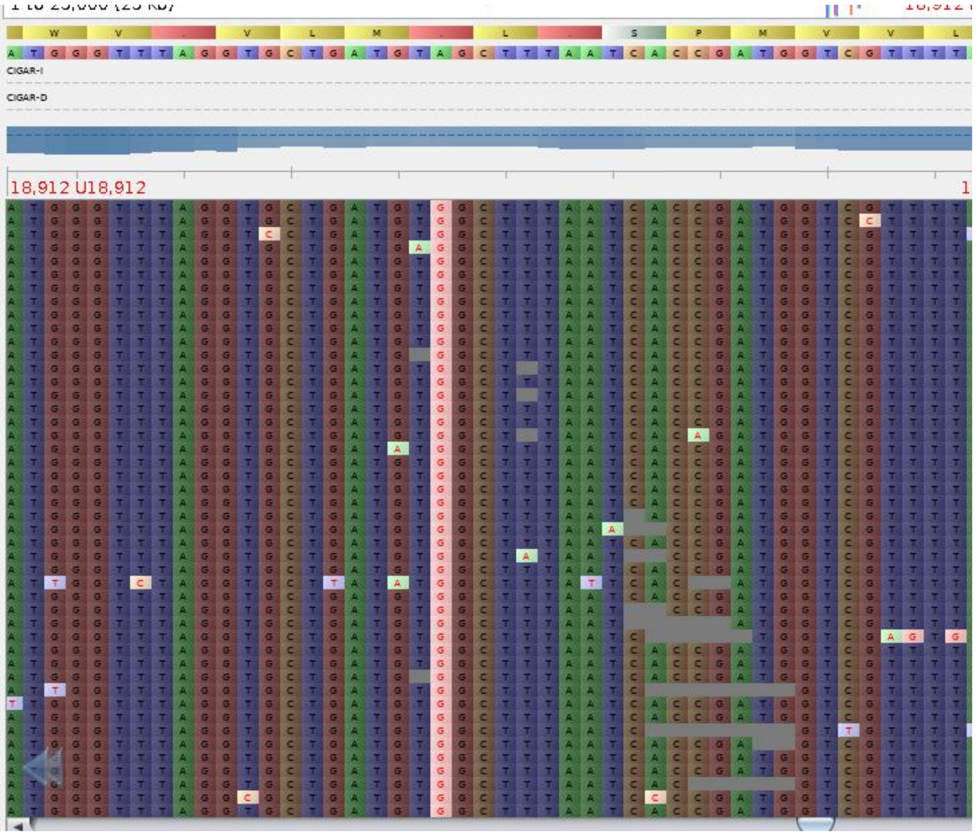


SOFTWARE
BactSNP
CFSAN
iVARCall2
ISG
kSNP
Lyve-Set
NASP
parsnp
PHEnix
Snippy
SPANDx

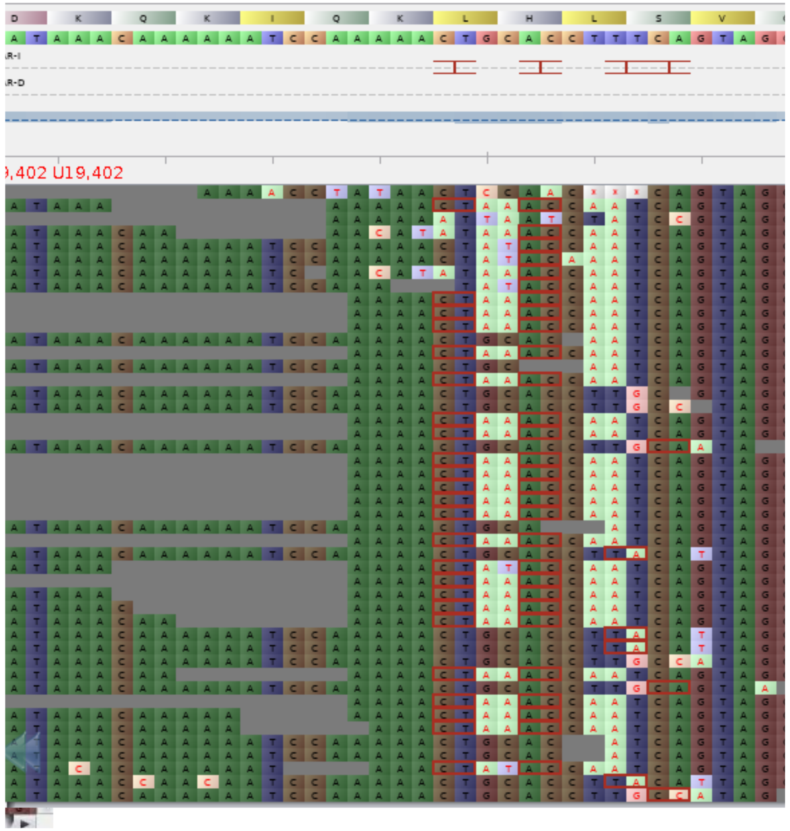
Online SNP pipelines



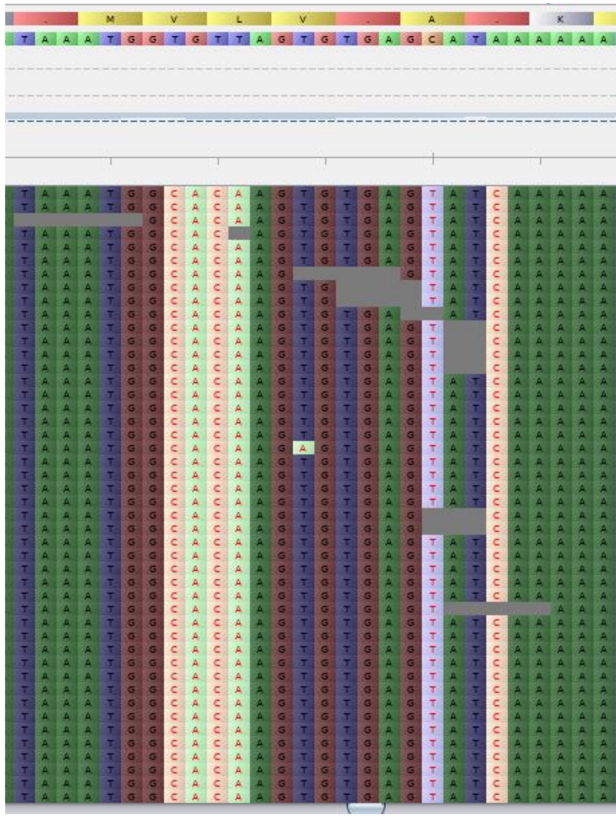
SOFTWARE
ARIES (includes e.g. **KSNP3, POPPUNK, FDA SNP PIPELINE)**
CSI Phylogeny
Enterobase
NDtree
RealPhy



An unambiguous SNP



A problematic region



Many SNPs or one larger mutation?

