SCIENTIFIC ADVICE

Expert Opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA

www.ecdc.europa.eu

ECDC SCIENTIFIC ADVICE

Expert Opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA

This report of the European Centre for Disease Prevention and Control (ECDC) was coordinated by Ivo Van Walle and produced by the FWD-NEXT Expert Group (ECDC expert group on introduction of next-generation typing methods for surveillance of food- and waterborne diseases).

Contributing authors and members of the expert group
Ariane Pietzka, Eva Møller Nielsen, Kristoffer Kiel, Ivelina Damjanova, Valeria Michelacci, Joël Mossong, Eelco Franz, Wilfrid van Pelt, Tomasz Wołkowicz, Vítor Borges, Cecilia Jernberg, Chris Lane, Ian Fisher, Tansy Peters, Joakim Ågren, Valentina Rizzi, Maria Teresa Da Silva Felicio, Marc Struelens, Daniel Palm, Johanna Takkinen, Ivo Van Walle

Acknowledgements
ECDC wishes to thank everyone for their valuable input. In some instances, members gathered further input from their own organisations in order to improve the quality of the report. In particular, substantial input was provided by Philip Ashton (United Kingdom) and Erik Alm (Sweden). The full draft of the report was shared with the entire EU/EEA FWD network for further comments, which were subsequently approved by the group and incorporated into the text.

Suggested citation: European Centre for Disease Prevention and Control. Expert Opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA. Stockholm: ECDC; 2015.

Stockholm, October 2015

ISBN 978-92-9193-723-3

doi 10.2900/453641

Catalogue number TQ-02-15-849-EN-N

Contents

Abbreviations iv
Definitions iv
Executive summary v
Expert group composition vi

1 Introduction 1
2 Sample provision and sequencing 2
2.1 Sampling frame 2
2.2 Current and future scenarios for diagnosis and typing 3
2.3 Whole genome sequencing 4
2.3.1 DNA extraction 4
2.3.2 Equipment and protocols 5
2.3.3 Data storage and sharing 6
2.3.4 Cost and time comparison of WGS with current typing techniques 7
3 WGS data analysis for public health 9
3.1 Raw read quality and filtering 9
3.2 Genome assembly 9
3.3 Alignment to a reference genome 10
3.4 Deriving phylogeny 11
3.5 Nomenclature assignment 13
3.6 Prediction of phenotypes and genotypes 15
3.7 Visualisation 16
4 Collaboration between organisations 18
4.1 Model process for data analysis at national level 18
4.2 Cluster detection and follow-up at the international level 22
4.3 Collaborative resources and databases 24
4.4 Quality assurance and transition to WGS 28
5 Conclusions 30
5.1 Sample selection 30
5.2 Sequencing, cost and timeliness 30
5.3 Data storage and analysis 30
5.4 Collaboration 31
5.5 Future steps 32

References 33

Figures

Figure 1. wgMLST principle 13
Figure 2. Taxonomical nomenclature principle based on SNP or wgMLST dendrogram 15
Figure 3. Databases and other resources used during the model process at national level 19
Figure 4. wgMLST allele identifier retrieval using a cached nomenclature database 26
Figure 5. Schematic overview of the minimum data to be stored in (a) a wgMLST nomenclature database and (b) a taxonomical nomenclature database 27

Tables

Table 1. Samples available for typing by country 3
Table 2. Techniques for diagnostics and typing 4
Table 3. General scenarios for molecular surveillance now and in the future 4
Table 4. WGS data storage requirements per isolate 7
Table 5. Cost and timeliness of different typing techniques including WGS, as assessed in June 2015 8
Table 6. Model process for data analysis, sharing and action at national level 18
Table 7. Relevant phenotypes and genotypes that could be predicted 21
Table 8. WGS data sharing options and corresponding analyses 24
Table 9. Minimum requirements for a wgMLST nomenclature database that must all be fulfilled 26
Table 10. Proposed future steps 32

Abbreviations

cgMLST	Core genome multilocus sequence typing
EFSA	European Food Safety Authority
EURL	European Union Reference Laboratory
FWD	Food- and waterborne diseases
MLST	Multilocus sequence typing
MLVA	Multiple-locus variable-number tandem-repeat analysis
PFGE	Pulsed-field gel electrophoresis
SNP	Single nucleotide polymorphism
wgMLST	Whole genome multilocus sequence typing
WGS	Whole genome sequencing

Definitions

Accessory genome	The subset of loci that is not common to all or the large majority of the strains of one species.
Contig	Contiguous sequence. The raw reads produced by a sequencer can be assembled into one or more non-overlapping contigs which together constitute the assembled genome.
Core genome	The subset of loci that is common to all or the large majority of the strains of one species.
Coverage	The average number of times each base in the genome is contained in individual raw reads. The raw reads are pre-processed first to remove bases where the quality is too low.
k-mers	Very short overlapping sequences typically of around 15 bases, derived in silico from raw reads or contigs.
N50	The length of the smallest contig among the set of the largest contigs that together cover at least 50% of the assembled genome.
Pan-genome	The set of all loci observed in all strains of one species, i.e. the core genome plus the accessory genome.
Raw reads	Sequence data generated by a DNA sequencer. When whole genome DNA from a pure isolate is sequenced, the raw reads are a set of short sequences that partially overlap with each other and together are intended to cover the entire genome of the cells from which the DNA was extracted. Each individual base in a raw read has a quality score assigned, calculated by the DNA sequencer.
Shared genome and shared accessory genome	Among a set of strains from the same species, the shared genome is the subset of loci they have in common. The shared accessory genome is within the shared genome the subset of loci that are not part of the core genome.

Executive summary

Molecular typing, and typing in general, has a long history of application to surveillance of food- and waterborne diseases (FWD) for public health purposes. The advent of new microbial typing and detection techniques, in particular whole genome sequencing (WGS) and culture-independent diagnostic methods, brings about a fundamental change in the way diagnostics and typing are performed for foodborne infections. To respond to the strategic needs related to the introduction of WGS into public health microbiology and facilitate the further development of EU/EEA-wide molecular-typing-enhanced surveillance for FWD, ECDC sent a call for interest to public health experts to form an expert group on introduction of next-generation typing methods for surveillance of food- and waterborne diseases (FWD-NEXT Expert Group), composed of microbiologists, epidemiologists and bioinformaticians.

This report, produced by the FWD-NEXT Expert Group, focuses on four pathogens: *Salmonella*, *Listeria monocytogenes*, verocytotoxin–producing *Escherichia coli* (VTEC) and *Campylobacter*. It is written from the country perspective and covers the entire process from sample provision and sequencing to data analysis, and finally data sharing and collaboration between different organisations. As such it serves to inform and support countries that are planning to, or are in the process of, implementing WGS for routine surveillance and outbreak investigation of FWD. The main audiences of the report are therefore national public health reference laboratories, national level epidemiologists and their immediate stakeholders. The report will also contribute to the ECDC strategy and roadmap for integration of molecular typing into European-level surveillance, response and epidemic preparedness, in particular for food- and waterborne diseases. In addition, IT departments of organisations that will be performing WGS on a routine basis may also find parts of the report relevant for their planning, especially on data storage and computing capacity requirements.

It is expected that WGS will eventually become the sole standard method for genotyping of FWD pathogens for public health purposes, with additional phenotypic tests such as antimicrobial resistance only being performed in situations where the phenotype cannot be reliably predicted from the sequence. In the meantime, laboratories that already perform WGS should ideally also still perform the current typing techniques if necessary on selected isolates, such as outbreak-related ones, so that data remain comparable across organisations and can be used for surveillance purposes and further validation during the transition phase. Largely independent of the typing method, it would be very beneficial to have isolates from positive samples (from food, food processing plants, animals, feed and the environment) available for real-time typing as well, since this allows complementing the traditional route of epidemiological investigation that is often unable to identify a potential vehicle or does so when the corresponding batches are no longer available for sampling. A legal framework, as already in place in some countries, can be very helpful, even if it is not required, to ensure that national reference laboratories receive a sufficient selection of both human and positive food samples for further typing.

A cost comparison between WGS and current typing methods indicates that already, on a per isolate basis, WGS can be less expensive than current typing methods for *E. coli* and *Campylobacter*. For *Listeria*, the cost is more or less the same, and for *Salmonella*, depending on the throughput, the cost can still be somewhat higher. The total time required for WGS is already comparable to that of current typing methods. As WGS technology is still evolving rapidly, the cost and total time can be expected to decrease further and become less expensive for all pathogens compared with current typing methods. Taking into account the higher accuracy of the method for delineating epidemiologically relevant clusters, the potential for preventing additional cases through earlier detection of clusters and outbreaks, as well as identification of the vehicle, is also higher than for current typing methods.

The actual laboratory work using WGS as the standard genotyping method will become simpler as only one genotyping method needs to be used per pathogen. In addition, the differences between pathogens are often small and limited to the DNA extraction process rather than to the preparation of the sequence library that forms the input material for the sequencer. It is therefore also easier to pool typing capacity. At the same time however, protocols for DNA library preparation provided by the manufacturer often benefit from some optimisation and hence are not yet standardised. The inter- and intralaboratory reproducibility of WGS results also needs to be assessed better.

WGS is different from current typing methods in the sense that in addition to the actual laboratory work also substantial subsequent data processing, storage and analysis is required in order to extract useful information from the large amount of generated data. The required storage capacity needs to be taken into account when planning the introduction of the method. It likely only becomes an important factor when reaching several thousands of isolates and consequently the terabyte range. Similarly, computing capacity also needs to be taken into account, but likely only becomes an important factor when processing more than one hundred isolates per week.

The routine analyses of WGS data for public health purposes are not yet standardised. However, a model process for routine analysis is described, including nomenclature assignment, cluster detection, prediction of relevant phenotypes and use of prior epidemiological knowledge. Many of these steps involve the usage of collaborative resources or databases, which are each described separately. The WGS nomenclature database for a particular

pathogen is the most important of these collaborative resources, enabling effective communication between organisations and standardised analyses. Since there is currently no agreed standard database for any pathogen, minimum requirements are proposed for any such database in order to be acceptable as a standard. In addition, two types of WGS nomenclature are described, one based on core genome multilocus sequence typing (cgMLST), and one that is a true hierarchical or taxonomical classification.

Collaboration and data sharing between organisations and countries is required due to the international dimension of FWD pathogens and food trade in particular. National public health reference laboratories form the first line of such collaboration, as they are usually the first to have sufficient information, i.e. the microbiological typing data, to allow linkage of cases at national level and subsequent detection of human clusters or outbreaks. Ideally, as soon as these typing data are available, they should also be sent to an international database to allow timely detection of microbiological clusters at the international level. The typing data have to be accompanied by some descriptive data, in particular a relevant date such as the date of sampling. For WGS typing data, several options are possible for what can be sent through, from raw reads to assembled genomes to nomenclature only. The most realistic option at present seems first to submit nomenclature, and then when a cluster is detected based on that, actual sequence data can be sent through as well to allow more detailed analysis.

Expert group composition

The composition of the expert group is given below, in alphabetical order by country.

Country	Organisation	Name
Austria	Agentur für Gesundheit und Ernährungssicherheit	Ariane Pietzka
Denmark	Statens Serum Institut	Eva Møller Nielsen and Kristoffer Kiel (alternating)
Hungary	Országos Epidemiológia Központ	Ivelina Damjanova
Italy	Istituto Superiore di Sanità	Valeria Michelacci
Luxembourg	Laboratoire National de Santé	Joël Mossong
The Netherlands	Rijksinstituut voor Volksgezondheid en Milieu	Eelco Franz
The Netherlands	Rijksinstituut voor Volksgezondheid en Milieu	Wilfrid van Pelt
Poland	Narodowy Instytut Zdrowia Publicznego	Tomasz Wołkowicz
Portugal	Instituto Nacional de Saúde Doutor Ricardo Jorge	Vítor Borges
Sweden	Folkhälsomyndigheten	Cecilia Jernberg
United Kingdom	Public Health England	Chris Lane, replaced by Ian Fisher
United Kingdom	Public Health England	Tansy Peters
N/A	EU Reference Laboratory for *Campylobacter*	Joakim Ågren
N/A	European Food Safety Agency	Valentina Rizzi and Maria Teresa Da Silva Felicio (alternating)
N/A	European Centre for Disease Prevention and Control	Marc Struelens
N/A	European Centre for Disease Prevention and Control	Daniel Palm
N/A	European Centre for Disease Prevention and Control	Johanna Takkinen
N/A	European Centre for Disease Prevention and Control	Ivo Van Walle (chair)

1 Introduction

Molecular typing, and typing in general, has a long history of application to surveillance of food- and waterborne diseases (FWD) for public health purposes, in particular for *Salmonella*, *Listeria monocytogenes* and *Escherichia coli*. It allows microbiologically meaningful linkage of cases, and subsequently timely detection of clusters and outbreaks. It also provides evidence for potential vehicles of infection during the foodborne outbreak investigation if isolates from food have been available for molecular typing. This in turn enables implementation of targeted control and prevention measures at the national or EU level. In addition, molecular typing makes it possible to identify persistent strains causing recurrent human infections that are likely originating from a continuous source, and would thus inform larger-scale policy actions such as introduction of EU-wide control programmes in the food chain.

The advent of new microbial typing and detection techniques, in particular whole genome sequencing (WGS) and culture-independent diagnostic methods, brings about a fundamental change in the way diagnostics and typing are performed for foodborne infections. Recent publications and outbreak reports have shown that WGS can provide higher resolution to typing to inform public health actions. The methodology has shown improved sensitivity, specificity and more timely resolution to outbreak clustering compared with traditional methods [1,2]. This will substantially impact the integration of typing into surveillance and outbreak detection/investigation of FWD both at the national and EU level. The typing techniques currently supported by ECDC as validated and standardised methods for the routine application in this regard are, depending on the pathogen, serotyping, multiple-locus variable-number tandem-repeat analysis (MLVA) and pulsed-field gel electrophoresis (PFGE).

While development of tools for analysis and interpretation of WGS data is rapidly progressing, there is at present much uncertainty and no consensus on translating them into practically useful information for public health purposes. In addition, culture independent diagnostic methods will require updating existing legislation on case definitions for EU level surveillance and likely also minimum requirements either for providing samples for characterisation to national public health reference laboratories or ensuring decentralised characterisation.

In accordance with the ECDC strategy and roadmap for integration of molecular typing into European level surveillance and epidemic preparedness, ECDC has piloted molecular surveillance for *Salmonella*, *Listeria monocytogenes* and *Escherichia coli* during the period of November 2012 – May 2014, using PFGE and MLVA as typing methods. The evaluation of the pilot was positive and countries unanimously supported the continuation of the integration of molecular typing data into EU level surveillance for FWD and encouraged ECDC efforts to develop a common European approach to applying WGS technology in close partnership with the European Food Safety Authority (EFSA).

The European Commission has requested both ECDC and EFSA to each establish a database for collection of molecular typing data of isolates from human and non-human origin, respectively. In addition, it requested both agencies to jointly analyse, in collaboration with the respective European Union Reference Laboratories (EURLs), molecular typing data on *Salmonella*, *Listeria monocytogenes* and *Escherichia coli* isolates from both human and food/animal origin, with *Campylobacter* and other FWD pathogens optionally to be covered later. The EFSA Biohazard Panel has issued two Opinions that further prioritise WGS as the optimal technology for the near future and call for concerted cross-sectorial and international harmonisation of methods and data management coordination. The main purpose of typing of foodborne pathogens in this context is to facilitate and support outbreak investigation, prevention and control activities, including source attribution studies.

To respond to the strategic needs related to the introduction of WGS into public health microbiology and facilitate the further development of EU/EEA-wide molecular surveillance for FWD in the light of technology changes and the cross-sectorial use of molecular typing data, ECDC sent a call for interest to public health experts to form an expert group on introduction of next-generation typing methods for surveillance of food- and waterborne diseases (FWD-NEXT), composed of microbiologists, epidemiologists and bioinformaticians.

The FWD-NEXT expert group focuses on four food- and waterborne pathogens: *Salmonella*, *Listeria monocytogenes*, verocytotoxin-producing *Escherichia coli* (VTEC), and *Campylobacter*. The group's main output is this report. It is written from the country perspective and covers the entire process surrounding typing and WGS in particular, from sample provision and sequencing to data analysis and finally data sharing and collaboration between different organisations.

This report serves to inform and support countries that are planning to, or are in the process of, implementing WGS for routine surveillance and outbreak investigation of FWD. The main audiences of the report are therefore public health reference laboratories, national level epidemiologists and their immediate stakeholders. The report will also contribute to the ECDC strategy and roadmap for integration of molecular typing into European-level surveillance, response and epidemic preparedness, and in particular for food- and waterborne diseases. In addition, IT departments of organisations that will be performing WGS on a routine basis may also find parts of the report, especially on data storage and computing capacity, relevant for their planning.

