Deliverable

D4.3 Analytical workflow for epidemiological handling of and response to food-borne outbreaks in Europe

Version: 1

Due: Month 48

Completed: Month 48

Contents

Deliverable description

This document has been compiled to describe how best to use Next Generation Sequencing (NGS), including Whole Genome Sequencing and amplicon based NGS, in the event of a foodborne disease outbreak in Europe. Whole genome sequencing provides a powerful discriminatory tool that can help identify which cases of disease are part of a defined outbreak linked to a foodborne source. This analytical workflow will suggest the ways Whole Genome Sequencing can be used for early detection of an outbreak, characterization and presentation of an outbreak and its epidemiological data. However, at present, Whole Genome Sequencing is not routinely used to analyze samples and so the selection of isolates for sequencing may bias results and potentially misinform the response to an outbreak. This document details how best to currently apply sequencing to outbreak detection and management and what may be possible in the future.

When working on the two deliverables, D4.3 and D7.4, the COMPARE working group on foodborne pathogens (WP4 and WP7) found that a more comprensive output would be obtained if the epidemiological handling and response to outbreaks (D4.3) is described together with the microbiological methods and criteria for cluster definition. Therefore, the sections of relevance mostly for D7.4 (Improved guidelines for interpretation criteria for defining clusters of disease and linking of isolates from various sources and reservoirs) is incorporated into this report on the “Analytical workflow for epidemiological handling of and response to food-borne outbreaks in Europe”.

1. Early detection of an outbreak

1.1 Use of cluster detection algorithms

Continuous typing of pathogens obtained from human infections is a very efficient and conclusive method of detecting outbreaks. Whole Genome Sequencing (WGS) is a superior typing technique with high-level discriminatory power, and usefulness at detecting genetic clusters of related strains rapidly. It is essential that the genome sequences are compared centrally for the region or country of surveillance, and preferably compared to other sectors as well (human samples compared to food or veterinary samples). Using WGS, cases belonging to an outbreak are identified from their degree of genetic similarity; thus the sensitivity of case detection will depend on the chosen cut-off. Typically two isolates with five or fewer SNP differences are classed as being from the same source of infection but above this cutoff as different strains.

A number of cluster detection methods were identified (Deliverable (D)4.2, section 3) that could be used effectively to detect outbreaks where isolates have been characterized by WGS. Two methods, in particular, have been shown to be of most value in outbreak detection and previous work has shown that known outbreaks were successfully detected using WGS data (D4.2 section 6, Case study 1.A). SaTScan™ (https://www.SaTScan.org) is free, easy-to-implement software that analyses spatio-temporal data. For a defined geographical region and study period, SaTScan scans over multiple start and end dates, and can detect both “alive clusters”, still in existence at the end date, and “historic clusters” that ceased to exist before the end of the study period, depending on the type of analysis specified. The Farrington method is a log-linear model that can adjust for over-dispersion, seasonality, secular trends and past outbreaks. The model calculates an expected value for the current time period based on historical data and a threshold above which an observed count is declared to be unusual. These two methods are used by a number of European surveillance teams, including Public Health England (PHE) and Robert Koch Institute (RKI), to run regular periodic analysis to provide early detection of an outbreak by raising a warning that the identified cluster requires further investigation.

In general, health agencies use simple outbreak definitions for early detection of outbreaks (see Section 1.2), which improves the timeliness of detection but may mean that ‘false-positive’ warnings are raised that are not linked to an actual epidemiological outbreak. The case studies included below and in D4.2 have shown how these methods could be used and adapted to different WGS outputs. In particular, WGS is typically without a stable historical dataset if SNP-based analysis is used, which can reduce the sensitivity of the Farrington algorithm. However, if other typing schemes are inferred from WGS then historical baselines are possible. The case studies have informed that testing the algorithm at a number of different time spans (e.g. 5 and 12 years of historical data) could be effective in allowing comparison between the outputs from analyses of longer timeframe datasets, with greater statistical power but potentially less stability in the number and type of isolates included, against shorter timeframe datasets. The results of the studies suggest that currently, in surveillance systems where WGS is not used as standard, it can be used to detect outbreaks but provides little advantage over the classical serotype or phagetype classifications unless outbreaks are geographically localized. However, exceedance methods (i.e. Farrington algorithm) lack the sensitivity to detect small increases in cases where strains are very common; thus WGS may narrow data resolution, making it easier to pick up an exceedance in a particular group of strains. As a routine tool for outbreak detection, SaTScan is likely to be too complex to implement and interpret compared to the Farrington but could be invaluable when looking at strain data to identify localized outbreaks. As the number of surveillance isolates that are sequenced increases and the

depth of historical WGS data increases, these methods could become more powerful at detecting outbreak clusters for early detection. However, it should be noted that descriptive analyses of surveillance data are also key to early identification of outbreaks, especially in fields where data are sparse and unlikely to be suitable for statistical analysis. The routine use of figures depicting the number of cases or cumulative cases over time and maps can be helpful in providing early evidence of spatial or temporal clusters.

Case study 1.A: Application of data derived from Whole Genome Sequencing to detection of a *Salmonella* Typhimurium DT104 outbreak in Britain

An outbreak of *Salmonella* Typhimurium DT104 in livestock and horses in a region of Britain, linked to an outbreak in people (unpublished data), provided an opportunity to compare WGS to existing methods of pathogen classification using sero- and phage-typing. A cluster of cases of DT104 in sheep and cattle was first identified in autumn 2016 and linked to an outbreak of DT104 in people with a specific SNP address (see Section 2.2 for more detail on the SNP address system). Subsequent enhanced surveillance and retrospective testing, including WGS and SNP address assignments, identified 42 STM DT104 cases from 36 separate incidents in livestock and horses between 2016 and 2018 linked to the human outbreak, but a check for related strains identified earlier livestock cases dating from 2014. An incident is defined as all isolations of the same serovars, or serovars and phagetype combination, of a particular *Salmonella* from an animal, group of animals or their environment on a holding. APHA sequenced and assigned SNP addresses to 323 *S*. Typhimurium isolates collected between 1992 and 2017. These were primarily DT104 complex but also included a few DT120 and 193 and a number of additional isolates sequenced in response to this outbreak. Two methods of outbreak detection were tested by creating sequential datasets that simulated the available outbreak data over time.

For spatio-temporal analysis, SaTScan was used and 15 data sets were created, one for each quarter (three months) starting with quarter 2 of 2014, when the first known outbreak case in animals occurred, until the end of 2017. Each dataset contained details of all existing cases and controls up to the end of that quarter. Prospective analyses was run on each dataset using a Bernoulli model, where cases and controls are defined and the method searches for areas where there is higher than expected risk of being a case. Using WGS, and defining cases as those with five or fewer SNP differences from the outbreak strain address, the models first detected a cluster at the end of Quarter 3 2015 after just three cases. This also occurred when controls were defined both as either all other sequenced isolates or all other submissions of STM DT104 taken from clinically diseased animals. In comparison, when cases were defined simply by phagetyping as STM DT104 from clinically diseased animals of any species, and controls as any other phagetype, the first date that a cluster was identified relating to the outbreak was the end of Quarter 2 2017. Although an earlier cluster in a separate geographical area was detected, isolates which had been sequenced from these incidents were more than 25 SNPs different from the outbreak strain, and thus thought to be unrelated. When the same analysis was run on clinical isolates only from cattle or sheep, no clusters were detected at all.

