DTU Food
National Food Institute

European
Union
Reference
Laboratory
Antimicrobial
Resistance

Protocol for whole genome sequencing and bioinformatic analysis of bacterial isolates related to the EU monitoring of antimicrobial resistance

Authored by the EURL-AR

Version 2.2 - December 2021

Rene S. Hendriksen & Jette S. Kjeldgaard

HISTORY OF CHANGES				
Version	Sections changed	Description of change	Date	Approval
2.2	Links	Update of links	03 Dec 2021	EURL-AR
2.1	AMR gene and point mutation prediction	Minor changes	February 2021	EURL-AR
2	Assessment of genome quality AMR gene prediction	Modifications and addition of specifications	15 Dec 2020	
1	All through the document	Minor modifications based on comments from the EURL Working group of NGS	10 Jan 2020	The EURL Working Group of NGS
Draft	New draft document	-	06 Dec 2019	Authors

Background

Application of whole genome sequencing (WGS) offers the possibility to rapidly transform and improve our understanding of the epidemiology of antimicrobial resistance (AMR) determinants. It can also enhance the prediction of antimicrobial susceptibility testing (AST), in view of the extensive information obtained with a single assay.

At present, antimicrobial susceptibility testing is performed by well-defined laboratory methods. In these, typically the minimum antimicrobial concentration at which bacterial growth is inhibited (MIC, minimum inhibitory concentration) are measured. The numerical values obtained (from the MIC) are then interpreted according to set criteria to classify isolates as wildtype/non-wildtype or susceptible/resistant, if the outcome of interest is the occurrence of acquired resistance determinants or prediction of clinical outcome, respectively. Although supported by long-term practice and by availability of comprehensive guidelines for performance and interpretation of AST (EUCAST, CLSI), this approach still suffers from serious drawbacks; *e.g.* when defining MICs/inhibition zone diameters for bacteriostatic antimicrobials. Furthermore, each bacterial isolate can only be tested for susceptibility to a limited number of antimicrobials to contain costs. In addition, an error rate is inherent to the methodology, which causes problems in result interpretations. This is especially ambiguous when MICs/inhibition zone diameters are close to breakpoint values for the definition of wildtype/non-wildtype or susceptibility/resistance.

Sequence-based technologies produce their outputs from the DNA fingerprint of the input material. It is of high importance how this data material is generated and quality assessed. The use of WGS for prediction of resistance genes is not a trivial protocol, as it covers a range of different laboratory processes and bioinformatics applications, which can be performed in various different ways. In relation to the Commission Implementing Decision 2020/1729 on the monitoring and reporting of antimicrobial resistance in zoonotic and commensal bacteria, and authorising the use of WGS as an alternative method for prediction of resistance in relation to the specific monitoring of ESBL- or AmpC- or carbapenemase-producing *E. coli* and *Salmonella*, the EURL-AR has produced the present protocol for guidance in these matters.

The whole genome sequencing (WGS) processes divides into three overall processes:

- Bacterial isolation, DNA preparation and DNA quality and quantity assessment
- Library preparation, library quality and quantity assessment and sequencing
- Bioinformatics analysis

Suggestions for commonly used technologies and methods within these areas are covered in this protocol.

Contents

Table of Contents

Important notes

I) Sequencing technologies and preparation protocol:
The sequencing technology used for generating DNA sequences is dependent on the equipment available by the NRL. The selected methodologies are of less importance, as long as the generated DNA sequences are of sufficient quality for the downstream sequence analysis. Likewise, the choice of methods/commercial kits for DNA extraction and library preparation, the type of flow cell used during the steps in the sequencing process is the choice of the NRL or as per recommended to use for the specific sequencing equipment.

II) Continuous improvements and updates are ongoing:
There is a continuous development in all parts of the WGS processes, and protocols and products are updated at regular intervals. Therefore, protocols for the specific processes are included as links to the manufacturer's homepage, and laboratories should always check if an updated protocol is available online. When using Illumina technologies, it is always recommended to sign up for the Illumina newsletter, which contains information on new products, protocols, product recalls and other relevant information. This protocol will contain examples of widely applied protocols, and is by no means a list of all available products and protocols. A list of abbreviations is included in Table 1 and all links to protocols, resources and tools are collected in Table 2.