2 Sample provision and sequencing

2.1 Sampling frame

Public-health action based on laboratory surveillance of pathogens found in samples of human origin benefits from a defined or recommended sampling frame. Ideally, this should be tailored to achieve specific surveillance objectives within a country or at EU level, and in particular with respect to what samples and data should be sent to the corresponding national reference laboratory. This allows, in spite of variations between countries, a relatively good understanding of the representativeness of the data, and ideally also defines the recommended minimum timeliness for sending through data and samples. How the data should be reported, in terms of standard variables, terms and what is mandatory, should also be specified. These data, along with typing data derived from the samples, can in turn be used to detect public health threat signals, assess their significance and subsequently can inform the decision whether to take action.

While defining a sampling frame is normally not an issue, in practice economical and technical factors usually determine how many samples are typed from which regions and with which timeliness. For the local primary laboratories that collect and receive the samples in the first instance, the main objective is clinical care. This includes identifying pathogenic species in the sample, if any, and sometimes also, for example, determining antimicrobial susceptibility in order to support physicians in diagnosis and treatment decision-making for individual patients. In the case of FWD pathogens, any other typing than diagnostic testing that would be relevant only for public health purposes, i.e. for the public in general, is normally not in the direct interest of primary laboratories, except in some cases for research purposes, such as in academic hospitals.

In contrast to primary laboratories, public health laboratories have as part of their tasks to perform typing for public health purposes. The primary laboratories therefore normally provide positive pure cultures of isolates, i.e. where a single colony is picked and grown, or specimens to the public health laboratory. Further characterisation and typing is then performed at the national level, mainly to try to detect similarities between isolates that are indicative of a common exposure of the patients in question and to monitor the emergence of new genetic variants. In the case of FWD pathogens, the exposure is usually to contaminated food. In a few countries, the provision of isolates to the public health laboratories is required by law, and a structural solution including coverage of the cost for sending the material therefore exists. In countries without such legal requirement, the transport costs, when not covered by the public health laboratory, frequently prevent a substantial part, or even any isolates, from being sent. In some cases, the provision of isolates is also dependent on the dedication of personnel within the primary laboratories. On the other hand, it is frequently an incentive for primary laboratories to provide isolates when they receive back the typing results from the reference laboratory.

An overview of the situation in the countries of the members of the Expert Group is given in Table 1. From these data, and taking into account all the factors mentioned above, it is clear that the sampling frame varies significantly and will continue to vary by country, at least in the near future. For identifying widespread clones or persistent sources however, an EU-wide sampling frame would be necessary – and this should be a long term goal.

Finally, having a suboptimal sampling frame does not imply that molecular surveillance will not lead to public health action. Since the main purpose is to detect signals that may indicate a common exposure, any signal detected can still lead to action. However, the number of such signals will be lower with fewer isolates or fewer representative isolates typed, and determining whether or not to take action when a signal is detected will be more difficult as there is less context that can be taken into account.

Table 1. Samples available for typing by country

Country	Reported confirmed cases in 2014				Proportion cases with sample provided to reference laboratory in 2014 (%)				Comment
	Salm.	*List.*	VTEC	*Camp.*	*Salm.*	*List.*	VTEC	*Camp.*	
Austria	1654	49	121	6514	100	100	82	100	Reported cases/provided samples vary for different reasons: one case with both Typhimurium and Enteritidis means two isolates provided but only one case reported, samples not sent to the reference laboratory, or an additional sample from the same patient.
Denmark	1124	92	257	3773	98	100	98	8	
Hungary	6946	39	18	8444	45	72	72	4	
Italy	3749	73	81	1252	unk	unk	100	unk	*E. coli* mainly from cases of bloody diarrhoea, haemorrhagic colitis and haemolytic uraemic syndrome.
Luxembourg	110	5	3	873	unk	unk	100	100	In 2014 and before, primary laboratories sent isolates for QC and typing, but stopped doing so in 2015. Legal framework to require automatic isolate transferral or primary sample in preparation.
The Netherlands	969	90	371	4159	100	100	100	0	Fixed coverage of 64% for *Salmonella* and 52% for *Campylobacter*. VTEC O157: 86 cases; non-O157: 286 cases
Poland	8392	86	5	650	2	unk	38	3	For *Salmonella*, some primary laboratories send rarer or more difficult to type serotypes.
Portugal	261	unk	unk	unk	59	(20)	(3)	(464)	Some primary laboratories send a fixed proportion of their isolates. Values for *Listeria*, VTEC and *Campylobacter* are absolute values rather than percentages as the total reported confirmed cases are not known.
Sweden	432	125	7	8288	25	96	16	0	For *Salmonella*, all domestic cases, around 20% of the total, can be provided free of charge. About 5-10% of travel related cases are provided, which must be paid for by the primary laboratory.
United Kingdom	8099	201	1326	66790	100	100	100	unk	Legal requirement to send all isolates for all notifiable diseases to reference laboratory. For *E. coli* non O157, clinical samples are provided.

2.2 Current and future scenarios for diagnosis and typing

The current practice of sending pure cultures from primary laboratories to the reference laboratories for FWD pathogens may undergo changes with the advent of new primary diagnostic and typing techniques, in particular new applications of PCR, mass spectrometry, WGS and metagenomics [3,4,5,6,7]. As a result, in some cases, and predominantly driven by cost reduction, there will no longer be a pure culture required for the primary laboratories' diagnostic testing that can then subsequently be sent for typing. Table 2 gives an overview of these techniques, their relative cost and whether they require a pure culture. In addition, Table 3 describes five general scenarios for analysis of typing data at the national level. Scenario A is currently the typical one in most countries, and some countries are moving to scenario B for selected pathogens. Scenario C is increasing due to the use of culture-independent diagnostics. Scenario D, where WGS typing is performed by the primary laboratory, is at present not known to be performed in any country. Finally, scenario E, where metagenomics is used for diagnosis, is at present not yet possible for routine purposes. For all scenarios, central collection of descriptive data on the isolates at the national level is assumed.

In any scenario, it is critical that typing data are centrally collected and analysed by the national reference laboratory, together with epidemiologists, since cases of disease resulting from the same event may not be locally clustered. This is due to the wide and frequently international distribution of food, which necessitates national- and international-level surveillance. The most realistic scenarios – at present and for the foreseeable future – are still scenarios A, B and C, i.e. scenarios where the primary laboratories provide pure cultures, or if not possible, clinical samples to the national reference laboratory, which then performs further typing. The advent of WGS would not immediately change that, but in some cases hospitals are already setting up their own sequencing facilities. Private operators also exist already.

In the future, scenario D – where WGS is not, or not always, performed in the national reference laboratory – may become more common, but there should still be an infrastructure in place to allow the timely analysis of the results at the national level. Finally, the increased use of culture-independent methods, in particular PCR, often reduces the number of isolates that are sent through to the national reference laboratory. Only stool samples are affected, as blood, urine and CSF samples are normally cultured, yet the former are responsible for the vast majority of the samples, at least for *Salmonella*, *E. coli* and *Campylobacter*.

Structural measures, including legal requirements for primary laboratories to send through samples or isolates – as well as in some cases a data privacy exception for accompanying data – would be very helpful and in some cases necessary to guarantee the provision of a minimum number of isolates to the national reference laboratory and thus support surveillance. In some countries, such as the United Kingdom and Austria, such legal requirements already exist.

Table 2. Techniques for diagnostics and typing

Technique	Cost per isolate	Culture independent	Comment
PCR	Low	Yes. Direct DNA extraction from clinical sample.	High sensitivity. Can detect non culturable variants, or pathogenic variants that are difficult to discriminate from non-pathogenic ones in, for example, stool samples, which is particularly important for *E. coli* and to some extent for *Campylobacter*. When more than one primer pair is used, there is some possibility that the amplified products, and the corresponding combination of genotypic traits, do not originate from the same cells which may complicate the establishment of genotype-phenotype associations.
Antimicrobial susceptibility testing	Depends on technique	No.	Can likely be predicted from WGS data with sufficient accuracy, depending also on the mechanism of resistance.
Mass spectrometry for identification	Very low	Depends on sample type. Culture required for faecal samples.	Has the potential to produce some subtype information in the future, but likely limited resolution.
Serotyping	Medium	No	Can likely be predicted from WGS data with sufficient accuracy.
Multiple-locus VNTR analysis	Medium	No	Cannot yet be derived from routine WGS data, potentially in the future.
Multilocus sequence typing	High	No	Can be derived unambiguously from WGS data.
Pulsed-field gel electrophoresis	High	No	Cannot be derived from routine WGS, to be determined if possible in the future.
WGS for subtyping	High	No	Costs are likely to fall and accuracy likely to improve substantially in the future.
Metagenomics for identification	High	Yes. Direct DNA extraction from clinical sample.	Has the potential to identify all species present in one experiment, including unknown ones. Substantial further research is needed in this area, including on high-quality DNA extraction from, for example, stool samples, and on species quantification in addition to identification.

Table 3. General scenarios for molecular surveillance now and in the future

Scenario	Primary laboratory	Reference laboratory
A	Perform diagnosis including isolation. Send the bacterial culture and additional descriptive data to the national reference laboratory.	Perform serotyping, PFGE, MLVA and/or other relevant typing methods. Analyse results at national level.
B	Perform diagnosis including isolation. Send the bacterial culture and additional descriptive data to the national reference laboratory.	Perform WGS and, where needed, phenotyping. In silico typing. Analyse results at national level.
C	Perform diagnosis without isolation, typically through PCR. If positive, send the remainder of the sample and additional descriptive data to the national reference laboratory.	Perform isolation, WGS or other genotyping and, where needed, phenotyping. In silico typing. Analyse results at national level.
D	Perform diagnosis including isolation and perform WGS typing. Send the sequence and additional descriptive data to the national reference laboratory.	Centrally collect sequence data. Analyse results at national level.
E	Perform diagnosis without isolation through next generation sequence typing on samples, i.e. metagenomics. Send the sequence and additional descriptive data to the national reference laboratory.	Centrally collect sequence data. Analyse results at national level.

2.3 Whole genome sequencing

This section describes the entire process of WGS in the laboratory, from DNA extraction to processing and storing the sequence data.

2.3.1 DNA extraction

The amount and quality of the DNA required for WGS is dependent on the sequencing platform that is being used. Obtaining the required amount of DNA is not expected to be a problem for most FWD pathogens, including *Salmonella*, *Listeria*, *E. coli* and *Campylobacter*, as they can be easily cultured. Most commercial kits can also

produce good-quality DNA, although recovery of plasmids may be an issue for kits based on precipitation, which is relevant because important phenotypic properties such as resistance or virulence are often encoded by plasmids. Sequencing platforms capable of producing very long reads may require more sophisticated extraction techniques to avoid DNA fragmentation. Finally, extraction is more difficult for gram-positive bacteria such as *Listeria* spp. than for Gram-negative bacteria such as *Salmonella* spp., *E. coli* and *Campylobacter* spp. due to the composition of the cell wall. An additional enzymatic step with lysozyme usually solves this problem.

Usually, only two generic protocols have to be used, rather than a separate protocol per pathogen: one for gram-positive and one for Gram-negative bacteria. The details of these two protocols will depend on the number of samples and available equipment. For smaller numbers of isolates, extraction is normally done entirely manually. For larger numbers of isolates, robotic extraction platforms using, for example, 96-well plates are advisable. Also, isolates from different species can normally be combined in a single DNA extraction run. After the extraction, an accurate measurement has to be done and where needed, the quality of the DNA should be assessed. The quality of the DNA, if not assessed at that point, can also, and should be, assessed based on the sequence data. If too low, the DNA extraction and sequencing should then be rerun. Standards should be set on concentration measurement and quality of the end product, in terms of fragmentation and presence of impurities, rather than on the actual protocol.

Training of laboratory technicians is not expected to be an issue, as the process is straightforward and nearly pathogen independent. Finally, it is possible to establish a DNA extraction pipeline that covers not only FWD pathogens but also other pathogens of interest. This has the potential to increase the efficiency and quality of the process, especially when the number of samples processed is otherwise low.

2.3.2 Equipment and protocols

At present, several WGS technologies and types of equipment useful for routine typing of FWD pathogens are commercially available. They typically require the construction of a 'sequence library' of short DNA fragments that are subsequently individually sequenced after pooling the libraries from different isolates. The protocols for constructing the sequence library are provided by the manufacturer, but they usually have to be optimised and adjusted, which also improves the quality of the process. There is a need for exchange of information and experience in this regard, for example through a user forum. The protocol for the 'sequencing run' itself is generally not an issue, as it is highly automated. However, quality controls can be included in each run to compare intra- and interlaboratory error rates. In general, there is a need for more quantitative results on intra- and interlaboratory reproducibility to determine the optimal balance between quality and cost.

Depending on the equipment and the desired minimum quality of the results, the running time and the number of samples that can be processed in parallel in the same sequencing run can vary. A major consideration in the purchase of equipment is therefore the required sequencing capacity, typically expressed in 'number of samples to process per week', and the desired minimum data quality. Because all techniques currently produce a set of overlapping reads, or raw reads, rather than a single fully contiguous chromosome plus potential extrachromosomal elements, it is not possible to express requirements as a single straightforward quality metric such as the expected number of wrongly sequenced bases in the entire genome. Instead, the main quality metric typically used is the coverage of the genome:

$$\text{Coverage} = \text{NumberOfRawReads} \times \text{AverageReadLength} / \text{GenomeSize}$$

Coverage is therefore the average number of times each base in the genome is contained in individual raw reads. The raw reads are filtered to remove bases of too-low quality. At the same time, coverage is also inversely proportional to the number of samples that can be processed in a single run and therefore crucial for capacity calculations as well as cost. Further details on coverage and cost per sample are given in the sections below.

Training of laboratory technicians is not a major issue once the protocol has been optimised and the desired coverage chosen. Work can be demanding during the set-up phase, however, especially when robotics are involved, which often requires a careful calibration of the equipment and workflow. Once set up, work can also be quite repetitive. As for DNA extraction, there is the potential to increase efficiency by centralising typing capacity at the national level, which does not necessarily have to be restricted to FWD pathogens. Timeliness of results is not expected to be significantly impacted by this. However, in practice it may often be easier to set up typing capacity within a single laboratory depending on organisational constraints.

2.3.3 Data storage and sharing

The amount of data generated per isolate by a sequencer is substantial. The raw read information consisting of both sequence and individual base quality scores is typically stored as a FASTQ file [8]. As a rule of thumb, the size of this file in megabytes (MB), and uncompressed, is the size of the genome in megabase (Mb) times the average coverage times two:

RawReadFileSize (MB, uncompressed) = GenomeSize (Mb) x Coverage x 2

If both forward and reverse reads are performed, this number combines the size of the data for both raw reads files if they are saved in separate files. At present, coverages from 30x upwards – this is a platform-dependent number – are used successfully in routine surveillance to evaluate whether isolates originate from an epidemiologically relevant common ancestor. An assessment of the impact of higher coverage should be made both per pathogen and per platform. An overview of file sizes for FWD pathogens is given in Table 4, for an average coverage of 50x. The value of 50x was used as a reasonable value for estimations on file size. For other coverage values, the expected file sizes should be adjusted linearly with the coverage.

The file size of a fully assembled genome in MB on the other hand is just equal to the size of the genome in Mb (see also Section 3.2). Individual base quality is normally not stored alongside the assembled genome, and hence the FASTA rather than FASTQ format is used for these files. For a coverage of 50x, the uncompressed file size would therefore be 100 times smaller than the raw reads file as indicated in Table 4. The current sequencing techniques however do not yet allow producing a fully assembled genome, unless at a substantial additional cost, though this may change in the future. Such partially assembled genomes may take up somewhat more storage than a fully assembled one, but only minimally.

As storage of the raw read data may well become a requirement for accreditation of the analysis pipeline, and likely also for litigation, it is very important to retain this information. Therefore, and also to allow later reanalysis, they will likely need to be kept under any scenario and remain the most important factor for estimating data storage requirements. Depending on the number of isolates that are sequenced, data storage may or may not be an issue, either technically or in terms of cost. The sequencing and storage of, for example, 10 000 *Salmonella* raw reads files, a substantial amount at present, would require around 5 terabyte of uncompressed data storage space according to Table 4. Taking into account backup requirements, this number has to be doubled at least, and if the data are to be available efficiently at all times for analysis, it should at least be tripled.

Data compression can reduce storage requirements substantially. A standard zipping of a FASTQ file reduces its size by a factor of 2 to 3. Compression algorithms developed specifically for raw reads also exist and are actively researched [9,10,11]. Especially algorithms that store the differences versus a reference sequence and use 'lossy' compression for the quality information, can reduce the required file size substantially, comparable to JPEG compression for images. For compression based on differences with a reference sequence, the decompression software needs to be able to make use of the exact same reference sequence. Finally, it is also possible to store data externally in cloud-based storage, where payment is essentially per terabyte, typically starting at around 30 euros per terabyte per month. Free storage facilities for sequence data, such as the European Nucleotide Archive or Genbank, can also be an option to store the data. In both cases, submission and retrieval of the raw read data can be facilitated through software, although it is likely that a local copy of the raw reads is still required [12,13]. The decision to use particular compression algorithms for locally stored data would probably be influenced by which algorithms are used in the large cloud storage facilities.