The Farrington algorithm was run in R using the farringtonFlexible function, with 5 and 12 years of historical data, and a half window of one month. A number of outcomes were investigated: outbreak strain defined using the SNP address was compared to a threshold exceedance of DT104 in any species; in cattle or sheep; only in cattle and only in sheep. When models were run on DT104 cases, a trend was included, as there was a clear downward trend in cases over the study period. No trend was included when only the sequenced isolates were used. A 2/3 power transformation was included to adjust for overdispersion where there were low counts of

cases; and the default threshold of no alarm if there were fewer than five cases in four weeks was included to avoid false positives with the sparse data. Using SNP address to define outbreak strain, the earliest a potential outbreak was flagged was September 2016 using both the 5 and 12 year models. The same alert was raised when using just DT104 in all species and a month later in the model using only cattle and sheep when 12 years of historical data were used. The 5 year models for DT104 had similar results regardless of whether all species or only cattle and sheep were used, raising alerts towards the end of 2014 and one in November 2015. Models run using only sheep or only cattle STM DT104 did not perform so successfully, with the first alerts not until December 2016 and April 2017, respectively. Using the SNP addresses, the model was also used to explore different levels of genetic similarity at which strains were defined as part of the outbreak. The model was run where only strains with 5 or fewer SNP differences were classified as outbreak cases; then progressively relaxed to allow isolates with up to 10, 25, 50, 100 and finally 250 SNP differences to be cases. The model detected a threshold exceedance first in the SNP25 and SNP50 clusters in August 2016, which was before that detected by the SNP5 and SNP10 clusters in September 2016. The SNP100 and 250 were almost identical to the 50.

WGS generates a huge amount of data that is unsuited to current epidemiological analysis methods and difficult to interpret in its raw form, but the SNP address provides a manageable output to describe the relatedness of strains. Two methods of outbreak detection have been explored here using the SNP address to define the outbreak strain of *Salmonella* Typhimurium DT104; other outputs such as core genome multilocus sequence typing could equally be used the same way to run the outbreak detection models. SaTScan quickly detected the outbreak strain defined by SNP address in the third quarter of 2015 after only three cases, which compared favorably to the cluster detection using DT104 against all other phagetypes, where the first cluster identified was much more geographically widespread and more than 18 months later. SaTScan showed very few other clusters in the data, suggesting good specificity. However, this outbreak was very geographically localized and outbreaks on a larger geographical scale may not have been detected so quickly using SaTScan.

The Farrington model provided a robust, easy-to-implement method of detecting a higher than expected number of cases and was easy to adapt to SNP address data. In these analyses, the Farrington algorithm performed well despite the lack of a good historical dataset and detected an exceedance after just a few cases towards the end of 2016. However, the exceedance was detected regardless of whether the DT104 phagetype or outbreak SNP address was modelled, indicating that with this dataset the WGS data did not provide any benefit to early detection.

In this study, the outbreak strain had already been identified, so epidemiological investigations had inflated the number of outbreak cases above what would probably occur in the early stages of an outbreak detected by routine surveillance. Identifying new strains to look for requires much time and effort, but is fundamental to the successful early detection of outbreaks using this method. The sequenced isolates in this study are not representative of *Salmonella* surveillance data as a whole; over-representing some animal species and time periods. Additionally, analyzing retrospective data does not realistically simulate an outbreak in real-time. However, this work usefully illustrates how WGS and SNP address data could be incorporated into surveillance and outbreak detection in future years, should sequencing become routine and methods standardized between institutions.

1.2 Use of simple case definitions

There are many different types of methods for analysis of WGS data which give different granularity and sensitivity related to defining cases as part of an outbreak. The WGS methods used to characterize isolates are described in Section 2, along with the utility of each type and suitability of each type for specific contexts (e.g. viral versus bacterial outbreaks). For the early detection of an outbreak, simple case definitions are suitable as the outbreak is unlikely to have been effectively characterized yet and it is preferable for any surveillance scheme to optimize sensitivity over specificity in order to improve timeliness of detection. In case study 1.A, it was still acceptable to use serotyping and phagetyping definitions to rapidly identify the outbreak, without the need for further WGS characterization. However, if the outbreaks were more clustered in space or were related to more common 'types' then WGS definitions would have been important to provide a case definition for early detection. This finding has been found to be true for both bacterial and virological pathogens.

To utilize broader case definitions for early detection of an outbreak, a greater number of SNP differences than typically used to define an outbreak could be used to ensure that all possible epidemiologically linked cases are likely to be included. For example, for SNP address outputs, it would be easy to adapt cluster detection methods so as to enable routine analysis, but at a higher resolution as detailed in case study 1.A. Some further work would be needed in order to decide on the groups (e.g. 50, 25, 10 SNP differences apart) on which to run any models and this may need a better understanding of the background population of strains. Some of the SNP cluster groups only have a few isolates in them and it may be appropriate to classify strains according to SNP level based on the frequency of strains within the groups.

1.3 Virological investigations

The outbreak examples detailed above have focused on bacteriological outbreaks. However, the epidemiology of viral foodborne infections is different from bacterial infections. In contrast to foodborne bacteria, for many foodborne viruses, such as hepatitis A virus (HAV), hepatitis E and norovirus (the pilot organism included in the COMPARE project), zoonotic transmission has not been observed or is not a major transmission route. Their main mode of transmission is person to person. Food is merely involved as a vehicle and is directly or indirectly contaminated by an infected person. Either directly by a foodhandler, or indirectly, such as when foodcrops are irrigated by sewage contaminated water. The type of food involved is very different from bacterial foodborne outbreaks, ranging from any restaurant or catered menu to berries and oysters. It should also be noted that, unlike bacteria, viruses are not able to replicate in these foods.

Hepatitis A is a reportable disease in many EU countries and fecal or serum samples of most suspected cases reach the lab, thus creating a large potential for outbreak detection based on typing data. Especially in diffuse outbreaks, where food which has been dispersed over a large area or a large time period is involved (e.g. frozen products) molecular data are necessary in order to detect the outbreak.

Norovirus gastroenteritis is not reportable and foodborne outbreaks are often very large, so here the detection will generally take place based on epidemiological clustering of human cases. Usually, typing will only be performed in a later stage when a foodborne transmission is suspected, and only on a small subset of the cases and may serve to confirm the role of an epidemiologically implicated food item.

In virology, partial sequence data has already been in use for decades to identify clusters of related human cases. Viral genomes are relatively small, 7000-8000 base pairs for HAV and norovirus, and until recently specified regions of the genome were sequenced for typing and clustering. The transition to complete genomes, made possible by NGS will enlarge the discriminative power of the sequence analysis. Another advantage of NGS over the classical sequencing of PCR fragments is the possibility to identify different strains in one sample, as occurs regularly especially in norovirus outbreaks, where sewage contamination is involved. A third advantage is that viral sequences will be found in samples where the viral RNA is degraded, where PCR remains negative due to fragmentation of the genome. In case of a non-targeted NGS approach, a fourth advantage is the detection of other pathogens in a sample.

