TECHNICAL REPORT

APPROVED: 29 April 2019

doi:10.2903/sp.efsa.2019.EN-1337

# 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

European Centre for Disease Control (ECDC), European Food Safety Authority (EFSA), Ivo Van Walle, Beatriz Guerra, Vitor Borges, João André Carriço, Guy Cochrane, Tim Dallman, Eelco Franz, Renata Karpíšková, Eva Litrup, Michel-Yves Mistou, Stefano Morabito, Joël Mossong, Erik Alm, Federica Barrucci, Chiara Bianchi, Giancarlo Costa, Saara Kotila, Iolanda Mangone, Daniel Palm, Luca Pasinato, Joana Revez, Marc Struelens, Daniel Thomas-López and Valentina Rizzi

## Abstract

EFSA and ECDC were requested by the European Commission to jointly evaluate the possible solutions for the collection and analysis of whole genome sequencing (WGS) data for at least *Listeria monocytogenes*, *Salmonella* and *Escherichia coli* by: (1) analysing the outcome of the surveys on the status of the use of WGS of food-borne pathogens in EU/EEA countries in both the food and public health sectors; (2) conducting a consultation of relevant actors to assess state-of-the-art pipelines for collecting and analysing WGS data in Europe; (3) involving relevant stakeholders to assess the needs and requirements for the analysis of WGS data and their comparability; and (4) preparing a technical report on the identification and comparison of potential solutions for the set-up and running of a joint EFSA–ECDC pipeline to collect and analyse WGS data. Logical components of the overall system were identified and technical requirements were prioritised and grouped according to functionality. Eleven platforms (solutions) that integrate the relevant functionalities for collecting, analysing and visualising WGS data were thoroughly described. The degree to which the requirements are met by the different solutions (as of 31 December 2018) was evaluated. The assessment made clear that no single solution meets all the critical requirements and each solution has significant gaps regarding the unmet critical requirements. Therefore, scenarios consisting of a combination of several solutions were considered. As there may be many suitable scenarios, and the choice among them will depend on strategic or financial elements that are not within the scope of this report, it contains the scientific and technical elements necessary to generate scenarios, rather proposing individual scenarios. A scenario builder is presented which includes the significant gaps for each solution/functionality and the limitations and risks to be considered when setting up a joint ECDC–EFSA database for the collection and analysis of WGS data.



**Key words:** whole genome sequencing, WGS, molecular typing, data collection, public health, food safety

**Requestor:** European Commission

**Question number:** EFSA-Q-2017-00397, ECDC-MR-2017-IN-1094

**Correspondence:** biocontam@efsa.europa.eu, fwd@ecdc.europa.eu

**Members of the ad hoc working group:** Vitor Borges, João André Carriço, Guy Cochrane, Tim Dallman, Eelco Franz, Renata Karpíšková, Eva Litrup, Michel-Yves Mistou, Stefano Morabito and Joël Mossong.

**ECDC staff:** Ivo Van Walle (co-chair, secretariat), Erik Alm, Saara Kotila, Daniel Palm, Joana Revez and Marc Struelens.

**EFSA staff:** Valentina Rizzi (co-chair), Beatriz Guerra (secretariat), Federica Barrucci, Chiara Bianchi, Giancarlo Costa, Luca Pasinato and Daniel Thomas-López.

**Acknowledgements:** ECDC and EFSA wish to thank the following for the internal support provided to this scientific output: Ernesto Liébana, Teresa da Silva Felicio, Mirko Rossi, Gina Cioacata, Henok Ayalew Tegegne and Johanna Takkinen. ECDC and EFSA also wish to acknowledge the hearing experts David Aanensen, Mark Achtman, Clara Amid, Bruno Gonçalves, Dag Harmsen, Keith Jolley, Ole Lund, Tomas Matthews and Hannes Pouseele. Also, ECDC and EFSA wish to thank the members of the EU EURL working group on next generation sequencing, the joint steering committee on the collection and management of molecular typing data from animal, food, feed and the related environment, and human isolates, the ECDC Food- and Waterborne Diseases and Zoonoses Network and the EFSA Zoonoses Monitoring Data Network for their support.

**Suggested citation:** ECDC (European Centre for Disease Prevention and Control), EFSA (European Food Safety Authority), Van Walle I, Guerra B, Borges V, Carriço JA, Cochrane G, Dallman T, Franz E, Karpíšková R, Litrup E, Mistou M-Y, Morabito S, Mossong J, Alm E, Barrucci F, Bianchi C, Costa G, Kotila S, Mangone I, Palm D, Pasinato L, Revez J, Struelens M, Thomas-López D and Rizzi V, 2019. 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. EFSA supporting publication 2019:EN-1337. 92 pp. doi:10.2903/sp.efsa.2019.EN-1337

**ISSN:** 2397-8325

## Summary

EFSA and ECDC were requested by the European Commission to jointly evaluate the possible solutions for the collection and analysis of WGS data for at least *Listeria monocytogenes*, *Salmonella* and *Escherichia coli* by: (1) analysing the outcome of the surveys on the status of use of WGS of food-borne pathogens in the Member States in both the food and public health sectors; (2) conducting a consultation of relevant actors to assess state-of-the-art pipelines for collecting and analysing WGS data in Europe; (3) involving relevant stakeholders to assess the needs and requirements for the analysis of WGS data and their comparability and to describe roles and responsibilities, taking into account that there are different types of WGS data (raw sequence reads, genome assemblies, wgMLST allele identifiers, strain nomenclature, phenotypic predictions), which may require an interface with externally hosted databases and applications; and (4) preparing a technical report on the identification and the comparison of potential solutions for the set-up and running of a joint EFSA–ECDC pipeline to collect and analyse WGS data, taking into account the deliverables set out in the terms of reference (ToRs 1, 2 and 3).

The development of a joint ECDC–EFSA WGS database is essential to ensure the integrated analysis of molecular typing data from food-borne pathogens across different countries and sectors. This project aims to improve crisis preparedness and management in the food and feed area in order to ultimately ensure a more effective and rapid containment of food- and feed-related emergencies and crises in the future.

EFSA and ECDC collected all the information needed to evaluate the possible scenarios for the collection and analysis of WGS data for at least *L. monocytogenes*, *Salmonella* and *E. coli*. Care was taken to make the assessment in each step of the process as objective as possible.

Member States in both the public health and food safety and veterinary sectors were consulted about their level of preparedness on the use of WGS to respond to the challenges posed by threats such as multinational food-borne outbreaks. A joint cross-sector analysis for EU/EEA countries was carried out (ToR 1).

Logical components of the Overall System were identified and technical requirements were prioritised (critical, medium or optional) independently of the existing solutions; they were accurately described and grouped according to functionality. The relevant stakeholders were consulted about them (ToR 3).

Eleven platforms (solutions) that integrate many of the functionalities for collecting, analysing and visualising WGS data or that are widely used in the public health, food and veterinary sectors and the wider scientific community were thoroughly analysed with the support of experts representing the various solutions ('hearing experts'). This analysis was followed by the assessment of these 11 individual solutions against 165 requirements (ToR 2) based on the situation as of 31 December 2018 and to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts.

In the absence of a specific methodology for the combined assessment of the outcomes of ToRs 1–3, in order to propose possible scenarios for the collection and analysis of WGS data, several approaches were discussed. Different methodological approaches were considered and an attempt was made to estimate the remaining work that would be needed in order to meet the critical requirements for each solution and each functionality (i.e. counting the number of critical requirements met per solution and functionality; estimating the complexity of implementing a requirement; quantitatively determining the remaining work for each existing solution per functionality). Eventually, the estimate was made by qualitatively determining the significant gaps (regarding the unmet critical requirements) per solution and per functionality.

The assessment of the individual solutions made clear that every single solution has a substantial number of gaps and none of them meet all the critical requirements. Therefore, scenarios were also considered that would consist of a combination of solutions. An attempt was made to enumerate such possible scenarios. However, there may be many suitable scenarios and the choice among them also depends on other strategic or financial elements that are not under the control of the joint working group that drafted this report.

Therefore, the outcome of the assessment was to set out the elements necessary to generate scenarios rather than to propose individual scenarios. These elements are summarised in the scenario builder that includes the significant gaps for each solution/functionality, as well as the limitations and risks to be considered when building the scenarios.

EFSA and ECDC will use this scenario builder together with other strategic elements to generate and propose suitable scenarios to the European Commission.

# Table of contents

# 1. Introduction

## 1.1. Background and Terms of Reference as provided by the requestor

### Background

Whole Genome Sequencing (WGS) has developed rapidly in recent years. Compared to other molecular typing analyses, WGS showed great potential to improve the ability not only to contribute to the epidemiological investigations of foodborne outbreaks and to the identification of emerging health threats, but also to perform correct identification of bacterial isolates and to identify virulence, antimicrobial resistance and other relevant genes in complex samples.

Member States (MSs) must ensure that the epidemiological investigations of foodborne outbreaks are carried out in accordance with the provisions of Directive 2003/99/EC. The Commission has mandated the European Food Safety Authority (EFSA) to collect directly from the MSs information on zoonoses and zoonotic agents and related antimicrobial resistance.

The European Centre for Disease Prevention and Control (ECDC) must, in coordination with the MSs, establish procedures for systematically searching for collecting, collating and analysing information and data with a view to the identification of emerging health threats which could affect the EU, in accordance with the provisions of Regulation (EC) No 851/2004 and Decision 1082/2013.

A Commission vision paper following the EHEC crisis was endorsed by the MSs in December 2012. Thereafter the Commission asked EFSA to provide technical support regarding the collection of molecular typing of food, feed and animal isolates of *Salmonella*, *Listeria monocytogenes* and Shiga toxin-producing *Escherichia coli*, and a similar request was made in parallel to ECDC on molecular typing data of human isolates. In addition, the Commission asked EFSA and ECDC to establish a joint database for molecular typing data of these foodborne pathogens of human and non-human origin. The joint EFSA-ECDC molecular typing database became functional for both human and non-human isolates at the end of 2015. The current molecular typing database is limited to the collection of PFGE data of *Salmonella*, *L. monocytogenes* and Shiga toxin-producing *E. coli* isolates, and MLVA data for *S.* Typhimurium isolates. As MLVA data produced by a standard method are also available for *S.* Enteritidis, this method-pathogen combination should be added to the scope of the joint database and also for non-human data.

Given the growing importance of WGS analysis in recent multinational foodborne outbreak investigations and in surveillance/monitoring fields, including antimicrobial resistance (AMR), and given the gradual increasing capacity of public health and food laboratories, the development of a WGS database, in the framework of the molecular typing project, is essential to ensure integrated analysis of molecular typing data from foodborne pathogens (across different countries and sectors).

This project is even more important given the focus of the Commission on improving crisis preparedness and management in the food and feed area in order to ultimately ensure a more effective and rapid containment of food and feed-related emergencies and crises in the future. The collection of WGS data would support risk managers to quickly respond to challenges posed by threats such as multinational foodborne outbreaks. Such threats, which may relate to accidental mismanagement within food production processes or even to intentional action such as bio-terrorist attacks, may seriously undermine the established high level of protection for consumers within the single market of the EU and put into question their confidence into the safety of the Overall System.

### Terms of Reference

**EFSA and ECDC are asked**, in accordance with article 31 of Regulation (EC) No 178/2002 and in the framework of the molecular typing project, **to provide technical support for the implementation and management of a database on relevant WGS data from foodborne pathogens isolated from animals, food, feed, food/feed environmental and human samples**. The database could initially be operational for the collection of WGS data from *Salmonella*, *L. monocytogenes* and *E. coli*

isolates, but eventually should be extended to include other foodborne pathogens such as *Campylobacter* and foodborne viruses, upon agreement between EFSA, ECDC, the relevant European Union Reference Laboratories (EURL) and the European Commission.

In particular EFSA and ECDC are requested to jointly evaluate the possible solutions for the collection and the analysis of WGS data for at least *L. monocytogenes*, *Salmonella*, *E. coli* by

(1) Analysing the **outcome of the surveys on the status of use of WGS of foodborne pathogens** in European Member States (MSs) in both food and public health sectors.

(2) Conducting a consultation of relevant actors and players to **assess the state of the art of pipelines** for collecting and analysing WGS data in Europe.

(3) Involving relevant stakeholders to **assess the needs/requirements for the analysis of WGS data and their comparability and to describe roles and responsibilities**, taking into account that there are different types of WGS data (raw sequence reads, genome assemblies, wgMLST allele identifiers, strain nomenclature, phenotypic predictions), which may require interfacing with externally hosted databases and applications.

(4) Preparing **a technical report on the identification and the comparison of potential solutions** for the set-up and running of a joint EFSA-ECDC pipeline for collecting and analysing WGS data, taking into account deliverables of ToRs 1, 2, 3.

The technical report should be delivered by April 2019.

Once the possible solutions have been evaluated, a subsequent mandate will be provided for setting up the chosen solution and starting the data collection and analysis, in line with the terms of reference of the previous mandate ref. Ares(2013)65361 and ref. Ares(2013)65450.

## 1.2. Interpretation of the Terms of Reference

It has been decided that, instead of using the term 'pipeline' for the future WGS system, we will refer to 'Overall System', the components of which will be described later. This is because the system to be designed based on the current report is a combination of different components, each with a distinct set of functionalities including data collection, analysis and visualisation (see Section 3.4). When referring to the current joint ECDC–EFSA molecular typing database, the term 'joint database' will be used (also referred to as the 'current joint database').

The technical report to reply to the above-mentioned mandate includes the identification and comparison of existing software and/or platforms (further referred to as 'solutions') with respect to the possibility to adapt to the needs of an EU system for collecting, analysing and visualising WGS data and the evaluation of the remaining development work for each of them. This assessment will support the definition of possible scenarios for the set-up and running of the Overall System for collecting, storing, sharing, analysing and visualising WGS data. The present mandate is not to already set up the Overall System. Rather, once the technical report is finalised, ECDC and EFSA will provide the EC with some proposed scenarios and additional considerations regarding the strategic and financial elements that will support the selection of the most suitable Overall System.

A subsequent mandate from the EC will be sent to both Agencies for setting up the chosen Overall System, by upgrading the current joint database to a joint system (including a database and application pipeline) for collecting, storing, sharing, analysing and visualising WGS data integrated with epidemiological data and starting the data collection and analysis.

This mandate is relevant from both the public health and food and veterinary points of view.

As described in ECDC's founding regulation (2004/851/EC) and the EC Decision on serious cross-border threats to health (2013/1082/EC), ECDC's tasks include searching for, collecting, evaluating and disseminating data relevant for the prevention and control of human communicable diseases, as well as monitoring, early warning of, and combating serious cross-border health threats. In the detection and verification of multicountry outbreaks, WGS-based methods have quickly become crucial. In the case of food- and waterborne disease outbreak investigations, it is often essential to establish a strong microbiological link between the human cases and the suspected food item. This requires the use of

highly discriminatory WGS-based methods, and subsequent joint data analysis with the sequences from the food, feed, animal or environment isolates.

ECDC's mandate includes other communicable diseases than food- and waterborne diseases. The surveillance and outbreak-related activities of these can in many cases also benefit from the inclusion of WGS data. Therefore, it is possible that ECDC will also use the public health components of the future Overall System for non-food- or waterborne communicable diseases in order to streamline its in-house WGS processes.

MSs must ensure that the epidemiological investigations of food-borne outbreaks are carried out in accordance with the provisions of Directive 2003/99/EC. The Commission has mandated EFSA to collect information on zoonoses and zoonotic agents, related antimicrobial resistance and food-borne outbreaks directly from the MSs.

The purpose of the Overall System was laid down in the background provided within the Commission's mandate according to Directive 2003/99/EC. The main objective would be to contribute to the epidemiological investigations and early detection of food-borne outbreaks and the identification of emerging health threats; however, it could also be used for source attribution and to monitor zoonoses and antimicrobial resistance.

Regarding the different Terms of Reference (ToR):

**ToR1: Outcome of the surveys on the status of use of WGS of foodborne pathogens in MSs in both food and public health sectors:**

Specific reports on the use of WGS in the EU, EEA and EFTA countries for the public health and food and veterinary sectors collected by ECDC and the EC /EFSA, respectively, have recently been published (ECDC, 2018; EFSA, 2018). For the present technical report, only comparable data were analysed. Thus the joint analysis, carried out as described in Sections 2.1.1 and 2.2.1 and presented in Section 3.1, considered only:

– data collected from EU/EEA countries

– data for *L. monocytogenes*, *Salmonella* spp., and Shiga-toxin-producing *E. coli* (STEC)

– for the food safety sector, only data provided by National Reference Laboratories

– for the public health sector, only data provided by National Public Health Reference Laboratories

– data regarding the use of WGS for surveillance and outbreak investigation.

**ToR2: Conducting a consultation of relevant actors and players to assess the state of the art of pipelines for collecting and analysing WGS data in Europe:**

Due to the rapid developments in the area of WGS analysis, the number of individual tools that could be considered would be unlimited. To answer this ToR, the ECDC–EFSA JWG members decided to focus on the assessment and description of available platforms (further referred to as 'solutions') that integrate relevant tools or functionalities for collecting, analysing and visualising WGS data. In addition, these solutions were also already used by organisations with goals similar to those of ECDC and EFSA, and likely to be suitable in terms of technical features for capacity, scalability, sustainability and acceptable ease of use for the purposes of the EU mandate. More information is provided in Sections 2.1.3 and 2.2.3.

**ToR3: Involving relevant stakeholders to assess the needs/requirements for the analysis of WGS data and their comparability and to describe roles and responsibilities, taking into account that there are different types of WGS data (raw sequence reads, genome assemblies, wgMLST allele identifiers, strain nomenclature, phenotypic predictions), which may require interfacing with externally hosted databases and applications**:

The stakeholders identified in the context of the present mandate are described in Section 2.2.2. The EU EURL working group on Next Generation Sequencing (NGS), the steering committee for the current joint database, the ECDC Food- and Waterborne Diseases and Zoonoses Network and the EFSA Zoonoses Monitoring Data Network were informed of the developments of the joint working group discussions. Data provided by public health laboratories in the MSs, existing collaboration agreement

for ECDC-EFSA Molecular Typing Database were considered for the discussions as described in Section 2.1.2.

## 1.3. Objectives of the data collection

The broad purpose of the current joint database is laid down in the background provided in the EC mandate to ECDC and EFSA requesting technical support on the collection of data on molecular testing in human and food/animal isolates of food-borne infections (Ref. Ares(2013)65450 and Ares(2013)65361). The purpose is to encourage the collation of data on molecular testing so that linkage of molecular typing data from humans to similar types of data from food and animals is possible. The final aim is to enable the rapid detection of clusters and outbreaks and suggest links with potential sources. Such linking allows faster epidemiological investigations, a better evaluation of the importance of certain foods and animal populations as sources of food-borne infections and outbreaks through case-by-case source attribution studies, and a better understanding of the ecology of food-borne infections. This purpose does not change with the advent of WGS as a new and more accurate typing method.

Specific objectives for the Overall System:

- Ensuring collection of comparable WGS data and performing standardised analyses on sequences from isolates from both the public health and food and veterinary sectors.
- Early detection of multicountry clusters of human cases.
- Generation of hypotheses about food vehicle and source of contamination to guide epidemiological investigations.
- Verification of the food vehicle and source of contamination.
- Verification of effectiveness of control measures.
- Source attribution.
- Detection of antimicrobial resistance (AMR) traits (future use for AMR monitoring).

## 1.4. Additional information

As discussed with the requestors (EC), the tasks and activities carried out to reply to the present mandate should start with and take advantage of the work done for the design and implementation of the current joint ECDC-EFSA Molecular Typing Database (EFSA, 2014a; Rizzi et al., 2017). In particular, the extensive discussions and agreement on the issues related to data sharing and visibility should be considered when identifying the potential solutions for the Overall System, as well as the roles of different users.

The current joint database collects molecular typing data produced with PFGE and MLVA methods for *Salmonella*, *L. monocytogenes* and STEC strains, and is physically hosted at ECDC, as part of The European Surveillance System (TESSy). Data are validated upon submission to the database, and PFGE profiles are curated by ECDC's contractors and the EURLs for the three pathogens for human and non-human isolates, respectively. Cluster analysis is performed only on isolates meeting the minimum requirements. The system offers the possibility for the data providers to consult, through the TESSy web interface, the joint database according to the differentiated access rights. The system also offers data providers the possibility of downloading the results of the curation process for their own isolates. Detailed information about the joint ECDC-EFSA Molecular Typing Database, in particular about its architecture, functionalities and uses, is available in Section 3.2.

To guarantee the confidentiality of data for the respective data owners, the microbiological information in the joint database is accompanied by a minimum subset of epidemiological data. Additionally, different rights for data accessibility are associated with each type of user, with particular restrictions for 'sensitive' data (i.e. country of sampling, laboratory identification code). These aspects will remain unchanged with the upgrade of the joint database to include WGS as a typing method.

In 2017, the PulseNet International network, which groups together public health organisations from around the world with respect to food- and waterborne diseases, published a 'Vision for the implementation of whole genome sequencing for global food-borne disease surveillance' (Nadon et al., 2017). As part of that vision, whole genome multilocus sequence typing (wgMLST) was put forward as

the primary method for typing based on WGS data. The upgraded joint database must therefore have the possibility to perform wgMLST analysis. The alternative to this analysis, based on single nucleotide polymorphisms or variants (SNPs/SNVs), should ideally be possible as well.

The ECDC Strategic Framework for integration of molecular and genomic data for EU surveillance and cross-border outbreak investigations 2019–2021' prioritises pathogens/diseases and outlines technical implementation options for the medium-term integration of molecular/genomic typing information into EU-level surveillance and multicountry outbreak investigations (ECDC, 2019). In the broad area of food- and waterborne diseases, it is foreseen that ECDC will support multicountry outbreak investigations enhanced through (whole genome or gene) sequence-based microbial typing data-sharing and analysis for *S. enterica*, STEC, *L. monocytogenes, Campylobacter jejuni/coli*, hepatitis A virus, *Legionella* spp., emerging multi- or extensively drug-resistant bacteria, outbreaks of new pathogens or those with new modes of transmission. Furthermore, the framework prioritises the gradual development of EU-wide sequence-based continuous surveillance for detection and delineation of multicountry outbreaks of *S. enterica* and STEC, while for *L. monocytogenes* this has been operational since March 2019. As part of these pilot surveillance projects, ECDC will, together with microbiology experts from EU surveillance networks and EUCAST[1] antimicrobial resistance experts, assess the technical specifications for genetic-based AMR prediction and molecular mechanisms of resistance identification for priority human pathogens. They will develop guidelines and protocols for further harmonised WGS-based public health surveillance of AMR for priority human pathogens in accordance with the Commission Implementing Decision (EU) 2018/945[2].

To enable these operations, ECDC is developing a set of technical solutions available in-house or externally provided for sharing, storing and analysing WGS typing data and the visualisation of integrated genomic and epidemiological data analysis outputs for public health risk assessment. These solutions encompass the following functionalities: data providers submit sequences and descriptive data about sequences and epidemiological data in a timely fashion using an easy-to-use solution for machine-to-machine communication. The WGS data submission process should ensure that every isolate submitted to the WGS solution has corresponding epidemiological data in TESSy. Based on MSs policies for public WGS data release, they can be uploaded to a protected, access-controlled, and reliable long-term storage solution. The WGS data are made public after an initial embargo period, but only if the data provider decides that they should be. Data are analysed, signals are detected, and visualisations are produced with a high level of automation.

In 2017, EFSA received a mandate from the EC for 'scientific and technical assistance on harmonised monitoring of AMR in bacteria transmitted through food'[3]. The mandate asked EFSA to review and update the technical specifications for AMR monitoring. Specifically, EFSA was asked to address the possible use of molecular typing methods (e.g. WGS) in the light of the latest scientific opinions on AMR, technological developments, recent trends in AMR and relevance for public health, as well as the audits assessing the implementation of Decision 2013/652/EU[4] performed by the EC in a number of MSs. The report will include proposals for implementing updated guidelines for further harmonised monitoring of AMR in food-producing animals and food and for providing continuity in following up further trends in AMR to possibly underpin the revision of existing legislation. Recommendations on the areas where WGS could or should be used by the MSs to complement or replace the phenotypic methods currently in use and the relative timelines are included (EFSA, 2019a).

EFSA is also currently working on a self-task mandate from its Panel on Biological Hazards to produce a scientific opinion on the 'application and use of NGS (including WGS) for risk assessment of food-borne microorganisms'. The ToRs for this mandate are: i) evaluate the possible use of NGS (e.g. WGS and metagenomic strategies) in food-borne outbreak detection and investigation, and hazard identification (e.g. generation of data on virulence and AMR genes, plasmid typing) based on the outcomes of the ongoing WGS outsourcing activities, experience from different countries and

[1] European Committee on Antimicrobial Susceptibility Testing.
[2] Commission Implementing Decision (EU) 2018/945 of 22 June 2018 on the communicable diseases and related special health issues to be covered by epidemiological surveillance as well as relevant case definitions. OJ L 170, 6.7.2018, p. 1–74.
[3] http://registerofquestions.efsa.europa.eu/roqFrontend/questionsListLoader?mandate=M-2018-0010
[4] 2013/652/EU: Commission Implementing Decision of 12 November 2013 on the monitoring and reporting of antimicrobial resistance in zoonotic and commensal bacteria (notified under document C(2013) 7145). OJ L 303, 14.11.2013, p. 26–39.

underlining the added value for risk assessment; and ii) critically analyse the advantages, disadvantages and limitations of existing NGS-based methodologies (including WGS) as compared with microbiological methods cited in the current EU food legislation (e.g. *Salmonella* serotyping, STEC monitoring, AMR testing), taking into account benchmarking exercises (deadline of 30 November 2019).

The present EC mandate has been addressed by EFSA within the activities of its WGS Umbrella Project which is one of the projects included in the EFSA Information Management Programme (EFSA, 2019b).

# 2. Data and methodologies

The ToRs are presented according to the order in which they were addressed by the JWG.

## 2.1. Data

### 2.1.1. WGS capacity surveys (ToR 1)

The data on the use of WGS in EU/EEA countries for the public health and the food and veterinary sectors were collected by ECDC and EC/EFSA, respectively, and analysed jointly for the present report. These are the latest data available as of August 2018. The public health sector data present the situation as of July 2017, while for the food and veterinary sector the situation as of December 2016 is presented.

**Public health sector survey**

The ECDC National Focal Points for Microbiology (NMFP) from 30 EU/EEA countries were asked about the capacity of the competent public health reference laboratories to use WGS-based typing methods for surveillance applications as of July 2017. The survey comprised 10 questions (for reference, see https://ec.europa.eu/eusurvey/runner/WGS_2017) and was opened from 16 October 2017 to 8 December 2017. Two reminders were sent for the data collection and NMFP received a draft of the data analysis for validation and clearance. The resulting report 'European Centre for Disease Prevention and Control. Monitoring the use of Whole-Genome Sequencing for infectious diseases surveillance in the European Union and European Economic Area, 2015-2017' was published in September 2018 (ECDC, 2018).

**Food/veterinary sector survey**

The survey was jointly developed by the EC (Directorate-General for Health and Food Safety, DG SANTE, G4) and EFSA, with the support of the EURLs. It was adressed to EURLs, National Reference Laboratories (NRLs), and the official laboratories of EU/EFTA countries (EEA and Switzerland). The questionnaire covered the following Networks: *Salmonella* spp., *E. coli* including STEC, *Campylobacter* spp., *L. monocytogenes*, Coagulase positive Staphylococci including *S. aureus* and viral and bacterial pathogens of live bivalve molluscs as well as Antimicrobial Resistance. The survey comprised 15 questions and was open between September and December 2016 (using the online survey tool EUSurvey, https://ec.europa.eu/eusurvey/). Respondents were further contacted by EFSA to provide clarifications. The report 'Outcome of EC/EFSA questionnaire (2016) on use of Whole Genome Sequencing (WGS) for food- and waterborne pathogens isolated from animals, food, feed and related environmental samples in EU/EFTA countries' was published in June 2018 (EFSA, 2018).

### 2.1.2. Assessment of the needs and requirements for the analysis of WGS data (ToR 3)

In assessing the specific requirements in line with the scope of the mandate, the following were taken into account:

- EU General Data Protection Regulation (GDPR)[5]
- Collaboration Agreement between ECDC, EFSA and the EURLs

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1–88.

- Joint ECDC–EFSA Molecular Typing Database Report (EFSA, 2014a)
- EC vision paper 2013 (EC, 2012).

### 2.1.3. Assessment of the state-of-the-art of pipelines (ToR 2)

Publicly available information on the various solutions was retrieved from the following sources for the individual solutions and/or references:

BIGSdb: https://pubmlst.org/software/database/bigsdb/; https://bigsdb.readthedocs.io/en/latest/; https://github.com/kjolley/BIGSdb

BioNumerics: http://www.applied-maths.com/bionumerics

CGE: https://cge.cbs.dtu.dk/services/

COMPARE: Amid et al. (2019); http://www.compare-europe.eu; https://compare.cbs.dtu.dk; https://www.ebi.ac.uk/ena/pathogens/home; https://www.biorxiv.org/content/10.1101/555938v1

European Nucleotide Archive: Harrison et al. (2019)

https://www.ebi.ac.uk/en; https://enadocs.readthedocs.io/en/latest/meta_01.html?highlight=metadata%20model; http://europepmc.org/abstract/MED/30395270.

