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

Bo Segerman, Hanna Skarin and Ásgeir Ástvaldsson

European Union Reference Laboratory for *Campylobacter*

Swedish Veterinary Agency, Uppsala, Sweden

Co-funded by the European Union

Co-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 European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor HaDEA can be held responsible for them.

Content

1. Glossary

Allele	Variant of a sequence. Every unique sequence is defined as a new allele.
Assembly	A merge of raw sequence reads into longer stretches of DNA aiming to reconstruct the original sequence.
BCF	A format to store genetic variants in nucleotide sequences (binary format)
cgMLST	Core genome multi locus sequence typing
Coverage	The average times a base is covered by a sequence read (100X = 100 times)
CRISPR	Clustered regularly interspaced short palindromic repeats (sequence elements used by the prokaryotic antiviral system)
ECDC	European Center for Disease Prevention and Control
EFSA	European Food Safety Authority
EURL	European Union reference laboratory
de Bruijn graph	A graph representation of overlaps between *k*-mers.
FASTA	A file format to store sequence data (no quality information)
FASTQ	A file format to store sequence data (with quality information)
INDEL	An insertion or deletion of bases
k-mer	A short sequence of the defined length k (e.g .if k=15, a 15-mer).
Mapping	To use a software that finds the best matching position of a sequence read in a reference sequence and gives an alignment for that match
MLST	Multi locus sequence typing
MST	Minimum spanning tree, a graph visualising distances
NRL	National reference laboratory
PCR	Polymerase chain reaction
SNP	Single nucleotide polymorphism
VCF	A format to store genetic variants in nucleotide sequences (text format)
WG	Working group
wgMLST	Whole genome multi locus sequence typing
WGS	Whole genome sequencing

2. Introduction

The continuous implementation of whole genome sequencing (WGS) has enabled new approaches for European surveillance and cross-country outbreak investigations. A new regulation will come into force in 2026, requiring EU and EFTA countries, as well as Northen Ireland (UK), to sequence the whole genome of *Campylobacter jejuni* (*C. jejuni*), *Campylobacter coli* (*C. coli*), *Escherichia coli* (*E. coli*), *Listeria monocytogenes* (*L. monocytogenes*), and *Salmonella enterica* (*S. enterica*) isolates from feed, animals, food, related environment linked to foodborne outbreaks, and to transmit the WGS results to EFSA[1]. Laboratories must make various decisions when implementing WGS analysis workflows, which can impact data interpretation and affect comparability. This document has been produced in the framework of the Inter-EURLs working group on next generation sequencing (Inter-EURLs WG on NGS). It aims to inform and support NRLs in the various choices of procedures for cluster analysis, where genetic distances between genomes are compared and visualised, enabling interpretation of the relatedness between genomes. The document focuses on the bacterial pathogens represented by the EURLs of the WG, as these methods are not yet applied to the same extent for parasites or viruses.

The two most widely used approaches for comparing bacterial genomes in cluster analysis are single nucleotide polymorphism (SNP) analysis and gene-by-gene analysis. In SNP analysis, individual mutations serve as distinct phylogenetic markers, whereas in gene-by-gene analysis, each gene is treated as the phylogenetic marker with comparisons based on variations in gene alleles. Gene-by-gene analyses, which are expanded versions of the multi-locus sequence typing (MLST) approach, are divided into core genome MLST (cgMLST), focusing on conserved core genes, and whole genome MLST (wgMLST), which considers all genes to assess genetic relatedness. Alternative comparison methods also exist that are based on *k*-mer distance estimation. Many of these methods are in practice quantifying SNPs and will be covered in section 2.1 and chapter 4 dedicated to SNP analysis. The SNP and cgMLST/wgMLST analyses are further detailed in the following sections (2.1 and 2.2). Chapter 3 provides an overview of the key distinctions between the different cluster analysis methods.

Cluster analysis involve several bioinformatic analysis steps that all can affect the end results. These steps may include e.g., read trimming, assembly, read-mapping, alignment, variant calling, allele calling and generation of dendrograms/trees. There are both freely available and commercial software solutions that perform these steps. Chapters 4-6 provide technical information on some of the most widely used methods for cluster analysis and cluster visualisation and list available software, including software that is used by the EURLs and/or the NRLs of the EURL-networks of the Inter-EURLs WG on NGS. The aim of this document is not to recommend a specific software but to provide an overview of available options.

It is important that the users have a solid understanding of the software and methodology they employ to ensure accurate, reliable and comparable results. Each step of the analysis procedure should be carefully evaluated for each specific pathogen and sequencing platform. 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. All deliverables produced by the Inter-EURLs WG on NGS can be reached from the EURL websites and Zenodo [2].

2.1. SNP analysis

Analysing WGS data by identifying SNPs is generally regarded as the method with the highest resolution for studying relatedness among bacterial isolates. SNPs can be very informative markers when analysed correctly. There are multiple software solutions available for identifying SNPs and many "SNP pipelines", which combine several bioinformatics tools into a workflow that generates an overview of SNP discrepancies and sometimes phylogenetic visualisation. Experienced bioinformaticians can develop their own customised SNP pipelines.

The most common approach is to determine SNPs by comparing WGS read data from isolates to a reference genome. However, there are also reference-free approaches, approaches that use several reference genomes, and approaches that use assembled genomes instead of sequence read data. A reference genome is ideally rather closely related to the strains under investigation to minimise alignment errors and enhance SNP calling accuracy. SNP identification is usually done by mapping the sequence reads to the reference using a read-mapping software. A variant calling software is then used to determine the SNPs relative to the reference. The analysis typically includes various quality filtering steps, which are very important to avoid calling of false SNPs. SNPs resulting from recombination rather than mutations can also be filtered out as they can cause closely related strains to appear more genetically distant. The variants for each of the isolates are then combined into a format that allows an analysis of phylogenetic relatedness, e.g. a multiple sequence alignment (msa) of core SNPs (SNP positions in conserved regions shared by all analysed isolates). The results are often visualised in a tree and/or a SNP distance matrix. Overall, the SNP approach is challenging to standardise due to the absence of consensus on pipeline design, reference genome criteria, uniform thresholds for quality filtering, and the handling of recombination. Analysing large datasets with SNP analysis can also be computationally demanding. A schematic view of the fundamental steps in the SNP approach is presented in Figure 1.