III) Most commonly used platforms:
The majority of laboratories performing sequencing are currently employing Illumina technologies (Illumina, Inc., San Diego, CA; **Link 1;** Table 2), which, dependent on the required throughput, could be Illumina MiniSeq, Illumina MiSeq, Illumina NextSeq, Illumina HiSeq, or Illumina NovaSeq, and thereby this protocol will be aimed at these technologies.

Examples of other widely used sequencing platform technologies available are Oxford Nanopore MinIon (**Link 2**) and ThermoFisher Ion Torrent and Ion Proton (**Link 3**) and performance indicators are furthermore described in the WHO landscape paper (WHO, 2018). There are still some limitations in the use of sequences generated by these platforms, including available tools for sequence quality control and compatibility of some analysis tools for the sequences generated, which hamper the use of these for the harmonised surveillance of AMR.

Currently, the Inter-EURLs Working Group (WG) is producing information and guidance documents about Next-generation sequencing (NGS), to be available on the EURLs' websites (e.g. **Link 4**). These include a document on bioinformatics tools for basic analysis of Next Generation Sequencing data (**Link 5**), and additional relevant documents are in the making.

Protocol

Bacterial isolation

The bacterial isolates should be obtained through the laboratory's standard procedure for isolation of the different organisms.

The EURL-AR recommended protocols for isolation of ESBL, ampC and carbapenemase-producing *E. coli* from fresh meat samples and from caecal samples are available at the EURL-AR website **(Link 6).**

For *Salmonella* spp., refer to the routine procedures used for isolation of *Salmonella* spp. for the national control programme. Procedures for preparation of samples are described in the EN-ISO 6887 series and for some specific products at the EURL-*Salmonella* website **(Link 7).** The international accepted method for the detection of *Salmonella* is described in EN-ISO 6579-1:2017, and for serotyping of *Salmonella* in CEN-ISO/TR 6579-3:2014. Alternative methods for detection as well as for serotyping of *Salmonella* can be used, provided they are validated in accordance with the relevant procedure described in one of the parts of EN-ISO 16140.

For *Campylobacter* spp. the EURL recommended protocol is available (currently as a draft version) on the EURL-*Campylobacter* website (**Link 8**)

The correct identification and purity of the bacterial isolates is crucial for obtaining DNA of the requested isolates only.

DNA preparation and quality assessment

DNA should be extracted from bacterial colonies without introducing contaminants or inhibitors, such as foreign isolates or chemicals like EDTA. The recommended extraction kit can be dependent on the preferred library preparation kit, but the laboratory's routine method for DNA extraction will often be applicable for WGS. A few issues can be considered; boiling lysates for extraction of DNA are generally not recommended to be used for WGS, and further; DNA extraction protocols based on salt and ethanol precipitation can cause the lack of plasmid extraction, which could be relevant for determining the presence of AMR genes.

One example of a widely used extraction kit is Invitrogen Easy-DNA^{TM} Kit (Invitrogen, Carlsbad, CA, USA) (**Link 9**).

DNA can also be extracted partly or fully automated using high-throughput robotic workstations. One example of this type of equipment is a MagNA Pure 96 Instrument. A wide range of instruments are also available (**Link 10**).

DNA quantity and quality (including purity) should be assessed before proceeding to library preparation. The quality can be assessed using spectrophotometric methods, the DNA quantity can be assessed by using fluorometric methods to measure the DNA concentrations. The required DNA quantity is dependent on the library kit specifications.

Many laboratories use the Qubit dsDNA assay kit (Invitrogen, Carlsbad, CA, USA) for quantification. Hereby an overview of applications (**Link 11**) and an example of a current protocol using Qubit 4 Fluorometer (**Link 12**).

