Deliverable

4.2 Algorithm for detection of informative (sub-) types for epidemiological analysis and RA for the main food-borne pathogens

Version: 1

Due Month: 32

Completed: Month 32

Contributors: Ethelberg S (SSI), Gillesberg Raiser S (SSI), Hald T (DTU), Smith RP (APHA), Arnold ME (APHA), Litrup E (SSI), Kroneman A (RIVM), Nielsen EM (SSI), Work package 4/7.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Table of content

Deliverable Description

Whole genome sequencing is a powerful technique, the application of which is leading to profound changes within the field of microbiology. It is increasingly being applied within diagnostics and, more recently, for epidemiological purposes. Its use for surveillance and the characterisation of circulating zoonotic and foodborne organisms provides new possibilities and new challenges in the analysis of data. This document aims to give an overview of how to apply whole genome sequencing data within different areas, to provide meaningful data for epidemiological analysis and risk assessment.

1. Reasons and strategies for analysis of surveillance data

1.1 Purpose and types of disease surveillance

Surveillance of infectious microorganisms and the diseases they cause, form the basis of national and international disease preparedness systems. Circulating and emerging disease-causing bacteria, viruses and parasites may be monitored from humans (generally as a bi-product of diagnostic practises), or through the monitoring of wild animals, production animals, the environment and crops, and the control programs and quality assessment of food products.

Microorganisms and the diseases they cause may be under surveillance for a number of different reasons, including:

- Detection of disease outbreaks
- Source tracing, e.g. in outbreaks involving imported foods
- Description of long-term trends and possibly emergence of 'success clones'
- Estimation of the impact of control measures such as interventions, food safety and other knowledge exchange campaigns
- Monitoring of antibiotic resistance traits
- Detection of changes in bacteria and viruses, e.g. the evolution of virulence factors or mechanisms to escape immune system responses
- Applications for research to understand (and ultimately prevent) risk factors and transmission routes of infectious diseases, such as foodborne infections
- Performing source attribution, i.e. a complete description of the relative importance of the combined sources of infection for a disease, to help guide surveillance strategies
- Identification of population groups with special risks of certain diseases, i.e. incidence according to age, gender, geography, or risk behaviour, which could inform risk-based surveillance schemes.

It is important that the disease surveillance systems are set up in order to best address these purposes (something which is sadly often not the case). The phrase 'Data for Action' is frequently used by epidemiologists to capture why surveillance is utilised, with data provided that directly serves the purpose of limiting the number of infections. Disease surveillance is sometimes conceptualised as a circular process to emphasise this fact, see Figure 1.1. Information is collected and analysed to provide data for risk assessments that, if they highlight a public health problem, are used to target an intervention, the result of which is monitored by the continued disease surveillance.

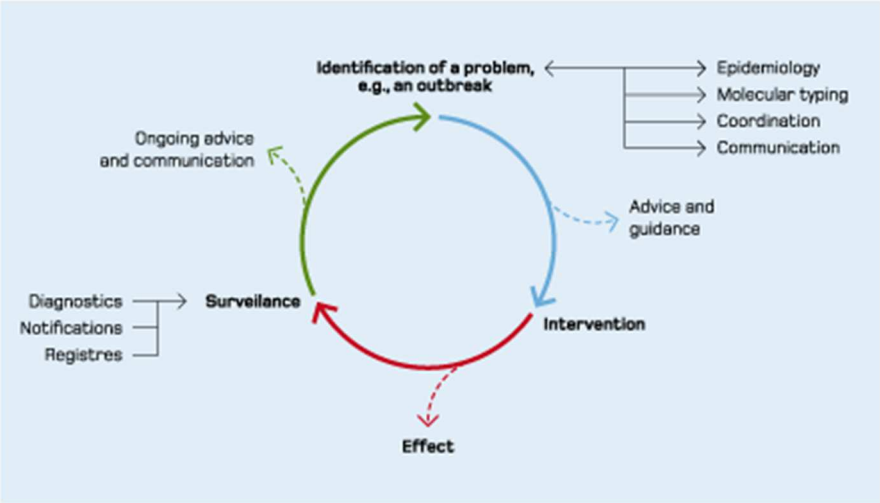


FIGURE 1.1: THE SURVEILLANCE LOOP. DATA FOR ACTION ARE BEING GENERATED THROUGH DISEASE SURVEILLANCE. SOURCE: WWW.SSI.DK . SEE TEXT FOR EXPLANATIONS.

A modern surveillance system therefore may make use of several instruments with which to collect and analyse data in a timely way and thereby provide results and knowledge to the authorities responsible for treatment, prevention and control.

1.2 Application of whole genome sequencing data for disease surveillance

For the purpose of surveillance, microorganisms causing foodborne diseases have historically been identified and characterised by various methods, determined to provide cost-efficient detection and characterisation specific to each type of organism. This involved phenotypic methods such as speciation and serotyping or phage typing but also, in more recent years, a number of genotypic methods. Although these were often pathogen specific, PFGE in particular has been applied successfully as a more generic typing method for foodborne bacteria, enabling epidemiologists to detect outbreaks dispersed over time or long geographical distances. In virology in recent decades, the prevalent typing method has been Sanger sequencing of PCR amplicons, often with different targeted genomic regions per laboratory. When new typing methods have been developed it has taken years to provide the hands-on experience needed to perform sound epidemiological analyses and new methods have often not been applied uniformly in different countries, making it impossible to do cross-border comparisons.

The generation of and analysis of whole genome sequencing (WGS) data is now increasingly becoming part of disease surveillance in many countries. A single typing method used by all confers obvious advantages for interpretation and comparison of results and additionally WGS holds the promise of being able to simultaneously fulfil all surveillance purposes listed above. The wealth of data generated, however, provides challenges for analysis and interpretation. Generally, it seems that the epidemiological (and microbiological) field is in the middle of a transformation process similar to what has been witnessed when new typing methods have previously been introduced. Tools for analysis, data storage and communication are needed, as well as experience on how to use and interpret data for routine surveillance purposes. This is becoming available for some organisms, *Listeria monocytogenes* being a good example and therefore treated in some depth in this document. For other organisms, e.g. *Campylobacter sp*, application is yet in its infancy.

1.3 Choice of analysis path depends on the aim of analysis

Though a single typing method, WGS, may now be used, it is not possible to devise a uniform analytical method for epidemiological results. One size does not fit all. Viruses and bacteria are fundamentally different. This will be reflected in the choices made for analysis of their sequence data. However, two different types of bacteria may also require very different analytical approaches to reach useful public health information. *Listeria* and *Campylobacter* for instance, have very different behaviours in terms of frequency and virulence, but also in terms of epidemiology and transmission routes. Furthermore, the analytic perspective also very much depends on the specific epidemiological field; disease in farm animals is surveyed and managed very differently from disease in humans. And from the point of view of food producers, a third set of methods might be emphasised.

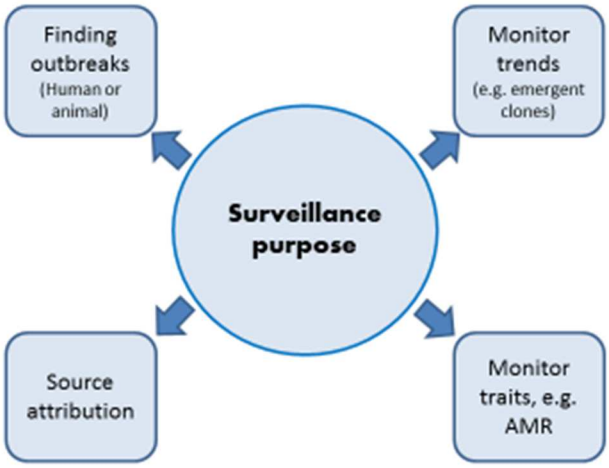


FIGURE 1.2: THE APPROACH TAKEN FOR ANALYSIS OF DATA DEPENDS ON THE PURPOSE OF THE INVESTIGATION.

The choice of analytical method for treating surveillance data also depends on the ultimate purpose of surveillance, as illustrated in Figure 1.2. Timely detection of *Salmonella* outbreaks with the aim of rapid control is different from trying to understand the dynamics in antibiotic resistances gene evolution and transmission between bacterial species. In this deliverable, we have tried to cover this aspect by describing different methods with an emphasis on outbreak detection and investigation and source attribution.

1.4 Finding and investigating outbreaks

One of the major aims of foodborne microorganism surveillance is detecting disease outbreaks. Using WGS for routine surveillance may find important use in detecting disease outbreaks, defining cases and comparing patient organism characteristics with those from organisms isolated from foods, production facilities or animals. Whereas the previously used typing methods for bacteriology (e.g. PFGE) would directly produce an epidemiological meaningful grouping of pathogens, with WGS data the definition of the genetic similarity within the grouping has to be agreed on and extracted from the rich sequence data using an appropriate analysis method. This step is what is referred to as A in Figure 1.3. Secondly, subtype information may then form the

basis of traditional cluster analysis detection methods (which may also rely on metadata), illustrated with B in Figure 1.3. Verification and investigation of the outbreak may then use a variety of epidemiological and microbiological methods, some involving data analysis. This is illustrated with process C in Figure 1.3. Here, collaboration between virologists and bacteriologist in the Compare consortium is very valuable; in virology cluster detection based on evolutionary distance cut-offs (based on partial genomic sequences) have been common practice for many years.

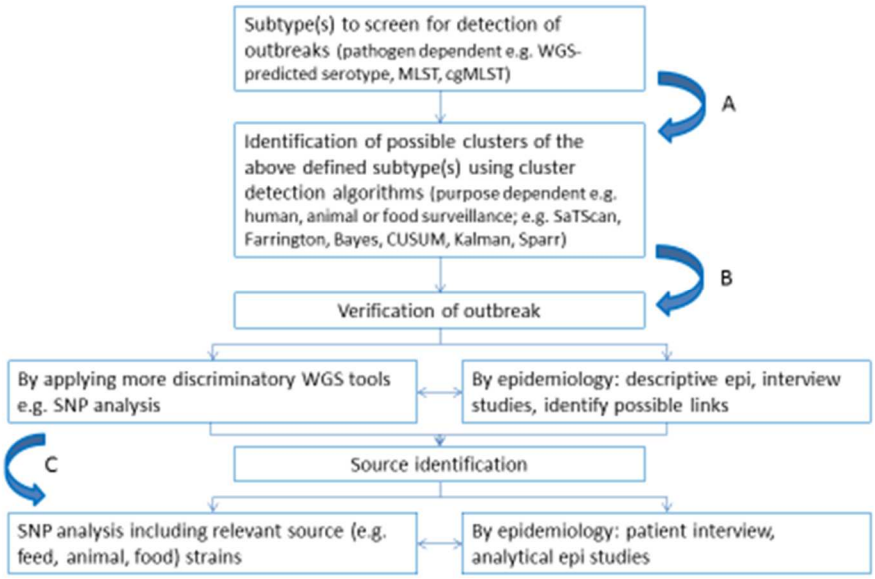


FIGURE 1.3: ALGORITHM FOR INVESTIGATION OF OUTBREAKS.