The gene-by-gene approach

Extended multilocus sequence typing (MLST) analysis, upscaled to include thousands of genes or alleles

No reference genome, instead this approach uses a pre-defined list or a database of target genes (called a scheme)

All sequenced genomes compared to the same list

Two main types of schemes

core genome MLST

(cgMLST)

Conserved core of target genes found in nearly all strains used to create the allele database

- Produces comparable results for almost any genome of the species
- Stable nomenclature
- Suitable for surveillance purposes

whole genome MLST

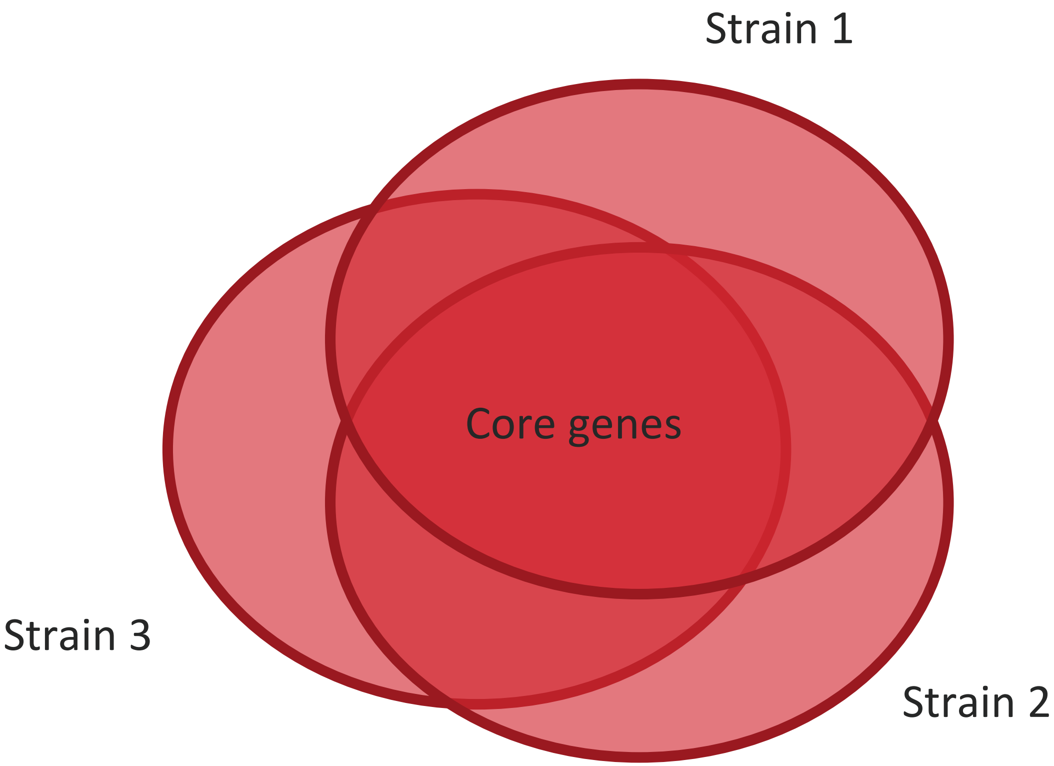
(wgMLST)

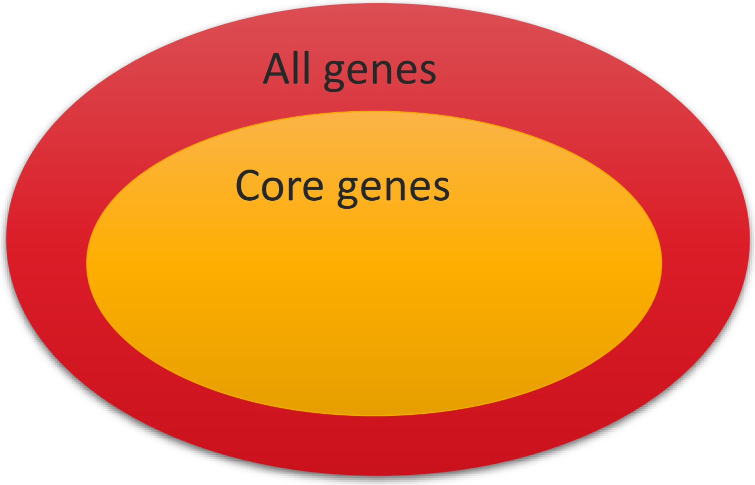
All genes found in the strains used to create the allele database (core genome + accessory genome)

- Not all genes presented in all sequenced genomes
- Higher number of alleles > higher resolution
- Resolution similar to a SNP analysis
- Useful for outbreak tracking