2. Outbreak definition

2.1 Definition of clusters

Surveillance programs, including systematic isolate collection and typing, have been established to detect clusters of microbiologically related cases, identify common sources of infection, and take appropriate control measures to reduce human illness and economic losses. Before WGS was implemented in routine surveillance, a multitude of pathogen specific phenotyping and less sophisticated genotypic methods were used (D7.2). Many of these methods can also be performed *in silico* using WGS, meaning that comparable results can be extracted from the genome sequences. Therefore, it is still possible to get data such as, serotypes, multi-locus sequence types and antimicrobial resistance data from the strains. These classic markers from phenotypes are still used in legislation especially in the veterinary sector, so it is important to have a very high correlation between classic phenotypes and the *in silico* derived types.

Cluster detection based on WGS provides a high resolution to detecting clusters of genetically related strains. Genetic clusters are clusters of closely related genome sequences and these can be detected in several ways. Genetic relatedness are often determined by either single nucleotide polymorphism (SNP) analysis or by gene-by-gene analysis (e.g. core genome multilocus sequence typing (cgMLST)). Both methods are described in more detail in D7.2. In short, the SNP analysis compares single nucleotide sites between all shared positions in a collection of genomes, whereas the gene-by-gene methods compares the sequences of a defined set of loci and assigns allele numbers to each loci. Regardless of which methods are used to define a genetic cluster, further investigations are needed to assess whether the patients/animals which the isolates originate from are likely to be epidemiologically linked and therefore constitute a possible outbreak of disease.

Therefore, a WGS-based cluster can be defined as two or more cases for which the whole genome sequences of the isolates are within a defined SNP distance from each other, or minimum of two isolates with the same cgMLST extracted from WGS. Both SNP-calling and gene-by-gene comparison have been accepted as valid outbreak investigation approaches and different initiatives have been launced by EFSA/ECDC intended to harmonise the WGS based methods.

In order to define clusters, a threshold must be set to determine the level of the genetic similarity between isolates at which they can be considered as an outbreak with a common source. This can then be characterised by the number of SNP differences or the similarity of the core genomes in conjunction with known epidemiological and geographical links.

2.2 SNP characterization for outbreak detection

The SNP based method for outbreak investigations adopted by the PHE (Ashton et al., 2015) was described in detail in D7.2. In this approach each isolate is first characterised by *in silico* Multi Locus Sequence Typing (MLST) from the WGS data to assign a profile or sequence type (ST) based on the 7-core MLST scheme. Sequence types generally cluster together into discrete, related groups called eBurst groups (eBGs) for *Salmonella* spp. (Achtman et al., 2012) or clonal complexes (CC) for *Escherichia coli*, *Shigella* spp. and *Listeria monocytogenes*.

Next, the core genome of the isolate is compared to an appropriate reference genome within the same eBG or CC to identify SNPs between the isolate and the reference genome. The 'genetic distance' (number of SNP differences) is calculated between all pairs of isolates within a specific eBG or CC. Hierarchical grouping of isolates into clusters based on increasing levels of similarity is performed by calculating the pairwise SNP distance for each pair of isolates in the analysis set and performing a single linkage clustering at 250, 100, 50, 25, 10, 5 and 0 SNPs from the deduced distance matrix. This approach is based on generating a large number of sequenced genomes to facilitate the development of a databases of variants to perform single linkage clustering.

Animal and Plant Health Agency (APHA) UK is taking the same approach in defining the SNP address of sequenced isolates of animal origin as typing scheme. This approach of assigning a SNP address to characterise animal isolates and describe their genetic relatedness to human isolates within the same cluster has been applied to a number of foodborne *Salmonella* spp outbreak investigations in the UK. In one of the investigated outbreaks, a cluster of cases of *Salmonella* Typhimurium DT104 in sheep and cattle that was first identified in autumn 2016 by WGS was linked to an outbreak in people using SNP address. The SNP address from this outbreak investigation was used for the development of the cluster detection methods described in the Case study 1.A.

2.2.1 SNP address

A single linkage clustering at 250, 100, 50, 25, 10, 5 and 0 SNPs performed by calculating the pairwise SNP distance for each pair of isolates in the analyzed set assigns a seven number SNP address representing case clustering at seven different levels of genetic relatedness, beginning with the least related at 250 SNP differences and ending with isolates that are genetically identical with 0 SNPs.

Isolates that fall into the same 5-SNP cluster will have no more than 5 SNPs that are different from any other given isolate in that cluster (i.e. the 'pairwise distance' from one isolate to at least one other isolate in the cluster is less than 5 SNPs). Case isolates that fall within a 0-SNP cluster have the strongest genetic evidence for being part of an outbreak with a common exposure or with direct person-to-person transmission over a short period of time. The numbers chosen for each level of the SNP address however, is not indicative of the relatedness of the isolates within and between cluster and therefore isolates within a 10-SNP cluster may not be directly linked or be part of the same outbreak. Clusters at the 10-SNP level may be indicative of transmission over an extended period of time, or persistent environmental contamination over an extended period of time from the same primary source. Validation studies conducted by PHE indicate that setting the threshold for cluster detection at the 5-SNP level is optimal for *Salmonella* spp. (Waldram et al., 2018) and Shiga-toxin producing *Escherichia coli* (STEC) (Dallman et al., 2015) both in terms of specificity, defined as level of discrimination within which common exposures are likely to be identified in an analytical investigation, and sensitivity, defined as detection of all cases likely to have had the same exposure.

2.2.2 Defining an outbreak based on SNP address

In response to a perceived increase in the number of *S.* Typhimurium DT104 incidents by epidemiological investigations (Zoonoses Order), data was extracted from the *Salmonella* database to look at all incidents of *S.* Typhimurium DT104 complex (DT104, DT104b, DT12 and U302) between 2007 and 2016 in Great Britain. The

most recent date of extraction was 21/12/2016. In the same time perion, in January 2017 PHE UK detected a cluster of human *S*. Typhimurium DT104 isolates with the SNP addres 60.11.15.31.458.459.4030. The FastQ files of the sequenced animal isolates that matched the epidemiology of the identified human clusters were transferred to PHE for SNP typing of which 18 isolates matched and had the same SNP address as the human isolates (60.11.15.31.458.459.4030). Subsequent enhanced surveillance and retrospective testing identified further human and animal isolates with less than 5 SNP difference (idenified as t459 cluster).

APHA investigated a total of 323 *S*. Typhimurium isolates collected between 1992 and 2017 by sequencing and assigned SNP addresses at APHA to the 323 isolates. Older *S.* Typhimurium DT104 complex isolates from the previous epidemic in 1990s were included in the analyses. The SNP address assigned to the 323 sequenced DT104 isolates and the extracted clusters from the dataset were used to investigate epidemiological methods to define outbreak based on WGS data using SaTScan and the Farrington algorithm described in Section 1. The SNP address defined by the APHA Snapper DB for the human outbreak related isolates was 2.2.2.2.7.7.10. As the SNP address is based on pairwise comparison of isolates within the dataset, the same SNP address could be derived only if isolates of interest are compared within the same database.

