Proposed protocol for whole genome sequencing-based analysis for detection and tracing of epidemic clones of antimicrobial resistant *Salmonella* and *Campylobacter*

- to be used for national surveillance and integrated outbreak investigations by NRLs for public health

8 July 2022

FWD AMR· RefLabCap

TABLE OF CONTENTS

ABBREVIATIONS

AD	Allelic differences
AMR	Antimicrobial Resistance
ARG	Antimicrobial resistance gene
AST	Antimicrobial Susceptibility Testing
cgMLST	Core genome multi-locus sequence typing
ECDC	European Centre of Disease Prevention and Control
EEA	European Economic Area
EQA	External Quality Assessment
EU	European Union
FWD AMR-RefLabCap	Food- and Waterborne Diseases Antimicrobial Resistance – Reference Laboratory Capacity
NRL	National Public Health Reference Laboratory
MLST	Multi-Locus Sequence Typing
PM	Point mutation
QC	Quality control
SNP	Single nucleotide polymorphisms
WGS	Whole Genome Sequencing

INTRODUCTION

The purpose of the Food- and Waterborne Diseases Antimicrobial Resistance – Reference Laboratory Capacity (FWD AMR-RefLabCap) project is to strengthen coordination, support and capacity building in national microbiology reference laboratory functions for testing and surveillance of antimicrobial resistance (AMR) in *Salmonella* and *Campylobacter* in human samples. The project aims at strengthening capacities for national surveillance and outbreak investigation of *Salmonella* and *Campylobacter* and improve the availability and quality of European-level molecular surveillance data. One of the main goals of the FWD AMR-RefLabCap project is to support implementation and alignment of whole genome sequencing (WGS)-based analysis methods in order to achieve those respective aims.

The present document suggests a standard protocol for AMR gene detection and its integration into routine WGS-based surveillance for use by the NRLs in public health participating in the FWD AMR-RefLabCap project.

This protocol describes how to perform WGS-based analysis of *Salmonella* and *Campylobacter*. The protocol covers the steps of obtaining high-quality DNA, performing library preparation and sequencing of the DNA and performing bioinformatics analysis (taxonomic analysis, bacterial typing, detection of genetic AMR determinants, cluster analysis). The protocol suggests specific quality control (QC) strategies, QC parameters and recommended thresholds and gives recommendations on bioinformatics tools and reference gene databases for the detection and prediction of AMR determinants in *Salmonella* and *Campylobacter.*

The recommendations are based on a review of existing national and international guidance on WGS-based monitoring of AMR, including available bioinformatics tools, bioinformatics development initiatives, and reference gene databases for the detection and prediction of AMR determinants in *Salmonella* and *Campylobacter*.

This protocol is directed towards short-read paired-end massive parallel synthesis sequencing, specifically using Illumina platforms (Illumina, Inc., San Diego, California, United States) such as MiSeq and NextSeq.

The FWD AMR-RefLabCap does not endorse nor is endorsed by any of the companies, brands or products referred in this document.

PROTOCOL

This protocol describes the different steps necessary to perform WGS of *Salmonella* and *Campylobacter* isolates. For each step different methods, kits or tools exist, but the choice should be carefully considered and take into account the laboratory's available methods, equipment and protocols. Different approaches can be employed as long as they are properly validated for the repeatability, the reproducibility, and the accuracy. The parameters and the acceptance criteria for validation and verification of WGS workflow(s) will be described in ISO 23418 (under development).

1.1. History of changes

Version	Sections changed	Description of change	Date	Approval
1.0	N/A	First version of protocol	08.07.2022	Eva Litrup

1.2. DNA Extraction and Quality Check

Only pure bacterial colonies should be employed to avoid contamination, and streaking out fresh culture from a single colony should be implemented as a routine. Carefully inspect all colonies for contaminations prior to DNA extraction and then collect a streak of colonies. DNA extraction can typically be performed by the routine method available in the laboratory, and a wide range of DNA extraction kits are applicable for WGS, e.g. ThermoFisher Easy-DNA gDNA Purification Kit or Qiagen DNeasy Blood & Tissue Kit. Be aware that extraction methods based on salt and ethanol precipitation can result in poor plasmid extraction, which can be problematic for determination of AMR genes, as these often reside on plasmids. A range of instruments exist for more automated high-throughput DNA extraction, one example being the *MagNa Pure 96* instrument. It is relevant to evaluate the quality of the DNA obtained, which can for instance be conducted using Nanodrop. If the DNA quality is inadequate (or if the QC performed on the WGS data reveals problems later on), it may be relevant to re-extract DNA.

