EURL CPS
European Union Reference Laboratory for Coagulase Positive Staphylococci
http://eurl-staphylococci.anses.fr

European Union Reference Laboratory Foodborne Viruses

EURL Salmonella

EURL Lm
European Union Reference Laboratory for *Listeria monocytogenes*
http://eurl-listeria.anses.fr

EURL Campylobacter

EU-RL VTEC

Guidance document for cluster analysis of WGS data

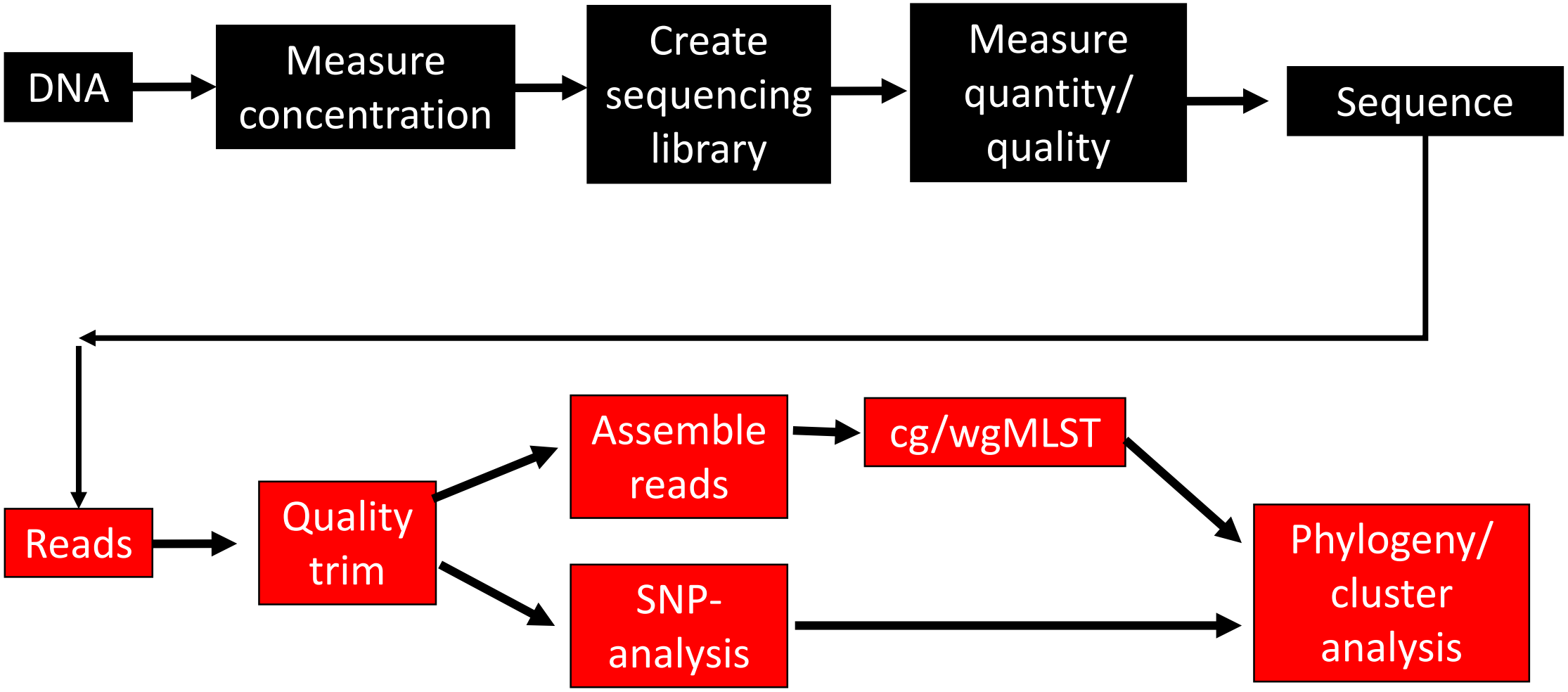
Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)