EnteroBase: https://enterobase.warwick.ac.uk; https://enterobase.readthedocs.io/en/latest/enterobase-terms-of-use.html; https://bitbucket.org/enterobase/enterobase-web; https://github.com/achtman-lab/GrapeTree; https://github.com/zheminzhou/EToKi; https://github.com/EnteroBaseGroup/EnteroBase

INNUENDO: Llarena et al. (2018);

https://github.com/TheInnuendoProject; https://github.com/B-UMMI

IRIDA: https://www.irida.ca, https://github.com/phac-nml/irida; https://irida.corefacility.ca/documentation/downloads/index.html

PathogenWatch: https://pathogen.watch

Seqsphere: https://www.ridom.de/seqsphere/

Information was also provided by hearing experts through PowerPoint presentations on the solutions, written answers to questions drafted by the JWG on each solution and additional evidence provided for the requirements' assessment.

## 2.2. Methodologies

### 2.2.1. WGS capacity surveys (ToR 1)

The reports mentioned above (ECDC, 2018; EFSA, 2018) present detailed information on the survey questionnaires and respective analyses that covered a broader scope than the one presented in the present report. For the joint cross-sector analysis, only those answers received from EU/EEA countries were considered. For the food and veterinary sector, the data presented are only those regarding the WGS capability of the NRLs for surveillance and/or outbreak investigation.

### 2.2.2. Assessment of the needs and requirements for the analysis of WGS data (ToR 3)

A classical business analysis approach was adopted to address this item. Stakeholders were identified, a number of whom were consulted on their needs to support the finalisation of requirements. To gather the most complete set of needs, some techniques have been applied such as survey and document analysis (BABOK – see IIBA, 2015). Some needs were considered *conditio sine qua non* and were therefore put as constraints.

First, high-level constraints on the envisaged Overall System were defined. Then, the logical components were identified. Finally, requirements were identified based on the data flow between the logical components as well as user interaction with relevant components and taking into account the operational systems and agreements between relevant actors at EU level. The requirements were broken down and formulated in such a way that each requirement covers a single functionality that can be either met or not met.

Each requirement was assigned a priority of 'critical', 'medium' or 'optional'. All critical requirements should be met by the Overall System, whereas medium and optional requirements contribute to the additional usability of the system (Cechich et al., 2003).

In order to get feedback on the above approach, relevant stakeholders were consulted; in particular, the following consultations were carried out:

- MSs' food safety and veterinary authorities from the MSs: EFSA's Scientific Network for Zoonoses Monitoring Data.
- MSs' public health authorities from the MSs: ECDC's Food- and Waterborne Diseases and Zoonoses Network (FWD-Net), National Microbiology Focal Points and National Surveillance Focal Points.
- The joint ECDC–EFSA Steering Committee for molecular typing data collection.
- The EURL WG on WGS.

### 2.2.3. Assessment of the state-of-the-art of pipelines (ToR 2)

The JWG identified which solutions would be assessed according to the approach described in Section 1.2 'Interpretation of the Terms of Reference'.

The members of the JWG provided a summary description of each solution as well as an initial assessment of how well each solution met each individual requirement, based on their personal experience (Expert Knowledge) or narrative literature reviews (public documents). Input from EFSA and ECDC experience on certain solutions, including participation in pilot studies (e.g. COMPARE and INNUENDO), was shared with the JWG. For each solution, requirements were assessed as 'met', 'not met' or 'unknown' (unclear to the JWG whether met or not). The deadline for meeting requirements was set at 31 December 2018.

In order to validate the assessment of the requirements for each solution, hearing experts representing different existing solutions were individually invited by the JWG. The process included a phase of preliminary assessment for each solution (from June to November 2018; see below, steps 1–3), and a second phase of validation of the assessment (from December 2018 to February 2019; see below, steps 4–7). The following steps were taken to make sure the experts at the hearing were able to fully provide their input:

1) The JWG drafted technical questions for the hearing experts on each requirement for which it was unclear whether it was met or not met by the corresponding solution. For consistency, the same questions were asked for every requirement, irrespective of the solution (June–November 2018).
2) These questions were shared with the hearing experts shortly before they were invited to the JWG meetings, as described in step 3. Hearing experts provided their answers in writing. The JWG reviewed their answers and formulated follow-up technical questions where needed (June–November 2018).
3) Hearing experts were invited individually to JWG meetings to briefly present their respective solutions and to answer the follow-up questions. The hearing experts' answers were written down on a shared screen during the hearing (June–November 2018).
4) Each hearing expert was given both the initial and follow-up questions and their answers for review and approval in order to correct any inaccuracies, update the assessment according to

the status at the end of 2018, and answer any follow-up questions that could not be answered during the hearing (December 2018–January 2019).

5) The JWG re-assessed the requirements per solution based on the hearing experts' approved answers (January 2019).

6) The updated overall assessment for each solution was provided to the individual hearing experts to allow them to provide their own assessment in terms of which requirements were met or not met. In the event of disagreement, the hearing experts were asked to provide corresponding additional evidence (end of January–February 2019).

7) The JWG again re-assessed the requirements per solution based on the hearing experts' own assessment and additional evidence. This constituted the final assessment (end of February 2019).

### 2.2.4. Identification and comparison of potential solutions (ToR 4)

The outcomes of the previous three ToRs were evaluated and collated in terms of requirements met and remaining work.

In order to estimate to what degree the requirements were met by the different solutions, they were categorised according to four main functionalities (low-resolution approach):

- Data Collection
- Data Analysis
- General user interaction and outputs
- Infrastructure.

In a high-resolution approach, more detailed functionalities were identified, and data analysis was split up further. In total, eight types of functionality were identified:

- Data Collection. Requirements were further subcategorised here, but eventually treated as one type of functionality.
- Sequence read data quality.
- Genome assembly.
- Inferring phylogenetic relationships. Requirements were further subcategorised here, but only the wgMLST subcategory was used further since it was the only one with critical requirements.
- Strain nomenclature.
- Genome characterisation.
- General user interaction and outputs.
- Infrastructure.

In the first instance, several approaches to evaluating the remaining work per individual solution and per functionality were considered:

1) Counting the number of critical requirements that are met.

2) Evaluating the complexity of requirement implementation, independent of any specific solution or corresponding design, as a proxy for the amount of work required in the event that it is not met.

3) Quantitatively estimating the remaining work for each existing solution per functionality.

4) Qualitatively determining significant gaps based only on unmet critical requirements, i.e. where substantial work would still be needed for each existing solution and per functionality.

The JWG then observed that no individual existing solution was able to comply with all critical requirements. Therefore, to satisfy all or almost all of the critical requirements, the overall solutions (further referred to as 'scenarios') would need to be a combination of several existing individual solutions. Any remaining critical requirements not met by such scenarios are defined as gaps that require

additional development or set-up work to close them. Medium and optional requirements were used to identify further advantages or the added value of each identified scenario.

At this scenario level, any work needed to integrate solutions, in the event that more than one is used, also has to be estimated and counted towards the total remaining work. No methodology was put forward to estimate this work because it depends on a very detailed analysis of the solutions, with many options for integrating them, to be done once an initial selection of scenario(s) has been made.

Risks, i.e. possible events with associated negative impacts, also apply to the scenario level. These were enumerated independently from any scenario and qualitatively assessed in terms of the severity of the impact (low, medium, high, very high) and possible mitigations.

Finally, there was a discussion on how all the possible scenarios are limited by the constraints of the Overall System. That is, the choice of a particular solution for a particular functionality may imply that one or more of the constraints is not met, and all these cases were listed.

All these methodological approaches were attempted in order to reach a comprehensive evaluation of the outcomes of the previous ToRs. But not all of them were feasible or successful. The detailed description of the work done, including the abandoned approaches and those that were eventually used for the final assessment, is included in Section 3.8.

The design of the envisaged Overall System does not depend solely on the technical characteristics of the combined solutions and the technical needs of the data collection system that have been compiled and thoroughly assessed by the JWG. It also depends on other strategic elements that will be identified by ECDC and EFSA. These elements could include, besides the obvious financial implications, for example:

1) Scientific fitness for purpose of the Overall System.
2) Ability of the system to keep pace with the development of methodologies in WGS.
3) Ability to use the system by ECDC for other diseases than food- and waterborne diseases.
4) Adherence to architecture best practice (business architecture and separation of concerns).
5) Openness of processes and codes for the analysis: publicly documented or available algorithms/schemas.
6) Feasibility and required resources from ECDC and EFSA to integrate the solution with existing and planned WGS-related systems.

# 3. Assessment

This section summarises the outcome of the different ToRs as set out in the mandate.

## 3.1. Cross-sector analysis of surveys on the use of WGS for food-borne pathogens in the MSs

As already explained in Section 2.1.1, data on the use of WGS in EU/EEA countries for the public health and the food and veterinary sectors were collected by ECDC (ECDC, 2018) and EC/EFSA (EFSA, 2018), respectively, and analysed jointly for the present report. The public health sector data presented the situation as of July 2017, while for the food and veterinary sector the situation as of December 2016 was presented.

A total of 30 EU/EEA countries replied to the public health sector and 29 to the food and veterinary sector surveys (no answer was received from Lithuania by EC/EFSA). The overall capability to use WGS-based typing for analysing food-borne pathogens is presented below (Table 1 and Figure 1), showing that *L. monocytogenes* was the most frequently used WGS-based typed pathogen for surveillance and outbreak investigation in both sectors. The overall capacity for WGS-based typing for either outbreak investigation, surveillance or both, was 20 out of 30 EU/EEA countries and 14 out of 29 EU/EEA countries, in the public health and food and veterinary sectors, respectively.

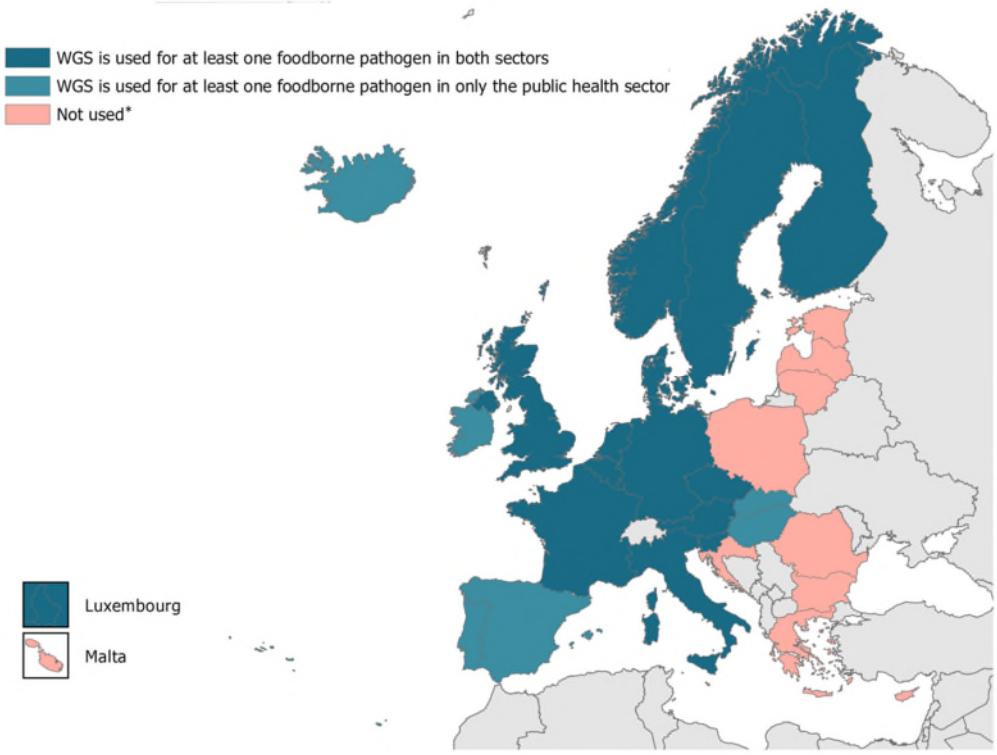
For the public health sector, the number of EU/EEA countries using WGS-based typing for surveillance and/or outbreak investigations was 18, 16 and 15 for *L. monocytogenes, S. enterica* and STEC,

respectively (Table 1). For the food and veterinary sector these numbers were lower, with 13 countries using it for *L. monocytogenes*, 9 for *S. enterica* and 9 for STEC.

**Table 1:** Number of EU/EEA countries with capacity to use WGS-based typing for surveillance and/or outbreak investigations and data completeness, by sector

| Food-borne pathogen and scope of the WGS-based typing use | Public health sector (30 countries) | Food/veterinary sector (29 countries*) |
|---|---|---|
| *Listeria monocytogenes* | | |
| Surveillance and outbreak investigations | 15 | 7 |
| Outbreak investigations only | 3 | 6 |
| No WGS-based typing capacity for surveillance or outbreak investigations | 12 | 12 |
| Data not available | 0 | 4 |
| *Salmonella enterica* | | |
| Surveillance and outbreak investigations | 7 | 5 |
| Outbreak investigations only | 9 | 4 |
| No WGS-based typing capacity for surveillance or outbreak investigations | 14 | 16 |
| Data not available | 0 | 4 |
| STEC | | |
| Surveillance and outbreak investigations | 10 | 5 |
| Outbreak investigations only | 5 | 4 |
| No WGS-based typing capacity for surveillance or outbreak investigations | 15 | 13 |
| Data not available | 0 | 7 |

*No reply was received from Lithuania.




*For Lithuania, this applies only to the public health sector as no reply to the EC/EFSA questionnaire was received.

**Figure 1:** Use of WGS-based typing by EU/EEA countries for surveillance and/or outbreak investigations, for at least one food-borne pathogen queried (*L. monocytogenes*, *S. enterica* and STEC; status December 2016 for the food sector, July 2017 for the public health sector)

Within the public health sector, a total of 8, 10 and 8 EU/EEA countries planned to implement WGS-based typing activities by the end of 2019 for *L. monocytogenes*, *S. enterica* and STEC, respectively. Regarding the food and veterinary sector, 7, 9 and 9 EU/EEA countries planned to implement it by 2019 for *L. monocytogenes*, *S. enterica* and STEC, respectively. Lack of funding, staff and expertise were the main reasons why the reference laboratories in these countries had not implemented WGS-based typing in their routine activities in both sectors, along with the fact that it was not required by legislation in the MS in question.

Regarding the laboratory and bioinformatic procedures it was observed that Illumina sequencers, mainly MiSeq series instruments, were the most frequently used in both sectors (ECDC, 2018). These data are in accordance with those collected in previous surveys (Revez et al., 2017).

The WGS-based bioinformatic analyses performed in the public health and food and veterinary sectors can be found in Table 2. The cgMLST, MLST and SNP analyses were the most widely used for typing *L. monocytogenes* in both sectors, together with serotyping for the public health sector. For *S. enterica* typing, the most frequently used bioinformatics analyses were *in silico* serotyping, MLST, SNP analysis, and detection of AMR genes in the food and veterinary sector, while in the public health sector cgMLST was also frequently used. The bioinformatic analysis most frequently used for STEC included the detection of genes associated with resistance (resistome), virulence (virulome) and genes implicated in horizontal gene transfer (mobilome), along with the *in silico* serotyping and MLST in both sectors.

Regarding the data storage, in the food and veterinary sector almost all NRLs stored the raw data exclusively in-house (locally or on a server of the institution) or in a cloud by outsourcing, a trend also observed in the public health sector with dedicated closed databases being the choice for the food-borne pathogens queried. Only in a very few cases was storage in public repositories reported.

**Table 2:** Number of EU/EEA countries using WGS-based typing for surveillance and/or outbreak investigations, by sector (status July 2017 for the public health sector, December 2016 for the food sector)

| **Applications** | **Public health sector** | | | **Food and veterinary sector** | | |
|---|---|---|---|---|---|---|
| **A. Surveillance and outbreak investigations** | *L. monocytogenes* (n=15) | *S. enterica* (n=7) | STEC (n=10) | *L. monocytogenes* (n=7) | *S. enterica* (n=5) | STEC (n=5) |
| cgMLST | 12 | 6 | 5 | 3 | 2 | 2 |
| SNP | 7 | 5 | 5 | 5 | 4 | 3 |
| Resistome prediction | 4 | 5 | 7 | 1 | 4 | 5 |
| wgMLST | 5 | 3 | 3 | 3 | - | |
| Virulome/mobilome prediction | 4 | 2 | 9 | 1 | - | 4 |
| MLST prediction | 12 | 6 | 8 | 4 | 3 | 3 |
| Serogroup/serotype prediction | 7 | 6 | 9 | 2 | 3 | 3 |
| Other(s) | 1 | 2 | 3 | 1 | 2 | 3 |
| **B. Outbreak investigations only** | *L. monocytogenes* (n=3) | *S. enterica* (n=9) | STEC (n=5) | *L. monocytogenes* (n=6) | *S. enterica* (n=4) | STEC (n=4) |
| cgMLST | 2 | 6 | 2 | 3 | - | 1 |
| SNP | 1 | 2 | 3 | 1 | 1 | 2 |
| Resistome prediction | - | 2 | 1 | 1 | 2 | 3 |
| wgMLST | - | 1 | 1 | 2 | 2 | 1 |
| Virulome/mobilome prediction | - | 2 | 2 | - | - | 4 |
| MLST prediction | 1 | 5 | 2 | 1 | 1 | 3 |
| Serogroup/serotype prediction | 1 | 2 | 4 | 1 | 3 | 4 |
| Other(s) | 1 | 2 | 1 | 1 | - | - |

MLST: multilocus sequence typing; cgMLST: core genome MLST; wgMLST: whole genome MLST; SNP: single nucleotide polymorphism.

## 3.2. Current joint ECDC–EFSA Molecular Typing Database

### 3.2.1. Architectural structure

The current joint ECDC–EFSA Molecular Typing Database collects molecular typing data produced with PFGE and MLVA methods for *Salmonella*, *L. monocytogenes* and STEC strains, and is physically hosted at ECDC as part of TESSy. The architecture of the current Joint Database is described here within the context of the existing data collection systems at EFSA and ECDC.

Typing data and descriptive epidemiological data on human strains are submitted to TESSy by public health authorities and laboratories in the MSs. Similar data from food, feed, animal and environmental samples (non-human data) are reported to the EFSA molecular typing data collection system ('EFSA database') by the food and veterinary authorities and laboratories (NRLs or other official laboratories). The EFSA database interfaces with and submits data to the joint database through TESSy (Figure 2).

An open standard, implemented in both XML and CSV formats, is used for data submission, so that data providers are free to use any software for compiling and providing the data. For PFGE data, there are no open standards available for describing image interpretation results (e.g. lane, band definitions, etc.). Here, the format implemented by the BioNumerics software is used, and the image interpretation data are included as a whole as the value of a single variable in the open XML/CSV standard. At present, ECDC has stopped actively collecting PFGE data from human isolates, although it is technically still possible to submit them.

EFSA's molecular typing data collection system includes the Data Collection Framework system software to collect data in XML format, a BioNumerics server and a database to store the information.

### 3.2.2. Data sharing and accessibility

To guarantee data confidentiality for the respective data owners, in the joint database the microbiological information (PFGE and MLVA typing data as well as serotype/serogroup) are accompanied by a minimum subset of epidemiological data; for example, for non-human isolates only the sample type category (i.e. food, feed, animal or environmental sample) is shared, but not the exact food type. All other human epidemiological descriptive data are stored in the same system (TESSy) but are not part of the Joint Database. All other non-human epidemiological descriptive data are stored in the EFSA database.

Additionally, nominated users are given credentials for accessing the database, but different rights for data accessibility are associated with each type of user. In particular, restrictions apply to 'sensitive' data that are visible only to the respective data providers and to all nominated authorised users from the same country. Data managers and data curators have access to all the data present in the joint database (Table 3).

To further protect the confidentiality of data, a collaboration agreement has been signed between the main actors in the database (ECDC, EFSA and EFSA's curators). In addition, to avoid any improper or non-authorised use of the data, all data providers are asked to sign an agreement with EFSA or ECDC, based on their area of competence, before any data submission or access to the database.

### 3.2.3. Data validation and analysis

The first data validation is performed upon submission to the database, both on the descriptive data and the molecular typing data (e.g. presence of reference lanes for PFGE, correct chronological relationship between dates, etc.). After passing the validation process, the data are inserted into a BioNumerics database that is part of TESSy and the Joint Database. Here, the PFGE profiles are curated and classified as either 'accepted' or 'rejected'. If the profile is accepted, a standardised sequential reference type is assigned to each indistinguishable PFGE pattern (zero bands difference is used as a criterion). The nomenclature of the reference types follows the TESSy nomenclature. There is a monthly reassessment of PFGE profiles and corresponding types across the entire database in order to guarantee consistency between the type assignment for human and non-human isolates.

The curation of human isolates is performed for non-human isolates by ECDC and by the EURLs for the three pathogens. There is no certification or accreditation for the curation process or for clustering.

Cluster analysis is performed and if clusters are found (manual curation), FWD-EPIS (Epidemic Intelligence Information System) generates and sends a weekly email to ECDC's FWD-EPIS users. Any scientific analyses (i.e. cluster detection) are only performed on isolates meeting the minimum requirements (PFGE curation).

### 3.2.4. Services for data providers

The system offers the possibility for the data providers to consult, through the TESSy web interface, the joint database according to their differentiated access rights. The system also offers data providers the possibility of downloading the results of the curation process, i.e. whether their molecular typing data were accepted and what reference types were assigned and, in this way, to synchronise their database with the joint database at the EU level.

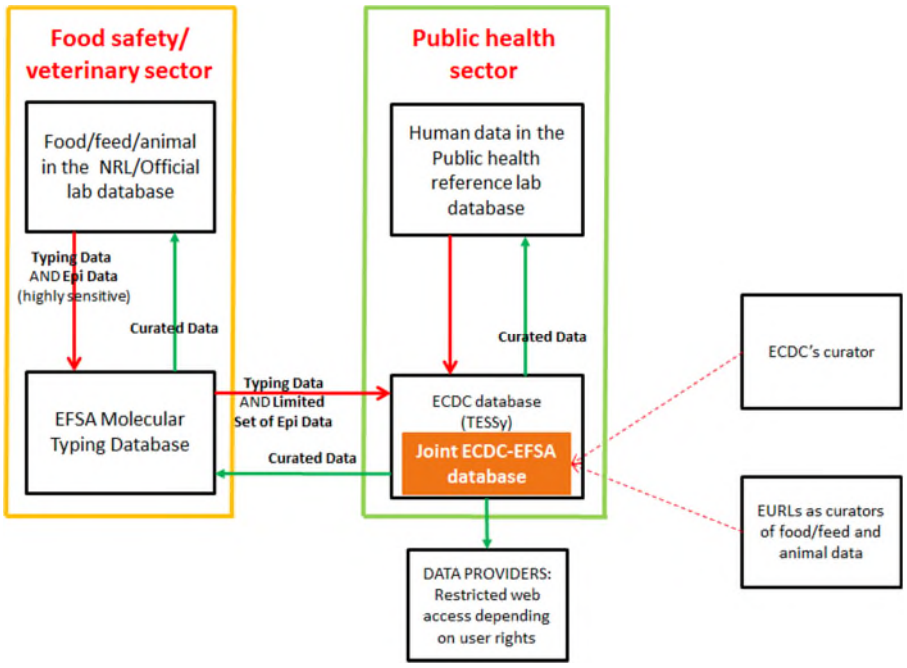



**Figure 2:** Structure of the current joint ECDC–EFSA Molecular Typing Database

Figure 2 presents an overview of the overall logical architecture and describes the main entities (systems or user groups) involved in the context of the current Joint Database, as well as the flow of the exchanged information.

**Table 3:** Data visibility in the current joint ECDC–EFSA Molecular Typing Database

| User group[a] | Non-human data (food, feed, animal, environmental data) | | | | Human data | | | |
|---|---|---|---|---|---|---|---|---|
| | **Country of sampling, laboratory identification code** | **Date of sampling/ sample type** | **Microbiological data[b]** | **Food, feed, animal or environmental descriptive data[c]** | **Country of sampling** | **Date of sampling/ sample type** | **Microbiological data[b]** | **Human descriptive data[d]** |
| EFSA | Yes | Yes | Yes | No (not in Joint Database) | Yes | Yes | Yes | No |
| ECDC | Yes | Yes | Yes | No (not in Joint Database) | Yes | Yes | Yes | Yes[e] |
| Users from MS food/veterinary side | Only if isolate is from the same country as the user | Yes | Yes | No (not in Joint Database) | Only if isolate is from the same country as the user | Yes | Yes | No |
| Users from MS human side | Only if isolate is from the same country as the user | Yes | Yes | No (not in Joint Database) | Yes | Yes | Yes | Yes[e] |
| Curators non-human data | Yes | Yes | Yes | No (not in Joint Database) | Yes | Yes | Yes | No |
| Curators human data | Yes | Yes | Yes | No (not in Joint Database) | Yes | Yes | Yes | Yes[e] |

MS: Member State.
(a): The EC has the right to receive upon request any data related to a specific event.
(b): PFGE and MLVA typing as well as serotype/serogroup.
(c): Detailed description of the sample, e.g. food category/animal population, origin. These are considered to be sensitive data and are not part of the Joint Database.
(d): More information on the patient, e.g. age, gender. These are considered to be sensitive data.
(e): These data are stored physically in the same system (TESSy) but are conceptually not part of the Joint Database.

## 3.3. Envisaged data flow and logical components of the Overall System

The envisaged system covers both the existing functionality, possibly re-implemented, and the new functionality for WGS. It may consist of physical components provided by different solution providers. In order to clearly describe the constraints and requirements for this envisaged system, it is necessary to break it up into separate logical components, each with a distinct set of functionalities. During the design phase, different physical implementations of these logical components, either by existing solutions or new designs, were assessed. There may be more than one physical component needed for one logical component, and vice versa, it is also possible that one physical component implements more than one logical component, or only parts of them.

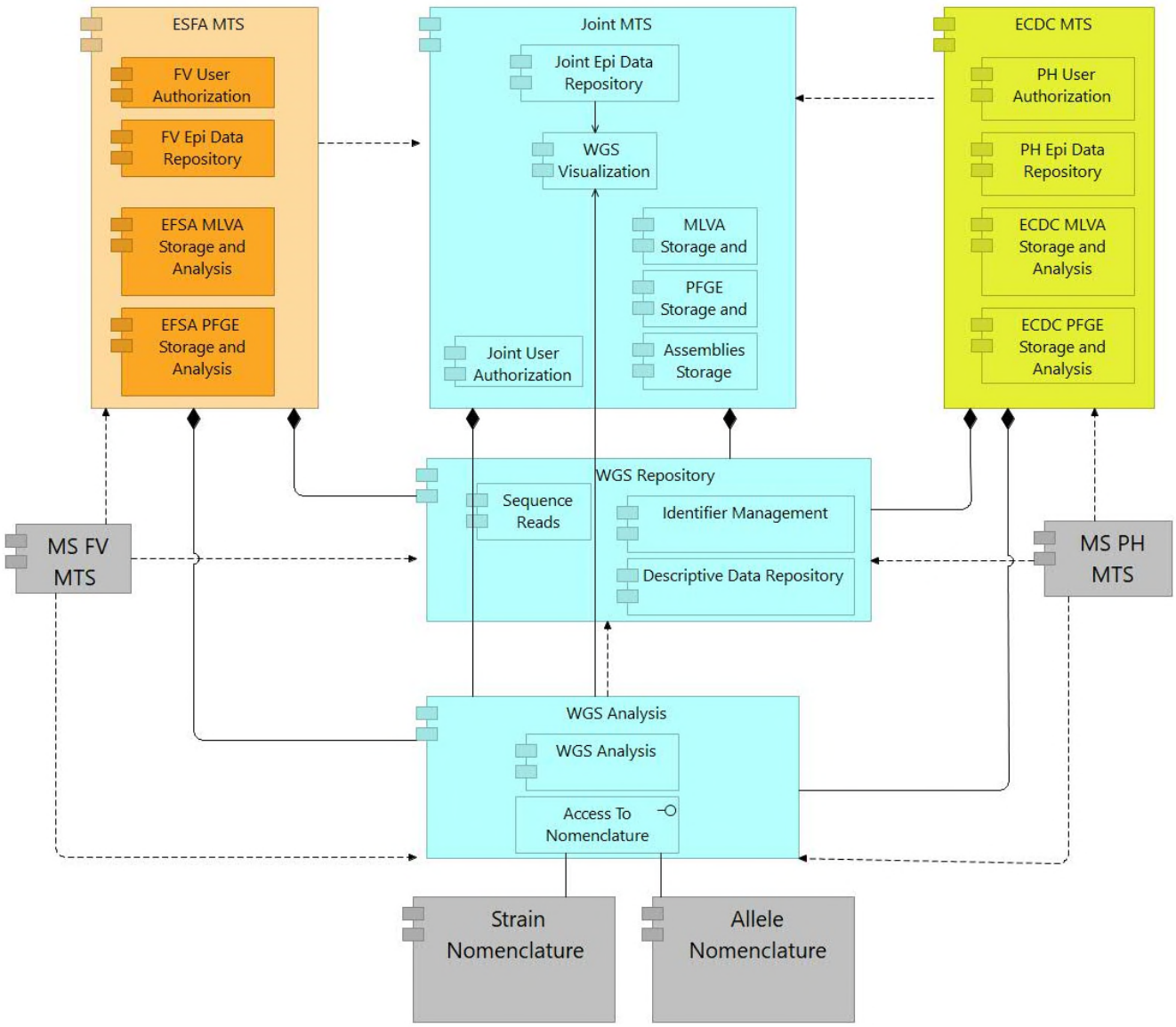
For this system, the logical components used to describe the constraints and requirements reflect (i) the challenges that WGS presents in terms of both storage and analysis of data, and (ii) the existence of two different sources of data from, respectively, the public health sector and the food and veterinary sector.