1.3. DNA Concentration and Dilution

The DNA quantity should be assessed, since a specific amount of DNA needs to enter the library preparation step (for instance a final concentration of 0.2 ng/µl if using the Nextera XT Library Preparation Reference Guide with input of 1 ng in 5 µl of each library). The concentration of DNA can be determined through different methods, e.g. using the Qubit fluorometer and the dsDNA reagent kits, or secondarily the Nanodrop spectrophotometer (not recommended for DNA concentrations <2 ng/µl). The initial DNA concentration should be measured and subsequently diluted with the adequate buffer to achieve a final DNA concentration in accordance to the library preparation protocol. The concentration of this diluted sample should further be confirmed by an additional measurement, for instance using the Qubit™ dsDNA High Sensitivity Assay Kit. If the initial DNA concentration is too low it will be necessary to re-extract DNA from the sample or concentrate the DNA preparation.

1.4. Library Preparation and DNA Sequencing

Various sequencing platforms exist and the method for library preparation depends on the chosen platform and selected chemistry. Currently, Illumina is the most widely used sequencing platform, and protocols with preparation guidelines for specific library kits and guidelines for sequencing on the specific sequencing platform are frequently updated and available on the Illumina website. Two commonly used library preparation kits are the Nextera XT DNA Library Preparation kit and the Illumina DNA Prep kit. Similarly, reference

guides and accessory documents exist for other sequencing platforms: e. g. decision tree for ION PGM users and a list of DNA library preparation kits for any Ion Torrent system in use. To the users of Nanopore technology we recommend to register to Nanopore Community where one can access online courses, plan the experiments and download relevant software and protocols. Furthermore, the companies (e.g. Illumina and Thermo Fisher Scientific) typically offer free guidance and training for procured instruments, equipment and associated protocols.

1.5. WGS Raw Data and QC

Following successful completion of sequencing, the raw data (i.e. the fastq files) can be retrieved either from a cloud solution, such as the Illumina sequence hub, or directly from the sequencing machine. Fastq files include sequence reads coupled with quality parameters (phred scores). The cloud solution also offer a range of visual QC parameters to evaluate the sequencing run.

The quality of the sequences should always be assessed, as poor quality sequences can lead to major errors in prediction of AMR genes and phylogenetic analysis.

Institutions having a high WGS throughput often have integrated sequence quality assessment into their bioinformatics workflows processing the data. The QC can, however, also be evaluated in less throughput systems, for instance using a program such as FastQC. Regardless of the QC system, parameters such as the *average read length*, *number of reads*, and *coverage* should as a minimum be determined. These QC parameters will reveal if there is sufficient amount of sequencing data.

- **Average read length:** Should correspond to that expected from the sequencing platform and kit (e.g. Illumina 150bp).
- **Number of reads:** Should be as high as possible. No assessed cut-off exists, but enough to obtain the desired coverage of the organism genome.
- **Coverage:** = Number of reads x (Average read length/Genome size). The coverage should as a minimum be 30x, and preferably even higher (many institutions use 50x as cut-off) (Ellington *et al.* 2017).

Raw data (fastq) should be examined for potential contaminations. For instance, contamination with a different species can be detected employing a version of a classification tool such as KRAKEN (https://ccb.jhu.edu/software/kraken/). Here the number of reads classified to other species than the target can be assessed for QC (e.g., if more than 5% reads match another species the sequence could be contaminated). Contamination checks can also be facilitated by tools such as KmerFinder or alternatively rMLST.

Only sequence data that pass the agreed QC parameters should be used for further analysis. If a proper quality is not obtained, the DNA should be re-sequenced (if the amount of data is too low), or re-extracted and then sequenced (if contaminated).