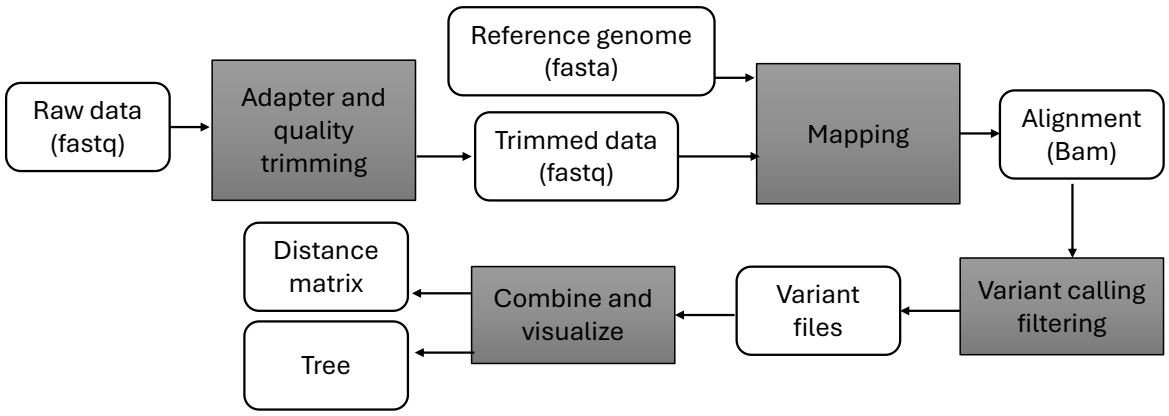

Figure 1. A typical outline of a SNP analysis.

There are some alternative comparison methods based on *k*-mers. *K*-mers are short, fixed-length sequences derived from a longer sequence, such as 15-mers or 20-mers. By slicing the original sequence into these smaller overlapping sub-sequences, *k*-mer analysis provides a computationally efficient way to analyse large datasets, enabling tasks like genome assembly and taxonomic read classification. Approaches using *k*-mers to directly infer phylogeny, often called alignment-free (AF) methods, can be based on comparing frequencies

of shared *k*-mers or comparing lengths of shared *k*-mers. The *mash* algorithm uses minhashing techniques to compare *k*-mer content and is commonly used to quickly estimate genetic distances (https://github.com/marbl/Mash). However, comparing *k*-mer frequencies alone often lacks the resolution needed to match the precision of SNP or cgMLST/wgMLST analysis. More intricate methods based on *k*-mer analysis have been developed that offer improved resolution. These resemble SNP analysis in that individual mutations are identified and used as phylogenetic markers. The KSNP method uses *k*-mer analysis to identify SNPs without a reference genome. PopPUNK (Population Partitioning Using Nucleotide K-mers) estimates SNP distances by quantifying and comparing *k*-mer matches at different *k*-mer lengths (since longer *k*-mers are more likely to contain SNPs) [3]. Another method is the split *k*-mer analysis (SKA) that analyses pairs of *k*-mers that are separated by one or more bases [4].

2.2. cgMLST and wgMLST analysis

The gene-by-gene approaches (cgMLST and wgMLST) are basically extended versions of the classical seven loci MLST procedure, incorporating up to thousands of loci (typically genes) [5]. In cgMLST analysis, the loci are restricted to a conserved set of genes present in nearly all strains of the species (core genes), whereas wgMLST analysis aims to cover as many genes as possible (core genes and accessory genes). The gene target list, which contains a set of numbered alleles (specific variants of a gene) for each locus, is known as the cgMLST or wgMLST scheme. cgMLST/wgMLST analysis usually takes assembled genomes as input, but some methods can accept raw reads. Allele calling is performed by comparing the targets defined in the scheme to the assembly, determining the isolate's combination of allelic sequences. If a new allele sequence is identified, it is assigned a new number. When multiple users contribute new alleles to a scheme, synchronisation becomes necessary. Therefore, central nomenclature servers can be used. Another approach is to use a hash algorithm to create a reproducible unique identifier for each allelic sequence [6]. The allele calling step can together with the assembly process be time consuming. However, if additional genomes are added to the analysis at a later stage, allele calling needs to be done only on the new genomes. For a cgMLST scheme, allele calling is expected to identify alleles for almost all loci in the target gene set. However, due to various factors, primarily related to data quality, some loci may not be found or may be flagged as problematic. The percentage of missing loci is a critical quality metric, as it reflects the completeness and reliability of the analysis. The result from a cgMLST/wgMLST analysis is a table with allele identifiers for each locus and analysed isolate. This table can be used to create a minimum spanning tree (MST) or be converted to a distance matrix, which gives an overview of the genetic relatedness. A schematic view of the fundamental steps in cgMLST/wgMLST analysis is presented in Figure 2.

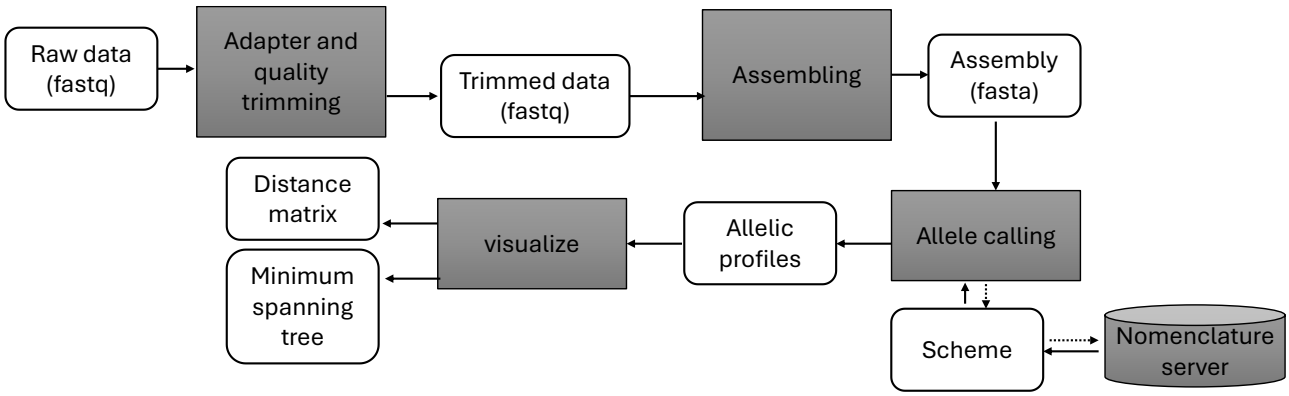

Figure 2. A typical outline of a cgMLST/wgMLST analysis.