The logical components are:

- **EFSA molecular typing system (EFSA MTS)**: this logical component stores both descriptive and PFGE/MLVA data on individual isolates of non-human origin, as well as derived WGS data such as assemblies, allele identifiers, genome characterisation results and quality control (QC) results. It performs analyses on PFGE/MLVA data, but not on WGS data, with the exception of data visualisation including WGS-based dendrograms annotated with descriptive data.
- **ECDC molecular typing system (ECDC MTS)**: this logical component stores both descriptive and PFGE/MLVA data on individual isolates of human origin, as well as derived WGS data such as assemblies, allele identifiers, genome characterisation results and QC results. It performs analyses on PFGE/MLVA data, but not on WGS data, with the exception of data visualisation including WGS-based dendrograms annotated with descriptive data.

- **Joint molecular typing system (Joint MTS)**: this logical component stores (limited) descriptive data and PFGE/MLVA data on individual isolates of both human and non-human origin, as well as derived WGS data such as assemblies, allele identifiers, predicted results and QC results. It performs analyses on PFGE/MLVA data, but not on WGS data, with the exception of data visualisation including WGS-based dendrograms annotated with descriptive data.
- **WGS Repository component(s)**: this logical component stores WGS sequence read data, (raw, trimmed, downsampled, etc.), assigns an identifier to them, and allows upload by, and sharing with, different authorised users, as well as interaction with the WGS Analysis component. This may be implemented as a single physical component or as several separate physical components, e.g. a WGS Repository component for non-human-origin isolates and one for human-origin isolates. In addition, implementations may also additionally store derived WGS data such as assemblies and allele identifiers.
- **WGS Analysis component(s)**: this logical component analyses WGS data, including assembly, allele calling, typing and genome characterisation. The WGS Analysis component should be able to handle sequence read data generated by the most commonly used technologies. It is worth noting that this flexibility implies that the component is able to run parallel (technology-specific) pipelines, at least from reads to the assemblies. From this point, the analyses can be done in the same way, regardless of the sequencing technology. It does not permanently store data on individual isolates, nor epidemiological data, but it might have dedicated storage to host temporary data required for analyses. This may be implemented as a single component or as several separate physical components, e.g. a WGS Analysis component for non-human-origin isolates and one for human-origin isolates.
- **Allele Nomenclature component(s)**: this logical component provides, per species, a unique identifier for each unique allele. The component must be publicly accessible to promote global use of the same nomenclature, and should ideally already exist. It does not store data on individual isolates. This may be implemented as a single physical component covering all species, or several separate physical components, per species or group of species.
- **Strain Nomenclature component(s)**: this logical component provides, per species, a 'type' for each isolate, i.e. a single short text code that classifies the isolate into a particular group of common ancestry. The component must be publicly accessible, to promote global use of the same nomenclature, and should ideally already exist. It may store limited data on individual isolates, such as allelic profiles. This may be implemented as a single physical component for all species or several separate physical components, per species or group of species.

The data flow between these logical components is described in Figure 3.




FV: food and veterinary; MLVA: multiple-locus variable-number tandem repeat analysis; MS: Member State; MTS: molecular typing system; PFGE: pulsed-field gel electrophoresis; PH: public health; WGS: whole genome sequencing.

**Figure 3:** Logical components of the envisaged joint EFSA-ECDC Overall System

## 3.4. Constraints

There are several constraints on the possible solutions for the envisaged system, imposed by ECDC and EFSA for legal or policy reasons. These are essentially absolute requirements that must be met by the Overall System, and also further shape the remainder of the requirements:

1) The submission, curation and analysis of PFGE and MLVA data must still be possible for the EFSA MTS, ECDC MTS and Joint MTS.
2) Storage of data in the system as well as overall access rights and access to different types of data for different types of previously existing users must remain the same, in compliance with the Collaboration Agreement (see Table 3).
3) The initial implementation must cover at least *Salmonella* spp., *L. monocytogenes* and *E. coli* (including STEC). The architecture of the system must allow the inclusion of other food-borne pathogens such as *Campylobacter* spp. and food-borne viruses.
4) The architecture of the system must be scalable to allow a higher volume of data in the future.

5) Short sequence reads will not be stored in the ECDC MTS, EFSA MTS or Joint MTS components but instead in the WGS Repository component(s). The WGS Repository component must support sequence reads generated by any platform and must be able to store data privately.
6) The analyses performed on WGS data must return equivalent, and thus comparable, results irrespective of whether the isolates were of human or non-human origin.
7) The analyses performed on WGS data must use the same allele and strain nomenclature.
8) The functionality provided by the Allele Nomenclature and Strain Nomenclature components should ideally already exist and must also be available to public users, to promote global use of the same nomenclature.
9) The existing platforms EPIS, the Rapid Alert System for Food and Feed (RASFF) and the Early Warning and Response System (EWRS) and their potential successors must be used for discussion about clusters among MS Public Health (PH)/Food/Veterinary (FV) users and ECDC/EFSA analysis users.
10) The data in the system are kept confidential unless permission to make data publicly available is given by the data owner or in order to comply with legal obligations.
11) For transparency, audit purposes and reproducibility of results, the system must have an audit trail for both isolate and sequence data, and the analyses performed on them.

## 3.5. Users

Users of the system can be persons or machines having the same role, and with the same authorisation.

Roles and responsibilities:

- MS PH user: representative of the MS public health sector who provides WGS data to ECDC in accordance with ECDC's Coordinating Competent Body policy[6].
- MS FV user: representative of the MS food and veterinary sector who provides WGS data to EFSA.
- ECDC admin user: manages user authentication and authorisation, manages parameter configurations for the ECDC MTS and Joint MTS, approves the application programming interface (API) of the Allele Nomenclature component, approves changes to the classification of stored isolates or changes to the classification algorithm.
- EFSA admin user: manages user authentication and authorisation, manages parameter configurations for the EFSA MTS and Joint MTS, approves the API of the Allele Nomenclature component, approves changes to the classification of stored isolates or changes to the classification algorithm.
- ECDC analysis user: uploads PH WGS data from the ECDC MTS to the Joint MTS, performs WGS analysis.
- EFSA analysis user: uploads FV WGS data from EFSA MTS to Joint MTS, performs WGS analysis.

To understand the needs of the Member States' PH and FV users, the relevant authorities were consulted through EFSA's Scientific Network for Zoonoses Monitoring Data and ECDC's Food and Waterborne Diseases and Zoonoses Network (FWD-Net), National Microbiology Focal Points and National Surveillance Focal Points (see Section 2.1.2). With respect to data sharing, for both sectors, slightly more than half of the responding countries expressed the view that, after upload to the WGS Repository, only a selection of these isolates' sequences (specifically, the assembly or allele identifiers) should be downloadable by other MSs. Most of the remaining countries thought that all of the uploaded isolates' sequences could be downloadable by other countries. Somewhat more MSs would upload raw reads rather than assemblies as sequence data.

With respect to the use of the Overall System after upload of data, it was assumed that MS PH/FV users should in general be able themselves to search across the whole database for matches closely related

[6] See https://ecdc.europa.eu/en/about-us/governance/competent-bodies

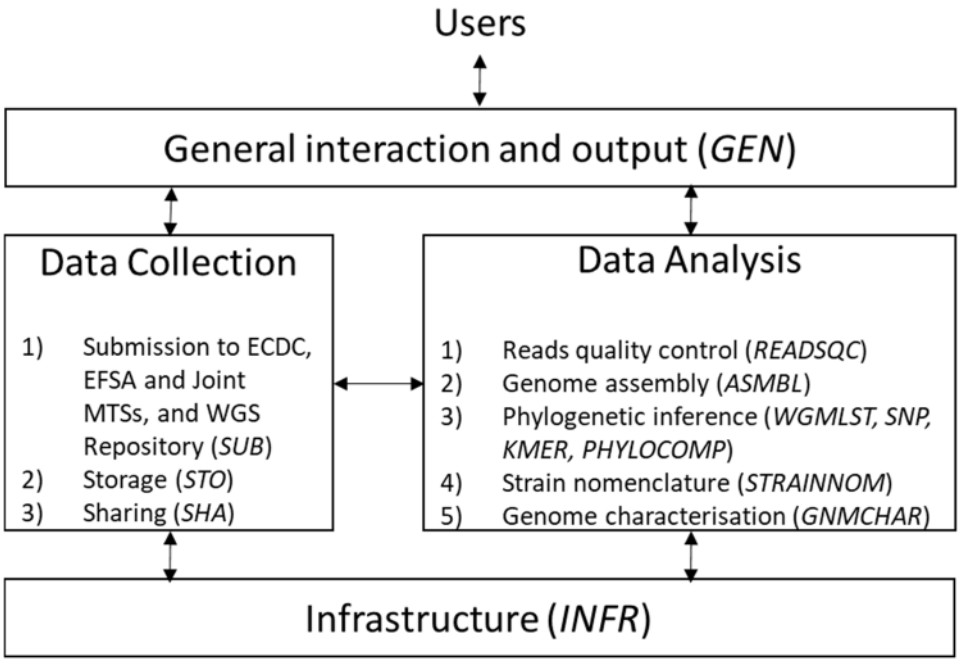
to their own uploaded isolates. In addition to that, a large majority of the MSs were interested in downloading different types of results derived from their own uploaded sequences. These include the assemblies, wgMLST identifiers, strain nomenclature, predicted AMR and other phenotype predictions such as serotype. The ideal maximum delay between data upload by MSs and having these derived results available varied between 12 and 72 hours.

Finally, many MSs are also interested in the possibility of having some or all of the sequence data uploaded to ECDC/EFSA further uploaded to public repositories such as the European Nucleotide Archive (ENA). For those data providers who already submit their data to such a public repository, there should be no need to submit them again to the Overall System because the system is able to retrieve them based on their identifiers.

## 3.6. Requirements analysis

Requirements were initially listed by going exhaustively over each logical component and the interactions between the components. They were subsequently refined after discussion, categorised into Data Collection (Submission, Storage and Sharing), Data Analysis (Sequence read data quality, Genome assembly, Phylogenetic inference, Strain nomenclature and Genome characterisation), General requirements and Infrastructure. Phylogenetic inference was further categorised into whole and core genome MLST, SNP analysis, k-mer-based distance estimates and comparing phylogenetic relationships. Finally, requirements were further broken down so that each requirement is essentially either met or not met. They were further prioritised according to the following classes: 'critical' (in bold in the text), 'medium' or 'optional'.

As further clarification or elaboration of the meaning of the requirements, general concepts have been added to the glossary. If applicable only to a specific set of requirements, separate technical clarifications were added to the section in question. These technical clarifications take into account such aspects as species specificity, current caveats about scientific research, security, performance, legal constraints and usability.




The codes in brackets in italics are used for the requirement description in the subsections below.

**Figure 4:** Overview of the different sets of requirements

## 3.6.1. Data collection

### 3.6.1.1. Submission

1) Interaction between MS PH/FV and ECDC/EFSA users and the WGS Repository component:

**(SUB.WGS.1)** *(Critical)* MS PH/FV users and ECDC/EFSA analysis users must be authenticated in the WGS Repository component before any upload.

**(SUB.WGS.2)** *(Critical)* As a MS PH/FV user or ECDC/EFSA analysis user, it must be possible to upload sequence reads to the WGS Repository component for one or more isolates with a minimum of effort.

(SUB.WGS.3) *(Medium)* Sequence reads must be formatted as single-end or paired-end FASTQ files compressed with gzip.

(SUB.WGS.4) *(Optional)* Additional formats, other than FASTQ, for sequence reads are available.

**(SUB.WGS.5)** *(Critical)* The WGS Repository component must guarantee the correct receipt of data.

(SUB.WGS.6) *(Medium)* The upload of sequence reads to the WGS Repository component must be possible through a user interface.

**(SUB.WGS.7)** *(Critical)* The upload of sequence reads to the WGS Repository component must be possible through file transfer protocol (FTP), secure file transfer protocol (SFTP), secure copy protocol (SCP) or another API.

**(SUB.WGS.8)** *(Critical)* The WGS Repository component must return a unique sequence identifier for each set of successfully uploaded sequence reads for later access and linkage with other descriptive data.

**(SUB.WGS.9)** *(Critical)* A disclaimer regarding the WGS Repository component must be accessible to all users of the component and include a detailed description of the warranties associated with its maintenance, availability and conditions of access, as well as the terms of liability associated with unexpected issues, unexpected disruption, force majeure and similar events.

(SUB.WGS.10) *(Optional)* As a MS PH/FV user, it must be possible to request, as part of the submission to the WGS Repository component, that sequence read data are automatically submitted further by the system to external public databases of the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/), such as the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena). The submission should include a subset of the data uploaded to the ECDC/EFSA MTS, agreed by the MS PH/FV user. The INSDC accession number should be automatically added to the ECDC/EFSA MTS.

2) Interaction between MS PH/FV users and the ECDC/EFSA MTS:

**(SUB.MTS.1)** *(Critical)* MS PH/FV users must be authenticated in the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users) before any upload or update.

**(SUB.MTS.2)** *(Critical)* As a MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: sequence read identifiers, descriptive data about sequence reads.

**(SUB.MTS.3)** *(Critical)* As a MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: assemblies (partial or complete), and descriptive data about assemblies.

**(SUB.MTS.4)** *(Critical)* As a MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: epidemiological data.

**(SUB.MTS.5)** *(Critical)* As a MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: PFGE image data.

**(SUB.MTS.6)** *(Critical)* As a MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: MLVA repeat number data.

**(SUB.MTS.7)** *(Critical)* Sequence reads identifiers uploaded to the ECDC/EFSA MTS must be either from the WGS Repository component, from the ENA or from the SRA. The ECDC/EFSA MTS must be able to differentiate the origin of the sequence identifier.

(SUB.MTS.8) *(Medium)* Partial or complete genome assemblies uploaded to the ECDC/EFSA MTS must be formatted as FASTA files and may be compressed with gzip.

**(SUB.MTS.9)** *(Critical)* The ECDC and EFSA MTS must guarantee the correct receipt of data, e.g. through the use of checksums.

(SUB.MTS.10) *(Medium)* The upload and update of data to the ECDC/EFSA MTS must be possible through a user interface.

**(SUB.MTS.11)** *(Critical)* The upload and update of data to the ECDC/EFSA MTS must be possible through an API.

**(SUB.MTS.12)** (*Critical)* As a MS PH/FV user, it must be possible to update in the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users) any data for their own existing isolates, including replacement of any previously uploaded data and upload of additional data. Changes to the data must be recorded in an audit trail.

(SUB.MTS.13) *(Medium)* The ECDC/EFSA/Joint MTS must, upon update of the sequence or any data that are used as input for the analysis (e.g. sequencing platform or pathogen) for one or more isolates, logically delete any previously derived data for each of these individual isolates. Any use of the logically deleted data must be documented. E.g. when the raw reads of an isolate are replaced, the assembly that was derived from the previous raw reads is logically deleted.

**(SUB.MTS.14)** *(Critical)* The ECDC/EFSA/Joint MTS must be able to validate the data uploaded for one or more isolates according to explicit rules and perform subsequent actions. E.g. in the event that the uploaded data are of insufficient quality, the ECDC/EFSA/Joint MTS must be able to reject them, so that they are not used further.

(SUB.MTS.15) *(Medium)* The ECDC/EFSA/Joint MTS must have an explicit mechanism for implementing changes to the metadata with respect to the variables that can be collected and their permitted values. E.g. there could be an admin interface where variables can be added or removed, and permitted values set.

(SUB.MTS.16) *(Optional)* The ECDC/EFSA/Joint MTS must have an explicit mechanism for implementing changes to the metadata with respect to ontologies and validation rules across variables. E.g. there could be an admin interface where ontologies and validation rules can be edited, added or removed.

3) Interaction between the ECDC/EFSA MTS and the Joint MTS:

**(SUB.JMTS.1)** *(Critical)* ECDC/EFSA analysis users must be authenticated in the Joint MTS before any upload or update.

**(SUB.JMTS.2)** *(Critical)* As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS with a minimum of effort, the following data for one or more isolates: sequence read identifiers and descriptive data about sequence reads.

**(SUB.JMTS.3)** *(Critical)* As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: partial or complete assemblies, and descriptive data about assemblies.

**(SUB.JMTS.4)** *(Critical)* As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: a subset of the epidemiological data in accordance with the collaboration agreement.

**(SUB.JMTS.5)** *(Critical)* As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: PFGE image data.

**(SUB.JMTS.6)** *(Critical)* As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: MLVA repeat number data.

(SUB.JMTS.7) *(Medium)* Data uploaded or updated by MS PH/FV users to the ECDC MTS (for MS PH users) or EFSA MTS (for MS FV users), may not be altered when uploaded or updated by EFSA/ECDC analysis users to the Joint MTS, except for specific mappings to align data semantics.

**(SUB.JMTS.8)** *(Critical)* Sequence identifiers uploaded to the Joint MTS must be those uploaded to the ECDC MTS or EFSA MTS (that could be either from the WGS Repository component, from the ENA or from the SRA). The Joint MTS must be able to differentiate the origin of the sequence identifier.

(SUB.JMTS.9) *(Medium)* Partial or complete genome assemblies uploaded to the Joint MTS must be formatted as FASTA files and may be compressed with gzip.

(SUB.JMTS.10) (*Medium)* The upload and update of data to the Joint MTS must be possible through a user interface.

**(SUB.JMTS.11)** *(Critical)* The upload and update of data to the Joint MTS must be possible through an API.

**(SUB.JMTS.12)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to update any data for existing isolates in the Joint MTS, including replacement of any previously uploaded data and upload of additional data. Changes to the data must be recorded in an audit trail.

**(SUB.JMTS.13**) *(Critical)* As an ECDC/EFSA analysis user, it must be possible to delete individual isolates and any data related to them in the Joint MTS. The deleted data must, however, still be present in an audit trail.

4) Application programming interface (API) of the relevant components:

**(SUB.API.1)** *(Critical)* The user interface and API of the WGS Repository component must be publicly accessible and require authentication.

(SUB.API.2) *(Medium)* The API of the WGS Repository component must be fully described in documentation, including sample code in Python, Perl or R for accessing them and file templates.

**(SUB.API.3)** *(Critical)* The API of the WGS Repository component must undergo a defined change management process, with major and minor versions, that maintains backwards compatibility at least within each major version.

**(SUB.API.4)** *(Critical)* The API of the ECDC MTS and EFSA MTS must be fully described in documentation, including sample code in Python, Perl or R for accessing them and file templates.

(SUB.API.5) *(Optional)* The API of the ECDC/EFSA/Joint MTS must undergo a defined change management process, with major and minor versions, that maintains backwards compatibility at least within each major version.

Technical clarifications:

- Formats: FASTQ, FAST5 (Oxford Nanopore technologies), HDF5 and BAM (PacBio file formats), CRAM. The WGS Analysis component should be able to handle sequence read data generated by the most commonly used technologies (see relevant constraint in Section 3.4). It is worth noting that this flexibility will imply that the system runs parallel, technology-specific pipelines, at least, from reads to the assemblies. From this point, the analyses can be done in the same way, regardless of the sequencing technology.

#### 3.6.1.2. Storage

Requirements:

1) Data stored in the WGS Repository component:

   **(STO.1)** *(Critical)* Any data related to individual isolates and sequences stored in the WGS Repository component may only be accessible to specific authorised MS PH/FV users and specific ECDC/EFSA analysis users.

   **(STO.2)** *(Critical)* The WGS Repository component must comply with applicable legal constraints on data protection.

2) Data stored in the MTS (ECDC, EFSA, Joint):

   **(STO.3)** *(Critical)* Any data related to individual isolates and sequences stored in the ECDC, EFSA and Joint MTS may only be accessible to authorised users.

   **(STO.4)** *(Critical)* The ECDC, EFSA and Joint MTS must comply with applicable legal constraints on data protection.

#### 3.6.1.3. WGS Data Sharing

(SHA.1) *(Medium)* As a MS PH/FV user, it must be possible to grant or deny other Member State public health or food and veterinary and ECDC/EFSA analysis users access to its own uploaded sequence data in the WGS Repository component.

(SHA.2) *(Medium)* It must be possible to alter the access rights of other users to the WGS Repository component through a user interface.

(SHA.3) *(Medium)* It must be possible to alter the access rights of other users to the WGS Repository component of other users through an API.

(SHA.4) *(Medium)* Downloading from the WGS Repository component the sequence data of other users that have granted access to their data must be possible through a user interface.

(SHA.5) *(Medium)* Downloading from the WGS Repository component the sequence data of other users that have granted access to their data must be possible through an API.

Technical clarifications:

- Granting access: Granting access to the components can be designed in several ways. Individual users or user groups can be granted access. In the latter case, there should be a clear and agreed structure for the user groups, e.g. (i) all users, (ii) users from the same country (PH, FV) plus ECDC and EFSA, (iii) FV users plus EFSA, (iv) PH users plus ECDC, or (v) users involved in an individual event. ECDC always has access to sequences uploaded by MS PH users, and EFSA to sequences uploaded by MS FV users.
- Update and deletion of an isolate: 'Updating data' in the ECDC/EFSA/Joint MTS means that a user uploads new or changed information for a previously uploaded isolate. The data used to run analyses are always those valid at the time of analysis. In the event that WGS data are updated, i.e. new sequence reads or a new assembly is provided, then all data derived from

the previous WGS data, and related only to the isolate in question, such as allelic profiles or genome characterisation results, must be either logically or physically deleted. An audit trail may be maintained of the changes. See Glossary, 'Permanent storage'.

## 3.6.2. Data analysis

### 3.6.2.1. Sequence read data quality

Requirements:

**(READQC.1)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform adapter removal and trimming of the raw sequence reads of selected isolates, stored in the WGS Repository component, and store the quality processed reads in the WGS Repository component.

**(READQC.2)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform QCs on both the raw and the quality processed sequence reads of selected isolates stored in the WGS Repository component.

**(READQC.3)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to receive the QC results on both raw and quality processed sequence reads in a standard machine-readable format for storage in the ECDC/EFSA/Joint MTS. The results must at a minimum contain a final outcome PASS/WARNING/FAIL (WARNING optional) for each applied QC, plus the read metrics that this classification was derived from.

**(READQC.4)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to inspect, through an interactive graphic interface, any QC results returned by the WGS Analysis component and store them in the ECDC/EFSA/Joint MTS.

**(READQC.5)** *(Critical)* The definition of each QC performed on reads, including algorithms, parameters and thresholds must be fully described in documentation.

(READQC.6) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to downsample the sequence reads of selected isolates stored in the WGS Repository component to a defined target coverage. When finished, the resulting downsampled reads must be stored in the WGS Repository component in any form that allows reconstruction of the original data.

(READQC.7) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to remove any reads that have a high probability of being of human origin from the sequence reads of selected isolates stored in the WGS Repository component. When finished, the resulting reads must be stored in the WGS Repository component for further analysis.

**(READQC.8)** *(Critical)* As an ECDC/EFSA admin user, it must be possible to update any parameters used in the ECDC (ECDC admin user), EFSA (EFSA admin user), Joint MTS or WGS Analysis component for the execution of the QC measures on reads.

**(READQC.9)** (*Critical*) As an ECDC/EFSA admin user, it must be possible to inspect an audit trail of changes to any parameters used in the ECDC (ECDC admin user), EFSA (EFSA admin user), Joint MTS or WGS Analysis component for the execution of the QC measures on reads.

Technical clarifications:

- The following QCs can be performed directly on sequence reads:
    - Species confirmation and contamination detection:
        - Determining the distribution of GC content across the individual sequence reads and comparing this with the reference distribution for the expected species.
        - Verifying the extent to which a defined set of single-copy core genome genes are present, by mapping the reads to reference alleles for each gene (see

Section 3.6.2.3 on whole genome MLST) and comparing them with a minimum threshold.

- Determining the frequency of all unique k-mers within the sequence reads and comparing this with the reference distribution for the expected species.

- Detection of same-species mixed cultures:
  - Searching for heterozygous positions by mapping against a defined set of single-copy genomic targets for the species. A maximum number of genes that are allowed to harbour heterozygous positions should be defined.
- Poor sequence read quality:
  - Determining the average genome coverage and comparing it with a minimum species or lineage-specific threshold.
  - Determining the average target coverage and comparing it with a minimum species or lineage-specific threshold (see Section 3.6.2.3 on core genome MLST).

#### 3.6.2.2. Genome assembly

Requirements:

**(ASMBL.1)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform genome assembly and post-assembly optimisations for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS. The input quality processed (and possibly downsampled) sequence reads must be retrieved from the WGS Repository component. The resulting assembly must be returned and stored in the ECDC/EFSA/Joint MTS.

**(ASMBL.2)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform QCs on the assemblies of selected isolates stored in the ECDC/EFSA/Joint MTS.

**(ASMBL.3)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to receive the QC results on assemblies in a standard machine-readable format for storage in the ECDC/EFSA/Joint MTS. The results must, at a minimum, contain a final outcome PASS/WARNING/FAIL (WARNING optional) for each applied QC, plus the assembly metrics that this classification was derived from.

**(ASMBL.4)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to inspect, through an interactive graphic interface, any QC results returned by the WGS Analysis component and store them in the ECDC/EFSA/Joint MTS.

**(ASMBL.5)** *(Critical)* The definition of each QC performed on assemblies, including algorithms, parameters and thresholds, must be fully described in documentation.

Technical clarifications:

- Thorough validation studies are needed to determine which assembly pipelines provide acceptable results.
- The following QCs can be performed directly on the assembly:
  - Species confirmation and contamination detection:
    - Verifying that the assembly length is within the expected range for the species under analysis. This also needs to take into account variation due to the presence of plasmids, which can add up to around 0.5 Mb to the assembly length.
    - Screening the contigs against a set of reference genomes for bacterial/viral species and comparing them with the expected extent of the matching regions.

- Verifying the number of contigs relying on a significantly lower depth of coverage than the average coverage of the contigs.

- Detection of same-species mixed cultures: verifying the number of single-copy core genome genes that are present more than once in the assembly, by mapping the contigs to the reference alleles of each core genome gene (see Section 3.6.2.3 on core genome MLST) and comparing them with a species-specific maximum threshold.
- Poor sequence read quality: verifying that the assembly length is below the expected range for the species under analysis.

#### 3.6.2.3. Inferring phylogenetic relationships

*Whole and core genome MLST*

Requirements:

**(WGMLST.1)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform allele calling against the entire pan genome on the assemblies of selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS. The allele identifiers, accompanied by allele sequence quality information for each locus, such as whether multiple alleles were called, and the overall QC results of the allele calling must be returned by the WGS analysis component and stored in the respective MTS.

**(WGMLST.2)** *(Critical)* The overall QC results must have a standard machine-readable format and, at a minimum, contain a final outcome PASS/WARNING/FAIL (WARNING optional) for each applied QC, plus the allele calling metrics that this classification was derived from.

**(WGMLST.3)** *(Critical)* Allele sequences must be converted into allele identifiers using the Allele Nomenclature component.

**(WGMLST.4)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to inspect, through an interactive graphic interface, any allele calling results and associated QC results stored in the ECDC/EFSA/Joint MTS.

**(WGMLST.5)** *(Critical)* The algorithms used for the allele calling and their parameters must be fully described in documentation.

**(WGMLST.6)** *(Critical)* Any authenticated user, including those not included as users of the system, must be able to upload individual allele sequences to the Allele Nomenclature component and retrieve their corresponding allele identifiers and allele sequence quality information, including those for putative new alleles. No interaction with the WGS Analysis component or any other component may be required for this.

(WGMLST.7) *(Medium)* Any authenticated user, including those not included as users of the system, must be able to upload the descriptive data about the assembly pipeline and the allele calling pipeline used to the Allele Nomenclature component. No interaction with the WGS Analysis component or any other component may be required for this.

**(WGMLST.8)** *(Critical)* The Allele Nomenclature component must be accessible through an API.

**(WGMLST.9)** *(Critical)* The Allele Nomenclature component must require authentication.

**(WGMLST.10)** *(Critical)* An internal allele nomenclature component that is part of the WGS Analysis component must be able to retrieve the external allele nomenclature from the Allele Nomenclature component as well, but not necessarily in real time.

(WGMLST.11) *(Medium)* The data in the Allele Nomenclature component must be publicly retrievable in bulk through an API. These must include at least (i) all the supported loci, together with their description and reference alleles, (ii) the definition of subsets of loci (schemas), and (iii) the individual allele sequences, their identifiers and any quality information associated with the individual alleles.

**(WGMLST.12)** *(Critical)* The Allele Nomenclature component may store only individual alleles by default with each new query. It may store information on isolates, such as isolate identifiers or allelic profiles, only if consent is given by the user.

**(WGMLST.13)** *(Critical)* The API of the Allele Nomenclature component must be publicly accessible and fully described in documentation, including sample code in Python, Perl or R for accessing them.

**(WGMLST.14)** *(Critical)* The API of the Allele Nomenclature component must undergo a defined change management process that includes approval by MS PH/FV and ECDC/EFSA admin users, with major and minor versions, that maintains backwards compatibility at least within each major version.

(WGMLST.15) *(Medium)* A reference implementation of the allele calling as executed by the WGS Analysis component, implementing the standard algorithms and parameters agreed by ECDC and EFSA must be publicly and freely available.

**(WGMLST.16)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to search for isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS that match at least one of a set of selected isolates available in the ECDC/EFSA/Joint MTS up to a given maximum number of allelic differences, and for a specific subset of loci.

**(WGMLST.17)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to perform clustering analysis on selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS based on different subsets of allele identifiers, i.e. schemas.

**(WGMLST.18)** *(Critical)* It must be possible, irrespective of any existing strain nomenclature, to apply a cluster definition, including a microbiological cluster cut-off and potentially a time limit, to enumerate clusters and store this information in the ECDC, EFSA or Joint MTS.

