Service Contract for the provision of EU networking and support for public health reference laboratory functions for antimicrobial resistance in priority healthcare associated infections

SC 2019 74 01

Agreed common WGS-based genome analysis methods and standard protocols for national CCRE surveillance and integrated outbreak investigations

Version n°: 1.0
Date: 01-09-2022

Health and Digital Executive Agency

Third EU Health Programme

Proposed common WGS-based genome analysis methods and standard protocols for national CCRE surveillance and integrated outbreak investigations

---

This report was produced under the EU Third Health Programme 2014-2020 under a service contract with the Consumers, Health, Agriculture and Food Executive Agency (Chafea) acting under the mandate from the European Commission. From 1 April 2021, a new executive Agency with name HaDEA (Health and Digital Executive Agency) is taking over all contractual obligations from Chafea. The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the Commission/Executive Agency. The Commission/Executive Agency do not guarantee the accuracy of the data included in this study. Neither the Commission/Executive Agency nor any person acting on the Commission's/Executive Agency's behalf may be held responsible for the use which may be made of the information contained therein.

EUROPEAN COMMISSION
Health and Digital Executive Agency (HaDEA)

Directorate-General Health and Food Safety (DG SANTE)
Directorate C — Public health
Unit C3 — Health security
European Commission
L-2920 Luxembourg
Email: **SANTE-CONSULT-C3@ec.europa.eu**

TABLE OF CONTENTS

1. INTRODUCTION

The EURGen-RefLabCap project is complementary to the European Centre of Disease Prevention and Control (ECDC) European Antimicrobial Resistance Genes Surveillance Network (EURGen-Net). The project aims at improving capacities of National Reference Laboratories (NRLs) in European countries for identification and for phenotypic and genotypic characterization of carbapenem-resistant *Enterobacterales* (CRE) and colistin-resistant CREs (CCRE), and other antimicrobial-resistant pathogens. Furthermore, the project aims at strengthening capacities for national surveillance and outbreak investigation of CRE/CCRE and improve the availability and quality of European-level molecular surveillance data. One of the main goals of the EURGen-RefLabCap project is to support modernisation of diagnostic and molecular typing tests using whole-genome sequencing (WGS) analytical methods in order to achieve those respective aims.

This protocol provides a framework to perform WGS directed towards short-read paired-end massive parallel synthesis sequencing, specifically using Illumina platforms (Illumina, Inc., San Diego, CA, USA) such as MiSeq and NextSeq. In addition, it presents the framework for bioinformatic analysis of CRE/CCRE using three pipelines to detect antimicrobial resistance determinants – particularly, resistance to colistin and carbapenems. The protocol covers the steps of obtaining high quality DNA, performing library preparation and sequencing of the DNA, performing bioinformatics analysis (taxonomic analysis, bacterial typing, detection of genetic determinants of antimicrobial resistance, cluster analysis) and adopting best practices for data management. Furthermore, this protocol defines specific quality control (QC) strategies, QC parameters and respective thresholds. Using other WGS platforms might yield results of equally good quality, but the bioinformatics tools and QC thresholds should be adapted accordingly.

Note: In most cases, WGS-based outbreak analysis cannot stand alone in outbreak investigations, but it is a powerful tool to guide directed epidemiological investigations.

The document will also be accompanied by a suite of supporting documents to aid novel users in becoming familiar with the most relevant WGS concepts and terms. These will also provide a review of available bioinformatics tools, bioinformatics development initiatives, and reference gene databases for the detection and prediction of relevant drug resistance determinants in CRE/CCRE, as well as scientific and technical background for each step of the workflow, links to other relevant resources which contain more detailed descriptions of certain steps, and related scientific literature. The suite of supporting documents will be available on the EURGen-RefLabCap website https://www.eurgen-reflabcap.eu/.