1.5 Content of this document

In this deliverable, we have taken to understand the term 'Algorithm' as a decision tree (not a mathematical singular algorithm). The deliverable aims to give an overview of how WGS data may be applied with an epidemiological purpose and to suggest the most appropriate methods. This deliverable describes the diversity of applications and divides the analysis into different situation dependent steps. These include the analysis of sequence data to provide useful organism groups (step A in Figure 1.3, Chapter 2) and the application of outbreak detections methods (step B in Figure 1.3, Chapter 3). It mentions different types of analysis tools, with the aim of finding the source of an outbreak (step C in Figure 1.3, Chapter 4). It further gives examples of a recent use in an epidemiological context within the Compare network, particularly involving work on *Listeria monocytogenes* (Chapter 5), but also on *Salmonella* (Chapter 6). It then goes on to treat the application for risk assessment and source attribution (Chapter 7).

2. Overview of methods to analyse sequence data

It is a requirement of the mathematical algorithms described in Chapter 3 that the WGS data are assigned to "types", i.e. that the genomes are grouped into a limited number of genetically closely related groups below the species level. Occurrence of these types is the data input in the algorithms described.

Different approaches can be used for analysis of WGS raw data and for the phylogenetically relevant grouping of genomes. Ongoing and planned work in the Compare project is evaluating these methods for the purpose of outbreak detection and linking sources to patients. Here, we present an overview of the presently most commonly used analysis methods and provide two relevant examples of the application of these methods. In the following section, the focus will be on the analysis of whole genome sequences of bacteria.

2.1 Subtyping by MLST

Original 7-locus MLST (Multi Loci Sequence Typing) is often used for initial typing of bacteria and gives a sequence type (ST) which yields a useful discriminatory level for grouping isolates into sub-types in most bacteria. In many cases, e.g. for *Salmonella*, the ST corresponds quite well to the serovar that was previously used as a sub-type for epidemiological surveillance (Achtman *et al* 2012). The 7- locus ST can rapidly and easily be deduced from the raw data produced by the sequencing run, and the ST is often the first step used to subdivide the strains into sub-groups for further phylogenetic analysis and cluster detection.

In bacteria that are less clonal, the Clonal Complexes (CC) are also often used as a subtype. The CC's are usually larger groups containing several STs, which share 5 or 6 loci in the MLST scheme.

Following the subtyping and division into smaller groups, the WGS data of isolates are then submitted to a more discriminatory analysis, which are based on larger fractions of the genome. The outcome of these analyses are genetic clusters, where isolates and the patients from which they were isolated can be subjected to epidemiological investigation. Description of these cluster detection methods are available in Chapter 4.

Reference

Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S; S. Enterica MLST Study Group. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog. 2012;8(6):e1002776. doi: 10.1371/journal.ppat.1002776.

BOX 2.A: A NATIONAL EXAMPLE: PUBLIC HEALTH ENGLAND'S APPROACH[1]

Whole genome sequencing (WGS) is now being performed by Public Health England (PHE) for routine surveillance of several gastrointestinal pathogens including *Salmonella* spp., *Shigella* spp., pathogenic strains of *Escherichia coli* and *Listeria monocytogenes*. This service provides resolution to the single nucleotide polymorphism (SNP) level. Typically, the genetic relationship between isolates (SNP differences) is represented on a phylogenetic tree; however in order to work with continuous phylogenetic data in an epidemiological context, a system to categorise the data into discrete groups is necessary. A hierarchical system for defining and naming clusters called the 'SNP address' has been developed and incorporated into PHE's gastrointestinal pathogen bioinformatics pipeline to address this need[1].

The initial bioinformatics step involves deriving the Multi Locus Sequence Typing (MLST) which is used to assign each isolate a sequence type (ST). STs are further grouped into discrete, phylogenetically related groups called eBurst groups (eBGs) or clonal complexes (CC). The eBG or CC is used to identify the appropriate reference genome for the second bioinformatics step. This involves the alignment of each isolate's genome to the appropriate reference genome and pairwise analysis to calculate the number of individual SNP differences (i.e. the 'genetic distance') between all pairs of isolates within a specific eBG or CC. Single linkage hierarchical clustering is performed at seven descending thresholds of SNP distance within each eBG or CC; these thresholds are 250, 100, 50, 25, 10 and 0 SNPs. This clustering results in a discrete seven-digit code where each number represents the cluster membership at each descending SNP distance threshold. The resultant SNP address provides an isolate-level nomenclature where two isolates with the same SNP addresses have 0 SNP differences. Ancestral relatedness of isolates with similar SNP addresses can then be ascertained using traditional methods.

All reference laboratory results for gastrointestinal bacteria are held within PHE's dedicated database, Gastro Data Warehouse (GDW). Several tools have been developed for cluster detection, extraction, characterisation and assessment to aid in the detection and prioritisation of outbreaks for investigation. These tools are used by teams at the local as well as the national level. They allow the extraction of line lists of cases matching the SNP address at the 5-SNP level, which are summarised by case and cluster-level demographics (age and sex distribution, temporal and geographic distribution, travel history outside the UK and presence of any food, water, environment or animal isolates in the cluster). These attributes are then used to perform an initial rapid assessment of each cluster. Clusters flagged for further more in depth assessment are then evaluated in the context of the wider phylogeny using more traditional methods (descriptive epidemiology and construction of a phylogenetic tree). Common attributes in genetically proximate isolates are used to aid hypothesis generation and to define the outer (or inner) limits of the cluster. Based on the outcome of this analysis, the cluster may be selected for full outbreak investigation. A partial SNP address (to the level chosen to define the outbreak) is then incorporated into the outbreak investigation case definition and used to monitor the cluster for new activity and following a public health intervention, to monitor the success of the intervention (if no further cases within the cluster limits are detected, the intervention is deemed successful).

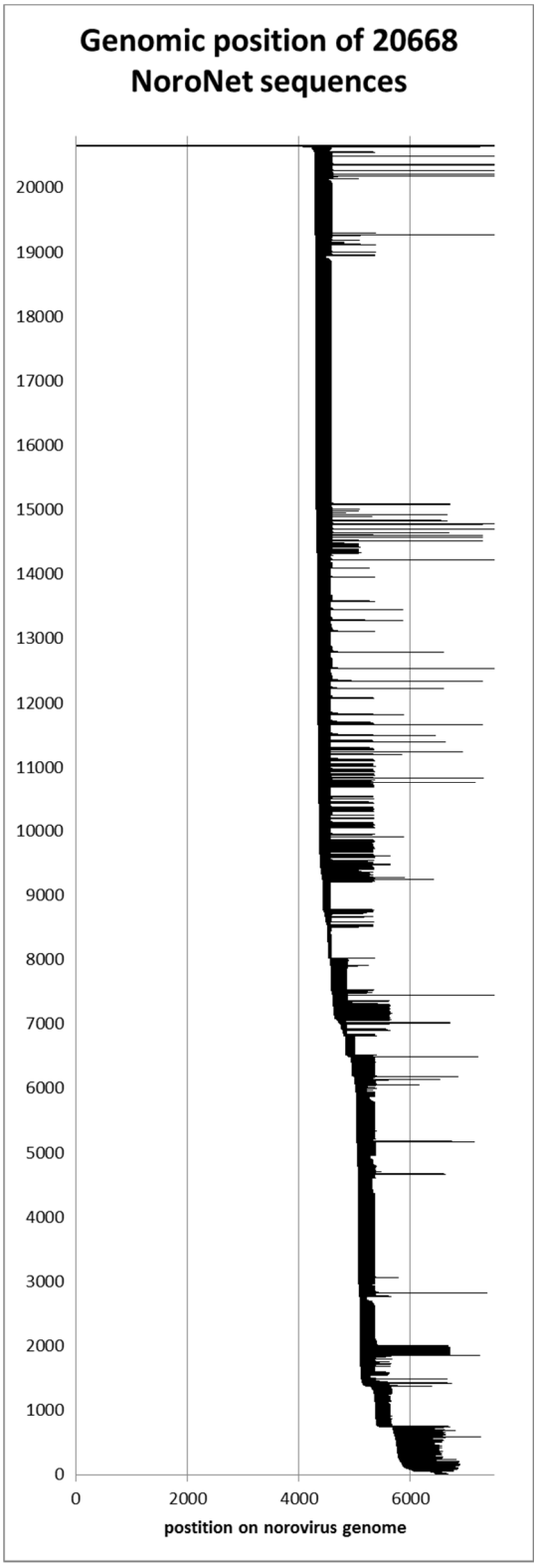
The current approach developed by PHE uses a SNP difference threshold to detect clusters with new activity, but thresholds are not used to define the limits of outbreak investigations as these are determined on a case-by-case basis to fit the context of the investigation. For most pathogens, the 5-SNP level is used as the cluster detection threshold as this was identified in validation studies[1–3] as the level within which outbreaks were most commonly contained and beyond which cases were less likely to be part of the same epidemiological event. An exception has been applied for *Shigella* sp. (detected at the 10-SNP level) as transmission for this pathogen is typified by long chains of person to person transmission over an extended period of time, during which more genetic variation is likely to occur.

References

1. Waldram, A., Dolan, G., Ashton, P. M., Jenkins, C. & Dallman, T. J. Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014. *Food Microbiol.* (2017). doi:10.1016/j.fm.2017.02.012
2. Dallman, T. J. *et al.* Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin. Infect. Dis.* **61,** 305–12 (2015).
3. Dallman, T. J. *et al.* Use of whole-genome sequencing for the public health surveillance of *Shigella sonnei* in England and Wales, 2015. *J. Med. Microbiol.* **65,** 882–884 (2016).