2.3 Gene-by-gene approach

The cgMLST approach is based on analysis of all genes present in all isolates of a species. An allelic profile produced by cgMLST is composed to all other (hundreds to thousands of different alleles depending on the genome size of the investigated species) within the allele database. The wgMLST, or the 'pan-genome approach', uses the full complement of genes found in species, including the core genome and the variable or accessory genome that contains the genetic information including integration and excision of mobile elements such as prophages and plasmids that often harbor virulence determinants and antimicrobial genes that influence virulence properties specific to single strains of the species. Both methods were fully described in D7.2.

2.3.1. Gene-by-Gene Approach for Surveillance and Outbreak Investigations for *Salmonella* at the German National Reference Centre

The German National Reference Centre (NRC) for *Salmonella* and other bacterial enteric pathogens is certified for subtyping of *Salmonella* by serotyping, PFGE and phage-typing. Recently, the Centre phased out PFGE and phage-typing and intensified NGS-based subtyping methods. Since there is no certification process for NGS-based methods (bioinformatics part) official subtyping reports based on NGS are not yet generated. Nonetheless, NGS-based methods are extensively used for surveillance and outbreak investigations. There is no guideline for utilizing WGS-based approaches in public health, although there are a number of relevant peer-reviewed publications and expert opinions from appropriate organizations, e.g. the World Health Organization and ECDC (WHO "Whole genome sequencing for foodborne disease surveillance" Landscape paper 2018, ECDC Scientific Advice 2015). The 2018 WHO landscape describes WGS as having a higher resolution than other subtyping techniques and representing an "all-in-one test" since *in silico* information like serotype and resistance profile may be extracted, which usually require different testing methods. WHO recommends allele-based typing (also known as gene-by-gene approach) as the best option for a nomenclature, which can be effectively integrated into surveillance networks. The other alternative, the SNP-based approach should be reserved for situations, where an even higher resolution is required. The ECDC paper "Expert opinion on the introduction of

next-generation typing methods for food-and waterborne diseases in the EU and EEA" from 2015 on the other hand does not advocate *per se* for a gene-by-gene-based method but also views the "SNP-address" used by PHE as a feasible method for food and waterborne diseases. The ECDC recommends to continue using classical subtyping methods in parallel to WGS-based techniques in order to ensure backward compatibility and phenotypic testing for instances where *in silico* methods may not be reliable yet (e.g. antimicrobial resistance). The gene-by-gene approach has been proven successful for retrospective outbreak investigations of salmonellosis (Simon et al., 2018). Additionally, with Enterobase (http://enterobase.warwick.ac.uk) there is a comprehensive database for *Salmonella* sequence and complex types, which also provides curated MLST, cgMLST and wgMLST schemes. Therefore, NRC has recently decided to perform analyses based on this cgMLST scheme (implemented in the Ridom SeqSphere+ software tool (commercially available from the Ridom GmbH Münster, Germany).

Sample provision

ECDC recommends defining what kind of samples should be send to a corresponding national reference laboratory. In Germany, there is a hierarchical approach based on the requirements defined in the Infection Protection Act (Infektionsschutzgesetz, IfSG). Additionally, Germany is in the process of developing a network of laboratories to send isolates (already typed at least at the genus level) to the national Centre. The laboratory network covers the whole of Germany and represents an average of 15-20% of nationally registered salmonellosis cases. All isolates are subtyped by classical serotyping and susceptibility testing to selected antimicrobials (for epidemiological purposes only). For serotypes Enteritidis and Typhimurium phage-typing is still performed to pre-define potential outbreaks (or rule them out). A subset of samples, which predominantly includes strains from suspected outbreak clusters (all serovars), is additionally subjected to WGS. The NRC is currently implementing a comprehensive WGS-based surveillance for *S*. Enteritidis and plan to expand this to *S*. Typhimurium in the near future.

WGS data analysis with the SeqSphere software

The NRC uses the Ridom SeqSphere+ software for analyzing WGS data in the context of surveillance and outbreak investigations. The Achtman 7-locus MLST scheme as well as the Enterobase cgMLST scheme have been embedded by the manufacturer. Since the software includes a QC pipeline, the raw reads are directly fed into it without prior quality control or trimming. SeqSphere creates distance matrices, minimum spanning trees or neighbor joining trees. The Ridom tool creates so-called 'complex types' for cases of very close relationship between strains. The threshold (distance to the closest neighbor) is 7-loci for all serotypes. This needs to be discussed and might be serovar-dependent. This distance is not strict enough in terms of a point source outbreak, where (in our experience) only strains with a maximum pairwise distance of 5-loci are considered as outbreak strains. For prolonged or more diffuse outbreaks other criteria might be necessary. In any case, epidemiological data need to be included. However, since Ridom SeqSphere uses a different allele calling mechanism than EnteroBase and established an independent allele database, the results are not comparable with EnteroBase (or BioNumerics), which is one of the major drawbacks of the otherwise very comfortable tool. Generated sequence data and metadata is shared on a national level on a case-by-case basis. A database for sharing sequencing data and metadata is being discussed at the federal level although it may take many more years to be established. Information about clustering isolates are shared with the unit for epidemiology at the RKI and the relevant state and local health authorities to decide which further epidemiological, microbiological

or traceback analysis along the food chain are carried out. A first example of using the complex type in an investigation is with an outbreak with *Salmonella* Enteritidis complex type (CT) 1734 which has been found in different federal states in Germany as well as in Norway, France, Luxembourg and Scotland. The outbreak investigation is ongoing. A vehicle of infection has not been identified so far. This outbreak may be used in future to learn more about the suitability of this nomenclature for outbreak investigations.

HierCC, the hierarchical clustering algorithm of cgMLST of enterobase

This genotyping tool is proposed in EnteroBase (https://enterobase.warwick.ac.uk/) (Alikhan et al., 2018). In this approach each cgST is clustered with all other cgSTs that differ in pairwise fashion by up to a discrete defined number of loci (HC0, HC2, HC5, HC10, … HC100 …) (see https://enterobase.readthedocs.io/en/latest/features/clustering.html for further explanation). In that perspective, the classification principle is similar to the one developed and used by PHE/APHA in Snapper DB except that allele differences rather than SNP difference are used to cluster isolates. Importantly, the cluster numbering is permanent and will not change. Cluster IDs are assigned on a First Come-First Served basis so when adding new genomes, the number of the first cluster which fulfills the criteria is assigned to that genome. So these numbers can be exchanged between the different actors involved in investigation facilitating communication and can also be referred to in case of publications. Depending on the epidemiological situations, different clustering groups can be used. For example, it is reasonable to think that clustering below 5 allelic differences (HC5) is relevant for short term surveillance or outbreak investigation, while the stringency can be relaxed for long term surveillance. Thus, the novelty of the tool does not yet allow to define strict rules of use that will be built and refined over time and with users feedback.

2.4 Attempt to define common rules for the interpretation of WGS data in the context of foodborne investigations

The definition of methodology used for interpreting WGS data is a crucial issue to make scientific conclusions compatible with the regulatory framework. In a recent paper focused on SNP-based isolate discrimination (Pightling et al., 2018) it was proposed to combine a set of four criteria: SNP-threshold; phylogenetic tree topology; bootstrap support; and epidemiological information, before reaching conclusions on whether isolates arose from the same source (Table 1). A limit of the present guidelines is that the SNP thresholds were defined on the basis of a limited number (17) of foodborne outbreaks taken from the literature and a mix of several pathogens *E. coli*, *Listeria monocytogenes* and *Salmonella enterica*. The values will certainly be subjected to revision or refinement, but they are precious to investigators, which can rely on them to build their conclusions.