While storage of data is not very likely to pose technical challenges, sharing data between organisations may well do so. For raw read data, the required bandwidth may be an issue because sending, for example, 20 *Salmonella* raw reads files over the internet – approximately 5 gigabytes, assuming 50% compression – normally takes a substantial amount of time. Depending on the solution, there may also be a risk of time outs, which then require restarting the procedure. Sending via email is not possible either, as the maximum attachment size is normally already exceeded by a single isolate's data. Other solutions such as (secure) FTP sites or file sharing services are therefore necessary, which may incur an extra cost to use or maintain.

In the future however, it may not be necessary for routine applications to share the raw reads, but rather only a fraction of the information. This could be the assembled genome or even a further reduction of the information in the form of a standardised nomenclature. Especially in the latter case, network bandwidth is no longer an issue, and a file-sharing solution is no longer required. Both for assembly and nomenclature however, standards are still in development (see Sections 3.2 and 3.5), so it is foreseen that data sharing will in many cases still imply sharing raw read information.

Table 4. WGS data storage requirements per isolate

Pathogen	Genome size (Mb)	Average coverage	Raw read 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
E. coli	4.6–5.4	50	460–540	4.6–5.4
Campylobacter spp.	1.6	50	160	1.6

2.3.4 Cost and time comparison of WGS with current typing techniques

Cost is often the decisive factor when considering switching to a new typing technique, in addition to, for example, discriminatory power, repeatability and timeliness. It is therefore essential to have a good idea of the cost factors to be taken into account. This will be different in each country, and thus a general conclusion on this cannot be made. However, an estimate of only the cost of consumables, equipment and work directly related to the typing is possible. For work, cost should be expressed as required operator time per isolate because wages differ between countries, and laboratory technicians are not paid per processed isolate.

An overview of costs of consumables, operator time and total time is given in Table 5. These data were provided by the FWD-NEXT members as representatives for their countries in June 2015 and were verified for accuracy and completeness. The three measures – cost of consumables, operator time and total time – are each given as minimum–median–maximum so as to give an idea of the variation between countries as well. The number of countries for which a value was available for any or all of the three measures is indicated in the 'Countries' column. All values are estimated for a single isolate. If the typing technique in question is normally performed per batch of isolates, values were converted from the cost of a full batch to the average cost for a single isolate. These values are therefore valid only under the assumption that a full batch is processed and thus optimal efficiency is gained. For the calculation of total time, waiting times incurred due to shipment to another laboratory are not included. Finally, the total cost per pathogen was calculated, as well as for WGS, in order to be able to compare the variable cost per isolate and the processing time. Depending on the pathogen, some of the current typing methods are not weighted at 100%, as indicated in the 'Weight' column and explained further in the footnotes, because not all variants are typically typed the same way.

Based on these data, the median cost for WGS consumables (91 euros), is roughly twice that of current typing techniques, whereas the median operator time for WGS (1.7 hours) is roughly half of the median for current typing techniques, except for *Salmonella*, where it is almost the same. It should be noted that there are no reliable data on cost and operator time for WGS data analysis because there is no established standard. Once WGS data analysis is sufficiently standardised, it is expected that analysis will be automated to the extent possible, similar to current typing methods.

Equipment costs are not included in these numbers. These costs can differ substantially, depending on the vendor and capacity, and it was not deemed useful to have aggregate values for this. An example, however, can be instructive: a total estimated cost of 100 000 euros for the sequencer, written off over five years, plus 10 000 euro maintenance per year, with sequencing of 40 isolates per week for 45 out of 52 weeks (i.e. 2 250 isolates/year) results in approximately 17 euros for additional equipment costs per isolate under good scheduling conditions.

The cost for data storage also has to be taken into account. This cost scales linearly, or close to linearly, with the number of isolates. For example, a *Salmonella* or VTEC isolate – VTEC and *Salmonella* constitute the vast majority of isolates and also have the largest storage requirements – requires around 0.5 gigabyte per genome (uncompressed). Assuming a compression rate of 50% and applying a factor of 3 to take into account backups, storing the corresponding 0.75 gigabyte over a period of 10 years at 30 euros per month per terabyte (0.03 euro per month per gigabyte) adds an additional 2.7 euros per isolate.

All in all, WGS for *E. coli* and *Campylobacter* can already be less costly than other typing techniques. For *Listeria* the cost is likely more or less the same and for *Salmonella* it is can still be substantially higher, depending on the throughput. Since both the cost and operator time for WGS are expected to decrease in the future, which is unlikely to be the case for other typing techniques, typing of *Listeria* and eventually *Salmonella* with WGS will almost certainly become less costly in the near future. It should also be kept in mind that the replacement of several typing techniques with a single one has additional advantages.

With respect to the total time required for typing, WGS is at par with other typing techniques and substantially faster for *E. coli*. Also, far less variation is expected between isolates compared with other typing techniques such as serotyping, and only one experiment has to be performed. As such, total time for the procedure is likely already favourable for WGS for all pathogens, assuming a well-established routine in the laboratory.

Apart from the cost per isolate and the total time, there are also several other factors to be considered when deciding to switch to WGS. Firstly, there is some cost associated with maintaining several different typing techniques per pathogen, in addition to the isolation and selected phenotyping methods that would still need to be performed. Secondly, some of the current typing techniques are difficult and require a lot of experience, and laboratories may not have a sufficient number of qualified technicians for this task. As a result, maintaining capacity over time for these techniques may be an issue.

In conclusion, it seems that WGS already now is a cost- and time-effective method for the typing of *E. coli* and *Campylobacter* samples. *Listeria* is not far behind, and *Salmonella* will most likely follow somewhat later, depending on the throughput. Costs and time, however, are not the only factors to consider because the diseases caused by these pathogens have substantially different epidemiology and the value of typing for informing public health action may vary correspondingly. Also, there is a question whether some capacity for the current typing techniques should be maintained after migrating to WGS, either in the same laboratory or at least in a laboratory specifically supported for that purpose (see Section 4.4). In particular for antimicrobial susceptibility testing, it may well be worthwhile to maintain such phenotyping capacity in the laboratory, if, for example, new mutations in resistance-associated genes are detected through WGS and whose impact is by definition unknown.

Table 5. Cost and timeliness of different typing techniques including WGS, as assessed in June 2015

Pathogen	Typing method	Weight	Countries	Samples per year	Consumables per isolate (EUR)	Operator time per isolate (h)	Total time per isolate (d)
Salmonella	Isolation, confirmation and pure culture	100%	5–7	18315–18465	1–**4**–15	0.1–**0.5**–1.5	1–**1.5**–4
	Serotyping agglutination[4]	100%	6–8	21430–21580	5–**25**–55	0.08–**0.5**–1	<1–**2**–17
	MLVA Typhimurium and Enteritidis	67%	6–8	2395–2430	7–**17.5**–30	0.67–**1**–3	1–**1.5**–2
	PFGE XbaI	33%	6–8	1165–1435	20–**25**–42	0.36–**0.9**–1	2–**3.5**–6
	Resistance testing	0%	5–6	4630–4730	1.5–**9**–20	0.17–**0.5**–1	1–**2**–3
	Total[1]	**N/A**	**N/A**	**N/A**	**17–49–104**	**0.7–2–4.8**	**3.3–5.7–24.3**
Listeria monocytogenes	Isolation, confirmation and pure culture	100%	4–5	700–775	1.5–**3**–20	0.1–**0.69**–3	1–**2**–6
	Serotyping agglutination	33%	4–5	143–148	10–**14**–40	0.25–**0.49**–1	<1–**1**–2
	Serotyping PCR	67%	5–6	290–360	1.2–**6.5**–35	0.33–**1**–3.5	0.09–**1**–2
	PFGE ApaI+AscI	100%	7–8	440–515	24–**30**–60	1–**2**–2.8	3–**3.5**–6
	Total[2]	**N/A**	**N/A**	**N/A**	**30–42–117**	**1.4–3.5–8.5**	**4.4–6.5–14**
Campylobacter	Isolation, confirmation and pure culture	100%	4–5	1015–1165	1.5–**3**–20	0.1–**0.69**–3	1–**2**–6
	Species typing	0%	4–6	2000–2150	1–**15**–30	0.07–**0.5**–2	<1–**1**–1
	MLST	100%	3–4	650	25–**61.6**–250	3–**3.5**–5	2–**3**–5
	Total	**N/A**	**N/A**	**N/A**	**27–65–270**	**3.1–4.2–8**	**3–5–11**
E. coli	Isolation, confirmation and pure culture	100%	6–7	2980–3150	1–**6**–35	0.1–**0.75**–3	1–**2**–3
	vtx1+vtx2+eae gene detection (not subtyping)	100%	7–8	4105–4275	7–**20**–40	0.7–**1**–2	0.5–**1**–2
	O-group typing agglutination	100%	6–8	2285–2375	5–**10**–85	0.2–**1**–2.5	1–**2**–7
	PFGE XbaI	100%	7–8	765–805	20–**25**–50	0.36–**0.9**–1.27	3–**4**–12
	ESBL production	0%	2–4	303–403	10–**13.5**–15	0.25–**0.5**–2	1–**1.5**–2
	Total[3]	**N/A**	**N/A**	**N/A**	**33–61–210**	**1.4–3.7–8.8**	**5.5–9–24**
WGS[5]	Isolation, confirmation and pure culture	100%	19–24	N/A	1–**6**–35	0.1–**0.5**–3	1–**2**–3
	DNA Extraction	100%	10–19	N/A	1.5–**5**–10	0.5–**0.67**–3	0.25–**1**–2
	Library preparation and sequencing	100%	6–8	N/A	48.3–**80**–151	0.09–**0.48**–1	2–**3**–12
	Data analysis	100%	-	-	-	-	-
	Total	**N/A**	**N/A**	**N/A**	**51–91–196**	**0.7–1.7–7**	**3.25–6–17**

[1] *MLVA typing: 67% of isolates corresponding to* Typhimurium *and* Enteritidis *incidence (EU/EEA-wide figures) and assuming same price for both as only* Typhimurium *MLVA prices were collected. PFGE typing: 33% of remaining isolates.*
[2] *Serotyping agglutination 67%, PCR 66%, based on number of samples per year.*
[3] *The O-group is often not determined if not one of the most common types.*
[4] *Minimum and median values are for* Enteritidis *and* Typhimurium *as well as other easy-to-determine serotypes; maximum values include also difficult cases.*
[5] *Values pooled across the four pathogens.*

3 WGS data analysis for public health

After performing the sequencing, the raw read information needs to be interpreted to be useful, for example with regard to public health action. This chapter describes the steps that can or need to be taken, how computationally intensive they are, and where further research is needed.

3.1 Raw read quality and filtering

The raw read data need to undergo a quality control before they can be used further so that parts with too low quality are filtered out. The first step is normally to verify that the produced raw read data meet general acceptance criteria, such as achieving the expected coverage. If not, the isolate in question should be sequenced again. The second step is normally to identify reads that are shorter than the minimum expected length, dependent on the technology used, and to discard these as a whole.

The third step is normally to 'trim' the reads at their beginning or end, removing bases of too low individual quality as calculated by the sequencer, and if needed also any adaptor sequence added as part of the library preparation (see section 2.3.2). Usually, the quality of individual bases is expressed as a 'PHRED' score, which is equal to 10 times the negative logarithm of the probability $p_{WrongBase}$ that the base is wrong:

$$PHRED = -10 \log(p_{WrongBase})$$

A PHRED score of 20, which is a typical minimum quality cut-off for the trimming, therefore implies a chance of 1 in 100 that that base is wrongly sequenced.

A fourth step may be to verify whether the DNA was indeed from the expected species as, for example, indicated by the provider of the sample, or if some contamination has occurred during the DNA extraction or library preparation process. This can already be done based on the raw reads, for example by deriving all 'k-mers' or short overlapping sequences of typically around 15 bases from the raw reads. The number of occurrences of each unique k-mer, i.e. their distribution, is species specific and can be compared with the expected distribution of the species in question as well as that of other species, so that any likely contamination can be detected [14]. The verification of contamination can also be done at a later stage because contamination is also clearly detectable during the interpretation of the sequence data. One example of this would be when de novo assembly (see next section) produces contigs that are larger than the expected genome size. It may be also useful to analyse the sequence of the contaminating organism in order to see if it matches other contaminations and thus identify it.

Both the removal of too short reads and their trimming are processes that can be fully automated and are computationally trivial, with insignificant impact on total computation time. The k-mer-based detection of contamination can be automated as well but tends to be more demanding, though still minor in comparison with subsequent analyses. The result after trimming is a smaller FASTQ file or equivalent, with somewhat fewer and shorter reads.

3.2 Genome assembly

After the raw read filtering to remove parts of too low quality, the next step is normally either to assemble them into a single genome, or to directly align them to a reference genome. In theory, genome assembly is the natural next step, since all the raw reads together are in effect a measurement of a single genome, including also extrachromosomal elements where applicable. There are currently three practical issues with genome assembly that need to be considered.

Firstly, the raw reads generated by current sequencing technologies do not allow, at least not in routine application, to produce a fully assembled genome. Instead, the end result is a number of contiguous sequences or contigs, with typically a few very large ones and many smaller ones. This is because the read lengths are typically too short to cover repetitive elements as a whole, as a result of which there is no clear single solution on how to assemble the reads in those areas. However, for most practical applications, including assigning nomenclature and deriving phylogeny, such a partially assembled genome is likely perfectly usable. At the same time, usability may depend on the sequencing technology used, as it is possible that technology-specific issues affect the quality of the assembly. Also, as sequencing technologies advance, particularly with respect to read length, genome assembly is expected to become more and more complete and it is not unimaginable that in the future full assembly will be possible on a routine basis. In the meantime, a quality metric that is used for genome assembly results is the N50 value, which is defined as the length of the smallest contig among the set of the largest contigs that together cover at least 50% of the assembly [15]. This metric is species- and sometimes strain-specific, as well as dependent on sequencing conditions, and mainly used to choose the best result among different assemblies.

Secondly, there are many algorithms available to perform genome assembly, each with its own set of parameters, which explains that they do not arrive at exactly the same partially assembled genome. The assembly algorithm

therefore is a source of variation in the analysis pipeline, and additional research is required to evaluate its impact. Eventually, a standard should be adopted for routine application to remove this source of variation, which is largely dependent on the used sequencing technology. On the other hand, as sequencing technology improves further, assembly algorithms may well become less and less important because variations between assembled genomes would be expected to decrease correspondingly. It may be worthwhile to add individual base quality information to the assembled genomes, for example based on coverage and sequence variation for each individual position in the sequence. This information can then be taken into account in the further analysis of the data, for example by disregarding bases that have too low quality. It could also be used as an acceptance criterion for sequence repositories. Storing this information would double the uncompressed file size of assembled genomes.

Thirdly, genome assembly is computationally a very demanding process, compared to other processes such as raw read filtering. The required computation time per isolate depends on the length and complexity of the genome (which is more or less fixed per species), on the number and length of raw reads, and the algorithm and parameters used. A typical genome assembly (i.e. for an FWD pathogen with a coverage of 30–50x) takes up to one hour per isolate on a contemporary processor with at least 8 GB of memory, although for the smaller genomes of *Listeria* and *Campylobacter* it can be significantly less. Depending on the number of samples that are sequenced per week in the laboratory and the number of processors on which genome assembly can be run, it may be a problem to perform this locally. Either the local computing capacity then has to be increased, or the calculation has to be performed in the cloud. For the latter, there is software available to do this without overly complicated technical requirements for setup, and payment is basically per use. If neither is an option, then direct alignment to a reference should be used, although this technique has several issues of its own (see Section 3.3). It should also be taken into account that processors will become more powerful, and that improvement of the sequencing technology, in particular with respect to the read length, is likely to substantially reduce the computational requirements.

In conclusion, genome assembly is likely to become a standard practice for FWD pathogens in the future, after filtering the raw reads. It can be seen as a natural data processing step, analogous to converting the ladders or fluorescent peaks of Sanger sequencing into a full sequence, although the raw read data is likely still needed for some analyses where assembly introduces too many issues. The output of the genome assembly, along with quality parameters such as the N50 value, is a FASTA file containing the sequence of each contig. Including quality scores for individual bases in this file could be a future improvement. In the ideal case of a fully assembled genome, the resulting file is 5 megabyte or less in size (see Table 4), which is very small compared with the raw read data. Regarding requirements for computational capacity, this is dependent on the number of isolates that are processed in the laboratory on a weekly basis, and computing time is expected to decrease as sequencing technologies improve. A single dedicated processor running for 24 hours will normally be able to assemble the genomes of at least 24 isolates sequenced with current technologies, which is already within the practical range for several countries.