Technical clarifications:

- The quality of allele sequences derived as a result of allele calling can be defined on two levels:
    - The individual allele sequence extracted from the input sequence data, without further context. This is the information sent to the Allele Nomenclature component, and, if a new allele not yet stored in its database is detected, it will perform the following checks before accepting: (i) verifying that no ambiguous bases are present, (ii) verifying that the allele length is within the normal range found in the target locus, (iii) verifying that both a start and a stop codon are present, and (iv) verifying that any indels are all in-frame.
    - The wider context of the allele sequence. This is information used by the WGS Analysis component, and includes two checks. First, if the allele is found close to the 5' or 3' ends of a contig, the sequence may be incorrect due to the presence of alternate start or stop codons. In such a case, no identifier should be assigned to the allele. Second, multiple allele sequences may be found for the same locus, and may indicate, e.g., a mixed culture. The number of loci with multiple alleles should be returned by the WGS Analysis component, and the locus in question can either have no allele identifier assigned or the identifier of one of the alleles.
- The following QCs can be performed on cg/wgMLST data:
    - Detection of same-species mixed cultures: verifying the number of single-copy core genome genes that are present more than once in the assembly, by mapping the contigs to reference alleles of each core genome gene, and comparing them with a species-specific maximum threshold.
    - Poor sequence read quality: core genome coverage. This is the number or proportion of core genome loci that were retrieved by allele calling, using the cgMLST schema. It is an indication of the overall quality of the sequence data that also takes into account the extent of local regions of low coverage. The thresholds for minimum core genome coverage are likely species-specific.

*Single nucleotide polymorphism analysis*

Requirements:

(SNP.1) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform a mapping of either the sequence reads or the assembly (if the reads are not available) to one or more reference assemblies, for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS. Input sequence reads must be retrieved from the WGS Repository component and input assemblies from the respective MTS. The alignments resulting from this SNP calling process must be retrievable from the WGS Analysis component and stored in the respective MTS.

(SNP.2) *(Medium)* The WGS Analysis component or the ECDC/EFSA/Joint MTS must be able to store reference assemblies for use in, e.g., SNP analysis.

(SNP.3) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform SNP filtering on an alignment to a reference, to filter out any SNPs that have a low probability of being true SNPs and, depending on the species (or lineage) and on the set of isolates under comparison, any true SNPs with a high likelihood of falling within recombination regions.

(SNP.4) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to perform clustering on selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS based on their alignment to a selected reference, and to visualise the result.

(SNP.5) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to generate a maximum parsimony or maximum likelihood tree for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS based on their alignment to a reference, and to visualise the result in accordance with the rules of data visibility established in the ECDC–EFSA–EURL collaboration agreement.

(SNP.6) *(Optional)* As an ECDC/EFSA analysis user, it must be possible to retrieve and combine, with a minimum of effort, filtered vcf files for selected isolates, in order to generate, e.g., distance matrices that can be used for descriptive statistics and statistical tests between groups of isolates.

*Seven- or eight-gene MLST*

Requirements: see the whole and core genome MLST section above with respect to allele calling, and the strain nomenclature section below with respect to sequence type assignment.

*K-mer-based distance estimation*

Requirements:

(KMER.1) *(Optional)* As an ECDC/EFSA analysis user it must be possible to perform an alignment-free estimation of genetic relatedness between isolates using k-mers.

*Comparing phylogenetic relationships*

Requirements:

(PHYLOCOMP.1) *(Optional)* As an ECDC/EFSA analysis user it must be possible to make statistical comparisons of dendrograms obtained with different methods based on topology and branch lengths.

(PHYLOCOMP.2) *(Optional)* As an ECDC/EFSA analysis user it must be possible to assess the correlation between pairwise distance matrices generated with different methods (e.g. based on wgMLST allele identifiers or SNPs) with statistical tests.

(PHYLOCOMP.3) *(Optional)* As an ECDC/EFSA analysis user it must be possible to compare (map) different partitions of the dataset based on categorical variables, e.g. mapping a serotype variable to a WGS-based strain nomenclature variable.

Technical clarifications:

- K-mer-based distance estimation is extremely fast, and scales only linearly with the number of isolates, rather than, e.g., quadratically for distance matrix-based methods.

#### 3.6.2.4. Strain nomenclature

Requirements:

**(STRAINNOM.1)** *(Critical)* Any authenticated user must be able to request the Strain Nomenclature component to return, for a given allelic profile of sufficient quality, a type in the form of a hierarchical numerical code.

(STRAINNOM.2) *(Optional)* The Strain Nomenclature component must return a partial type if the allelic profile could not be classified fully, or no type if it could not be classified at all. If the allelic profile could be assigned to several types of the same hierarchical level, more than one type must be returned, including the probability of matching each of these types.

(STRAINNOM.3) *(Optional)* The Strain Nomenclature component may only store the submitted allelic profile and who submitted it with the explicit consent of the submitter.

(STRAINNOM.4) *(Optional)* Any authenticated user must be able to request the Strain Nomenclature component to return for a given allelic profile, and in addition to a hierarchical numerical code, an equivalent human-readable label. E.g. the equivalent human-readable label for hierarchical numerical code '1.2.4' is 'STEC sprouts outbreak 2011'.

(STRAINNOM.5) *(Optional)* Any authenticated user, including those not included as users of the system, must be able to request the Strain Nomenclature component to return, for a given allelic profile, and in addition to a hierarchical numerical code, the 7-gene MLST sequence type and clonal complex.

**(STRAINNOM.6)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to retrieve the type from the Strain Nomenclature component for selected isolates in the ECDC/EFSA/Joint MTS.

**(STRAINNOM.7)** *(Critical)* Any required manual curation of the Strain Nomenclature data must be guaranteed over time.

**(STRAINNOM.8)** *(Critical)* The Strain Nomenclature component must add new types fully automatically, whereas for adjustments to previous types a manual approval step must be included to avoid, e.g., unnecessary changes to types that have already been used in outbreak investigations. Such manual approval should include consultation with the main users of the nomenclature and a maximum duration of the process from identification of the need to change to approval or rejection should be agreed.

**(STRAINNOM.9)** *(Critical)* The Strain Nomenclature component must maintain an exact history of the type classification and any adjustments. These adjustments include a type that disappears, a type that splits into several new types, several types that merge into a single new one, a type that becomes a subtype of another existing type at the same hierarchical level and a type at the lowest hierarchical level that gets subtypes below it.

**(STRAINNOM.10)** *(Critical)* The Strain Nomenclature component must be accessible through an API.

**(STRAINNOM.11)** *(Critical)* The Strain Nomenclature component must require authentication.

**(STRAINNOM.12)** *(Critical)* An internal strain nomenclature component that is part of the WGS Analysis component must be able to retrieve the external strain nomenclature from the Strain Nomenclature component as well, but not necessarily in real time.

**(STRAINNOM.13)** *(Critical)* The API of the Strain Nomenclature component must be publicly accessible and fully described in documentation, including sample Perl, Python or R code for accessing them.

**(STRAINNOM.14)** *(Critical)* The API of the Strain Nomenclature component must undergo a defined change management process that includes approval by MS PH/FV and ECDC/EFSA admin

users, with major and minor versions, that maintains backwards compatibility at least within each major version.

**(STRAINNOM.15)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to search for isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS that match the type, possibly only partially, of at least one of a set of selected isolates.

Technical clarifications:

- 'Strain classification based on cgMLST/wgMLST allelic profiles': this classification is derived from allele identifiers and relies on the establishment of sublevels of classification defining distinct degrees of phylogenetic relatedness within the diversity of the species under study. While the strain nomenclature should be fundamentally derived from cgMLST loci (which should be carefully chosen), the possibility of generating an additional level of classification based on the genetic similarities found in the accessory genome (i.e. using loci from the wgMLST schema that do not belong to the cgMLST schema) cannot be excluded. Such additional accessory genome-based classification may substantially strengthen the cgMLST-based classification, depending on the accuracy of the locus selection. For instance, a lineage- or pathotype-specific (or particular phenotype-specific) locus may be very informative about the phylogenetic position or pathogenicity of a given isolate, respectively. The pathotyping of *E. coli* based on pathotype-specific loci, such as *stx* genes, is the classic example of how accessory genome loci can be used for nomenclature purposes. Although it is known for some species that phylogeny is largely congruent with specific geno/phenotypic features (e.g. serotypes, presence of specific virulence genes, plasmids), there is currently intensive research on unveiling loci whose presence is (nearly) specific for particular lineage- or pathotypes, which may open avenues for reinforcement of a wide wgMLST-based nomenclature. It is worth noting that, in contrast to SNP-address-based nomenclature, cgMLST/wgMLST-based nomenclature per se does not reflect phylogenetic relationships (i.e. the numerical proximity of two identifier codes within the same sublevel does not reflect genetic relatedness), similarly to what currently happens with seven-loci MLST-based classification (e.g. ST1 isolates can be closer related to ST1000 isolates than to ST2 isolates). Strain classification based on cgMLST/wgMLST can computationally be done very efficiently.
- 'Strain classification based on SNP-based dendrograms/trees': This classification is derived from SNP distances and relies on the establishment of distinct sublevels of classification based on increasing SNP distance cut-offs defined for the SNP-based 'species' phylogenetic tree/dendrograms (ECDC, 2015). In this case, the nomenclature relies on the sequential combination of identifiers obtained for each sublevel (from the highest to the lowest sublevel), so the final codes reflect not only phylogenetic relationships per se, but also provide clues about the degree of diversity between closely related isolates (e.g. isolates 1.1.1 and 1.1.2 share the same parent, 1.1, and the number of differences between them is at least *x* and maximally *y*, with *x* and *y* reflecting the sequential cut-off used).

### 3.6.2.5. Genome characterisation

The requirements GNMCHAR.1–14 in this section cover different types of genome characterisation for different species. For readability, only a short description is given for each requirement. The full requirement can be phrased as 'As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to return and store in the respective MTS predicted "XXX" results for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS'. The input for the prediction is the quality processed (potentially downsampled) sequence and/or the assemblies, whereby 'XXX' is to be replaced by the specific types of genome characterisation.

Requirements:

**(GNMCHAR.1)** *(Critical)* It must be possible to predict the serotype for *Salmonella* spp.

(GNMCHAR.2) *(Medium)* It must be possible to predict the serotype for *E. coli*.

(GNMCHAR.3) *(Medium)* It must be possible to predict the serogroup/serotype for *L. monocytogenes*.

**(GNMCHAR.4)** *(Critical)* It must be possible to detect antimicrobial resistance genes in *Salmonella* spp.

**(GNMCHAR.5)** *(Critical)* It must be possible to detect antimicrobial resistance genes in *E. coli.*

(GNMCHAR.6) *(Optional)* It must be possible to detect antimicrobial resistance genes in *Campylobacter* spp.

**(GNMCHAR.7)** *(Critical)* It must be possible to detect mutations associated with antimicrobial resistance for *Salmonella* spp.

**(GNMCHAR.8)** *(Critical)* It must be possible to detect mutations associated with antimicrobial resistance for *E. coli*.

(GNMCHAR.9) *(Optional)* It must be possible to detect mutations associated with antimicrobial resistance for *Campylobacter* spp.

(GNMCHAR.10) (*Medium)* It must be possible to detect virulence genes for *E. coli*.

(GNMCHAR.11) *(Optional)* It must be possible to detect virulence genes for *L. monocytogenes*.

(GNMCHAR.12) *(Optional)* It must be possible to detect persistence genes for *L. monocytogenes*.

(GNMCHAR.13) *(Optional)* It must be possible to predict the MLVA type for *S*. Typhimurium and *S*. Enteritidis.

(GNMCHAR.14) (*Optional)* It must be possible to predict the mobilome (i.e. plasmids, insertion sequences, integrons, phages, etc.) for at least *Salmonella* spp., *E. coli* and *L. monocytogenes*.

**(GNMCHAR.15)** *(Critical)* The algorithms, parameters and thresholds used for each method must be fully documented.

**(GNMCHAR.16)** *(Critical)* The reference databases used for searching must be publicly available.

**(GNMCHAR.17)** *(Critical)* Any authenticated user must be able to request the WGS Analysis component to return, for a given sequence of sufficient quality, a set of genome characterisations, including in the form of a report.

**(GNMCHAR.18)** *(Critical)* The genome characterisations performed by the WGS Analysis component must be maintained over time to accommodate new algorithms and data.

**(GNMCHAR.19)** *(Critical)* The WGS Analysis component must have an architecture that allows new types of genome characterisations to be added with a minimum of effort from a cost and development point of view.

*Technical clarifications of targeted* in silico *typing/genome characterisation:*

- 'Overview of tools for targeting *in silico* typing/genome characterisation': Intensive research efforts have been carried out in recent years towards the development and implementation of *in silico* typing tools that allow a straightforward determination/prediction of relevant genotypic/phenotypic features from WGS data. Most of the available *in silico* tools focus on capturing/predicting the typing data (e.g. *S. enterica* serotypes, *E. coli* pathotypes, *L. monocytogenes* serogroups/lineages) routinely provided by the traditional pheno- and genotyping techniques. In general, high levels of concordance have been reached between the traditional and the WGS-based *in silico* methods, leading to good perspectives that backwards compatibility with 'historical' typing data will be ensured during the methodology transfer process. In addition, an increasing number of *in silico* approaches have been designed to provide a wider perspective over the genetic backbone of isolates (and ultimately over their virulence/antibiotic resistance properties) by focusing on determining/predicting specific genetic traits, such as plasmids, phages, pathogenicity islands, CRISPR loci, AMR-associated genes and SNPs, etc.

- 'Selection of targets/database construction': Regardless the bioinformatics approach behind the currently available tools, and depending on the trait that will be predicted, a critical upstream step for tool development is the careful selection of the targets (e.g. loci, SNPs) or panel of targets that define a given phenotype (e.g. the definition of the combination of marker genes that define *L. monocytogenes* serogroups; the choice of differential presence/absence profiles or combination of pathotype-specific markers that can guide the *in silico* classification of *E. coli* pathotypes; or the precise definition of repertoire of SNPs mediating a specific antibiotic resistance phenotype).

- 'Targeted *in silico* typing tools' (i.e. tools capturing/predicting the typing data routinely provided by the traditional pheno- and genotyping assay): The large majority of state-of-the-art *in silico* typing tools rely on determining the presence or absence of specific (or a combination of) loci, by taking reads and/or assemblies as input to query well-defined reference databases. Once targets (or a combination thereof) have been well selected, two main approaches can be followed:
    - i) read mapping against reference target sequences;
    - ii) screening of assemblies for a specific target. For both approaches, it should always be ensured that input data (reads or assemblies) are generated following well-defined quality assurance (QA)/QC procedures.

- Important technical aspects when performing read mapping against reference target sequences include:
    - 'Mean depth of coverage per nucleotide': this is the number of times each nucleotide is present in individual reads.
    - 'Sequence region used as reference for mapping': the addition of flanking regions both upstream and downstream of the target region is normally beneficial for potentiating a non-heterogeneous depth of coverage throughout the whole target loci (mapping directly against the target sequence usually yields a coverage plot following an inverted U shape).
    - 'Minimum percentage identity': the minimum proportion of exact nucleotides matching the reference region.
    - 'Minimum percentage horizontal coverage': the minimum proportion of the target region covered by a depth of coverage threshold.
    - 'Read mapping quality': the probability of the alignment being incorrect.
    - If SNPs or indels are reported, other aspects should be inspected: the minimum number of reads covering the variant position, the minimum base quality at variant position, the minimum proportion of reads at variant position differing from the reference to consider a position as homozygous, strand bias, neighbouring base quality.

- Important technical aspects when screening assemblies for specific targets:
    - Minimum percentage identity.
    - Minimum percentage horizontal coverage.
    - Verification whether or not identified genes are complete, truncated or fragmented.
    - Provide positional location of identified loci within the assembly (i.e. contig and bp range).
    - Provide annotation for coding sequences.

- 'Tools for enhanced genome characterisation' (i.e. tools providing a wider perspective over the genetic backbone of isolates): there are currently plenty of easy-to-use command line or web tools for *in silico* detection and characterisation of specific genome elements, such as antimicrobial resistance genetic determinants, plasmids, phages or phage-like elements, CRISPR, etc. Such tools mainly use BLAST-based approaches to query assemblies against extensive sequence databases, although some may also take reads as input for subsequent mapping. K-mer-based strategies are also available. Regardless of the approach, tools should

use high-quality reads or assemblies generated following well-defined QA/QC procedures as input.

### 3.6.3. General user interaction and outputs

Requirements:

**(GEN.1)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to request the ECDC/EFSA/Joint MTS and/or the WGS Analysis component to rapidly generate a clustering visualisation based on cgMLST/wgMLST allele identifiers for a selection of isolates that can be visually browsed and interactively explored according to different properties of the selected isolates.

(GEN.2) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to request the ECDC/EFSA/Joint MTS and/or the WGS Analysis component to rapidly generate a clustering visualisation based on core SNPs for a selection of isolates that can be visually browsed and interactively explored according to different properties of the selected isolates.

**(GEN.3)** *(Critical)* It must be possible to generate both rooted (dendrograms/phylograms) and unrooted trees (minimum spanning trees) or graphs.

**(GEN.4)** *(Critical)* It must be possible to generate statically generated visualisations such as dendrogram images.

**(GEN.5)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to visually browse and interactively create a selection of isolates in the ECDC/EFSA/Joint MTS according to different properties for further processing, analysis and visualisation. Properties can be, e.g., isolate descriptive data, QC results on reads, assemblies and cgMLST/wgMLST; the results of any specific software or analysis that is applied to the reads or assembly, such as specific subtyping, serotyping/pathotyping, or the presence/absence of antimicrobial resistance genes or specific virulence factors; membership of a cluster or any other grouping; and specific cgMLST/wgMLST alleles from selected loci.

**(GEN.6)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to combine and store results from individual QCs into a single overall category such as 'Accepted', 'AcceptedForOutbreak' or 'Rejected'.

**(GEN.7)** *(Critical)* As an MS PH/FV or ECDC/EFSA analysis user, it must be possible to annotate trees, regardless of the method used to calculate them, with any variable/property stored in the ECDC/EFSA/Joint MTS that exists in a one-to-one relationship with an isolate.

**(GEN.8)** *(Critical)* As an MS PH/FV or ECDC/EFSA analysis user, it must be possible to annotate trees, with only the variables/properties stored in the ECDC/EFSA/joint MTS that the user is allowed access to in accordance with data visibility restrictions.

**(GEN.9)** *(Critical)* As an MS PH/FV or ECDC/EFSA analysis user, it must be possible, in the case of rooted trees, to show the variables/properties aligned underneath each other in columns for ease of interpretation.

**(GEN.10)** (Critical) As an MS PH/FV user, it must be possible to download derived data of isolates uploaded by that user, including assemblies, allele nomenclature, strain nomenclature and genome characterisations.

**(GEN.11)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to visually select subsets of isolates from (different) trees for inclusion in further analyses.

**(GEN.12)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to visually access all epidemiological data in the ECDC/EFSA/Joint MTS related to a single isolate through a single interface.

**(GEN.13)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to visually access all descriptive data and QC results on sequence reads in the ECDC/EFSA/Joint MTS related to a single isolate through a single interface.

**(GEN.14)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to visually access all descriptive data and QC results on an assembly in the ECDC/EFSA/Joint MTS related to a single isolate through a single interface.

**(GEN.15)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to export a static figure of any generated visualisation plus any annotations.

(GEN.16) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to export the tree in a text-based format such as Newick.

(GEN.17) *(Medium)* As an ECDC/EFSA analysis user, it must be possible to export either the distance matrix or the core SNP alignment used to generate the tree.

(GEN.18) *(Medium)* The Joint MTS must have an API that allows the retrieval of a selection of isolates belonging to a defined cluster.

**(GEN.19)** *(Critical)* The Joint MTS must have an API that allows the retrieval of a dendrogram representation based on at least wgMLST/cgMLST plus isolate data in accordance with the rules of data visibility established in the ECDC–EFSA–EURL collaborative agreement.

**(GEN.20)** *(Critical)* As a MS PH/FV user, it must be possible to search for isolates in the Joint MTS that match at least one of a set of selected isolates available in the Joint MTS up to a given maximum number of allelic differences, and for a specific subset of loci. This functionality must be available through a web browser.

**(GEN.21)** *(Critical)* As a MS PH/FV user, it must be possible to perform clustering analysis on a set of query isolates in the Joint MTS based on different subsets of allele identifiers, i.e. schemas, and their matches, with the results shown in accordance with the rules of data visibility established in the ECDC–EFSA–EURL collaborative agreement. This functionality must be available through a web browser.

**(GEN.22)** *(Critical)* A user support service function must be available for each component providing training and helpdesk services appropriate for the level of usage.

**(GEN.23)** *(Critical)* As an ECDC/EFSA analysis user, it must be possible to retrieve all the versions of the software stack used in each analysis and all parameters used by each software to generate stored results.

(GEN.24) *(Medium)* All versions from databases and reference sequences, be it complete genomes, draft genomes or specific loci of the genome, used by the software need to be available to the user.

**(GEN.25)** *(Critical)* As an ECDC/EFSA admin user, it must be possible to update any parameters used in the ECDC (ECDC admin user), EFSA (EFSA admin user) or Joint MTS.

### 3.6.4. Infrastructure

This section describes non-functional requirements for hosting software packages which need to comply with analysis and storage requirements. They are related to computational capacity, bandwidth, storage capacity and business continuity in terms of disaster recovery. If the solution is offered in the form of services, this paragraph does not apply to the infrastructure but instead to the ability to support functional requirements.

Requirements:

**(INFR.1)** *(Critical)* Any data related to individual isolates and sequences stored in the WGS Repository component must be stored permanently.

**(INFR.2)** *(Critical)* The web interface and API of the ECDC MTS and EFSA MTS must be accessible through authenticated end points with a public IP address.

**(INFR.3)** *(Critical)* Any data related to individual isolates and sequences stored in the WGS Repository component must be protected by appropriate disaster recovery.

(INFR.4) *(Medium)* The WGS Repository component must be able to add storage capacity with minimal disruption to the system.

**(INFR.5)** *(Critical)* Any data related to individual isolates and sequences stored in the ECDC, EFSA and Joint MTS must be stored permanently.

**(INFR.6)** *(Critical)* Any data related to individual isolates and sequences stored in the ECDC, EFSA and Joint MTS must be protected by appropriate disaster recovery.

**(INFR.7)** *(Critical)* The WGS Analysis component must be able to add computational capacity with minimal disruption to the system and to perform parallel processing of jobs.

(INFR.8) *(Medium)* The Allele Nomenclature component must be able to temporarily or permanently block the IP addresses of users deemed to have malicious intent, e.g. by submitting artificial allele sequences or engaging in a denial-of-service attack.

(INFR.9) *(Medium)* The Strain Nomenclature component must be able to temporarily or permanently block the IP addresses of users deemed to have malicious intent, e.g. by submitting artificial allele sequences or engaging in a denial-of-service attack.

(INFR.10) *(Medium)* It must be possible to apply different phylogenetic/clustering methods and handle datasets of up to 1,000 isolates and 20,000 loci.

**(INFR.11)** *(Critical)* Any component not implemented at or under the direct control of ECDC or EFSA must have sufficient guarantees in terms of long-term sustainability.

Technical clarifications:

- Worst-case peak throughput in terms of number of isolates, for the foreseeable future. If all *Salmonella*, *Campylobacter*, STEC and *L. monocytogenes* human cases in the EU are sequenced, that would be around 13,000+30,000+800+300≈45,000 isolates in the peak month (ECDC, online). That could be doubled to ~90,000 to include all the non-human-origin isolates in a peak month. Divided by 22 x 8 working hours in a month, gives around 500 isolates per hour.
- Reads (FASTQ) file size is a function of the genome size and average genome coverage. Compression rates using gzip can be assumed to be the same across species and reduce the size to around 40% of the original. Table 4 gives an overview of the file sizes for different species, based on an average genome coverage of 50x – the size varies linearly with the actual coverage.
- Genome assembly is likely the most computationally demanding step, requiring 2–5x the initial read size in memory, and a minimum of about 15 GB of available space. The process takes up to 30' on an 8-core machine with 16 GB RAM, for a 5 Mb genome (*Salmonella*, *E. coli*) and a coverage of up to 100x. Smaller genomes such as *Listeria* and *Campylobacter* take less time.
- The computational capacity needed for processing the worst-case peak throughput of 500 isolates per hour was assessed during hearings for different solutions. Estimates for the assembly of 500 isolates per hour included 10 nodes with 16 cores/64 GB RAM, 250 nodes with 4 cores/32 GB RAM, 50 nodes with 32 cores, 250 nodes with 8 cores/61 GB RAM and 2,000 cores. Considering only computation and not memory, and a linear dependency on the number of isolates, this translates to, respectively, 0.32, 2, 3.2, 4 and 4 cores per isolate per hour. Allele calling estimates varied at 0.02, 0.064, 0.07, 0.08 and 0.128 cores per isolate per hour. Other computations such as strain classification, genome characterisation and hierarchical clustering calculation were considered to be less computationally intensive than allele calling, with the possible exception of a strain classification that also included an update of the entire distance matrix used for strain classification.

**Table 4:** Data storage requirements per isolate

| **Pathogen** | Genome size (Mb) | Average coverage | Raw reads FASTQ file size, uncompressed (Mb) | Fully assembled genome FASTA file size, uncompressed (Mb) |
|---|---|---|---|---|
| *Salmonella* spp. | 5.1 | 50 | 510 | 5.1 |
| *Listeria monocytogenes* | 2.9 | 50 | 290 | 2.9 |
| *Escherichia coli* | 4.6–5.4 | 50 | 460–540 | 4.6–5.4 |
| *Campylobacter* spp. | 1.6 | 50 | 160 | 1.6 |

Source: Expert Opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA (ECDC, 2015).

## 3.7. Existing solutions

As explained in Sections 1.2, 2.1.3 and 2.2.3, 11 available platforms ('solutions') that integrate different tools or functionalities for collecting, analysing and visualising WGS data, and that are widely used in the public health, food and veterinary sectors and in the scientific community, were evaluated: BIGSdb, BioNumerics, CGE, COMPARE, Cloud Services, ENA, EnteroBase, INNUENDO, IRIDA, PathogenWatch, SeqSphere. Information on these selected solutions and their main characteristics are shown in Table 5.