3. Comparison between cluster analysis methods

The European Food Safety Authority (EFSA) and the European Centre for Disease Prevention and Control (ECDC) manage a One Health WGS system. It is composed of the EFSA WGS System and the ECDC EpiPulse Case, which exchange cgMLST profiles of *E. coli*, *L. monocytogenes* and *S. enterica* (*C. jejuni* and *C. coli* next in line) to detect multi-country clusters [7]. Also the PulseNet International network, which includes public health organisations from around the world with respect to food- and waterborne diseases, supports the use of cgMLST as the primary method for surveillance of *Salmonella* and *Campylobacter* outbreak clusters [8, 9], but also applies SNP and wgMLST data analysis methods for outbreak investigations.

Different reports have been published comparing the performance of SNP and gene-by-gene approaches and show that despite the differences between the methods, they generally group isolates into the same clusters. Evaluation studies of outbreak detection using whole genome data from *Campylobacter, E. coli*, *Listeria*, and *Salmonella,* show that regardless of analysis methodology, the results from the different approaches are usually concordant and comparable to each other [8, 10-16].

Regardless of the method, a thorough validation using reference datasets from confirmed outbreaks should be performed to be able to trust the chosen pipeline/software/parameters etc. This is further described in the 'Guidance document for WGS-benchmarking'. Despite the relatively small differences observed in performance, there are other differences between the approaches that should be considered when choosing method for analysis. These are summarised below in sections 3.1 and 3.2.

3.1. Differences in resolution

A cgMLST/wgMLST analysis is restricted to coding regions whereas SNP analysis also include intergenic regions. A gene that contains several mutations will in cgMLST/wgMLST be collapsed to one new allele and only counted as one change, but in a SNP analysis every mutation is counted. However, the accumulation of several SNPs in close proximity may have arisen during the same evolutionary event (e.g. a recombination) and quantifying them individually without correction for this may overestimate the genetic distance. Short insertions/deletions (INDELs) will not be counted by all SNP approaches. In cgMLST/wgMLST analysis, INDELs

will produce new alleles, unless they completely disrupt the gene and the locus may be classified as "missing" in the analysis.

The resolution of analysis of all clustering methods is directly related to the proportion of data included in the comparison. SNP analysis is generally restricted to what is present in the reference genome. Thus, a closely related reference genome increases the resolution of the analysis. Similarly, cgMLST/wgMLST analysis is restricted to the set of loci present in the schemes. The higher number of loci in the scheme, the higher the resolution of the analysis. Since cgMLST analysis is restricted to core regions, some resolution is lost, but comparability is on the other hand improved.

Generally, the resolution of analysis also depends on data quality, as SNPs or allele targets are discarded if they fail to meet the set quality thresholds (see 4.5 and 5.3). Quality factors include number of high-quality bases and read length after quality and adapter trimming, presence of contamination (both interspecies and intraspecies), assembly quality, and the efficiency of allele or SNP calling. In a cgMLST analysis, almost all loci are expected to be found, and the percentage of missing loci is often used as a quality measure.

Some technological factors affect the quality of results. Sequence data from low GC-content species such as *C. jejuni, C. coli, Staphylococcus aureus* (*S. aureus*), and *L. monocytogenes* is significantly affected by GC-dependent coverage bias when using Illumina Nextera XT library preparation. As a result, a larger amount of input data is required compared to other library preparation methods to ensure sufficient coverage over the entire genome [17]. Further, different NGS technologies can produce different types of errors, which is important to consider when choosing downstream methods for analysis. Illumina is today the most used technology for WGS based cluster analysis, but alternatives exist (e.g., BGI, Element Biosciences, Ion torrent, Oxford Nanopore, PacBio). The IonTorrent/proton technologies are prone to produce errors determining homopolymer lengths, which may lead to incorrect frameshifts when annotating the genomes resulting in false pseudogenes [18]. For this reason, a proper validation is needed when you want to compare Ion Torrent and Illumina data (see 'Guidance document for WGS-benchmarking'). SNP-based analyses tend to be more resilient to technical noise in sequencing data due to their quality filtering mechanisms. These filters can remove erroneous variants without disrupting the analysis in unaffected regions of the gene.

3.2. Comparability of results and nomenclature

The results from SNP analysis performed at different laboratories can be compared if the SNP calling was performed using the same reference genome and the same SNP pipeline and parameters [19]. However, the results have been considered more difficult to communicate between laboratories than those produced by the gene-by-gene approach, since there is no general approach for nomenclature when doing SNP-analysis. Public Health England (PHE) has developed the system of SNP addresses as unique identifiers within a given dataset [20]. However, this system requires using the same database (SnapperDB) to be able to identify new SNP addresses. Public Health Agency of Canada has developed an application called BioHansel, which uses canonical SNP genotyping schemas (including selected phylogenetically informative SNPs) for genotyping of some *Salmonella* serovars [21]. This application of SNP data enables the use of nomenclature, providing that the cooperating laboratories uses the same application.

If several laboratories perform an analysis using the same cgMLST/wgMLST-scheme and the allele identifiers are accessible, they can directly compare and communicate the results, even if the analysis is run by different software solutions. This also means that results from different analysis run at the same laboratory can be compared without having to call the alleles again. If a cgMLST/wgMLST scheme is updated locally (i.e., different laboratories or users implement their own updates independently), it can affect comparability across datasets. Allele hashing is a method that can be used to compare results between laboratories without the need of a central allele nomenclature [6].

4. SNP-analysis methods and software

A single nucleotide polymorphism (SNP) is a nucleotide difference in a specific position of a genome compared to another genome/reference genome. Some SNP analysis software also collects information about short INDELs. There are several "pipelines" publicly available for running a SNP analysis. Most of them depend on bioinformatic tools developed and maintained by other research groups for making the core analysis steps. Many pipelines also offer the possibility to choose between different tools for performing the necessary analytical steps. Chapter 4 lists some SNP pipelines and briefly describes the most common analysis steps included in the pipelines.