3.3 Alignment to a reference genome

An assembled genome or even raw reads of an isolate are frequently aligned to a reference genome as a first step in the interpretation of a sequence. The reference genome should preferably be fully sequenced and closed, i.e. a circular, high-quality genome. It is typically also annotated with loci and chosen to be as close as possible to the isolate in question. As such, alignment to a reference identifies the loci present in both the reference and the isolate. From there, differences in sequence versus the reference or versus other isolates can be enumerated. If no fully closed reference sequence is available, any close regular isolate's partially assembled sequence can be used as a reference, although this implies that genome assembly is performed on at least some of the sequences.

Computationally, alignment to a reference genome is a much less demanding process than genome assembly, requiring typically between 1 and 15 minutes per isolate on a contemporary processor for FWD pathogens. When computing capacity is an issue, it is for example possible to skip assembly altogether and directly align the raw reads. The result, also when aligning the assembled genome to the reference rather than raw reads, is essentially the reference genome with its sequence replaced by either the consensus of matching raw reads or that of the assembled contig. Any genome rearrangements or additional elements in the actual isolate sequence are lost.

The latter is an important issue with alignment to a reference: any loci present only in the isolate and not in the reference, for example plasmids, will not be detected and thus cannot be considered in further analyses. Therefore, to reduce the number of such missed loci, it is important to use a reference sequence that is as closely related to the isolate or isolates as possible. Detecting the closest reference genome can be done using the same k-mer approach as mentioned in Section 3 for the detection of contamination, which is computationally not very demanding. At present, only a few fully assembled, high-quality and annotated genomes are available for FWD pathogens, and further investments in the sequencing of reference genomes is recommended. In addition, it may be worthwhile to expand the methodology so that it does not only detect matches against a reference genome but also, for example, perform match operations against a – not yet existing – standard library of extrachromosomal elements, so that the probability of missing relevant loci is lowered.

Regardless of the issue of potentially missed loci, the alignment of several isolates to the same reference genome does produce a multiple alignment of these sequences, from which differences between isolates can readily be derived and used, for example for deriving phylogeny based on single nucleotide polymorphisms (SNPs, see Section 3.4). Also, absent or present loci that are associated with a particular phenotype such as resistance or virulence can be detected based on alignment to an annotated reference.

In conclusion, the process of aligning to a reference is very useful for practical applications and computationally not very demanding. The methodology could be expanded further to also include extrachromosomal elements, and there is a need for more reference genomes to capture the diversity within species and reduce the number of loci that are missed when aligning to a less closely related reference.

3.4 Deriving phylogeny

3.4.1 Fundamental aspects

One of the primary purposes of typing of FWD pathogens is to assess whether two or more isolates originate from an epidemiologically relevant common ancestor. This is frequently done by comparing information about their genotype in a dendrogram and then applying a practical similarity cut-off. With the advent of WGS, where the majority, and in the future perhaps the entirety, of a genome is available for such comparisons, it is expected that the accuracy of dendrograms to describe evolutionary relationships will be much higher than that of any other technique that is currently available. However, the process of constructing and interpreting a dendrogram based on WGS data is not straightforward for several reasons.

Firstly, it is not yet understood what degree of sequence diversity is likely to imply with strong likelihood – and without much additional epidemiological evidence – that two isolates are not linked to an epidemiologically relevant common ancestor. This is a typical situation where epidemiologists must decide whether to take further action on a cluster, including the collection of additional epidemiological evidence. Fundamental to the understanding of the problem is the fact that the amount of diversity that is generated in one unit of time depends on the rate of reproduction and the evolutionary pressure that organisms are under. Both factors depend on the specific environment of these organisms. For FWD pathogens, this can be quite diverse, depending also on the different conditions along the food production chain such as cold or frozen storage versus room temperature, and the presence of disinfectants. It is therefore highly unlikely, even as data and research accumulate, that, for example, the number of single nucleotide polymorphisms (SNPs) difference found between two isolates can be reliably related to the number of days that have passed since they diverged from a common ancestor. Furthermore, distantly related strains may still be epidemiologically linked in a meaningful way, for example in the case of several contamination foci along the same food chain. There is a clear research need to study the epidemiological relevance of sequence diversity of different FWD pathogens by sequencing isolates from known outbreaks, plus sporadic cases from different countries. The impact of the methodology used to interpret the data, as described in the paragraphs below, should also be assessed.

Secondly, there are several ways to express sequence diversity given the same WGS data, each further dependent on parameter settings. Ideally, it should be expressed as the number of evolutionary events of different types that likely occurred between two isolates. Subsequently, the expected frequency of these events, which are pathogen specific, should be taken into account to arrive at a single 'genetic distance' between these two isolates [16]. Such types of events are, for example, point mutations (SNPs), insertions, deletions, translocations including different types of transposons, inversions, loss or acquisition of bacteriophages, and acquisition of plasmids or other horizontal transfer events. For SNPs it is also important to consider that they consist of both transitions (A↔G and C↔T) and transversions (all other combinations), with the latter being significantly less frequent than the former.

Finally, it should be noted that there are also algorithms that directly use the sequence to create a dendrogram, such as maximum parsimony, maximum likelihood, and Bayesian phylogenetic inference. These are likely to produce more accurate results than methods such as neighbour-joining and UPGMA, which use a single distance between each pair of isolates. However, they are also extremely computationally intensive and therefore very unlikely to be usable in a routine setting.

3.4.2 Phylogeny based on SNPs

Among the different evolutionary events, only SNPs are, for the foreseeable future, useful for calculating genetic distance because they occur with a relatively well understood frequency – albeit not a uniform one – throughout the genome due to different selective pressure and can be derived from the alignment to a reference genome (see Section 3.3). In practice though, it is not straightforward to derive true SNPs between two isolates: sequence differences may also be due to a sequencing error, an artefact of assembly or alignment to a reference, or another type of evolutionary event. Sequencing errors are often dependent on the technology used, which may hinder comparison of sequences produced by different technologies, and even when sequences are produced by the same

technology, the expected error rate should be known to be able to properly interpret genetic distances in terms of number of SNPs difference.

Filtering out false SNPs and keeping as much as possible of the true SNPs would reduce the number of errors and is to some extent already applied in practice. This can, for example, be done based on the number of nucleotides between SNPs because it is statistically less likely to have several consecutive SNPs, or on quality values of individual bases if available. In addition, the region at the end of a contig is more prone to alignment or assembly errors, and any SNPs in there could also be filtered out. More research is clearly needed, however, to improve enumeration of only true SNPs. As sequencing technology improves and more reference genomes become available, filtering out false SNPs may become less of an issue.

The enumeration of SNPs and the calculation of genetic distances based on SNPs is computationally demanding. It requires comparing the full sequence of one isolate to the other and filtering out false SNPs. As the number of pairwise genetic distances to calculate for a dendrogram – together making up the 'distance matrix' – increases with the square of the number of isolates, calculating a large dendrogram based on SNPs may be prohibitive. Moreover, the set of SNPs to consider for the calculation of the distance matrix also depends on the exact isolates that are included, as typically only SNPs in genome regions present in all included isolates are considered to make sure that the scale of the genetic distance values in the matrix remains the same. In that case, it is also not possible to maintain a pre-computed distance matrix of all isolates based on SNPs, which would otherwise substantially reduce the computational requirements.

Taking all these elements together, it is unlikely that SNPs can be used in the near future in a standard way to calculate genetic distances and dendrograms across any set of isolates. In theory though, when only true SNPs are considered, they should provide a very good estimate of evolutionary distance. In the case of closely related isolates however, as opposed to any random set of isolates, it is expected that the proportion of false SNPs will be much smaller, especially after filtering out false SNPs. For such sequences, which, for example, may have been identified through a cgMLST-based dendrogram (see next section), a SNP-based dendrogram would likely provide both high accuracy and close to the highest possible resolution extractable from the sequence data. This is especially relevant, since in practice closely related sequences may be linked to the same public health event, and the higher the resolution and expected accuracy of the dendrogram, the better it is suited to inform follow-up public health actions.

3.4.3 Phylogeny based on allele differences (wgMLST)

The main alternative approach to SNPs for calculating evolutionary distances between isolates is based on the multilocus sequence typing (MLST) method, applied to whole genomes and therefore also termed 'whole genome' or wgMLST (Figure 1a) [17]. First, individual loci or targets – the term loci will be used further – are detected on each isolate's sequence and the corresponding allele sequences extracted. As such, this technique only considers coding regions and therefore does not consider point mutations occurring outside of coding regions, which may be under less selective pressure. The extraction of the allele sequences can be done, for example, by aligning each isolate's raw reads or assembled genome to a reference sequence that is annotated with those loci. Further research is needed here to determine the best way for this 'allele calling', as several other approaches (e.g. aligning per locus to a consensus sequence, a sequence profile, or even a hidden Markov model) could be expected to be, at least in theory, more accurate because they do not depend on any single reference, but rather capture the available sequence diversity in increasing levels of detail. Also, aligning the raw reads to the reference seems to give better results than using the assembled genome. Regardless of the exact process used for allele calling, there may be loci that are not found in the isolate due to the quality of its sequence data, but which are actually not missing. Such missing loci may affect the accuracy of wgMLST-based phylogeny.

After the allele calling, the extracted allele sequences from the same locus are compared as a whole between each pair of isolates: when two isolates have a different allele sequence for the same locus, the distance between those two isolates for that locus is equal to one, regardless of the actual number of differences between the two allele sequences. When the allele sequences are identical, or the locus is missing in either or both sequences, the distance for that locus is zero. The sum of all the individual locus distances is then the distance between the two isolates, which can be used for a dendrogram (Figure 1b). As such, MLST reduces the information used (i) by looking only at a part of the genome, i.e. the loci or coding regions, and (ii) by considering allele sequences to be either identical or different, regardless of the number of actual evolutionary events. While this does reduce the resolution that can be obtained by this technique, it also reduces the frequently disproportional impact on the distance matrix of false SNPs that are actually due to other evolutionary events such as insertions or deletions, alignment or assembly artefacts, or sequencing errors. Therefore, the technique is probably more robust than SNPs when comparing isolates that are not, or not known to be, closely related, as is normally the case with newly sequenced isolates.

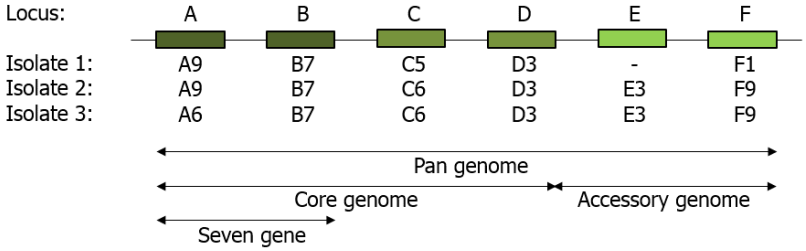
The wgMLST approach can be implemented in a computationally efficient manner, as opposed to the SNP approach. To this end, a database is kept of all unique allele sequences for each locus ever observed among the full set of isolates, and a unique allele identifier, typically an integer number, is assigned to each of these allele

sequences. When a new isolate is processed, its allele sequences can first be converted into the corresponding allele identifiers retrieved from this database, at the same time adding any new allele sequences to it. The construction of a distance matrix for any set of isolates only requires the comparison of their precomputed allele identifiers, rather than the entire sequences. This is a fundamentally faster computation method which should not require a prohibitively long time on a contemporary processor, even for large numbers of isolates.

A crucial choice that needs to be made as part of the wgMLST approach is which set of loci will be used for the distance calculation. The set of all loci ever observed within a species is termed the pan-genome (Figure 1a), but this is not usable because it will also contain loci that are the result of horizontal transfer events [17]. In theory, the set of loci that are not the result of horizontal transfer events at any point in time, i.e. that have been purely transmitted vertically and evolved further through clonal expansion, should be used to achieve the highest possible resolution for a dendrogram that represents ancestral relationships. However, this information about the evolutionary origin of each locus is not known and if it were, it would not necessarily be the useful in practice. Instead, the largest set of loci that can be chosen is the shared genome, i.e. the set of loci that is common to all the selected isolates, but by definition these loci will vary depending on the selected isolates. A smaller but more relevant set is that of all loci that are observed to be common to all or nearly all isolates, also called the core genome (Figure 1a). Importantly, core genome MLST or cgMLST – as well as other MLST variants, including the traditional seven-gene MLST – can also be used as the basis for nomenclature, which is further described in Section 3.5. Determining the loci that should be part of the core genome is an active area of research, as it requires a wide range of sequences to be available; it also requires the exclusion of problematic loci that are similar to one another, which could result in errors during the allele calling.

Figure 1. wgMLST principle

a) Assignment of unique allele identifiers



b) Distance between isolates according to different schemes

Isolate pair	Seven gene	Core genome	Pan genome
1-2	0	1	3
1-3	1	2	4
2-3	1	1	1

c) Assignment of sequence types

Isolate	Seven gene	Core genome	With sublevels (AB / CD)
1	23 (A9-B7)	115 (A9-B7-C5-D3)	23-9 (A9-B7 / C5-D3)
2	23 (A9-B7)	267 (A9-B7-C6-D3)	23-15 (A9-B7 / C6-D3)
3	47 (A6-B7)	349 (A6-B7-C6-D3)	47-93 (A6-B7 / C6-D3)

Note: (a) Allele identifier assignment. A dash indicates the locus is not present in the isolate; (b) genetic distance calculation; (c) sequence type assignment for different variants

3.5 Nomenclature assignment

Nomenclature is in its essence a technique to reduce the amount of available information by assigning a short, yet still informative human readable code to isolates. Where two isolates share the same code, it implies that they have the same properties as defined by the nomenclature scheme that is assumed to be commonly understood and adhered to. The *Salmonella* serotype is an example of this, where each combination of somatic and flagellar antigens is assigned a name, typically of the location where the first isolate was detected. These names are easy to remember, communicate, and when different between two isolates, imply a strong likelihood of them not being part of the same event. This is also important in cases of litigation, where nomenclature is often used in the proceedings as evidence for a causal relationship between a vehicle and human cases.

For WGS data one or more agreed nomenclature schemes are also required, since efficient communication between organisations is a prerequisite due to the international dimension of FWD pathogens. However, sharing sequence data as part of such communication is acceptable within the practical constraints of, for example, file size.

3.5.1 wgMLST nomenclature

The approach currently under wide discussion for nomenclature is based on wgMLST, whereby a fixed set of loci for a particular species or genus is agreed upon by all parties that use the scheme, and unique alleles for these loci are given a unique identifier in a global nomenclature database accessible to all (see also Section 4.3.1). The choice of loci is an active area of research. They can be chosen as those that are common to all known isolates of the species, i.e. the core genome, or simply all loci ever observed within isolates of the species or genus, i.e. the pan-genome (see also Section 3.4.3). Depending on the species, there may be a large difference between both. To give an example: *E. coli* and *Campylobacter* are well known to undergo significant horizontal gene transfer, and in those cases the pan-genome will be much larger than the core genome. The former also continuously grows whereas the latter shrinks as more isolates are included, with some loci moving from core genome to accessory genome. This movement is expected to reduce over time, along with the number of new alleles found.

An additional step in assigning allele identifiers to a particular set of loci, which also further reduces the information to a degree that it can be used effectively for human communication, is to assign an additional unique identifier to each combination of alleles observed within a single genome. This is referred to as the sequence type, although that term should ideally be reserved for the original seven-gene MLST. An approach with sublevels could be adopted here, whereby sequence types are assigned first, covering the traditional seven housekeeping loci, and then – within each of these sequence types – sub-sequence types are assigned, covering the next N most stable loci. As an example, an isolate with sequence type 23–15 would have allele combination 23 for the traditional seven housekeeping loci, and then within sequence type 23, it has allele combination 15 for the next N most stable loci (Figure 1c). In addition, independent sequence types for the accessory genome could be developed as well, particularly for *E. coli* and likely also *Campylobacter*, where this part of the genome is important for properly identifying closely related isolates. Several sets of loci could be chosen to cover different genes of interest such as pathogenicity islands.

However, the sequence-type-based nomenclature, even with sublevels, is fundamentally not one that describes phylogenetic relationships. Consider two isolates that have identical sequences, except for one SNP, in one of the seven housekeeping loci. In a sequence type with two levels as described in the previous paragraph, these isolates would, for example, be assigned sequence types 23–15 and 47–93, since the top-level sequence type is different and the sublevel numbering restarts within each top level. They would thus be placed in completely different parts of the sequence type 'tree'. While this is not expected to occur frequently because the seven housekeeping loci at the top level are chosen for their slow mutation rate compared with other loci, it is bound to occur, and more so at the next level of the sequence type 'tree'.

Therefore, when using sequence types for communication, the receiving party still has to verify that any of its own isolates that do not exactly match the full sequence type, are in fact still a close match. The main purpose of nomenclature, i.e. to enable efficient communication on similarities between isolates, is undermined by this. Nonetheless, when two isolates have exactly the same sequence type, they are guaranteed to be similar, and thus sequence-type-based nomenclature may still be useful in such cases. In addition, it can also serve as a further compression of the information, whereby the combination of hundreds or thousands of alleles is condensed into one agreed code.

