|  |
| --- |
|  |



**Report on requirement analysis for an “OH Harmonisation Infrastructure Hub”**

**JIP1 - ORION - AI1 - 1st Call**

**Responsible Partner: NVI**

**GENERAL INFORMATION**

|  |  |
| --- | --- |
| **European Joint Programme full title** | **Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards** |
| **European Joint Programme acronym** | **One Health EJP** |
| **Funding** | **This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No 773830.** |
| **Grant Agreement** | **Grant agreement n° 773830** |
| **Start Date** | **01/01/2018** |
| **Duration** | **60 Months** |

**DOCUMENT MANAGEMENT**

|  |  |
| --- | --- |
| **JIP/JRP Deliverable** | **D-2.2 Report on requirement analysis for an “OH Harmonisation Infrastructure Hub”** |
| **Join Integrative/Research Project** | **JIP1 - ORION - AI1 - 1st Call** |
| **JIP/JRP Leader** | **Matthias FILTER (BfR)** |
| **Contributors** | Work package leader: Karin Lagesen (NVI) Work package deputy: Mohammed Umaer Naseer (NIPH) Other participants: **Thomas Haverkamp (NVI), Jeevan Karloss (NVI), Wonhee Cha (SVA)**, **Olov Svartström (FOHM)**, **Mia Torphdahl (SSI), Eva Litrup (SSI), Katrine Grimstrup Joensen (SSI), Lesley Larkin (PHE)**, Taran Skjerdal (NVI), Tim Dallman (PHE), Eelco Franz (RIVM), Tasja Buschhardt (BfR), Matthias Filter (BfR) |
| **Due month of the deliverable** | **M12** |
| **Actual submission month** | **M14** |
| **Type**  ***R: Document, report***  ***DEC: Websites, patent filings, videos, etc.***  ***OTHER*** | Report |
| **Dissemination level**  ***PU: Public***  ***CO: confidential, only for members of the consortium (including the Commission Services)*** | **PU** |

**EJP ORION — WP2NGS — Year report 2018**

Work package leader: Karin Lagesen (NVI)

Work package deputy: Mohammed Umaer Naseer (NIPH)

Other participants: **Thomas Haverkamp (NVI), Jeevan Karloss (NVI), Wonhee Cha (SVA)**, **Olov Svartström (FOHM)**, **Mia Torphdahl (SSI), Eva Litrup (SSI), Katrine Grimstrup Joensen (SSI), Lesley Larkin (PHE)**, Taran Skjerdal (NVI), Tim Dallman (PHE), Eelco Franz (RIVM), Tasja Buschhardt (BfR), Matthias Filter (BfR)

## 

## **Executive summary**

The main goal for this work package is to create a One Health Knowledge Hub (OHKH) for NextGen analytical data, methods, analyses and systems. The goal of the tasks in this work package in 2018 has been to discover the state-of-play within the field. This has been done through literature reviews, surveys, discussions and visits with relevant partners, and also discussions with EFSA and ECDC. Through this work, we have come up with what we believe to be a good framework and suggestions for what should be included in the OHKH within this area. This year report contains information based on our work within work package in 2018, and describes the draft framework of the NextGen OHKH.

## 

## **Defining the scope of WP2-NGS**

The stated goal for WP2 of this project is to create a One Health Knowledge Base (OHKH) - a cross-domain inventory of currently available resources for performing integrated One Health (OH) surveillance generation, data analysis and interpretation. The process of creating this inventory is divided into three parts - a knowledge discovery phase, a pilot development and testing phase, and a review and evaluation phase. This current document is part of the knowledge discovery process, and will help focus the work to be done in the pilot development and testing phase.

The focus for the WP2NGS work package is on NextGen analytical data, methods, analyses and systems. The inventory generated from this work package will also include best practices for data management and handling approaches for bacterial identification techniques. It should also include current experiences on practical issues such as data storage (both short and long term) and exchange capacity, data exchange platforms, harmonised data, harmonised terminology, as well as the need for bioinformatics expertise.

Through the first discussions of the work to be done within this work package, the scope was limited to bioinformatics methods for bacterial typing and in silico characterization, and usage of such data for surveillance purposes. These are methods that have received quite a bit of attention in recent years. Both EFSA and ECDC are recommending that public, food and veterinary health institutions implement such methods as part of their work processes. However, while public health institutions have come quite far in this process, food and veterinary institutions have not made the same kind of progress. Considering that we in the ORION team have participants from institutions that span the spectrum on where they are regarding NGS implementation, we see this as an area where this work package could contribute with guidelines and suggestions for how to implement such methods in a OH perspective.

### **Defining NGS based One Health Surveillance**

The management of complex health matters across human, animal and environment health sectors necessitates integrated policies and a transdisciplinary approach to health. In turn, this requires the engagement of a wide range of stakeholders from different professional sectors and decision-makers. The OH concept is such an approach that promotes collaborative efforts across sectors and disciplines to attain optimal health for humans, animals and their environment. Close collaboration between health systems is required in particular in the surveillance of health hazards involving humans, animals and their environment.