[1] Public Health England is not a participating organisation in the COMPARE project

BOX 2.B: EXAMPLE: ELEVATIONS TOOL IN NORONET



In international data sharing databases with sanger sequences of viruses, the genomic regions of the sequences often do not (completely) overlap. For instance in the Noronet database, sequence lengths range from 30 nt to 7570 (i.e. complete genomes), and there are several different regions on the genome which are being used for PCR and sequencing (see figure). This makes clustering algorithms based on multiple alignment less useful. Also clustering based on pairwise similarity will only lead to clusters of overlapping sequences. Often molecular clustering is used to support or disprove the identification of clusters based on epidemiological criteria.

In the Noronet database, a clustering algorithm (elevations module) has been designed in which a cluster is identified on a combination of criteria, both epidemiological (time and place) and molecular (genetic distance).

As in NoroNet we are interested in the detection of international dispersed (foodborne) clusters the criteria of the elevation module are set as follows:

A new cluster is identified when:

- Two sequences are found with a pairwise similarity of at least 99.5% (momentarily based on BLAST score): This means that 1 mutation is accepted in sequences which overlap by at least 200 nt.
- The two sequences have reporting dates which are at most one year apart: As norovirus can be transmitted via frozen foods, we kept this period this long
- The two sequences have been reported by two different countries: because we are interested in dispersed/diffuse/international outbreaks

This elevation analysis is run daily on all newly submitted sequences.

A new sequence is added to an existing cluster if:

- It has at least 99.7% similarity with one of the sequences in the previously identified clyster: this way two non-overlapping sequences can end up in one cluster if one longer sequence is available which overlaps with both.
- Is has a reporting date which is at most one year apart from one of the sequences in the cluster.
- It is reported by a different country from one of the other sequences in the cluster.

The software of the NoroNet database includes a notification system, in which an email is sent to the curator(s) when a new elevation has been identified.

As norovirus is a pilot pathogen in Compare, several partners are performing WGS on norovirus. Once a substantial amount of norovirus WGS data have been collected, the Noronet elevation algorithm will be redesigned and tested for cluster detection based on complete genomes.

3. Inventory of available cluster detection algorithms

Cluster detection algorithms can be used for the identification of disease outbreaks and may thereby contribute to their early detection. This chapter assesses five cluster/anomaly detection algorithms used for disease monitoring and their appropriateness for use of whole genome sequencing (WGS) data. This chapter is based on the previous Compare Milestone report within WP4, M16.

The five algorithms selected represent distinct types of mathematical methods to determine spatial, temporal or spatio-temporal clusters which could be used for disease outbreak detection. This chapter summarizes the methods of each algorithm and its applicability for WGS datasets, based on the experience of the authors. The methods represent the main classes of these algorithms, however, it should be noted that many modifications of these methods or methods that utilize similar approaches exist which are not presented here. It is a requirement of the algorithms that the WGS data are already 'pre-analysed' so that strains are put into groups for epidemiological analysis (Chapter 2).

The methods are described and discussed below.

3.1 Spatial/ temporal cluster detection

3.1.1 SaTScan temporal, spatial or spatio-temporal analysis

Method: The scan statistic within the SaTScan freeware (https://satscan.org; Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) locates the sites of the most likely pathogen clusters that are not randomly distributed. The method can be used for spatial, temporal or spatio-temporal cluster detection, with a time precision of day, month or year. The test compares the relative risk of being a case within an area/ time period in comparison with the risk outside of the area/ period. Clusters of cases are identified when the relative risk is above that expected and the cluster has the maximum likelihood of representing the study population. A main cluster with the highest relative risk is identified while secondary clusters are detected in this method that do not overlap the main cluster. The method can use either a Bernoulli model, where the WGS output is ordered into cases and controls, a Poisson-based model, assessing the number of cases in an area against a known background population at risk, or a space-time permutation model, where only the cases are necessary. The choice of model is often constrained by the lack of data on the background population.

This methodology has been used in a number of situations to investigate possible areas of disease outbreaks, such as to assess the most likely spatial-temporal cluster of horse cases of *Salmonella* Krefeld in a veterinary hospital (Pare et al., 1996). Other examples include the assessment of acute respiratory disease in Norwegian cattle herds (Norstrom, Pfeiffer and Jarp, 2000) and the detection of the most likely temporal cluster of different *Salmonella* serotypes in dairy cattle (Sato et al., 2001).

Retrospective Space-Time analysis
scanning for clusters with high rates using the Poisson model.

SUMMARY OF DATA
Study period.....: 2008/6/1 - 2012/6/30
Total population.....: 14
Total number of cases....: 834
Annual cases / 100000....: 1462794.1

MOST LIKELY CLUSTER
Location IDs included.: All
Time frame.....: 2010/3/1 - 2010/5/31
Number of cases.....: 77
Expected cases.....: 52.35
Annual cases / 100000.: 2151773.0
Observed / expected...: 1.471
Relative risk.....: 1.519
Log likelihood ratio..: 5.455924
Monte Carlo rank.....: 74/1000
P-value.....: 0.074

FIGURE 3.1: EXAMPLE OF SATSCAN ANALYTICAL OUTPUT

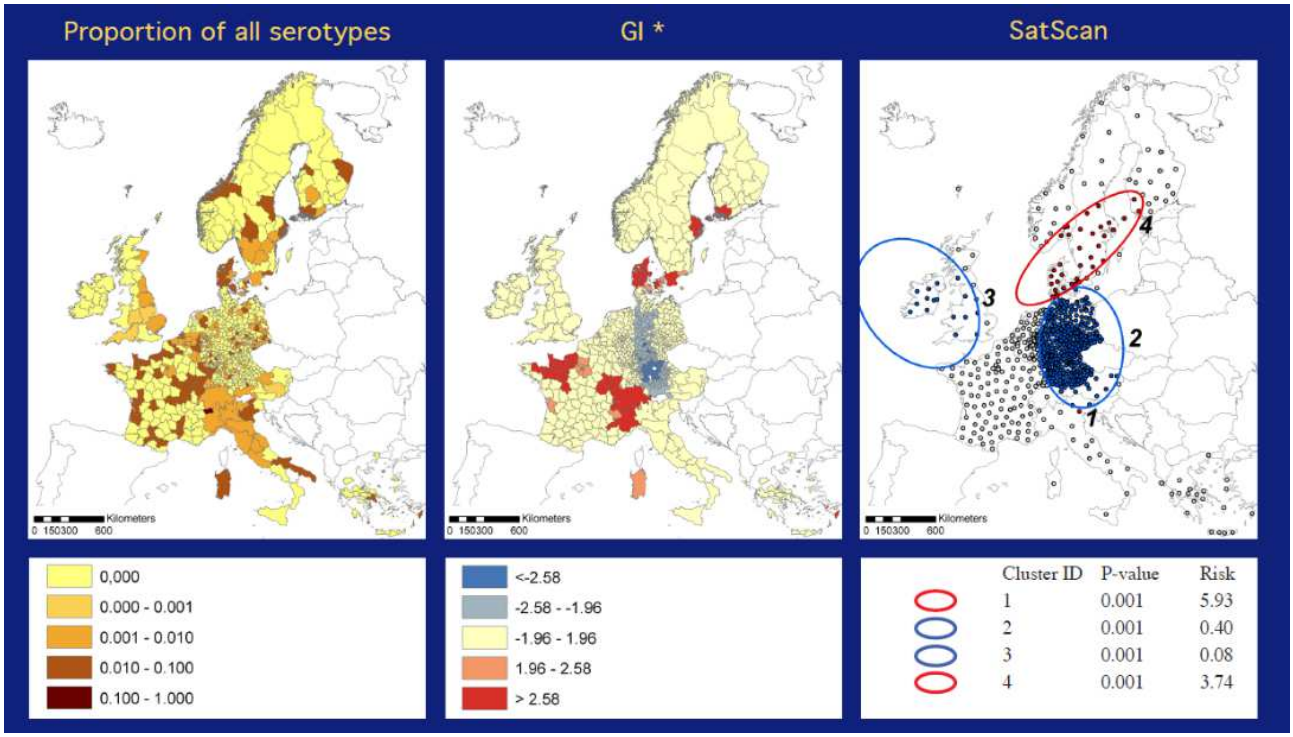


FIGURE 3.2: EXAMPLE OF THE USE OF SATSCAN TO ANALYSE EUROPEAN SURVEILLANCE DATA. THE EXAMPLE USES PAN-EUROPEAN DATA FROM 2002 COLLECTED VIA THE ENTER-NET NETWORK ON SALMONELLA BOVISMORBIFICANS. THREE TRANS-NATIONAL CLUSTERS WERE DETECTED USING ELLIPSOID ANALYSIS, PANEL 3. (PEDERSEN ET AL, 2009)

Pros: The scan statistic can adjust for changes in population size over time, take the underlying population into account, and can also incorporate a limited number of categorical covariates.

Cons: the power to detect clusters is reduced for long and narrow clusters, as the test searches for circular clusters, and for where clustering occurs throughout the study region. The test has also been amended to look at different shapes of clusters (e.g. oval), although these are computationally complex. The temporal-spatial analysis greatly increases the run time of the analysis and so this may not be suitable for weekly assessments. The outputs of the analysis are not presented on a map.

Expected suitability for use with WGS data: SaTScan has been shown to have suitable power in detecting localised epidemiological clusters which would be useful. However, the statistical power has been shown to be reduced for long and narrow spatial clusters, as the test searches for circular clusters, and for where clustering occurs throughout a study region.

3.1.2 Sparr spatial cluster analysis

Method: The 'Sparr' package (SPAtial Relative Risk) developed for the R statistics platform provides a method for detecting spatial 'hot spot' of high or low relative risk (https://cran.r-project.org/web/packages/sparr/index.html; Davies, Hazelton and Marshall, 2010). The method requires data to be organised into a binomial outcome and allows for the estimation of both fixed and adaptive kernel-smoothed relative risk surfaces via a density-ratio method. Most applications of the relative risk function utilise plotting the

relative risk within the study region (especially for an inspection of tolerance contours), and analytical outputs can attribute a risk value and a P-value to each point in a grid that overlaps the study region. The resolution of the grid can be adjusted. The benefit of the Sparr method is that the data-adaptive techniques allow for high intensity analysis of data-rich geographical areas and low intensity analysis of data-sparse areas, whereas most spatial analyses require a single intensity level (bandwidth) and so are limited by any areas with few cases. This method was shown to be useful in analysing the presence of H5 and H7 avian 'flu subtypes from surveillance across the European Union (Hillman et al., in prep).