3.5.2 Taxonomical nomenclature

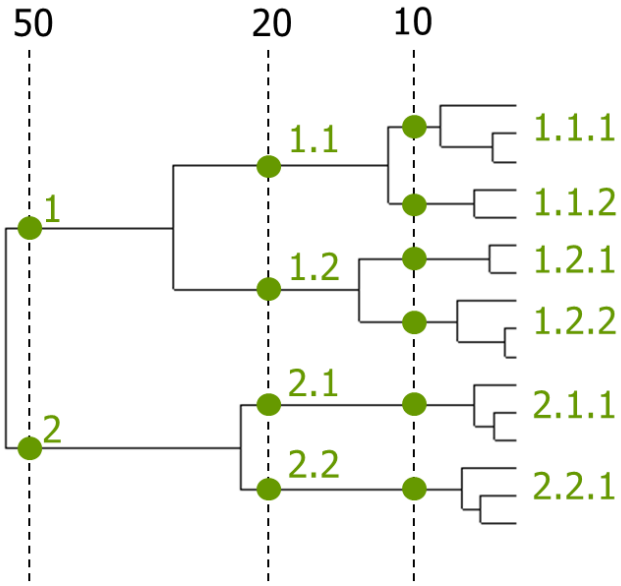
A nomenclature that truly describes phylogenetic relationships between isolates, i.e. a taxonomical nomenclature, would be very beneficial for communication as it allows determining with reasonable accuracy how similar isolates are based on a single code. One such approach is currently under development. It is based on a hierarchical description of where a particular isolate fits in a dendrogram that spans all isolates.

To this end, a simplified dendrogram is derived from the actual dendrogram by merging nodes at different heights. Wherever an edge crosses one of the selected heights, a node is created in the simplified dendrogram, and nodes exactly on the same height also remain nodes themselves. Then, starting from the root node of the simplified dendrogram, each child node is given an incremental number that is suffixed to that of its parent node, so that a hierarchical 'address' is built up for the actual isolates that are children of the leaf nodes of the simplified dendrogram. Figure 2 demonstrates the principle using three different heights at 10, 20 and 50 SNPs or alleles difference, depending on which is used to create the dendrogram. This difference should be interpreted as the number of different SNPs or alleles that are at least to be found between members of two taxa that have the same parent node. In the figure, taxa 1.1.1 and 1.1.2 share the same parent, 1.1, and between the members of these two taxa, the number of differences is at least 10 and maximally 20. For some individual pairs of isolates, the actual number of differences may be slightly below 10 or above 20, due to the fact that a dendrogram reduces the

available information in the distance matrix to an extent that no longer guarantees this exact cut-off. However, such cases are expected to be small both in number and in actual deviation from the boundaries.

The approach can indeed be based on any actual dendrogram, either SNP- or MLST-based, or even other variants. At present, it is being researched at Public Health England using SNP-based dendrograms, and the nomenclature is called the 'SNP address'. One could also envisage that over time, some of the nodes in the simplified dendrogram are given an additional easy-to-remember name, analogous to those of the serotypes for *Salmonella*, which in fact label a combination of O and H antigens, or of subspecies in higher-level taxonomy. On the other hand, to make this type of nomenclature possible, each new isolate has to be fitted into the entire dendrogram of all isolates, which may prove computationally expensive. In addition, new isolates will introduce subtle changes in the dendrogram as a result of which earlier isolates may have to be reclassified under a different address. As an alternative to this process, one or more representatives are defined per taxon, and isolates are matched only to them. This reduces the computational requirements substantially and at the same time reduces the reclassification issue, provided that the choice of representatives, which together imply the taxonomy, is well maintained over time. In any case, more research is required to evaluate the different variants of this type of nomenclature and its practical utility.

Figure 2. Taxonomical nomenclature principle based on SNP or wgMLST dendrogram



Note: The height cut-offs of 50, 20 and 10 correspond to the average number of SNPs or alleles difference.

3.5.3 Combined nomenclature

In conclusion, the wgMLST-based nomenclature using sequence types is the most widely discussed and also straightforward to implement (see also Section 4.3.1). It has the drawback that it is inherently not expressing the evolutionary relationships between isolates, although to some extent this can be captured by using hierarchical sequence types that use different levels of conservation between loci as an approximation. At the same time, the sequence types are also a means of compressing data of hundreds or thousands of alleles into a single identifier.

The 'SNP address' nomenclature on the other hand, is truly hierarchical and could be considered as an extension of the current taxonomy. As a consequence of this, it requires comparing each new isolate to all the previous ones, or, potentially, to a representative subset. These two types of nomenclature are not mutually exclusive, however. In fact, the 'SNP address' can also be based on wgMLST allele identifiers and could in that case be implemented straightforwardly within a single global nomenclature database.

3.6 Prediction of phenotypes and genotypes

It will clearly be possible to use WGS to derive phylogeny and assess whether two or more isolates may be originating from an epidemiologically relevant common ancestor or, ideally, a common source, thereby fulfilling the primary purpose of molecular surveillance for FWD pathogens. Apart from this, however, there are other potential uses of WGS data, including prediction of phenotypic properties such as resistance, virulence and serotype. If this can be done reliably, the need to perform such phenotypic tests, at least for public health purposes, may be reduced drastically. The entire area of phenotype prediction for FWD pathogens is currently one of active research,

with no defined standards yet in place. Where research is conclusive, a global agreement on the standard to use for predicting each phenotype would be recommended.

The complexity of phenotype prediction depends on the cellular processes involved. In the simplest case, a single gene is involved, with little-known variation between alleles. This can be the case for some forms of acquired resistance, where a single enzyme, typically located on a plasmid, breaks down the antimicrobial in question. The presence of this single gene, which is readily inferred from the sequence, will therefore predict resistance with high accuracy for all the antimicrobials that it can break down. In addition, if the information is used for clinical care, it would likely be appropriately cautious not to treat with antimicrobials that the pathogen has a high probability of being resistant to.

Most phenotypes are however much more complex than this simple, single-gene case. For *E. coli* the presence of vtx1 and/or vtx2 genes is already an indication of virulence and is indeed used for diagnosis. It is also known that different alleles of these genes are associated with different degrees of virulence, and that the presence of the eae gene or other genes involved in the colonisation process can be important for the pathogenic potential of a verocytotoxin-producing *E. coli* strain. Similarly, for antimicrobial resistance conferred through mutations in the target gene, an association, ideally quantitative rather than qualitative, between these mutations and resistance has to be established. In addition, resistance due to new mutations can likely not be predicted. More complex phenotypes such as the *Salmonella* O-antigen depend on a series of genes and variation within them, and are thus more complex to predict. In general, for these cases, a substantial and diverse dataset of isolates for which both WGS and the associated phenotypic results are available, is needed to be able to construct a useful predictive engine.

Apart from resistance and virulence, which have a clear public-health and even clinical-care application, there are also phenotypes and genotypes that have historically been used for typing of FWD pathogens, for example the *Salmonella*, *Listeria* and *E. coli* serotype, MLVA and PFGE. These phenotypes and genotypes have little importance per se: the *Salmonella* flagellar and somatic antigens together form the serotype but are not by themselves implied in pathology. However, there is a very substantial amount of knowledge associated with them, particularly in terms of the potential vehicles they have been associated with in the past. For this reason, it is very important to be able to predict these 'historical' phenotypes with good accuracy so that this associated information can still be used in ongoing events for hypothesis generation. To ensure broad access to this knowledge, the creation of a global database is recommended with information between serotype and associated potential vehicles, available to everyone as a hypothesis generator.

Unfortunately, both MLVA and PFGE, even though they are in essence genotypes, cannot be predicted from current WGS data because repetitive elements are exactly where current genome assembly fails in routine application, and a fully assembled genome is required to derive restriction fragment lengths. In the future however, it may be possible to predict MLVA, provided that read lengths become sufficiently long to cover the entire repetitive sequence. It may also be predictable with reasonable accuracy by matching the non-repetitive part of the sequence to sequences with known MLVA pattern, i.e. based on genetic association. For PFGE on the other hand, even when a fully assembled genome is available so that restriction fragment lengths can be derived, it may still be hard to compare with an actual gel profile because the gel migration and interpretation also have to be taken into account. Finally, serotype prediction is possible in practice for *Salmonella* and *E. coli* [18,19,20]. For *Listeria,* the PCR serotype at least can be predicted and likely also the phenotypic one.

3.7 Visualisation

The visualisation of typing data is an extremely important aspect of the interpretation as it forms the main bridge between the microbiological information and epidemiologists that have to decide on further actions. Any available WGS data therefore have to be reduced to a manageable extent without losing the relevant information. At present, there are no established practices for that. In addition, it is recommended that microbiologists or bioinformaticians interpret the information together with epidemiologists. Several different ways of visualising WGS data for public health action can be envisaged.

Firstly, line lists of cases or isolates can be expanded with a 'column' containing a standardised dendrogram. To interpret the dendrogram, it could be annotated with a relevant threshold for similarity above which an epidemiological link becomes likely. At present however, there is no standard yet for the dendrogram, and similarity thresholds are not yet established for FWD pathogens. There is also a distinct possibility that in practice there will be no single threshold possible, and this is an area where validation studies are required. That leaves direct interpretation of the dendrogram, which should ideally be done by epidemiologists and microbiologists together. Over time, epidemiologists may well learn just by experience what clusters would be relevant to investigate further. Initial experience in some countries shows that this is possible, and that follow-up of events is often limited by available resources rather than by issues with interpreting the dendrogram. A particular issue here is the type of dendrogram, which can be an unrooted minimum spanning tree, a rooted tree that assumes the same distance to the root for all isolates (such as based on UPGMA), or a rooted tree that does not have that assumption (such as

based on neighbour-joining). For the minimum spanning tree, which does not express evolutionary relationships, it should be kept in mind that this representation is only appropriate for closely related isolates, i.e. those that are likely part of the same epidemiologically relevant cluster.

Secondly, in addition to the dendrogram column, it may also be worthwhile to add columns to the line list with relevant predicted (or actual) phenotype information. Information about vehicles that have been associated with closely related isolates in the past could also be included (see also Section 3.6 on predicted phenotypes).

Finally, in order to maintain an overview of the entire set of available isolates rather than just a subset that contains isolates related to a single event, a zoomable view should be possible. This could be in the form of an unrooted tree where the zoom function represents the similarity threshold applied. The isolates lumped together by applying the threshold can be represented as circles whose size is proportional to the number of isolates. This is already a representation available in many software packages. In addition however, clicking on the nodes in the tree could trigger either further zooming in on only that node, or opening the line list for all the isolates in that node. As such it would be possible to quickly navigate through the entire available dataset by means of genetic similarity.

4 Collaboration between organisations

This chapter describes the different aspects of collaboration between microbiologists, epidemiologists and bioinformaticians from different organisations and countries. First, a model process for analysis of data and communication between organisations is put forward in Section 4, in order to structure the matter in a way that is closely aligned with routine practice. The international equivalent of this process is described in Section 4.2. Any collaborative resources required during these processes, such as global databases, are described in further detail in Section 4.3. Finally, in order to produce quality results throughout this process, that are comparable between different organisations, it is necessary, at a minimum, to assess the ability of each organisation of doing so, as described in Section 4.4 on quality assurance and transition to WGS.

4.1 Model process for data analysis at national level

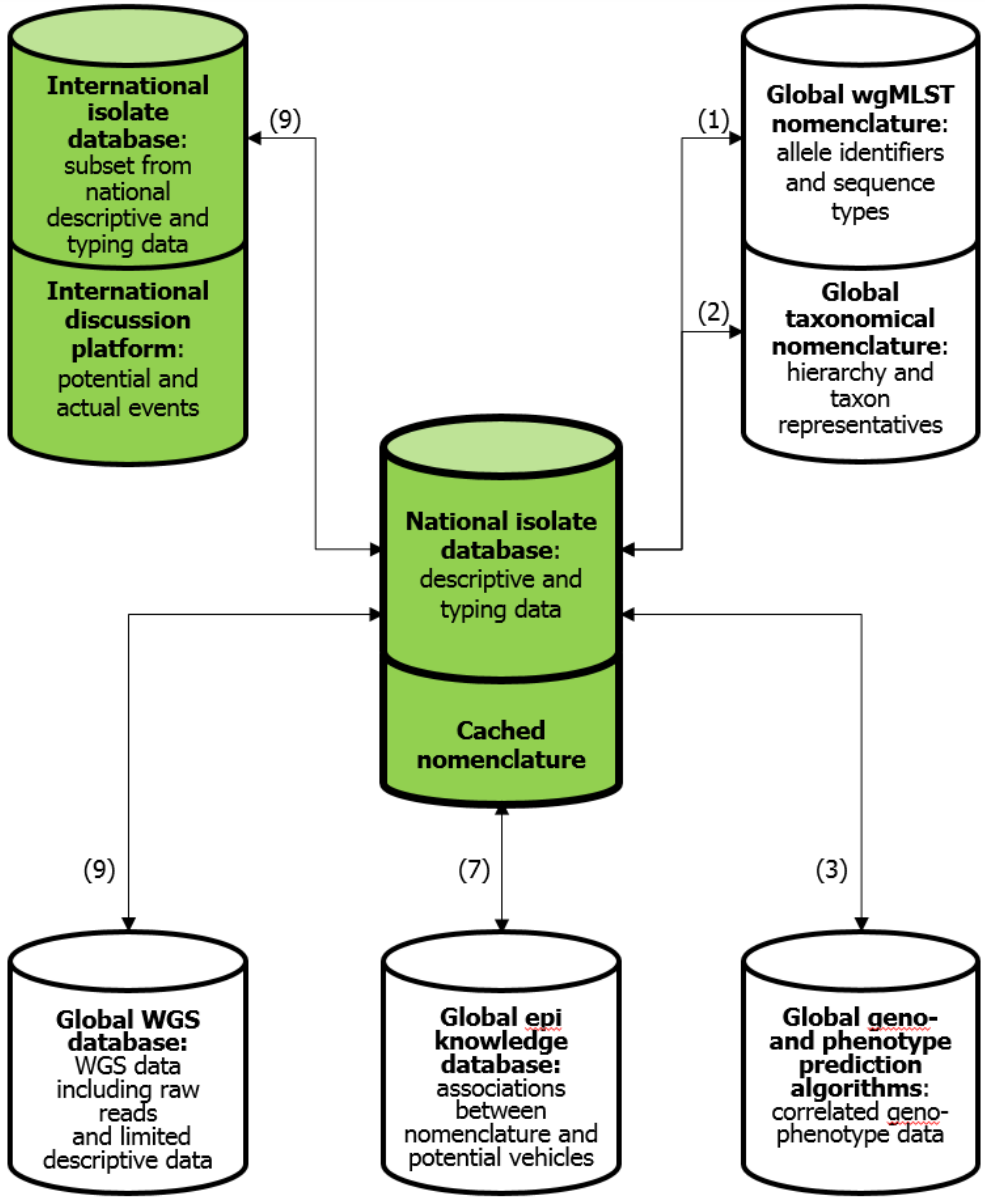
The primary objectives of typing for public health, i.e. identifying a potential common exposure of cases and monitoring the emergence of new genetic variants, either for prevention or response, are not altered due to the advent of WGS. If anything, those objectives can be better achieved as there is much more relevant information available. Correspondingly, the challenge will be, much more than before, to share and communicate this information effectively between people with different backgrounds, i.e. mainly microbiologists and epidemiologists, and between different organisations.

To structure this complex topic, a model process for data analysis and sharing will be presented, as summarised in Table 6. This process is not intended to be a standard for national reference laboratories to follow, but merely to capture the essential steps in a logical order. The process is written from the point of view of national reference laboratories, where most of the initial work is performed. It starts with a new batch of isolates that have been typed, at a minimum, with WGS, and that have been verified to be of the expected genus or species. Any collaborative resources that are required are added as well. A schematic overview of those interactions is given in Figure 3. Each step of the process is detailed further in subsections.

Table 6. Model process for data analysis, sharing and action at national level

Step	Description	Collaborative resources
1	Assign wgMLST or cgMLST allele identifiers and nomenclature to all isolates of the batch.	Global wgMLST nomenclature database
2	Assign taxonomical nomenclature to all isolates of the batch, based on wgMLST allele identifiers of step (1).	Global taxonomical nomenclature database
3	Predict relevant phenotypes and genotypes such as serotype, presence of virulence associated genes, food chain survival factors and antimicrobial resistance. Perform actual phenotyping where prediction is unreliable.	Global database with correlated geno- and phenotype data. Standard phenotype and genotype prediction algorithms
4	Find similar isolates in the national database based on cgMLST allele identifiers of step (1) or taxonomical nomenclature of step (2).	
5	Delineate individual microbiological clusters among the new batch of isolates plus the similar ones found in step (4), with high sensitivity for inclusion of isolates.	
6	For each microbiological cluster of step (5), create a high resolution dendrogram based on SNPs, or on wgMLST including common accessory genome loci.	Global wgMLST nomenclature database (potentially)
7	For each cluster of step (6), associate prior epidemiological knowledge such as potential vehicles from earlier events based on nomenclature of steps (1) and (2), and predicted genotypes or phenotypes of step (3).	Global or international database that associates nomenclature, including from earlier typing methods such as serotype, with for example potential vehicles, veterinary source and place and time previously observed.
8	For each cluster of step (7), decide on follow-up actions, if any, at the national level.	
9	Submit isolate data to international isolate database(s).	Standard database for real-time surveillance at international level. Global database for WGS data including raw reads.
10	Perform the equivalent of steps 1–8 at international level. Note: this step is added for completeness as it involves many other actors, and is further described in Section 4.2.	Standard database for real-time surveillance at international level. Standard communication platform for potential and actual events at international level.