Inter EURLs WG on NGS - tasks

- Task 1 - Proficiency tests (EURL AMR)
- Task 2 - WGS laboratory procedures (EURL Parasites)
- Task 3 - Bioinformatics tools – (EURL VTEC)
- Task 4 - WGS Cluster Analysis (EURL Campylobacter)
- Task 5 - Bench marking (EURL Listeria)
- Task 6 - Trainings on NGS (EURL CPS)
- Task 7 - Reference and confirmatory testing using NGS (EURL Salmonella)
- Task 8 - Follow-up of ISO activities on WGS (All)

The documents from the tasks will be available at the EURL webpages late 2020

A typical WGS workflow



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

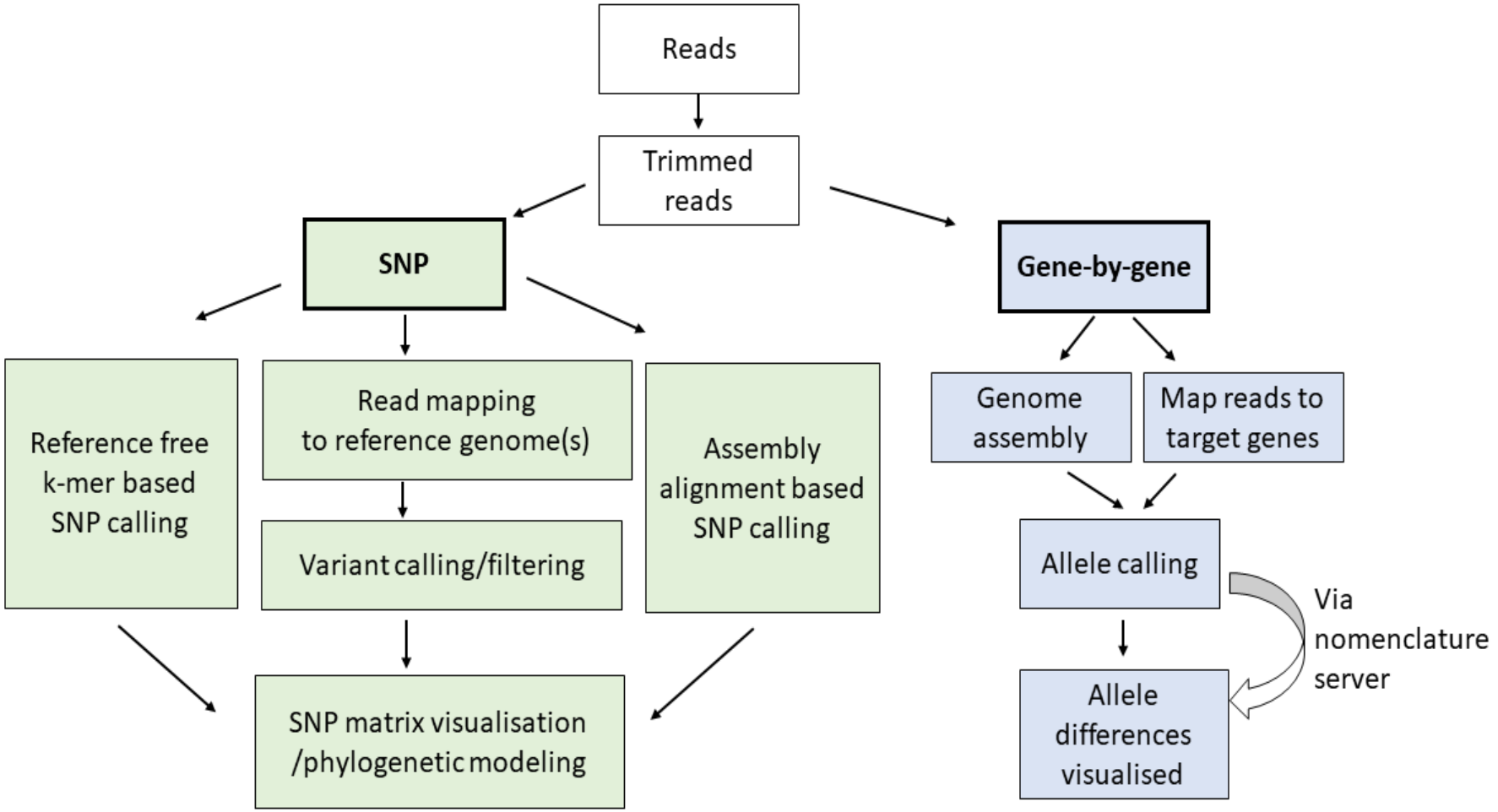
How to perform WGS cluster analysis

- Broadly, the most common comparison approaches can be divided into (i) the single nucleotide polymorphism (SNP) approach where individual mutations are used as separate phylogenetic markers and (ii) the gene-by-gene approach, where each variant of a gene is considered an allele.