4.1. SNP pipelines

Several pipelines exist that combines the required steps to do SNP analysis in bacterial sequences. Some of them can be found in Table 1.

Table 1. SNP and *k*-mer based pipelines.

SOFTWARE	LINK TO SOFTWARE
BactSNP	http://platanus.bio.titech.ac.jp/bactsnp
CFSAN	https://github.com/CFSAN-Biostatistics/snp-pipeline
iVARCall2	https://github.com/afelten-Anses/VARtools/tree/master/iVARCall2
ISG	https://github.com/TGenNorth/ISGPipeline
kSNP	https://sourceforge.net/projects/ksnp/
Lyve-Set	https://github.com/lskatz/lyve-SET
NASP	https://github.com/TGenNorth/NASP
parsnp	https://github.com/marbl/parsnp
PHEnix	https://github.com/phe-bioinformatics/PHEnix
PopPUNK	https://github.com/bacpop/PopPUNK
Snippy	https://github.com/tseemann/snippy
SKA	https://github.com/simonrharris/SKA
SPANDx	https://github.com/dsarov/SPANDx

Some pipelines are also available as online services and they are summarised in Table 2.

Table 2. SNP pipelines available as online services.

SOFTWARE	LINK TO SOFTWARE
ARIES (Galaxy server that includes e.g. KSNP, PopPUNK, FDA SNP pipeline)	https://www.iss.it/site/aries
CSI Phylogeny	https://cge.food.dtu.dk/services/CSIPhylogeny/
Enterobase	https://enterobase.warwick.ac.uk/
GALAXY@SCIENSANO	https://galaxy.sciensano.be/
NDtree	https://cge.food.dtu.dk/services/NDtree/
RealPhy	https://realphy.unibas.ch/realphy/

Most SNP pipelines are built by joining several analysis steps that often are similar between the pipelines. In some pipelines it is also possible to choose between different software solutions for some of the analysis steps. It is important to read the documentation of the pipeline so that proper parameter settings are used. Below, some of the main analysis steps typically used in the pipelines are described.

4.2. Read-mapping

Many SNP pipelines use unassembled reads as input, which may have been subjected to some quality trimming and adapter removal. The reads are commonly mapped to a reference genome sequence with a mapping program. There are also pipelines that use more than one reference genome (e.g., RealPhy). It is important to choose a reference genome representative of the pathogen or of a subset of the pathogen studied in order to maximise the resolution of the analysis. Mapping programs position reads on a reference genome and provide alignment information for the mapped region.

The most commonly used mappers are bowtie2 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml) and BWA (https://bio-bwa.sourceforge.net). The read alignment is usually stored in file formats called BAM or SAM (which is a text version of the binary BAM format). SAMtools (http://www.htslib.org/) is often required by the pipelines to convert and sort/manipulate BAM/SAM files. Picard tools (https://broadinstitute.github.io/picard/) is sometimes used to remove duplicate reads from the analysis.

4.3. Non-read-mapping based solutions

Some SNP pipelines require, or can optionally also use, assembled genomes as input. The genomes are then compared to the reference genome with whole genome alignment programs such as MUMmer/Nucmer (http://mummer.sourceforge.net/), mugsy (http://mugsy.sourceforge.net/) or mauve (http://darlinglab.org/mauve/mauve.html) and the SNPs are extracted from these alignments. A disadvantage with SNP identification from assembled genomes is that the quality values of the underlying read bases cannot be used in the evaluation of a SNP.

Some SNP pipelines (e.g., kSNP) do not use reference genomes, but instead compare all *k*-mers present in the assembled genomes/sequence read files to identify SNPs.

In addition, some variant calling software solutions, e.g. Cortex, use an approach that loads the reads into a de Bruijn graph (http://cortexassembler.sourceforge.net/index_cortex_var.html).

4.4. Variant calling

From the BAM/SAM alignment files, variants can be called by several variant calling software solutions. This may include using SAMtools to convert the BAM/SAM file to a “pileup” file format, which describes the alignment nucleotide position-by-position rather that read-by-read. Variants are typically stored in the variant calling format (VCF) and/or its binary counterpart BCF. The bcftools (http://www.htslib.org/) is often required to manipulate the VCF/BCF files. Most variant calling software were originally designed to work with diploid genomes but can be used for haploid genomes as well.

4.5. Variant filtering and merging of results

Incorrect SNPs/variants may be called for various 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 (under a certain threshold) or where reads are only mapped in one direction may be excluded/masked.

- Genomic regions with coverage much larger than the average coverage may be excluded (possibly repetitive).
- Threshold for how large fraction of reads must support the allele. If more than one allele in the same position is indicated by the alignment, the SNP may be discarded, as bacterial unrepetitive genes normally should fall out as homozygous.
- 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 contigs ends may be excluded.
- Duplicate regions or CRISPR regions in the reference sequence may be excluded/masked.
- Regions where many SNPs are found in close proximity to each other may be excluded (possible recombination or misaligned reads).
- Duplicate reads in the alignment may be removed (possible PCR duplicates, not true unique sequenced fragments).

Finally, the variants identified in each isolate need to be combined into a SNP matrix or a FASTA file summarising the SNPs. The combined data often includes only polymorphic regions but may alternatively also include monomorphic positions (conserved). Including monomorphic positions may be beneficial for inferring phylogeny but increases the computational requirements drastically. Visualisation of data is further described in chapter 6. There are also tools that can annotate a SNP result matrix (e.g. snpEff, http://snpeff.sourceforge.net/).

5. cgMLST/wgMLST analysis methods and software

Chapter 5 describes the different steps of analysis included in the cgMLST/wgMLST analysis and lists commonly used pipelines and software.

5.1. cgMLST/wgMLST-schemes

The first step in gene-by-gene analysis is selecting a cgMLST/wgMLST scheme that defines the target genes for comparison across sequenced genomes. If a suitable scheme is unavailable, most cgMLST/wgMLST solutions allow the creation of custom schemes. A cgMLST-scheme is relatively stable and should produce comparable results for almost any genome of the species. This enables a stable nomenclature and is suitable for surveillance purposes. A wgMLST-scheme can provide a higher resolution than a cgMLST-scheme and can be useful for outbreak investigations and similar studies.