Figure 3. Databases and other resources used during the model process at national level



Note: The numbers refer to the steps in the process. Green resources have restricted access.

4.1.1 Assign wgMLST or cgMLST allele identifiers and nomenclature

The starting point for this first step is when the raw read data for all isolates of the same species in a batch are finalised and converted into partially assembled genomes. Depending on the evolution of WGS technology and cost, it may well be that fully assembled genomes become the norm in the future. The assembly should in any case be done according to an agreed standard to enhance comparability with data from other organisations by reducing or eliminating variation due to the assembly process.

Allele identifiers can then be assigned to all isolates in the batch, which in turn forms the basis for many further analyses. The assignment should be done using a single global standard for allele identifiers that covers at least the core genome of the species in question (cgMLST), and potentially also the accessory genome (wgMLST). This ensures that every organisation uses the same allele identifier for the same allele sequence so that these data remain comparable across all organisations. A single global wgMLST nomenclature database – the term will be used further to also comprise cgMLST nomenclature database – must therefore exist for the pathogen in question, and that will be the authoritative source for allele identifiers (see Section 4.3.1). Any new allele sequence must be submitted to this database so that it can have a new identifier assigned.

In addition to wgMLST allele identifiers, MLST-based nomenclature can also be assigned to the new isolates. A single-sequence type, potentially with sublevels as described in Section 3.5.1, can be assigned to each

combination of alleles within a set of loci, thereby further reducing the data so that it can be communicated more easily. This assignment must be done by the global MLST nomenclature database, so that there is only one sequence type for each unique combination of alleles and the receiving party can unambiguously decode the sequence type again into alleles, for example to construct a dendrogram. The existing seven-gene sequence type can also be used, potentially as the first level of a sequence type with sublevels.

Whenever possible, any isolates of food, feed, animal or environmental origin that have also been sequenced, should be included in the analysis to increase the chance of identifying potential vehicles or sources. Depending on the country, these isolates are processed within the same laboratory or elsewhere, and the timeliness of their availability also varies greatly.

4.1.2 Assign taxonomical nomenclature

An important issue that arises with the sequence-type nomenclature assigned in the previous step is that it is fundamentally not describing phylogenetic relationships, even if the sequence type has sublevels. For example, sequence types 23–15 and 47–93 could be assigned to two sequences that differ by only one SNP in one of the seven housekeeping loci, as described in Section 3.5.1. Therefore, it will not always be the most efficient in communication between organisations on similarities between isolates, as the actual similarity cannot directly be derived from only the sequence type unless it is identical.

A truly hierarchical or taxonomical nomenclature that guarantees to the receiving party how close its own isolates match to it without further having to verify this should therefore also be considered. As described in Section 3.5.2, research is being performed in this direction. Essentially, isolates are classified in a simplified phylogenetic tree and assigned a corresponding hierarchical number sequence, e.g. 4.12.81.76.8. When this code is used in communication, the receiving party that has, for example, an isolate classified as 4.12.81.76.13, will know without further need for verification that this isolate matches the sender's fairly closely, differing only in the last level of the hierarchy. A difference in only the last level of the hierarchy would then correspond to, for example, an average of 10 alleles or SNPs difference, as defined separately for each species by the global taxonomical nomenclature database (see Section 4.3.2).

4.1.3 Predict relevant phenotypes and genotypes

The previous two steps used the WGS data to derive information on phylogenetic relationships. In this step, the WGS data is used to predict relevant phenotypic and genotypic characteristics as described in more detail in Section 4.3.3. Table 7 gives an overview per pathogen of what are likely the most relevant characteristics to predict. Clearly, antimicrobial resistance is one of the most important phenotypic characteristics for many pathogens, although it may not be possible to predict it with high accuracy, depending on the mechanism of resistance. If the predicted phenotype for a particular isolate is not considered reliable, the actual phenotypic test should be performed. With the exception of antimicrobial resistance and the *E. coli* virulence genes, the other listed phenotypic and genotypic characteristics are almost exclusively for backwards comparability, including using knowledge from prior events (see Section 4.3.4).

The prediction of phenotypic and genotypic characteristics should be done according to standardised algorithms, ideally derived from a global database that stores correlated genotype and phenotype data, so that predicted characteristics can be assumed to be the same when shared between organisations. These algorithms can be available online, or be implemented according to the standard in locally used software, provided that the parameters and algorithms are kept up to date. In the former case, the sequence data, or at least the relevant part required for the prediction, have to be sent to the online resource, but at the same time any changes to the standard are immediately available to everyone. For predictive algorithms, and especially antimicrobial resistance, this is an important aspect as it is expected that the parameters for the algorithms change over time as more data become available.

Table 7. Relevant phenotypes and genotypes that could be predicted

Pathogen	Phenotypes	Genotypes
Salmonella	Antimicrobial resistance, serotype (O and H antigens)	Typhimurium and Enteritidis MLVA pattern[1], PFGE pattern[1]
E. coli	Antimicrobial resistance, O-antigen, H-antigen	Vtx1, vtx2, eae and other virulence associated genes' presence and allelic variants, PFGE pattern[1]
Listeria	Serotype, disinfectant resistance[1]	Seven gene MLST sequence type, pathogenicity associated genes, PFGE pattern[1]
Campylobacter	Antimicrobial resistance	Seven gene MLST sequence type, PFGE pattern[1]

[1]*Not currently possible, but potentially in the future.*

4.1.4 Find similar isolates

After wgMLST allele identifiers and nomenclature have been assigned, the existing database of isolates can be queried to find isolates that are similar to those in the new batch. Such a search should emphasise sensitivity over specificity, to have a very high probability that any earlier isolate that may have an epidemiological link with some of the new ones is picked up. Other criteria than microbiological ones, for example a maximum time difference between new and matching isolates, should not yet be applied at this stage.

The search can be done by comparing the cgMLST allele identifiers from the new isolates against the database and keeping only the matches with at most N differences to at least one of the new isolates. This threshold of N differences will have to be determined per pathogen. Alternatively, if a taxonomical nomenclature is being used (see Section 4.1.2), the search can be done based on this, keeping only the matches on all but the last level of the hierarchy, for example. This has the advantage that it is extremely fast, which in the case of large databases may be an advantage compared to conducting a search based on cgMLST allele identifiers.

4.1.5 Delineate microbiological clusters

Once the new isolates have been matched with similar isolates in the database, individual microbiological clusters, i.e. clusters purely based on microbiological criteria, should be either defined for the first time or, if already existing, expanded. A dendrogram must therefore be created for all these isolates. This should be based on cgMLST as this has the advantage that the distances observed have the same scale compared to analyses of previous batches because they are based on the same set of loci. That makes it easier to gain experience and to establish general rules than is the case when using a distance metric that has a different scale for each dendrogram generated. Another possibility could be based on k-mer distributions, although experience on this in a routine public-health context is very limited at present.

Clusters can subsequently be defined by applying a cut-off on the dendrogram in terms of maximum number of allele differences. As in the previous step, sensitivity should be still be emphasised over specificity at this point in order to ensure a very high probability that any isolate that may have an epidemiological link with another one is included in the same cluster.

4.1.6 Create high-resolution dendrogram for each microbiological cluster

For each of the microbiological clusters delineated in the previous step, a high-resolution dendrogram should be created to show the phylogenetic relationship between them in more detail. This dendrogram can be based on wgMLST, using all loci common to all isolates in the cluster, or SNP based. If SNPs are used, these can be derived by selecting one of the isolates in the cluster as the reference and aligning all others against its de novo assembly. As such, by definition only the part of the sequence that is common to all isolates in the cluster will be considered. At present, aligning the raw reads to the reference rather than the assembled sequence is very likely to yield substantially higher quality results.

For isolates that are already preselected to be very similar, as is the case here, a dendrogram based on SNPs, potentially still using filtering, may well be more accurate than one based on wgMLST. This is because SNPs are real evolutionary events, and there will be fewer other evolutionary events among these sequences, such as acquisition or loss of mobile genetic elements, that could be misinterpreted as (consecutive) SNPs. In addition, the reference against which the isolates are aligned to enumerate SNPs can also be chosen much closer to all isolates, especially when it is chosen among the isolates themselves. If accuracy is too low, the wgMLST approach can be used. In that case, the global nomenclature database for the pathogen must also cover the accessory genome. Alternatively or complementary, comparison software could perform allele calling on the accessory genome and assign temporary allele identifiers to allow the full comparison.

The importance of this high-resolution dendrogram versus the cgMLST dendrogram of the previous step will likely depend on the species. A cgMLST dendrogram by itself should already be much more accurate and precise than

any of the current typing techniques. It does, however, not take into account the accessory genome, which especially for *E. coli* and likely also *Campylobacter* will be important, if not required, to correctly assess the similarities between isolates. Further research is required to determine whether this is the case. Finally, for more forceful follow-up actions such as control measures that may result in litigation, it may be advisable to make use of all the available information as recommended in this step.

4.1.7 Associate prior epidemiological knowledge with each cluster

With the microbiological clusters defined in detail, any available prior epidemiological knowledge, in particular on potential vehicles, should be collected for each of them. If the cluster includes isolates that were already in the database, any potential vehicles associated with these earlier isolates should be taken into account for the new isolates as well in the next step.

Ideally, a database is also available at the international level, either globally or EU/EEA-wide, that associates nomenclature with potential vehicles so that knowledge in this area gained by other organisations can be transferred efficiently. The current knowledge on this is mainly based on serotype, PFGE and MLVA, where the latter two can unfortunately not be derived from WGS. However, at least the knowledge associated with different serotypes could be pooled in such a database, and subsequently queried with the WGS-predicted serotype(s) for each microbiological cluster. Any new knowledge on potential vehicles or the veterinary source, where WGS was used for typing, could be stored as well and made queryable based on predicted serotype as well as cgMLST pattern and taxonomical nomenclature. A further description of such a database is given in Section 4.3.4.

4.1.8 Decide on follow-up actions for all clusters

The previous steps result in a detailed microbiological description of each cluster, including phylogenetic relationships, actual and predicted phenotypic characteristics and associated epidemiological information. From this point, epidemiologists and microbiologists work together to also include additional epidemiological information already available such as dates, age, gender, travel relatedness and place of notification of cases. Where available, statistical information on the frequency of occurrence of isolates of the same assigned nomenclature being above normal, should be included.

With all this information available, there is essentially no difference with respect to current typing methods for the remaining process of deciding on follow-up actions, e.g. case interviews. This process is led by epidemiologists with input from microbiologists. A remaining challenge is to present the information in a format understandable to everyone, such as a line list enhanced with a dendrogram or a minimum spanning tree, as described in Section 3.7.

4.1.9 Submit data to international isolate databases

Collaboration between organisations and countries is required due to the international dimension of FWD pathogens and food trade in particular. National reference laboratories form the first line of such collaboration, as they are normally the first to have sufficient information, i.e. the microbiological typing data, to allow the direct detection of clusters. As soon as typing data are available, they should be sent to an international database, along with some general information and a relevant date, e.g. the date of sampling or the date of receipt in the laboratory. In practice, and depending on the level of automation and available time, this will likely be done with some delay.

Data from all participating countries are then pooled in an international database and analysed in order to detect clusters at an international level. In practice, two different international isolate databases should be distinguished: those with restricted access for real-time cluster detection and follow-up (which includes detailed descriptive and other typing data), and those for public use where WGS data are made available for public use, potentially with a time delay and only limited additional data.

4.2 Cluster detection and follow-up at the international level

This process is equivalent to the national process described in the previous section, although with some important differences. Firstly, it requires an international database to which participating organisations agree to submit, in real time, their isolate data as well as epidemiological data and data from isolates of food, feed, animal and environmental origin. This has to be done through formal agreements, stipulating the rights and duties of the organisations that submit the data and the organisation responsible for operating the database, including funding. One fundamental aspect of this agreement is that the database is accessible only to authorised users, with well-defined access rights for different groups of users and a policy on public requests or data sharing. In practice, at the EU/EEA level, this has been implemented for PFGE and MLVA since 2012 in The European Surveillance System (TESSy) hosted by ECDC. A truly global database, although ideal because of the global nature of the food trade and travel, is at present not yet feasible, but should remain a long-term goal [21].

Secondly, the analysis of WGS data (e.g. cluster detection and further follow-up including communication) conducted at the international level should be as equivalent as possible to the analysis conducted at the national level. A very important question is therefore which WGS data are submitted to the international database. The relevant options are given in Table 8, together with the main types of analyses that can be conducted. Option 1, submission of the raw read data, would allow assessing the quality of the raw reads and perform the assembly again in a way that is guaranteed to be standardised, as well as all subsequent analyses. In addition, it allows creating high-resolution, SNP-based dendrograms that are of significantly higher quality than those based on assembled genomes. In addition, when sequence data are derived from different sources, using raw read data for allele calling, without assembly, may at present be the only way to generate results that are comparable across different sequencing technologies. This approach has high data storage needs, requires a high bandwidth and potentially also a high level of computing capacity if this is to be done for all isolates. This option is therefore only realistic when raw reads are only provided for some isolates, for example when they have been identified as belonging to a relevant cluster. This would be based on prior submission of either the assembled genome or wgMLST data. The raw read data quality control and standardised assembly for all isolates can be replaced by a standardised approach implemented by the data provider.

Within the remaining options 2–4, it is clear that the assembled genome allows not only the highest number of analyses, but also copes with the practical issue of new accessory genome loci only being available at a later time, after manual curation, in the standard nomenclature database. The latter is crucial to maintain consistency in the accessory genome description over time, as earlier submitted isolates that have a newly confirmed locus should retrospectively be assigned the corresponding allele identifier. Therefore, in order to have a high-resolution dendrogram (also when based on SNPs), which is especially relevant for *E. coli* and likely also *Campylobacter*, only options 2 or 1 are realistic, i.e. providing the assembled genome or raw reads. The main technical argument against option 2 could be storage requirements, although as described in Section 2.3.3, a partially assembled genome for the four major FWD pathogens requires between 1.6 and 5 megabytes per isolate (uncompressed). Combined with data compression, these amounts should be manageable for any international database. A potential alternative could be a separate international database with restricted access, dedicated to sequence storage and analysis, to which the database with the descriptive data has access in order to retrieve, for example, high-resolution dendrograms. In either case, the technical aspects seem less of an issue for an international database than pathogen-specific needs, i.e. what types of data provide sufficient information, also retrospectively, and agreements on data sharing.

Thirdly, due to the large number of people involved, a standard communication platform is needed where potential and actual events can be discussed. As for the database, this communication platform may be accessible only to authorised users, with well-defined access rights for different groups of users. In practice, this has been implemented at the EU/EEA and international level since 2010 in the Epidemic Intelligence Information System for FWD diseases (EPIS-FWD) hosted by ECDC. There are also other systems, e.g. the Early Warning and Response System (EWRS), which shares information on risk management measures in Member States, and the Rapid Alert System for Food and Feed (RASFF). Both systems serve broader needs than real-time surveillance on FWD diseases. At the international level, there are the INFOSAN alert system and the PulseNet International forum. Integration between the communication platform and the database is also important, so that relevant data are easily accessible during the discussion. In the case of WGS data, the most immediately relevant data in a discussion would be the nomenclature, as it allows to quickly assess genetic similarity between isolates. For the MLST-based nomenclature, the hierarchical approach in terms of progressively less conserved loci is likely the most useful, compared to a single sequence type based on a single set of loci that may be either too discriminatory or not discriminatory enough, depending on the event. If a nomenclature can be defined for the accessory genome, this should be included as well. Taxonomical nomenclature should be included if available. It should also be possible to easily zoom in on different levels of similarity based on the nomenclature.

Finally, the collection of non-human origin data, i.e. of food, feed, animal and environmental origin, and their use together with human-origin data, is more complex at the international level. Different competent authorities are responsible within the country to collect non-human-origin data, and the economic impact when potential sources, especially food, are identified and communicated, can be significant. Clear procedures and actors must therefore be in place for risk assessment and risk management so that sufficient trust can be built for successful action across countries. At the EU/EEA level, the European Commission is responsible for risk management, whereas ECDC and EFSA are responsible for risk assessment. Currently, ECDC and EFSA are setting up a collaboration, at the request of the European Commission, to share molecular typing data of human and non-human origin. The latter will be collected by EFSA from the Member States, and a portion of these data, across all isolates, will be sent to a joint database (an ECDC TESSy module) for analysis. This will be done in formal agreement with all actors in this process.