Per definition, surveillance is the systematic and ongoing collection, collation, and analysis of health-related information communicated in a timely manner to decision makers so that they can direct action. Currently there is no consensual definition of an OH surveillance system. Stärk et al. proposed the following definition: “*One Health surveillance describes the systematic collection, validation, analysis, interpretation of data and dissemination of information collected on humans, animals and the environment to inform decisions for more effective, evidence- and system-based health interventions*” [1](http://f1000.com/work/citation?ids=522820&pre=&suf=&sa=0). Importantly, the surveillance system must allow for the detection of unusual disease patterns and trigger disease control and prevention efforts. The overall goal of infectious disease surveillance is to help reduce the incidence and prevalence of infectious diseases by providing relevant public health information and knowledge to public health professionals, health care professionals, and decision makers to promote actions that result in the timely prevention and control of infectious disease.

Depending on the surveillance context and objective, the required collaborative efforts across sectors and disciplines might differ. They might be implemented for various activities throughout the surveillance process and engage different combinations of sectors, disciplines and decision-making scales. Collaboration must be tailored to the surveillance systems objectives and context, characterized by a wide range of factors including: epidemiological, microbiological, ecological, economic, social and environmental considerations, and must consider the constraints and expectations of all surveillance actors and end-users. In that sense, Next Generation Sequencing (NGS) data could improve current collaboration efforts in the OH perspective, since it is data that can be analyzed independently by each network participant and results generated by different participants should in theory be highly comparable. In addition, NGS differs from other molecular typing methods, since it can be generated from most isolates, and with higher discriminatory resolution. Thus when NGS data is used for molecular typing of pathogens it allows different sectors and disciplines to increase collaboration efforts by improving the detection of unusual disease patterns.

### **Molecular typing and NGS**

Molecular typing of pathogens that cause infectious diseases complements the traditional epidemiological surveillance by providing appropriate discriminatory analyses to aid the rapid and early detection of outbreaks, to detect and investigate transmission chains and the relatedness of strains. It also supports studies to trace-back the source of an outbreak and identify new risk factors, as strains can be linked more accurately to epidemiological and clinical data.

Molecular typing data can also be used to study characteristics of pathogen in different hosts, its spread over time and space, transmission dynamics of its disease, genetic shifts of strains over time, and the development of drug resistance across multiple generations. This information helps in understanding the disease mechanisms and improves infectious disease prevention and control measures.

Laboratory techniques have developed quickly over the past years, and public health investigations of infectious disease is rapidly shifting towards the use of Whole Genome Sequencing (WGS) as method for molecular typing. In many countries, WGS-based typing is in the trial implementation phase for use as the routine first-line or second-line typing method for national surveillance of a number of bacterial and viral diseases.

The advantages of NGS-based typing over other pathogen typing methods include [2](http://f1000.com/work/citation?ids=6491259&pre=&suf=&sa=0):

* the optimal resolution of the near-complete genomic sequence comparison for measuring inter-genomic sequence similarity, and inferring the most probable phylogenetic lineages of descent between isolates to infer the direction and route of pathogen transmission, from environmental, animal or human sources and reservoirs;
* the in silico prediction of phenotype and, in particular, acquired antimicrobial resistance mechanisms, pathogenicity and virulence determinants as well as correlates of epidemiological/ecological fitness associated with epidemic spread, also described as ‘high-risk clones.

## 

## **Knowledge discovery and problem definition**

The work done in 2018 has consisted of literature review, survey work, discussion meetings, and discussions with ECDC and EFSA.

The literature review has focused on papers which describe the use of sequencing for either clinical, surveillance or genomic epidemiology purposes. In line with what was described in the project proposal, we have mainly focused on the foodborne pathogens *Campylobacter*, *Listeria monocytogenes*, *Escherichia coli* and *Salmonella*. It is our aim to provide a curated list of papers that can work as a resource within the OHKH.

Based on the literature review, we created a survey that was sent out to all EJP participating institutions (n=40). The survey questions are attached as appendix I to this report. The aim of this survey was to discover what is actually in use in public, food and veterinary institutions in Europe, and also to discover potential barriers to implementation of NextGen technology for surveillance purposes. Only six of the institutions participating in the EJP participated in 2018. As a consequence, we invite more institutions to complete the survey, as well as perform interviews and site visits to selected institutions during Q1 of 2019.

However, the survey results collected so far showed that all but one of the responding institutions had started implementing NGS. For the institution that had not started yet, the reasons for not having implemented NGS were: financial costs, lack of personal, equipement and computing infrastructure, but also internal management decisions. All institutes that implemented NGS use it for WGS of bacterial isolates in light of different goals such as surveillance, outbreak responses or for research purposes. Some labs use additional techniques such as: amplicon sequencing (16S rRNA / ITS), transcriptomics (RNA viruses) and metagenomic shotgun sequencing analyzing feces or environmental samples, but it is unclear if these techniques are part of their surveillance programs.. Several labs indicate that they use other molecular typing methods (e.g. PCR, MLST, PFGE, MALDI-TOF) in addition to NGS. The survey respondents differ when it comes whether sequencing is done in-house. Some labs collaborate with other national institutions to perform this step. For the processing and analysis of the sequence data the labs use either in-house pipelines, open source software and/or commercial packages. Analysis of genes involved in AMR resistance or virulence is performed by all labs, but the species for which this is analyzed can differ between labs.