The benefit of using online publicly available databases with stable schemes is the possibility to compare isolates to a high number of other deposited genomes or allele profiles. This is a prerequisite for continuous surveillance of pathogens and detection of cross-country outbreaks. Table 3 lists databases and schemes available for several food-borne pathogens.

Table 3. Public databases and cgMLST/wgMLST-schemes available for the bacterial food-borne pathogens represented by EURLs of the working group.

PATHOGEN	SITE	REFERENCE
C. jejuni and *C. coli*	PubMLST: https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_seqdef&page=schemes Ridom: https://www.cgmlst.org/ncs/schema/schema/Cjejuni382/ Innuendo: https://zenodo.org/record/1322564 https://chewbbaca.online/species/4	[22] [23]
E. coli (including STEC)	Enterobase: https://enterobase.warwick.ac.uk/species/index/ecoli Innuendo (curated version of Enterobase scheme): https://zenodo.org/record/1323690#.XzvSEOgza72 https://chewbbaca.online/species/5	[15] [24]
L. monocytogenes	Institute Pasteur: https://bigsdb.pasteur.fr/listeria https://chewbbaca.online/species/6 Ridom: https://www.cgmlst.org/ncs/schema/Lmonocytogenes360/	[25]
S. enterica	Enterobase: https://enterobase.warwick.ac.uk/species/index/senterica Innuendo: https://chewbbaca.online/species/8	[15]
S. aureus	Ridom: www.cgMLST.org/ncs/schema/141106/ PubMLST: https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=schemes	[26]

Standalone cgMLST/wgMLST solutions sometimes comes with a dedicated nomenclature server. Ridom SeqSphere+ hosts a nomenclature server called cgMLST.org (https://www.cgmlst.org/). Chewie-NS is a nomenclature server that integrates with chewBBACA (https://chewbbaca.online/).

5.2. Assembly

The most common input format for cgMLST/wgMLST analysis is a genome assembly. Read trimming is a critical step before assembly to remove low-quality sequences and adapter contamination. Examples of trimming software are Trimmomatic (https://github.com/usadellab/Trimmomatic) and fastp (https://github.com/OpenGene/fastp). It is strongly recommended to downsample FASTQ files when coverage is excessively high (files are often capped at 100X), as this helps prevent low-level contaminants from reaching sufficient coverage to be assembled. The most used assemblers for Illumina data are SPAdes https://github.com/ablab/spades) and SKESA (https://github.com/ncbi/SKESA).

Metrics that can be used for quality control of the assembly are assembly length, GC-content, N50, and number of contigs. The Quast tool is commonly used to evaluate assembly quality (https://github.com/ablab/quast). Assembly quality can also be assessed using CheckM, which evaluates the completeness and contamination of microbial genomes by analysing marker genes specific to a given lineage (https://github.com/Ecogenomics/CheckM). A poor assembly will often have a negative impact on the result of downstream analysis. There are assembly correcting software, e.g. Pilon (https://github.com/broadinstitute/pilon), that by mapping reads back to contigs can correct the assembly from errors created in the assembly process. However, assembly correction can sometimes introduce new errors due to misalignment or incorrect mapping of reads. The tools for assembly correction need to be properly benchmarked in each laboratory. Finally, applying strategies to filter out short and low-coverage contigs from the assembly is highly beneficial, as it removes contamination that otherwise can interfere with downstream analyses. When using SPAdes, this may involve adjusting the --cov-cutoff parameter.

Shovill (https://github.com/tseemann/shovill) is an assembly pipeline which combines downsampling, adapter trimming, assembly with SPAdes (also supports SKESA), assembly correction, and filtering of short and low-coverage contigs.

Unicycler is an alternative wrapper around SPAdes that optimise the assemble process and filters low-depth contigs (https://github.com/rrwick/Unicycler).

Assembly, trimming reads, correcting assemblies and calculating assembly metrics are often performed in command-line based software, which requires some basic Linux and bioinformatics knowledge. However, there are pipelines and commercial software available that reduce the need for extensive command-line input or provide a graphical user interface (GUI), making the analysis process more user-friendly (see below).

Since the type of errors produced by different sequencing platforms differ from each other, a proper validation should be performed when using assembled contigs derived from different sequencing platforms in the same gene-by-gene comparison.

Some methods, such as MentaLiST and cgMLSTFinder uses FASTQ files as input. MentaLiST uses *k*-mers to match alleles from the scheme directly with the FASTQ files whereas cgMLSTFinder uses a read-mapping

approach. These assembly-free approaches may face limitations such as quality issues and difficulties in detecting new alleles, making thorough validation essential during implementation.

5.3. Allele calling

The allele calling step is carried out by specialized software (an allele caller), which often employs an alignment tool such as Basic Local Alignment Search Tool (BLAST) (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) to compare the genome analysed with the cgMLST/wgMLST scheme. Different online and standalone solutions that performs cgMLST/wgMLST analysis can be powered by the same allele caller and they can span the whole cluster analysis chain or different sub-parts of it. The called alleles are presented in a results table with allele numbers/identifiers. These numbers/identifiers can either be sequential values assigned to each allele in the nomenclature server or scheme, or hashes (calculated with algorithms such as CRC32, MD5, or SHA-256) produced directly from the DNA sequence of the identified allele. Failed allele calling can be due to several reasons including missing or incomplete gene targets, assembly quality problems or contaminations.

Online allele calling services and tools linked to nomenclature servers utilise a shared database to assign allele identifiers. When a novel allele is identified, a new identifier is automatically generated and deposited into the database, ensuring consistent allele tracking. When using a local approach (i.e., not working towards a nomenclature server) the alleles will be designated local allele identifiers. The allele calling step can be computationally intense since many BLAST comparisons are made. The allele differences can be visualised in a minimum spanning tree (MST) or a distance matrix which are two ways to visualise the number of allele differences (ADs) between the isolates in the analysis. See chapter 6 for examples of MSTs and how to interpret them.

