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 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 Control
EFSA	European Food Safety Authority
ENGAGE	European project “Establishing next generation sequencing ability for genomics analysis in Europe”
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)
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
MiSeq	A benchtop sequencing instrument from Illumina
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

1. Introduction

The continuous implementation of whole genome sequencing (WGS) by different laboratories in the EU has enabled new approaches for European surveillance and cross-country outbreak investigations. There are many different choices the laboratories are facing with the analysis of WGS data. Some of the choices will affect the end results and other will affect practical aspects of the application of results, for example when data is not comparable or when there are no tools or conformity to communicate the data. 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 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. 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 viruses or parasites.

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 are introduced in the next two sections, 2.1 and 2.2, and chapter 3 describes the main differences between them. Both approaches involve several steps of analysis, each depending on bioinformatic scripts or software, that all can affect the end results. These steps may include e.g., read trimming, assembly, read-mapping, alignment, variant calling, allele calling and dendrogram/tree production. There are both freely available and commercial software solutions that perform these steps. Which tools or software the laboratories choose to use will rely heavily on previous experiences as well as national and financial preferences. Chapter 4 and 5 provide technical information on each approach and list software, including those used by the EURLs and/or the NRLs of the EURL-networks of the WG on NGS, but does not discriminate between the different software. An alternative comparison approach is based on estimation of *k*-mer distances. This is summarised in section 2.3.

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. All deliverables produced by the Inter-EURLs WG on NGS can be reached from the EURL websites.

2.1. The SNP approach

Analysing WGS data by identifying SNPs that vary among isolates is generally regarded as the method with the highest resolution for relatedness studies. SNPs can be very informative markers when analysed correctly. Several solutions exist for identifying SNPs and many so-called “SNP pipelines”, which typically combine standalone bioinformatics tools into a workflow that generates a compilation of SNP differences

and sometimes also include phylogenetic visualisation. For experienced bioinformaticians, it is possible to build customised SNP pipelines. The most common approach is to determine SNPs by comparing WGS data from isolates to a reference genome. However, there are also approaches that do not use a reference genome and procedures that use several reference genomes. 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) and the variants for each of the isolates are then combined into a format that allows an analysis of phylogenetic relatedness. Some approaches can use/require assembled genomes instead of sequence reads as input. There are typically some quality filtering steps, which are very important to avoid calling false SNPs. Lack of a consensus in how to apply these filtering criteria and the multitude of read mappers, aligners, variant callers and tree-producing algorithms make SNP analysis difficult to standardise. Analysing a large dataset with SNP analysis can be computationally intensive and may therefore be time consuming depending on the available computational capacity. A schematic view of the fundamental steps in the SNP approach is presented in Figure 1.

2.2. The gene-by-gene approach

The gene-by-gene approach is basically a multilocus sequence typing (MLST) analysis upscaled to include up to thousands of genes or parts of genes [1]. This extended MLST is often referred to as core genome (cg) MLST (using a conserved core of target genes found in nearly all strains of a species) or whole genome (wg) MLST (using all genes found in the strains used to create the allele database). For the gene-by-gene approach, instead of a reference genome, the user supplies a gene target list, which is usually called the cg/wgMLST-scheme. This is either a list of conserved core genes (cgMLST) or both conserved and accessory genes (wgMLST). The gene-by-gene method usually accepts assembled genomes as input. Analysis is performed by aligning the gene targets (from the cg/wgMLST-scheme) to the assembly and extracting the isolate's allelic sequences. An alternative strategy is to skip the assembly step and identify alleles by mapping reads directly to the target genes. When a new allele sequence has been identified, it receives an integer, which is increased by 1 for each new allele. This is referred to as allele calling and can, together with the assembly process, be time consuming depending on the computational capacity. However, once the allele calling is done, it does not have be performed again on those isolates. Thus, if the user wants to add additional genomes to the analysis at a later stage, allele calling will only be done on the new genomes. The result from a cg/wgMLST run is a table with integers or a dissimilarity matrix, which makes the following cluster analysis computationally trivial. A schematic view of the fundamental steps in the gene-by-gene approach is presented in Figure 1.