The results from the NGS pipelines survey will be documented and used as a basis for describing current practices in the OHKH. They will also be used for describing solutions that are in use for each of the different steps in current NGS pipelines that are available.

Throughout 2018, we also conducted several meetings, both within the work package and with stakeholders. The work package meetings mainly focused on defining the scope of this work package, as well as resource and knowledge discovery. We also had discussions with representatives from EFSA and ECDC. These discussions helped us be aware of the resources that these institutions have developed already, and also helped us define the scope of the work to be done in this work package.

### **Problem definition**

Through the work within this work package, we decided to divide the problem into four components, as illustrated in Figure 1. These components will form the framework for the OHKH for this work package. A closer description of the focus area and likely contents for each component, as well as the introduction, is included below.

Infrastructure

Sequencers

Compute systems

Storage/databases

Data sharing

Personnel

Pipelines

Workflow systems

Software/tools

Typing systems

Characterization schemes

xMLST

SNP analyses

AMR/virulence detection

Surveillance

Distance measures

Clustering methods

Cluster thresholds

Phylogeny

**Figure 1**: This figure describes the four components into which we have divided the work within this work package. The blue components are species agnostic, while the pink ones are not.

## **Preliminary draft of the structure of the OHKH/WP2-NGS**

The aim of the OHKH/WP2-NGS is to create a document that can be used as a guide for an institution that is planning on using NGS data and bioinformatics analysis methods as a basis for surveillance and outbreak detection and investigation. We are aware that there are many possibilities within this field, and that there are many local and national constraints that might come into play. Thus we will not in any way attempt to cover all possible options.

It should also be noted that there are inherent differences between what can be covered for the various components due to varying maturity levels within the fields, and also due to the extent that each component is affected by local and national conditions. This is especially relevant for the Surveillance component. While for the other components knowledge is gradually coalescing into a set of best practices, it is difficult to give general guidelines regarding how to use NGS data for surveillance. This is due to the fact that countries differ with respect to bacterial diversity of pathogens. There might also be human, animal and environmental health practices that require country specific surveillance procedures which will also affect NGS data usage and interpretation.

The OHKH is planned as a document that can be updated by others, including people outside of the original ORION consortium. For this purpose, we have established a github repository which will form the basis for a ReadTheDocs website, which will at least in the development phase be the chosen platform for this handbook.

### **- Preamble -**

This section will contain the introduction and problem definition of what will go into the OHKH. The main focus here is to enable users of the OHKH to obtain a quick overview over what is happening in the field, and who the main actors are. Contents in this introductory section is likely to be:

Curated literature list  
There is a lot of literature published today containing useful information regarding NextGen sequencing for typing, outbreak detection and investigation, and surveillance. This section will contain a curated list of papers that can be of assistance within this field. This list will be crowdsourced from the community.

Overview over ongoing work within EFSA and ECDC  
These institutions are doing a lot of work within this field. This section will link to and describe relevant initiatives and reports. This section will be created with input from the EFSA and ECDC representatives, to ensure that we include all relevant information.

Overview over other initiatives and resourcesThere is a lot of other institutions and groups that are doing work within this field, both generally, and for specific species. This include work done within GenomeTrackr, PulseNet, FoodNet, UNSGM, GMI as well as the CDC and WHO. This section will list and describe these for easy reference.

**- Component 1: Resource and infrastructure considerations -**

Implementing sequencing for surveillance within an institution requires access to resources such as sequencing and compute infrastructure, as well as human competence and funding. Within this section we aim to give an overview over available infrastructure, as well as link to or describe differences between various options. In addition, we will describe what kind of competence is needed for the various options.

Within this project, we are very aware that there are great differences between the institutions. They may vary not only in their access to resources, but their local and national surveillance goals might also be different. Thus it is a primary concern for this work package to provide options that can work under many different local conditions and scenarios.

**Infrastructure considerations**The main goal for this section is to describe the infrastructure that is needed to get sequencing data, as well as the infrastructure necessary to analyze the results. We will here link to or describe the various elements that make up the full analysis pathway, from laboratory information management systems (LIMS), sequencers, storage and compute resources, as well as data management and analysis platforms (DMAPs). Not all all elements may be necessary or feasible for certain local conditions or scenarios, however, we will include them to show what is possible and available.