- Both approaches involve several steps of analysis, each depending on bioinformatic scripts or software, 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 of these steps.
- It is important that the users have a solid knowledge of the software and methodology in order to produce correct and comparable results.
- Further, the different steps of analysis should be evaluated for each pathogen, sequencing machine and software intended for use when setting up the method.

- 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

The SNP-approach



The SNP approach

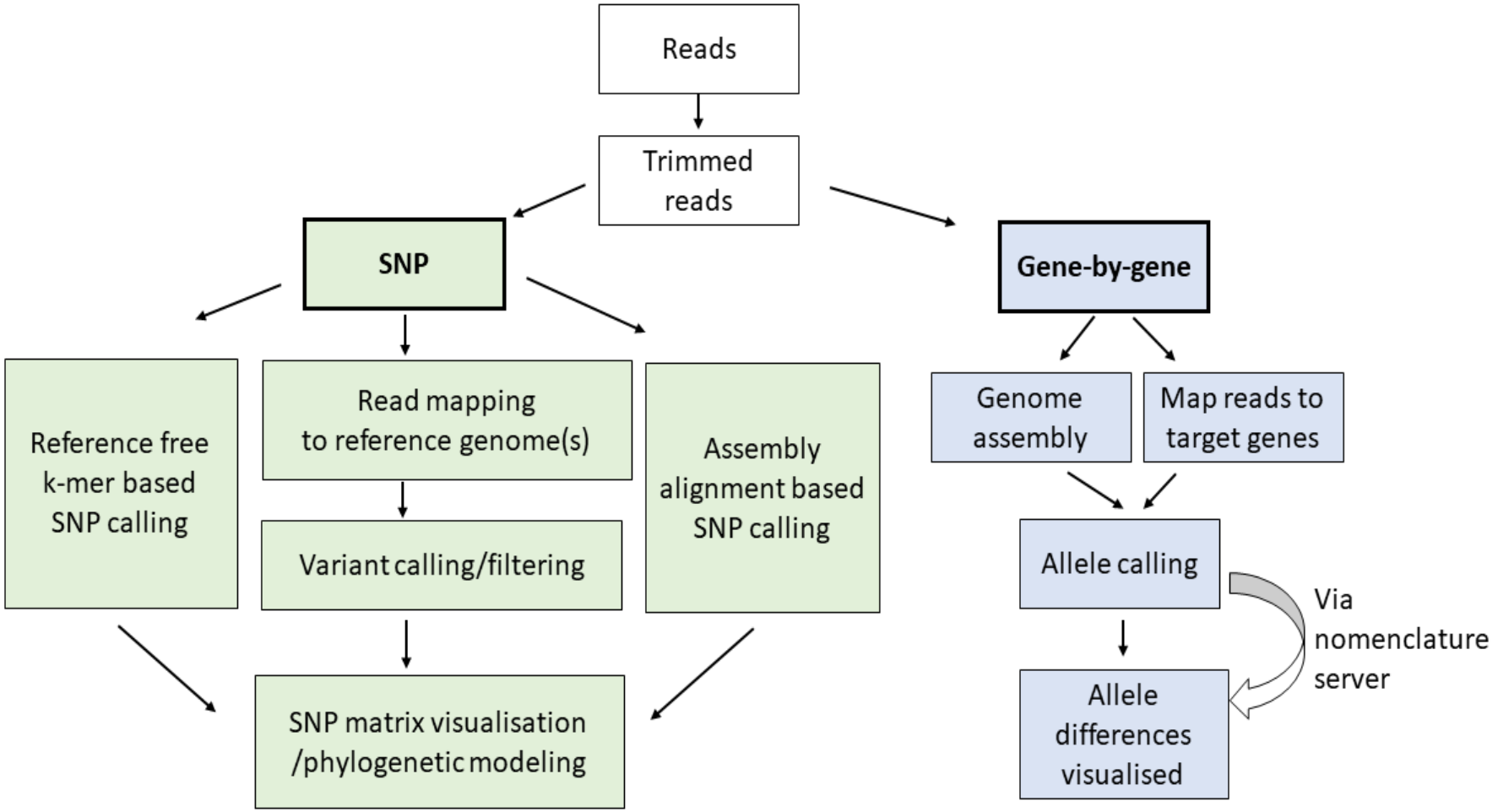
- Highest resolution
- “SNP pipelines”
- Reference genome
- Mapping the sequence reads to the reference
- Variant calling
- Quality filtering
- Difficult to standardize
- Can be computationally intensive