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

2. PROTOCOL

This protocol describes the different steps necessary to perform WGS of CRE/CCRE. Furthermore, it describes how to perform bioinformatic analysis of CRE/CCRE through open-source, curated bioinformatics tools and databases. Users might opt to employ different approaches as long as these are properly validated for the purpose. For each step, different methods, kits, and tools exist, thus it is important to carefully consider and take into account the laboratory's existing and available consumables, kits, and equipment that can be applied to this procedure. It is plausible that, as the laboratory's expertise increases, some of the proposed steps can be streamlined or skipped entirely. However, it is useful to start by using the complete protocol with as many QC steps as possible, to gain maximum understanding of the WGS process and troubleshooting possibilities.

Procedure	**Theory/ Comments**
DNA extraction and QC	
1. From a primary culture, select one single isolated colony to prepare a subculture.	Streaking out a fresh culture from a single colony should be implemented as a routine.
2. Inspect the subculture carefully to ensure purity. If the culture is not pure, prepare a new subculture.	Do not extract DNA from cultures that are not pure.
3. Extract bacterial DNA using in-house protocols or commercial kits.	Examples of commercial kits are, among others, ThermoFisher Easy-DNA gDNA Purification Kit and Qiagen DNeasy Blood & Tissue Kit. A range of instruments exists for more automated high-throughput DNA extraction, one example being the *MagNa Pure 96* instrument. Be aware that extraction methods based on salt and ethanol precipitation can result in poor plasmid extraction, which can be problematic for determination of antimicrobial resistance (AMR) genes, as these often reside on plasmids.
4. Measure UV 260/280 absorbance ratio values of the DNA samples to confirm that they are in the interval 1.8 – 2.0. If absorbance ratio values are outside the interval, the DNA should be re-extracted. If QC thresholds are not achieved, the DNA should be re-sequenced.	This can be done by using, for example, Nanodrop or Bioanalyser.

DNA Concentration and Dilution	
5. Measure the concentration of the undiluted DNA samples. If below the necessary concentration, re-extract the DNA. If above the necessary concentration, dilute the DNA with the adequate buffer to achieve a final concentration in accordance with the library preparation protocol.	The DNA quantity should be assessed, since a specific amount of DNA needs to be used for the library preparation step. Many laboratories use Qubit fluorometer and respective dsDNA reagent kits. Alternatively, Nanodrop spectrophotometer may be used to quantify DNA, using only 1-2 µl of sample volume. *Example: a final concentration of 2 ng/µl is needed if using the Nextera XT Library Preparation Reference Guide, with input of 5 µl of each library.*
6. Confirm the DNA concentration of the diluted samples.	This may be done using, for example, the Qubit fluorometer and the Qubit™ dsDNA High Sensitivity Assay Kit.
7. The DNA dilution and confirmation of the DNA concentration should be repeated until the desired concentration is achieved.	In case the initial DNA concentration is too low it will be necessary to re-extract DNA from the sample or concentrate the DNA solution.
Library preparation and DNA sequencing	
8. Perform library preparation. Various methods for library preparation exist and they depend on the chosen sequencing platform.	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 machinery are frequently updated and available on the Illumina website. Examples are the Illumina Nextera XT Reference guide or Illumina DNA Prep Reference Guide. Other library preparation kits and protocols can be used. According to the choice, other reference guides and accessory documents might be needed.