Laboratory information management systems  
Laboratory information management systems are software solutions for keeping track of sample processing in a lab. Such systems are likely to include tracking where a sample is in a processing workflow, which other samples it was processed with, what reagents were used for processing that sample, etc. There are many such generic systems on the market, however there are a number of systems that have specialized in supporting NGS activities. Within this section, we will link to and briefly describe some of the systems that are commonly used for supporting sequencing workflows.

Sequencing Technologies   
There are many sequencing platforms available today. There are two short read platforms in common use, Illumina miSeq (IM) and Life Technologies Ion Torrent (IT). In addition, long read technologies such as PacBio and Oxford Nanopore Technologies (ONT) may be used. Quite a bit of literature concerning how the various platform compare already exists. In addition, these matters have been the focus for several projects such as the COMPARE project. There is also quite a bit of hands-on knowledge among the partners participating in the ORION project. Within this section, we will link to or describe the differences between the platforms and to what extent using one as compared to another may have on the outcome of analyses.

Storage and Compute infrastructures   
Storing (short term and long term) and analyzing sequencing data produced from sequencers can be cost and compute resource intensive. This section will list and describe the important things to explore when acquiring the needed storage and computing resources.

A primary concern that has to be discussed is where storage and analysis will take place. There are three likely scenarios for this: 1. an in-house storage and compute resource will be set up. 2. a cloud solution will be acquired. This might be in collaboration with other institutes or with public e-infrastructure providers, or possibly with a commercial organization. 3. storage and analysis will be done via a european or international provider of such services such as the solution being established by COMPARE. Several different factors will guide what solution is selected, and some institutions might go for a mixed solution. There might be legal, regulatory and financial considerations that might exclude some solutions, while access to expertise might exclude other solutions. Another major consideration is how much data will be produced and what analyses are planned.

Within this section of the OHKH, we will link to or describe what a storage and analysis set-up could look like under the various scenarios mentioned above. We are aware that there are many possibilities within this field, and that there are many local and national constraints that might come into play within this field, thus we will not in any way attempt to cover all possible options. However, we believe that describing some possible set-ups will enable local institutions to figure out what the local situation is so that they can tailor a solution that can fit their own situation.

Data management and analysis platforms   
Data management and analysis platforms (DMAPs) are platforms where a user can upload raw sequence data securely and get the desired reports as outputs after all the necessary analyses. A number of DMAPs have been developed and they are available and actively being used. Two such systems are the INNUENDO and the IRIDA systems. These are locally-installable systems that have such features as a user friendly graphical user interface, data management and access control, secure logins, analyses pipelines, tracking of how results have been produced, and analysis reports. In addition, the COMPARE project is producing a cloud based DMAP which can be used for analysis. Within this section of the OHKH, we will link to or describe how the various available DMAPs work, and describe the similarities and differences between the various DMAPs so that local institutions can evaluate whether any of these may work for their local situation.

**- Component 2 - Analysis pipelines -**

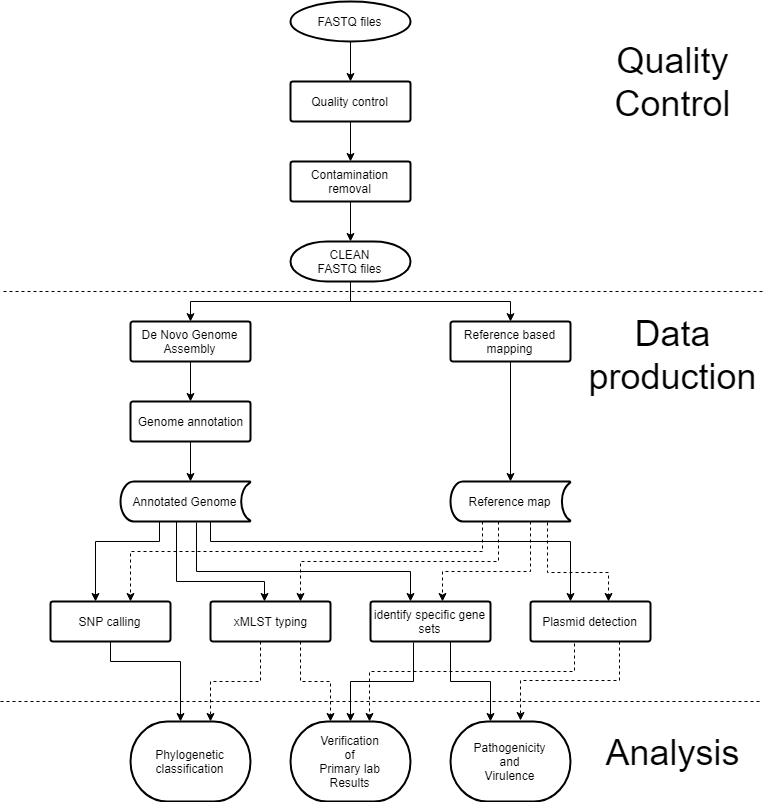
In recent years next generation sequencing has matured and with that a solidification on the required methods for WGS projects [3,4](http://f1000.com/work/citation?ids=3963817,5641017&pre=&pre=&suf=&suf=&sa=0,0) has occurred. Nonetheless, the current practice of microbial WGS is mostly confined to the academic world, and has not yet resulted in large scale introduction of WGS in clinical settings [5,6](http://f1000.com/work/citation?ids=5717913,4695443&pre=&pre=&suf=&suf=&sa=0,0) with the exception of a few countries. However, while it is clear that as the technological developments in bioinformatics proceed at a rapid pace, a consensus has formed on which steps are required to produce WGS datasets that can be used for surveillance and outbreak detection and investigation (Figure 2). These steps involve quality control of the raw sequence data and a subsequent usage of clean data to produce an annotated genome or reference map. Such “products” can then be used to obtain typing data, which is needed for the final analysis under surveillance or outbreak scenarios.