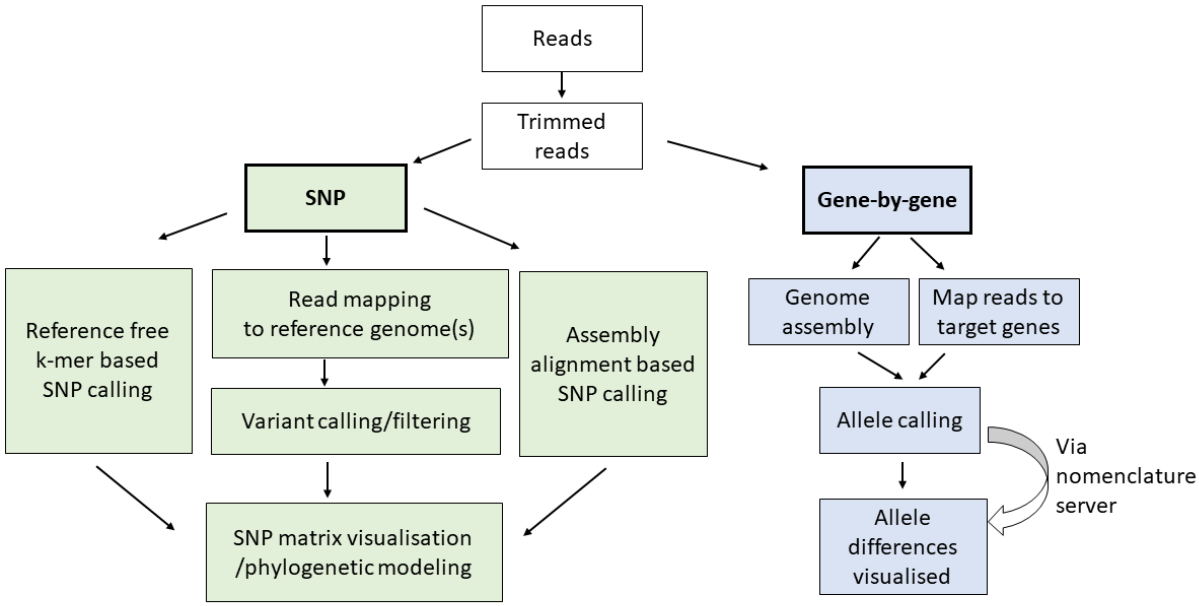

Figure 1. A simplified schematic view of the fundamental steps in cluster analysis, either by the single nucleotide polymorphism (SNP) approach (green) where individual mutations are used as separate phylogenetic markers, or by the gene-by-gene approach (blue) where each variant of a gene is considered an allele.

2.3. *k*-mer approaches

Counting *k*-mers is computationally fast and can be used to identify SNPs or to detect MLST alleles. Dividing sequences into *k*-mers is an approach that also is widely used in genome assembly programs and taxonomic classification programs. The 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. It is also possible to indirectly estimate SNP distances by quantifying and comparing *k*-mer matches at different *k*-mer lengths (since longer *k*-mers are more likely to contain SNPs). This approach is implemented in PopPUNK (Population Partitioning Using Nucleotide *K*-mers) [2]. Since methods relying only on quantifying *k*-mer matches of a defined length generally have a poorer precision than SNP and gene-by-gene methods, this document will not go into further details on these approaches.

2. SNP vs gene-by-gene approach

The PulseNet International network, which includes public health organisations from around the world with respect to food- and waterborne diseases, has recommended wgMLST as the most suitable approach for bacterial food-borne disease surveillance [3]. ECDC and EFSA have agreed that the joint database for disease surveillance in Europe using WGS data should be possible with both SNP and gene-by-gene analysis [4, 5].

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 concordant and comparable to each other [6-12].

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 in sections 3.1 and 3.2.