Table 1: Conditions used to determine when whole-genome sequence analyses can support the conclusion that two or more genomes arise from the same source (taken from Table 2, Pightling et al., 2018).

	Support	**Neutral**	**Does not support**
SNP-distance	<21	21-100	>100
Bootstrap support	>0.89	0.8-0.89	<0.8
Tree topology	monophyletic	paraphyletic	polyphyletic

It is certain that these rules will be refined as WGS-based investigative experiences increase. To this end, it would be desirable to set up databases dedicated to the deposit of WGS data on isolates from outbreaks with which complete epidemiological data would be associated. In all cases the authors insist that epidemiological and traceback evidences are of paramount importance to infer source and link isolates.

2.5 WGS viral foodborne outbreak detection

The use of WGS leads to increased resolution in outbreak detection in viral foodborne outbreaks. Now that more and more complete genomes become available, our research question is whether we are able to identify 1 transmission cycle using HAV and norovirus sequences. As for both viruses the route of transmission is fecal/ oral, and food is only a vehicle in foodborne infections, for this analysis there is no need to restrict it to foodborne outbreak samples. As there is a large difference in evolutionary rate between HAV and norovirus, the resulting discriminatory level by using complete genomes is not be the same.

In a large ongoing outbreak of HAV in the Netherlands among men having sex with men (MSM) in 2018, a small foodborne cluster (n=9) was suspected based on epidemiological criteria: the possible index worked in the food industry (Figure 1). Sequence analysis based on the standard typing region of the HAV could not confirm the cluster. Analysis based on WGS of 8 sequences: 3 of the possible foodborne outbreak (red), including the possible index (blue) and 5 sequence of the ongoing outbreak among MSM, revealed 5 unique SNP's of the foodborne outbreak strains compared to the others. The left tree is based on WGS sequences, the right tree is based on the standard typing region of HAV.

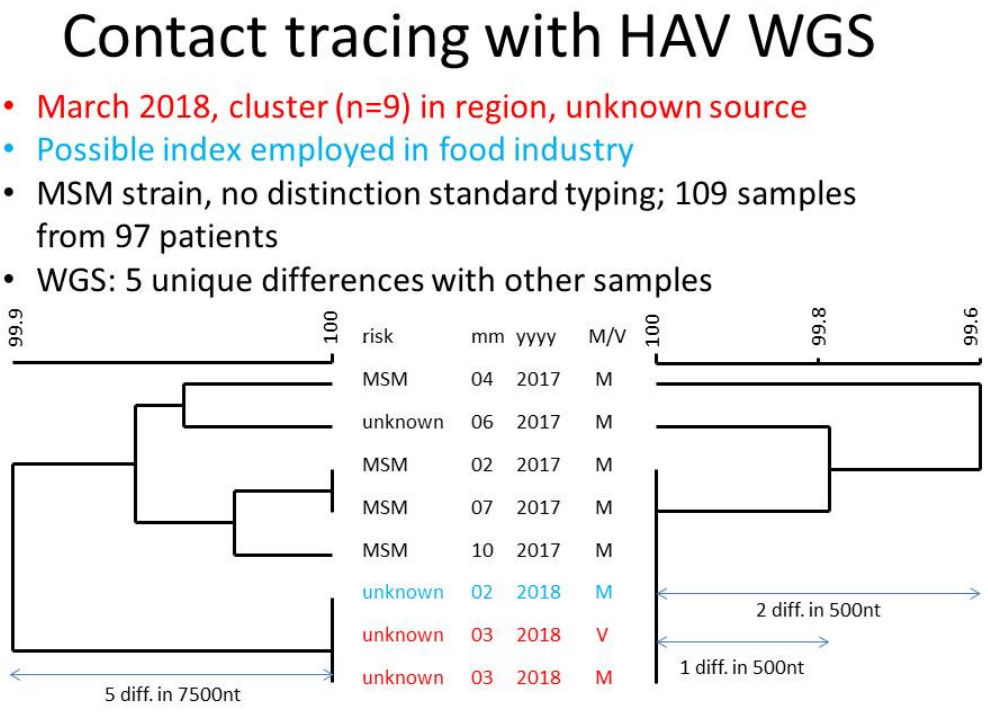

Figure 1: Example Hepatitis A: Contact tracing with HAV WGS.

3. Review of surveillance after initial outbreak characterization

3.1 Examine biases in sequenced population

After the initial detection of an outbreak and characterization of the form and epidemiology of the outbreak, a review is needed to infer what influence the selection of samples for WGS to generate the 'sequenced population' may have had on these outputs. Biases represent any deviation away from random selection, which form the basic assumption for most epidemiological analysis. Biases within the sequenced population could lead to an incorrect case definition, reducing the sensitivity of case detection and incorrectly informing ongoing surveillance. For example, if all sequenced isolates were from cattle and sheep samples, then this may misinform surveillance to prioritize detecting outbreak cases in these animal species, whereas outbreak cases may also be present and spreading in pigs and poultry. Even in instances where all isolates received at a laboratory are sequenced, an analysis of potential bias is helpful, as not all infected or diseased cases within a population present to doctors or vets, or have samples submitted to laboratories and so those sequenced will still only be a subset of the total case population and may be biased (Figure 2).

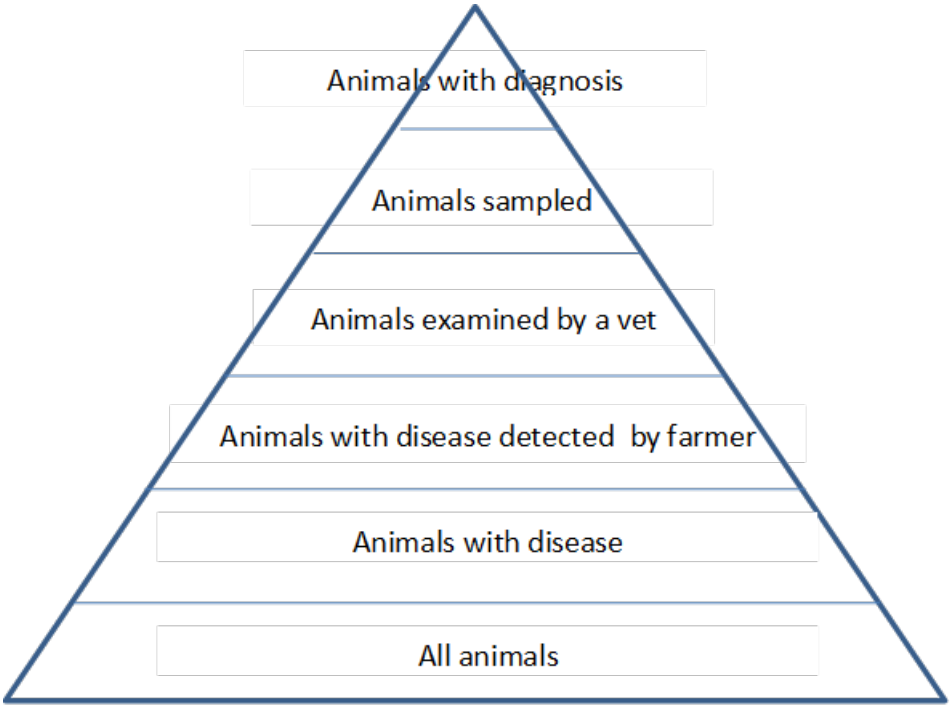

Figure 2: Example of the veterinary disease surveillance pyramid.