Quality control  
The first step in a WGS bioinformatics workflow is to perform quality control of the raw sequence data (Figure 2). Without this step downstream analysis can be significantly affected due to the presence of incorrect or contaminating DNA sequence. Such artefacts can then cause incorrect identification or characterization of a pathogen (e.g. virulent vs non-virulent strain detected), and incorrect assumptions can be made on the seriousness of the threat. Thus it is required to remove sequences of poor sequence quality (e.g. PHRED quality score < 20), or contaminants such as PhiX or human sequences. In the OHKB we will describe the background for these procedures, how this step is performed and what methods are available.

Data production  
Once the sequence data is clean, it can then be used in two different processes. i) Genome assembly, which is the computational process of combining all the shotgun DNA sequences in order to recreate the original genome sequence of an organisms. ii) Reference mapping, which is the alignment of the shotgun sequence fragments to a chosen reference genome, in order to identify differences between the reference and the investigated isolate (Figure 2). Most genome sequencing projects will choose one process for the analysis of their pathogenic isolates. In the OHKB we will link to or describe these processes to show their differences as well as possible reasons for selecting one method over another in order to analyze the pathogen of interest. For instance, when using the reference mapping approach selection of the right reference pathogen genome is very important because that will affect the analysis phase results and hence the interpretation. The OHKB will link to or describe the different steps in the data production process and how they affect the final products that can be generated.

In silico WGS based typing methods  
These final products could thus be i) a collection of genome single nucleotide polymorphisms (DNA sequence differences between the isolate and a reference genome), ii) a multilocus sequence typing (MLST) profile — e.g. the differences found in a set of genes shared between all isolates of the same species (core-genome MLST) or differences found in all genes that are present in isolate of the same species (whole genome MLST). iii) The differences in a number of specific genes involved in for example genes involved in virulence, antimicrobial resistance or toxins. iv) It could be a description of the identified plasmids or other mobile elements (Figure 2). Each of these final products can be generated in a specific way, and we will describe the methods that are used to generate these products in the OHKB. The final products can then used for phylogenetic analysis, the verification of primary laboratory results, the detection of genetically related isolates that may indicate a potential outbreak event or in the analysis of pathogenicity / virulence of the isolate. Such analyses are needed for epidemiological / outbreak investigations which are further described in component 4.

Virulence and antimicrobial resistance detection methods  
For all of the species included in this project, finding virulence genes and detecting antimicrobial resistance is of importance. These methods have commonly been developed to be able to detect virulence and antimicrobial resistance genes regardless of species. Additionally, in many cases, the methods in question can be used to detect both virulence and antimicrobial resistance genes by adjusting what external data source is used by the program.

****

**Figure 2**: Flowchart showing a standardized WGS bioinformatic analysis workflow. The standardized workflow can be divided into three sections: Quality control, Data production, and Data analysis. xMLST typing: The collection of MLST typing approaches including cgMLST, wgMLST

**- Component 3 - In silico WGS based characterization -**

In silico WGS based characterization is the means and methods by which sequencing data is converted to analyzable and actionable data. The methods and tools used will to a large extent be described in Component 2, while actions taken based on results from these analyses are described in Component 3.

The characterization done commonly focuses on two main issues: species and strain typing, and characterizing specific features found in the genomes. Both issues commonly make use of external data in the process of characterizing the genomes. Within this section of the handbook we will aim to refer to or describe methods used for characterization, as well as the external data sources used with those methods.

Species and strain typing methods  
One of the most pressing issues is to discover what kind of species and subtype the isolate in question is. Quite a lot has been done on this issue for the species we are considering in this project. Within this part of the handbook we will thus be focusing on the following:

Literature and other reference materials: We will for each of the species in focus go through the literature and other relevant materials, such as ECDC/EFSA reports, in order to show where the current state of the art is for the species in question.

Typing methods: Generally, there are three types of methods used for typing: some version of a Multi Locus Sequence Typing (MLST) scheme, various ways of using SNPs for classification, either absolute in relation to a reference, or relative to a set of genomes, and also k-mer based typing schemes, which aim to classify the genome based on k-mer composition. These methods will be linked to or described in component 2. Within this section, we will for each species in question specify what bioinformatic typing methods are commonly used for that species, as gathered from literature, other reference materials and from the community. An especially important consideration is to what extent these typing methods yield results that are backwards compatible in some form. For instance, a lot of information has been gathered about specific sequence types (STs) for these species, thus it would be useful if historical information can be connected to results produced with novel methods.