**Table 5:** Summary of main features of the solutions[a]

| Description | BIGSdb | BioNumerics | CGE | Cloud Services | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | Pathogen Watch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *General aspects* | | | | | | | | | | | |
| Main responsible organisation(s) | University of Oxford | Applied Maths (Biomerieux) | Danish Technical University | Different operators, e.g. Amazon, Microsoft, Google | European Bioinformatics Institute, Danish Technical University | European Bioinformatics Institute | University of Warwick | University of Lisbon | Public Health Agency of Canada | Wellcome Sanger Institute | Ridom GmbH |
| Links | https://pubmlst.org/software/database/bigsdb/ | http://www.applied-maths.com/bionumerics | http://www.genomicepidemiology.org | https://aws.amazon.com https://azure.microsoft.com | http://www.compare-europe.eu/ | https://www.ebi.ac.uk/ena | https://enterobase.warwick.ac.uk/ | https://innuendo.readthedocs.io/en/latest/index.html | http://www.iridia.ca/ | https://pathogen.watch/ | https://www.ridom.de/seqsphere/ |
| Installation type (instance) | Instances managed by University of Oxford for *Campylobacter* and by Pasteur Institute for *L. monocytogenes*. Local instance possible as well. | Locally installable instance. Single Calculation Engine and allele nomenclature server instance managed by Applied Maths, both also locally installable. | Single instance managed by Danish Technical University. Several tools can be run locally. | Several instances (cloud computing centres) in the EU managed by the respective companies. | Single instance managed by European Bioinformatics Institute with additional functionality managed by Danish Technical University. | Single instance managed by European Bioinformatics Institute. | Single instance managed by University of Warwick. | Locally installable instance. | Locally installable instance. | Single instance managed by Wellcome Sanger Institute. | Locally installable instance. Single allele and strain nomenclature server instance managed by Ridom. |
| *Data collection and sharing* | | | | | | | | | | | |
| Raw reads | No | Yes | Yes | Yes, general file/object storage only | Yes | Yes | Yes | Yes | Yes | No | Yes |
| Assemblies | Yes | Yes | Yes | Yes, general file/object storage only | Yes | Yes | No | No | No | Yes | Yes |
| Epidemiological data and descriptive data about sequences | Yes. Predefined set only for public instance. | Yes | Yes, predefined set only | Yes, general file/object storage only | Yes, predefined set only | Yes | Yes | Yes, predefined set only | Yes, predefined set only for descriptive data about sequences | Yes, predefined set only | Yes, predefined set only |
| Submission, import and/or linking to data | Web interface and API. Import possible from ENA/SRA. | Reads: linking to local files or to ENA/SRA. Assemblies and epidemiological data: client application and API. | Web interface. | Upload of any data via SFTP, API or local client application. | Web interface and API. Storage is done on ENA. | Web interface and API. | Web interface and API. Import possible from ENA/SRA. | Reads: API (SFTP). Epidemiological data: web interface. Import possible from ENA/SRA. | Reads: Illumina MiSeq and NextSeq instruments and web interface. Epidemiological data: web | Web interface. | Reads: linking to local files or submission via client application. Assemblies and epidemiological data: client |

| Description | BIGSdb | BioNumerics | CGE | Cloud Services | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | Pathogen Watch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | interface and API. | | application. Import possible from ENA/SRA. |
| Submission to ENA/SRA from the solution | No | Yes | Yes | No | Yes | Yes | No | No | Yes | No | Yes |
| Data sharing and public accessibility | Public instances: all data become public without embargo period, except for private 'projects'. Local instance: determined by ECDC/EFSA. | Data shared only with local users. Further restrictions possible for user groups on sets of isolates, individual epidemiological data variables and/or individual 'experiment' data. | No public or restricted sharing. | General functionalities available for sharing, either publicly or with a restricted set of users. Use determined by ECDC/EFSA. | All data become public after embargo period. During embargo, access is granted to a set of users determined by ECDC/EFSA that has access to the 'data hub' in question. | All data become public after embargo period. | Subset of epidemiological data and all sequence data become public without embargo period. Remaining epidemiological data do not become public and can be shared with a restricted set of users determined by ECDC/EFSA. | Data shared only with local users. | Data shared only with local users. Further access control possible. | Data only become public if consent is given. Restricted access possible through 'private collections'. | No, closed environment. Allele nomenclature data are publicly accessible read-only. |
| Search uploaded/ accessible data | Yes | Yes | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Data analysis* | | | | | | | | | | | |
| Genome assembly | No | Yes, including post-assembly optimisation | Yes, including post-assembly optimisation | No | Yes, including post-assembly optimisation | No | Yes, including post-assembly optimisation | Yes, including post-assembly optimisation | Yes, without post-assembly optimisation | No | Yes, including post-assembly optimisation |
| Allele calling[(b)] | Yes. Public instances for *Campylobacter* and *L. monocytogenes*, hosting the respective standard schemas. Local instance can use any schema for any species. | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Standard schemas used, plus own accessory/pan genome schema for *L. monocytogenes* . User-defined | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Standard schemas used. | No | Yes, for *Salmonella*. Standard schema used. | No | Yes, for *Salmonella* and *E. coli*. Hosts the respective standard schemas. | Yes, for *Campylobacter, E. coli*, and *Salmonella*. Own schemas used at present. | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Only alleles already existing in public databases can be called. | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Standard schemas used. | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Public schemas used except for *L. monocytogenes*. Additional own schema for *Campylobacter*. Possible to create own schemas. |

| Description | BIGSdb | BioNumerics | CGE | Cloud Services | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | Pathogen Watch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | subsets of loci can be made. | | | | | | | | | |
| Serotype/serogroup prediction | No | Yes, for *E. coli* and *Listeria*. | Yes, for *E. coli* and *Salmonella*. | No | Yes, for *E. coli* and *Salmonella*. | No | Yes, for *E. coli* and *Salmonella*. | Yes, for *E. coli* and *Salmonella*. | Yes, for *Salmonella*. | Unknown[(c)] | Yes, for *L. monocytogenes*. |
| AMR identification | Yes, for *Campylobacter*. | Yes, for *E. coli*. | Yes, for *Campylobacter, E. coli* and *Salmonella*. | No | Yes, for *Campylobacter, E. coli* and *Salmonella*. | No | No | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Species-independent method. | No | Unknown[(c)] | No |
| Virulence gene identification | No | Yes, for *E. coli* (STEC). | Yes, for *E. coli* (STEC). | No | Yes, for *E. coli* (STEC). | No | No | Yes, for *Campylobacter, E. coli, L. monocytogenes* and *Salmonella*. Species-independent method. | No | No | No |
| Strain nomenclature | Yes | Yes | No | No | No | No | Yes | Yes | No | No | Yes |
| Phylogenetic analysis | Outputs can be produced directly by BIGSdb or by plugins. Trees based on hierarchical clustering, distance matrix, splitstree, locus presence frequency. | Trees based on hierarchical clustering, maximum parsimony and maximum likelihood. Distance matrix, polymorphic loci. | Trees based on hierarchical clustering and maximum likelihood. | No | No | No | Trees based on hierarchical clustering and maximum likelihood. | Trees based on hierarchical clustering. | Trees based on hierarchical clustering. | Trees based on hierarchical clustering. | Trees based on hierarchical clustering. |
| *Infrastructure* | | | | | | | | | | | |
| Scalable processing power | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No, all calculations run on local computer. |

| Description | BIGSdb | BioNumerics | CGE | Cloud Services | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | Pathogen Watch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Permanent storage | Public instances: University of Oxford for *Campylobacter*, Pasteur Institute for *L. monocytogenes*. Local instance: determined by ECDC/EFSA. | Isolate and sequence data: determined by ECDC/EFSA, except for read data linked from ENA/SRA. Allele nomenclature data: determined by ECDC/EFSA if locally installed, otherwise by Applied Maths. | Danish Technical University. | Cloud Services operator. | European Bioinformatics Institute (on ENA) | European Bioinformatics Institute. | University of Warwick. | Determined by ECDC/EFSA, except for read data linked from ENA/SRA. | Determined by ECDC/EFSA. | Wellcome Sanger Institute. | Isolate and sequence data: determined by ECDC/EFSA. Allele and strain nomenclature data: Ridom. |
| Sustainability | Depends on further funding. | Company has no known plans to discontinue the product. | Depends on further funding. | Companies have no known plans to discontinue cloud service products. | Developed through a research project that ends in 2019. Depends on further funding from 2020 onwards. | Institute has no known plans to discontinue the product. Very high degree of sustainability. | Depends on further funding from 2021 onwards. | Developed through a research project that ended in 2018. No further funding from 2018 onwards. | Institute has no known plans to discontinue the product. | Institute has no known plans to discontinue the product. | Company has no known plans to discontinue the product. |

(a): To the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

(b): Standard schemas, core and accessory/pan genome unless stated otherwise: *L. monocytogenes* (Pasteur Institute BIGSdb instance, core genome only); *Salmonella* and *E. coli* (University of Warwick – EnteroBase); *Campylobacter* (University of Oxford – pubMLST BIGSdb instance).

(c): Unknown to the Joint Working Group if this functionality was available by the date of the assessment.

## 3.8. Possible scenarios

### 3.8.1. Individual solutions

Each of the 11 existing solutions listed in Section 3.7 were assessed against the requirements outlined in Section 3.6, using the methodology described in Section 2.2.3. The resulting requirements assessment matrix is shown in Table 6. For PathogenWatch, no feedback was received from the hearing expert in step 4 of the process (corrections to hearing expert answers). For SeqSphere, the hearing expert was unable to contribute to the phase related to the additional evidence and comments on the assessment, i.e. step 6 of the process, and communicated that he would rely on the assessment made by the JWG.

**Table 6:** Requirements assessment for selected solutions[(a)]

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data collection | Data collection | **SUB.WGS.1** | **Critical** | MS PH/FV users and ECDC/EFSA analysis users must be authenticated in the WGS Repository component before any upload. | **1** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **0** |
| Data collection | Data collection | **SUB.WGS.2** | **Critical** | As a MS PH/FV user or ECDC/EFSA analysis user, it must be possible to upload sequence reads to the WGS Repository component for one or more isolates with a minimum of effort. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **UNK** | **0** |
| Data collection | Data collection | SUB.WGS.3 | Medium | Sequence reads must be formatted as single-end or paired-end FASTQ files compressed with gzip. | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| Data collection | Data collection | SUB.WGS.4 | Optional | Additional formats, other than FASTQ, for sequence reads are available. | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| Data collection | Data collection | **SUB.WGS.5** | **Critical** | The WGS Repository component must guarantee the correct receipt of data. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **UNK** | **0** |
| Data collection | Data collection | SUB.WGS.6 | Medium | The upload of sequence reads to the WGS Repository component must be possible through a user interface. | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| Data collection | Data collection | **SUB.WGS.7** | **Critical** | The upload of sequence reads to the WGS Repository component must be possible through FTP, SFTP, SCP or another API. | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **0** | **0** |
| Data collection | Data collection | **SUB.WGS.8** | **Critical** | The WGS Repository component must return a unique sequence identifier for each set of successfully uploaded sequence reads for later access and linkage with other descriptive data. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **0** | **1** |
| Data collection | Data collection | **SUB.WGS.9** | **Critical** | A disclaimer regarding the WGS Repository component must be accessible to all users of the component and include a detailed description of the warranties associated with its maintenance, availability and conditions of access, as well as the terms of liability associated with unexpected issues, unexpected disruption, force majeure and similar events. | **1** | **0** | **0** | **0** | **1** | **1** | **1** | **0** | **0** | **UNK** | **0** |
| Data collection | Data collection | SUB.WGS.10 | Optional | As a MS PH/FV user, it must be possible to request, as part of the submission to the WGS Repository component, that sequence read data are automatically submitted further by the system to external public databases of the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/), | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | such as the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena). The submission should include a subset of the data uploaded to the ECDC/EFSA MTS, agreed by the MS PH/FV user. The INSDC accession number should be automatically added to the ECDC/EFSA MTS. | | | | | | | | | | | |
| Data collection | Data collection | **SUB.MTS.1** | **Critical** | MS PH/FV users must be authenticated in the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users) before any upload or update. | **1** | **1** | **1** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.MTS.2** | **Critical** | As an MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: sequence read identifiers, descriptive data about sequence reads. | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **0** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.MTS.3** | **Critical** | As an MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: assemblies (partial or complete), and descriptive data about assemblies. | **1** | **1** | **0** | **0** | **1** | **1** | **0** | **0** | **0** | **1** | **1** |
| Data collection | Data collection | **SUB.MTS.4** | **Critical** | As an MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: epidemiological data. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.MTS.5** | **Critical** | As an MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: PFGE image data. | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| Data collection | Data collection | **SUB.MTS.6** | **Critical** | As an MS PH/FV user, it must be possible to upload to the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users), with a minimum of effort, the following data for one or more isolates: MLVA repeat number data. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.MTS.7** | **Critical** | Sequence read identifiers uploaded to the ECDC/EFSA MTS must be either from the WGS Repository component, from the ENA or from the SRA. The ECDC/EFSA MTS must be able to differentiate the origin of the sequence identifier. | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **1** | **0** | **0** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data collection | Data collection | SUB.MTS.8 | Medium | Partial or complete genome assemblies uploaded to the ECDC/EFSA MTS must be formatted as FASTA files and may be compressed with gzip. | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Data collection | Data collection | **SUB.MTS.9** | **Critical** | The ECDC and EFSA MTS must guarantee the correct receipt of data, e.g. through the use of checksums. | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **0** |
| Data collection | Data collection | SUB.MTS.10 | Medium | The upload and update of data to the ECDC/EFSA MTS must be possible through a user interface. | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Data collection | Data collection | **SUB.MTS.11** | **Critical** | The upload and update of data to the ECDC/EFSA MTS must be possible through an API. | **1** | **1** | **0** | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **0** |
| Data collection | Data collection | **SUB.MTS.12** | **Critical** | As an MS PH/FV user, it must be possible to update in the ECDC MTS (for MS PH users) or the EFSA MTS (for MS FV users) any data for their own existing isolates, including replacement of any previously uploaded data and upload of additional data. Changes to the data must be recorded in an audit trail. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **1** | **0** | **1** |
| Data collection | Data collection | SUB.MTS.13 | Medium | The ECDC/EFSA/Joint MTS must, upon update of the sequence or any data that is used as input for the analysis (e.g. sequencing platform or pathogen) for one or more isolates, logically delete any previously derived data for each of these individual isolates. Any use of the logically deleted data must be documented. E.g. when the raw reads of an isolate are replaced, the assembly that was derived from the previous raw reads is logically deleted. | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | UNK | 0 |
| Data collection | Data collection | **SUB.MTS.14** | **Critical** | The ECDC/EFSA/Joint MTS must be able to validate the data uploaded for one or more isolates according to explicit rules and perform subsequent actions. E.g. in the event that the uploaded data are of insufficient quality, the ECDC/EFSA/Joint MTS must be able to reject them, so that they are not used further. | **0** | **1** | **0** | **0** | **0** | **1** | **1** | **1** | **1** | **0** | **1** |
| Data collection | Data collection | SUB.MTS.15 | Medium | The ECDC/EFSA/Joint MTS must have an explicit mechanism for implementing changes to the metadata with respect to the variables that can be collected and their permitted values. E.g. there could be an admin interface where variables can be added or removed, and permitted values set. | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | UNK | 0 |
| Data collection | Data collection | SUB.MTS.16 | Optional | The ECDC/EFSA/Joint MTS must have an explicit mechanism for implementing changes to the metadata | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | UNK | 0 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | with respect to ontologies and validation rules across variables. E.g. there could be an admin interface where ontologies and validation rules can be edited, added or removed. | | | | | | | | | | | |
| Data collection | Data collection | **SUB.JMTS.1** | **Critical** | ECDC/EFSA analysis users must be authenticated in the Joint MTS before any upload or update. | **1** | **1** | **1** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.JMTS.2** | **Critical** | As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: sequence read identifiers and descriptive data about sequence reads. | **0** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.JMTS.3** | **Critical** | As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: partial or complete assemblies, and descriptive data about assemblies. | **1** | **1** | **0** | **0** | **1** | **1** | **0** | **0** | **0** | **1** | **1** |
| Data collection | Data collection | **SUB.JMTS.4** | **Critical** | As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: a subset of the epidemiological data in accordance with the collaboration agreement. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.JMTS.5** | **Critical** | As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: PFGE image data. | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| Data collection | Data collection | **SUB.JMTS.6** | **Critical** | As an EFSA/ECDC analysis user, it must be possible to upload from the ECDC MTS (for ECDC analysis users) or the EFSA MTS (for EFSA analysis users) to the Joint MTS, with a minimum of effort, the following data for one or more isolates: MLVA repeat number data. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Data collection | Data collection | SUB.JMTS.7 | Medium | Data uploaded or updated by MS PH/FV users to the ECDC MTS (for MS PH users) or EFSA MTS (for MS FV users), may not be altered when uploaded or updated | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | by EFSA/ECDC analysis users to the Joint MTS, except for specific mappings to align data semantics. | | | | | | | | | | | |
| Data collection | Data collection | **SUB.JMTS.8** | **Critical** | Sequence identifiers uploaded to the Joint MTS must be those uploaded to the ECDC MTS or EFSA MTS (that could be either from the WGS Repository component, from the ENA or from the SRA). The Joint MTS must be able to differentiate the origin of the sequence identifier. | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **1** | **0** | **0** | **1** |
| Data collection | Data collection | SUB.JMTS.9 | Medium | Partial or complete genome assemblies uploaded to the Joint MTS must be formatted as FASTA files and may be compressed with gzip. | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Data collection | Data collection | SUB.JMTS.10 | Medium | The upload and update of data to the Joint MTS must be possible through a user interface. | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Data collection | Data collection | **SUB.JMTS.11** | **Critical** | The upload and update of data to the Joint MTS must be possible through an API. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **0** |
| Data collection | Data collection | **SUB.JMTS.12** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to update any data for existing isolates in the Joint MTS, including replacement of any previously uploaded data and upload of additional data. Changes to the data must be recorded in an audit trail. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **1** | **UNK** | **1** |
| Data collection | Data collection | **SUB.JMTS.13** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to delete individual isolates and any data related to them in the Joint MTS. The deleted data must, however, still be present in an audit trail. | **0** | **1** | **0** | **0** | **0** | **1** | **1** | **0** | **1** | **1** | **1** |
| Data collection | Data collection | **SUB.API.1** | **Critical** | The user interface and API of the WGS Repository component must be publicly accessible and require authentication. | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **UNK** | **0** |
| Data collection | Data collection | SUB.API.2 | Medium | The API of the WGS Repository component must be fully described in documentation, including sample code in Python, Perl or R for accessing them and file templates. | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | UNK | 0 |
| Data collection | Data collection | **SUB.API.3** | **Critical** | The API of the WGS Repository component must undergo a defined change management process, with major and minor versions, that maintains backwards compatibility at least within each major version. | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **1** | **0** | **UNK** | **0** |
| Data collection | Data collection | **SUB.API.4** | **Critical** | The API of the ECDC MTS and EFSA MTS must be fully described in documentation, including sample code in Python, Perl or R for accessing them and file templates. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **0** | **1** | **UNK** | **0** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data collection | Data collection | SUB.API.5 | Optional | The API of the ECDC/EFSA/Joint MTS must undergo a defined change management process, with major and minor versions, that maintains backwards compatibility at least within each major version. | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | UNK | 0 |
| Data collection | Data collection | **STO.1** | **Critical** | Any data related to individual isolates and sequences stored in the WGS Repository component may only be accessible to specific authorised MS PH/FV users and specific ECDC/EFSA analysis users. | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** |
| Data collection | Data collection | **STO.2** | **Critical** | The WGS Repository component must comply with applicable legal constraints on data protection. | **1** | **1** | **0** | **1** | **1** | **1** | **0** | **1** | **1** | **UNK** | **1** |
| Data collection | Data collection | **STO.3** | **Critical** | Any data related to individual isolates and sequences stored in the ECDC, EFSA and Joint MTS may only be accessible to authorised users. | **1** | **1** | **0** | **0** | **1** | **0** | **0** | **1** | **1** | **UNK** | **1** |
| Data collection | Data collection | **STO.4** | **Critical** | The ECDC, EFSA and Joint MTS must comply with applicable legal constraints on data protection. | **1** | **1** | **0** | **1** | **1** | **1** | **0** | **1** | **1** | **UNK** | **1** |
| Data collection | Data collection | SHA.1 | Medium | As an MS PH/FV user, it must be possible to grant or deny other MS PH/FV and ECDC/EFSA analysis users access to its own uploaded sequence data in the WGS Repository component. | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| Data collection | Data collection | SHA.2 | Medium | It must be possible to alter the access rights of other users to the WGS Repository component through user interface. | 1 | 0 | 0 | UNK | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Data collection | Data collection | SHA.3 | Medium | It must be possible to alter the access rights of other users to the WGS Repository component through an API. | 0 | 0 | 0 | UNK | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Data collection | Data collection | SHA.4 | Medium | Downloading from the WGS Repository component the sequence data of other users that have granted access to their data must be possible through a user interface | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | UNK | 1 |
| Data collection | Data collection | SHA.5 | Medium | Downloading from the WGS Repository component the sequence data of other users that have granted access to their data must be possible through an API. | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | UNK | 0 |
| Data analysis | Reads QC | **READQC.1** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform adapter removal and trimming of the raw sequence reads of selected isolates, stored in the WGS Repository component, and store the quality processed reads in the WGS Repository component. | **0** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data analysis | Reads QC | **READQC.2** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform QCs on both the raw and the quality processed sequence reads of selected isolates, stored in the WGS Repository component. | **0** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Reads QC | **READQC.3** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to receive the QC results on both raw and quality processed sequence reads in a standard machine-readable format for storage in the ECDC/EFSA/Joint MTS. The results must at a minimum contain a final outcome PASS/WARNING/FAIL (WARNING optional) for each applied QC, plus the read metrics that this classification was derived from. | **0** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Reads QC | **READQC.4** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to inspect, through an interactive graphic interface, any QC results returned by the WGS Analysis component and store them in the ECDC/EFSA/Joint MTS. | **0** | **1** | **JNK** | **0** | **JNK** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Reads QC | **READQC.5** | **Critical** | The definition of each QC performed on reads, including algorithms, parameters and thresholds must be fully described in documentation. | **0** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Reads QC | READQC.6 | Medium | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to downsample the sequence reads of selected isolates stored in the WGS Repository to a defined target coverage component. When finished, the resulting downsampled reads must be stored in the WGS Repository component in any form that allows reconstruction of the original data. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | UNK | 1 |
| Data analysis | Reads QC | READQC.7 | Medium | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to remove any reads that have a high probability of being of human origin from the sequence reads of selected isolates stored in the WGS Repository component. When finished, the resulting reads must be stored in the WGS Repository component for further analysis. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | UNK | 0 |
| Data analysis | Reads QC | **READQC.8** | **Critical** | As an ECDC/EFSA admin user, it must be possible to update any parameters used in the ECDC (ECDC admin user), EFSA (EFSA admin user), Joint MTS or WGS | **0** | **1** | **0** | **0** | **1** | **0** | **0** | **1** | **1** | **UNK** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Analysis component for the execution of the QC measures on reads. | | | | | | | | | | | |
| Data analysis | Reads QC | **READQC.9** | **Critical** | As an ECDC/EFSA admin user, it must be possible to inspect an audit trail of changes to any parameters used in the ECDC (ECDC admin user), EFSA (EFSA admin user), Joint MTS or WGS Analysis component for the execution of the QC measures on reads. | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **1** | **UNK** | **0** |
| Data analysis | Assembly | **ASMBL.1** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform genome assembly and post-assembly optimisations for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS. The input quality processed (and possibly downsampled) sequence reads must be retrieved from the WGS Repository component. The resulting assembly must be returned and stored in the ECDC/EFSA/Joint MTS. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **0** | **UNK** | **1** |
| Data analysis | Assembly | **ASMBL.2** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform QCs on the assemblies of selected isolates stored in the ECDC/EFSA/Joint MTS. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Assembly | **ASMBL.3** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to receive the QC results on assemblies in a standard machine-readable format for storage in the ECDC/EFSA/Joint MTS. The results, must at a minimum, contain a final outcome PASS/WARNING/FAIL (WARNING optional) for each applied QC, plus the assembly metrics that this classification was derived from. | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Assembly | **ASMBL.4** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to inspect, through an interactive graphic interface, any QC results returned by the WGS Analysis component and store them in the ECDC/EFSA/Joint MTS. | **0** | **1** | **0** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** |
| Data analysis | Assembly | **ASMBL.5** | **Critical** | The definition of each QC performed on assemblies, including algorithms, parameters and thresholds, must be fully described in documentation. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **0** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data analysis | wgMLST | **WGMLST.1** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform allele calling against the entire pan genome on the assemblies of selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS. The allele identifiers, accompanied by allele sequence quality information for each locus such as whether multiple alleles were called, and the overall QC results of the allele calling must be returned by the WGS Analysis component and stored in the respective MTS. | **1** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **0** | **1** | **1** |
| Data analysis | wgMLST | **WGMLST.2** | **Critical** | The overall QC results must have a standard machine-readable format and, at a minimum, contain a final outcome PASS/WARNING/FAIL (WARNING optional) for each applied QC, plus the allele calling metrics that this classification was derived from. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **0** | **1** |
| Data analysis | wgMLST | **WGMLST.3** | **Critical** | Allele sequences must be converted into allele identifiers using the Allele Nomenclature component. | **1** | **1** | **1** | **0** | **0** | **0** | **1** | **1** | **0** | **1** | **1** |
| Data analysis | wgMLST | **WGMLST.4** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to inspect, through an interactive graphic interface, any allele calling results and associated QC results stored in the ECDC/EFSA/Joint MTS. | **0** | **1** | **UNK** | **0** | **0** | **0** | **1** | **1** | **0** | **1** | **1** |
| Data analysis | wgMLST | **WGMLST.5** | **Critical** | The algorithms used for the allele calling and their parameters must be fully described in documentation. | **1** | **0** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | wgMLST | **WGMLST.6** | **Critical** | Any authenticated user, including those not included as users of the system, must be able to upload individual allele sequences to the Allele Nomenclature component and retrieve their corresponding allele identifiers and allele sequence quality information, including for putative new alleles. No interaction with the WGS Analysis component or any other component may be required for this. | **1** | **1** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **1** | **1** |
| Data analysis | wgMLST | WGMLST.7 | Medium | Any authenticated user, including those not included as users of the system, must be able to upload the descriptive data about the assembly pipeline and the allele calling pipeline used to the Allele Nomenclature component. No interaction with the WGS Analysis component or any other component may be required for this. | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | UNK | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data analysis | wgMLST | **WGMLST.8** | **Critical** | The Allele Nomenclature component must be accessible through an API. | **1** | **1** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **UNK** | **0** |
| Data analysis | wgMLST | **WGMLST.9** | **Critical** | The Allele Nomenclature component must require authentication. | **1** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **0** | **1** | **1** |
| Data analysis | wgMLST | **WGMLST.10** | **Critical** | An internal allele nomenclature component that is part of the WGS Analysis component must be able to retrieve the external allele nomenclature from the Allele Nomenclature component as well, but not necessarily in real time. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **UNK** | **0** |
| Data analysis | wgMLST | WGMLST.11 | Medium | The data of the Allele Nomenclature component must be publicly retrievable in bulk through an API. These must include at least (i) all the supported loci, together with their description and reference alleles, (ii) the definition of subsets of loci (schemas), and (iii) the individual allele sequences, their identifiers and any quality information associated with the individual alleles. | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | UNK | 1 |
| Data analysis | wgMLST | **WGMLST.12** | **Critical** | The Allele Nomenclature component may store only individual alleles by default with each new query. It may store information on isolates, such as isolate identifiers or allelic profiles, only if consent is given by the user. | **1** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **0** | **UNK** | **1** |
| Data analysis | wgMLST | **WGMLST.13** | **Critical** | The API of the Allele Nomenclature component must be publicly accessible and fully described in documentation, including sample code in Python/Perl or R for accessing them. | **1** | **0** | **1** | **0** | **1** | **0** | **1** | **1** | **0** | **0** | **0** |
| Data analysis | wgMLST | **WGMLST.14** | **Critical** | The API of the Allele Nomenclature component must undergo a defined change management process that includes approval by MS PH/FV and ECDC/EFSA admin users, with major and minor versions, that maintains backwards compatibility at least within each major version. | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **1** |
| Data analysis | wgMLST | WGMLST.15 | Medium | A reference implementation of the allele calling as executed by the WGS Analysis component, implementing the standard algorithms and parameters agreed by ECDC and EFSA must be publicly and freely available. | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | UNK | UNK |
| Data analysis | wgMLST | **WGMLST.16** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to search for isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS that match at | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **1** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | least one of a set of selected isolates available in the ECDC/EFSA/Joint MTS up to a given maximum number of allelic differences, and for a specific subset of loci. | | | | | | | | | | | |
| Data analysis | wgMLST | **WGMLST.17** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to perform clustering analysis on selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS based on different subsets of allele identifiers, i.e. schemas. | **1** | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **UNK** | **0** |
| Data analysis | wgMLST | **WGMLST.18** | **Critical** | It must be possible, irrespective of any existing strain nomenclature, to apply a cluster definition, including a microbiological cluster cut-off and potentially a time limit, to enumerate clusters and store this information in the ECDC, EFSA or Joint MTS. | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **1** | **1** |
| Data analysis | Other | SNP.1 | Medium | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform a mapping of either the sequence reads or the assembly (if the reads are not available) to one or more reference assemblies, for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS. Input sequence reads must be retrieved from the WGS Repository component and input assemblies from the respective MTS. The alignments resulting from this SNP calling process must be retrievable from the WGS Analysis component and stored in the respective MTS. | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Data analysis | Other | SNP.2 | Medium | The WGS Analysis component or the ECDC/EFSA/Joint MTS must be able to store reference assemblies for use in, e.g., SNP analysis. | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| Data analysis | Other | SNP.3 | Medium | As an ECDC/EFSA analysis user, it must be possible to request the WGS Analysis component to perform SNP filtering on an alignment to a reference, to filter out any SNPs that have a low probability of being true SNPs and, depending on the species (or lineage) and on the set of isolates under comparison, any true SNPs with high likelihood of falling within recombination regions. | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Data analysis | Other | SNP.4 | Medium | As an ECDC/EFSA analysis user, it must be possible to perform clustering on selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MTS based on their alignment to a selected reference, and to visualise the result. | | | | | | | | | | | |
| Data analysis | Other | SNP.5 | Medium | As an ECDC/EFSA analysis user, it must be possible to generate a maximum parsimony or maximum likelihood tree for selected isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS based on their alignment to a reference, and to visualise the result in accordance with the rules of data visibility established in the ECDC–EFSA–EURLs collaboration agreement. | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| Data analysis | Other | SNP.6 | Optional | As an ECDC/EFSA analysis user, it must be possible to retrieve and combine with a minimum effort filtered vcf files for selected isolates, in order to generate, e.g., distance matrices that can be used for descriptive statistics and statistical tests between groups of isolates. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Data analysis | Other | KMER.1 | Optional | As an ECDC/EFSA analysis user it must be possible to perform an alignment-free estimation of genetic relatedness between isolates using k-mers. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Data analysis | Other | PHYLOCOMP.1 | Optional | As an ECDC/EFSA analysis user it must be possible to compare dendrograms obtained with different methods through statistical comparisons based on topology and branch lengths. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Data analysis | Other | PHYLOCOMP.2 | Optional | As an ECDC/EFSA analysis user it must be possible to assess the correlation between pairwise distance matrices generated with different methods (e.g. based on wgMLST allele identifiers or SNPs) with statistical tests. | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Data analysis | Other | PHYLOCOMP.3 | Optional | As an ECDC/EFSA analysis user it must be possible to compare (map) different partitions of the dataset based on categorical variables, e.g. mapping a serotype variable to a WGS-based strain nomenclature variable. | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Data analysis | Strain nomenclature | **STRAINNOM.1** | **Critical** | Any authenticated user must be able to request the Strain Nomenclature component to return for a given allelic profile of sufficient quality a type, in the form of a hierarchical numerical code. | **1** | **0** | **1** | **0** | **0** | **0** | **1** | **1** | **0** | **0** | **0** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data analysis | Strain nomenclature | STRAINNOM.2 | Optional | The Strain Nomenclature component must return a partial type if the allelic profile could not be classified fully, or no type if it could not be classified at all. If the allelic profile could be assigned to several types of the same hierarchical level, more than one type must be returned, including the probability of matching each of these types. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Data analysis | Strain nomenclature | STRAINNOM.3 | Optional | The Strain Nomenclature component may only store the submitted allelic profile and who submitted it with the explicit consent of the submitter. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Data analysis | Strain nomenclature | STRAINNOM.4 | Optional | Any authenticated user must be able to request the Strain Nomenclature component to return for a given allelic profile, and in addition to a hierarchical numerical code, an equivalent human-readable label. E.g. the equivalent human-readable label for hierarchical numerical code '1.2.4' is 'STEC sprouts outbreak 2011'. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Data analysis | Strain nomenclature | STRAINNOM.5 | Optional | Any authenticated user, including those not included as users of the system, must be able to request the Strain Nomenclature component to return, for a given allelic profile, and in addition to a hierarchical numerical code, the 7-gene MLST sequence type and clonal complex. | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | UNK | 1 |
| Data analysis | Strain nomenclature | **STRAINNOM.6** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to retrieve the type from the Strain Nomenclature component for selected isolates in the ECDC/EFSA/Joint MTS. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **0** | **1** |
| Data analysis | Strain nomenclature | **STRAINNOM.7** | **Critical** | Any required manual curation of the Strain Nomenclature data must be guaranteed over time. | **0** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **0** | **1** |
| Data analysis | Strain nomenclature | **STRAINNOM.8** | **Critical** | The Strain Nomenclature component must add new types fully automatically, whereas for adjustments to previous types a manual approval step must be included to avoid, e.g., unnecessary changes to types that have already been used in outbreak investigations. Such manual approval should include consultation with the main users of the nomenclature and a maximum duration of the process from identification of the need to change to approval or rejection should be agreed. | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **0** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data analysis | Strain nomenclature | **STRAINNOM.9** | **Critical** | The Strain Nomenclature component must maintain an exact history of the type classification and any adjustments. These adjustments include a type that disappears, a type that splits into several new types, several types that merge into a single new one, a type that becomes a subtype of another existing type at the same hierarchical level and a type at the lowest hierarchical level that gets subtypes below it. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **1** |
| Data analysis | Strain nomenclature | **STRAINNOM.10** | **Critical** | The Strain Nomenclature component must be accessible through an API. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** |
| Data analysis | Strain nomenclature | **STRAINNOM.11** | **Critical** | The Strain Nomenclature component must require authentication. | **1** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **0** | **0** | **0** |
| Data analysis | Strain nomenclature | **STRAINNOM. 12** | **Critical** | An internal strain nomenclature component that is part of the WGS Analysis component must be able to retrieve the external strain nomenclature from the Strain Nomenclature component as well, but not necessarily in real time. | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| Data analysis | Strain nomenclature | **STRAINNOM. 13** | **Critical** | The API of the Strain Nomenclature component must be publicly accessible and fully described in documentation, including sample Perl, Python or R code for accessing them. | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **0** | **0** |
| Data analysis | Strain nomenclature | **STRAINNOM. 14** | **Critical** | The API of the Strain Nomenclature component must undergo a defined change management process that includes approval by MS PH/FV and ECDC/EFSA admin users, with major and minor versions, that maintains backwards compatibility at least within each major version. | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| Data analysis | Strain nomenclature | **STRAINNOM. 15** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to search for isolates in the ECDC (ECDC analysis user), EFSA (EFSA analysis user) or Joint MTS that match the type, possibly only partially, of at least one of a set of selected isolates. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** |
| Data analysis | Genome characterisation | **GNMCHAR.1** | **Critical** | It must be possible to predict the serotype for *Salmonella* spp. | **0** | **0** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **0** |
| Data analysis | Genome characterisation | GNMCHAR.2 | Medium | It must be possible to predict the serotype for *E. coli*. | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | UNK | 0 |
| Data analysis | Genome characterisation | GNMCHAR.3 | Medium | It must be possible to predict the serogroup/serotype for *L. monocytogenes*. | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | UNK | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data analysis | Genome characterisation | **GNMCHAR.4** | **Critical** | It must be possible to detect antimicrobial resistance genes in *Salmonella* spp. | **0** | **0** | **1** | **0** | **1** | **0** | **0** | **1** | **0** | **UNK** | **0** |
| Data analysis | Genome characterisation | **GNMCHAR.5** | **Critical** | It must be possible to detect antimicrobial resistance genes in *E. coli*. | **0** | **1** | **1** | **0** | **1** | **0** | **0** | **1** | **0** | **UNK** | **0** |
| Data analysis | Genome characterisation | GNMCHAR.6 | Optional | It must be possible to detect antimicrobial resistance genes in *Campylobacter* spp. | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | UNK | 0 |
| Data analysis | Genome characterisation | **GNMCHAR.7** | **Critical** | It must be possible to detect mutations associated with antimicrobial resistance for *Salmonella* spp. | **0** | **0** | **1** | **0** | **1** | **0** | **0** | **1** | **0** | **UNK** | **0** |
| Data analysis | Genome characterisation | **GNMCHAR.8** | **Critical** | It must be possible to detect mutations associated with antimicrobial resistance for *E. coli*. | **0** | **0** | **1** | **0** | **1** | **0** | **0** | **1** | **0** | **UNK** | **0** |
| Data analysis | Genome characterisation | GNMCHAR.9 | Optional | It must be possible to detect mutations associated with antimicrobial resistance for *Campylobacter* spp. | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Data analysis | Genome characterisation | GNMCHAR.10 | Medium | It must be possible to detect virulence genes for *E. coli*. | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Data analysis | Genome characterisation | GNMCHAR.11 | Optional | It must be possible to detect virulence genes for *L. monocytogenes*. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Data analysis | Genome characterisation | GNMCHAR.12 | Optional | It must be possible to detect persistence genes for *L. monocytogenes*. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Data analysis | Genome characterisation | GNMCHAR.13 | Optional | It must be possible to predict the MLVA type for *S.* Typhimurium and *S.* Enteritidis. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Data analysis | Genome characterisation | GNMCHAR.14 | Optional | It must be possible to predict the mobilome (i.e. plasmids, insertion sequences, integrons, phages, etc.) for at least *Salmonella* spp., *E. coli* and *L. monocytogenes*. | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Data analysis | Genome characterisation | **GNMCHAR.15** | **Critical** | The algorithms, parameters and thresholds used for each method must be fully documented. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **0** | **1** |
| Data analysis | Genome characterisation | **GNMCHAR.16** | **Critical** | The reference databases used for searching must be publicly available. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **UNK** | **1** |
| Data analysis | Genome characterisation | **GNMCHAR.17** | **Critical** | Any authenticated user must be able to request the WGS Analysis component to return, for a given sequence of sufficient quality, a set of genome characterisations, including in the form of a report. | **0** | **1** | **1** | **0** | **1** | **0** | **0** | **1** | **1** | **1** | **1** |
| Data analysis | Genome characterisation | **GNMCHAR.18** | **Critical** | The genome characterisations performed by the WGS Analysis component must be maintained over time to accommodate new algorithms and data. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **0** |
| Data analysis | Genome characterisation | **GNMCHAR.19** | **Critical** | The WGS Analysis component must have an architecture that allows new types of genome | **1** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | characterisations to be added with a minimum of effort from a cost and development point of view. | | | | | | | | | | | |
| General | General | **GEN.1** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to request the ECDC/EFSA/Joint MTS and/or the WGS Analysis component to rapidly generate a clustering visualisation based on cgMLST/wgMLST allele identifiers for a selection of isolates that can be visually browsed and interactively explored according to different properties of the selected isolates. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **1** | **1** |
| General | General | GEN.2 | Medium | As an ECDC/EFSA analysis user, it must be possible to request the ECDC/EFSA/Joint MTS and/or the WGS Analysis component to rapidly generate a clustering visualisation based on core SNPs for a selection of isolates that can be visually browsed and interactively explored according to different properties of the selected isolates. | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| General | General | **GEN.3** | **Critical** | It must be possible to generate both rooted (dendrograms/phylograms) and unrooted trees (minimum spanning trees) or graphs. | **1** | **1** | **1** | **0** | **0** | **0** | **1** | **0** | **1** | **UNK** | **1** |
| General | General | **GEN.4** | **Critical** | It must be possible to generate statically generated visualisations such as dendrogram images. | **1** | **1** | **1** | **0** | **0** | **0** | **1** | **1** | **1** | **1** | **1** |
| General | General | **GEN.5** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to visually browse and interactively create a selection of isolates in the ECDC/EFSA/Joint MTS according to different properties for further processing, analysis and visualisation. Properties can be, e.g., isolate descriptive data, QC results on reads, assemblies and cgMLST/wgMLST; results of any specific software or analysis that is applied to the reads or assembly such as specific subtyping, such as serotyping/pathotyping, or the presence/absence of antimicrobial resistance genes or specific virulence factors; membership of a cluster or any other grouping; and specific cgMLST/wgMLST alleles from selected loci. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **1** | **1** | **1** |
| General | General | **GEN.6** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to combine and store results from individual QCs into a single overall category such as 'Accepted', 'AcceptedForOutbreak' or 'Rejected'. | **0** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **0** | **1** |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| General | General | **GEN.7** | **Critical** | As an MS PH/FV or ECDC/EFSA analysis user, it must be possible to annotate trees, regardless of the method used to calculate them, with any variable/property stored in the ECDC/EFSA/Joint MTS that exist in a one-to-one relationship with an isolate. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **UNK** | **1** |
| General | General | **GEN.8** | **Critical** | As an MS PH/FV or ECDC/EFSA analysis user, it must be possible to annotate trees, only with variables/properties stored in the ECDC/EFSA/Joint MTS that the user is allowed access to in accordance with data visibility restrictions. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **1** |
| General | General | **GEN.9** | **Critical** | As an MS PH/FV or ECDC/EFSA analysis user, it must be possible, in the case of rooted trees, to show the variables/properties aligned underneath each other in columns for ease of interpretation. | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **1** | **1** |
| General | General | **GEN.10** | **Critical** | As an MS PH/FV user, it must be possible to download derived data of isolates uploaded by that user, including assemblies, allele nomenclature, strain nomenclature and genome characterisations. | **1** | **1** | **0** | **UNK** | **0** | **0** | **1** | **1** | **1** | **UNK** | **UNK** |
| General | General | **GEN.11** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to visually select subsets of isolates from (different) trees for inclusion in further analyses. | **1** | **1** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **1** | **1** |
| General | General | **GEN.12** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to visually access all epidemiological data in the ECDC/EFSA/Joint MTS related to a single isolate through a single interface. | **1** | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| General | General | **GEN.13** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to visually access all descriptive data and QC results on sequence reads in the ECDC/EFSA/Joint MTS related to a single isolate through a single interface. | **0** | **1** | **0** | **0** | **0** | **1** | **1** | **1** | **1** | **UNK** | **1** |
| General | General | **GEN.14** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to visually access all descriptive data and QC results on an assembly in the ECDC/EFSA/Joint MTS related to a single isolate through a single interface. | **1** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **1** |
| General | General | **GEN.15** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to export a static figure of any generated visualisation plus any annotations. | **1** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **1** |
| General | General | GEN.16 | Medium | As an ECDC/EFSA analysis user, it must be possible to export the tree in a text-based format such as Newick. | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| General | General | GEN.17 | Medium | As an ECDC/EFSA analysis user, it must be possible to export either the distance matrix or the core SNP alignment used to generate the tree. | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | UNK | 1 |
| General | General | GEN.18 | Medium | The Joint MTS must have an API that allows the retrieval of a selection of isolates belonging to a defined cluster. | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | UNK | 0 |
| General | General | **GEN.19** | **Critical** | The Joint MTS must have an API that allows the retrieval of a dendrogram representation based on at least wgMLST/cgMLST plus isolate data in accordance with the rules of data visibility established in the ECDC–EFSA–EURL collaborative agreement. | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| General | General | **GEN.20** | **Critical** | As an MS PH/FV user, it must be possible to search for isolates in the Joint MTS that match at least one of a set of selected isolates available in the Joint MTS up to a given maximum number of allelic differences, and for a specific subset of loci. This functionality should be available through a web browser. | **1** | **1** | **0** | **0** | **1** | **0** | **1** | **1** | **0** | **1** | **1** |
| General | General | **GEN.21** | **Critical** | As an MS PH/FV user, it must be possible to perform clustering analysis on a set of query isolates in the Joint MTS based on different subsets of allele identifiers, i.e. schemas, and their matches, with the results shown in accordance with the rules of data visibility established in the ECDC–EFSA–EURL collaborative agreement. This functionality should be available through a web browser. | **1** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| General | General | **GEN.22** | **Critical** | A user support service function must be available for each component providing training and helpdesk services appropriate for the level of usage. | **0** | **1** | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| General | General | **GEN.23** | **Critical** | As an ECDC/EFSA analysis user, it must be possible to retrieve all the versions of the software stack used in each analysis and all parameters used by each software to generate stored results. | **0** | **1** | **1** | **0** | **1** | **0** | **1** | **1** | **1** | **1** | **1** |
| General | General | GEN.24 | Medium | All versions from databases and reference sequences, be it complete genomes, draft genomes or specific loci of the genome, used by the software need to be available to the user. | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| General | General | **GEN.25** | **Critical** | As an ECDC/EFSA admin user, it must be possible to update any parameters used in the ECDC (ECDC admin user), EFSA (EFSA admin user) or Joint MTS. | **0** | **1** | **1** | **0** | **1** | **0** | **0** | **1** | **1** | **UNK** | **1** |
| Infrastructure | Infrastructure | **INFR.1** | **Critical** | Any data related to individual isolates and sequences stored in the WGS Repository component must be stored permanently. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **0** | **0** | **1** | **1** |
| Infrastructure | Infrastructure | **INFR.2** | **Critical** | The web interface and API of the ECDC MTS and EFSA MTS must be accessible through authenticated end points with a public IP address. | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Infrastructure | Infrastructure | **INFR.3** | **Critical** | Any data related to individual isolates and sequences stored in the WGS Repository component must be protected by appropriate disaster recovery. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **0** | **0** | **UNK** | **0** |
| Infrastructure | Infrastructure | INFR.4 | Medium | The WGS Repository component must be able to add storage capacity with minimal disruption to the system. | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| Infrastructure | Infrastructure | **INFR.5** | **Critical** | Any data related to individual isolates and sequences stored in the ECDC, EFSA and Joint MTS must be stored permanently. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **0** | **0** | **1** | **0** |
| Infrastructure | Infrastructure | **INFR.6** | **Critical** | Any data related to individual isolates and sequences stored in the ECDC, EFSA and Joint MTS must be protected by appropriate disaster recovery. | **0** | **0** | **1** | **1** | **1** | **1** | **1** | **0** | **0** | **UNK** | **1** |
| Infrastructure | Infrastructure | **INFR.7** | **Critical** | The WGS Analysis component must be able to add computational capacity with minimal disruption to the system and to perform parallel processing of jobs. | **0** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| Infrastructure | Infrastructure | INFR.8 | Medium | The Allele Nomenclature component must be able to temporarily or permanently block the IP addresses of users deemed to have malicious intent, e.g. by submitting artificial allele sequences or engaging in a denial-of-service attack. | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| Infrastructure | Infrastructure | INFR.9 | Medium | The Strain Nomenclature component must be able to temporarily or permanently block the IP addresses of users deemed to have malicious intent, e.g. by submitting artificial allele sequences or engaging in a denial-of-service attack. | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Infrastructure | Infrastructure | INFR.10 | Medium | It must be possible to apply different phylogenetic/clustering methods and handle datasets of up to 1,000 isolates and 20,000 loci. | 1 | 1 | 1 | UNK | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