In some cases it may be relevant to perform trimming of the reads for adaptors and low-quality regions, which can be facilitated with tools such as Bbtools or Trimmomatic. If the sequence quality is OK it is generally not a problem to skip the trimming step for data obtained from platforms, e.g. Illumina, where adaptor trimming is integrated and the sequence quality is high. However, for other platforms, e.g. Ion Torrent, it may be essential in order to obtain reads of high enough quality for further data processing.

1.6. Genome Assembly and QC

Assembly of reads into contigs can be performed using numerous tools. Depending on the sequencing platform in use, the objective of sequencing, the properties of the genomes and a number of other parameters an assembler can be chosen as described by Dominguez Del Angel (2018). Most assembly programs can be installed locally, and many institutions performing WGS routinely have this step incorporated into their bioinformatics workflow. Among a number of assembly softwares, SPAdes is the most commonly used assembler since 2016. Other frequently used assemblers were listed and described by Segerman (2020).

Assembly programs output is a fasta file containing the assembled sequences (contigs), and further output QC parameters that can be evaluated:

- **Size of assembled genome:** (= number of nucleotides in the contig file) Should be within the range for the targeted organism. *Salmonella*: 4.4 Mb-5.8 Mb, *Campylobacter*: 1.5 Mb -1.9 Mb. A larger genome size than expected can indicate that the sample was contaminated (also for contaminations with the same species), while a smaller can be due to an insufficient amount of sequence data.
- **N50**: Indicates the size of the contigs in general. The higher N50, the better as it indicates longer contigs. There is no general cut-off, but >30,000 bp has been suggested (Bortolaia *et al*. 2020).
- **Total number of contigs**: We recommend less than 500 (excluding contigs <300 bp), depending on methods etc. (Bortolaia *et al*. 2020). *Campylobacter* will typically be assembled into less than 100 contigs and *Salmonella* to less than 300 contigs. A higher number than normal may point to poor sequence quality or to contamination (also with same species, which is not always detectable from species prediction on raw data).

If the QC thresholds are not achieved consider resequencing the isolate.

1.7. Typing of Bacterial Isolates

Further bioinformatics analysis can be carried out for sub-typing of isolates.

Species

Species detection for *Campylobacter* can be performed using the same classification tools as employed for contamination checks, e.g. KRAKEN or KmerFinder. Here, almost all reads should be classified to the target species (some hits may overlap with closely related species). Subspecies detection on *Salmonella* can be performed using the same tools.

MLST

Multi-locus sequence typing (MLST) profiles can readily be extracted, typically from assembled data, based on the MLST schemes available and curated for *Campylobacter jejuni* (*C. jejuni*)/ *Campylobacter coli* (*C. coli*) at PubMLST and for *Salmonella* spp. at Enterobase.

MLST types can also be predicted online by employing e.g. the Center for Genomic Epidemiology (CGE) MLST tool (MLST 2.0 (dtu.dk)). In order to perform MLST typing, it is necessary to know the target species to select the proper scheme. An MLST type can only be assigned if perfect allelic matches are present at all (seven) loci, while new alleles and profiles need to be submitted to PubMLST or Enterobase.

In case of many mismatches on the MLST loci, or failure to detect a specific loci, it should be considered if contamination is the reason.

Serotyping

Serotype prediction can be performed for some species, for instance for *Salmonella* spp. using SeqSero2, while WGS-based serotyping of *Campylobacter* is not established.

Prediction of resistance traits

Several bioinformatic solutions exist for prediction of acquired genes/proteins and/or point mutations (PMs) relevant for AMR, e.g. ResFinder, CARD-RGI or AMRFinderPlus. Only sequences that have passed the previous quality steps should be further analysed (Hendriksen *et al.* 2019). As with any other databases, it is important to take note of the version being used, and to make sure to follow any updates and implement them into existing workflows. The databases of the above mentioned bioinformatic solutions are often integrated into different software, including commercially available softwares e.g. AMRFinderPlus in SeqSphere. It is important to be aware of how often the databases are updated in the different software, as updates of the above mentioned bioinformatic solutions may not be integrated or integrated with delay in the different software.

When submitting sequences for analysis, default parameters (% identity and length) for the hits are provided by the tools, and they are typically a good guideline, while they can be adjusted if needed. ResFinder, for instance, reports all hits covering at least 60% of the gene length with 90% identity or above.