The gene-by-gene approach

Core genome vs accessory genome





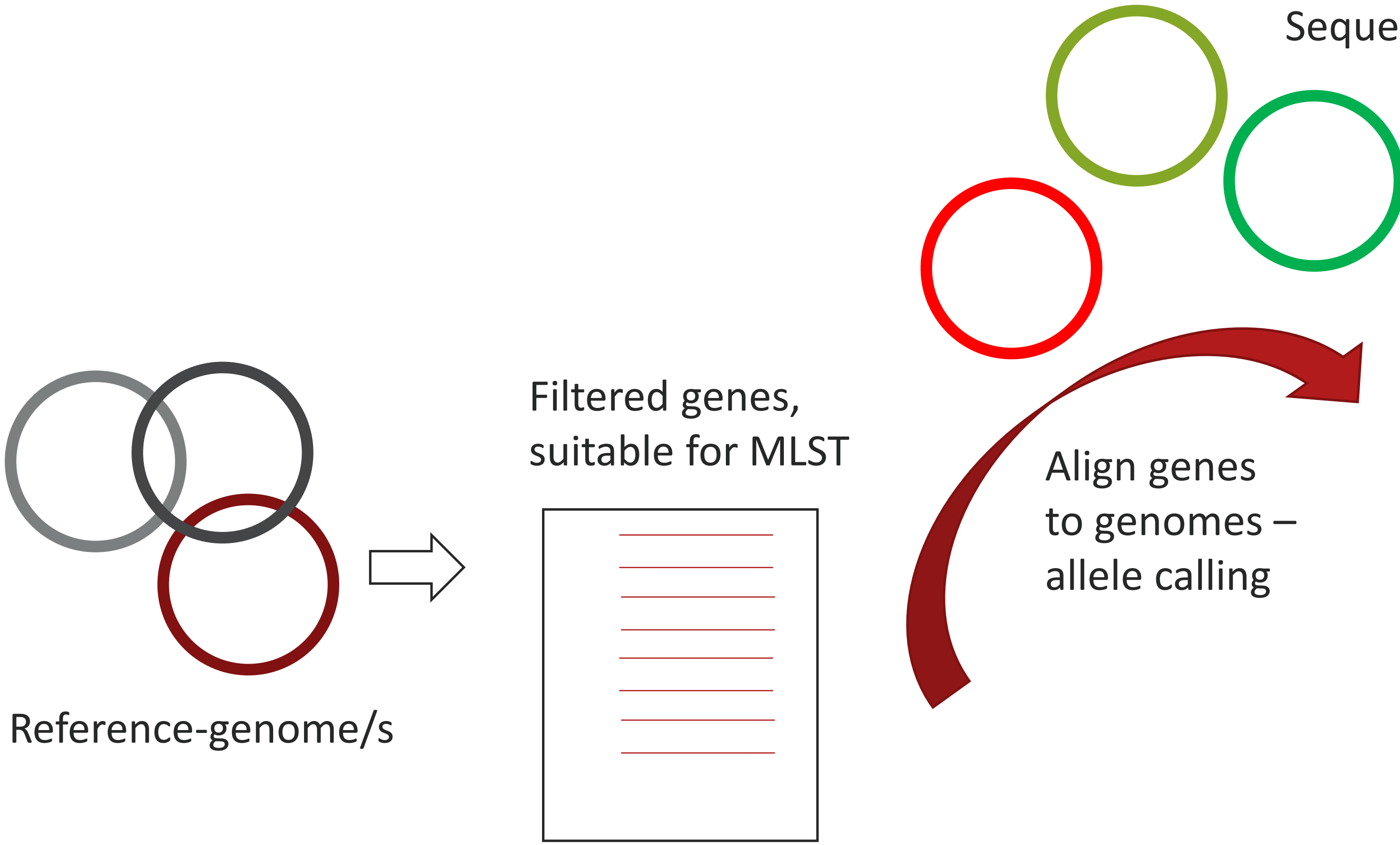
Core genes: Present in all isolates compared
Example of core genes: genes necessary for survival
(housekeeping genes)

Accessory genes: NOT present in all isolates compared
Example of accessory genes: genes for strain specific
adaptation (e.g. AMR, plasmids, metabolic …)

The gene-by-gene approach

- Assembled genomes most commonly used as input
  - Read mapping to the target genes can be used instead
- Analysis performed by aligning the gene targets (from the scheme) to the assembly and extract the isolate´s allelic sequence
- Allele calling can be time consuming
- New genomes can be added at later stages

The gene-by-gene approach



Sequenced isolates

Reference-genome/s

Filtered genes, suitable for MLST

MLST scheme

Align genes to genomes – allele calling

	Gene 1	Gene 2	Gene 3	Gene 4
Isolate1	1	2	1	1
Isolate2	4	3	-	3
Isolate3	4	1	1	1
Isolate4	5	1	1	1
Isolate5	22	4	24	13
Isolate6	4	3	4	5

Allele identifiers – each number matches a certain DNA-sequence of that gene

The gene-by-gene approach

Validated cg/ wgMLST-schemes available for food-borne pathogens

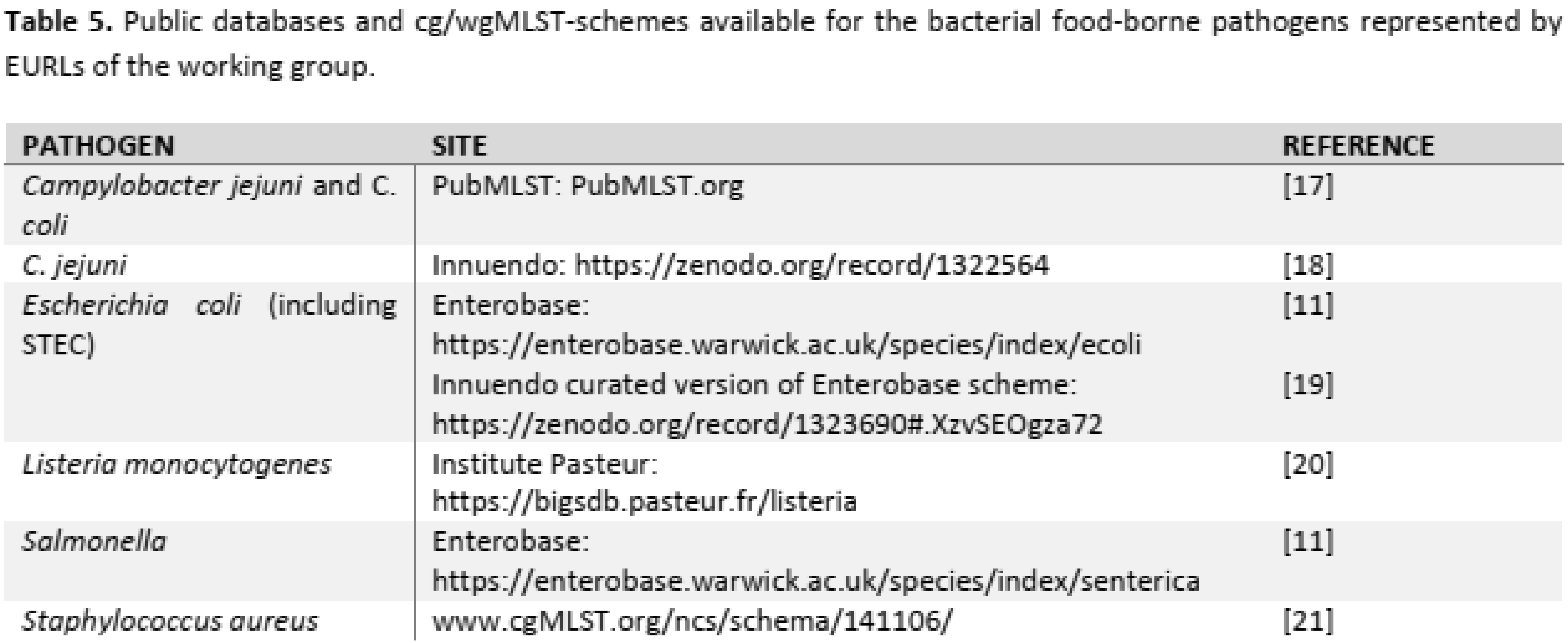


Table 5. Public databases and cg/wgMLST-schemes available for the bacterial food-borne pathogens represented by EURLs of the working group.