Outbreak detection/investigation methods: Outbreak detection and investigation is of particular importance for the foodborne pathogens in question. We will here link to or describe current considerations regarding how genomes can be characterized for use for outbreak detection and investigation, as well as any considerations that have to be taken into account when using in silico characterization for such purposes.

Species specific feature detection: Several of the species in question have specific features that are of importance for evaluating their pathogenicity level. For instance, for Shiga toxin producing *E. coli*, it is important to establish which shiga toxin genes are present. We will here describe such features, and link to or describe methods that can be used to detect such features. Of particular importance here is to describe to what extent such in silico methods can replace phenotypic characterization.

External data sources  
Data from online databases is in many cases used as input for several of the methods described in the previous sections. Such resources can be databases containing various MLST type schemes (7-gene schemes as well as core and whole genome schemes), and databases containing antimicrobial resistance genes or virulence genes. Within this section of the handbook we will link to or describe which resources are available, and to the extent possible, the differences between them and to what extent using one source as compared to another may have an effect on the outcome of the results.

**- Component 4 - NGS for surveillance -**

The ultimate goal is to use the data produced through the three components described above for surveillance, and in this context, for One Health surveillance as discussed in the beginning of the report. However, the focus for this part of the handbook will not be on providing implementation options, but more focused on what considerations have to be taken into account when planning a One Health surveillance program based on NGS data. The main reason for this is that the local and national situation regarding issues such as the prevalence of a bacteria, what subtypes that are seen, and how often outbreak events occur will differ from country to country. Thus, for this part of the handbook we will rather aim to describe some examples of procedures that are in use for using NGS data and the integration of this data with epidemiological data, and to give some guidance regarding what factors have to be taken into consideration for developing such procedures.

**Using NGS data for cluster definition and outbreak investigations**

One of the main uses of NGS data in our context is for outbreak detection and investigation. How to translate genomic data into meaningful information for public health decision-making is still not a settled issue. The phylogeny derived from genomic data can be used to infer likely transmission events. A first step is typically to assign cases to clusters; for infectious diseases, a cluster is a group of closely related isolates derived from cases that is usually interpreted as resulting from transmission of infection from a common source of contamination. The simplest way to determine sequence relatedness is to count the number of single nucleotide polymorphisms (SNPs) or alleles that differ between two sequences. The SNP or allele threshold approach places two individuals in the same putative transmission cluster if there are fewer than a threshold number of SNPs or allelic differences between their sequenced pathogen genomes. However, there is little agreement in the literature as to what such a threshold should be.

By itself, the number of SNP or allelic differences between genomes may not directly imply a probability of recent transmission, although probability of a common source. The Appropriate cut-off for inferring a common source cluster has to take into account many sources of uncertainties, including nucleotide mutation rates of the pathogen, which may vary at different stages of an infection, and subject to the effects of selection pressure. Furthermore, mutation rates are not constant over time and stable in different environmental conditions and hosts.

Once a cluster of related isolates from a probable common source has been identified, the process of identifying a potential outbreak cluster that would require an investigation must be determined. On a national level, this has to take into account the baseline of occurrence of that particular pathogen, whether sporadic or endemic. What are the known local reservoirs, what is the time between identification of cases, and the severity of the disease. Once such background information has been established, it can be taken into account when interpreting the results. Consequently, epidemiological links between isolates in terms of time and space is required to guide the final identification and investigation of an outbreak cluster.

**References**

[1. Stärk, K. D. C. *et al.* One Health surveillance - More than a buzz word? *Prev. Vet. Med.* **120,** 124–130 (2015).](http://f1000.com/work/bibliography/522820)

[2. *European Centre for Disease Prevention and Control. Expert opinion on whole genome sequencing for public health surveillance*. (ECDC, 2016).](http://f1000.com/work/bibliography/6491259)

[3. Lynch, T., Petkau, A., Knox, N., Graham, M. & Van Domselaar, G. A primer on infectious disease bacterial genomics. *Clin. Microbiol. Rev.* **29,** 881–913 (2016).](http://f1000.com/work/bibliography/3963817)

[4. Carriço, J. A., Rossi, M., Moran-Gilad, J., Van Domselaar, G. & Ramirez, M. A primer on microbial bioinformatics for nonbioinformaticians. *Clin. Microbiol. Infect.* **24,** 342–349 (2018).](http://f1000.com/work/bibliography/5641017)

[5. Balloux, F. *et al.* From theory to practice: translating whole-genome sequencing (WGS) into the clinic. *Trends Microbiol.* **26,** 1035–1048 (2018).](http://f1000.com/work/bibliography/5717913)

[6. Sekse, C. *et al.* High throughput sequencing for detection of foodborne pathogens. *Front. Microbiol.* **8,** 2029 (2017).](http://f1000.com/work/bibliography/4695443)

Appendix I: NGS Survey Questions

**Questionnaire**

Purpose of survey: collect information on the availability and implementation status of Next Generation Sequencing (NGS) methods for the purposes of routine surveillance and outbreak detection for pathogens within Europe. Please note, a responder might need to request information from several people at their institution to complete the questionnaire.