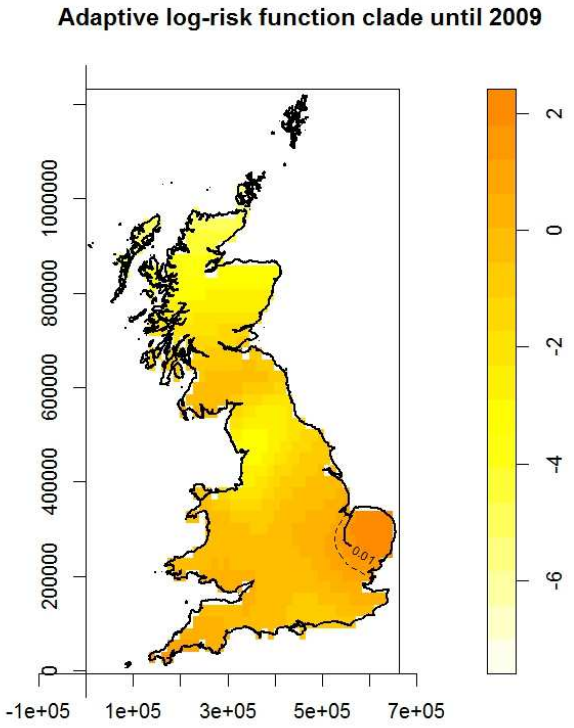


FIGURE 3.3: EXAMPLE OF 'HEAT' MAP OF RELATIVE RISK FROM SPARR ANALYSIS (CONTOURS INDICATE P-VALUE FOR AREAS OF STATISTICALLY SIGNIFICANT RISK).

Pros: The analytical output is typically summarised in an easy to interpret map of smoothed log-relative risk, displayed as 12 increasing 'heat' colours, with two sets of contour lines representing the upper 5th and 25th percentile of risk. The ability to account for areas of sparse data without limiting the analysis of other areas could be very useful for large-scale analysis of WGS, for example for comparing across regions with large amounts of relevant WGS data and areas with few data points.

Cons: The method may be oversensitive to indicating areas of significant risk when data is scarce, which would likely be a concern with its use for WGS data. The analysis is also computationally demanding to run.

Expected suitability for use with WGS data: As WGS of isolates is not completed as standard and data coverage may be patchy, a method that can account for areas of sparse data may be very important for completing analysis of WGS surveillance data. The method would also highlight the size and shape of spatial areas of high risk that could be targeted for enhanced surveillance.

3.2 Temporal anomaly detection

3.2.1 Farrington algorithm temporal cluster detection

Method: A Farrington algorithm uses a rolling window (typically weekly or monthly) to assess whether the number of cases in that time period are significantly above the expected number derived from previous comparable periods (e.g. weeks or months in the past years) (Farrington et al., 1996). The algorithm has been used to produce 'alarms' or signals whenever the number of cases in a week/ month exceeds that expected so that these anomalies can be investigated. These alarms should instigate an appropriate manual evaluation as detailed in Vial and Berezowski, 2015 (https://www.ncbi.nlm.nih.gov/pubmed/25475688) and described further in chapter 4.

The algorithm has been modified a number of times (e.g. Noufaily et al., 2013) and can be set to select a threshold value based on different distributions e.g. log-linear. Reference data selection can be optimised to enable any seasonality and trends in the data to be incorporated and thus limit the number of false alarms. In addition, the method incorporates an attempt to adjust for over-dispersion and for the occurrence of outbreaks in the reference period. The algorithm typically requires stable historic data, typically three to five years, to use as reference data. The Farrington algorithm has been used to monitor disease surveillance data by many organisations (APHA, PHE, SSI etc, see for instance Hulth et al., 2010). The algorithm is available through the 'Surveillance' package for the R statistics platform (https://cran.r-project.org/web/packages/surveillance/index.html).

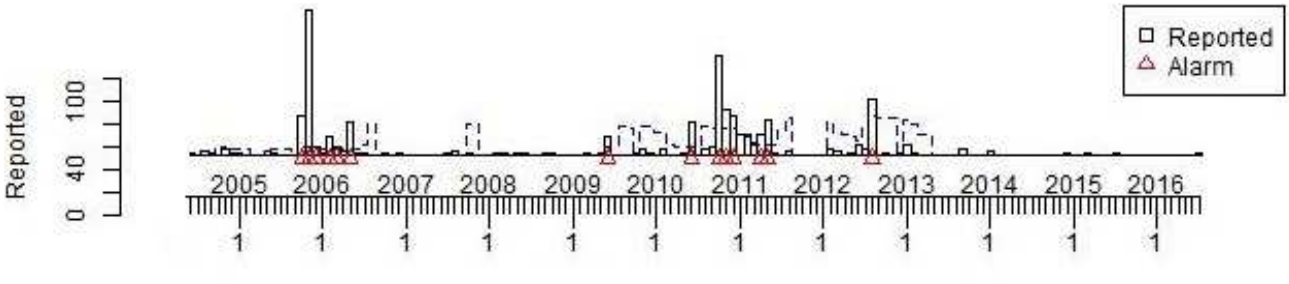


FIGURE 3.4: EXAMPLE OF FARRINGTON ANALYSIS RESULTS PLOTTED IN A GRAPH FOR MONTHLY DISEASE INCIDENCE DATA.

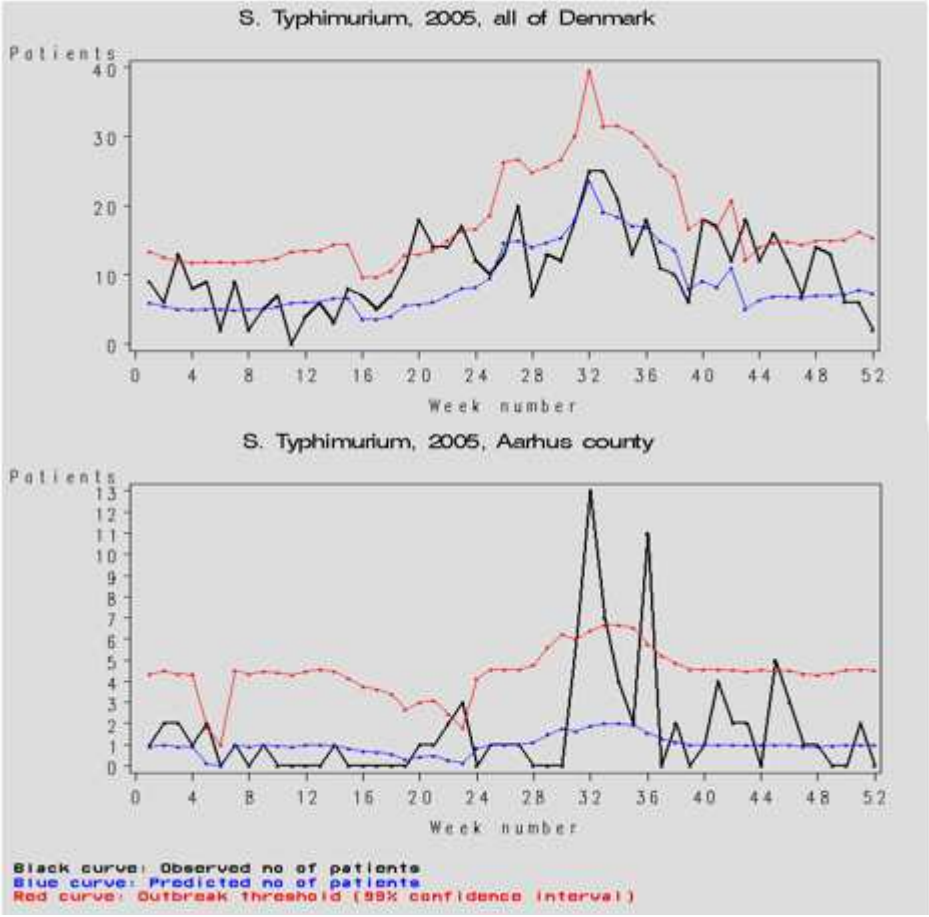


FIGURE 3.5: EXAMPLE OF THE APPLICATION OF THE FARRINGTON METHOD IN ROUTINE NATIONAL SURVEILLANCE OF SALMONELLA SEROTYPE DATA, DENMARK. IN THE EXAMPLE SHOWN, WEEKLY ANALYSES OF AN ENTIRE YEAR IS COMPILED; AN OUTBREAK WITH SALMONELLA TYPHIMURIUM WAS DETECTED USING REGIONAL ANALYSIS (ETHELBERG ET AL 2006).

Pros: Simple, fast analysis that accounts for seasonality. These analyses have been shown to have high detection sensitivity in simulations (0.76-1.00, Arnold et al., unpublished data).

Cons: Benefits from a consistent historical dataset to compare against, which may be unlikely at present. In scenarios when the number of cases per month is small, then experience shows that when there are less than five cases following a period of zero cases then alarms can be ignored as alarms can be triggered with just a single case as the algorithm predicts zero cases as the threshold. Due to this, the method may not be useful for rare events. Detection sensitivity and time to detection have been shown in simulations to vary according to length and size of outbreaks.

Expected suitability for use with WGS data: One foreseeable problem is that this method assumes a stable surveillance system exists which have given rise to a comparable historical data, whereas currently in most organisations WGS is used inconsistently, to identify rare or unusual cases or during known outbreaks.

3.2.2 Bayes temporal cluster detection

Method: It is assumed that the data follow a negative binomial distribution, with parameters based on previous case numbers. The threshold value at which an outbreak alert is triggered is set according to a user-specified percentile i.e. an alert is triggered if the probability of observing the data is less than the user-specified value. Bayes can be implemented through the R package 'Surveillance' (https://cran.r-project.org/web/packages/surveillance/index.html). In comparisons of outbreak detection performance when applied to simulated data, the Bayes algorithm produced the highest sensitivity and shortest time to detection in outbreak simulations (APHA unpublished data), although there was only a small difference between the Bayes and Farrington algorithms.

Pros: Has been shown to have comparable sensitivity to Farrington in previous comparisons (Arnold et al., unpublished data).

Cons: Will have similar data requirements to the Farrington algorithm. Much less widely used than the Farrington, so less evidence available of satisfactory performance.

Expected suitability for use with WGS data: As detailed under the Farrington method, the sensitivity of this method may suffer due to the lack of a stable and consistent historical dataset.