PATHOGEN	SITE	REFERENCE
Campylobacter jejuni and *C. coli*	PubMLST: PubMLST.org	[17]
C. jejuni	Innuendo: https://zenodo.org/record/1322564	[18]
Escherichia coli (including STEC)	Enterobase: https://enterobase.warwick.ac.uk/species/index/ecoli Innuendo curated version of Enterobase scheme: https://zenodo.org/record/1323690#.XzvSEOgza72	[11] [19]
Listeria monocytogenes	Institute Pasteur: https://bigsdb.pasteur.fr/listeria	[20]
Salmonella	Enterobase: https://enterobase.warwick.ac.uk/species/index/senterica	[11]
Staphylococcus aureus	www.cgMLST.org/ncs/schema/141106/	[21]

The gene-by-gene approach

Genome assembly

Genome assembly is most commonly used for the gene-by-gene approach

Poor assemblies can have a negative impact on allele calling

Steps for assembly:

Adapter and quality trimming of reads

Trimmomatic, Sickle, Trim Galore, fastp

Assembly of reads

SPAdes, Velvet, SKESA

Assembly correction and polishing

Pilon

Check assembly quality metrics

length, GC%, N50, no of contigs

All tools need to be properly optimized using proper validation datasets for each pathogen in every laboratory

The gene-by-gene approach

Allele calling

- Alignment tools such as BLAST returns the allele sequences of the genome analyses
- Receives allele identifiers if connected to online databases
- If a allele sequence is novel, a new identifier is assigned and is deposited to the database
- Commercial software, open source software and online services available

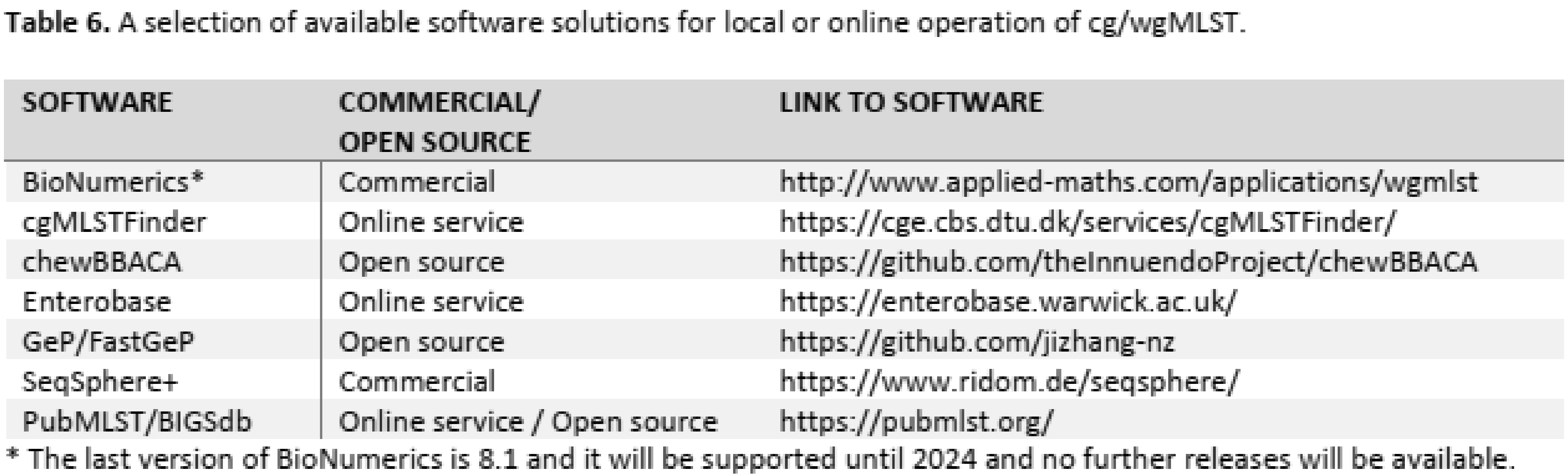


Table 6. A selection of available software solutions for local or online operation of cg/wgMLST.