WGS raw data extraction and QC	
9. Extract the raw data and store them locally.	The raw reads are in the *fastq* file format, which also includes quality parameters (phred scores).
The raw data might be located on the Illumina sequencer or from a cloud solution such as Illumina sequence hub.	The cloud solution also offers a range of visual QC parameters to evaluate the sequencing run.
10. Perform QC of the sequences.	The quality of the sequences should always be assessed, as poor quality sequences can lead to major errors in prediction of genes and phylogenetic analysis. FastQC is an example of a tool that can be used for this purpose.
Minimum QC parameters should be determined, specifically the *average read length*, *coverage* and *number of reads*.	Average read length should be equal to the expected read length from the sequencing platform. Depth of coverage should be as high as possible. No harmonised cut-off exists, but a minimum coverage of 30X is often used as standard. Lower coverage values may interfere with later analysis and prevent comparison of inter-laboratory data. Thus, these should not be implemented routinely, even if they might be accepted for specific internal analysis. Number of reads should be sufficient to ensure a coverage of at least 30X, using the formula: "Coverage = Number of reads x (Read length / Genome size)".
Raw data should preferably be examined for potential contaminations.	For instance, KRAKEN can be used to quantify the number of reads assigned to other species than the target species. The percentage of reads assigned to other species should be residual (for example less than 5%). Contamination checks can also be facilitated by tools such as KmerFinder or rMLST.
Raw reads should be trimmed for adaptors and low quality regions.	Using tools such as Bbtools or Trimmomatic.

If QC thresholds are not achieved, the DNA should be re-sequenced or re-extracted.	
Genome assembly and QC	
11. Assemble the reads into contigs (*fasta* files).	Genome assembly may be performed using SPAdes, Unicycler or other preferred assembly tools. Most assembly programs can be installed locally, and many institutions performing WGS routinely have this step incorporated into their analysis pipeline.
12. Perform QC of the assembly.	A tool that can be used for this purpose is QUAST. Other available public QC and assembly pipelines, such as BIFROST, exist on Github or other repositories.
Minimum QC parameters should be determined, specifically the *number of contigs, N50, coverage* and *genome size*. The proposed QC thresholds should, in principle, guarantee that results obtained with assembled data are comparable with results obtained with raw data. Furthermore, using benchmarking datasets ensures that the selected assembly tool and QC thresholds yield accurate results.	Most assembly QC parameters are dependent on the sequencing platform and bacterial species. If using Illumina platforms to analyse CRE/CCRE: Number of contigs should be less than 500. A higher number may point to poor sequence quality or to contamination (also with isolates belonging to the same species, which is not always detectable with species identification tools or through analysis of raw data). N50 should be as high as possible, and larger than 15,000. Depth of genome coverage should be at least 30X. Genome size should be within 10% of deviation of the expected genome size. A larger genome size can indicate that the sample was contaminated (including with isolates belonging to the same species), while a smaller genome size can be due to poor DNA extraction or insufficient amount of sequenced data. *Enterobacterales* genome size is generally between 4.5 – 5.5 million bps.
If QC thresholds are not achieved, the DNA should be re-sequenced or re-extracted.	

Bacterial species identification and QC	
13. Use a curated bioinformatics tool to perform species identification. If using KmerFinder, the QC parameters should be confirmed as follows: - at least 90% of template and of query coverage when summing up the several hits from the same species; - low number of individual hits; - high score (naturally occurring when both previous parameters are fulfilled); - absence (or very low percentage) of hits belonging to different species. If using rMLST, the QC parameters should be confirmed as follows: - at least 96% of support: - absence of hits belonging to different species. If QC parameters of the primary and secondary tools are not fulfilled, the DNA should be re-extracted.	Examples of commonly used species identification tools are KmerFinder and rMLST. Other tools are available for bacterial species identification. Independent of the tool used, it is of critical importance to fulfil the QC parameters specific for the selected tool. If the QC parameters of your chosen tool are not fulfilled, species can be determined with a second tool.
Bacterial isolate typing	
14. Use a species-specific MLST typing scheme such as PubMLST.	If a sequence type (ST) is not assigned, a different scheme may be used, if available. If no schemes successfully assign a ST, the target isolate might also represent a new ST. However, the possibility of contamination with isolates belonging to the same species should be considered when troubleshooting. Tools such as MLST can be used to perform the analysis in user friendly interfaces. Other species-specific tools or pipelines also include bacterial typing; one example is Kleborate, a tool to screen genome assemblies of *Klebsiella pneumoniae* species complex.