3.3 Statistical process control

3.3.1 CUSUM (CUmulative SUMs) temporal cluster detection

Method: This conceptually simple method assesses the sequential cumulative sum of the deviation between a reference value and observed values. An alarm is raised if the cumulative sum equals or exceeds a pre-specified boundary (such as a higher than expected number of cases), which can be a single atypical observation or repeated exceedances (O'Farrell, 2015). In typical process control environments, CUSUM would be used to detect both positive and negative exceedance of the boundary (e.g. products that are too big and too small), whereas in disease surveillance we are only interested in positive exceedance.

Examples of its use in the veterinary field include assessment of daily egg production (Mertens, 2009) or lameness in cattle based on weight at milking (Pastell and Madsen, 2008), whereas it has been used in a variety of public health applications to monitor disease surveillance data (Woodall, 2006). CUSUM is available through the 'qcc' package for the R statistics platform (https://cran.r-project.org/web/packages/qcc/qcc.pdf).

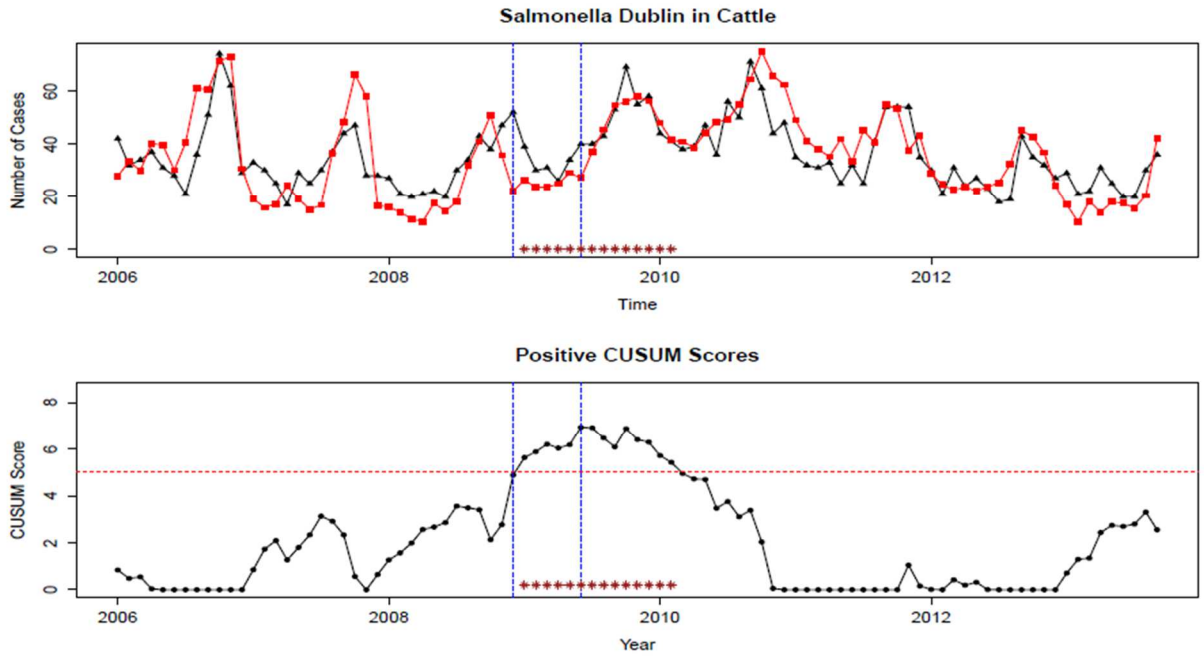


FIGURE 3.6: EXAMPLE OUTPUT OF CUSUM USE, APPLIED TO BRITISH SALMONELLA DUBLIN DATA (O'FARRELL, 2015).

Pros: The algorithm does not rely on large amounts of historical data for threshold prediction.

Cons: An issue with the use of CUSUM to monitor biological data is that typically the outcome is somewhat subjective, and not as rigidly standardised as in a statistical process control environment, which could affect the sensitivity of the algorithm.

Expected suitability for use with WGS data: As for most disease monitoring systems, WGS has not been used for a relatively large number of years, CUSUM benefits from not needing a large amount of historical data to detect temporal anomalies. However, CUSUM may have problems accounting for seasonal variation that is likely to be present for many pathogens.

3.4 Summary

The five typical anomaly detection algorithms discussed here could be used for spatial or temporal outbreak detection of sequenced isolates. However, a number of specific issues have been identified, specifically those related to the non-standard application of sequencing and lack of a stable, historical dataset to which current results could be compared to. The remaining chapters discuss which genetically similar subtypes could be used as outcomes in these analyses and which algorithms would be most suitable for sequenced isolates. This chapter was also delivered as Milestone 16.

References

Davies, T., Hazelton, M., Marshall, J. (2010) Sparr: analysing spatial relative risk using fixed and adaptive kernel density estimation in R. J. Stat. Softw. 39, 1-14.

Ethelberg S, Hanon FX, Peter Gerner-Smidt P, Olsen KEP, Wegener HC & Mølbak K (2006): Five Years use of an Outbreak Detection Algorithm for Diarrheagenic Bacteria, Denmark. Emerging infectious diseases conference, Atlanta, USA

Farrington, C.P., Andrews, N.J., Beale, A.D. and Catchpole, M.A. (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. J. R. Stat. Soc. Series A, Vol. 159, No. 3, 547-563.

Hulth, A., Andrews, N., Ethelberg, S., Dreesman, J., Faensen, D., van Pelt, W., Schnitzler, J. (2010). Practical usage of computer-supported outbreak detection in five European countries. Euro Surveillance: European Communicable Disease Bulletin, 15(36). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20843470.

Kulldorff, M., Nagarwalla N. (1995) Spatial disease clusters: Detection and Inference. Statistics in Medicine, 14: 799-810.

Kulldorff, M. (1997) A spatial scan statistic. Communications in Statistics: Theory and Methods, 26: 1481-1496.

Mertens, K. (2009) An intelligent system for optimizing the production and quality of consumption eggs based on synergistic control. PhD Thesis.

Norstrom, M., Pfeiffer, D.U., Jarp, J. (2000) A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. Preventive Veterinary Medicine, 47 (1-2): 107-119.

Noufaily, A., Enki, D.G., Farrington, P., Garthwaite, P., Andrews, N., Charlett, A. (2013). An improved algorithm for outbreak detection in multiple surveillance systems. Statistics in Medicine, 32(7), 1206–1222. https://doi.org/10.1002/sim.5595.

O'Farrell, H. (2015) Temporal Modelling of Disease Outbreaks using State Space and Delay Differential Equations. PhD thesis. http://epubs.surrey.ac.uk/809649/1/Final_thesis_24_12_15_HOF.pdf

Pare, J., Carpenter, T.E., Thurmond, M.C. (1996) Analysis of spatial and temporal clustering of horses with Salmonella in an intensive care unit of a veterinary hospital. Journal of the American Veterinary Medical Association, 209 (3): 558.

Pastell, M. and Madsen, H. (2008) Application of CUSUM charts to detect lameness in a milking robot. Expert Systems with Applications, Vol. 35, No. 4, 2032-2040.

Pedersen A, Bødker R, Fisher I, Ethelberg S (2009): Spatio-temporal cluster analysis of ten selected salmonella serotypes in Europe, 2001-2006. Report (Deliverable) part of WP6, GIS in epidemiology, Med-Vet-Net.

Sato, K., Carpenter, T.E., Case, J.T., Walker, R.L. (2001) Spatial and temporal clustering of Salmonella serotypes isolated from adult diarrheic dairy cattle in California. Journal of Veterinary Diagnostic Investigation, 13 (3): 206-212.

Woodall, W.H. (2006) The use of control charts in health-care and public-health surveillance (with discussion). J. Qual. Technol., Vol. 38, No. 2, 89-104.

4. Methods for outbreak verification and source identification

The World Health Organization (WHO) defines a foodborne outbreak as either the occurrence of an increased number of cases of disease than what would normally be expected in a defined community, geographical area or season, or the occurrence of two or more cases of a similar foodborne disease resulting from the ingestion of a common food (WHO 2008). The following will describe the steps for the management of an outbreak of infectious diseases in humans in order to verify the outbreak and identify the source or vehicle. This is, however, not an outbreak management manual[2]. The aim is to illustrate the level of work and collaboration that takes place in such a situation and that the algorithm shown in Figure 1.3 (Chapter 1) should always be seen in relation to the pathogen, its characteristics, transmission routes and epidemiology.

4.1 Human food- or waterborne outbreaks

A total of 4,362 food- and waterborne outbreaks involving 45,874 human cases were reported from 26 EU member states in 2015. A strong link between human cases and vehicles was found in 422 foodborne outbreaks (EFSA and ECDC 2016). Different microbiological and epidemiological methods can be applied in order to identify a link between an outbreak and its source.

An outbreak or outbreak signal can be detected using several surveillance methods. One method is a cluster detection algorithm applied to laboratory data, as described above. An outbreak can also be identified by a defined signal threshold or by mandatory notifications to public- or veterinary health authorities by doctors or notifications from citizens. Once an outbreak or outbreak signal is detected further laboratory typing and epidemiological investigations can be initiated. The aim is to verify if the genetic cluster/outbreak signal is an outbreak or not, to define which patients can be included as cases in the outbreak and if they are possible, probable, or confirmed cases. Cases are defined using a time, place and person description. The case definition is key to generating a hypothesis and identifying the source but is, however, adjustable according to the development of the outbreak or as additional information arises (CDC 2011, Heymann 2008) e.g. change of the geographical area or inclusion of more strains as occasionally seen (Gillesberg Lassen et al 2013).

Once a case definition is made, active case finding can be initiated and cases can be described epidemiologically, as a minimum by time, place and person. Active case finding can, for example, be conducted by identifying if the outbreak strain has been seen in other geographical areas or by use of questionnaires asking cases if cases know more people with the same symptoms or describing symptoms in order to detect cases that have not been assessed by a doctor (US CDC 2011, Heymann 2008). Having comparable typing information across laboratories is generally a necessity for this process leading to hypothesis generation for source identification.