Careful evaluation of the results always needs to be performed. In AMRFinderPlus, the prediction is based on hierarchical clustering of protein sequences, and additionally the tool reports if non-functional proteins (stop codons) are detected. For tools making prediction on nucleotide level, hits that are not identical (100% length and 100% identity) to the database content may pose problems, perhaps producing non-functional proteins (if this is not reported), or it could just be silent mutations resulting in the same phenotype. Thus all hits with less than 100% identification should always be translated to amino acid sequence and evaluated further, e.g. check if there is a premature stopcodon in your protein and check if the correct variant is reported (in case of equally good matches).

The detection of less than full-length genes should also be further evaluated, for instance using a mapping-based method as part of a missing gene could be caused by an assembly issue. Also, it should be noted that for all these tools that rely on databases: hits are only produced if the target of interest is included in the database, and therefore it is crucial to be aware of the database content. However, the AMR databases of the different tools are in some cases overlapping, including and updating the same genes and variants.
The appendices contains tables of significant AMR determinants detected in *Salmonella* and *Campylobacter*. This data is a snapshot of the current situation, as AMR is a moving target and genes are exchanged between organisms continuously.

1.8. Cluster Analysis

Cluster analysis is typically performed to distinguish very closely related (outbreak) isolates from sporadic cases. Cluster analysis should be performed on sequences that have passed the prior QC parameters; sequences that did not meet the QC parameters could disturb downstream analysis and results including these should be interpreted with extreme care.

In general, two overall approaches exist. Single nucleotide polymorphism (SNP)-based analysis where raw reads are typically the input data, and core genome multi-locus sequence typing (cgMLST)-based analysis which is often based on assemblies.

SNP-based cluster analysis

SNP analysis is a reference-based method where raw sequence data from multiple isolates are mapped to a reference genome to evaluate the SNP differences. Several tools exist for this type of analysis, some of which can be installed and run locally such as Snippy and FastTree, but SNP analysis can also be performed online using for instance CSIPhylogeny.

It is important to choose a reference that resembles the isolates included in the analysis, for instance one of the same type (MLST), or even a de novo assembled genome from the

dataset being analysed. If the query genomes are too different from the reference it will reduce the percentage of genome content to be included in the analysis and thus decrease resolution. A good rule of thumb is that at least 90% of the reference genome should be covered by each query genome.

cgMLST-based cluster analysis

cgMLST analysis is another approach for establishing phylogenetic relationships between isolates. For *Campylobacter* (Cody *et al.* 2017) and *Salmonella* (Achtman *et al.* 2021) established and curated schemes exist based on 1343 loci and 3002 loci, respectively. cgMLST can be performed in commercial solutions such as BioNumerics (Whole genome multi-locus sequence typing (wgMLST) | BIONUMERICS) or SeqSphere (Ridom SeqSphere+ - Overview). Non-commercial solutions such as chewBBACA or MentaLiST also exist. *Salmonella* cgMLST can furthermore be performed online through Enterobase, however, with the uploaded genomes made public after an agreed release period. All the mentioned softwares call alleles in different ways and it should be noted that even if users employ the same cgMLST scheme there will be minor variations in the alleles detected and reported.

When performing cgMLST analysis, the percentage of the core genome found in each query sequence can be evaluated as a final QC step, ideally aiming for >95% core (coverage of the scheme loci). However, for *Campylobacter* the scheme is built for *C. jejuni* meaning that *C. coli* will most often produce lower core percentage (around 80%).

Identifying genetic clusters

It is essential to evaluate that the identified genetic clusters are in accordance to the epidemiological data of time, place and person to assure proper outbreak definitions.

There are no well established cut-offs for interpretation of the allelic differences (AD) or SNP distances between isolates, and it can not be recommended defining outbreak clusters only based on such a set cut-off as the genetic diversity of epidemiologically related isolates can vary dependent on the nature of the outbreak.