3.1. Data differences and resolution

The type of data included in the analysis differs between the different clustering methods. A cg/wgMLST scheme will not include intergenic regions as opposed to the SNP analysis. And a gene that contains several mutations will be collapsed to a new allele number and only counted as one change. As an example, a gene containing three SNPs will be counted as one difference in cg/wgMLST but as three differences in a SNP analysis. However, the accumulation of several SNPs in close proximity may have arisen during the same evolutionary event (e.g. recombination) and quantifying them individually without correction for this may overestimate the genetic distance. Furthermore, small INDELs will not be counted by all SNP approaches but will always be counted as a new allele by the cg/wgMLST method.

The resolution of analysis of all clustering methods is directly related to the proportion of data included in the comparison. Reference-based SNP analysis is restricted to what is present in the reference genome, so the closer related the reference is to the studied strains, the higher the resolution of the analysis. Cg/wgMLST analysis is restricted to the alleles present in the schemes, so the higher number of alleles in the scheme, the higher the resolution of the analysis. Further, cgMLST is also restricted to core regions of the analysed genomes. Potentially useful information in the accessory genome is therefore often discarded. In contrast, SNP analysis and the wgMLST approach also includes information in the accessory genome. Generally, the resolution of analysis is also affected by the quality of the data input since both SNPs or allele targets can be discarded if the data does not reach the quality threshold set for the particular analysis (see 4.5 and 5.3). The amount of data needed to achieve the highest resolution possible depends on the quality of the data but also

on the clustering method applied. The 'Guidance document for WGS-benchmarking' describes a data filtering method that can be applied to validate amount of data (coverage) needed for the different approaches.

Further, the different sequencing techniques generating sequence data can produce different types of errors, which is important to consider when choosing downstream methods for analysis. For example, 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 [13]. For this reason, a proper validation is needed when you want to compare Ion Torrent and Illumina data through an assembly-based approach (see 'Guidance document for WGS-benchmarking').

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 [14]. 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 [15]. 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 [16]. 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 cg/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.

3. 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 insertions/deletions (INDELs). There are several "pipelines" publicly available for making 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 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
Snippy	https://github.com/tseemann/snippy
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 (includes e.g. **KSNP3, POPPUNK, FDA SNP PIPELINE)**	https://www.iss.it/site/aries
CSI Phylogeny	https://cge.cbs.dtu.dk/services/CSIPhylogeny/
Enterobase	https://enterobase.warwick.ac.uk/
NDtree	https://cge.cbs.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 removing of adapters (such as with Trimmomatic, http://www.usadellab.org/cms/index.php?page=trimmomatic). The reads are commonly mapped to a reference genome sequence with a mapping program. There are also solutions 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. Some examples of read-mapping software solutions are presented in Table 3.

Table 3. Read-mapping software.

SOFTWARE	LINK TO SOFTWARE
bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA	http://bio-bwa.sourceforge.net/
Maq	http://maq.sourceforge.net/
novoalign	http://www.novocraft.com/products/novoalign/
SMALT	https://www.sanger.ac.uk/science/tools/smalt-0

The most commonly used mappers are bowtie2 and BWA (BWA-mem). 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., kSNP3) 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 a number of 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. Some variant calling software solutions are listed in Table 4. 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.

Table 4. Variant calling software

SOFTWARE	LINK TO SOFTWARE
Freebayes	https://github.com/ekg/freebayes
GATK	https://gatk.broadinstitute.org/hc/en-us
SAMtools	http://www.htslib.org/
SolSNP	https://sourceforge.net/projects/solsnp/
VarScan	http://varscan.sourceforge.net/

4.5. Variant filtering and merging of results

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 (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 that 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 contig ends may be excluded.
- Duplicate regions /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 (may be PCR duplicates, not true unique sequenced fragments).

Finally, the variants identified in each isolate needs 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/).