Generating and testing a hypothesis on possible outbreak source(s) can be done by applying laboratory and epidemiological methods. It is preferable to apply both concurrently. The current outbreak strain can be compared after typing with previous outbreak strains with a known source or with strains derived from food

[2] For outbreak management manuals we can refer to CIFOR “Guidelines for Foodborne Disease Outbreak Response – toolkit”, WHO “Foodborne disease outbreaks: guidelines for investigation and control” from 2008, US CDC “Self-Study Course SS1978 Principles of Epidemiology in Public Health Practice, Third Edition An Introduction to Applied Epidemiology and Biostatistics” from 2006 and updated in 2011 or national toolkits available in individual countries.

isolates but with no genetic match. Epidemiologically, hypothesis generation can be derived from the descriptive epidemiology of cases and by case interviews (US CDC 2011).

The hypothesis can also be tested using both laboratory and epidemiological methods. A strong link between a source or vehicle and human cases is found when either a food or environmental isolate and a human isolate match or an analytical epidemiological study, such as a retrospective cohort study, a case-control study or case-case study, can show a statistically significant higher risk of disease for people exposed to the source or vehicle. It is preferable to have the laboratory, environmental and the epidemiological link between exposure and illness (US CDC 2011).

4.2 Methods to detect and delineate a genetic cluster

As described in Chapter 2, the initial analysis of WGS-data for use in cluster detection algorithms should focus on the grouping of isolates, e.g. for bacterial illnesses at the level of 7-locus MLST. For further analysis of the possible epidemiological links between patients and between patients and suspected sources, a higher resolution is needed. At present, two different main approaches, SNP and gene-by-gene analyses, are in use as described below.

4.2.1 Single Nucleotide Polymorphisms (SNP) analysis

Single Nucleotide Polymorphisms (SNP) are nucleotide variants in a specific isolate called against a chosen reference genome. In the analysis of a larger dataset, i.e. a collection of strains from a suspected outbreak, all the positions shared between the strains and the reference are analysed for SNPs to the chosen reference. This is a core genome SNP analysis and this kind of phylogenetic analysis is considered one of the most accurate phylogenetic analyses. However, the SNP analysis has the drawback that any genomic feature not present in the chosen reference will not be analysed even if it is present in the remaining population of suspected outbreak strains. Furthermore, the core genome and SNP dataset has to be re-calculated for each addition of a new strain as the core genome might change when new strains are added to the analysis. Additionally, this new calculation can then again change the already determined distance between genomes.

The core genome SNPs are then used for phylogenetic analysis and grouping of the strains in smaller clusters determined by the branches in the tree. The detection of a genetic cluster and the delimitation of the cluster is now possible. The method of calculating SNPs and trees can be somewhat time-consuming if the hardware of the calculating computers is not powerful enough. In addition, the possibilities to compare the results of these analyses between laboratories are not straightforward as e.g. a nomenclature of such analyses are not immediately obtained. Public Health England has attempted to overcome this drawback of SNP analysis by introducing “SNP addresses” (see Box 2.1, Chapter 2).

4.2.2 Gene-by-gene analysis (MLST, cgMLST, wgMLST)

The gene-by-gene approach is a reference-free method that determines any number of present loci and the sequence of these. In original MLST, it is often 7 loci that are detected, whereas in the whole genome MLST (wgMLST) several thousand loci, including those only present in some strains, are included in the analysis (Maiden et al., 2014). Core genome MLST (cgMLST) aims at including only the loci that are expected to be present in almost all strains of a certain species. For many organisms, a cgMLST scheme has been developed, e.g. for *Listeria monocytogenes* (Maura et al, 2016) and *Campylobacter jejuni/coli* (Sheppard et al, 2012). The level of

discrimination differs greatly between MLST and cg/wgMLST. However, all methods are suitable for determining groups of relatedness within a strain population, with the latter two having much more discriminatory power.

The gene-by-gene approach of analysing a population are more suited for a nomenclature scheme as every locus is assigned a number, and then it is merely a text string of numbers that needs to be exchanged between laboratories to determine if strains are genetically closely related. The format of the gene-by-gene approach also means that once the alleles has been called, the comparison of cg/wgMLST of hundreds of strains is quite fast and genetic clusters are very easily and quickly re-calculated on a daily basis to determine if new cases or suspected sources are part of the detected cluster.

To allow for a 'global' nomenclature, the gene-by-gene approach demands a public, curated and maintained database that everyone can use for the allele calling. Many such databases already exists for foodborne bacteria (http://bigsdb.pasteur.fr/listeria/listeria.html; http://pubmlst.org/campylobacter; https://enterobase.warwick.ac.uk), and there seems now to be consensus among the public health institutes organised in PulseNet International that the allele-based analysis is the way forward for public health (Nadon et al, 2017).

References

EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control) (2016). The European Union summary repo rt on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. EFSA Journal 2016;14(12):4634,231 pp. doi:10.2 903/j.efsa.2016.4634

Gillesberg Lassen S, Soborg B, Midgley SE, Steens A, Vold L, Stene-Johansen K, Rimhanen-Finne R, Kontio M, Löfdahl M, Sundqvist L, Edelstein M, Jensen T, Vestergaard HT, Fischer TK, Mølbak K, Ethelberg S (2013). Ongoing multi-strain food-borne hepatitis A outbreak with frozen berries as suspected vehicle: four Nordic countries affected, October 2012 to April 2013. Euro Surveill. 2013;18(17):pii=20467. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20467

Heymann, David L. (Editor) (2008). Control of Communicable Diseases Manual, 19th Edition.

Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728-36. 10.1038/nrmicro3093

Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nat Microbiol. 2016;2:16185.

Nadon et al. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017 Jun 8;22(23). pii: 30544. doi: 10.2807/1560-7917.ES.2017.22.23.30544.

Sheppard SK, Jolley KA, Maiden MC. A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter. Genes (Basel). 2012;3(2):261-77.

US CDC (2011). Self-Study Course SS1978 Principles of Epidemiology in Public Health Practice, Third Edition An Introduction to Applied Epidemiology and Biostatistics. 2006, updated in 2011. Lesson 6; Section 2: Steps of an Outbreak Investigation. URL: https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson6/section2.html

World Health Organization (2008). Foodborne disease outbreaks: guidelines for investigation and control, World Health Organization, Geneva, Switzerland

5. CASE: Use of WGS in surveillance and outbreak management of human listeriosis

The introduction of new typing methods such as prospectively used WGS for routine surveillance of pathogens can bring with it a number of advantages but may also raise new questions and challenge previously defined concepts. The surveillance of *Listeria monocytogenes* (*Lm*) in food and invasive infections in humans has been one of the central starting points for applying WGS and getting experiences using WGS in routine national surveillance. The prospective application of WGS for national surveillance of listeriosis has been described for countries such as USA, Australia and Denmark (Jackson et al 2016, Kwon et al 2016, Kvistholm Jensen et al 2016a, Gillesberg Lassen et al 2016). The following will describe the application of WGS as part of an enhanced listeriosis surveillance programme in Denmark and highlight key lessons learned.

5.1 Enhanced surveillance of listeriosis in Denmark

In Denmark, listeriosis is a laboratory notifiable disease and all isolates from the local clinical microbiological departments are sent to the national reference laboratory at Statens Serum Institut for typing. In addition, listeriosis is epidemiologically notifiable when causing meningitis or seen to be a foodborne infection (BEK nr 277). In September 2013, real-time WGS replaced PFGE as first-choice typing method in the national surveillance of human *Lm* isolates. All patients, since January 2014 have been followed up clinically upon notification and when possible interviewed with a questionnaire on risk factors such as travel and food consumption history. In June 2014, real-time WGS comparison of human isolates with food isolates taken from national surveillance or in outbreak situations began in collaboration with the Danish Veterinary and Food Administration and the Danish National Food Institute (Kvistholm Jensen 2016a, Gillesberg Lassen 2016).

Epidemiological follow-up of patients and multi-agency collaboration, integrating laboratory and epidemiological surveillance on both human and veterinarian side is also described for Australia and USA (Kwong et al, 2016; Jackson et al, 2016). This aspect necessary for the comparison of human and food isolates in real-time and to confirm laboratory links epidemiologically or confirm epidemiological links with laboratory testing.

5.2 Outbreak management

In Denmark in 2005-2013, prior to the introduction of WGS in routine surveillance, three confirmed and four non-confirmed outbreaks of listeriosis were registered in the Danish national database of Food- and Waterborne Outbreaks (FUD). One of these outbreaks had a known source (Smith et al 2010). A retrospective analysis of trends in listeriosis in Denmark in 2002-2012, using PFGE and partly MLST identified 29 clusters, when a cluster was defined as at least three patients with indistinguishable pulsotypes within 14 weeks. Four of the 29 clusters were only identified once in the period. One of these was the identified outbreak with known source (Kvistholm Jensen et al 2016b). Since 2014, nine confirmed outbreaks have been registered in the FUD-database with four having a known source. The food sources were found using both epidemiological linking and a laboratory WGS-link between food/environmental isolates and human isolates. In addition, it has been possible in some situations to link a single case to a specific food source (i.e., solve an 'outbreak' consisting of one case only). The first outbreak registered after the initiation of the enhanced surveillance of listeriosis was detected in June 2014. This is to date the largest outbreak of listeriosis in Denmark and the largest in Europe for many years, comprising 41 cases and 17 deaths. Cases were confined by <3 SNP difference (Kvistholm Jensen 2016a), while one occuring patient with an isolate with the same ST (ST224) but 60 SNPs difference was defined as a non-case. By use of

WGS the outbreak was traced to and proven to stem from a single production plant making a pork based cold-cut product and subsequently retracted from 6000 down-stream businesses, following which the outbreak stopped. This WGS-success was follow by the identification of two additional outbreaks, where cases were defined by <5 SNPs difference, both linked to consumption of smoked fish. Two cases within one month was only seen twice in the three years these two relatively small outbreaks with 10 cases each (and 8 deaths) occurred (Gillesberg Lassen 2016). These two outbreaks would very likely not have been identified using previous typing methods, and even if they had been so, not all cases would have been interviewed in due time. These outbreaks were also largely solved by performing inspections and taking samples for testing at the two fish production plants and finding WGS matching *Lm* isolates, including in the production environments. Without the stringency of WGS (rather than for instance PFGE), a food source link would not have been established with sufficient strength for the Danish authorities to act on. These three outbreaks have been described in scientific journals with a focus on the benefits of using WGS for outbreak investigations within the Compare project (Compare is acknowledged in both papers) (Kvistholm Jensen 2016a, Gillesberg Lassen et al 2016).