Library preparation

The methods used for library preparation are dependent on the intended sequencing platform. For library preparation from bacterial DNA, the kits recommended by Illumina, and most commonly used, are the Nextera DNA Flex and Nextera XT Library Preparation Kits. The specific protocols for sample preparations can be found using the Illumina NexteraXT™ Guide 15031942 **(Link 13)** or Illumina Nextera™ DNA Flex Library Prep Reference Guide (1000000025416) **(Link 14).** These protocols are regularly updated on the Illumina website.

Accurate quantification and proper quality check of bacterial DNA libraries are crucial for a successful sequencing run. Different methods for quantification and quality control are recommended depending on the sequencing library kit being used **(Link 15)**.

For DNA quantification, fluorometric methods like Qubit or bead-based normalization are recommended, but qPCR is also an option. For quality control, Illumina recommend Bioanalyzer or FragmentAnalyzer. The quality control is optional according to Illumina when using bead-based normalization, however this can be problematic if a number of samples have low DNA concentration after library preparation.

The choice of flow cell for the sequencing run is dependent on the laboratory's need for capacity regarding both the amount of isolates sequenced, the amount of data coverage generated per sequencing run and the desired coverage for each bacterial genome. Flow cells are available for both single isolates or for several hundred isolates per sequencing run. A full list of the current reagents and flow cells available for the different sequencing platforms is regularly updated by Illumina (**Link 16**).

Sequencing

Examples of current protocols of how NexteraXT libraries are loaded onto an Illumina MiSeq (Guide 15039740; **Link 17**) or NextSeq (Guide 15048776; **Link 18**) reagent cartridge using Reagent Kit v2 or v3 are available. Protocols are depending on the sequencing equipment, and are regularly updated on the Illumina website. Protocols for other Illumina sequencing platforms are also available on Illumina website.

When acquiring Illumina sequencing platforms, Illumina offers free guidance and training for their equipment and protocols. Illumina software offers a range of QC parameters for each sequencing run. These are explained in the experiment manager software guide

(Guide 15031335; **Link 19**). Similar training services and QC parameters are also offered by the providers of other sequencing platforms (e.g. Thermo Fisher Ion Torrent).

Assessment of the genomic sequence quality

Quality control (QC) is essential to guarantee accuracy and precision of any laboratory test results, including WGS. Genomic sequences (raw reads (fastq) or contigs (fasta)) of poor quality can lead to major errors in prediction of AST by failing to reveal AMR genes or mutations. Other sources of error may derive from contamination of the DNA or from erroneous data handling. At present, different QC parameters are available to control and standardize WGS procedures (Ellington et al. 2017). Only datasets that pass agreed QC metrics should be used in prediction of antimicrobial resistance.

File format

The raw sequences generated through Illumina sequencing are in the file format *.fastq*. These can often be used directly in many downstream analysis tools, although trimming of adaptors and low quality reads are recommended. Thus, many of the available criteria for genomic sequence QC are relying on the assembly process, generating a set of sequence contigs in the file format *.fasta*. For this reason, it is recommended to include the assembly of .fastq files as a part of the quality control.

Available tools for trimming and quality control

Several tools are available for sequence quality control, either as QC pipelines or web-based tools. Tools often used for trimming include (but are not limited to) Cutadapt (**Link 20**) and Trimmomatic (**Link 21**). The Illumina software performs some extent of adaptor trimming, and as such, trimming is not essential for the sequence quality.

More general information on available tools can be found in the Inter-EURL WG on NGS document (**Link 5**) and on the ENGAGE website, which also include a guide for genome quality assessment with a description of important criteria (**Link 22**).

Contamination of genomic sequences

Contamination with unwanted DNA can occur in any stage of the laboratory preparation, including carry-over of traces of DNA during sequencing, and can be caused by either foreign or similar bacterial species. A test for contamination with foreign species can be performed by using e.g. KmerFinder (**Link 23**) or similar tools for species determination. Some QC parameters like size of assembled genome and total number of contigs can give an indication of contamination with both foreign or similar bacterial species.