To analyse bias, the sequenced population should be compared against the characteristics of the wider population of interest (the background population) to identify how it differs e.g. compare the isolates selected for sequencing against all surveillance isolates presenting at the testing laboratory during a set time period. This comparison can be concluded statistically or descriptively against knowledge of the background population (e.g. farm type, animal species, house location). Additionally, information on how samples were selected for sequencing should be evaluated to inform whether this may have led to introducing bias into the selection, e.g.

isolates that present with more exotic antimicrobial profiles may have been selected for sequencing, which would reduce the population of sequenced isolates with more standard profiles. Understanding the biases within the population can allow greater interpretation of what is known and unknown about a potential outbreak, which can then inform better outbreak characterization.

3.2 Sampling strategies

Sampling is crucial for the pertinence/performance of the genomic analysis carried out. Several strategies of sampling are available and the analyst must adapt to the global objective they have. At least two different strategies can be envisaged. If the objective is to characterize the strains that contribute the more to consumer exposure, the sampling effort must be put on strains isolated in the food product from the main food companies (or issued from the main food producing regions). Stratified sampling, if commonly applied, is such a situation. If the objective is to explore the diversity of strains circulating in a country, the analyst can use the metadata associated to strain to reach that objective. When there is more than two categories of metadata describing the strains, the selection requires an algorithm of selection.

A three step process for selecting strains based on metadata information (e.g. region, type of animal) is originally proposed. This method relies on the Gower's coefficient (1971), which is a metric expressing a dissimilarity: the “distance” between two units is the sum of all the variable-specific distances (associated to metadata categories). GC metric is capable of combining numeric and categorical data. GC offers the opportunity to the analyst to select weights for each individual variable, effectively altering the importance of each metadata categories (region more important than year). A three step process is proposed:

- calculating dissimilarity matrix based on Gower distance,
- applying the clustering method on dissimilarity matrix with hierarchical clustering. (agglomerative (bottom-up approach of clustering),
- assessing clusters with the “Silhouette” method: the silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters.

The approach is illustrated in Figure 3 below. An Rscript (strain_metaselect.r) is available. It takes a csv file that includes strains ID and metadata information. It provides as output a csv file of selected strains.

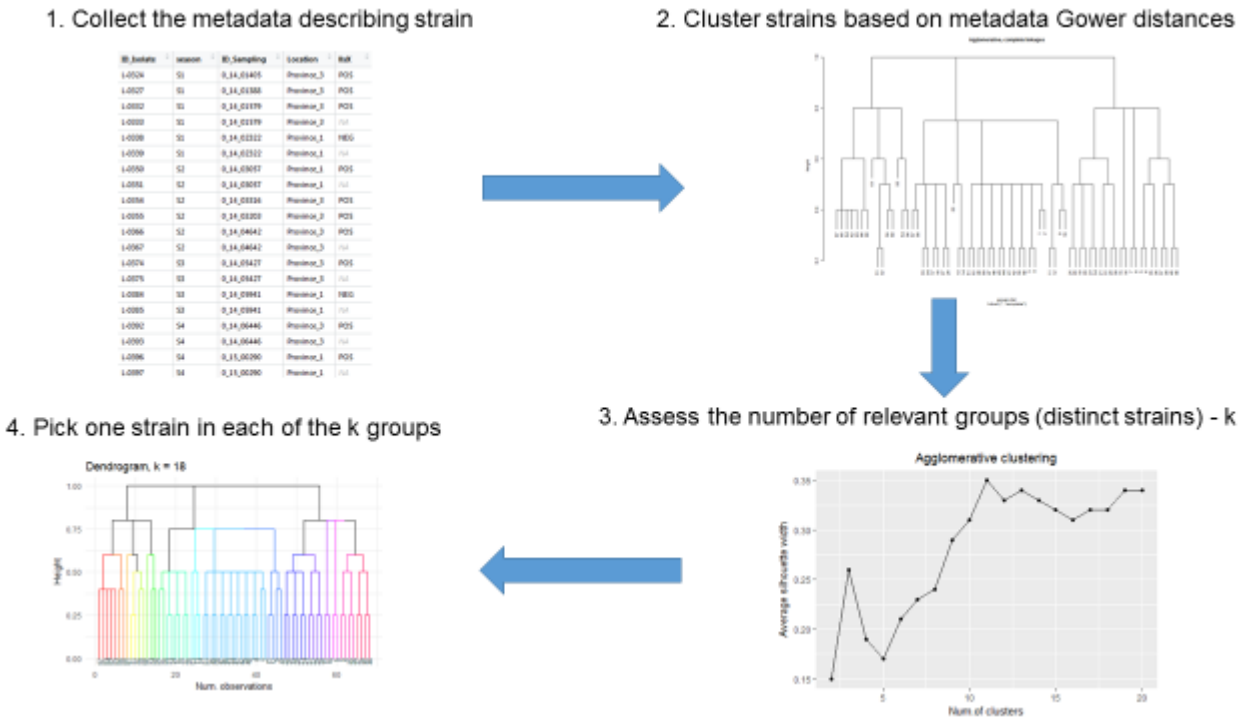

Figure 3: Method for selecting strains based on their metadata for diversity characterization.

3.3 Adjustment of surveillance schemes to improve case detection

After the initial detection of a potential outbreak, surveillance schemes are typically informed and enhanced so as to prioritize the ongoing detection of new outbreak cases. The efficiency of any enhanced surveillance is based upon a strong case definition and a good understanding of how much is known and unknown about the outbreak. Risk-based surveillance (RBS) design is used to improve surveillance efficiency (Schwermer, Reding and Hadorn, 2009). RBS optimizes cost-efficient case detection by placing the majority of surveillance resource in the areas of highest risk, such as geographical regions of greater risk of incursion or animals species more susceptible to infection. RBS design is informed by quantifiable knowledge of the difference in risk within subpopulations. A well designed investigation of an outbreak could greatly inform RBS to improve upon the detection of new cases, whereas a heavily biased evaluation of an outbreak could reduce the sensitivity of case detection.

3.4 Retrospective sequencing

On detection of an outbreak, it is typical to review historical isolates and identify further samples for sequencing to help identify additional cases and potentially identify earlier cases which may inform the original food source or incursion of the pathogen into the food source. Retrospective sequencing can also be used to alleviate known biases within the sequenced population by targeting sequencing in areas that were under sequenced or by using isolate selection that would bring the population closer to a randomized selection.

Typically only single isolates are sequenced from each incident (i.e. from each occurrence of positive cases at a single location). However, in a recent analysis of an outbreak there were 23 incidents with WGS data from 2 or

more isolates, providing an opportunity to look at within-farm strain diversity. In 18 of these incidents, SNP addresses isolates differed by five SNPs or fewer, with up to six isolates of the same strain being sequenced in one case. However, in five incidents, samples of the same phagetype but more than five SNP differences were collected. Whilst this dataset was small, it was interesting to note that isolates collected from the same incident but on different dates had more SNP differences than those collected on the same date. This finding, highlights the need for more work on understanding the diversity of strains within an incident to inform how many isolates should be required to be sequenced from each known incident, which could be followed up retrospectively to characterize the outbreak after the initial detection.