5.3 An outbreak or a genetic cluster?

For *Lm*, a pathogen that can survive in the environment and has a long incubation time (Heymann 2008), the use of WGS has enabled countries to identify prolonged outbreaks and their sources. The previously described three outbreaks in Denmark and the experiences from USA and Australia illustrate this (Jackson et al 2016, Kwon et al 2016, Kvistholm Jensen et al 2016a, Gillesberg Lassen et al 2016). The practical application of the outbreak definition by WHO (WHO 2008) (see Chapter 4) may therefore be challenged when WGS is applied as typing method in routine surveillance. However, this will depend on the pathogen, source and vehicle.

One question that arises when applying the WHO outbreak definition using WGA in surveillance of foodborne pathogens is: when can we say that two patients are most likely to have the same source of infection? When do we have a genetic cluster? In Denmark, it has been decided not to set an absolute SNP threshold for defining a genetic cluster. Instead, genetic clusters are defined using phylogenetic comparisons with historical strains. This is in line with one of the advantages of using real-time WGS for surveillance described by Kwong et al (2016); that applying real-time WGS in surveillance provides an overall sense of the genetic diversity of local strains and thus the background material for making phylogenetic comparisons.

A second practical question that rises when applying the WHO outbreak definition when using WGS in routine surveillance is: when does a genetic cluster constitute an outbreak? Investigating each genetic cluster as an outbreak may simply not be feasible. Experience of when a genetic cluster is or can develop into an outbreak is therefore important in order to set a threshold for when to trigger a signal to the epidemiologists that there is a genetic cluster and possible outbreak. In Denmark, based on a descriptive analysis on genetic clusters with two or more listeriosis patients form 2013-2015 (data not shown) as well as the incubation time and shelf life of known sources of listeriosis outbreaks, it was decided to look at a genetic cluster epidemiologically when

- Two patients diagnosed with invasive *Lm* had laboratory date within 12 weeks AND clustering on WGS
  OR
- Three patients diagnosed with invasive *Lm* clustering on WGS
  OR
- A food isolate clusters on WGS with one or more patient isolates.

Applying this algorithm in Denmark since 2015, we have identified eight confirmed outbreaks, six genetic clusters where epidemiological follow-up is conducted and two genetic clusters of two patients where no epidemiological follow-up has been conducted. An additional 11 genetic clusters of 2-3 patients have been identified prior to the threshold being defined (Figure 5.1). An additional five signals following the identification of *Lm* in food isolates were also followed up epidemiologically. A single human case was identified for one of these signals.

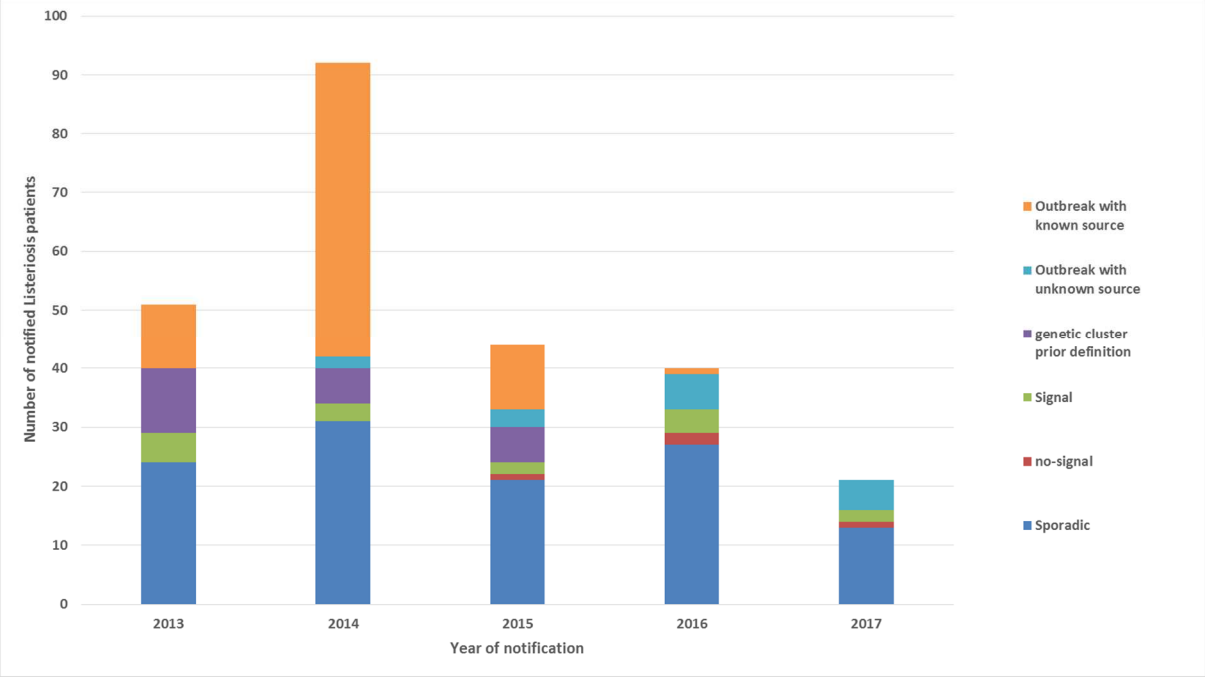


FIGURE 5.1: NUMBER OF LISTERIOSE CASES BY CASE-TYPE, DENMARK, JANUARY 2013- MAY 2017(N=248)

Based on the experience acquired in the USA, Jackson et al (2016) summarise six key ways in which the application of real-time WGS in national surveillance has improved outbreak investigations and give examples of each. These correspond well to the experiences seen in Denmark and described above. The six key points were that WGS could:

- Delineate clusters with diverse PFGE patterns
- Determine the source of "cold cases"
- Demonstrate that certain PFGE-defined clusters did not consist of highly related isolates
- Refine outbreak case definitions
- Link sporadic illnesses to contaminated food
- Confirm outbreaks following product testing.

5.4 Public health implications

In Denmark, the introduction of enhanced surveillance including WGS, and the identification of outbreaks and their sources, have resulted in a higher awareness of listeriosis and more information for public health agencies to act on. The national recommendations for risk groups have been updated based on a risk assessment where identified outbreak sources, consumption and risk behaviour has been taken into account (Fødevarestyrelsen 2016). The Danish Veterinary and Food Administration initiated a series of campaigns aiming at food production

companies producing products with risk of containing listeria, e.g. producers of smoked fish products in 2015 and 2017 and kitchens and producers of ready-to-eat food in 2016. This has led to a standardised testing scheme for and intensified dialogue with food producers and kitchens producing food for high-risk groups (e.g. hospital kitchens). It is too soon to evaluate the impact of the new initiatives; however, the number of listeriosis patients was below 45 per year in 2015 and 2016, which has otherwise not happened since 2005.

References

BEK nr 277, Bekendtgørelse om lægers anmeldelse af smitsomme sygdomme m.v. Acceded June 2017: URL: https://www.retsinformation.dk/Forms/R0710.aspx?id=21406

Fødevarestyrelsen (2016), Revurdering af anbefalinger om at undgå at blive syg af listeria – herunder afklaring af risikogrupper, Fødevarestyrelsen, Glostrup, Danmark (in Danish)

Gillesberg Lassen S, Ethelberg S, Bjorkman JT, Jensen T, Sorensen G, Kvistholm Jensen A, et al. (2016) Two listeria outbreaks caused by smoked fish consumption-using whole-genome sequencing for outbreak investigations. Clinical Microbiology and Infection, 22(7):620-4.

Heymann, David L. (Editor) (2008). Control of Communicable Diseases Manual, 19th Edition.

Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, et al. (2016). Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clinical Infectious Diseases, 63(3):380-6.

Kvistholm Jensen A, Nielsen EM, Bjorkman JT, Jensen T, Muller L, Persson S, et al. (2016a) Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. Clinical Infectious Diseases, 63(1):64-70.

Kvistholm Jensen A, Bjorkman JT, Ethelberg S, Kiil K, Kemp M, Nielsen EM (2016b). Molecular Typing and Epidemiology of Human Listeriosis Cases, Denmark, 2002-2012. Emerging infectious diseases.22(4):625-33.

Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. (2016) Prospective Whole-Genome Sequencing Enhances National Surveillance of Listeria monocytogenes. Journal of Clinical Microbiology, 54(2):333-42.

Smith B, Larsson JT, Lisby M, Müller L, Madsen SB, Engberg J, Bangsborg J, Ethelberg S, Kemp M (2010), Outbreak of listeriosis caused by infected beef meat from a meals-on.wheels delivery in Denmark 2009., Clinical Microbiological Infections, 2011; 17:50-52

World Health Organization (2008). Foodborne disease outbreaks: guidelines for investigation and control, World Health Organization, Geneva, Switzerland

6. CASE: Outbreak detection methods for Salmonella in Ducks

This veterinary science example relates to the detection of outbreaks of *Salmonella* in ducks in the UK. It illustrates use of several of the methods described in Chapter 3.

6.1 Aim

This pilot study was completed to assess outbreak detection methods to determine the usefulness of these methods for genome sequenced data.

6.2 Dataset context

Salmonella outbreaks which started in 2010 in Great Britain, Northern Ireland and the Republic of Ireland were linked to a specific phage type of *S.* Typhimurium (DT8), which is associated with ducks. Whole Genome Sequencing (WGS) and Multi-locus variable-number tandem repeat analysis (MLVA) was carried out on 295 *S.* Typhimurium strains from a collection of duck, human and other species' isolates from Great Britain, Northern Ireland, Republic of Ireland, France and Italy. The strains were collected between 1993 and 2013 and included those associated with the 2010 outbreak as well as a stratified random selection of other isolates to provide a representative selection in each calendar year. The MLVA and WGS work identified a group of closely related strains (hereafter referred to as the outbreak clade), which included mainly DT8 and DT30 isolates that emerged in association with the 2010 outbreaks.

6.3 Material and Methods

From the full dataset of sequenced isolates, only 142 duck isolates with a UK origin were selected. Post codes were matched to this file from farm information held at APHA and these were converted to x and y coordinates from a total of 111 isolates with relevant information.