| Main functionality | Functionality | Code | Priority | Description | BIGSdb | BioNumerics | CGE | CloudServices | COMPARE | ENA | EnteroBase | INNUENDO | IRIDA | PathogenWatch | SeqSphere |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Infrastructure | Infrastructure | **INFR.11** | **Critical** | Any component not implemented at or under the direct control of ECDC or EFSA must have sufficient guarantees in terms of long-term sustainability. | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **1** | **0** | **1** |

API: application programming interface; FTP: file transfer protocol; FV: food and veterinary; MLST: mulitlocus sequence typing; MLVA: multiple loci variable-number tandem repeat analysis; MS: Member State; MTS: molecular typing system; PH: public health; QC: quality control; SCP: secure copy protocol; SFTP: secure file transfer protocol; SRA: Sequence Read Archive; WGS: whole genome sequence; UNK: unknown to the Joint Working Group by the date of the assessment.
Bold: critical requirement.
(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

For further assessment and comparison, requirements were grouped into different sets or functionalities. Both a low-resolution approach (data collection, analysis, general and infrastructure) and a high-resolution approach (data collection, sequence read data quality, genome assembly, inferring phylogenetic relationships/wgMLST, strain nomenclature, genome characterisation, general and infrastructure) were considered. The high-resolution approach was selected. For inferring phylogenetic relationships, only the wgMLST subset had critical requirements, and therefore wgMLST was used instead of this functionality.

The different methods mentioned in Section 2.2.4 to estimate the remaining work for each solution per functionality to meet critical requirements were subsequently applied:

1) **Counting the number of critical requirements that are met**. This approach was rejected because for this type of system there is a very large variation in the complexity of implementing each (critical) requirement. The number of unmet requirements is therefore not proportional to the work needed to implement them: it can happen that the solution meeting the most critical requirements may require substantially more work to close the remaining gap than another that meets fewer requirements. Regardless of this issue, Table 7 shows the number of critical requirements met, as a general reference. The number of medium and optional requirements met are shown in Appendix A, along with graphical representations.

2) **Estimating the complexity of (implementing) a requirement**, irrespective of any specific solution and corresponding design, as a proxy for the amount of work required in the event that it is not met. This approach has the advantage that there is no bias possible towards a specific solution when estimating the remaining work. At least two experts independently estimated the complexity of each requirement as 'low', 'medium' or 'high'. After that, any differences were discussed and a resolution was attempted. However, it was found that generating a consensus was not possible because (i) the complexity also depends on the design of the actual solution, and (ii) the complexity is not additive across requirements because, e.g., more than one requirement may be solved through one single development. This approach was therefore abandoned as well, though the estimations were used to facilitate the approach in point 4 below.

3) **Quantitatively estimating the remaining work for each existing solution per functionality**. In this approach, which is common in the Scrum approach to iterative software development [7], 'story points' are given to the functionality in question, reflecting the development work needed. The story points are first independently estimated by different experts and then a consensus is reached on the final value. While this approach has the advantage that it is very thorough, it is also for that reason very time consuming. In addition, this assessment requires intricate knowledge of every solution in question, which the JWG experts felt they did not always have for all solutions. This approach was therefore abandoned as well.

4) **Qualitatively determining significant gaps regarding unmet critical requirements**, i.e. where substantial work would still be needed, but not further quantitatively estimated, for each existing solution and per functionality (i.e. requirement set). The experts of the JWG were asked to evaluate different functionalities and to list the significant gaps for each solution. They were given the list of unmet critical requirements for that functionality, which was sorted by the complexity assessment made under (2) above to facilitate the work. The results were subsequently reviewed and updated by all JWG experts.

[7] See https://en.wikipedia.org/wiki/Scrum_(software_development)

**Table 7:** Number and proportion of critical requirements met by each solution and per functionality[a]

| Solution | Data collection (n=34) | Reads QC (n=7) | Assembly (n=5) | wgMLST (n=15) | Strain nomenclature (n=11) | Genome characterisation (n=10) | General (n=20) | Infra-structure (n=7) |
|---|---|---|---|---|---|---|---|---|
| BIGSdb | 22 (65%) | 0 (0%) | 0 (0%) | 11 (73%) | 10 (91%) | 1 (10%) | 13 (65%) | 1 (14%) |
| BioNumerics[b] | 26 (76%) | 6 (86%) | 5 (100%) | 13 (87%) | 6 (55%) | 6 (60%) | 20 (100%) | 3 (43%) |
| CGE | 11 (32%) | 0 (0%) | 4 (80%) | 8 (53%) | 1 (9%) | 10 (100%) | 4 (20%) | 6 (86%) |
| Cloud Services | 10 (29%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 7 (100%) |
| COMPARE | 30 (88%) | 4 (57%) | 4 (80%) | 8 (53%) | 1 (9%) | 10 (100%) | 7 (35%) | 6 (86%) |
| ENA | 30 (88%) | 0 (0%) | 2 (40%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (15%) | 7 (100%) |
| EnteroBase | 24 (71%) | 5 (71%) | 5 (100%) | 11 (73%) | 9 (82%) | 5 (50%) | 16 (80%) | 6 (86%) |
| INNUENDO | 22 (65%) | 7 (100%) | 5 (100%) | 11 (73%) | 6 (55%) | 10 (100%) | 14 (70%) | 2 (29%) |
| IRIDA | 26 (76%) | 7 (100%) | 4 (80%) | 2 (13%) | 0 (0%) | 6 (60%) | 12 (60%) | 3 (43%) |
| PathogenWatch | 16 (47%) | 0 (0%) | 1 (20%) | 8 (53%) | 0 (0%) | 3 (30%) | 11 (55%) | 4 (57%) |
| SeqSphere | 21 (62%) | 6 (86%) | 5 (100%) | 11 (73%) | 4 (36%) | 4 (40%) | 18 (90%) | 5 (71%) |

QC: Quality control.

(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

(b): Two requirements met for Data Collection refer to PFGE data.

### 3.8.2. Combinations of solutions

It was observed that every single solution has a substantial number of significant gaps depending on the type of functionality (Table 7). Therefore, scenarios were also considered that would consist of a combination of existing solutions.

An attempt was made to enumerate such possible scenarios based on the unmet critical requirements per solution and per functionality. For this, a template was made that allowed one or more solutions to be selected for each functionality, and then automatically showed which requirements of that functionality were not met, based on the requirements assessment, and along with their complexity analysis. In addition, the template contained free-text fields for (i) the estimated remaining work for individual functionalities, (ii) the estimated work for the integration of solutions, (iii) the main types of costs, and (iv) the constraints, if any, that were not met. Finally, the template listed all the risks and allowed their associated probability of occurrence to be filled in for the scenario in question, along with relevant possible mitigations.

All JWG experts were provided with the template and asked to generate at least two scenarios they considered relevant based on the remaining work, associated risks and meeting of the constraints. The scenarios were then put together and discussed by the whole JWG. Inconsistencies and inaccuracies were removed, missing data added and any highly similar scenarios were merged into one.

However, the point was then raised that there might be many suitable scenarios, in particular when taking into account the fact that ECDC and EFSA (i) have the ability to negotiate with the individual solution providers to get to an agreement as a result of which some of the currently unmet constraints could be met, and (ii) might have the financial resources to select a scenario that perhaps requires more remaining work, but provides a better fit with the overall strategic elements or provides more risk reduction.

It was therefore agreed that, rather than aiming to propose individual scenarios, this report should contain the elements necessary to generate scenarios. These elements were summarised in the Scenario Builder which consists of three parts:

1) The individual solutions' significant gaps for each of the eight detailed functionalities, derived from the individual requirements assessment (Table 8, see Section 3.8.1).
2) Limitations to be considered when generating scenarios by selecting a solution for each main functionality. The same solution can be selected for more than one functionality. These limitations are listed in Table 9.
3) The risks that should be assessed for their probability of occurring with the scenario in question. These are listed in Table 10 along with their severity and potential mitigations to reduce either the probability or the severity.

**Table 8:** Significant gaps of each solution for each functionality[a]

**Part A: Data collection, sequence read data quality, genome assembly and whole and core genome MLST**

| Solution | Data collection[b] | Sequence read data quality | Genome assembly | Whole and core genome MLST |
|---|---|---|---|---|
| BIGSdb | No support for raw reads (assemblies only). No audit trail on deleted isolates. | Lacks all main functionality in this area. | Lacks all main functionality in this area. | No change management of the API. |
| BioNumerics | No own raw reads management system. No public user interface or API. | No audit trail on parameter changes. | (no significant gaps) | API not public. Algorithms not fully publicly described. |
| CGE | No support for epidemiological data. No update of data. No audit trail. No upload API. | No graphical interface for QC inspection. No possibility for parameter changes unless locally installed. No audit trail on parameter changes. | No graphical interface for QC inspection. | No change management of the API. No functionality for multiple concurrent schemas. No searching for wgMLST matches. |
| COMPARE | Controlled access for a time period. Policy that all data must ultimately be made public. | Unclear if graphical interface for QC inspection. No possibility for parameter changes unless locally installed. No audit trail on parameter changes. | No graphical interface for QC inspection. | No change management of the API. No functionality for multiple concurrent schemas. No searching for wgMLST matches. |
| Cloud Services | No data management system for any data type. | Lacks all main functionality in this area. | Lacks all main functionality in this area. | Lacks all main functionality in this area. |
| ENA | Access not restricted to authorised users only (policy that all data must be made public, before which no access is given other than for update purposes). | Lacks all main functionality in this area. | Lacks all main functionality in this area. | Lacks all main functionality in this area. |
| EnteroBase | No support for assemblies (raw reads only). Policy that all sequence data be made public immediately. | No possibility for parameter changes. No audit trail on parameter changes. | (no significant gaps) | Not possible to upload individual allele sequences. No change management of the API. No functionality for multiple concurrent schemas. |
| INNUENDO | No support for assemblies (raw reads only). No audit trail. Access not restricted to authorised users only (all data accessible to all users of the system). | (no significant gaps) | (no significant gaps) | Not possible to upload individual allele sequences. No API. |
| IRIDA | No support for assemblies (raw reads only). No support for PFGE. | (no significant gaps) | No post-assembly optimisation | Not possible to upload individual allele sequences. Not possible to detect new alleles. No API. No functionality for multiple concurrent schemas. |
| PathogenWatch | No support for raw reads (assemblies only). No audit trail. Unknown if API will exist. Unclear legal compliance. | Lacks all main functionality in this area. | Unknown if genome assembly and post-assembly optimisations can be performed. Unknown if graphical interface for QC inspection.[c] | No API. No functionality for multiple concurrent schemas. |
| SeqSphere | No own raw reads management system. No API. | No audit trail on parameter changes. | (no significant gaps) | API not public. Algorithms not fully publicly described. |

**Part B: Strain nomenclature, genome characterisation, general and infrastructure**

| Solution | Strain nomenclature | Genome characterisation | General user interaction and outputs | Infrastructure |
|---|---|---|---|---|
| BIGSdb | (no significant gaps) | Not possible to detect antimicrobial resistance genes. Not possible to detect mutations associated with resistance. Not possible to predict *Salmonella* serotype. | No analysis audit trail. Not possible to change analysis parameters if not locally installed. No or very limited support service. Not possible to inspect all QC data on a single isolate in a single interface. | Very limited or no long-term sustainability of the software. No permanent storage/disaster recovery if installed locally. |
| BioNumerics | No API. | Not possible to detect antimicrobial resistance genes. Not possible to detect mutations associated with resistance. Not possible to predict *Salmonella* serotype. | (no significant gaps) | No permanent storage/disaster recovery. |
| CGE | Lacks all main functionality in this area. | (no significant gaps) | Not possible to generate trees based on cg/wgMLST. Not possible to interactively create and visualise subsets of isolates. Admin user cannot change parameters if not locally installed. Very limited or no support service. Trees cannot display selected annotations. Not possible to inspect all QC data on a single isolate in a single interface. Not possible to search for matches up to a maximum number of allelic differences. | Very limited or no long-term sustainability of the software. No permanent storage/disaster recovery if installed locally. |
| COMPARE | Lacks all main functionality in this area. | (no significant gaps) | Not possible to generate trees based on cg/wgMLST. Not possible to interactively create and visualise subsets of isolates. Trees cannot display selected annotations. Not possible to inspect all QC data on a single isolate in a single interface. Not possible to search for matches up to a maximum number of allelic differences. | Very limited or no long-term sustainability for elements not fully integrated into ENA. |
| Cloud Services | Lacks all main functionality in this area. | Lacks all main functionality in this area. | Lacks all main functionality in this area. | (no significant gaps) |
| ENA | Lacks all main functionality in this area. | Lacks all main functionality in this area. | Lacks all main functionality in this area, except support service and visual inspection of epidemiological data. | (no significant gaps) |
| EnteroBase | No change management for API. | Not possible to detect antimicrobial resistance genes. Not possible to detect mutations associated with resistance. | Not possible to generate trees based on generic subsets of loci. Admin user cannot change parameters if not locally installed. | Very limited or no long-term sustainability of the software. No permanent storage/disaster recovery if installed locally. |

**Part B: Strain nomenclature, genome characterisation, general and infrastructure**

| Solution | Strain nomenclature | Genome characterisation | General user interaction and outputs | Infrastructure |
|---|---|---|---|---|
| INNUENDO | No API. | (no significant gaps) | Not possible to generate trees based on generic subsets of loci. Very limited or no support service. | Very limited or no long-term sustainability of the software. No permanent storage/disaster recovery. |
| IRIDA | Lacks all main functionality in this area. | Not possible to detect antimicrobial resistance genes. Not possible to detect mutations associated with resistance. | Not possible to generate trees based on cg/wgMLST. Trees cannot display selected annotations. Not possible to search for matches up to a maximum number of allelic differences. | No permanent storage/disaster recovery. |
| PathogenWatch | Lacks all main functionality in this area. | Unknown if possible to detect antimicrobial resistance genes. Unclear if possible to detect mutations associated with resistance. Unknown if possible to predict *Salmonella* serotype.(c) | Not possible to generate trees based on generic subsets of loci. Admin user cannot change parameters if not locally installed. | Very limited or no long-term sustainability of the software. |
| SeqSphere | Nomenclature not hierarchical. No public API. | Not possible to detect antimicrobial resistance genes. Not possible to detect mutations associated with resistance. Not possible to predict *Salmonella* serotype. | | Limited permanent storage/disaster recovery. |

API: application programming interface; MLST: multilocus sequence typing; QC: quality control.

(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

(b): Collection of PFGE and MLVA data is not included in this assessment.

(c): Unknown to the JWG if this functionality was available by the date of the assessment.