QC parameters

A set of QC parameters for draft genome assembly and their explanation has been listed by the EUCAST committee (Ellington et al. 2017). Some of the more commonly used are; number of reads, average read length, depth of coverage, size of assembled genome, total number of contigs and N50, for which the following recommendations are given:

- The **number of reads** and **depth of coverage** should be as high as possible. There is no assessed cut-off numbers for these. The higher number of reads and depth of coverage indicate high amount of raw reads data to start with. The depth of coverage (C) can be calculated as the length of the reads (L) divided by the genome size (G) multiplied by the number of reads (N); **C = N * (L/G)**.
- The **average read length** should be similar to the expected read length from the selected sequencing platform.
- The **size of assembled genome** should deviate more than 0.5 million base-pairs from the expected genome size. If the deviation in the size of the assembled genome is greater than 0.5 million based-pair of expected genome size, this is an indication that the genome sequences are either contaminated, not the expected species or poor sequencing quality.
- The **total number of contigs** (after assembly) should be less than 500 contigs. A higher number of contigs indicates poor sequencing quality.
- **N50** indicates size of contigs in general. The higher N50 indicates the longer contigs in a genome. There is no general cut-off for N50, but some suggest using a N50 of >30 000 bp (Bortolaia et al. 2020).

Assembly

Assembly of raw reads into contigs can be performed using various tools, including SPAdes (**Link 24**) among many others (For additional information see **Link 5**). For harmonisation purposes, the EURL recommends to use **SPAdes v 3.14** or newer. This assembly tool is available as an open-access web server on the CGE website, for single isolate upload (**Link 25**) or for local installation (**Link 24**). The web server assembler can run on Illumina paired and single end files. It is recommended to choose the 'Careful' option, which tries to reduce the number of mismatches and short indels.

The SPAdes 3.14 tool will output the contigs file (.fasta) and additionally a .txt file with some basic statistics and QC parameters.

The output file contains data on:

- Input files :
  - Total number of reads
  - Total number of bases
- Contigs file :
  - Number of contigs
  - Number of bases (assembled genome size)
  - N50

Using this output, it is also possible to calculate the average read length= Number of bases/Number of reads (input files).

AMR gene and point mutation prediction

Genomic sequences assessed to pass the QC demands are further analysed for the predicted presence of acquired AMR genes and chromosomal point mutations using the open access web-based tool ResFinder (also including PointFinder; Bortolaia et al. 2020). A recent review that describes with examples available tools and databases for antimicrobial resistance detection has been published (Hendriksen et al. 2019).
The EURL-AR recommends using the ResFinder tool v4.1 or newer, which is available from the CGE website (Developed, owned and curated by DTU; **Link 26**; and for local installation (**Link 27**)(Bortolaia et al. 2020).

For harmonisation of the AMR data reported by different laboratories, it is important to use the defined settings. The EURL-AR recommends running the ResFinder analysis on the contigs assembly files (.fasta) using the following settings, which are set as default:

<u>For chromosomal point mutations:</u>

- Select threshold for % ID: 90 %
- Select minimum length: 60 %

<u>For acquired antimicrobial resistance genes</u>:
Select all antimicrobial databases (default setting)

- Select threshold for % ID: 90 %
- Select minimum length: 60 %

Select species: as appropriate
Select type of your reads: Assembled genome/Contigs

Beyond the sampling and isolate data, the results reported in relation to Decision 2020/1729 should include:

- Date of sequencing
- Sequencing technology used
- Library preparation used
- Version of the predictive tool
- AMR-conferring genes data:
  - Gene name
  - Output information on % identity
  - Output information on % coverage (length)
- Date of ResFinder analysis

The data submission system is still pending. EFSA is building a platform for facilitating analysis and harmonised data output.