Detection of antimicrobial resistance genes (ARGs) and chromosomal point mutations (PMs) mediating antimicrobial resistance in CRE and CCRE	
15. Use a curated bioinformatics tool to perform detection of genetic AMR determinants. If using ResFinder, the default analysis thresholds of minimum 90% of identity and minimum 60% of length are recommended. If using AMRFinderPlus, the default analysis thresholds of minimum 90% of identity and minimum 50% of length are recommended. If using CARD-RGI, the analysis parameters of "perfect and strict hits only" and "include nudge [nudge ≥95% identity Loose hits to strict]" are recommended.	Currently there are three main tools with associated specific databases to detect ARGs and PMs in WGS data: ResFinder, AMRFinderPlus and CARD-RGI. CARD is not recommended for non-experienced users as the output requires in-depth knowledge for correct interpretation. Sensitivity and specificity of the tools for detection of ARGs and PMs may be modified by adjusting the thresholds and/or parameters used for the analysis.
16. Evaluate the results (also called "hits") obtained with the chosen tool.	Be aware of which ARG and PMs are included in your chosen database: lack of hits might be due to real absence of the genes or mutations in the query genome, but might also be due to absence of those in the database. It is also possible to combine more than one tool and/or database for detection of ARGs and PMs, which requires careful evaluation of the results obtained. For ARGs: Length and identity of the gene(s) in the query genome (i.e. the genome you sequenced) should be equal to 100% of the gene(s) in the database used by the tool. If length < 100% and identity ≤ 100%, it should be verified if the gene is artificially truncated due to being positioned at the beginning or end of a contig or if it is truly a partial gene. If identity is < 100% and length ≤ 100%, it should be confirmed by searching other databases or literature if that variant has been described; if not, the impact of the nucleotide mutation(s) on the amino-acid sequence may be assessed. Silent mutation: this scenario is consistent with a predicted phenotype of

	microbiological resistance to the relevant antimicrobial(s) (with few exceptions). Other type of mutation: it is recommended not to predict an AMR phenotype but to report the detected gene variant and its attributes. The presence of multiple genes from the same gene family should be carefully evaluated to determine if it is an artefact of the tool/database used (which is revealed by observing if the genes are placed at the same positions in the same contig) or if it is a true occurrence. Generally, this scenario is consistent with a predicted phenotype of microbiological resistance to the relevant antimicrobial(s). For chromosomal PMs: Specific PMs or combinations of PMs in selected genes and bacterial species are known to mediate resistance to specific antimicrobials. If these "known" mutations are detected, the isolate likely exhibits microbiological resistance to the specific antimicrobial(s). If detecting "unknown" mutations (mutations for which a role in AMR has not been elucidated yet), results should be reported but the phenotype cannot be predicted. Bear in mind that PMs mediating AMR are generally species-specific.
Cluster analysis and respective QC	
17. Design a general approach regarding frequency of cluster analysis.	For outbreak investigation purposes, cluster analysis can be initiated as soon as there are suspicions of an outbreak, and repeated as often as needed once new isolates are collected. Warning signs that might suggest that cluster analysis should be conducted are, for example, increase in incidence of a certain species or a certain sero- or sequence type, or observing unexpected antimicrobial resistance profiles. For routine surveillance purposes it may be decided to perform the analysis every last Friday of each month, as an example.