The following cut-offs are guiding suggestions. For *Campylobacter* isolates, 5 or less ADs/SNPs can be considered as genetic clusters (Brehony et al. 2021, Joensen et al. 2021), while for *Salmonella* this depends on the serovar, with 2-3 ADs/SNPs in clonal serovars and up to 5 AD/SNP in other serovars (Payne et al. 2021, Gymoese et al. 2017). It is important to consider that the amount of SNPs between specific isolates can change depending on the dataset being analysed, e.g. inclusion of more distant isolates will decrease the resolution of the SNP-analysis. This is not an issue for allele-based analyses.

1.9. Data storage and sharing

It is recommended to store raw sequence data or assemblies for a period of at least five years. It requires extensive server space to store raw sequence data locally, so uploading to public repositories is an advantage and data can be pseudonymised before upload. Storing only assemblies has drawbacks as possible novel analysis cannot always be carried out on assemblies.

Raw and/or assembled data and certain metadata can be shared in public repositories if allowed by national data protection legislation. Reporting of WGS data and metadata in a closed and confidential setting at European level are relevant in outbreak situations and available in European surveillance portal for infectious diseases (EpiPulse) at European Centre of Disease Prevention and Control (ECDC). To share WGS data with ECDC for European Union/European Economic Area (EU/EEA) level surveillance and within outbreak

situations is recommended by ECDC. Furthermore, sharing WGS data is of great importance to the collaboration and investigation of cross-border outbreaks.

SURVEILLANCE OF AMR AND OUTBREAK DETECTION

Investigation of WGS data for AMR in *Campylobacter* and *Salmonella* in near real-time, together with epidemiological data, is vital for detecting the emergence of high-risk clones, monitoring of time and spatial trends, detection and investigation of outbreaks, sources of transmission and prevention and control measures. Routine WGS-based surveillance provides a cornerstone in both local, regional and national epidemic preparedness, enabling typing of sporadic and outbreak-related *Salmonella* and *Campylobacter* isolates and identification and monitoring of resistant epidemic clones.

In order to establish a WGS-based AMR surveillance, laboratories should agree on the isolate/sample testing frequency for the surveillance of AMR and the highest possible number of isolates should be submitted and further characterised by WGS. To ensure detection of rare and/or emerging resistant *Salmonella* and *Campylobacter* clones, special *Salmonella* serovars and *Campylobacter* species/types with certain AMR phenotype characteristics, and strains causing severe disease outcomes should be considered for WGS. Also, a representative proportion of human samples/isolates associated with a suspected outbreak should be further characterised by WGS ensuring detailed knowledge of the outbreak strain.

WGS-based surveillance of AMR includes steps for detection of genetic determinants of antimicrobial resistance, mainly acquired antimicrobial resistance genes (ARGs) and chromosomal PMs in specific target genes. Either of these mechanisms can lead to decreased susceptibility towards antimicrobials of relevance in public health settings.

In addition to the investigation of genetic AMR mechanisms, selected isolates from a defined site (such as a country, a region, or a hospital or healthcare facility) can be further analysed by WGS to determine the genetic relatedness between isolates. This requires the use of a suite of genomic typing tools, including but not limited to MLST, cgMLST and SNP-based analysis. These bacterial typing and cluster analysis strategies are able to detect genetic clusters and support epidemiological analysis aimed at monitoring the introduction and expansion of high-risk multidrug-resistant clones, transmission events and detection of outbreaks.

The WGS analysis and AMR detection should be designed to meet the identified quality parameters and recommendations in this protocol. Sequencing and bioinformatics approaches that produce high-quality results of AMR detection and cluster analysis are of great significance for the final result. Thus, to ensure quality of data and comparability of WGS results among laboratories, participation in External Quality Assessments (EQAs) is essential. ECDC is offering a phenotypic AMR EQA (AST-EQA) and the FWD AMR-RefLabCap is offering WGS-based AMR EQAs for EU and EEA countries. The strains received in these EQAs would serve as excellent reference material when setting up WGS-based AMR detection. WGS data for surveillance and outbreak detection should always be accompanied by clinical epidemiological data, species characteristics and nature of the outbreak.

Finally, by uploading sequence data with associated metadata to online databases, investigations can be extended to assess multinational outbreak events.