An essential aspect of outbreak detection based on molecular data is a complete set of reference sequences. This has been one of the aims of establishing the NoroNet and HAVnet databases. As these contain mainly partial genomes, in order to make them useful for WGS analyses, these need to be supplemented with complete genomes. NGS on norovirus and HAV is performed directly on clinical material, without a culturing step, which significantly decreases the sensitivity of the method compared to bacterial NGS on isolates. The first part of the COMPARE project has been dedicated to improving and fine tuning the laboratory procedures, which are now in a stage in which we can apply them on scarce material and samples with lower viral load. Thus, besides NGS to retrieve complete genomes of specific outbreak strains, a set of samples containing different genotypes and variants of both norovirus and HAV is also included in the NGS experiments.

4. Verification of outbreak and guidance on when to respond to an outbreak

It is essential that the genome sequences are compared centrally for the region or country of surveillance and preferable if the sequences can be compared to other sectors as well (human samples compared to food or veterinary samples). It is common for foodborne outbreaks to be widely distributed across regions, and or countries. WGS provides a unique opportunity to share sequence data to better identify these broadly distributed outbreaks.

The decision on whether to act on a genetic cluster should be taken in close consultation between microbiologists/ virologists and epidemiologists. The decision may be based on specific information, such as the number of affected patients, geography, date of disease onset, demography as well as the strain characteristics and knowledge on the previous occurrence of the strain. Further considerations on how to verify and investigate outbreaks described in the following subsections.

4.1. Confidence in the data quality

The first step prior to performing any kind of analysis of WGS data is to check the quality of the sequence obtained from the sequencing instrument. Statens Serum Institute (SSI) has a custom pipeline that checks a range of quality parameters for each sample and also provides a few basic analysis answers e.g. 7-locus MLST and resistance genes found in each sample.

The pipeline is freely available (https://github.com/ssi-dk/bifrost) and is developed for quality control of WGS of bacterial samples collected for surveillance. Bifrost is a QC platform which runs a QC pipeline based on de-novo assembly (skesa), with contaminant check (kraken), remapping of reads (minimap2) and variant calling (variant_caller in bbmap) for checking contaminants. The resulting values are collected and compared to species specific value ranges as informed by expert opinion from Statens Serum Institut and against previous successful samples run through the pipeline. The resulting output is a graphical interface where samples can be sorted according to several parameters including supplying lab and species.

Once the data has been approved for further analysis, cgMLST analysis is done using BioNumerics. The loci is called in a dual mode, both by blasting an assembly and by mapping the reads, and a consensus decision is made. There are a range of quality parameters reported in the BioNumerics software, and the user can then decide which cut-off that are acceptable for the situation at hand. The genome size of the strain is important and should be as expected for the pathogen in question. The percentage of core loci reported is also quite important, if this is too low there are enough missing data to influence your analysis outcome by creating false short distances between strains. Also the number of multiple alleles should be observed, if there are several multiple alleles detected this could be a sign of species-species contamination.

The clustering algorithm that are usually applied to cgMLST data are single linkage clustering. Depending on which species, subspecies or even clone that are analysed, there are different considerations to be made to determine where to set the cut-off in a specific cluster to decide whether a strain is part of a possible outbreak. Some species are more clonal than others and some outbreaks are more diverse than others depending on vehicle and distribution.

4.2 Data visualisation - epidemiological data and phylogenies

To keep the overview in surveillance for outbreak deteion and investigations, it is an advantage to easily visualise analytical results in relation to a number of information fields (metadata) that are usually kept in the surveillance databases.

Several tools can visualize phylogenies. This includes both open source and commercial solutions. In the COMPARE consortium, we have observed that while most academic partners use custom made and/or open source solutions; the public health institutions often use for their routine work commercial solutions that are tailored for the specific needs of continuous surveillance. Another reason might be that commercial systems are easy to use by microbiologists reponsible for the surveillance and outbreak detection.

COMPARE has identified the following visualization tools in use in the WP4 consortium :

Tool	**Used by**	**Open source / commercial**
Bionumerics	SSI / ANSES	C
Seqsphere	ANSES / RKI	C
CSI-phylogeny	DTU	OS
Figtree	DTU/APHA/SSI	OS
Geneious	DTU / RKI	C
Itol	DTU / ANSES/APHA	OS
Mega	DTU/APHA/SSI	OS
R	DTU/APHA/SSI	OS
Micro-react	APHA	OS
Enterobase /Grapetrees GUI	ANSES/APHA/SSI	OS
Phandango	APHA	OS

COMPARE has also developed Notebook solutions, tools hat should prove valuable in the next future to build directly on sequencing data for more rapid initial sorting of data.

4.3 Linking of isolates from various sources and reservoirs

Two examples of the use of WGS data to investigate origin of human contamination

In some situations, genomic data processing can provide major contribution to the attribution of sources. This is particularly evident when conventional typing methods are weakly discriminatory. As an example, *Salmonella* Derby in France, which ranks among the 10 most frequent *Salmonella* serotypes isolated in humans, is also one of the most prevalent serotypes in pork and poultry meat. The genetic diversity of a large collection of French *S*. Derby isolates representative of the pork and fowl sectors was analyzed by SNP. The *S*. Derby isolates were spread across four different genetic lineages found to be associated with specific animal hosts: pork or poultry (Sevellec et al., 2018). The inclusion of human strains in the phylogenetic analysis revealed that 98% of human strains belonged to the swine strain clusters, demonstrating that this sector was the main cause of human infections in France. This kind of result provides the risk manager with actionable information that allows him to target a specific sector to take appropriate management actions.

Another example is the exploitation of WGS data to identify host-segregating genetic markers (i.e. genes) to perform source-attribution studies for campylobacteriosis (Thépault et al., 2017). The work was based on a pan-genome analysis of 884 *Campylobacter jejuni* genomes which provides a set of 1,810 loci that were analysed for their host-segregating potential. This made it possible to retrieve fifteen loci that were used to attribute the source of French and British *Campylobacter* isolates using STRUCTURE software, a Bayesian model-based clustering method designed to infer population structure and attribute individuals to populations using multilocus genotype data (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461096/). Interestingly, the results of the study confirmed chicken as a major source of campylobacteriosis in the United Kingdom and that ruminants contributed an equal part of the human contamination in France. This unexpected result highlighted the necessity to investigate potential transmission route of *Campylobacter* from ruminant to humans. Additionally, it was found that the source of campylobacteriosis is subject to annual variations and therefore surveillance must not be relaxed (Thépault et al., 2018). Further source attribution methods currently developed in COMPARE project are described in D4.4.