**Table 9:** Limitations to be considered when generating scenarios[a, b]

| Solution | Constraints not met | Other issues |
|---|---|---|
| BIGSdb | (5) when used as the only solution for COLLECTION, READQC, ASMBL or GEN due to no data collection and analysis support for sequence reads. (11) when used for COLLECTION due to no audit trail on deleted isolates. | |
| BioNumerics | (5) when used as the only solution for COLLECTION due to no data collection functionality for non-public sequence reads. | If used for WGMLST, it cannot be combined for that with other solutions due to the assignment of unique allele identifiers. If used for STRAINNOM, it cannot be combined for that with other solutions due to the use of a single linkage tree. |
| CGE | (2), (11) when used as the only solution for COLLECTION due to no data collection functionality. | |
| COMPARE | (2), (10) when used as the only solution for COLLECTION, READQC, ASMBL, WGMLST, STRAINNOM and GNMCHAR due to provider's policy for data to become public. (3) when used for *Listeria* due to no public instance available at present. | Not possible to have a separate instance under ECDC/EFSA control. |
| Cloud Services | (5) when used as the only solution for READQC, ASMBL, WGMLST, STRAINNOM, GNMCHAR or GEN due to no analysis functionality. | No specific functionality for the system. |
| ENA | (2), (10) when used as the only solution for COLLECTION, READQC, ASMBL, WGMLST, STRAINNOM and GNMCHAR due to provider's policy for data to become public. | Not possible to have a separate instance under ECDC/EFSA control. Very high sustainability independent of ECDC/EFSA. |
| EnteroBase | (2), (10) when used for READQC, ASMBL, WGMLST, STRAINNOM, GNMCHAR due to provider's policy for sequence data to become public. (3) when used for *Listeria* due to no public instance available at present. | Not possible to have a separate instance under ECDC/EFSA control. If used for WGMLST, it cannot be combined for that with other solutions due to the assignment of unique allele identifiers. If used for STRAINNOM, it cannot be combined for that with other solutions due to the use of a single linkage tree. |
| INNUENDO | (11) when used for COLLECTION due to no audit trail on epidemiological data on isolates. (3) when used for *Listeria* due to no public instance available at present. | If used for WGMLST, it cannot be combined for that with other solutions due to the assignment of unique allele identifiers. If used for STRAINNOM, it cannot be combined for that with other solutions due to the use of a single linkage tree. |
| IRIDA | | If used for WGMLST, it cannot be combined for that with other solutions due to the assignment of unique allele identifiers. |
| PathogenWatch | Large degree of uncertainty at present about meeting constraints. | Not possible to have a separate instance under ECDC/EFSA control. Large degree of uncertainty at present about available functionality. |
| SeqSphere | (5) when used as the only solution for COLLECTION due to no data collection functionality for non-public sequence reads. | If used for WGMLST, it cannot be combined for that with other solutions due to the assignment of unique allele identifiers. |
| General | (1) only met when BioNumerics is included in the scenario for PFGE. (5) not met by all solutions, as no solution currently supports sequence reads generated by any platform. | |

COLLECTION: data collection; READSQC: sequence reads data quality; ASMBL: genome assembly; WGMLST: wgMLST; STRAINNOM: strain nomenclature; GNMCHAR: genome characterisation; GEN: general user interaction and outputs.

(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

(b): Numbers in brackets refer to the constraints listed in Section 3.4.

**Table 10:** Risks and mitigation actions

| Event | Impact | Maximum severity of the impact | Possible mitigations |
|---|---|---|---|
| Loss of critical data, i.e. no data recovery following a disaster | System not available for use | High | 1. Implement data recovery strategy compliant with ECDC and EFSA IT policies.<br>2. Use cloud storage services with acceptable service level agreement. |
| Security breach, e.g. leak of sensitive data | Potential liability for damages, reputational damage | (Very) High | 3. Do not store sensitive data on infrastructure not directly controlled by ECDC or EFSA (e.g. ENA, PathogenWatch, EnteroBase, CGE).<br>4. Store only relevant WGS data on infrastructure not directly controlled by ECDC or EFSA (e.g. ENA, PathogenWatch, EnteroBase, CGE) with pseudonymised isolate and/or sequence identifiers.<br>5. Use only existing ECDC and EFSA authentication mechanisms. |
| No timely user support | System not available for use, reputational damage | High | 6. Make specific arrangements for user support, either by ECDC and/or EFSA or outsourced. |
| No timely maintenance available | System not available for use, incorrect data in system | High | 7. Conclude acceptable service level agreement with each software provider.<br>8. Have in-house capacity to perform maintenance not related to off-the-shelf software itself.<br>9. For open-source software, ECDC and EFSA to become contributors to any further development. |
| No upgrades available due to critical loss of developer | No new functionality implementable | (Very) High | 10. Rely on more than one software provider with (partially) overlapping functionality as a fallback. |
| Critical system component down | System not available for use | High | 11. Implement system availability compliant with ECDC and EFSA IT policies.<br>12. Conclude acceptable service level agreement with each infrastructure provider. |
| Introduction of unwanted changes due to no full control over the development | Reduced usability, incompatibility with other systems or data | Per solution | 13. Rely on more than one software provider with (partially) overlapping functionality as a fallback. |
| Delays in deployment due to no full control over the infrastructure | System not available for use, new functionality not available to end-user | Per solution | 14. Do not use of software that cannot be locally installed and/or where an acceptable service level agreement is not possible.<br>15. Conclude acceptable service level agreement with each infrastructure provider. |
| Delays in deployment due to high frequency of deployments | New functionality not available to end-user | Per solution | 16. Give authority for high frequency deployments to development team. |
| Low or no reproducibility of the analysis outside the system due to lack of transparency regarding algorithms and parameters | Reduction of trust in the quality of the results | Medium | 17. Require disclosure of algorithms used by each software.<br>18. Do not use software with insufficient transparency regarding algorithms and parameters. |
| Data not available timely after submission (dependent on the infrastructure available to run the solution) | Delay in availability of data | Medium | 19. For submission to external systems, have service level agreement and include maximum delay in data availability after submission in performance requirements.<br>20. For submission to internal systems, include performance requirements in general in development process. |
| Use of technologies not directly supported by overall ECDC or EFSA IT policy | Delays or cancelling of the implementation, fewer people able to carry out the maintenance. | Medium | 21. Use only technologies either compliant with ECDC and EFSA IT policies or for which sufficient general knowledge and evidence of reliability is available. |

| Event | Impact | Maximum severity of the impact | Possible mitigations |
|---|---|---|---|
| Substantial overall technical complexity | Increased maintenance cost | Medium | 22. Have clearly separated (sub)systems and a well-documented interface for each.<br>23. Have full knowledge of the system in-house rather than outsourced.<br>24. Put strict documentation requirements that allow reproducibility of actions. |

# 4. Discussion

*EFSA and ECDC were requested to jointly evaluate the possible solutions for the collection and the analysis of WGS data for at least* L. monocytogenes*,* Salmonella *and* E. coli *by:*

*(1) analysing the outcome of the surveys on the status of use of WGS of food-borne pathogens in MSs in both food and public health sectors;*

*(2) conducting a consultation of relevant actors and players to assess state-of-the-art pipelines for collecting and analysing WGS data in Europe;*

*(3) involving relevant stakeholders to assess the needs and requirements for the analysis of WGS data and their comparability and to describe roles and responsibilities, taking into account that there are different types of WGS data (raw sequence reads, genome assemblies, wgMLST allele identifiers, strain nomenclature, phenotypic predictions), which may require interfacing with externally hosted databases and applications.*

A joint ECDC–EFSA WGS database is essential to ensure that the analysis of molecular typing data from food-borne pathogens is integrated across different countries and sectors. This project aims to improve crisis preparedness and management in the food and feed area in order to ultimately ensure more effective and rapid containment of food and feed-related emergencies and crises in the future.

EFSA and ECDC collected all information needed to evaluate the possible scenarios for the collection and analysis of WGS data for at least *L. monocytogenes*, *Salmonella* and *E. coli*. Care was taken to make the assessment in each step of the process as objective as possible.

MSs in both the public health and the food safety and veterinary sectors were consulted about the status of their preparedness on the use of WGS to respond to challenges posed by threats such as multinational food-borne outbreaks. A joint cross-sector analysis for EU/EEA countries was carried out (ToR 1).

Logical components of the Overall System were identified, and technical requirements were prioritised ('critical', 'medium' or 'optional') independently of existing solutions; they were accurately described and grouped into areas of functionality. Relevant stakeholders were consulted about them (ToR 3).

The platforms (solutions) that integrate many functionalities for collecting, analysing and visualising WGS data and/or that are widely used in the scientific community were thoroughly analysed with the support of hearing experts representing the solutions; this analysis allowed an assessment of whether the solutions met the requirements identified (ToR 2). The evaluation of the various solutions against the requirements presents the situation as of 31 December 2018.

*(4) Preparing a technical report on the identification and the comparison of potential solutions for the set-up and running of a joint EFSA–ECDC pipeline for collecting and analysing WGS data, taking into account the deliverables described in ToR 1, 2 and 3.*

In the absence of a specific methodology for the combined assessment of the outcomes of the previous ToRs and to propose possible scenarios for the collection and analysis of WGS data, several approaches were discussed.

Different methodological approaches were considered which attempted to estimate the remaining work required to meet the critical requirements for each solution and per functionality (i.e. counting the number of critical requirements met per solution and functionality; estimating the complexity of implementing a requirement; quantitatively determining the remaining work for each existing solution per functionality). Eventually the estimate was made by qualitatively determining significant gaps based on unmet critical requirements per solution and per functionality.

The assessment of the individual solutions made clear that every single solution has a substantial number of gaps due to not meeting all the critical requirements. Therefore, scenarios were also considered that would consist of a combination of solutions.

An attempt was made to enumerate such possible scenarios. However, there may be many suitable scenarios and the choice among them also depends on other strategic or financial elements that are not under the control of the JWG, such as hardware infrastructure or workforce required to develop and maintain the selected scenario.

Therefore, the outcome of the assessment contains the elements necessary to generate scenarios rather than individual proposed scenarios. These elements are summarised in the Scenario Builder that includes the significant gaps regarding the unmet critical requirements for each solution/functionality, limitations and risks to be considered when building the scenarios (Tables 8, 9 and 10).

As a general consideration about the data collection, the decision to submit data to a repository that requires them to be made public can be taken only by the data owner. ECDC and EFSA must be able to work with each data owner regardless of that decision, which implies that a closed environment should always be made available in parallel to a public repository. In this regard, ENA currently offers for free many of the data collection functionalities with respect to sequence data, and with a high degree of sustainability. However, it has a requirement that the data be made publicly available before they can be shared with anyone. As a result of the COMPARE project, controlled access for a defined time period may be introduced through 'private data hubs', after which time the data also become publicly available. At the same time, it is technically feasible to combine ENA, with or without the temporarily private data hubs, as public storage with private storage through Cloud Services.

As a general consideration about the data analysis, the different solutions evaluated (BIGSdb, BioNumerics, CGE, COMPARE, EnteroBase, INNUENDO, IRIDA, PathogenWatch and SeqSphere) offer different types of scientific data analysis. The present assessment addresses the needs related to the analysis of WGS data for *L. monocytogenes*, *Salmonella* and *E. coli*. Eventually it should be extended to include other food-borne pathogens such as *Campylobacter* (included in the current report for antimicrobial resistance) and food-borne viruses, upon agreement between EFSA, ECDC, the relevant EURL and the European Commission.

# 5. Conclusions

General:

- No individual existing solution complies with all the critical requirements and constraints. Therefore, scenarios that consist of a combination of solutions were considered.
- Several single solutions or combinations of two solutions meet many of the critical requirements. Additional resources are needed to close the gaps and/or integrate solutions with one another.
- Data Collection and Infrastructure functionalities are largely independent of the Data Analysis functionalities in terms of design of the Overall System. Exceptions to this are at least EnteroBase and COMPARE, where there is a tight integration between the two sets of functionalities.
- An audit trail on the data collection and analysis is instrumental to provide evidence of any changes to epidemiological data, descriptive data about sequences, and how the analyses are carried out. This will ensure transparency of the process and the reproducibility of results.

Data collection:

- The Overall System must be able to work with all data owners/providers, regardless of their decision to make their data publicly available, which implies that a closed environment should be made available in parallel to a public repository.
- For the Data Collection functionality, ENA with or without the extra functionality developed in COMPARE, currently offers for free much of this functionality but with the requirement to make

data publicly available immediately upon submission or after a defined time period. At the same time, it is technically feasible to combine ENA as public storage with private storage through Cloud Services.

- For those data providers who submit their data to a public repository such as ENA, there is no need to submit them again to the Overall System because the system will be able to retrieve them based on their identifiers.
- The need to maintain a solution for PFGE data collection and analysis, which is based on BioNumerics, does not imply that analysis of WGS data must be done through the same solution. Different scenarios can be envisaged that combine two fully separated systems for PFGE and WGS data.

Data analysis:

- For the Data Analysis functionality, it needs to be considered whether the solution needs to be under the exclusive control of ECDC/EFSA, i.e. a local or cloud installation, or whether limited or no control by the Agencies is acceptable. Some solutions can be installed under the exclusive control of ECDC/EFSA: BioNumerics, INNUENDO, IRIDA and SeqSphere. BIGSdb can be used under both modes. Several solutions cannot be installed or used under the exclusive control of ECDC/EFSA: at least ENA, CGE, COMPARE and EnteroBase. For at least CGE, it is at present possible to install single components under the exclusive control of ECDC/EFSA.
- Different solutions offer different types of scientific data analysis. The categorisation ('critical', 'medium', or 'optional') of certain requirements on data analysis could change if the scope of the WGS data collection and analysis is extended, such as the need for SNP analysis for viruses.
- One of the scientific aspects that has a substantial impact on the design of the system is microbiological cluster detection, which can be done either through SNV analysis or cgMLST/wgMLST. The choice is made here to use at least cgMLST/wgMLST (Nadon et al., 2017), and this is implemented in most of the solutions.
- Strain nomenclature, as well as antimicrobial resistance gene or mutation detection and phenotype prediction, still require scientific development and/or international standardisation. Strain nomenclature may also have a substantial impact on the Overall System design. In the meantime, a working strain nomenclature can be used by food safety and public health stakeholders within the EU.
- Many additional strategic and financial elements need to be taken into account to select a few individual scenarios, such as hardware infrastructure or workforce required to develop and maintain the selected scenario.

Other considerations:

- The assessment of the different existing solutions against the requirements represents the situation as of 31 December 2018. All described solutions have since this time been further developed and improved or are due for further development and improvement in the future. The positive engagement of the solution providers with whom the working group had contact, and their enthusiasm towards future developments is also noted.
- This report contains the elements necessary to generate scenarios, rather than aiming to propose individual scenarios. A Scenario Builder is presented, listing significant gaps in terms of critical requirements not met per solution, limitations on scenarios with regard to constraints not met for particular combinations of solutions, and possible risks (Table 8, 9 and 10).

Next steps:

- EFSA and ECDC will use the Scenario Builder provided by the JWG together with other strategic elements, to generate and propose suitable scenarios to EC. In addition, it should also be recognised that integration of solutions brings many challenges, e.g. technical, financial and strategic.

- New developments, improvements or changes in policies of the assessed solutions that would influence the development of the Overall System should be taken into account when exploring possible scenarios.
- The maintenance and further developments of the new Overall System should be guaranteed (e.g. financial investment, strategic planning, qualified human resource availability) to allow the collection and the analysis of WGS data and ultimately to support the investigation of public health events.

## References

Adam MP, Ardinger HH, Pagon RA and Wallace SE (eds.), online. GeneReviews. Seattle (WA), University of Washington, Seattle, 1993-2019. Available online: https://www.ncbi.nlm.nih.gov/books/NBK1116

Amid C, Pakseresht N, Silvester N, Jayathilaka S, Lund O, Dynovski LD, Pataki BA, Visontai D, Cotten M, Xavier BB, Alako B, Belka A, Cisneros JJL, Haringhuizen GB, Harrison PW, Hoeper D, Holt S, Hundahl C, Hussein A, Kaas RS, Malhotra-Kumar S, Leinonen R, Nieuwenhuijse DF, Rahman N, dos S Ribeiro C, Skiby JE, Steger J, Szalai-Gindl JM, Thomsen MCF, Csabai I, Koopmans M, Aarestrup F and Cochrane G, 2019. The COMPARE Data Hubs. bioRxiv pre-print. doi: 10.1101/555938

Cechich A, Piattini M and Vallecillo E (eds.), 2003. Component-Based Software Quality: Methods and Techniques. Heidelberg, Springer-Verlag, 366 pp.

Cock PJ, Fields CJ, Goto N, Heuer ML and Rice PM, 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acid Research, 38, 1767-1771. doi: 10.1093/nar/gkp1137

EC (European Commission), 2012. Vision paper on the development of data bases for molecular testing of foodborne pathogens in view of outbreak preparedness. http://ec.europa.eu/food/safety/docs/biosafety-crisis-vision-paper_en.pdf

ECDC (European Centre for Disease Prevention and Control), online. Surveillance Atlas of Infectious Diseases. Available online: http://atlas.ecdc.europa.eu

ECDC (European Centre for Disease Prevention and Control), 2015. Expert Opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA. Stockholm, ECDC. Available online: https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/food-and-waterborne-diseases-next-generation-typing-methods.pdf

ECDC (European Centre for Disease Prevention and Control), 2018. Monitoring the use of whole-genome sequencing for infectious diseases surveillance in Europe 2015-2017. Stockholm, ECDC. Available online: https://ecdc.europa.eu/en/publications-data/monitoring-use-whole-genome-sequencing-infectious-disease-surveillance-europe

ECDC (European Centre for Disease Prevention and Control), 2019. The ECDC Strategic framework for integration of molecular and genomic data for EU surveillance and cross-border outbreak investigations 2019-2021. Stockholm, ECDC. Available online: https://ecdc.europa.eu/sites/portal/files/documents/framework-for-genomic-surveillance.pdf

EFSA (European Food Safety Authority), 2013. Standard Sample Description ver. 2.0. EFSA Journal 2013;11(10):3424, 114 pp. doi: 10.2903/j.efsa.2013.3424

EFSA (European Food Safety Authority), 2014a. Technical specifications for the pilot on the collection of data on molecular testing of food-borne pathogens from food, feed and animal samples. EFSA supporting publications 2014;11(12):EN-712, 58 pp. doi: 10.2903/sp.efsa.2014.EN-712

EFSA (European Food Safety Authority), 2014b. Guidance on Data Exchange version 2.0. EFSA Journal 2014; 12(12):3945, 173 pp. doi: 10.2903/j.efsa.2014.3945

EFSA (European Food Safety Authority), García Fierro R, Thomas-López D, Deserio D, Liébana E, Rizzi V and Guerra B, 2018. Outcome of EC/EFSA questionnaire (2016) on use of Whole Genome Sequencing (WGS) for food- and waterborne pathogens isolated from animals, food, feed and related environmental samples in EU/EFTA countries. EFSA supporting publication 2018:EN-1432, 49 pp. doi: 10.2903/sp.efsa.2018.EN-1432

EFSA (European Food Safety Authority), Aerts M, Battisti A, Hendriksen R, Kempf I, Teale C, Tenhagen B-A, Veldman K, Wasyl D, Guerra B, Liébana E, Thomas-López D and Beloeil P-A, 2019a. Scientific report on the technical specifications on harmonised monitoring of antimicrobial resistance in zoonotic and indicator bacteria from food-producing animals and food. EFSA Journal 2019;17(5):5709, 121 pp. https://doi.org/10.2903/j.efsa.2019.5709

EFSA (European Food Safety Authority), 2019b. Programming document 2019–2021. Parma, EFSA, 181 pp. Available online: http://www.efsa.europa.eu/sites/default/files/corporate_publications/files/amp1921.pdf

European Commission, 2012. Vision paper on the development of data bases for molecular testing of foodborne pathogens in view of outbreak preparedness. Available online: http://ec.europa.eu/food/safety/docs/biosafety-crisis-vision-paper_en.pdf

Gartner, online. 23 May 2018 release. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide. Available online: https://www.gartner.com/en/documents/3875999

Harrison PW, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Holt S, Hussein A, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martínez-Villacorta J, Milano A, Pakseresht N, Rajan J, Reddy K, Richards E, Rosello M, Silvester N, Smirnov D, Toribio AL, Vijayaraja S and Cochrane G, 2019. The European Nucleotide Archive in 2018. Nucleic Acids Research, 47(D1), D84-D88. doi: 10.1093/nar/gky1078

IIBA (International Institute of Business Analysis), 2015. A Guide to the Business Analysis Body of Knowledge® (BABOK). Toronto, Ontario, International Institute of Business Analysis, BABOK Guide v3. Available online: https://www.iiba.org/standards-and-resources/babok/

Koonin EV, 2005. Orthologs, Paralogs, and Evolutionary Genomics. Annual Review of Genetics, 39, 309-338. doi: 10.1146/annurev.genet.39.073003.114725

Li H, Ruan J and Durbin R, 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 18, 1851-1858. doi: 10.1101/gr.078212.108

Llarena A-K, Ribeiro-Gonçalves BF, Nuno Silva D, Halkilahti J, Machado MP, DaSilva MS, Jaakkonen A, Isidro J, Hämäläinen C, Joenperä J, Borges V, Viera L, Gomes JP, Correia C, Lunden J, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Bikandi J, San Millan R, Martinez-Ballesteros I,Laorden L, Mäesaar M, Grantiņa-Ieviņa L, Hilbert F, Garaizar J, Oleastro M, Nevas M, Salmenlinna S, Hakkinen M, Carriço JA and Rossi M, 2018. INNUENDO: A cross-sectoral platform for the integration ofgenomics in the surveillance of food-borne pathogens. EFSA supporting publication 2018:EN-1498, 142 pp. doi: 10.2903/sp.efsa.2018.EN-1498

Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, Smith AM, Man Kam K, Perez E, Trees E, Kubota K, Takkinen J, Nielsen EM, Carleton H and the FWD-NEXT Expert Panel, 2017. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. EuroSurveillance, 22(23). pii: 30544. doi: 10.2807/1560-7917.ES.2017.22.23.3054.

Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ and the National Microbiology Focal Points and Experts Group, 2017. Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015–2016. Frontiers in Public Health, 5:347. doi: 10.3389/fpubh.2017.00347

Rizzi V, Da Silva Felicio T, Felix B, Gossner CM, Jacobs W, Johansson K, Kotila S, Michelon D, Monguidi M, Mooijman K, Morabito S, Pasinato L, Björkman JT, Torpdahl M, Tozzoli R and Van Walle I, 2017. The ECDC-EFSA molecular typing database for European Union public health protection. Euroreference 2 – March 2017. Available online: https://euroreference.anses.fr/sites/default/files/17%2003%20ED%20ER%2002%201_RIZZI.pdf

WHO (World Health Organization), 2011. Laboratory quality management system: handbook. Lyon, WHO, 248 pp. Available online: https://www.who.int/ihr/publications/lqms/en/

# Glossary and abbreviations

## Glossary

| Term | Description |
|---|---|
| Adapter removal | The removal of adapter sequences added during library preparation from the raw sequence reads. This must be done prior to any further processing of the reads. |
| Allele calling | The overall process of allele sequence extraction from either sequence reads or an assembly, followed by allele identifier assignment. |
| Allele identifier assignment | The process of assigning a unique identifier, per locus to each observed allele. |
| Allele nomenclature | Names and definitions of all loci included in the wg/cgMLST schema, as well as correspondence between allele sequences and allele identifiers. It is intended to generate allelic profiles to be used to determine genetic similarity between isolates (Nadon et al., 2017). |
| Allele sequence | The defined open reading frame or part thereof identified for a particular locus in a particular isolate. |
| Allele sequence extraction | The process of extracting from the assembly, or sequence reads (if a mapping approach is used), for each locus in a schema, the corresponding allele sequence. If the locus is not present, no allele is found, and if it is present more than once or the reads come from a mixed culture or a culture contaminated with a related species, more than one allele may be found for the same locus. In general, this is done by first extracting the defined coding sequences in an assembly and then assigning them to a specific locus of the schema by aligning them to one or more reference alleles for the locus in question, together with a pre-defined nucleotide or amino acid similarity threshold. Thresholds for allele sequence length variation may also be set to avoid, for instance, the identification of gene fusions (alleles longer than usually observed) or pseudogenes (shorter alleles resulting from the introduction of an earlier stop codon). |
| Allelic distance | A measure of genetic relatedness between two isolates derived from their allelic profile. It is calculated by counting the number of loci that have a different allele identifier among the loci of a given schema that are present in both isolates. Loci that are present in only one of the two isolates are normally not considered as contributing to the pairwise distance. For a set of N isolates, a distance matrix containing all N(N-1)/2 possible pairwise distances, can be computed and subsequently used as input for clustering methods. Distances between allelic profiles can also be calculated by defining a set of loci that are present in all the isolates in the comparison, which reduces the discrimination between isolates but avoids the bias that can be introduced in pairwise analyses. |
| Allelic profile | The set of allele identifiers for all loci observed in a particular isolate. The allelic profiles can be used to determine a pairwise distance between two isolates, the allelic distance. |
| Assembler | Algorithm that assembles short nucleotide sequences into larger contiguous sequences or contigs. The set of all contigs for a given strain defines a draft genome for that strain. |
| Assembly metrics | Metrics that can be derived directly from the assembly, including N50, number of contigs and total assembly length. These can be evaluated both before and after post-assembly optimisations, and can also be used to assess the impact of these optimisations. |
| Assembly pipeline | The combination of the assembler and parameter settings used, and any pre-assembly operations done on the reads (quality assessment and/or read trimming) or post-assembly operations related to contig sequence evaluation/validation. |
| Assembly | For bacteria, this comprises the bacterial chromosome and any plasmids each either partially assembled into several contiguous sequences (contigs), or fully assembled into a single contiguous sequence. Partial assemblies occur due to the assembler not having enough information to connect the remaining contigs, either due to insufficient sequence reads available in the connecting regions or due to repetitive regions that are longer than the length of the reads. |
| Audit trail | A security-relevant chronological record, set of records, and/or destination and source of records that provide documentary evidence of the sequence of activities that have affected at any time a specific operation, procedure, or event. Source: https://en.wikipedia.org/wiki/Audit_trail. |
| Average genome coverage | Also referred to as depth of coverage. The average number of times each base in the genome is contained in individual reads (ECDC, 2015). It is usually estimated as the number of nucleotides in the reads divided by either the expected genome size or the size of the genome assembled from the same reads. |

| Term | Description |
|---|---|
| Average target coverage | The number of nucleotides mapping against defined genomic targets for the species of interest and divided by the total length of those targets. |
| Change management process | In the context of information technology, a process to manage changes to infrastructure or software. For software, the term 'change control' is also used. See further https://en.wikipedia.org/wiki/Change_management_(ITSM). |
| Cloud computing | Cloud computing is a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service using internet technologies (Gartner, online). Cloud service elasticity: in the service provider's view, cloud service elasticity is the ability to increase or decrease the amount of system capacity (e.g. CPU, storage, memory and input/output bandwidth) that is available for a given cloud service on demand, in an automated fashion. |
| Cluster analysis | Cluster analysis is an indirect way to perform phylogenetic inference between strains. Several clustering algorithms can be applied to the distance matrices generated by the comparison of allelic profiles. Such clustering algorithms can be hierarchical or non-hierarchical in nature. Examples of hierarchical clustering methodologies that can be used are single linkage, complete linkage and weighted or unweighted pair group method with arithmetic mean (WPGMA or UPGMA). Non-hierarchical methods include neighbour joining and minimum spanning trees or graphic matroids. |
| Contamination detection | Verifying whether a substantial fraction of the reads or assembly does not originate from the expected or detected species (e.g. contamination from other bacterial or viral species that contaminated the sample, from the host or from a human operator). Used as a quality control. |
| Contig | See Assembly. |
| Core genome | The set of loci that is nearly always present in all strains of a given species. In theory the core genome can shrink over time as new strains are sequenced and the frequency of occurrence of some loci becomes lower than the threshold used. In practice, when sufficient strains are initially available, the core genome can also be defined as a fixed set. In addition, technically problematic loci, such as close paralogues or highly repeatable regions, are normally removed to improve the performance of cgMLST. |
| Core /whole genome MLST (cgMLST/wgMLST) | Extension of the concept of multilocus sequence typing (MLST), or gene-by-gene approach, to the core genome loci (cgMLST) or pan genome loci of a given species or genus (wgMLST). By greatly extending the number of target loci, the discriminatory power of such methodologies is also greatly superior to that of MLST. |
| Core/whole genome MLST (cgMLST/wgMLST) schema | A fixed set of core (for cgMLST) or pan (for wgMLST) genome loci, including one or more reference allele sequences per locus, plus a similarity threshold on nucleotide or amino acid level to determine whether a particular allele sequence matches a reference sequence for a particular locus and consequently belongs to the same locus. For each species, a globally accepted schema should be used. |
| Data owner | Legal or natural person that has legal rights over the respective data. |
| Deletion | See Permanent storage |
| Descriptive data about assemblies | Data about how the assembly was generated and further processed. Examples:<br>o Assembler used, including version<br>o Assembler parameter settings<br>o Post-assembly optimisation used. |
| Descriptive data about sequence reads | Data about how the sequence was generated and further processed. Examples:<br>o DNA extraction method<br>o Library preparation protocol<br>o Sequencing technology and equipment<br>o Post-sequencing read modification: trimming, filtering, downsampling. |
| Disaster recovery (for data storage) | Set of policies, tools and procedures to enable the recovery or continuation of vital technology infrastructure and systems following a natural or human-induced disaster.<br>The disaster recovery can be dedicated to storage and also to machines dedicated to computational capacity.<br>This should include at least the generation of backups with an appropriate frequency and their storage in a separate physical location. |