APPENDIX 1 – AMR genes and point mutations in *Salmonella*

Table 1. Significant antimicrobial resistance genes and chromosomal point mutations detected in Salmonella *spp. The data in the table is based mainly on AMRFinder nomenclature.*

Antimicrobial class	Antimicrobial resistance genes (ARG) and point mutations (PM)
Penicillins	*blaTEM-1, blaCARB-2*
1st , 2nd, 3rd and 4th generation cephalosporins	*blaSHV-12, blaCMY-2, blaCTX-M-1, blaCTX-M-65, blaHER-3, blaCTX-M-3, blaCTX-M-9, blaCTX-M-14, blaCTX-M-15, blaCTX-M-55,* *blaNDM-1, blaVIM, blaKPC, blaOXA-48, blaDHA*
Fluoroquinolones	*aac(6')-Ib-cr; ramR_G25A, ramR_T18P, qnrA1, qnrB19, qnrB2, qnrS2, oqxA, oqxB* Single PM in *gyrA_D87G, gyrA_D87N, gyrA_D87Y, gyrA_S83F, gyrA_S83Y, gyrB_S464F* Multiple PMs in *gyrA_D87G, gyrA_D87N, gyrA_D87Y, gyrA_S83F, gyrA_S83Y, gyrB_S464F*
Macrolides	*mef(B), mef(C), mph(A), mph(E), mph(G), msr(E), ant(2'')-Ia, aph(3')-Ia, aph(3'')-Ib, aph(3')-IIa, aph(4)-Ia, aph(6)-Ic, aph(6)-Id, ere(A)*
Phenicols, Lincosamides, Pleuromutilins and Oxazolidinones	*lnu(G)*
Bleomycin	*bleO*
Aminoglycosides	*aadA1, aadA2, aadA5, aadA6, aadA7 aadA22, aac(2')-IIa, aac(3)-II, aac(3)-IIa, aac(3)-IId, aac(3)-IIg, aac(3)-IV, aac(3)-Iva, aac(3)-VIa, aac(6')-Ib3, aac(6')-Ib4, aac(6')-Ib-cr, aac(6')-IIc, rmtE1, armA*
Phenicols	*catA1, catA2, cmlA1, cmlA5, floR*
Fosfomycin	*fosA3, fosA7*
Polymyxins	*mcr-1, mcr-2, mcr-3, mcr-4, mcr-5, mcr-9*
	PMs in *pmrA*, *pmrB*
Sufonamides	*sul1, sul2, sul3*
Trimethoprim	*dfrA1, dfrA12, dfrA14, dfrA15, dfrA19, dfrA34, drfA5*
Tetracyclines	*tet(A), tet(B), tet(C), tet(D), tet(G), tet(M)*

APPENDIX 2 – AMR genes and point mutations in *Campylobacter*

Table 2. Significant antimicrobial resistance genes and chromosomal point mutations detected in Campylobacter *spp. The data in the table is based on AMRFinder nomenclature.*

Antimicrobial class	Antimicrobial resistance genes (ARG) and point mutations (PM)
Fluoroquinolones	Single or multiple PMs in *gyrA_T86A, gyrA_T86I, gyrA_T86V, gyrA_D90N, gyrA_D90Y, gyrA_P104S, gyrA_T86A, gyrA_T86I, gyrA_T86K*
Macrolides	*erm(B), erm(N)* PM in *23S_A2075G, 50S_L22_G86E*
Phenicols, Lincosamides, Pleuromutilins and Oxazolidinones,	*cfr(C), lnu(C)*
Tetracyclines	*tet(O)*
Aminoglycosides	*ant(6)-Ia, aph(2'')-If, aph(2'')-Ig, aph(2'')-Ih, aph(2'')-IIIa, aph(3'), aph(3')-IIIa, aph(3')-VIIa, aadE, aad9, sat4* PM in *rpsL_K43R, rpsL_K88R* *aadE-sat4-aphA-3* Gene Cluster
Beta-lactams	*blaOXA-61, blaOXA-184, blaOXA-193, blaOXA-447, blaOXA-449, blaOXA-460, blaOXA-461, blaOXA-466, blaOXA-489, blaOXA-493, blaOXA-577, blaOXA-578, blaOXA-580, blaOXA-594, blaOXA-595, blaOXA-603, blaOXA-623*