18. Choose a general approach regarding isolates to include in the cluster analysis.	Inclusion criteria may be "all isolates from the species", "all isolates belonging to the same MLST", "all isolates collected in the last three months", etc. Epidemiological information is necessary for understanding the significance of cluster analysis results (especially for detecting outbreaks).
19. Perform SNP-based phylogenetic analysis with CSIphylogeny, FastTree or other tool. Analysis should be performed with raw WGS data and an adequate reference should be selected (an isolate with predicted high genetic relatedness).	The results should be interpreted: a SNP distance under 5 suggests relatedness of isolates, but slightly higher thresholds should not be discarded (e.g. up to 25 SNPs) depending on the nature of the given outbreak. Be aware of method limitations: at least 90% of each query genome should have been included in the alignment to create the distance matrix; lower percentages of alignment directly suggest limited relatedness of the isolates or that a non-optimal reference was used for mapping.
20. Additionally, or instead, choose another clustering approach such as using species-specific core-genome MLST schemes.	Clustering approaches can also be used and there are online interfaces, like cgMLSTFinder, that facilitate their use. cgMLST approaches may provide lower resolution than SNP-based analysis. On the other hand, a well-designed and thoroughly validated cgMLST scheme may produce more robust comparisons than SNP analysis, especially for bacterial species which undergo rapid recombination events. cgMLST is also suitable for long-term surveillance as computations generally scale better with dataset size. The results should be interpreted: the threshold of 0.0030 dissimilarity has been proposed for inferring genetic relatedness among *Enterobacterales* isolates. The conversion of percentage of dissimilarity into number of alleles depends on the cgMLST scheme of each species. *Example: 0.0030 x 2,358 loci included in* K. pneumoniae *scheme = maximum of 7 different alleles* In general terms, differences of 5 alleles suggest close genetic relatedness, but higher values should not be discarded (e.g. up to 10).

Data and metadata storage	
21. Store raw sequence data perpetually, either in private or public databases.	Raw sequence data must be accompanied by minimum metadata parameters. Examples of minimum fields are: - metadata of the isolate: collection date; geographical origin; source; sample type; expected species; storage location - details on DNA extraction: date of extraction; kit used; DNA concentration; storage location - details on library preparation protocol: date of preparation; kit used; DNA concentration of each input library; layout of the microtiter plate; normalization and dilution approaches - sequencing platform and sequencing run: platform name; sequencing run number; sequencing start date; sequencing end date; sequencing yield - raw data QC: average read length; coverage; number of reads
22. Store trimmed and assembled data likewise.	If storing assembled data, information on the assembly approach and QC should be included.
23. Store bioinformatic results, if feasible.	If storing bioinformatics results, at least the following details should be stored: information on the workflow, QC results, date of the analysis and/or version of the bioinformatics tools and databases used, and interpretation guidelines that were used.

3. SURVEILLANCE OF CRE/CCRE AND OUTBREAK INVESTIGATION

Analysis of WGS data for CRE/CCRE, together with epidemiological data, is vital for detecting the emergence of high-risk clones/plasmids, monitoring of time and spatial trends, detection and investigation of outbreaks in both community and healthcare settings and for the identification of high-risk populations, sources of transmission and prevention and control measures.

WGS-based routine and/or sentinel genomic surveillance of healthcare priority pathogens provide a cornerstone in both local, regional and national epidemic preparedness. As a first step, laboratories should implement a local sampling strategy, laboratory and clinical case definitions aligned with EUCAST guidance and EU case definitions for communicable diseases, and selection criteria for performing WGS.

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

It is important to note that one isolate harbouring ARGs or PMs that mediate resistance towards a class of antimicrobials can express different phenotypes to the individual agents included in that antimicrobial class. Also, different gene variants within the same gene family can lead to different phenotypes. Finally, there can be situations where the presence of an ARG will not lead to phenotypic resistance, due to variation in gene expression, possible simultaneous changes in expression of efflux pumps, and potential porin loss. Similarly, not all PMs in target genes will lead to phenotypic resistance. However, due to incomplete knowledge regarding the effects of all possible mutations in target genes, and the possibility that these PMs have a cumulative effect in the expression of resistance phenotypes (as extensively described in literature for PMs in the gyrase and topoisomerase genes, associated with fluoroquinolone resistance), these should be kept under surveillance.