Variant filtering (SNP)

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:

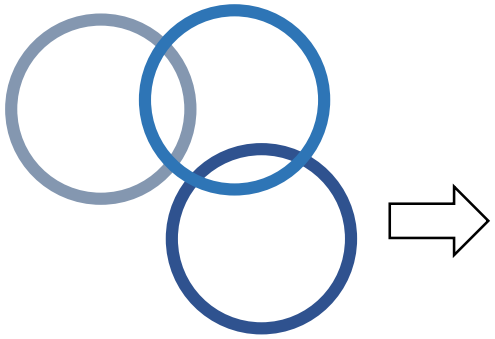
- Genomic regions with low coverage.
- Genomic regions with coverage much larger than the average coverage (possibly repetitive).
- Threshold for how large fraction of reads that must support the allele.
- Minimum quality values for the base calling of the reads at the SNP position.
- Minimum quality 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).

Gene-by-gene


Reads
Trimmed reads
SNP
Gene-by-gene
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
Genome assembly
Map reads to target genes
Allele calling
Via nomenclature server
Allele differences visualised

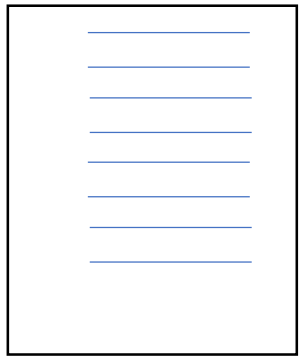
The gene-by-gene approach

- Multilocus sequence typing (MLST)
- Core genome (cg) MLST
- Whole genome (wg) MLST
- cg/wgMLST-scheme
- Input is usually assembled genomes
- Allele calling
- Easy to add new genomes
- The more genes the higher the resolution
- cgMLST - Suitable for surveillance
- wgMLST - Suitable for outbreak tracking