4. Gene-by-gene analysis methods and software

The first step in gene-by-gene analysis is to define a target list of genes that all sequenced genomes will be compared against. This is often referred to as the cg/wgMLST-scheme. Generally, a scheme with a higher number of alleles gives a higher resolution of the analysis. If a conserved core genome MLST scheme is to be created, the gene targets included must be present in all, or at least close to all, genomes of a species. 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. The other alternative is to also include genes that are part of the accessory genome. These genes will not be present in all sequenced genomes to be analysed but can elevate the resolution of the MLST-approach to levels similar of a SNP-analysis. Since a wgMLST-scheme can provide a higher resolution than cgMLST, it can be useful for outbreak tracking and similar studies. Chapter 5 describes the different steps of analysis included in the gene-by-gene analysis and lists commonly used pipelines and software.

5.1. cg/wgMLST-schemes

The benefit of using publicly available databases with stable online schemes is the possibility to compare isolates to a high number of other deposited genomes or allele profiles. This is a prerequisite for a continuous surveillance of pathogens and detection of cross-country outbreaks. Table 5 lists such databases and schemes available for a number of 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]

Furthermore, there are commercial software packages that have implemented their own wgMLST and cgMLST schemes. If there is no available scheme for a certain pathogen it is possible to create an *ad hoc* scheme. For example, the commercial software SeqSphere+ (Ridom) has this function and it is called 'Core Genome MLST Target Definer'. It accepts any reference genome as seed for the scheme and then performs tests to evaluate the MLST-suitability of the genes. Genes that are found in all query genomes, also added to the target definer, will constitute the cgMLST targets and genes that are not present in all genomes will make

up the accessory targets. The accessory targets can also be added to the analysis if a higher resolution is needed. An example of a an open-source and free software package is chewBBACA, in which one can create whole-genome or core-genome gene-by-gene typing schemes and perform the allele calling from assembled genomes [22].

5.2. Assembly

It is possible to map sequence reads directly to the gene targets using a short reads aligner like KMA or in software such as SeqSphere+ and Mentalist. However, the most commonly used data input for gene-by-gene analysis is in the format of genome assemblies. It is recommended that reads are trimmed based on quality before assembly. Examples of trimming software are Trimmomatic [23], Sickle (https://github.com/najoshi/sickle) and Trim Galore (https://github.com/FelixKrueger/TrimGalore). There are several published assembly programs suitable for bacterial genomes that are freely available for use. Three popular assemblers for Illumina data are SKESA [24], Velvet [25] and SPAdes [26], the latter which today is the most frequently used assembly software in RefSeq [13]. Of the laboratories participating in the WG-WGS survey for bioinformatics tools conducted in the EURL networks for AMR, *Campylobacter*, *E. coli, Salmonella*, in 2018, the majority of the NRLs used SPAdes for genome assembly, followed by Velvet (see the summery 'Bioinformatics tools for basic analysis of Next Generation Sequencing data' produced by the Inter-EURLs WG on NGS). SKESA is a fast assembler, which in comparison with SPAdes has been shown to produce assemblies of higher quality [24]. However, it should be noted that the higher assembly contiguity of SKESA has the drawback of producing less complete genes, which may be a disadvantage for gene-by-gene approaches [27]. Benchmarking of SPAdes 3.9 and Velvet 1.2, was performed in the ENGAGE project [28]. The results showed that SPAdes generated longer contigs than Velvet and the accuracy of predicting the correct MLST and serovar in *Salmonella* genomes was higher using SPAdes (100%) in comparison to using Velvet (94%).

For quality control of the assembly, metrics such as assembly length, GC-content, N50, and number of contigs can be used. A poor assembly will often have a negative impact of the result in downstream analysis. There are assembly correcting software (e.g., Pilon [29]) that by mapping reads back to contigs can correct the assembly from errors created in the assembly process. The user should be aware that Pilon sometimes extends the length of the contigs and includes some ambiguous nucleotides (i.e., "N") in the end of the sequences, which can have a negative impact on allele calling. The assembly can also be improved by using gap closing and scaffolding software steps. The tools for assembly correction need to be properly benchmarked in each laboratory.

Shovill (https://github.com/tseemann/shovill) is a pipeline which uses the preferred assembler at its core (supports SKESA, Velvet and SPAdes), but alters the steps before and after the primary assembly step to get similar results in less time. This pipeline for example trims adapters, correct sequencing errors, pre-overlap paired-end reads and correct assembly errors and remove contigs that are too short or have too low coverage.