Table 8. WGS data sharing options and corresponding analyses

ID	WGS data provided	Main analyses that can be done	Comment
1	Raw reads	Raw read data quality control and standardised assembly. All others from options 2–4.	High-resolution dendrograms based on all shared SNPs (see option 2) are at present of significantly higher quality when based on raw reads than on an assembled genome. At present, raw reads are also the only source for a reliable dendrogram when data are generated by different technologies. Large storage capacity, high bandwidth and potentially high computing capacity are required when raw reads are submitted for all isolates.
2	Assembled genome	Assembled genome quality control. High-resolution dendrogram based on all shared SNPs or wgMLST loci. Phenotype predictions that require non-coding parts of the genome. All others from options 3–4.	Allows retrieval of MLST alleles for new loci in the accessory genome as they become available after manual curation in the standard nomenclature database.
3	Whole genome MLST alleles	High-resolution dendrogram based on all shared loci. Phenotype predictions that only require coding parts of the genome. All others from option 4.	Assumes that every locus in the accessory genome is available through the standard nomenclature database at moment of submitting.
4	Core genome MLST alleles	Core genome MLST dendrogram and nomenclature. Taxonomical nomenclature. Phenotype predictions that only require coding parts that are in the core genome. Associations with prior epidemiological knowledge.	

4.3 Collaborative resources and databases

The model process for data analysis and sharing described in Sections 4 and 4.2 makes use of several collaborative resources, including nomenclature databases and predictive engines that are agreed to be the standard. The most critical success factor for any such standard resource is that data owners, typically national reference laboratories, are guaranteed how the data under their responsibility are used after they have been submitted, so that trust can be built and maintained. The extent to which individual countries and organisations will consider sending their data to a third party that hosts the resource will depend strongly on their preferences. In any case, the conditions on data use and publication must be very clear and also meet the countries' own legislation on data sharing. This has to be respected if the system as a whole is to work properly. In addition, changes to the standard must also be formally agreed by the public health community. A formal governance and terms of use, accepted by the public health community, is therefore a minimum requirement for all online resources that are to be used as a standard.

The second most critical success factor is the usability of the resource. This includes not only the specific functionality offered by the resource, but also its availability, stability and maintenance. It requires, at a minimum, that a stable source of funding and formal change management systems are in place. In the subsections below, the different resources are described, essentially in order of their importance for routine public health applications. It should be noted that several, if not all, of these resources could technically be combined into a single comprehensive resource that combines microbiological and epidemiological data for isolates of both human and non-human origin on a global scale. However, in practice this is not yet deemed possible due to the agreement that would then have to be reached on a global scale on the use of this data. Technically, it would also be a far greater challenge. For the wgMLST nomenclature, which is the most critical resource, a more detailed description is given on minimum requirements.

4.3.1 wgMLST nomenclature

The authoritative source for all wgMLST allele identifiers is the most important resource to implement as it enables efficient communication across organisations. The database must provide allele identifiers for the entire core genome of the pathogen in question, and potentially also for the accessory genome. Users must be able to access this resource through a range of scenarios corresponding to their needs, using either machine-to-machine communication or a browser. In the main scenario, either raw reads or an assembled genome are submitted by a public user through machine-to-machine communication, and subsequently allele identifiers for all known loci are returned. This must be accomplished without manual intervention from a curator, also for new alleles, so that this step does not cause additional work or a substantial delay during the daily routine use of this information (see Section 4.1.1). Sequence types for one or more schemes can also be returned at the same time.

In a second scenario, users must be able to make use of a cached nomenclature database to improve performance and, where desired, reduce or delay the amount of data that is submitted to the global nomenclature database. Especially when raw reads are used instead of assembled genome sequences, performance can be a substantial issue due to bandwidth limitations. Such a cached or local nomenclature database would also have to contain an

implementation of the standard allele calling algorithm, so that users can be sure that either their raw read or assembled genome data is analysed in the same way as in the public database. This implies that recent copies of the nomenclature database must be made available for public download as well as a specification of the allele calling algorithm, from which the cached nomenclature database can be constructed. In addition, it implies that single allele sequences that do not yet have an identifier in the cached database, must be submittable to the global database to retrieve a new allele identifier. An overview of this second scenario, which uses a cached nomenclature database, is given in Figure 4. In addition, these cached nomenclature databases may make it possible, if the global database is unavailable, to continue operations, provided that a mechanism is in place to work with temporary local identifiers for new alleles.

The database itself will need to use an agreed 'allele calling' algorithm to identify loci on a newly submitted sequence and extract the corresponding allele sequences. A list of unique alleles per locus and their identifiers, and the different sets of loci or schemes, must also be maintained. In addition, the wgMLST sequence type-like nomenclature for unique combinations of alleles each extracted from a single genome must be stored as well. A schematic overview of the data that this resource should contain is given in Figure 5.

Quality control and maintenance or curation of the database are also important. Users should get fully automated feedback on the quality of their submission before finalising it, or alternatively, receive feedback on the reasons for rejection, for example that the genome sequence does not seem to belong to the expected species, which could be due to a mix up or contamination. Beyond that, a manual control mechanism for the quality of new allele sequences is not required, except perhaps in the case of loci that have very similar sequences. In those cases, alleles may be assigned to the wrong locus before being compared to the alleles of that locus, which in turn may negatively impact allele calling for subsequent sequences. This behaviour does not occur when alignment to a fixed reference is used for allele calling, which is the currently used approach. As mentioned in conjunction with the main scenario above, manual curation would delay users in their routine work. Furthermore, having lower-quality allele sequences in the database has no effect on the assignment of unique allele identifiers and sequence-type nomenclature. If the database is also intended to be used for other purposes such as research, alleles of low quality can be identified and excluded from the analysis. In such a case, working with user accounts would make it possible to select only data from users that agree to be part of a particular research project. These users would also have to agree then on further terms on their submitted data.

There is, however, at least one clearly manual task left for a curator, and this is the identification of new loci in the accessory genome, in case this option is provided by the database. Newly submitted sequences may well have genes not previously observed in other isolates, and when constructing a high-resolution dendrogram with all loci common to a small subset of similar isolates (see Section 4.1.6), the addition of any new loci will increase its accuracy. Over time, there may also be a need to define additional schemes, i.e. sets of loci, which is also a manual task to be done by the curator.

Based on the above, Table 9 describes the minimum requirements for any wgMLST nomenclature database. That is, all of these requirements must be fulfilled for a database to be usable as a standard for MLST nomenclature. In addition, these minimum requirements can be used as a starting point for a discussion on a more comprehensive standard, as there are many other issues to agree on, such as the allele-calling algorithm, whether raw reads can be submitted or not, and whether there should be a distinction between good-quality and low-quality sequence types.

Table 9. Minimum requirements for a wgMLST nomenclature database that must all be fulfilled

ID	Description	Rationale
1	There must be a formal governance of the database that includes also the main contributors, i.e. in first instance national reference laboratories or public health authorities in general, and which sets the terms of use as well as agrees on fundamental changes.	Without formal governance it is not possible to guarantee what happens with the data over time, and likely not possible to justify funding for the database by main contributors.
2	The database must have a minimum guaranteed uptime.	Since it is a global resource, the impact of the resource being offline is severe and this risk must therefore be mitigated.
3	The maintenance of the database, including curation such as identification of new loci, must be guaranteed.	As no system is without flaws and there is also nearly always a need for additional functionality, it must be possible to make changes in a reasonable time period.
4	There must be an open interface for machine-to-machine communication that covers all of the publicly available functionality.	This allows any person or organisation to develop solutions for the resource, making it used more widely and more useful.
5	Exports of the full database, including at least the loci, schemes, unique alleles, sequence types and data required for the allele-calling algorithm must be made publicly available, as must be the specifications for the allele-calling algorithm.	This, together with (4), allows any person or organisation to create a local version of the database for their own purpose, in particular a cached nomenclature database.
6	The retrieval of allele identifiers for new and existing alleles that belong to a known locus, as well as new and existing sequence types, must be done without manual curation steps. This functionality must be publicly available.	Without fully automated results for allele identifiers, the timeliness depends on the availability and priorities of the curator, which would cause a significant delay.
7	In addition to full genome sequences, it must be possible to submit single allele sequences and retrieve allele identifiers for them. This functionality must be publicly available.	This functionality allows retrieval of allele identifiers for new loci in the accessory genome, and only those, at a later time, after they have been confirmed through curation. In addition, this functionality is required to enable cached nomenclature databases. Finally, it accommodates those that do not wish to have their full sequence data stored in the nomenclature database, thereby increasing global adoption of the database. In any case, for those that do not want to submit their actual data, it is possible to simulate this scenario by repeatedly submitting the same genome sequence each time with one allele replaced by the actual allele.

Figure 4. wgMLST allele identifier retrieval using a cached nomenclature database

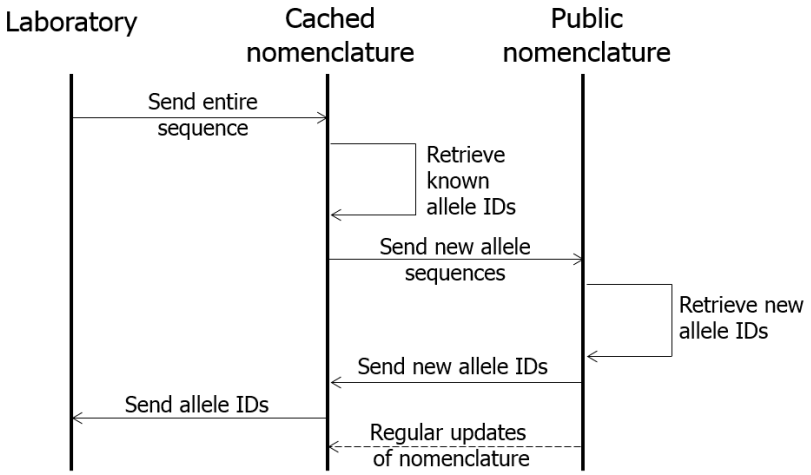


Figure 5. Schematic overview of the minimum data to be stored in (a) a wgMLST nomenclature database and (b) a taxonomical nomenclature database

a) wgMLST schemes

Scheme	Loci
PanGenome	A-B-C-D-E-F
CoreGenome	A-B-C-D
SevenGene	A-B
Sublevels	A-B / C-D

wgMLST locus definitions

RefGenome	Locus	Location
Ref1	A	50001-51123
Ref1	B	250001-251320
Ref1	C	300001-301566
Ref1	D	321285-322284

wgMLST allele sequences

Locus	Allele	Sequence
A	A1	ATGCATTAT…
A	A2	ATGCAATAG…
B	B1	ATGGGCCTG…
B	B2	ATGGGCTTG…

wgMLST sequence-type nomenclature

Scheme	SeqType	Alleles
CoreGenome	115	A9-B7-C5-D3
CoreGenome	267	A9-B7-C6-D3
SevenGene	23	A9-B7
Sublevels	23-15	A9-B7 / C6-D3

b) Taxonomical nomenclature

Taxon	RefGenome
1.3.4.1	A9-B7-C5-D3
1.3.4.1	A9-B7-C6-D3
1.3.1.7	A6-B7-C6-D3

4.3.2 Taxonomical nomenclature

The taxonomical nomenclature as described in Section 3.5.2 is in a much earlier stage of development than the wgMLST nomenclature, and a broad consensus on whether and how it should be used is still to be gained. However, the main scenario for the use of this resource can already be defined. Users would submit a full genome sequence through machine-to-machine communication and receive back a hierarchical identifier that represents the taxon to which the sequence belongs. As for the wgMLST nomenclature, a scenario with a cached database must also be possible here, to improve performance and, where desired, reduce or delay the amount of data that is submitted to the global nomenclature database.

The matching of a sequence against the simplified dendrogram that represents all the known taxa would presumably be done based on cgMLST allele identifiers, rather than SNPs, as the core genome remains constant over time and the required computing resources are much lower. Under this assumption, a new sequence would be matched against representatives of each taxon, and if a match is found within N alleles difference, the sequence would be considered to belong to that taxon. In this approach (as opposed to using SNPs), taxonomical nomenclature and wgMLST nomenclature would likely be implemented together within the same resource for efficiency.

This process may require a manual curation step in three different situations, as a result of which the retrieval of taxonomical nomenclature may be slower. Firstly, when no match is found and the sequence is thus likely the first member of a new taxon, it should be verified if this is indeed the case. Almost certainly, this will require that the raw read data are made available to the curator, so that the sequence quality can be assessed as well. In the meantime, a partial nomenclature up to the parent node of the putative new taxon could be assigned still in real time during submission. For example, the isolate in question is assigned code 1.1 in real time because it does not belong to the two known taxa 1.1.1 and 1.1.2. Later, when it is decided to create a new 'child' taxon 1.1.3 to which this sequence would belong, its nomenclature can be updated by resubmitting the cgMLST allele identifiers. Secondly, as the number of members for a particular taxon grows, it will become more apparent which ones are good representatives. The choice of representatives will therefore likely have to be adjusted over time to improve the matching of new sequences. Finally, as time progresses each taxon will evolve further and the thresholds originally set to correspond to an average number of differences with its closest neighbour will start to correspond slowly to larger numbers. Over time therefore, an additional hierarchical level may have to be added, either to all or some taxa.

4.3.3 Genotype and phenotype prediction

The creation and maintenance of standard algorithms for genotypic and phenotypic prediction would benefit substantially from having a global database in which correlated geno- and phenotype data are stored and pooled. For antimicrobial resistance, both the WGS data and the phenotypically measured quantitative result would be stored. Based on this, parameterised algorithms can be created that produce a quantitative or qualitative prediction for the phenotype in question. These algorithms can then be made available to users for online access both through a web browser and through machine-to-machine communication. A local implementation, using the same standard parameters should also be possible, implying that both the algorithms and their parameters, including updates, should be made available publicly. A very important practical aspect is whether the predictive algorithms require raw reads as input or can work with an assembled sequence, and whether there is a substantial difference in quality between the two. If raw reads are preferred or required, bandwidth requirements for an online implementation will increase substantially.

Predictive algorithms usually benefit substantially from being trained on representative and up-to-date data, and also require substantial research to develop. For this reason, it is likely that for any resource that is to be agreed to be the standard for a particular genotype or phenotype prediction, additional agreements will have to be in place with data providers so that well characterised data are provided and peer-reviewed publications are made with their consent. Finally, for phenotypes that are driven by horizontally transferable elements, algorithms could benefit from pooling data across different species, so that relevant plasmids observed in one species are also by default detected in other species.

4.3.4 Prior epidemiological knowledge

This resource would allow users to associate isolates, through typing information or derived nomenclature, with prior epidemiological knowledge. For example, for a newly detected cluster of *Salmonella* isolates, users could query this resource based on the wgMLST type, taxon or predicted serotype. The query result could consist of a list of potential vehicles that have been associated with this type, and when this occurred, as well as a list of relevant publications.

Such a resource would likely not be very challenging to set up from a technical point of view, and could be part of ongoing initiatives such as the COMPARE project, which is funded by the European Commission's Horizon 2020 framework programme for research and innovation. However, it would require substantial and continuous manual effort to keep up to date with new findings. One way to facilitate this could be to agree on a standard, machine-readable format for reports from outbreaks that describe the different associations found. Finally, potential vehicles are very sensitive information that may be misinterpreted by non-professionals. It is therefore likely that such a resource would only be accessible to authorised users and its content would need to undergo a clear and agreed upon approval process.

4.3.5 WGS data repository

WGS data can also have much wider potential uses than real-time surveillance. There is therefore also a need to share these data publicly. At present, there are several global sequence repositories where these data can be submitted to: the European Molecular Biology Laboratory's European Nucleotides Archive (EMBL ENA) and the National Center for Biotechnology Information's Genbank (NCBI GenBank), two established public databases, and some countries already send their routine WGS data there [12,13]. The data could additionally be submitted to the standard nomenclature database or even an additional database dedicated to sequence analysis.

The main questions in this context are how timely these data should be submitted or made public, what additional descriptive data (e.g. information on time, place, person or type of food) should be submitted, and under what legal agreement, if any, this should be done. Each national reference laboratory, or in general the organisation that owns the data, should define this for themselves. The sampling frame (Section 1) is also important in this context, as it should define what data are to be sent to the national reference laboratory, and hence constrains what data could be forwarded to the public WGS data repository. Over time, as WGS becomes more common, a guideline can be developed to ensure that public sharing of data is handled consistently.

4.4 Quality assurance and transition to WGS

In order to reach and maintain an acceptable level of quality throughout the entire process, starting from the generation of WGS data and then through the entire process of analysis and sharing, several measures can be taken. Firstly, training is needed for laboratory personnel to perform WGS. This includes data processing up to, and including, genome assembly, and such training courses are already available from several sources. Microbiologists will also have to be trained on the proper use of software tools and pipelines for data analysis, as well as on the use of collaborative resources, unless this task is not already taken care of by bioinformaticians. Finally, epidemiologists and likely also risk managers and policy makers will also require training on WGS and on how they

can best make use of the information. IT departments of organisations that are, or will be performing, WGS should likewise be informed on the requirements for their infrastructure.