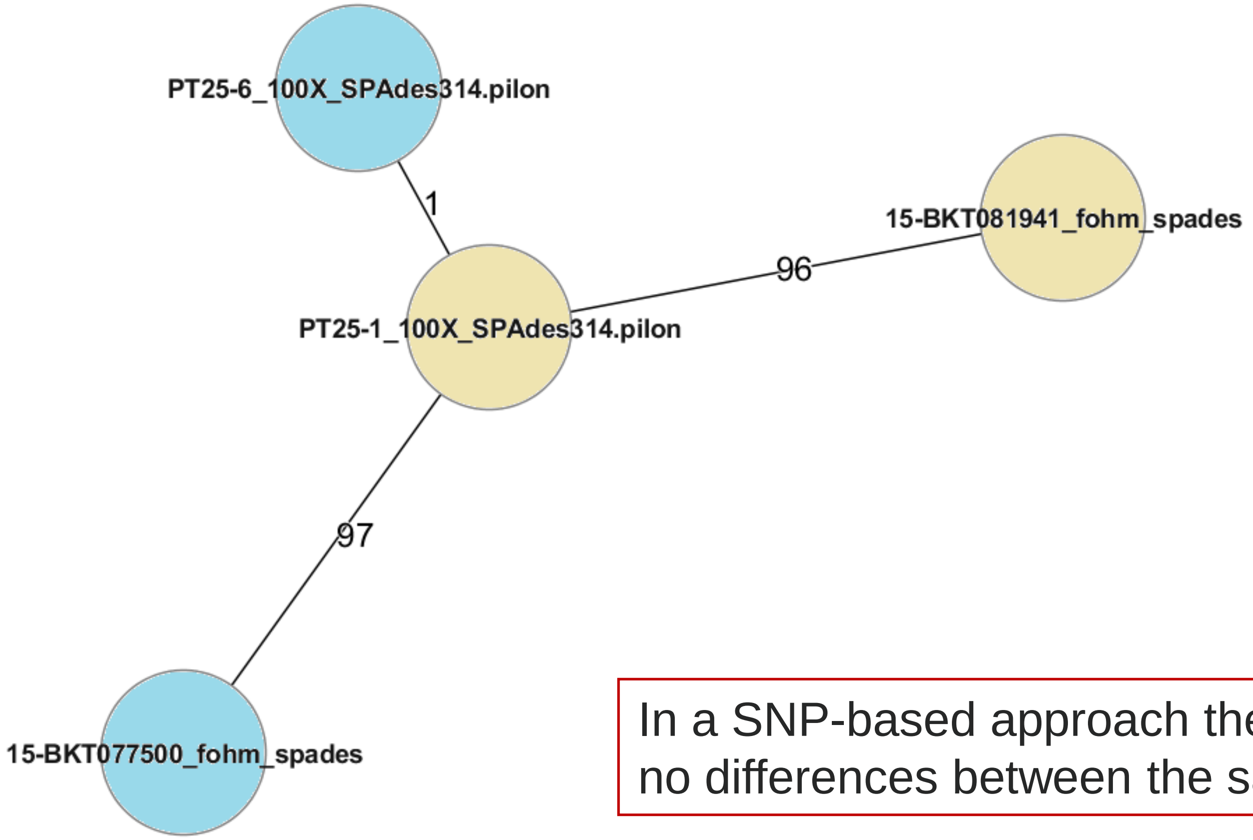
SOFTWARE	COMMERCIAL/ OPEN SOURCE	LINK TO SOFTWARE
BioNumerics*	Commercial	http://www.applied-maths.com/applications/wgmlst
cgMLSTFinder	Online service	https://cge.cbs.dtu.dk/services/cgMLSTFinder/
chewBBACA	Open source	https://github.com/theInnuendoProject/chewBBACA
Enterobase	Online service	https://enterobase.warwick.ac.uk/
GeP/FastGeP	Open source	https://github.com/jizhang-nz
SeqSphere+	Commercial	https://www.ridom.de/seqsphere/
PubMLST/BIGSdb	Online service / Open source	https://pubmlst.org/

* The last version of BioNumerics is 8.1 and it will be supported until 2024 and no further releases will be available.

The gene-by-gene approach

Illumina vs Ion Torrent

Errors produced by Illumina and Ion Torrent differ, therefor a proper validation should be performed when using assemblies derived from different platforms in the same gene-by-gene comparison



- Trimmomatic
- 100X coverage
- SPAdes 3.14

In a SNP-based approach there were no differences between the samples.

SNP vs gene-by-gene approach

- Generally group isolates into same clusters
- Results from the methods are most often comparable
- Validation using reference datasets should be performed for chosen pipeline/software/parameters etc.

Differences between the methods:

- Intergenic regions not included in gene-by-gene approach
- Several mutations and indels in a gene collapsed and only counted as 1 change using gene-by-gene approach
    - E.g. a gene has 3 mutation, counted as 1 change using gene-by-gene approach and 3 changes using SNP approach
- Small INDELs not counted by all SNP approaches but always counted as new allele using a gene-by-gene approach
- SNP restricted to reference genome, needs to be closely related for high resolution
- MLST restricted to genes in scheme
- Both for SNP and gene-by-gene, the input data quality affects the end result (but perhaps more for the assembly based methods)

Software for SNP and gene-by-gene

- **Online services**
  - Dependency on service provider
  - Downtimes of server
  - Long waiting times

  + No cost

  + Easy to perform

- **Local operation**
  - Often requires bioinformatics/Linux knowledge
  - Computer power
  - Comercial software expensive

  + Full control of analysis

  + Not dependant on external provider

A selection of available software solutions for SNP and cg/wgMLST are listed in the guidance document. You can find both commercial and free software, local and online software.

Visualisation of clustering data

Number of SNPs or allele differences can be directly derived from a table and converted into a distance matrix describing the pairwaise distances

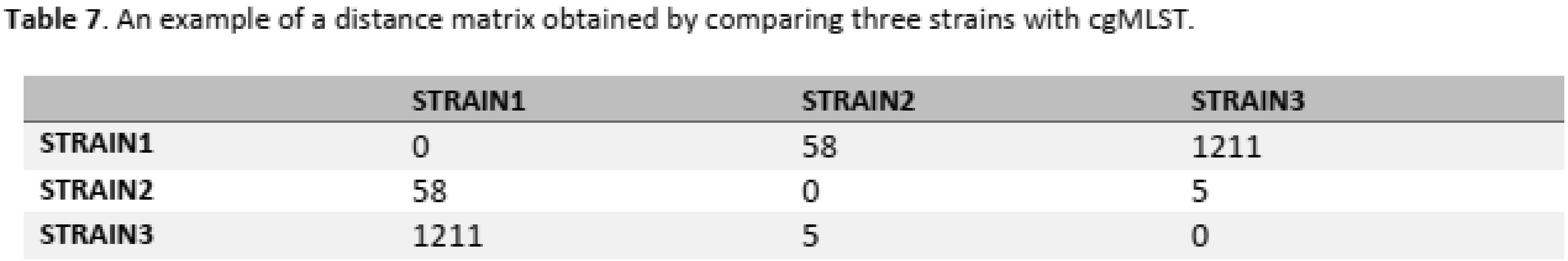


Table 7. An example of a distance matrix obtained by comparing three strains with cgMLST.