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 commercial software with graphical user interfaces that can do most, if not all, of these steps including cg/wgMLST analysis (e.g., BioNumerics and SeqSphere+).

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

5.3. Allele calling

The allele calling step often utilises an alignment tool such as Basic Local Alignment Search Tool (BLAST) (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) that returns the allele sequences of the genome analysed. If the user is working with an online allele database or in a tool connected to such a database, the alleles will receive their respective allele identifiers. If the allele sequence is a novel one, a new identifier will be assigned, which will then be deposited at the online database. When using a local approach (i.e., not working towards an allele identifier server) the alleles will be designated allele identifier integers, starting from 1 and counting upwards for each new allele. This process can be computationally heavy but once completed, the following cluster analysis is computationally trivial since the allele identifiers is a simple list of integers. The allele differences can be visualised in a minimum spanning tree (MST), which shows the number of allele differences between the isolates in the analysis. See chapter 6 for examples of MSTs and how to interpret them.

There are also free online services that perform the cg/wgMLST-analysis. Disadvantages of this approach include dependency on the service provider, downtimes of server, long waiting times (days sometimes) and a lack of control of the analysis. Online services may also have a disclaimer for ownership of the data, which can be considered a disadvantage. Therefore, to be able to respond to an outbreak or to have a consistent surveillance, bioinformatic analysis is preferably also locally available. Online servers include PubMLST (https://pubmlst.org/), Enterobase (https://enterobase.warwick.ac.uk) and the cgMLSTFinder (https://cge.cbs.dtu.dk/services/).

For local operation (and connection to online databases in some cases) there are both commercial and free software available. A selection of available software solutions for cg/wgMLST are listed in Table 6.

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 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*, *Salmonella*, *Vibrio*, and *Yersinia*. cgMLSTFinder covers: *Campylobacter* (using the PubMLST scheme), *Clostridioides* (Enterobase scheme), *Escherichia coli* (Enterobase scheme), *Listeria 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 Institute 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.

The called alleles are presented in a results table. Failed allele calling can be due to a missing target or target of the wrong length, both which can be effects of assembly or sequencing errors or true differences between isolates.

5. Visualisation of clustering data

The number of SNPs or allele differences can be directly derived from a table and converted into a distance matrix describing the pairwise distances (Table 7), or the results can be visualised in for example a minimum spanning tree (MST) (Figure 2).

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. In the example given in Table 7, the results of a cgMLST analysis gave a total of 58 allelic differences between STRAIN1 and STRAIN2, 1211 allelic differences between STRAIN1 and STRAIN3, and 5 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 2, isolates A and C are separated by 9 allelic differences, which means that out of the 1,340 genes investigated in this analysis, only 9 genes showed differing 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.

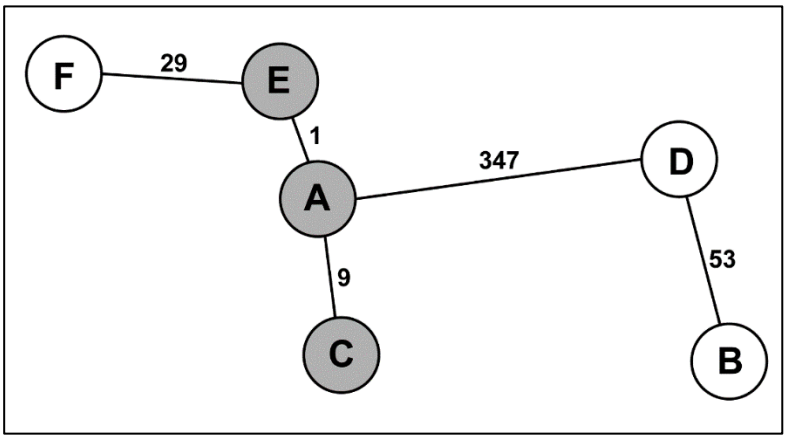

Figure 2. A cgMLST result for six genomes visualised in a minimum spanning tree. The numbers between the sample names represent the number of allelic differences between the samples. The line lengths are not proportional to the number of differences. The total number of gene targets compared in this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

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 wg/cgMLST) is the tool PHYLOViZ [30, 31]. 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 8.