There are free online services that can perform cgMLST/wgMLST-analysis. Disadvantages of this approach include dependency on the service provider, downtimes of server, long waiting times and a lack of control of the analysis. Online servers include PubMLST, Enterobase and the cgMLSTFinder (Table 4).

Systems for local operation may have the capability to connect to a nomenclature server. There are both commercial and free software available. Ridom SeqSphere+ is one of the most widely used commercial solutions for cgMLST/wgMLST analysis. It also contains some functionality for trimming and assembly. chewBBACA is a comprehensive, free, open-source solution for cgMLST/wgMLST analysis. ChewieSnake is a pipeline built around chewBBACA that integrates several steps, including trimming (using fastp), assembly (using Shovill), allele calling (using chewBBACA), and a server-free, hashing-based nomenclature approach for allele identification.

The EFSA One Health WGS system is another online tool developed and used by EFSA for rapid detection and management of multi-country foodborne outbreaks by running a cgMLST analysis on uploaded genomic data. The online system is restricted to officially authorised users, but the source code of the analytical pipeline is openly available for download, enabling local implementations. This End-to-end pipeline, designed for analysis of illumina and ion torrent reads, incorporates various open-source tools for QC (fastp, confindr, CheckM) and assembly (Shovill) before running the cgMLST analysis using chewBBACA and hashing-based nomenclature approach from ChewieSnake. It will also run tools for AMR, MLST and other types of analysis

depending on the organisms being analysed. The pipeline is currently available for *L. monocytogenes*, *Salmonella* and *E. coli* (STEC). Analysis for *C. jejuni* and *C. coli* is scheduled to become available by 2026.

A selection of available software solutions for cgMLST/wgMLST analysis is presented in Table 4.

Table 4. A selection of available software solutions for local or online operation of cgMLST/wgMLST analysis.

SOFTWARE	DESCRIPTION	LINK TO SOFTWARE
cgMLSTFinder	Online service: Standalone version:	https://cge.food.dtu.dk/services/cgMLSTFinder/ https://bitbucket.org/genomicepidemiology/cgmlstfinder/src/master/
chewBBACA	Standalone allele calling engine	https://github.com/theInnuendoProject/chewBBACA
chewieSnake	Pipeline built on chewBBACA	https://gitlab.com/bfr_bioinformatics/chewieSnake
Enterobase	Online service: Source code: Allele caller (etoki):	https://enterobase.warwick.ac.uk/ https://bitbucket.org/enterobase/workspace/repositories/ https://bitbucket.org/enterobase/etoki_enterobase/src/master/
GeP/FastGeP	Standalone pipeline	https://github.com/jizhang-nz
SeqSphere+	Commercial suit (built in allele caller)	https://www.ridom.de/seqsphere/
PubMLST/BIGSdb	Online service (built in allele caller)	https://pubmlst.org/
EFSA WGS One Health pipeline	End to end pipeline using chewBBACA	https://dev.azure.com/efsa-devops/EFSA/_git/efsa.wgs.onehealth
MentaLiST	Standalone assembly-free pipeline	https://github.com/WGS-TB/MentaLiST

The online services do not cover the exact same pathogens so one service cannot be used for all types of species. Enterobase covers the following pathogens: *Clostridioides*, *Escherichia/Shigella*, *Helicobacter*, *Moraxella*, *Mycobacterium tuberculosis, Salmonella*, *Streptococcus*, *Vibrio*, and *Yersinia*. cgMLSTFinder covers: *Campylobacter* (PubMLST v1 scheme), *Clostridioides* (Enterobase scheme), *E. coli* (Enterobase scheme), *L. monocytogenes* (Institut Pasteur scheme), *Salmonella* (Enterobase scheme) and *Yersinia* (Enterobase scheme). PubMLST covers a multitude of pathogenic species except for *L. monocytogenes*, which is instead available at the Institut Pasteur's BIGSdb instance. *E. coli* and *Salmonella* can be analysed in PubMLST but alleles and scheme definitions for these pathogens are synchronised from Enterobase and all submissions must be performed to Enterobase. This means that Enterobase should be the preferred choice for these two species since no new alleles can be assigned via PubMLST.

cgMLST/wgMLST analysis is also available from some Galaxy servers such as:

- ARIES (https://www.iss.it/site/aries) using chewBBACA to call alleles for *E. coli* (Innuendo scheme) and *L. monocytogenes* (Pasteur scheme).
- GALAXY@SCIENSANO (https://galaxy.sciensano.be/) using BLAST, KMA, or SRST2 to call alleles for *C. jejuni/C. coli* (PubMLST scheme), *E. coli* (Enterobase or Innuendo scheme), *L. monocytogenes* (Pasteur scheme), *S. aureus* (PubMLST scheme), *S. enterica* (Enterobase scheme) and other species.

6. Visualisation of clustering data

The number of allele differences (ADs) or SNPs can be directly derived from a table and converted into a distance matrix describing the pairwise distances (Table 5), or the results can be visualised in for example a minimum spanning tree (MST) (Figure 3).

Table 5. An example of a distance matrix generated 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 allelic differences or SNPs detected among each pair of strains analysed. In the example given in Table 5, the results of a cgMLST analysis gave a total of 58 allelic differences between STRAIN1 and STRAIN2, 1,211 allelic differences between STRAIN1 and STRAIN3, and five allelic differences between STRAIN2 and STRAIN3.

An MST is an undirected graph that shows the shortest distances between individual analysed components. In the MST shown in Figure 3, isolates A and C are separated by nine allelic differences, which means that out of the 1,340 genes investigated in this analysis, only nine genes showed differing differences in their sequences. This indicates that they are genetically similar and share a recent common ancestor. The same is true for isolate E, which is even more closely related to isolate A, likely sharing an ancestor even closer in time. In contrast, the high number of allelic differences between D and A indicate that they did not recently originate from the same source.