	STRAIN1	STRAIN2	STRAIN3
STRAIN1	0	58	1211
STRAIN2	58	0	5
STRAIN3	1211	5	0

The distance matrix lists the number of SNPs or allelic differences detected among each pair of strains analysed

Visualisation of clustering data

A minimum spanning tree (MST) is a common way to visualise SNPs or allelic differences

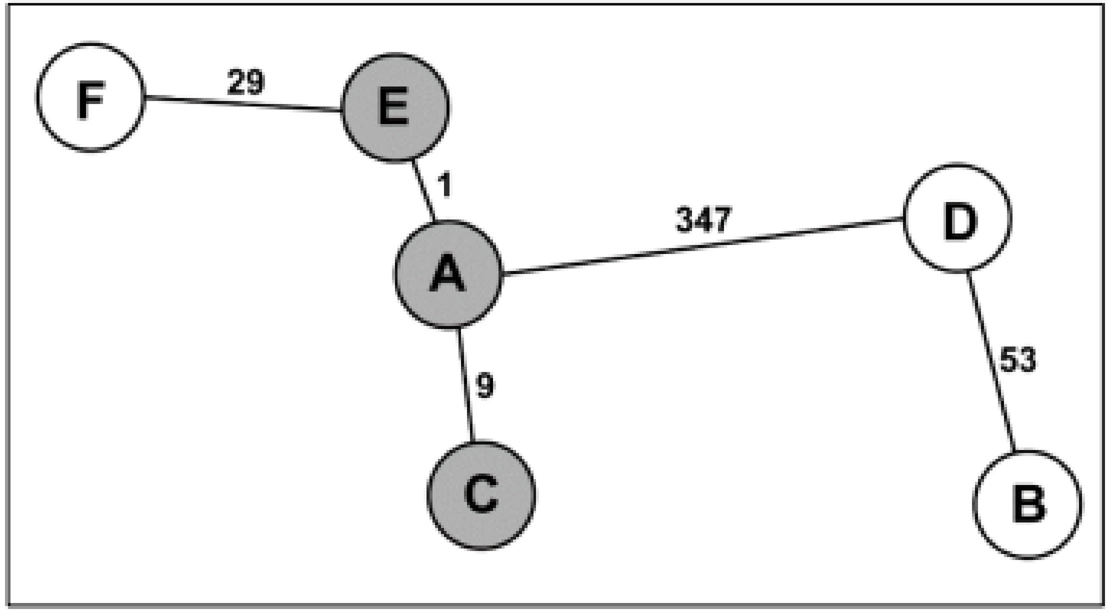


Figure 2. A cgMLST result for six genomes visualised in a minimum spanning tree. The numbers between the sample names represent the number of allelic differences between the samples. The line lengths are not proportional to the number of differences. The total number of gene targets compared in this analysis is 1,340. The identified cluster has been highlighted in grey, with a cluster definition set to ≤ 10 alleles differences.

If many genomes – two step analysis. Elevates resolution of the identified clusters and neighbouring isolates since the shared genome will be larger when only closely related genomes are analysed

Visualisation of clustering data

The results of cluster analysis can also be visualised in a phylogenetic tree, rooted or unrooted.