Estimated time to fill in the questionnaire when all information is available is 60 - 90 minutes.

By using the button op top of this survey: “Resume Later”, you can save the questionnaire and return at a later time point to continue answering the questions. The saved questionnaire can be retrieved using the button: “Load unfinished survey”.

**Questionnaire Participant details:**

Institution name: \_\_\_\_\_\_\_\_

Country:\_\_\_\_\_\_\_\_

Due to the newly implemented EU regulations regarding privacy, we are not able to ask for your email address through this form. However, if you are willing to discuss the issues brought up in this survey further with the ORION project, we ask you to send an email to karin.lagesen@vetinst.no.

**About your institution:**

Focus for your institution (tick all that apply):

* Food and feed
* Human/public health
* Veterinary and animal health
* Environmental
* Governmental
* Academic institution
* Other, please specify : \_\_\_\_\_\_\_\_\_\_

Size of institution:

* Upto 500 employees
* 500 - 1000 employees
* 1000- 2000 employees
* More than 2000 employees

Are Next Generation Sequencing (NGS) methods being used to analyse pathogens at your institution?

* Yes (if so, go to **General questions**)
* No (if so, go to **Questions on possible hindrances**)

**General questions:**

Comment: these questions are focused on what is being done at your institution. Questions later in the survey will ask whether such analyses are being done in-house or not.

What is the goal of NGS analysis of pathogens at your institute? (multiple)

* Case investigation
* Surveillance
* Outbreak response
* Analysis of representative strains, and if so:

What is the selection criterion? Please specify:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* Only for research purposes
* Other: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Which NGS methods are used in your institute for the analysis of pathogens? (multiple)

* Whole Genome Sequencing (WGS), please specify sample types: \_\_\_\_\_\_\_\_\_\_\_\_\_
* Amplicon sequencing (16S rRNA / other genes), please specify sample types: \_\_\_\_\_\_\_\_\_
* Other methods? If so, specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Are other molecular typing methods used in parallel with NGS analysis of microbial organisms?

* Yes, for all strains. If so, please describe methods below
* Yes, for a subset of strains. If so, please describe methods below
* No
* I do not know.

Please add your comment here:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Which pathogens are analyzed using NGS at your institution? (multiple)

* Salmonella
* E.coli, if so, specify pathotypes: \_\_\_\_\_\_\_\_\_\_\_
* Listeria monocytogenes
* Campylobacter
* Hepatitis A/E virus
* Other bacteria, if so, please specify species: \_\_\_\_\_\_\_\_\_\_\_
* Other viruses, if so, please specify species: \_\_\_\_\_\_\_\_\_\_\_

Our institution uses NGS methods to analyze (multiple) :

* Antimicrobial resistance
* Virulence
* Serotyping

Which kinds of categories of sources do your samples originate from? Examples could be animals, environmental swabs, food, fecal samples, host body fluids, etc.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Questions on sample storage and laboratory methods**

This section should be answered provided your institution does laboratory work for the purposes of NGS analyses of microbes.

Approximately how many people are involved in the laboratory work relating to NGS analyses?

* 1-4
* 5-14
* 15-24
* 25+

Approximately how many samples are processed in your institution for the purposes of NGS analyses each year? \_\_number\_\_

Do you have a computer information system (i.e. lab management software) at your institution containing information on samples?

* Yes, using a commercial solution, please describe this below:
* Yes, using an in-house created solution, pleae describe this below:
* No

Please add your comment here:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Is cultivation performed on pathogen samples after arrival at your institution to generate enough material for DNA/RNA extraction and/or -80 storage?

* Yes
* No
* I do not know

Are isolates of microbial cultures stored at -80 C after arrival at your institution?

* We store all isolates. Please indicate for how long below:
* We store representative isolates. Please indicate for how long below:
* No
* I do not know

Please add your comment here:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Which biological molecules are extracted from the samples? (multiple)

* DNA
* RNA

Please specify where the DNA extraction is performed? (multiple options)

* In house
* Outsourced (commercial partner)
* In a different laboratory in your country (collaboration network)
* Other, please specify: \_\_\_\_\_\_

Please indicate where sequencing libraries are made? (multiple options)

* In house
* Outsourced (commercial partner)
* In a different laboratory in your country (collaboration network)
* Other, please specify: \_\_\_\_\_\_

Please indicate where the sequencing is performed? (multiple options)

* In house
* Outsourced (commercial partner)
* In a different laboratory in your country (collaboration network)
* Other, please specify: \_\_\_\_\_\_

Which sequencing platform do you use? (multiple options possible)

* Illumina series (HiSeq and other larger scale instruments)
* Illumina series (MiSeq and other smaller scale instruments)
* Ion Torrent / Ion Proton Torrent
* Pacific Biosciences
* Nanopore minION / GridION / Promethion
* Other, please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Do you store the raw unprocessed NGS data? (multiple options possible)

* No
* Yes, in house on local file storage in our institution.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Yes, in a public archive (e.g. NCBI-SRA, ENA). \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Yes, outsourced in the cloud. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

There might be some human DNA reads in the raw NGS data stemming from the lab or other people that handled the samples. Are human contaminating reads removed before archiving?