| Term | Description |
|---|---|
| | In addition, organisations also implement precautionary measures with the objective of preventing a disaster (i.e. mirrors of systems, antivirus, use of an uninterruptible power supply, fire prevention). |
| Epidemiological data | Data about the place and time related to the isolate as well as further characterisation. The data are submitted to the ECDC and EFSA MTS and from there a subset is sent to the Joint MTS, in accordance with the collaboration agreement. Examples:<br>- Isolates of any origin: date of sampling, date of arrival in the reference laboratory, reporting country, existing conventional typing information (e.g. serotype or pathotype) and other molecular data (e.g. MLVA, presence of genes).<br>- Human-origin isolates: age, gender, travel history. |
| FASTA file format | A text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. FASTA formatted files are normally used to represent assemblies. See further https://en.wikipedia.org/wiki/FASTA_format |
| FASTQ file format | A text-based format similar to FASTA for storing both a biological sequence (usually a nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single character. FASTQ formatted files are normally used to represent sequence reads. See further https://en.wikipedia.org/wiki/FASTQ_format. (Cock et al., 2010) |
| File transfer protocol (FTP) | The FTP is a standard network protocol used for the transfer of computer files between a client and server on a computer network. See further https://en.wikipedia.org/wiki/File_Transfer_Protocol. |
| Genome assembly | The process of constructing one or more contigs from sequence reads, in order to reconstruct the original complete genome to the extent possible. The parameters of the assembly must be adapted to the sequencing technology in terms of read length, and paired/single-end library preparations. The output is normally a FASTA format file containing the contigs, which should ideally be subjected to further post-assembly optimisation. |
| Indel | An insertion or deletion of bases in the genome of an organism. Source: https://en.wikipedia.org/wiki/Indel. Some sequencing technologies and bioinformatic steps (assembly or SNP/indel calling) are prone to generating/reporting false-positive indels, which usually demands their post-correction (see Post-assembly optimisation) or exclusion from the analysis (see Data analysis section ). |
| K-mer | When applied in the field of DNA/RNA sequencing, a k-mer commonly represents all the possible sub-sequences of length *k* from a read or other sequence. |
| Microbiological cluster | A set of isolates that have similar or identical genotypic or phenotypic properties. |
| Mixed culture | A non-pure culture contaminated only with other strains from the same species. |
| Mixed culture detection | Verifying that a culture is pure or mixed. Used as a quality control. |
| Mobilome | The set of genes associated with horizontal gene transfer. |
| Multiple-locus variable-number tandem repeat analysis (MLVA) | Typing method that measures the number of tandem repeats at specific loci. This is used in routine analysis for *Salmonella* Enteritidis and *Salmonella* Typhimurium in the EU/EEA, each using a schema of five loci. The raw output consists of peak data, from which the number of repeats is deduced per locus. A distance between two isolates can be defined as the number of loci that have a different number of repeats. Given the low number of loci used, a type can also be assigned by simply concatenating the number of repeats in each locus into a single code. |
| Non-human-origin isolates | Isolates originating from food, feed, animal and environmental samples. |
| Open reading frame (ORF) | A continuous stretch of codons that contain a start codon (usually AUG) and a stop codon (usually UAA, UAG or UGA). An ATG codon within the open reading frame (ORF) (not necessarily the first) may indicate where translation starts.<br>Source: https://en.wikipedia.org/wiki/Open_reading_frame; Adam et al. (online). |
| Pan genome | The set of loci present in at least one strain of a given species. In theory, the pan genome can always expand, each time a new locus is found in a new strain of the same species. In practice, when sufficient strains are initially available, the pan genome can also be defined |

| Term | Description |
|---|---|
| | as a fixed set of loci. In addition, technically problematic loci such as close paralogues, may be removed as well to improve the performance of wgMLST. |
| Paralogue | Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life. Homologous sequences are paralogous if they were created by a duplication event within the genome. For gene duplication events, if a gene in an organism is duplicated to occupy two different positions in the same genome, then the two copies are termed paralogues (Koonin, 2005). |
| Partial type | See Strain nomenclature. |
| Permanent storage | Data are never deleted and are available using the same interface, regardless of their age. In addition, data that are intended to be deleted may still be stored, and not used further, i.e. only logically rather than physically deleted. Older data can, however, be stored, e.g., on less performant but cheaper hardware. In addition, more performant but slower (lossless) compression can be used for different types of data such as older data or subsets of data. Examples of the latter are mapping reads to their assembly and storing only the mapping, or only describing which reads have been removed or shortened as part of read trimming or filtering. |
| Post-assembly optimisation | Changes to the contigs produced through genome assembly, in order to improve their quality. Contaminant DNA (originated during either sample DNA preparation or sequencing steps) can generate contigs that will pass the assembler filtering options. In addition, very small or obsolete contigs (e.g. contigs composed by undefined bases or a single long homopolymeric tract) can be generated due to assembler-derived artefacts. Optimisations can include:<br>- Contig removal: contigs can be removed in their entirety based on defined criteria: minimal size, GC content range, minimal k-mer coverage and minimum endpoint coverage per nucleotide (obtained by mapping the reads back to the assembly) to exclude possible contigs originating either from contamination (during sample DNA preparation or sequencing steps; e.g. DNA barcode bleaching) or from assembly artefacts.<br>- Assembly correction by mapping: the input reads are mapped to the contigs and based on this alignment the contigs are corrected for possible false SNPs/indels, e.g. by taking the consensus nucleotide for each position. |
| Pulsed-field gel electrophoresis (PFGE) | Typing method that measures the length of genomic fragments obtained by cleaving the genome at specific sites with a restriction enzyme. The raw output consists of an image of the gel in which the fragments ('bands') have been separated and stained. From this, the length of the fragments is determined by interpolating with a standard set of fragments of known size run on the same gel. The interpretation requires substantial manual intervention to determine, e.g., if a band is present or not, if more than one fragment is contributing to the same band or which exact point within the band should be used to determine the length. A distance between two isolates can be defined as the number of fragments that are different between them, given some tolerance in their exact length and usually excluding fragments below a particular minimum length. More than one restriction enzyme can be used to independently generate a second set of fragment lengths to increase resolution. Typically, a code (type) is manually assigned to each unique pattern of fragment lengths, which also requires regular revision as more isolates are typed due to the non-transitivity of the distance metric. |
| Quality assurance and quality control (QA/QC) procedures | Quality assurance (QA) procedures refer to proactive processes aimed at measuring and assuring quality, while Quality Control (QC) procedures are a set of reactive activities focused on identifying and correcting defects towards quality improvement. The combination of these two elements is key for any laboratory quality management system focused on providing valid and correct results, as well as on promoting inter-laboratory concordance of results and harmonisation of the interpretation of data. See, e.g., WHO (2011). |
| Resistome | The set of genes within a genome that are associated with antimicrobial resistance. |
| Secure copy protocol (SCP) | The SCP is a network protocol that supports file transfers between hosts on a network. An SCP uses Secure Shell (SSH) for data transfer and uses the same mechanisms for |

| Term | Description |
|---|---|
| | authentication, thereby ensuring the authenticity and confidentiality of the data in transit. See further https://en.wikipedia.org/wiki/Secure_copy. |
| Sequence read data formats | FASTQ, FAST5, HDF5, BAM and CRAM. |
| Raw sequence reads | Reads that are not further processed other than adapter removal. Ideally, these are the sequence reads that should be submitted to the system. |
| Reference database | A dataset, typically in the form of a relational database, containing the information required to perform a particular analysis. Typically, a reference database refers to panels of genetic markers (usually loci or SNPs) defining a traditional type (i.e. types obtained by traditional genotyping methods) or correlating to a given phenotype such as resistance. Examples include: the loci and reference alleles that define a core genome for cgMLST, marker genes and alleles that define *L. monocytogenes* serogroups, the combination of pathotype-specific markers that can guide the *in silico* classification of *E. coli* pathotypes, and the repertoire of SNPs mediating a specific antibiotic resistance phenotype. The construction of a reference database (i.e. the choice of the presence/absence profiles or repertoire of SNPs defining a given pathotype or resistance profile) is a critical step for a reliable analysis regardless of the bioinformatics approach applied (e.g. reads mapping against reference databases or direct screening of assemblies). |
| Reference implementation | An implementation of a specification (e.g. an algorithm) to be used as a definitive interpretation for that specification.<br>See further https://en.wikipedia.org/wiki/Reference_implementation. |
| Secure file transfer protocol (SFTP) | The SSH file transfer protocol (also secure file transfer protocol, or SFTP) is a network protocol that provides secure file access, file transfer, and file management over any reliable data stream.<br>See further https://en.wikipedia.org/wiki/SSH_File_Transfer_Protocol. |
| Single nucleotide polymorphism (SNP)/single nucleotide variants (SNV) | SNP or SNV are variations on a single position of a sequence, usually obtained by comparing the reads to a reference sequence that can be used for phylogenetic inference. |
| SNP/SNV calling and filtering | The process of aligning the reads or assemblies of one isolate to one reference genome and subsequently extracting the list of true SNPs for that isolate. The accuracy of short-read mapping is important, especially in regions prone to generate SNP-calling errors such as indels and repeat regions. Criteria to validate SNPs include: minimum mapping quality at variant position, minimum number of reads covering the variant position, minimum base quality at variant position, minimum proportion of reads at variant position differing from the reference to consider a position as homozygous, strand bias and neighbouring base quality (Li et al., 2008). Any SNPs called in the resulting alignment to the reference genome that have a low probability of being true SNPs based on the defined criteria must be filtered out. Also, true SNPs with a high likelihood of falling within recombination regions may have to be ignored/excluded before proceeding with phylogenetic inferences, depending on the species (or lineage) and on the set of isolates under comparison. It is worth noting that SNPs are normally called based on reads as input, but it can also be done based on an assembly, provided that the assembly is of good quality. If a very strong confidence in SNPs is desirable, two independent programmes can be used and only the variants called by both retained. However, if the purpose is to cluster isolates, which are epidemiologically linked, a very stringent SNP discovery procedure will tend to favour grouping and may lead to erroneous epidemiological interpretations. |
| Standards for descriptive data about sequences | The Global Alliance for Genomics and Health (GA4GH), GMI defined standards (ENA/SRA required data), NGSOnto (https://bioportal.bioontology.org/ontologies/NGSONTO). |
| Standards for epidemiological data | ECDC TESSy metadata, EFSA Standard Sample Description 2 (EFSA, 2013), EFSA Guidance on Data Exchange version 2.0 (EFSA, 2014b). Standards for transfer of epidemiological data: csv, xml. |
| Strain nomenclature | Strain nomenclature is intended to provide a classification of isolates according to their (phylo)genetic relatedness within the diversity of the species, which is pivotal for simple and rapid communication among the different players involved in routine surveillance and/or outbreak investigation (Nadon et al., 2017). Operationally this can be done by defining types, which are short human-readable codes that represent a set of strains that have a minimum genetic relatedness to each other. This genetic relatedness can be |

| Term | Description |
|---|---|
| | defined, e.g., through a clustering algorithm applied to allelic distances (cgMLST/wgMLST) or SNP distances. A hierarchical strain nomenclature can also be defined by, e.g., establishing several similarity cut-offs and a corresponding naming scheme for the types. To establish a strain nomenclature, a comprehensive strain database for each species should be compiled including several well-defined and confirmed outbreak-related strains. This constitutes the basis for the determination of the cut-off thresholds needed for defining well-known lineages/clonal complexes for the species in analysis or that represent outbreak level. Each threshold of similarity then defines a nomenclature identifier that should correspond to well-defined lineages of a given species. The combination of identifiers resulting from different cut-offs (i.e. sublevels) constitutes the strain nomenclature. However, depending on the threshold level chosen for the classification, the nomenclature can become unstable due to clone diversification, leading to the joining of different sublevels. The probability of such occurrence is higher near the leaves of a hierarchical dendrogram. As such, while high-level nomenclature has longer term stability, this is not the case for low-level nomenclature. This caveat should be carefully communicated, so that its operational implications are well understood by users. It is also worth noting that the stability and accuracy of such nomenclature is dependent on the species, it being currently not well-established whether such low-level nomenclature can ever be applied for some species (or lineages within species). Other solutions should be envisaged for these cases, although high-level nomenclature (like sequence types or clonal complexes for seven-gene MLST) can always be derived (ECDC, 2015; Nadon et al., 2017). |
| Species confirmation | Verifying that the species detected from the reads or assembly corresponds to the expected species or genus. Used as a quality control to detect, e.g., a sample mix-up. |
| Trimming | Removing entire reads or sections at either edge of the read based on the PHRED quality score of each position and accepted minimal read length (see Li at el., 2008). The trimming to be performed is dependent on the sequencing platform used to generate the reads. |
| Type | In phylogeny, a strain type is defined as a monophyletic group or a clade in a phylogenetic tree. Depending on the species this can also be defined as a lineage. Strain nomenclature approaches aim to assign an identifier to each type. See Strain nomenclature. |
| VCF file format | The variant calling file (VCF) format allows for the storage of gene sequence variation in text format. It is usually obtained after the process of SNP/SNV calling. See further https://en.wikipedia.org/wiki/Variant_Call_Format. |
| Virulome | The set of genes within a genome associated with virulence. |
| wgMLST | See Core/whole genome MLST (cgMLST/wgMLST) |

## Abbreviations

| | |
|---|---|
| cgMLST | Core genome multilocus sequence typing |
| ECDC | European Centre for Disease Prevention and Control |
| EFSA | European Food Safety Authority |
| ENA | European Nucleotide Archive |
| EPIS | Epidemic Intelligence Information System |
| EURL | European Union Reference Laboratory |
| EWRS | Early Warning and Response System |
| FTP | File transfer protocol |
| FV | Food and veterinary |
| MLST | Multilocus sequence typing |
| MLVA | Multiple loci variable-number tandem repeat analysis |
| MS | Member State |
| MTS | Molecular Typing System |
| NGS | Next generation sequencing |
| NMFP | National Focal Points for Microbiology |
| NRL | National Reference Laboratory |
| ORF | Open reading frame |
| PFGE | Pulsed-field gel electrophoresis |
| PH | Public health |
| QA | Quality assurance |
| QC | Quality control |
| RASFF | Rapid Alert System for Food and Feed |
| SCP | Secure copy protocol |
| SFTP | Secure file transfer protocol |
| SNP | Single nucleotide polymorphism |
| SNV | Single nucleotide variant |
| SRA | Sequence Read Archive |
| SSH | Secure shell |
| STEC | Shiga-toxin-producing *E. coli* |
| TESSy | The European Surveillance System |
| ToR | Term of Reference |
| UPGMA | Unweighted Pair Group Method with Arithmetic Mean |
| VCF | Variant calling file |
| wgMLST | Whole genome multilocus sequence typing |
| WGS | Whole genome sequencing |

# Appendix A – Requirements assessment summary

The number and proportion of requirements met by each solution for each functionality, i.e. set of requirements, is shown in Table A1 (critical requirements, identical to Table 7 in Section 3.8.1), Table A2 (medium requirements) and Table A3 (optional requirements). Figures A1 and A2 display these numbers graphically as spider graphs. As discussed in Section 3.8.1, counting the number of requirements met was not selected as an appropriate methodology to determine the most suitable solution or combination of solutions, and therefore the numbers below should be interpreted with care.

**Table A1:** Number and proportion of critical requirements met by each solution and per functionality[(a)]

| Solution | Data collection (n=34) | Reads QC (n=7) | Assembly (n=5) | wgMLST (n=15) | Strain nomen-clature (n=11) | Genome character-isation (n=10) | General (n=20) | Infra-structure (n=7) |
|---|---|---|---|---|---|---|---|---|
| BIGSdb | 22 (65%) | 0 (0%) | 0 (0%) | 11 (73%) | 10 (91%) | 1 (10%) | 13 (65%) | 1 (14%) |
| BioNumerics[(b)] | 26 (76%) | 6 (86%) | 5 (100%) | 13 (87%) | 6 (55%) | 6 (60%) | 20 (100%) | 3 (43%) |
| CGE | 11 (32%) | 0 (0%) | 4 (80%) | 8 (53%) | 1 (9%) | 10 (100%) | 4 (20%) | 6 (86%) |
| Cloud Services | 10 (29%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 7 (100%) |
| COMPARE | 30 (88%) | 4 (57%) | 4 (80%) | 8 (53%) | 1 (9%) | 10 (100%) | 7 (35%) | 6 (86%) |
| ENA | 30 (88%) | 0 (0%) | 2 (40%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (15%) | 7 (100%) |
| EnteroBase | 24 (71%) | 5 (71%) | 5 (100%) | 11 (73%) | 9 (82%) | 5 (50%) | 16 (80%) | 6 (86%) |
| INNUENDO | 22 (65%) | 7 (100%) | 5 (100%) | 11 (73%) | 6 (55%) | 10 (100%) | 14 (70%) | 2 (29%) |
| IRIDA | 26 (76%) | 7 (100%) | 4 (80%) | 2 (13%) | 0 (0%) | 6 (60%) | 12 (60%) | 3 (43%) |
| PathogenWatch | 16 (47%) | 0 (0%) | 1 (20%) | 8 (53%) | 0 (0%) | 3 (30%) | 11 (55%) | 4 (57%) |
| SeqSphere | 21 (62%) | 6 (86%) | 5 (100%) | 11 (73%) | 4 (36%) | 4 (40%) | 18 (90%) | 5 (71%) |

QC: quality control.

(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

(b): Two requirements met for Data Collection refer to PFGE data.

**Table A2:** Number and proportion of medium requirements met by each solution and per functionality[(a)]

| Solution | Data collection (n=15) | Reads QC (n=2) | Assembly (n=0) | wgMLST (n=3) | Strain nomen-clature (n=0) | Genome character-isation (n=3) | General (n=5) | Infra-structure (n=4) |
|---|---|---|---|---|---|---|---|---|
| BIGSdb | 11 (73%) | 0 (0%) | n.a. | 2 (67%) | n.a. | 0 (0%) | 4 (80%) | 3 (75%) |
| BioNumerics | 8 (53%) | 0 (0%) | n.a. | 1 (33%) | n.a. | 3 (100%) | 4 (80%) | 2 (50%) |
| CGE | 6 (40%) | 1 (50%) | n.a. | 1 (33%) | n.a. | 2 (67%) | 5 (100%) | 4 (100%) |
| Cloud Services | 6 (40%) | 0 (0%) | n.a. | 0 (0%) | n.a. | 0 (0%) | 0 (0%) | 3 (75%) |
| COMPARE | 12 (80%) | 0 (0%) | n.a. | 1 (33%) | n.a. | 2 (67%) | 2 (40%) | 2 (50%) |
| ENA | 10 (67%) | 0 (0%) | n.a. | 0 (0%) | n.a. | 0 (0%) | 1 (20%) | 2 (50%) |
| EnteroBase | 10 (67%) | 0 (0%) | n.a. | 2 (67%) | n.a. | 1 (33%) | 4 (80%) | 4 (100%) |
| INNUENDO | 5 (33%) | 2 (100%) | n.a. | 1 (33%) | n.a. | 2 (67%) | 3 (60%) | 0 (0%) |
| IRIDA | 12 (80%) | 0 (0%) | n.a. | 0 (0%) | n.a. | 0 (0%) | 4 (80%) | 0 (0%) |
| PathogenWatch | 6 (40%) | 0 (0%) | n.a. | 0 (0%) | n.a. | 0 (0%) | 3 (60%) | 3 (75%) |
| SeqSphere | 9 (60%) | 1 (50%) | n.a. | 2 (67%) | n.a. | 1 (33%) | 3 (60%) | 2 (50%) |

n.a.: not applicable. QC: quality control.

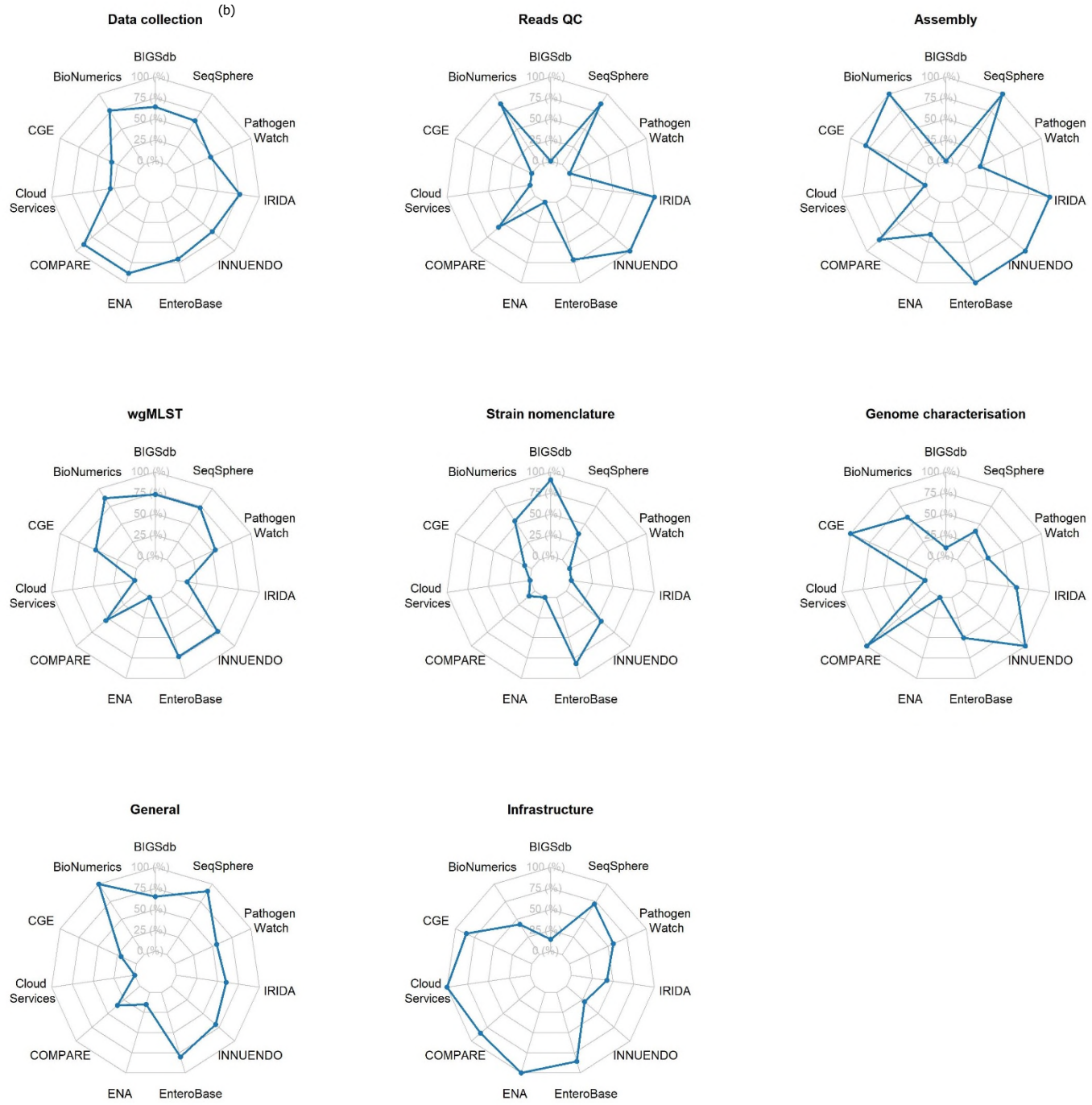
(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.

**Table A3:** Number and proportion of optional requirements met by each solution and per functionality[(a)]

| Solution | Data collection (n=4) | Reads QC (n=0) | Assembly (n=0) | wgMLST (n=0) | Strain nomen-clature (n=4) | Genome character-isation (n=6) | General (n=0) | Infra-structure (n=0) |
|---|---|---|---|---|---|---|---|---|
| BIGSdb | 1 (25%) | n.a. | n.a. | n.a. | 3 (75%) | 0 (0%) | n.a. | n.a. |
| BioNumerics | 2 (50%) | n.a. | n.a. | n.a. | 1 (25%) | 0 (0%) | n.a. | n.a. |
| CGE | 1 (25%) | n.a. | n.a. | n.a. | 0 (0%) | 3 (50%) | n.a. | n.a. |
| Cloud Services | 1 (25%) | n.a. | n.a. | n.a. | 0 (0%) | 0 (0%) | n.a. | n.a. |
| COMPARE | 4 (100%) | n.a. | n.a. | n.a. | 0 (0%) | 3 (50%) | n.a. | n.a. |
| ENA | 4 (100%) | n.a. | n.a. | n.a. | 0 (0%) | 0 (0%) | n.a. | n.a. |
| EnteroBase | 1 (25%) | n.a. | n.a. | n.a. | 1 (25%) | 0 (0%) | n.a. | n.a. |
| INNUENDO | 1 (25%) | n.a. | n.a. | n.a. | 0 (0%) | 5 (83%) | n.a. | n.a. |
| IRIDA | 2 (50%) | n.a. | n.a. | n.a. | 0 (0%) | 0 (0%) | n.a. | n.a. |
| PathogenWatch | 0 (0%) | n.a. | n.a. | n.a. | 0 (0%) | 0 (0%) | n.a. | n.a. |
| SeqSphere | 2 (50%) | n.a. | n.a. | n.a. | 3 (75%) | 0 (0%) | n.a. | n.a. |

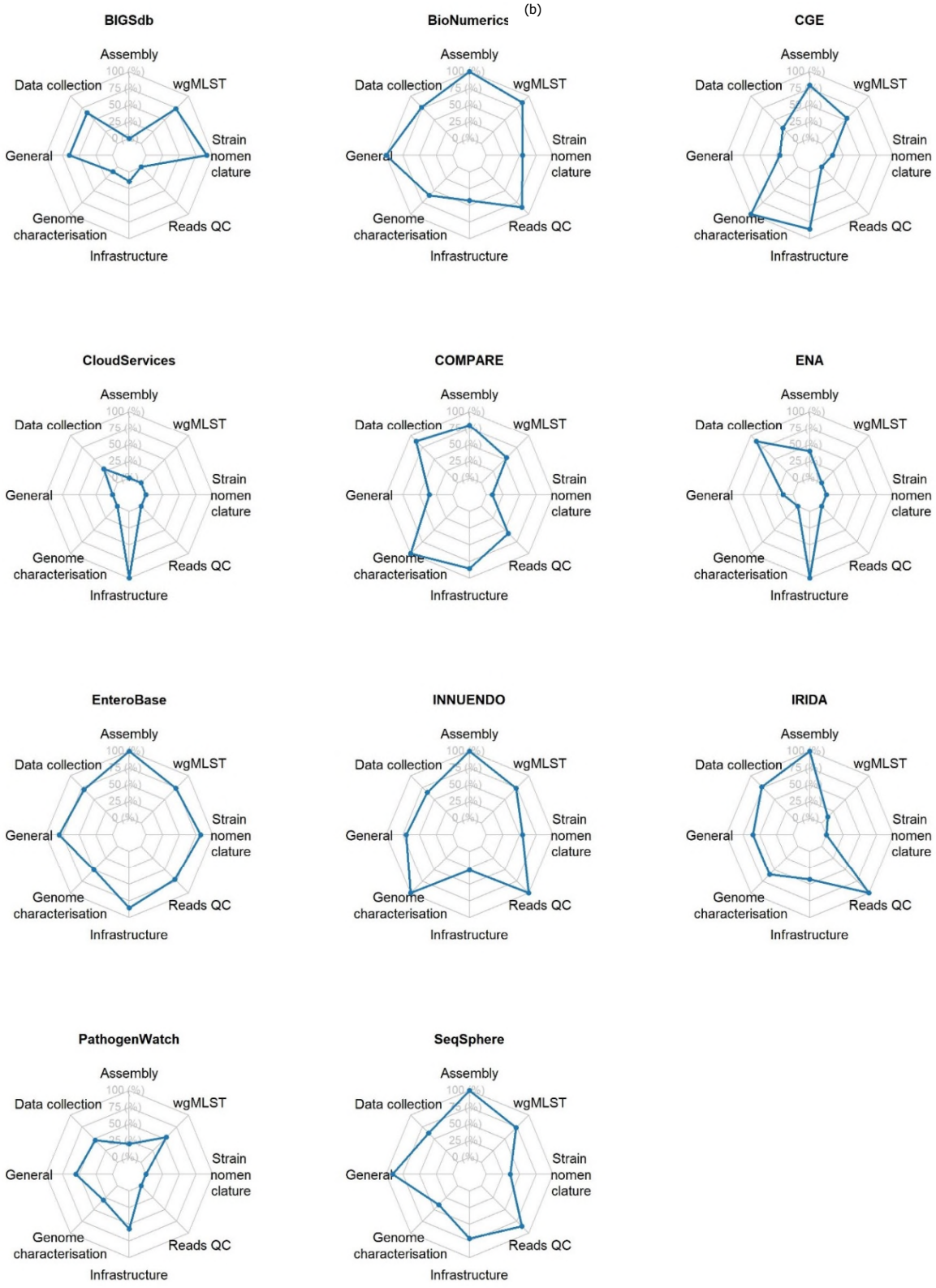
n.a.: not applicable. QC: quality control.

(a): Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.




(a) Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.
(b) Two of those requirements met for Data Collection refer to PFGE data.

**Figure A1:** Proportion of critical requirements met per functionality for all solutions[(a)] (total number of critical requirements per functionality indicated in Table A1)




(a) Assessment made to the best knowledge of the JWG experts, based on their own expertise, publicly available information and information provided by hearing experts as explained in Sections 2.1.3, 2.2.3 and 3.8.1. Status as of 31 December 2018. Further developments to the different solutions since that date are not presented here.
(b) Two of those requirements met for BioNumerics Data Collection refer to PFGE data.

**Figure A2:** Proportion of critical requirements met per solution for all functionalities[(a)] (total number of critical requirements per functionality indicated in Table A1)