Can be produced from distance matrix or directly from the SNP alignment data

Software solutions for visualisation of clustering data

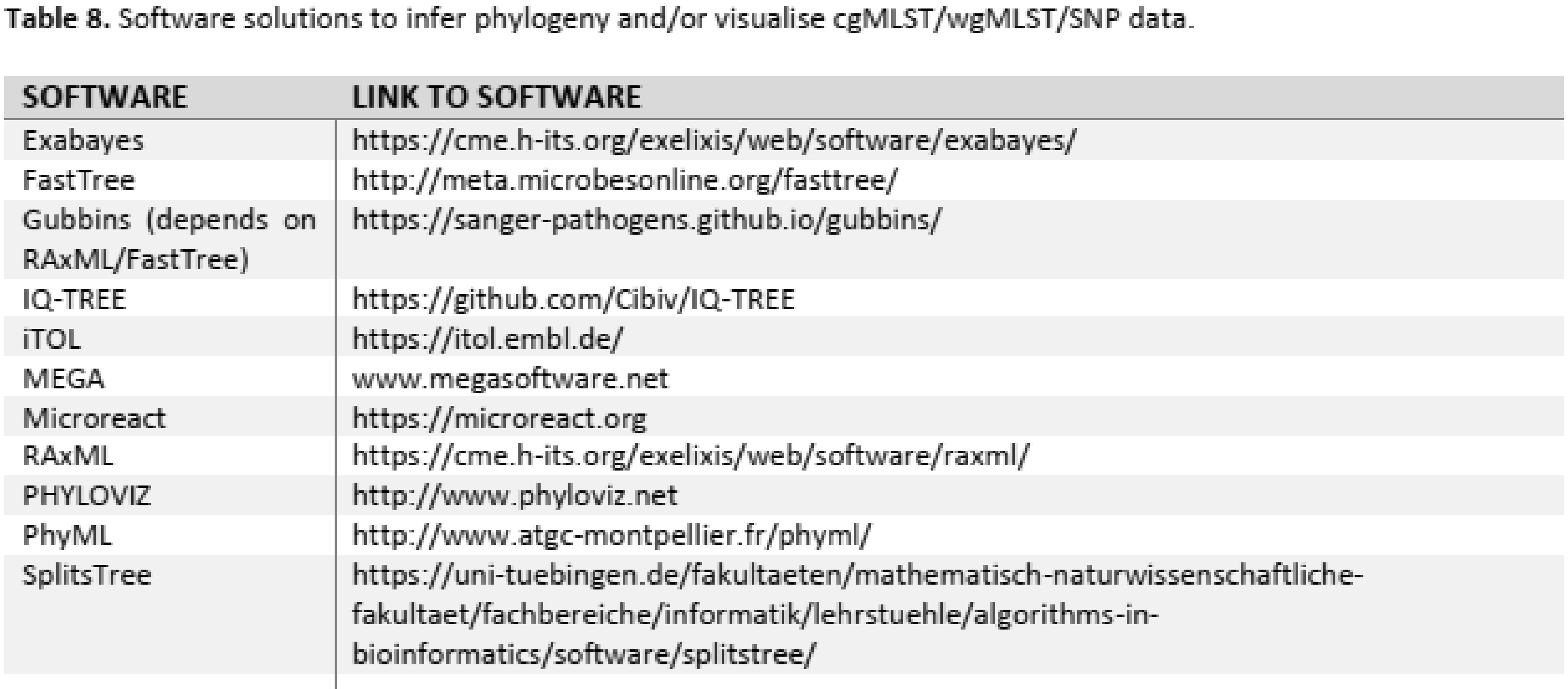


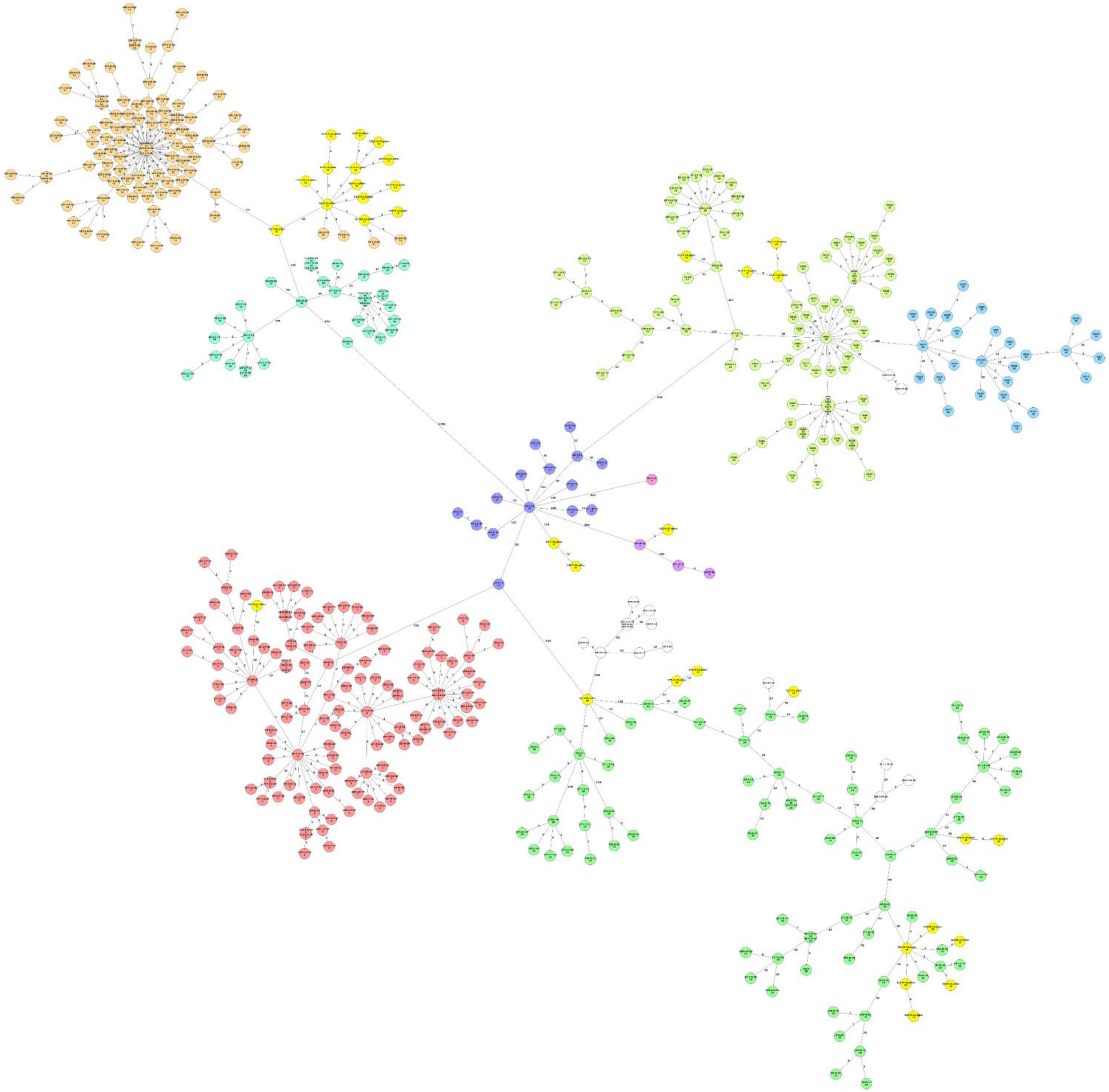
Table 8. Software solutions to infer phylogeny and/or visualise cgMLST/wgMLST/SNP data.

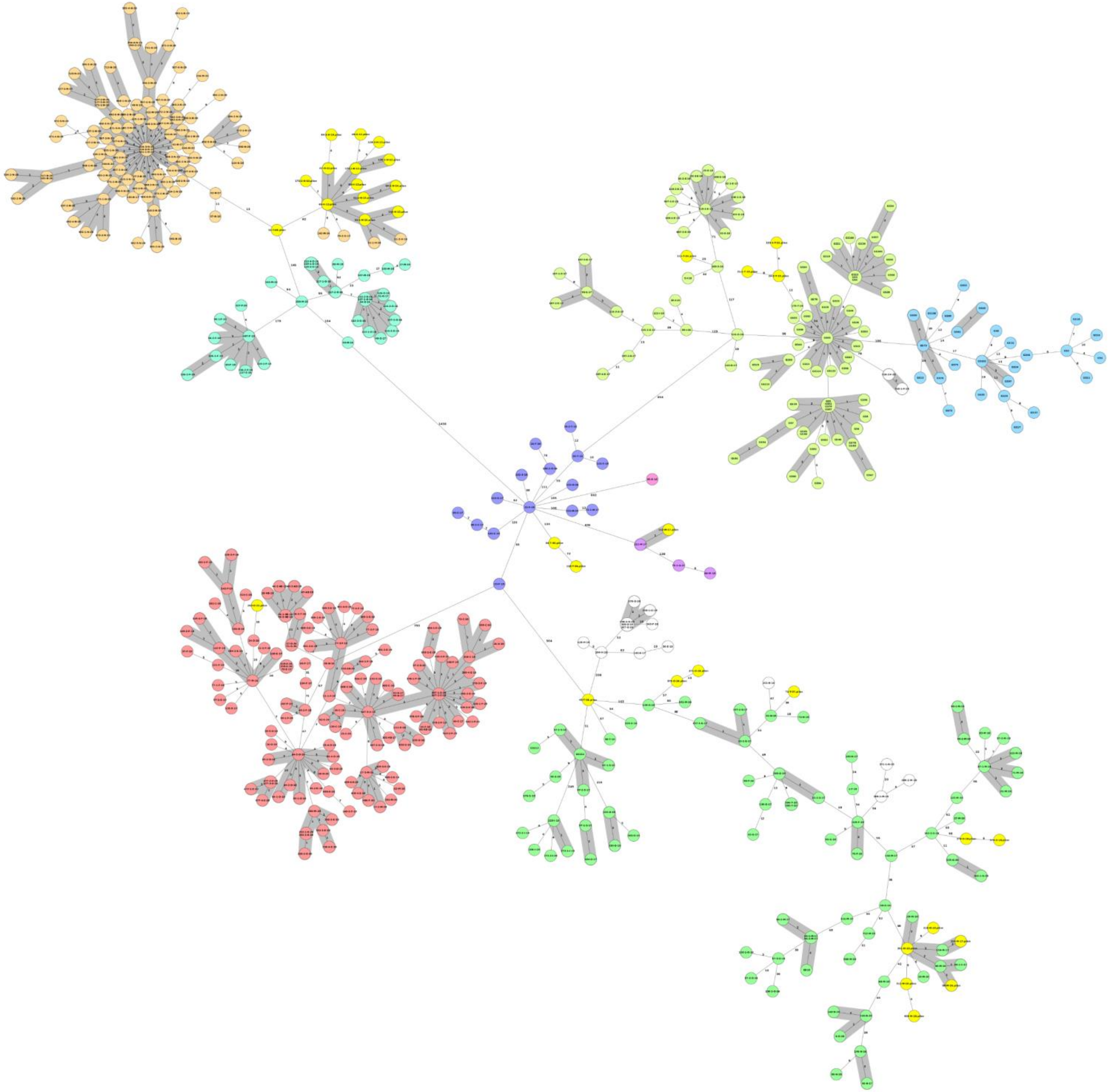
SOFTWARE	LINK TO SOFTWARE
Exabayes	https://cme.h-its.org/exelixis/web/software/exabayes/
FastTree	http://meta.microbesonline.org/fasttree/
Gubbins (depends on RAxML/FastTree)	https://sanger-pathogens.github.io/gubbins/
IQ-TREE	https://github.com/Cibiv/IQ-TREE
iTOL	https://itol.embl.de/
MEGA	www.megasoftware.net
Microreact	https://microreact.org
RAxML	https://cme.h-its.org/exelixis/web/software/raxml/
PHYLOVIZ	http://www.phyloviz.net
PhyML	http://www.atgc-montpellier.fr/phyml/
SplitsTree	https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree/