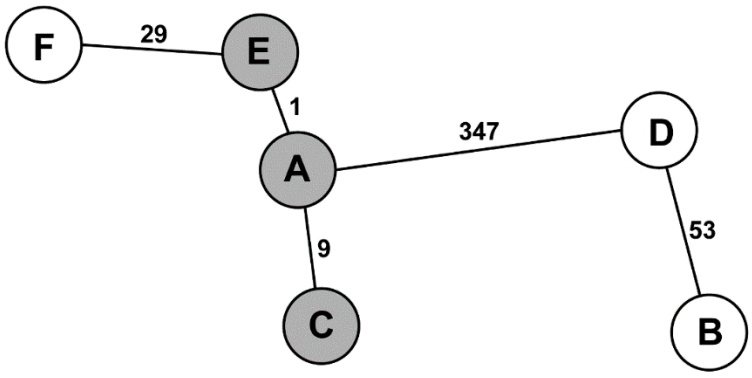

Figure 3. A cgMLST analysis result for six genomes visualised in a minimum spanning tree (MST). The numbers between the sample names represent the number of allelic differences between the samples. The line lengths are not proportional to the number of differences. The total number of gene targets compared in this analysis is 1,340. The identified cluster has been highlighted in grey, with a cluster definition set to ≤ 10 alleles differences.

The results of a cluster analysis can also be visualised in the form of a phylogenetic tree, rooted or unrooted. Rooted trees often use an outgroup, which infers the oldest point in the tree, i.e., identifies a most recent

common ancestor (MRCA) for the isolates. This gives information on the direction of the evolutionary changes. The robustness of phylogenetic trees can be estimated by bootstrapping, which is a statistical procedure that creates many simulated replicates by resampling with replacement. Phylogenetic trees may be produced from a distance matrix or directly from the SNP alignment data. Phylogenetic trees built from distance matrix data use clustering methods such as Neighbour-joining (NJ) and UPGMA (Unweighted Pair Group Method with Arithmetic mean). One commonly used software solution applying these algorithms that can provide both MST and phylogenetic trees from molecular epidemiological data (such as SNP and cgMLST/wgMLST) is the tool PHYLOViZ [27, 28]. Phylogeny inferred from distance matrix-based methods (NJ and UPGMA) involves fitting all characters to the tree at once whereas more advanced methods fit individual characters to the tree individually. These methods include maximum parsimony, maximum likelihood and Bayesian methods. These methods use not just the pairwise distance data but the whole alignment data. Maximum parsimony minimises the total number of evolutionary steps in the tree whereas maximum likelihood and Bayesian methods use statistical models to determine the tree.

Phylogeny can be inferred and visualised by a number of software solutions. A selection is listed in Table 6.

Table 6. 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/
GrapeTree	https://github.com/achtman-lab/GrapeTree
Gubbins (depends on RAxML/FastTree)	https://github.com/nickjcroucher/gubbins
IQ-TREE	https://github.com/Cibiv/IQ-TREE
iTOL	https://itol.embl.de/
MEGA	https://www.megasoftware.net
Microreact	https://microreact.org
PHYLOViZ	http://www.phyloviz.net https://online2.phyloviz.net/index
PhyML	http://www.atgc-montpellier.fr/phyml/
RAxML	https://cme.h-its.org/exelixis/web/software/raxml/
ReporTree	https://github.com/insapathogenomics/ReporTree
SplitsTree	https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree/ https://github.com/husonlab/splitstree6
SPREAD (based on GrapeTree)	https://github.com/genpat-it/spread

7. Interpretation of clustering data

The interpretation of the results from the SNP-based or gene-by-gene approaches means identification of clusters of genomes and deductions on whether two or more isolates are closely related. Determining if two isolates are “related or not” is a difficult question to answer since all isolates of a species are likely to share origin at some time point in history, thus being “related”. However, when put into the context of an outbreak and preferably also in relation with other isolates not connected to the outbreak, at least the relative relatedness can be determined.

If faced with a high number of genomes in a cluster analysis, a two-step analysis can be performed. This means that all genomes are included in the first comparative analysis to determine possible clusters. The second step is to re-analyse the genomes identified in, or close to, the individual clusters. This makes the result images easier to view and the resolution is often increased since assembly-errors in cgMLST/wgMLST analysis increase with the number of genomes analysed. Also, when running a wgMLST or SNP analysis, the shared genome will be larger when only closely related genomes are analysed, thus increasing the resolution.

The method used to calculate the number of allelic differences among the strains should also be carefully considered. For example, in practice, not all loci of a cgMLST scheme are called for every analysed strain. Therefore, the percentage of missing loci should be taken into account when interpreting results and as a key quality parameter. The user needs to consider if a locus missing only in some strains should be maintained in the analysis or not. A pairwise comparison considering all the loci shared between each pair of strains would allow obtaining more detailed information, but it is not the default option for some of the tools used to compare the allelic tables. This step, as well as all the rest of the sequencing and the use of analytical pipelines, should be evaluated by each laboratory using different procedures through benchmarking exercises.

The number of allele differences or SNPs that can be expected in an outbreak situation is dependent on the evolutionary processes that govern the bacterial populations in question, so it is crucial that pathogen-specific knowledge is acquired before a correct interpretation of a real outbreak dataset is performed. There are attempts to create guidelines for what constitutes relatedness between genomes and a summary of some of them can be found in Schürch et al. [29]. In the paper by Schürch et al., the relatedness thresholds or cluster cut-off values are suggested to be as low as ≤ 2 SNPs for *Francisella tularensis* and ≤ 15 SNPs for *C. jejuni,* which illustrates the species-specific differences. For some species, including for example *E. coli*, different levels of variation can be observed for different serotypes, which should be considered when choosing the cluster cut-off values. The EFSA-ECDC One Health WGS System uses two levels of thresholds (core and extended) for cluster detection, specifically designated for each pathogen, where the highest extended threshold is ≤ 10 AD. More care needs to be taken to use thresholds in an outbreak investigation. As an example, a retrospective analysis was performed on *L. monocytogenes* strains from nine different outbreaks. There was a maximum of 21 SNPs difference between isolates in one outbreak, but the majority of outbreaks had a maximum pairwise distance of ≤ 10 SNPs [30].

Instead of, or in combination with, counting SNPs or allelic differences between genome sequences, the creation of phylogenetic trees may provide a more robust interpretation of evolutionary relationships. The

framework for interpreting WGS analyses used by the Food and Drug Administration's Center for Food Safety and Applied Nutrition (CFSAN) combines SNP counts with phylogenetic tree topologies and bootstrap support. In this framework there is strong support for a match when there are 20 or fewer SNPs and the phylogenetic analysis shows a monophyletic relationship with bootstrap support of 0.90 or higher [31].