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/

6. 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 cg/wgMLST 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 elevating 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 the loci of a cgMLST scheme are called for all the analysed strains. 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 a pathogen-specific knowledge is acquired before a correct interpretation of a real outbreak dataset is performed. There are ongoing attempts to create guidelines for what constitutes relatedness between genomes and a summary of some of them can be found in Schürch et al. [32]. 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 *Campylobacter jejuni,* which illustrates the species-specific differences. These thresholds will be more reliable as more and more confirmed outbreaks are investigated, but although desirable, it is not likely that we will be able to use a fixed threshold for cluster definition for each pathogen. As an example, a retrospective analysis was performed on *Listeria 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 pair-wise distance of ≤ 10 SNPs [33].

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 [34].

Phylogeny-independent solutions based on statistical tests have also been used to separate between strains connected to outbreaks or not [35]. 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 is 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.

7. References

1. 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.
2. Lees, J.A., et al., *Fast and flexible bacterial genomic epidemiology with PopPUNK*. Genome Res, 2019. **29**(2): p. 304-316.
3. Nadon, C., et al., *PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance*. Euro Surveill, 2017. **22**(23).
4. ECDC and EFSA, *EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human, animal, food, feed and food/feed environmental samples in the joint ECDC–EFSA molecular typing database*. 2019: EFSA supporting publication.
5. ECDC, *ECDC strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigations*. 2019.
6. Henri, C., et al., *An Assessment of Different Genomic Approaches for Inferring Phylogeny of Listeria monocytogenes*. Front Microbiol, 2017. **8**: p. 2351.
7. 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 publication*. 2018, Technical University of Denmark - National Food Institute
8. 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.
9. 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).
10. 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.
11. Alikhan, N.F., et al., *A genomic overview of the population structure of Salmonella*. PLoS Genet, 2018. **14**(4): p. e1007261.
12. 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).
13. 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.
14. 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.
15. Dallman, T., et al., *SnapperDB: a database solution for routine sequencing analysis of bacterial isolates*. Bioinformatics, 2018. **34**(17): p. 3028-3029.
16. Labbé, G., et al., *Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel*. bioRxiv, 2020.
17. 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.
18. Rossi, M., et al., *INNUENDO whole genome and core genome MLST schemas and datasets for Campylobacter jejuni*. Zenodo, 2018.
19. Rossi, M., et al., *INNUENDO whole genome and core genome MLST schemas and datasets for Escherichia coli*. Zenodo, 2018.
20. Moura, A., et al., *Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes*. Nat Microbiol, 2016. **2**: p. 16185.
21. 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.
22. Silva, M., et al., *chewBBACA: A complete suite for gene-by-gene schema creation and strain identification*. Microb Genom, 2018. **4**(3).

23. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. **30**(15): p. 2114-20.
24. Souvorov, A., R. Agarwala, and D.J. Lipman, *SKESA: strategic k-mer extension for scrupulous assemblies*. Genome Biol, 2018. **19**(1): p. 153.
25. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs*. Genome Res, 2008. **18**(5): p. 821-9.
26. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing*. J Comput Biol, 2012. **19**(5): p. 455-77.
27. *Center for Algorithmic Biotechnology: http://cab.spbu.ru/benchmarking-tools-for-de-novo-microbial-assembly/*. St. Petersburg State University.
28. Hendriksen, R.S., et al., *Final report of ENGAGE - establishing next generation sequencing ability for genomic analysis in Europe*, in *EFSA supporting publication* 2018.
29. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement*. PLoS One, 2014. **9**(11): p. e112963.
30. 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.
31. Francisco, A.P., et al., *PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods*. BMC Bioinformatics, 2012. **13**: p. 87.
32. Schurch, 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.
33. 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 Publication.
34. 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.
35. 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.