4.4 Conclusions

WGS outputs may identify new strains on which models could be run on an ad hoc basis but routine surveillance would probably produce too many new strains for this to be efficient. However, the work here has raised some questions about the level at which cases should be defined. Previous studies have found that while the numbers of SNP differences between isolates within outbreaks were usually small (2-12 SNP differences) larger differences of up to 249 (Leekitcharoenphon et al., 2014) may exist and may be dependent on the serovar and also on the used analytical tool. Different SNP pipelines could produce different number of SNPs between isolates and therefore the exact number of SNPs might not be comparable. The SNP differences between strains may also depend on the SNP based sub-typing workflows used (Saltykova et al., 2018), stressing the need for thresholds at which to include or exclude cases to be considered on an outbreak-by-outbreak basis. Likewise temporal changes in the SNP address where strains have changed over time need to be considered and defining outbreak cases by SNP 10 or 25 level should be considered. Where SNP address is too discriminatory, cases may be wrongly excluded. However, if too inclusive then much time is spent investigating cases that are not related. The use of WGS and SNP address in case definition will also depend on the strain in question. A new or rare strain would be valuable in defining cases, whereas a more common strain that is present over much of the population would be less so. This same caveat can be applied at the serotype or phagetype level too. Epidemiological information and proven transmission links need to be also taken into account. Based on only a

small amount of data and descriptive analysis this work has suggested that strains with the same or similar SNP addresses are epidemiologically linked but this seems to only hold when strains are geographically or temporally close and not so much where there are large distance in time and space between the isolates. Using data from the Gastrointestinal Bacteria Reference Unit (GBRU) at PHE, Waldram et al., (2018) only found a significant epidemiological link for 17 out of 32 clusters of isolates they looked at and Ågren et al., (2016) showed that herds with known epidemiological contacts generally showed smaller SNP differences between isolates than where no known links were found. More information is needed about the rates of change of SNPs over time and what it means to have identical strains 10+ years apart. Leekitcharoenphon et al., (2014) was unable to find an association between time of isolation and the number of SNP differences and suggested the existence of groups of isolates that comprise single clonal haplotypes with virtually no genetic change over time. Knowing more about changes in strains over time is important for outbreaks spanning many years.

SNP address is one way of describing the differences between strains and classifying them into groups. There is an urgent need for some consensus about how WGS data is used. Currently SNP address is run per institution but addresses are not comparable between institutions unless as different reference strains are used so for a multinational outbreak SNP address would not be useful. There are other methods such as cgMLST and there is an urgent need for collaboration between labs to agree on a method that can be used on a large scale.

5. References

- Achtman, M., Wain, J., Weill, F.X., Nair, S., Zhou, Z., Sangal, V., Krauland, M.G., Hale, J.L., Harbottle, H., Uesbeck, A., Dougan, G., Harrison, L.H., Brisse, S. (2012) *S*. Enterica MLST Study Group. Multilocus sequence typing as a replacement for serotyping in *Salmonella* enterica. PLoS Pathog. 8(6):e1002776.
- Alikhan, N.F., Zhou, Z., Sergeant, M.J., Achtman, M. (2018) A genomic overview of the population structure of *Salmonella*. PLoS Genet. 14(4):e1007261. doi: 10.1371/journal.pgen.1007261.
- Ågren, E.C., Wahlström, H., Vesterlund-Carlson, C., Lahti, E., Melin, L., Söderlund, R. (2016) Comparison of whole genome sequencing typing results and epidemiological contact information from outbreaks of *Salmonella* Dublin in Swedish cattle herds. Infect Ecol Epidemiol. 6:31782.
- Ashton, P., Nair, S., Peters, T., Tewolde, R., Day, M., Doumith, M., Green, J., Jenkins, C., Underwood, A., Arnold, C., de Pinna, E., Dallman, T., and Grant, K. (2015) Revolutionising Public Health Reference Microbiology using Whole Genome Sequencing: *Salmonella* as an exemplar. bioRxiv preprint Nov 29, 2015; https://doi.org/10.1101/033225
- Dallman, T.J., Byrne, L., Ashton, P.M., Cowley, L.A., Perry, N.T., Adak, G., Petrovska, L., Ellis, R.J., Underwood, A., Green, J., Hanage, W.P., Jenkins, C., Grant, K., Wain, J. (2015). Whole genome sequencing for national surveillance of shiga-toxin producing *Escherichia coli* O157. Clinical Infectious Diseases 61(3): 305–312.
- ECDC Scientific Advice (2015) Expert opinion on the introduction of next-generation typing methods for food-and waterborne diseases in the EU and EEA. https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/food-and-waterborne-diseases-next-generation-typing-methods.pdf.
- Leekitcharoenphon, P., Nielsen, E.M., Kaas, R.S., Lund, O., Aarestrup, F.M. (2014) Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella* enterica. PLoS One.9(2):e87991.
- Pightling, A.W., Pettengill, J.B., Luo, Y., Baugher, J.D., Rand, H., Strain, E. (2018) Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. Front. Microbiol. https://doi:10.3389/fmicb.2018.01482 .
- Saltykova, A., Wuyts, V., Mattheus, W., Bertrand, S., Roosens, N.H.C., Marchal, K., De Keersmaecker, S.C.J. (2018) Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to *Salmonella* enterica serotype Typhimurium and serotype 1,4,[5],12:i:. PLoS One. 13(2):e0192504.
- Schwermer, H., Reding, I., Hadorn, D.C. (2009) Risk-based sample size calculation for consecutive surveys to document freedom from animal diseases. Prevent. Vet. Med. 92: 366–372.
- Sévellec, Y., Vignaud, M.L., Granier, S.A., Lailler, R., Feurer, C., Le Hello, S., Mistou, M.Y., Cadel-Six, S. (2018) Polyphyletic Nature of *Salmonella* enterica Serotype Derby and Lineage-Specific Host-Association Revealed by Genome-Wide Analysis. Front Microbiol. 9:891. doi:10.3389/fmicb.2018.00891.
- Simon, S., Trost, E., Bender, J., Fuchs, S., Malorny, B., Rabsch, W., Prager, R., Tietze, E., Flieger, A. (2018) Evaluation of WGS based approaches for investigating a food-borne outbreak caused by *Salmonella* enterica serovar Derby in Germany. Food Microbiol.71:46-54. doi: 10.1016/j.fm.2017.08.017.
- Thépault, A., Méric, G., Rivoal, K., Pascoe, B., Mageiros, L., Touzain, F., Rose, V., Béven, V., Chemaly, M., Sheppard, S.K. (2017) Genome-Wide Identification of Host-Segregating Epidemiological Markers for Source Attribution in *Campylobacter jejuni*. Appl Environ Microbiol. 83, 7.doi: 10.1128/AEM.03085-16.

- Thépault, A., Rose, V., Quesne, S., Poezevara, T., Béven, V., Hirchaud, E., Touzain, F., Lucas, P., Méric, G., Mageiros, L., Sheppard, S.K., Chemaly, M., Rivoal, K. (2018) Ruminant and chicken: important sources of campylobacteriosis in France despite a variation of source attribution in 2009 and 2015. Sci Rep.8(1):9305. doi:10.1038/s41598-018-27558-z.
- Waldram, A., Dolan, G., Ashton, P.M., Jenkins, C., Dallman, T.J. (2018) Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014. Food Microbiol.71:32-38. doi: 10.1016/j.fm.2017.04.005.
- WHO whole genome sequencing for foodborne disease surveilance, Landscape paper 2018: http://apps.who.int/iris/bitstream/handle/10665/272430/9789241513869-eng.pdf?ua=1.