Additional analysis and sub-typing

Further bioinformatics analyses, such as species determination, sub-typing of a range of different bacteria and cluster analysis, can be conducted based on other EURL-AR recommended pipelines and are available from the CGE website (**Link 28**). These include determination of MLST, cgMLST, serotyping, identification of virulence factors and plasmids, and phylogenetic SNP analysis. The Inter-EURLs WG on NGS is finalising a document on cluster analysis on SNP and cgMLST, which is expected to be published early 2021 (**Link 4**)

Proficiency test

It is advisable for laboratories to participate in proficiency tests or ring trials regarding WGS and identification of AMR genes, to be able to evaluate the laboratory performance. One example of this is the DTU Genomic Proficiency Test 2021, an inter-laboratory performance test, which is organised to run in 2021, and there will be announcements of new rounds of this proficiency test on the web site (**Link 29**).

Online training

DTU provides free online courses in both WGS and AMR. Visit the course home pages at Coursera for more details: Whole genome sequencing of bacterial genomes - tools and applications (**Link 30**) and Antimicrobial resistance - theory and methods (**Link 31**).

References

Bortolaia, V et al. (2020) ResFinder 4.0 for predictions of phenotypes from genotypes, *Journal of Antimicrobial Chemotherapy*, Volume 75, Issue 12.
DOI: https://doi.org/10.1093/jac/dkaa345

CEN-ISO/TR 6579-3:2014. Microbiology of the food chain – Horizontal method for the detection, enumeration and serotyping of *Salmonella* - Part 3: Guidelines for serotyping of *Salmonella* spp.

Ellington, MJ et al. (2017) The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clinical Microbiology and Infection*, Volume 23, Issue 1.
DOI: https://doi.org/10.1016/j.cmi.2016.11.012

EN-ISO 6579-1:2017. Microbiology of the food chain – Horizontal method for the detection, enumeration and serotyping of Salmonella - Part 1: Detection of *Salmonella* spp.

EN-ISO 6887:2010-2017. Microbiology of the food chain - Preparation of test samples, initial suspension and decimal dilutions for microbiological examination - all parts.

EN-ISO 16140: 2016-2020. Microbiology of the food chain – Method validation – all parts.

Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM and McDermott PF (2019) Using Genomics to Track Global Antimicrobial Resistance. *Front. Public Health* 7:242.
Doi: 10.3389/fpubh.2019.00242

WHO (2018) Whole genome sequencing for foodborne disease surveillance: landscape paper. Geneva: World Health Organization; 2018. Licence: CC BY-NC-SA 3.0 IGO

Abbreviations and acronyms

Table 1: Commonly used abbreviations and definitions

Abbreviation	Explanation
AMR	Antimicrobial resistance
AST	Antimicrobial susceptibility testing
CGE	Centre for Genomic Epidemiology
cgMLST	Core genome multi locus sequence type
DTU	Technical University of Denmark
ENGAGE	Project: Establishing Next Generation sequencing Ability for Genomic analysis in Europe
ESBL	Extended Spectrum Beta-Lactamase
EUCAST	The European Committee on Antimicrobial Susceptibility Testing
EURL(-AR)	European Union Reference Laboratory (for antimicrobial resistance)
CLSI	The Clinical and Laboratory Standards Institute
MIC	Minimal inhibitory concentration
MLST	Multi locus sequence type
N50	Quality control parameter; N50 is defined as the sequence length of the shortest contig at 50% of the total genome length
NGS	Next –generation sequencing
NRL	National reference laboratory
QC	Quality control
qPCR	Quantitative PCR - a method of quantifying DNA based on PCR
SNP	Single nucleotide polymorphism
WGS	Whole genome sequencing
WHO	World Health Organisation

Links

Table 2: Collection of links referred to in the protocol, including last date of accession