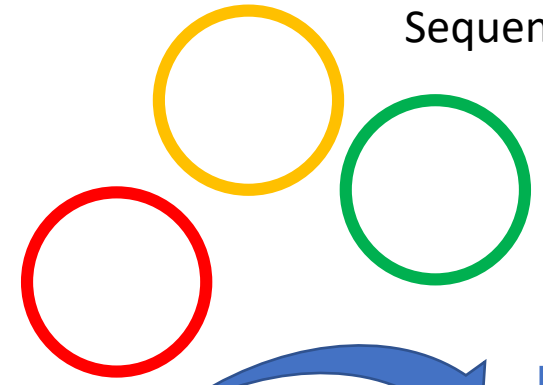


Reference-genome/s

Filtered genes, suitable for MLST



MLST scheme



Sequenced isolates

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

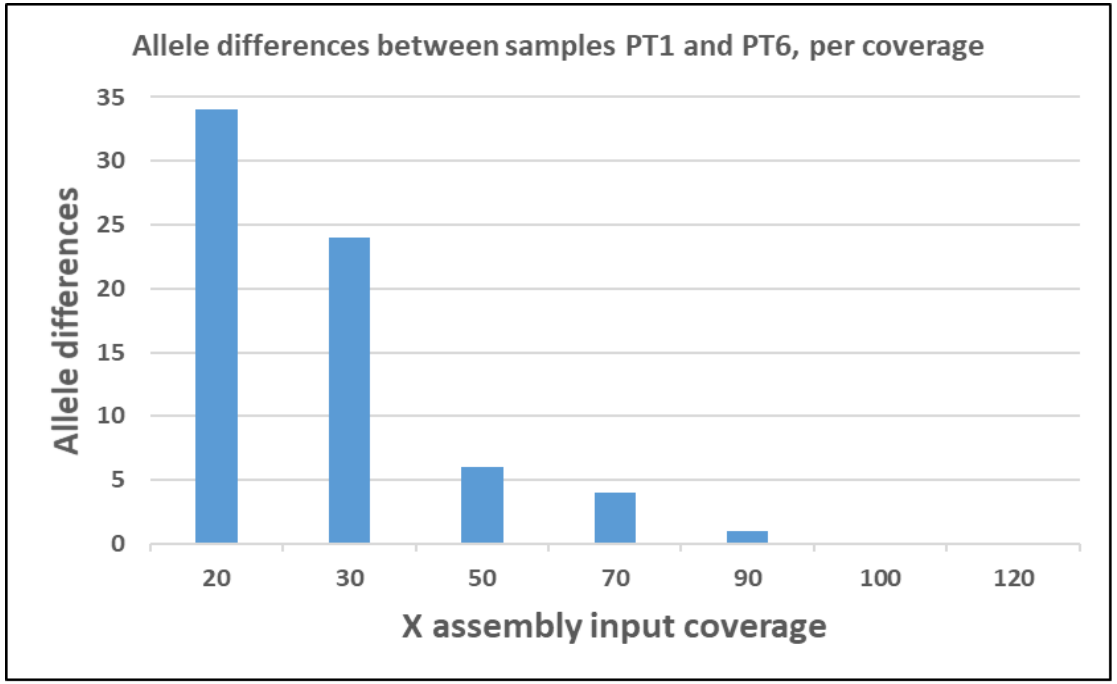
Validated cg/ wgMLST-schemes available for food-borne pathogens

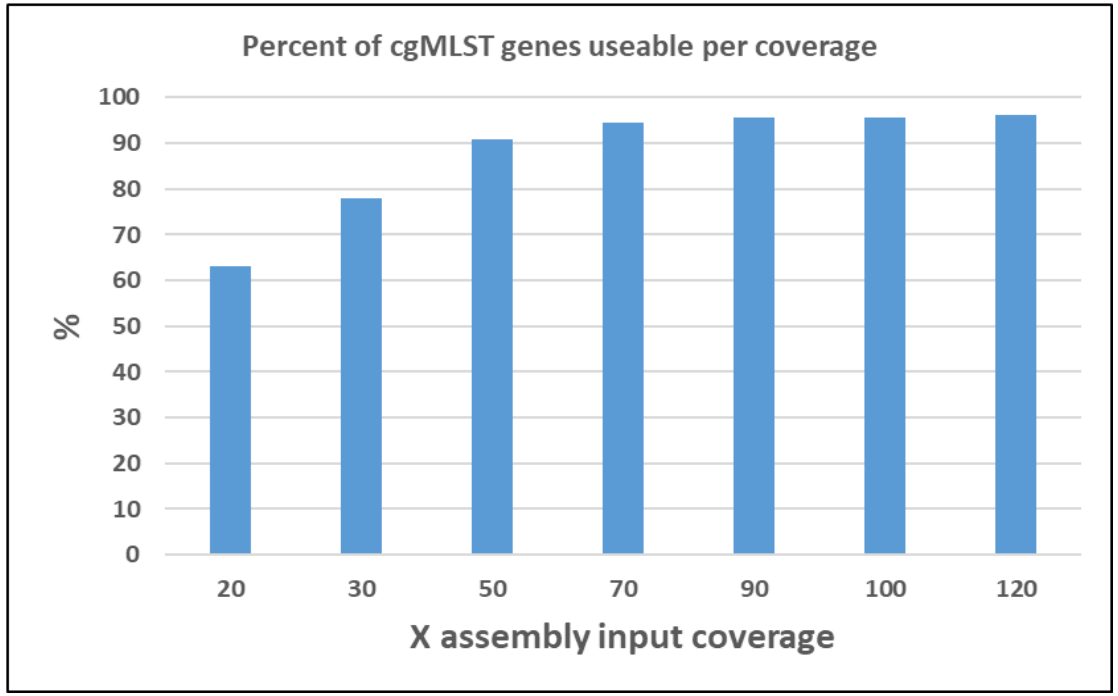
Pathogen	Site	Reference
Campylobacter	PubMLST.org	[15]
Salmonella	Enterobase *https://enterobase.warwick.ac.uk*	[10]
Escherichia coli	Enterobase *https://enterobase.warwick.ac.uk* Innuendo curated version of Enterobase scheme https://zenodo.org/record/1323690#.XzvSEOgza72	[10] [16]
Listeria monocytogenes	Institute Pasteur https://bigsdb.pasteur.fr/listeria	[17]
Staphylococcus aureus	www.cgMLST.org	[18]