In addition to the investigation of ARGs and PMs, selected isolates from a defined site (such as a hospital or healthcare facility, the community, a region or country) 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 multi-locus sequence typing (MLST), core genome multi-locus sequence typing (cgMLST), and phylogenetic single nucleotide polymorphisms (SNP)-based analysis. Furthermore, plasmid content and presence of genes encoding virulence factors may also be determined using WGS data. These bacterial typing and cluster analysis strategies are able to support epidemiological analysis aimed at monitoring the introduction and expansion of high-risk multidrug resistant clones, transmission events and detection of clusters and outbreaks.

The analytical WGS pipeline should be designed to meet the identified characterization and cluster analysis needs, by using sequencing and bioinformatics approaches that produce standardized results. Thus, to ensure comparability of WGS results among sites, agreement should be reached on the minimum quality control parameters and respective thresholds. These threshold parameters should be established with caution and always be used in combination with clinical epidemiological data, population and species characteristics.

Finally, by uploading raw sequence data with associated metadata to international databases, such as the European Nucleotide Archive and the National Center for Biotechnology Information, and by actively engaging in participation in the upcoming ECDC portal EpiPulse, investigations can be extended to assess cross-border transmission.

4. REFERENCES AND FURTHER READING

Description of QC parameters

Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 2019;20(6):356-370. https://doi.org/10.1038%2Fs41576-019-0108-4. PMID: 30886350.

Clausen PT, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother 2016;71(9):2484-8. https://doi.org/10.1093/jac/dkw184. PMID: 27365186.

Ellington MJ, Ekelund O, Aarestrup FM, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect 2017;23:2–22. https://doi.org/10.1016/j.cmi.2016.11.012. PMID: 27890457.

European Food Safety Authority. EFSA statement on the requirements for whole genome sequence analysis of microorganisms intentionally used in the food chain. EFSA J 2021;19. https://doi.org/10.2903/j.efsa.2021.6506. PMID: 34335919.

Rooney AM, Raphenya AR, Melano RG, Seah C, Yee NR, MacFadden DR, et al. Performance characteristics of next-generation sequencing for antimicrobial resistance gene detection in genomes and metagenomes. BioRxiv 2021:2021.06.25.449921. https://doi.org/10.1101/2021.06.25.449921.

Timme RE, Wolfgang WJ, Balkey M, et al. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Heal Outlook 2020;2:20. https://doi.org/10.1186/s42522-020-00026-3. PMID: 33103064.

World Health Organization. GLASS whole-genome sequencing for surveillance of antimicrobial resistance. Geneva, Switzerland: 2020. Available from: https://www.who.int/publications/i/item/9789240011007.

Thresholds for genetic relatedness in cluster analysis

Bush SJ, Foster D, Eyre DW, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience 2020;9:1–21. https://doi.org/10.1093%2Fgigascience%2Fgiaa007. PMID: 32025702.

Dallman TJ, Greig DR, Gharbia SE, et al. Phylogenetic structure of Shiga toxin-producing Escherichia coli O157:H7 from sub-lineage to SNPs. Microb Genomics 2021; 7:mgen000544. https://doi.org/10.1099/mgen.0.000544. PMID: 33720818.

David S, Reuter S, Harris SR, et al. Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. Nat Microbiol. 2019;4(11):1919-1929. https://doi.org/10.1038/s41564-019-0492-8. PMID: 31358985.

Jamin C, De Koster S, van Koeveringe S, et al. Harmonization of whole-genome sequencing for outbreak surveillance of Enterobacteriaceae and Enterococci. Microb Genomics 2021;7:000567. https://doi.org/10.1099%2Fmgen.0.000567. PMID: 34279213.

Kluytmans-van den Bergh MFQ, Rossen JWA, Bruijning-Verhagen PCJ, et al. Whole-Genome Multilocus Sequence Typing of Extended-Spectrum-Beta-Lactamase-Producing Enterobacteriaceae. J Clin Microbiol 2016;54:2919–27. https://doi.org/10.1128/jcm.01648-16. PMID: 27629900.