* Yes
* No
* I do not know

**Questions on bioinformatic analysis:**

This section should be answered provided your institution does any kind of bioinfomatics processing for the purposes of NGS analyses of microbes.

Approximately how many people are involved in the bioinformatics work relating to NGS analyses in your institution?

* 1-4
* 5-14
* 15-24
* 25+

Approximately how many samples are processed using bioinformatics methods for the purposes of NGS analyses at your institution each year? \_\_number\_\_

Do you perform quality control of the NGS sequence data:

* Yes
* No
* I do not know

What software / method do you use for quality control (e.g. adapter trimming, low quality base detection etc) of your raw NGS sequences. Please specify:

* Commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Non-commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* In house created pipeline consisting of: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Online methods in the cloud: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Other, please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Not included in our analyses.

If contaminants (e.g. reads from PhiX / human / animal host ) are removed, what software / methods do you use to remove those reads? Please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What methods / software do you use to perform genome assembly using WGS sequences?

* Commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Non-commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* In house created pipeline consisting of: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Online methods in the cloud: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Other, please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Not included in our analyses.

Which approach is used for typing of pathogens based on sequence data?  
 (multiple options possible, each square is a tickbox)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type/Organism | Salmonella | E. coli | L. monocytogenes | Campylobacter | Hepatitis A/E |
| MLST |  |  |  |  |  |
| wgMLST |  |  |  |  |  |
| cgMLST |  |  |  |  |  |
| SNP analysis |  |  |  |  |  |
| Serotyping |  |  |  |  |  |
| Antimicrobial resistance genes |  |  |  |  |  |
| Virulence genes |  |  |  |  |  |
| Toxin genes (e.g. shiga toxin, other toxins) |  |  |  |  |  |
| Genotyping |  |  |  |  |  |

What methods / software are used for typing of microbial organisms?

* Commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Non-commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* In house created pipeline consisting of: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Online methods in the cloud: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Other, please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Not included in our analyses.

What methods / software are used for the analysis of:

* Antimicrobial resistance, please specify: \_\_\_\_\_\_\_\_\_
* Virulence, please specify: \_\_\_\_\_\_\_\_\_
* Serotyping, please specify: \_\_\_\_\_\_\_\_\_
* Not included in our analyses.

What methods / software are used to perform cluster analysis using typing results:

* Commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Non-commercial software package, namely: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* In house created pipeline consisting of: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Online methods in the cloud: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Other, please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Not included in our analyses.

Is microbial typing data linked to information concerning the source of the sample (patient, animal, other sample source) in your bioinformatic analysis workflow?

* Yes, if so, please describe at which step below
* No
* I don’t know

Please add your comment here:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What source information is included in the bioinformatic workflow? (multiple options)

* Species
* Age
* Sex
* Date of onset of symptoms
* Sampling date
* Travel of origin source (patient, animal, other sample source)
* Location of origin source (patient, animal, other sample source)
* Location of where sampling was performed (hospital, slaughterhouse, farm, etc)
* Other data, please specify below:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Questions on sequence data storage**

How do you store / archive results after bioinformatic analysis? Please specify:

* Public database (e.g. NCBI, PubMLST, or other):\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* In house database: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Online in the cloud: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Questions on possible hurdles**

Have you in the past considered the implementation of NGS for routine surveillance of pathogens?

* Yes
* No
* I do not know

What would be the main factors that has or could stall the implementation of NGS for routine surveillance of pathogens at your institution (multiple options possible)?

* Financial requirements
* Lack of sequencing equipment
* Lack of laboratory expertise (personnel)
* Lack of bioinformatic expertise (personnel)
* Lack of other personnel
* Lack of collaboration between relevant groups
* Lack of computational infrastructure
* Internal management decisions?
* Legal issues
* Regulatory issues
* Being done by collaborating institution
* WGS is not mature enough
* WGS is not harmonized enough
* Other, please specify:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_free text\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Are you currently considering the implementation of NGS for routine surveillance of pathogens at your institution? If so, when do you think NGS will be used for routine surveillance at your institution:

* We are currently not considering it.
* It will be implemented within 2019
* It will be implemented within 2020
* It will be implemented within 2022
* It will be implemented within 2025
* It will be implemented, but we don’t have a fixed time frame

What factors are important for your institution to decide if and when to implement WGS for routine surveillance of pathogens?

Please specify: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_