Genome assembly

- SKESA, Velvet and SPAdes
- Benchmarking
- Bacterial de-novo-assembly from IonTorrent
- Assembly QC
- Command-line based
- Commercial software with GUI

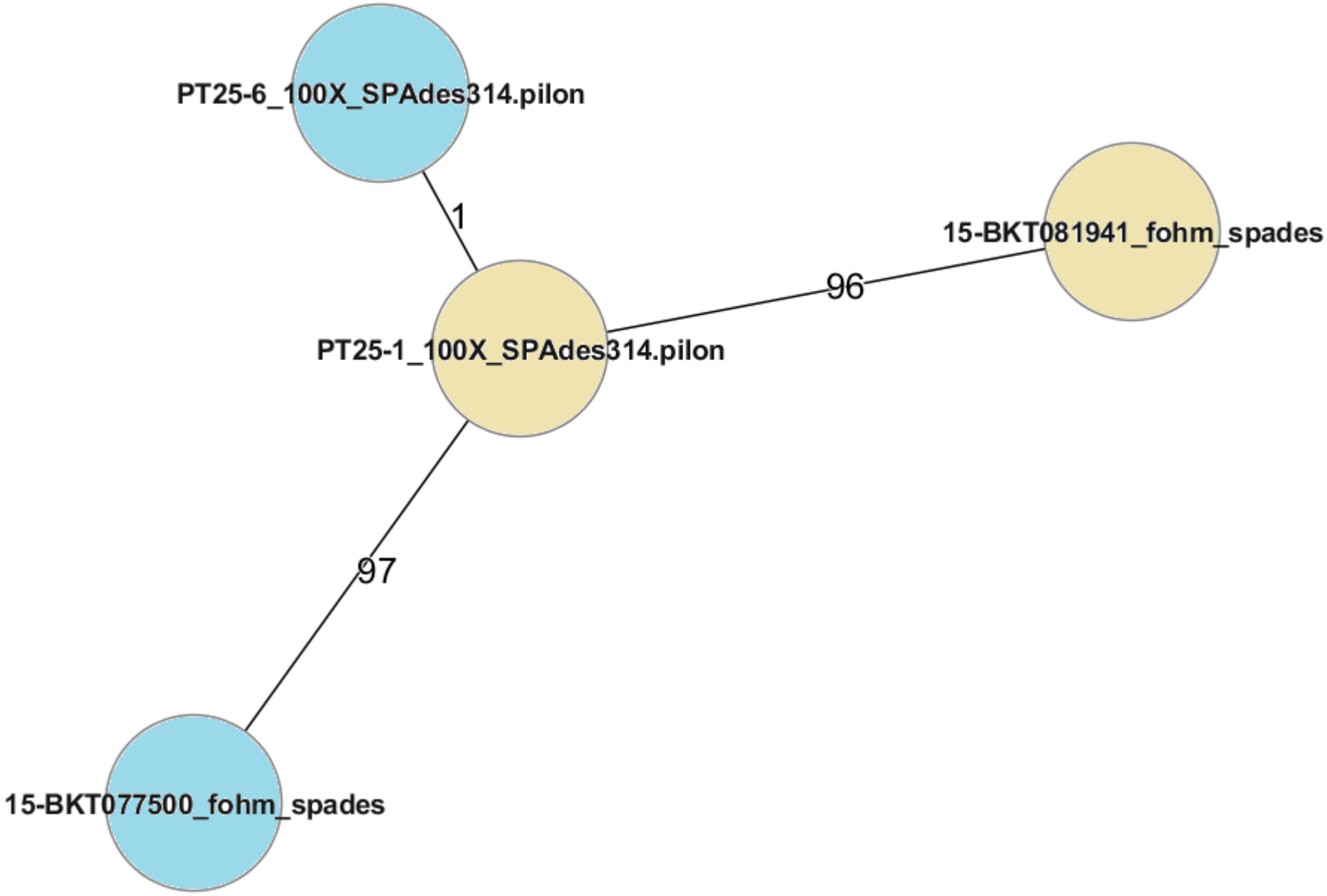
NexteraXT – Miseq, Spades, 2 isolates





Assembly of IonTorrent data

- Trimmomatic
- 100X coverage
- SPAdes 3.14



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
  - Costly software

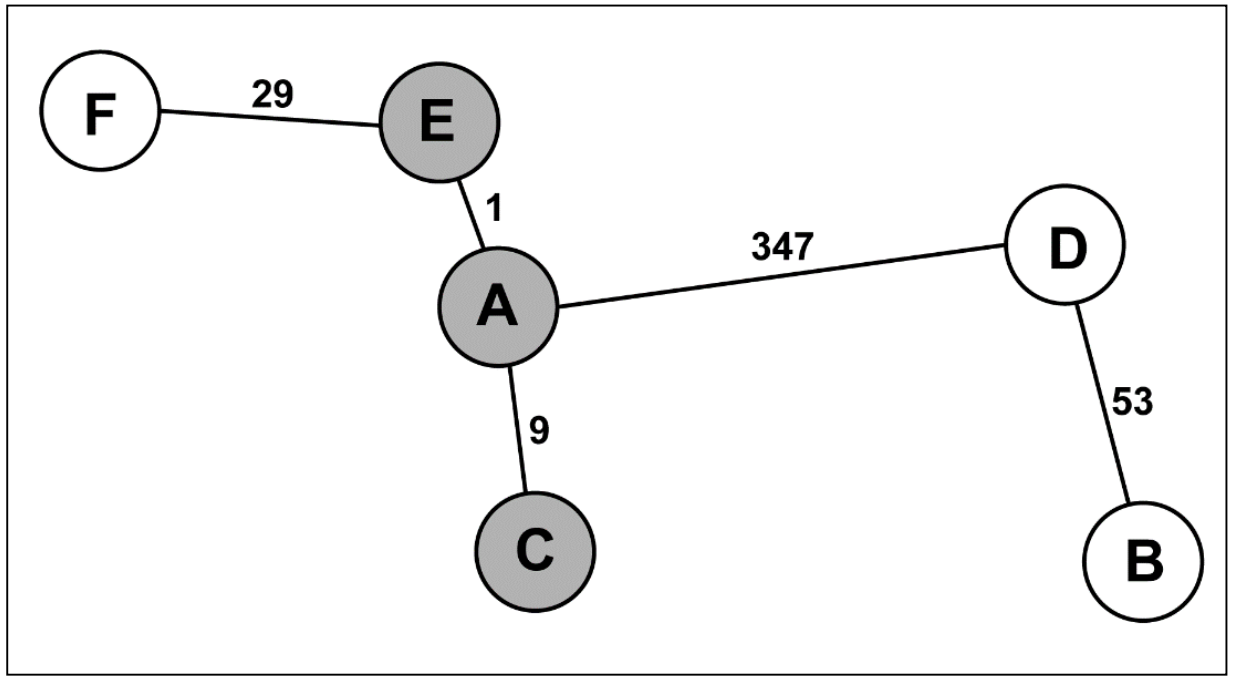
  + Full control of analysis
  + Not dependant on external provider

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. The software are presented in no specific order.

Comparability and differences

- SNP- and gene-by-gene-based methods often give comparable results
- Validation using confirmed outbreak data
- Some differences:
    - Intergenic regions
    - Collapsing mutations and indels
    - SNP restricted to reference genome
    - 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)

Interpretation of clustering data



A cgMLST result for six genomes visualised in a minimum spanning tree. The numbers represent the number of allele 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.

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