Pightling AW, Pettengill JB, Luo Y, et al. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. Front Microbiol 2018;9:1482. https://doi.org/10.3389/fmicb.2018.01482. PMID: 30042741.

Schürch AC, Arredondo-Alonso S, Willems RJL, 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;24:350–4. https://doi.org/10.1016/j.cmi.2017.12.016. PMID: 29309930.

Bioinformatics tools

AMRFinderPlus

- Feldgarden M, Brover V, Gonzalez-Escalona N, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1):12728. https://doi.org/10.1038/s41598-021-91456-0. PMID: 34135355.

Bbtools

- No publication; documentation available online for bbduk, bbmap, others.

BIFROST

- No publication.

CARD-RGI

- Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 2019;48:517–25. https://doi.org/10.1093/nar/gkz935. PMID: 31665441.

cgMLST

- Zhou Z, Alikhan NF, Mohamed K, et al. The EnteroBase user's guide, with case studies on 323 Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res 2020;30:138–52. https://doi.org/10.1101/gr.251678.119. PMID: 31809257.

cgMLSTFinder

- Leekitcharoenphon P, Johansson MHK, Munk P, et al. Genomic evolution of antimicrobial resistance in 376 Escherichia coli. Sci Rep 2021;11:15108. https://doi.org/10.1038/s41598-021-93970-7. PMID: 34301966.

CSIphylogeny

- Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the Problem of Comparing Whole 378 Bacterial Genomes across Different Sequencing Platforms

Friedrich A, ed. PLoS One 2014;9:e104984. https://doi.org/10.1371/journal.pone.0104984. PMID: 25110940.

FastQC

- No publication; documentation available online.

FastTree

- Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009;26(7):1641-50. https://doi.org/10.1093/molbev/msp077. PMID: 19377059.

Kleborate

- Lam MMC, Wick RR, Watts SC, et al. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat Commun 2021;12:4188. https://doi.org/10.1038/s41467-021-24448-3. PMID: 34234121.

KmerFinder

- Larsen M V., Cosentino S, Lukjancenko O, et al. Benchmarking of Methods for Genomic Taxonomy. J Clin Microbiol 2014;52:1529–39. https://doi.org/10.1128/JCM.02981-13. PMID: 24574292.

- Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 2018;19:307. https://doi.org/10.1186/s12859-018-2336-6. PMID: 30157759.

KRAKEN

- Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46. PMID: 24580807.

MLST

- Larsen M V., Cosentino S, Rasmussen S, et al. Multilocus sequence typing of total-genome-sequenced 317 bacteria. J Clin Microbiol 2012;50:1355–61. https://doi.org/10.1128/jcm.06094-11. PMID: 22238442.

QUAST

- Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5. https://doi.org/10.1093/bioinformatics/btt086. PMID: 23422339.

ResFinder

- Bortolaia V, Kaas RS, Ruppe E, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother. 2020;75(12):3491-3500. https://doi.org/10.1093/jac/dkaa345. PMID: 32780112.

rMLST

- Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. Microbiology 2012;158:1005–15. https://doi.org/10.1099/mic.0.055459-0. PMID: 22282518.

SPAdes

- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021. PMID: 22506599.

Trimmomatic

- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. https://doi.org/10.1093/bioinformatics/btu170. PMID: 24695404.

Unicycler

- Wick RR, Judd LM, Gorrie CL, et al. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6):e1005595. https://doi.org/10.1371/journal.pcbi.1005595. PMID: 28594827.

Other supporting documentation

This document will be accompanied by supporting documents to aid novel users in becoming familiar with the most relevant WGS concepts and terms. These will include reviews of available bioinformatics tools, scientific and technical background for each step of the workflow, links to other relevant resources which contain more detailed descriptions of certain steps, and related scientific literature.

The suite of supporting documents will be available on the EURGen-RefLabCap website https://www.eurgen-reflabcap.eu/.

[Catalogue number]

Publications Office

doi:[number]