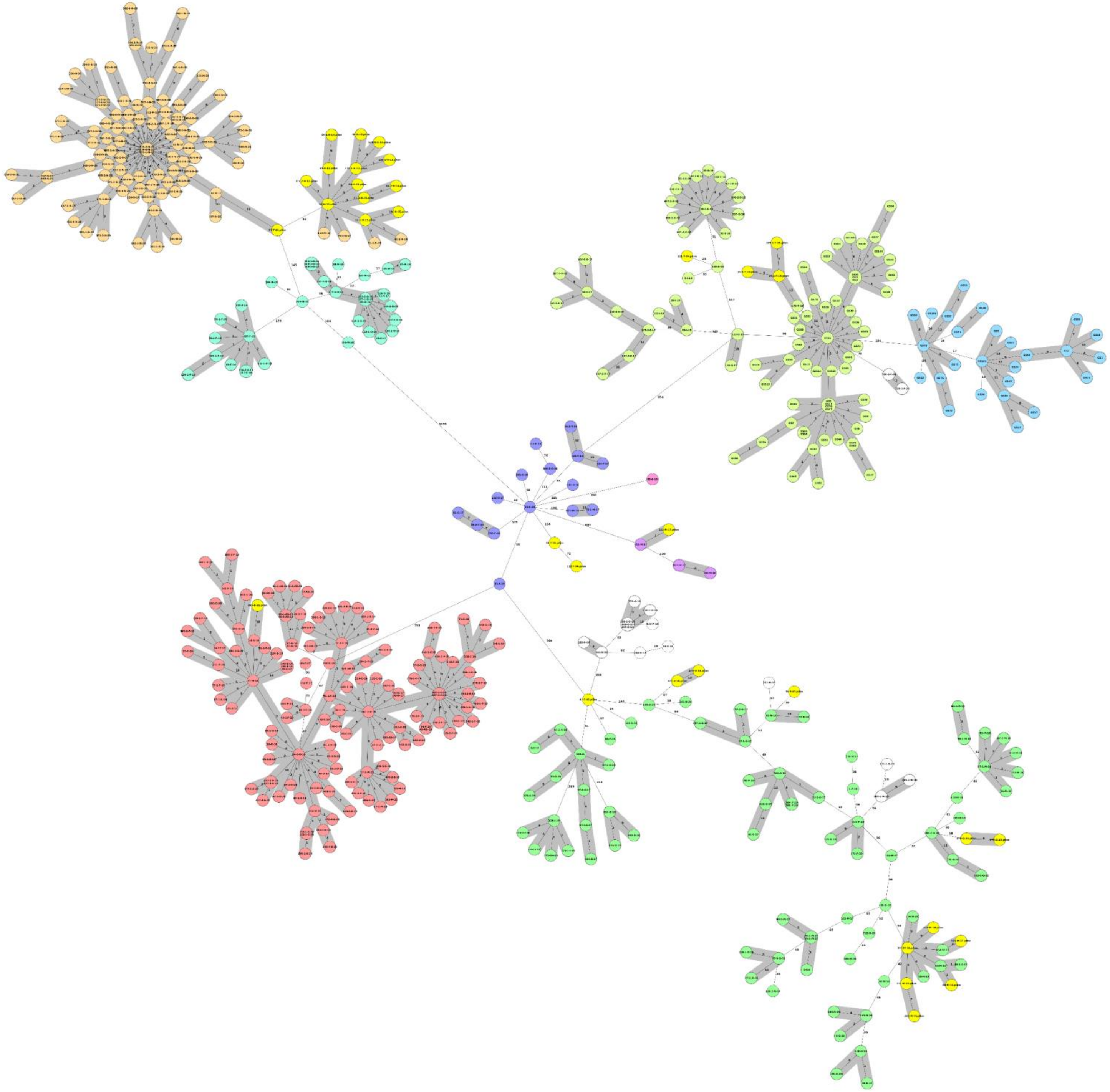
Interpretation of clustering data

Identification of clusters of genomes and deductions on whether two or more isolates are related or not

- Difficult question, all isolates of a species share a common ancestor
- Needs to be put into context to an outbreak and in relation with other isolates
- Number of allelic differences should be carefully considered
  - Not all alleles are called for all strains, should it be in the analyses if missing in some of the strains
  - Pairwise comparison considering all the alleles obtains more detailed information
- Pathogen-specific knowledge is required before a correct interpretation of a real outbreak is performed
  - Cluster cut-off is very species-specific, e.g. ≤ 2 SNPs for *Francisella tularensis* and ≤ 15 for *Campylobacter jejuni*
- Results of SNPs or allelic differences can be combined with phylogenetic tress for more robust interpretation of evolutionary relationship
- For outbreak investigation, it is crucial to include epidemiology and traceback evidence, not rely on clustering data alone







Questions?

Funded by
the European Union