Secondly, external quality assessment programs (EQAs) will be needed to verify both the laboratory work for generating the raw read data, and the subsequent software pipeline for further processing the data into an assembled genome, including its analysis. The first such programme is already running, organised by the Global Microbial Identifier (GMI) project. In the United States, PulseNet is also performing EQAs and certification. An ISO standard on whole genome sequencing for typing and genomic characterisation of food products is under development. A significantly more demanding step that also needs to be repeated annually would be accreditation of the laboratory for the whole process, which is at present virtually uncharted territory. In the food sector, the general principle is that all official laboratories performing official controls have to be accredited, but in the public health sector this is not necessarily the case, and in the absence of a legal requirement, each laboratory should decide for itself if it wants to be accredited for WGS. One of the main questions here is whether the software itself should then be accredited as well, which would likely raise its price. If the software is considered to be intrinsic to the generation of the result, it may be considered as part of the system and thus should be evaluated as well. Given that software usually continuously evolves, this may be complex and time-consuming. Potentially, the software could implement a particular standard, and maintain a record of what processing was applied to which isolates. Another relevant issue is that accreditation generally requires a validated and published protocol to which a comparison can be made. For the analysis step, this would normally also have to include the use of a standard nomenclature, which is not yet in place. Therefore, an accreditation of the analysis step is currently very difficult.

Thirdly, there is a need for validation studies for the different collaborative resources. For the wgMLST nomenclature, for example the definition of different schemes, i.e. locus sets such as the core genome, would have to be made based on sufficient supporting evidence. Another example would be the cut-offs in terms of average number of alleles or SNPs difference that mark the boundaries between the levels of the taxonomical nomenclature, and the selection of representatives for each taxon. Also, the accuracy of genotype and phenotype predictive algorithms has to be well established before it can be applied. Such studies could be conducted in collaboration between the public health sector and the food safety sector, and with other initiatives, such as the COMPARE project.

Finally, laboratories that already perform WGS on a routine basis should retain capacity for, and ideally still perform, current typing methods that would otherwise be phased out, until there is sufficient capacity for WGS available to other countries. This would still allow producing comparable data when needed. At the same time, for every current typing method that is phased out, residual capacity should be left in at least one or a few national reference laboratories. These would agree to perform such typing at the request of other laboratories, in order to allow for a transition period which would minimise the risk of gaps in typing capacity. Ideally, this should be coordinated on a global scale. In addition, laboratories that do not yet perform WGS can work together with other laboratories that already have WGS capacity in order to have their isolates sequenced.

5 Conclusions

It is expected that WGS will eventually become the sole standard method for genotyping of FWD pathogens for public health purposes, with additional phenotypic tests, e.g. on antimicrobial resistance, being performed in situations where the phenotype cannot be sufficiently reliably predicted from the sequence. In the meantime, laboratories that already perform WGS should ideally also still perform the current typing techniques, at least on selected isolates, such as outbreak-related ones, so that data remain comparable across organisations and can be used for further validation as well. After that, an efficient solution should be in place to retain some residual capacity for the current typing techniques in only a few laboratories, to which other laboratories can send their isolates if needed.

Setting up WGS as a new typing method for FWD pathogens is far from trivial, in particular when it is intended for routine use for public health. This report describes in relative detail many of the aspects that should be taken into account, systematically covering the entire process from sample provision and sequencing to data analysis and collaboration with other organisations. As such, it is meant as a guide for countries that are planning to introduce WGS for routine public health purposes.

5.1 Sample selection

Regardless of the typing method, the selection of samples that is available for typing, and the descriptive data collected for them, largely determine what public health objectives can be achieved and to what extent, primarily then the identification of a potential common exposure of patients and the monitoring of the emergence of new genetic variants. For human samples, this may be impacted by increased usage of culture-independent diagnostics.

Also, it is likely that WGS will not always be performed only in national reference laboratories but over time in hospitals as well, particularly in those that are also engaged in research. In those cases, the corresponding data, rather than the physical samples, must still be pooled at the national level so that geographically dispersed signals do not remain undetected.

In addition to this, it would be very beneficial to have positive food samples available for real-time typing as well, since this allows complementing the traditional route of epidemiological investigation that is often unable to identify a potential vehicle or does so when the corresponding batches are no longer available for sampling. A legal framework, as already in place in some countries, can be very helpful to ensure that national reference laboratories receive a sufficient selection of human samples for further typing. The equivalent framework for positive food samples would substantially increase the probability of identifying potential vehicles. [21].

5.2 Sequencing, cost and timeliness

The actual laboratory work using WGS as the standard genotyping method will become simpler as only one genotyping method needs to be used per pathogen. In addition, the differences between pathogens are often small and limited to the DNA extraction process rather than to library preparation. It is therefore also easier to pool typing capacity. At the same time however, protocols for DNA library preparation provided by the manufacturer often benefit from some optimisation and hence are not yet standardised. Thus, the inter- and intralaboratory reproducibility of WGS results also needs to be assessed better.

From a cost perspective, on a per isolate basis, WGS can be less expensive than current typing methods for *E. coli* and *Campylobacter*. For *Listeria*, the cost is more or less the same, and for *Salmonella*, depending on the throughput, the cost can still be somewhat higher. The total time required for WGS is already comparable to that of current typing methods. As WGS technology is evolving rapidly, the cost and total time can be expected to decrease further. Overall, the method is expected to become less expensive for all pathogens compared with current typing methods. Taking into account also the higher accuracy of the method for delineating epidemiologically relevant clusters, the potential for preventing additional cases through earlier detection is also higher than for current typing methods.

5.3 Data storage and analysis

WGS is different from current typing methods because it requires – in addition to the actual laboratory work – substantial data processing, storage and analysis to extract useful information from the large amount of generated data. The required storage capacity needs to be taken into account when planning the introduction of the method. Storage needs only become an important factor when reaching several thousands of isolates and consequently the terabyte range. Storage requirements are almost solely driven by the raw read data, rather than the assembled genome, and it is likely that these data need to be retained for litigation purposes as well. Similarly, computing

capacity also needs to be taken into account, but likely only becomes an important factor when processing more than one hundred isolates per week.

The routine analyses for public health purposes are not yet standardised. However, a model process for routine analysis is proposed, starting from raw reads. Genome assembly, rather than alignment to a reference genome, is likely to become standard practice and is already used in many cases. An exception to this is SNP-based analysis, which is likely of substantially higher quality when based on raw reads rather than assembled genomes. The assembly algorithms vary at present, and it remains to be seen what the impact will be from this source of variation. Over time however, as sequencing technologies, and read lengths in particular, improve, assembly is likely to become less and less a source of variation.

The assessment of genetic similarity between isolates, typically through construction of a dendrogram, can be done using two main approaches. The wgMLST approach extends the established MLST concept to the entire genome, or typically the core genome shared by all members of the same species (i.e. cgMLST), and looks at differences at the locus level. The SNP-based approach on the other hand looks at individual point mutations not only within loci but across the entire genome shared by the isolates that are being compared. Likely, the cgMLST approach will be used to create an initial, standardised dendrogram for isolates with unknown similarity, whereas the SNP or shared-genome MLST approach will subsequently, if needed, be used for a higher resolution dendrogram of a cluster delineated by cgMLST that can then also include the shared accessory genome. The latter is important for pathogens with a substantial accessory genome that may need to be compared in order to have sufficient discriminatory power. This is almost certainly the case for *E. coli* and likely also for *Campylobacter*, whereas for *Salmonella* and *Listeria* it is at present unclear. Also, for more forceful follow-up actions such as control measures that may result in litigation, it may well be advisable to make use of all the available typing information.

Additional analyses include the prediction of relevant genotypes or phenotypes such as antimicrobial resistance and the presence of virulence genes. Phenotypes, such as the serotype, that are relevant because they have substantial prior epidemiological knowledge associated with them can be predicted from the sequence data. Over time, a standard for each of these predictive algorithms, with well-defined accuracy, should be selected or developed.

5.4 Collaboration

Collaboration between organisations within and across countries is essential for FWD pathogens, because exposure is frequently not geographically restricted due to the international dimension of food trade. A common nomenclature is required to enable efficient communication between organisations. This is currently under development by several organisations, in particular in the form of cgMLST or wgMLST nomenclature databases.

A second type of nomenclature is a taxonomical nomenclature, which is in an earlier stage of development. However, such a truly hierarchical nomenclature is much preferred from an epidemiological point of view and may well become the standard over time. It is compatible, if not synergistic, with wgMLST nomenclature databases since both nomenclatures can be based on wgMLST allele identifiers. For both types of nomenclature, clear agreements will have to be in place, particularly on where the database is hosted, what functionality it has, how it is managed, and how data submitted to it may be used. Corresponding minimum requirements are proposed that would have to be fulfilled to allow international adoption. It is not expected that such nomenclature databases would require manual curation for the assignment of new alleles, but curation would be required for the identification of new loci or new reference sequences.

As is the case for the current typing methods, an international database is also required where data on actual isolates, both from human and non-human origin, can be shared in real time and microbial clusters are detected across multiple countries. This not only includes the WGS nomenclature, but also the partially assembled genome itself and potentially raw read data so that dendrograms can be constructed. Any other predicted or actual phenotypic measurements and descriptive data, such as relevant dates, travel relatedness, age and gender, can be included as well. It is not expected that such a database can be created on a global scale in the near future. At the EU/EEA level however, ECDC's TESSy database, together with the EPIS-FWD communication platform, has taken up this role since 2012 for PFGE and MLVA-based typing. It should be explored if this can be expanded to include WGS information. At present, TESSy is likewise being expanded to also become a joint database in which EFSA's molecular typing data collection on non-human origin isolates is shared and analysed together with the human origin isolates. Likely, submission of WGS data to any international database would in the first instance comprise cgMLST data or nomenclature, or taxonomical nomenclature for all isolates. The assembled genome could be included as well, or only after a relevant cluster is detected. The raw read data may at that point also be required for the isolates in the cluster to enable accurate high-resolution dendrograms, in particular when based on SNPs. Analyses of WGS data could also be carried out by one or more separate resources, as proposed in the COMPARE project funded by the European Commission's Horizon 2020 framework programme for research and innovation.

A final resource that would be helpful on a global scale, and which would likely not be technically complex to build, would be a database that correlates the designated serotype or other current typing methods, as well as later on also WGS nomenclature, with prior epidemiological knowledge. This would enable experts to quickly look up what

potential vehicles have been associated with a particular serotype and possibly retrieve a list of relevant publications.

5.5 Future steps

A number of steps are proposed that should be taken in the future, in order to guide and improve the introduction of WGS as the routine genotyping method for public health. These are listed in Table 10, grouped by area of application.

Table 10. Proposed future steps

ID	Area	Description
1	Sequencing	Develop standards for raw read quality and coverage.
2	Sequencing	Assess intra- and interlaboratory reproducibility of WGS sequencing, taking into account the dependency between cost and coverage as well as different technologies.
3	Sequencing	Develop a discussion forum for WGS laboratory issues.
4	Sequencing	Assess further the variation introduced by genome assembly as well as partial assembly and its effect on downstream analyses, including allele calling and dendrogram construction.
5	Analysis	Develop software that can annotate (partially) assembled genomes with quality information per nucleotide, expressing coverage and sequence variability per position, and subsequently use this information in further analyses.
6	Analysis	Compare different methods for allele calling during wgMLST nomenclature assignment, including alignment to a consensus sequence and to a sequence profile in addition to alignment to a reference sequence with or without the use of a substitution matrix. Develop software for standardised allele calling.
7	Analysis	Compare different methods for filtering out of (false) SNPs as a pre-processing step before constructing a dendrogram.
8	Analysis	Assess further the epidemiological relevance of sequence diversity per pathogen, by sequencing isolates from known outbreaks, plus controls, from different countries. Determine per pathogen the relevance of the accessory genome for assessing the similarity between isolates. Develop relevant thresholds for taxonomical nomenclature.
9	Collaboration	Agree on a single standard global wgMLST and taxonomical nomenclature database for each pathogen that complies with all of the minimum requirements described in this report, and that is endorsed and contributed to by public health authorities around the world.
10	Collaboration	Agree on standard predictive engines for relevant genotypes and phenotypes for each pathogen and that are endorsed and contributed to by public health authorities around the world.
11	Collaboration	Explore whether and how investment in the sequencing of high-quality, closed-reference genomes should be made.
12	Collaboration	Explore whether a single global or EU/EEA-wide database can be established for each pathogen that contains epidemiological knowledge such as associations between potential vehicles and serotype or later on WGS nomenclature, and that is endorsed and contributed to by public health and food safety authorities around the world.
13	Collaboration	Explore whether ECDC's TESSy system should be expanded to receive raw read data, assembled genome data and/or wgMLST data from EU/EEA Member States, and/or be able to retrieve externally generated analysis data such as dendrograms, including those based on the entire shared genome within a cluster. Correspondingly, explore whether the EPIS-FWD system should be expanded to facilitate communication on WGS results, including the use of nomenclature.

References

1. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-Genome Sequencing for National Surveillance of Shiga Toxin-Producing Escherichia coli O157. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2015;61(3):305-12.
2. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. PloS one. 2014;9(2):e87991.
3. de Boer RF, Ott A, Kesztyus B, Kooistra-Smid AM. Improved detection of five major gastrointestinal pathogens by use of a molecular screening approach. Journal of clinical microbiology. 2010;48(11):4140-6.
4. Cronquist AB, Mody RK, Atkinson R, Besser J, Tobin D'Angelo M, Hurd S, et al. Impacts of culture-independent diagnostic practices on public health surveillance for bacterial enteric pathogens. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2012;54 Suppl 5:S432-9.
5. Iwamoto M, Huang JY, Cronquist AB, Medus C, Hurd S, Zansky S, et al. Bacterial enteric infections detected by culture-independent diagnostic tests--FoodNet, United States, 2012-2014. MMWR Morbidity and mortality weekly report. 2015;64(9):252-7.
6. Patel R. MALDI-TOF MS for the diagnosis of infectious diseases. Clinical chemistry. 2015;61(1):100-11.
7. Li J, Feng Q. Analysis of Gut Microbiome and Diet Modification in Patients with Crohn's Disease. SOJ Microbiol Infect Dis 2014:2(3): 1-4.
8. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research. 2010;38(6):1767-71.
9. Brandon MC, Wallace DC, Baldi P. Data structures and compression algorithms for genomic sequence data. Bioinformatics (Oxford, England). 2009;25(14):1731-8.
10. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome research. 2011;21(5):734-40.
11. Pathak S, Rajasekaran S. LFQC: a lossless compression algorithm for FASTQ files. Bioinformatics (Oxford, England). 2014.
12. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic acids research. 2011;39(Database issue):D28-31.
13. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic acids research. 2013;41(Database issue):D36-42.
14. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, et al. Benchmarking of methods for genomic taxonomy. Journal of clinical microbiology. 2014;52(5):1529-39.
15. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(6):315-27.
16. Ashton PM, Peters T, Ameh L, McAleer R, Petrie S, Nair S, et al. Whole Genome Sequencing for the Retrospective Investigation of an Outbreak of Salmonella Typhimurium DT 8. PLoS currents. 2015;7.
17. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nature reviews Microbiology. 2013;11(10):728-36.
18. Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, et al. Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. PLoS pathogens. 2012;8(6):e1002776.
19. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, et al. Salmonella serotype determination utilizing high-throughput genome sequencing data. Journal of clinical microbiology. 2015;53(5):1685-92.
20. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. Journal of clinical microbiology. 2015;53(8):2410-26.
21. EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), 2014. Scientific Opinion on evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 2 (surveillance and data management activities). EFSA Journal 2014;12(7):3784.

European Centre for Disease Prevention and Control (ECDC)

Postal address:
Granits väg 8, SE-171 65 Solna, Sweden

Visiting address:
Tomtebodavägen 11A, SE-171 65 Solna, Sweden

Tel. +46 858601000
Fax +46 858601001
www.ecdc.europa.eu

An agency of the European Union
www.europa.eu

Subscribe to our monthly email
www.ecdc.europa.eu/en/publications

Contact us
publications@ecdc.europa.eu

Follow us on Twitter
@ECDC_EU

Like our Facebook page
www.facebook.com/ECDC.EU

ECDC is committed to ensuring the transparency and independence of its work

In accordance with the Staff Regulations for Officials and Conditions of Employment of Other Servants of the European Union and the ECDC Independence Policy, ECDC staff members shall not, in the performance of their duties, deal with a matter in which, directly or indirectly, they have any personal interest such as to impair their independence. Declarations of interest must be received from any prospective contractor(s) before any contract can be awarded.
www.ecdc.europa.eu/en/aboutus/transparency

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or
  calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

Publications Office