Link#	Method or content	Last accessed
Link 1	**Illumina website** https://emea.illumina.com/	**November 2021**
Link 2	**Oxford Nanopore website** https://nanoporetech.com/products	**November 2021**
Link 3	**Thermofisher website** https://www.thermofisher.com/dk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-products-services.html	**November 2021**
Link 4	**EURL-AR website – Inter-EURLs WG on NGS** https://www.eurl-ar.eu/inter-eurls-working-group-on-ngs.aspx	**November 2021**
Link 5	**Document on bioinformatics tools for basic analysis of Next Generation Sequencing data** https://www.iss.it/documents/20126/0/Bioinformatics_tools_for_basic_analysis_of_Next_Generation_Sequencing_data_Del4.pdf/02c8f77b-db2c-6b8d-e2ba-416144f89f7e?t=1602603602556	**November 2021**
Link 6	**Methods for isolation of ESBL, ampC and carbapenemase-producing *E. coli* from meat and caecal samples** https://www.eurl-ar.eu/protocols.aspx	**November 2021**
Link 7	**Method for detection of *Salmonella* in food and animal feed** https://www.eurlsalmonella.eu/publications/analytical-methods	**November 2021**
Link 8	**Method for detection of Campylobacter** https://www.sva.se/en/about-us/eurl-campylobacter/laboratory-procedures/	**November 2021**
Link 9	**DNA extraction protocol EasyDNA** https://assets.thermofisher.com/TFS-Assets/LSG/manuals/easydna_man.pdf	**November 2021**
Link 10	**Automated DNA extraction Magna Pure** https://lifescience.roche.com/en_dk/products/magna-pure-96-instrument-382411-1.html	**November 2021**
Link 11	**Overview of applications of Qubit** https://www.thermofisher.com/dk/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit.html	**November 2021**
Link 12	**Protocol for Qubit 4 DNA quantification** https://assets.thermofisher.com/TFS-Assets/BID/manuals/MAN0017210_Qubit_4_Assays_QR.pdf	**November 2021**
Link 13	**Library prep Nextera XT** http://support.illumina.com/downloads/nextera_xt_sample_preparation_guide_15031942.html	**November 2021**

Link 14	**Library prep Nextera DNA Flex** https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/illumina-dna-prep-reference-guide-1000000025416-09.pdf	**November 2021**
Link 15	**Guide for quantification and QC of library prep** https://emea.support.illumina.com/bulletins/2016/05/library-quantification-and-quality-control-quick-reference-guide.html	**November 2021**
Link 16	**Illumina instrument-specific sequencing reagents, flow cells, cluster generation reagents** https://emea.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents.html	**November 2021**
Link 17	**MiSeq** https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-10.pdf	**November 2021**
Link 18	**NextSeq** https://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/nextseq/nextseq-500-550-denature-dilute-libraries-guide-15048776-16.pdf	**November 2021**
Link 19	**Illumina experiment manager software guide** https://support.illumina.com/downloads/illumina-experiment-manager-user-guide-15031335.html	**November 2021**
Link 20	**Cutadapt Trimming tool** https://cutadapt.readthedocs.io/en/stable/guide.html	**November 2021**
Link 21	**Trimmomatic Trimming tool** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/	**November 2021**
Link 22	**Guide to available web tools for assessment of genome quality** http://www.engage-europe.eu/resources/protocols-and-training	**November 2021**
Link 23	**KmerFinder tool** https://cge.cbs.dtu.dk/services/KmerFinder/	**November 2021**
Link 24	**SPAdes Assembly website** http://cab.spbu.ru/software/spades/	**November 2021**
Link 25	**SPAdes EURL web-tool** https://cge.cbs.dtu.dk/services/SPAdes-3.14/	**November 2021**
Link 26	**ResFinder 4.1 web-tool** https://cge.cbs.dtu.dk/services/ResFinder/	**November 2021**
Link 27	**ResFinder – BitBucket** https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/	**November 2021**
Link 28	**CGE web tools** https://cge.cbs.dtu.dk/services/	**November 2021**

Link 29	**DTU Genomic Proficiency Test 2020** https://www.globalsurveillance.eu/projects/genomic-proficiency-test-2021	**November 2021**
Link 30	**WGS online course on Coursera** https://www.coursera.org/learn/wgs-bacteria	**November 2021**
Link 31	**AMR online course on Coursera** https://www.coursera.org/learn/antimicrobial-resistance	**November 2021**