Phylogeny-independent solutions based on statistical tests have also been used to separate between strains connected to outbreaks or not [32]. Further, it is wise to keep in mind that there will likely be a genetic variation also within the population of isolates causing a single outbreak. If possible, it is advisable to sequence multiple isolates from the potential source of an outbreak (e.g. a suspected food item) to capture this variability.

It is crucial to keep in mind that the interpretation of clustering data cannot only rely on cut-off values or phylogenetic trees; epidemiology and traceback evidence are also needed to link isolates to each other and even more strikingly when a causative link has to be established between a case or an outbreak and the suspected source of infection. The epidemiological context becomes a major point to be considered given the large variability observed in almost all the various steps composing all the bioinformatic workflows aiming at producing strains signatures, regardless of whether these are allele or SNPs-based. As described in this document, each and every passage is in fact subjected to a number of parameters to be fine-tuned depending on e.g. the depth and quality of sequencing and variations in the final result can be introduced at any of these steps, making the assignment of a 100% reliable causative link not possible when only the cluster analysis data are considered.

8. References

1. European Union. *EURL-Lex. Document Ares(2024)5951694*. 2024; Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=pi_com%3AAres%282024%295951694.
2. Inter Biorisks-EURLs WG on NGS. *Zenodo community Inter Biorisks-EURLs WG on NGS*. Available from: https://zenodo.org/communities/eurls-biorisks_wg_on_ngs.
3. Lees, J.A., et al., *Fast and flexible bacterial genomic epidemiology with PopPUNK*. Genome Res, 2019. **29**(2): p. 304-316.
4. Harris, S.R., *SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology*. bioRxiv, 2018.
5. Sheppard, S.K., K.A. Jolley, and M.C. Maiden, *A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter*. Genes (Basel), 2012. **3**(2): p. 261-77.
6. Deneke, C., et al., *Decentralized Investigation of Bacterial Outbreaks Based on Hashed cgMLST*. Front Microbiol, 2021. **12**: p. 649517.
7. EFSA., et al., *Guidelines for reporting Whole Genome Sequencing-based typing data through the EFSA One Health WGS System*. EFSA Supporting Publications, 2022.
8. Joseph, L.A., et al., *Evaluation of core genome and whole genome multilocus sequence typing schemes for Campylobacter jejuni and Campylobacter coli outbreak detection in the USA*. Microb Genom, 2023. **9**(5).
9. Leeper, M.M., et al., *Evaluation of whole and core genome multilocus sequence typing allele schemes for Salmonella enterica outbreak detection in a national surveillance network, PulseNet USA*. Front Microbiol, 2023. **14**: p. 1254777.
10. Henri, C., et al., *An Assessment of Different Genomic Approaches for Inferring Phylogeny of Listeria monocytogenes*. Front Microbiol, 2017. **8**: p. 2351.
11. Leekitcharoenphon, P., et al., *Comparative genomics of quinolone-resistant and susceptible Campylobacter jejuni of poultry origin from major poultry producing European countries (GENCAMP)*, in *EFSA Supporting Publications*. 2018, Technical University of Denmark - National Food Institute
12. Rumore, J., et al., *Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157:H7 from the Canadian perspective*. BMC Genomics, 2018. **19**(1): p. 870.
13. Coipan, C.E., et al., *Concordance of SNP- and allele-based typing workflows in the context of a large-scale international Salmonella Enteritidis outbreak investigation*. Microb Genom, 2020. **6**(3).
14. Pearce, M.E., et al., *Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak*. Int J Food Microbiol, 2018. **274**: p. 1-11.
15. Alikhan, N.F., et al., *A genomic overview of the population structure of Salmonella*. PLoS Genet, 2018. **14**(4): p. e1007261.
16. Luth, S., et al., *Translatability of WGS typing results can simplify data exchange for surveillance and control of Listeria monocytogenes*. Microb Genom, 2021. **7**(1).
17. Segerman, B., et al., *The efficiency of Nextera XT tagmentation depends on G and C bases in the binding motif leading to uneven coverage in bacterial species with low and neutral GC-content*. Front Microbiol, 2022. **13**: p. 944770.
18. Segerman, B., *The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases*. Front Cell Infect Microbiol, 2020.
19. Gardner, S.N. and B.G. Hall, *When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes*. PLoS One, 2013. **8**(12): p. e81760.
20. Dallman, T., et al., *SnapperDB: a database solution for routine sequencing analysis of bacterial isolates*. Bioinformatics, 2018. **34**(17): p. 3028-3029.
21. Labbé, G., et al., *Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel*. bioRxiv, 2020.
22. Cody, A.J., et al., *Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates*. J Clin Microbiol, 2017. **55**(7): p. 2086-2097.
23. Rossi, M., et al., *INNUENDO whole genome and core genome MLST schemas and datasets for Campylobacter jejuni*. Zenodo, 2018.

24. Rossi, M., et al., *INNUENDO whole genome and core genome MLST schemas and datasets for Escherichia coli*. Zenodo, 2018.
25. Moura, A., et al., *Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes*. Nat Microbiol, 2016. **2**: p. 16185.
26. Leopold, S.R., et al., *Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes*. J Clin Microbiol, 2014. **52**(7): p. 2365-70.
27. Nascimento, M., et al., *PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods*. Bioinformatics, 2017. **33**(1): p. 128-129.
28. Francisco, A.P., et al., *PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods*. BMC Bioinformatics, 2012. **13**: p. 87.
29. Schürch, A.C., et al., *Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches*. Clin Microbiol Infect, 2018. **24**(4): p. 350-354.
30. Møller Nielsen, E., et al., *Closing gaps for performing a risk assessment on Listeria monocytogenes in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis*. 2017: EFSA Supporting Publications.
31. Pightling, A.W., et al., *Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations*. Front Microbiol, 2018. **9**: p. 1482.
32. Radomski, N., et al., *A Simple and Robust Statistical Method to Define Genetic Relatedness of Samples Related to Outbreaks at the Genomic Scale - Application to Retrospective Salmonella Foodborne Outbreak Investigations*. Front Microbiol, 2019. **10**: p. 2413.