The SatScan scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) locates the sites of the most likely clusters that are not randomly distributed. The test compares the relative risk of being a case within an area in comparison with the risk outside of the area. Clusters of cases are identified when the relative risk is above that expected and the cluster has the maximum likelihood of representing the study population. A main cluster is identified with the highest relative risk and secondary clusters are detected that do not overlap the main cluster. The test applied here used a space-time permutation model, which uses only the observed cases to detect spatio-temporal clusters. As applied as standard, a maximum size for a detected cluster was set as 50% of the studied population. To evaluate the scan statistics performance in 'real time', SaTScan was applied prospectively to the data for the outbreak clade outcome, starting with using data only from the first year in the dataset and increasing by one year for each analysis. Due to time constraints, the analysis did not include individual yearly assessments before 2005.

A Farrington algorithm was used, with a 'rolling window' used so as when each year of data was entered into the analysis it was added to the reference dataset for the following year. A monthly approach (current month, plus one month either side of the current month) was used for selecting reference data from previous years. If there were less than five cases following a period of zero cases then alarms generated were ignored that were based on a single case due to issues with generating an expected value within the analysis. The analysis started in year 1996 as the algorithm required a full three year 'run in' with reference data.

The Sparr (SPAtial Relative Risk) package provides functions to estimate fixed and adaptive kernel-smoothed relative risk surfaces via the density-ratio method and perform subsequent inference. Cases were defined as those within the outbreak clade and those remaining were controls. The analysis was used to plot smoothed risk values as a heart map along with P value contours onto the study region. In order to compare the risk values with the output of SaTScan a prospective analysis was run, starting with samples from 1993 to 1996 and adding each subsequent year to the Sparr analysis. For the analysis described in this report, R statistical package version 3.1.1. was used with the following packages: ggplot2, Rcpp, sp, rgeos, rgdal, sparr, maptools, raster, plyr and tidyr.

6.4 Results

Table 6.1 summarises the detection of significant spatial, temporal or spatio-temporal clusters for each rolling year of data. Significant (P<0.05) clusters from the SaTScan analysis were detected when running the model for 2010, 2011 and 2012. However, the clusters did not contain many farms; in fact, the primary cluster for 2010 only contained a single farm, which had seven cases between 2008 and 2010. The 2011 primary cluster included 6 cases, all from 2011, whereas the 2012 cluster included 5 cases from 2011 and 2012. Interestingly both the SaTScan and Farrington analyses agreed on the 2010-2012 outbreak periods, with the Farrington method also producing an alarm in mid 2009 before the 2010 known outbreak. The Farrington analysis also produced some earlier alarms (2003 and 2006) which appeared to be unrelated to the known outbreak. The Sparr analysis detected significant hotspots of risk in all but one year, which differed greatly in terms of size, from small clusters around a single point location to widespread increased risk including most of England, Wales and Scotland.

TABLE 6.1: SUMMARY OF RESULTS FROM THREE CLUSTER DETECTION METHODS APPLIED PROSPECTIVELY FROM 1996 TO 2013. DETECTION OF A SIGNIFICANT CLUSTER IS INDICATED BY Y.

Time period assessed	SaTScan	Farrington	Sparr
1993-1996	NA	N	Y
1993-1997	NA	N	Y
1993-1998	NA	N	Y
1993-1999	NA	N	Y
1993-2000	NA	N	Y
1993-2001	NA	N	Y
1993-2002	NA	N	Y
1993-2003	NA	Y	Y
1993-2004	NA	N	Y
1993-2005	N	N	N
1993-2006	N	Y	Y
1993-2007	N	N	Y
1993-2008	N	N	Y
1993-2009	N	Y	Y
1993-2010	Y	Y	Y
1993-2011	Y	Y	Y
1993-2012	Y	Y	Y
1993-2013	N	N	Y

NA: Not attempted
Y: Yes
N: No

6.5 Discussion

The study has tested the use of three different cluster detection methods (Sparr, SaTScan and Farrington algorithm). The methods differed greatly, with Farrington applying a monthly assessment of temporal clusters, whereas Sparr and SaTScan applied spatio-temporal methods and so were limited to only a reduced dataset that could be linked to geographical coordinates. Both spatial methods were run yearly, with the Sparr method detecting hotspots of significant risk for the selected time period, whereas the SaTScan method detected the most likely cluster in both space and time. Both spatio-temporal methods utilised a binary outcome of whether sequenced isolates belonged to the outbreak clade, with all other isolates defined as controls, whereas the Farrington analysis only used counts of cases without a denominator.

The epidemiologically known outbreak in 2010 was successfully detected by all three methods, with clusters being detected in 2010, 2011 and 2012 from samples from holdings associated with the outbreak. However, other clusters were detected by the Farrington algorithm in earlier years and also detected regularly by the Sparr methods. These results may indicate over-sensitivity in the method. However, due to the historical nature of the dataset and lack of epidemiological information, it cannot be discounted that some of these clusters were real outbreaks. The methods are generally used to produce alerts for potential outbreaks, which require further investigation to confirm. The Sparr analysis should have benefited from a method that uses an adaptive bandwidth, so that higher levels of detection intensity can be used in data rich areas and lower intensities used in sparse areas. It would appear that the method detected a significant high risk area in the dataset from 1993-1996 which was generally detected in subsequent years. The identification of spatial risk in historical data may not be useful for surveillance and so it would be preferable to limit Sparr analyses to data only from years of current interest.

The relative risks for the most-likely clusters in SaTScan were quite high and the numbers of cases quite low, which provides an example of the effect of the relatively small sequenced dataset and small number of cases. As detailed by Muellner et al (2016) “if observations become too few, the epidemiological analysis of the data generated can become very difficult, which impacts negatively on the ability to detect disease trends” and this has to be weighed against the additional discriminatory power that sequencing provides in ensuring cases related to an outbreak are correctly defined (Höfler, 2005; Petersen et al., 2011).

References

Höfler, M., 2005: The effect of misclassification on the estimation of association: a review. Int. J. Methods Psychiatr. Res. 14, 92–101.

Kulldorff, M., Nagarwalla N. (1995) Spatial disease clusters: Detection and Inference. Statistics in Medicine, 14: 799-810.

Kulldorff, M. (1997) A spatial scan statistic. Communications in Statistics: Theory and Methods, 26: 1481-1496.

Muellner P, Stärk KD, Dufour S, Zadoks RN. (2016) 'Next-Generation' Surveillance: An Epidemiologists' Perspective on the Use of Molecular Information in Food Safety and Animal Health Decision-Making. Zoonoses Public Health. 2016 Aug;63(5):351-7.

Petersen, R. F., E. Litrup, J. T. Larsson, M. Torpdahl, G. Sorenson, L. Muellner, and E. Nielsen, 2011: Molecular characterization of *Salmonella* Typhimurium highly successful outbreak strains. Foodborne Pathog. Dis. 8, 655–661.

7. Source attribution

Source attribution is defined as the partitioning of the human disease burden of one or more foodborne infections to specific sources, where the term source includes animal reservoirs and vehicles, e.g. food products (Pires et al., 2009). In contrast to the investigation of foodborne outbreaks, where the focus is on identifying the single (food) source causing the outbreak, source attribution works at higher population level trying to associate all human cases, including sporadic cases, of a particular foodborne pathogen with specific sources. Quantifying the most important food sources and animal reservoirs enables a targeted control of the foodborne pathogen and provides support for risk managers for their allocation of resources to control the disease. Source attribution is therefore regarded as an important decision-support tool in prioritizing effective food safety interventions (Havelaar et al., 2007).

There exist several approaches for conducting source attribution, including microbial subtyping (e.g. de Knegt et al, 2016), comparative exposure assessment (e.g. Pintar et al., 2016), meta-analysis of case-control studies (e.g. Domigues et al., 2012), summarization of outbreak data (e.g. Pires et al., 2012), intervention studies (e.g. Tustin et al., 2011) and structured expert judgement (e.g. Hald et al., 2016). All methods present strengths and limitations, and the applicability and usefulness of each depend on the question addressed, data availability, pathogen characteristics, and the type of intervention aimed for (Batz et al., 2005; Pires, 2013). In this chapter, we will focus on the use of microbial subtyping including the application of WGS data for source attribution of food-borne hazards.

7.1 Microbial subtyping

Microbial subtyping involves the characterisation of pathogen isolates by phenotypic and/or genotypic subtyping methods. The principle is to compare the distribution of subtypes in potential sources (e.g. animals and food) with the subtype distribution in humans. The approach is facilitated by the identification of strong associations between some of the dominant subtypes and specific reservoirs, providing a heterogeneous distribution of subtypes among the sources (Pires et al, 2009; Hald & Pires, 2011).

For many years, subtyping of food-borne bacteria has relied on phenotypic methods such as biotyping, serotyping or phage typing (bacteriophage susceptibility). For certain food-borne pathogens, phenotyping has enabled the identification of the main reservoirs for human infections (e.g. *S*. Enteritidis in poultry, *E. coli* O157 in ruminants, and *Y. enterocolitica* O:3 in pigs). Although, these methods are still valid, they are increasingly being replaced by or supplemented with molecular methods based on the characterisation of bacterial DNA (EFSA BIOHAZ Panel, 2013).

The most commonly applied molecular methods generate banding patterns (e.g. PFGE, AFLP), characterise variations at predefined loci (e.g. MLVA, MLST) or are based on whole genome sequences (e.g. wgMLST, SNP). Often these methods have been developed specifically to characterize very closely related isolates (e.g. SNP analysis in outbreak investigations) or, in contrast, to compare very distantly related isolates (e.g. MLST in evolutionary studies) (Barco et al., 2013). The former methods normally investigate fast-evolving genes, whereas the latter methods target the conserved and slowly evolving core genes.

For source attribution studies, the appropriate level of discrimination typically lies somewhere in between these two applications and will differ between pathogens depending among others on clonality (Son et al., 2013). This

is because surveillance systems only rarely are sufficiently fine-meshed to let us identify direct links between sporadic human cases and the responsible sources. We, therefore, need a process that allows for some genetic diversity between strains from human and food sources, but only to the degree so that it can still be assumed that they are epidemiologically related.

As mention above the microbial subtyping approach works best for pathogens that are heterogeneously distributed among the reservoirs, i.e. it should be possible to identify some host-associated subtypes. This also means that human cases are attributed to the reservoir level, whereas the actual transmission route through which humans are finally exposed is not revealed. As an example, cattle are the main reservoir for *S*. Dublin, but the relative importance of different pathways within this reservoir (e.g. dairy products, beef or direct contact) cannot be estimated based only on subtyping.

In addition to the application of discriminatory typing methods, the microbial subtyping approach requires a collection of temporally and spatially related isolates, preferably collected through integrated surveillance, so that the source data represent what the human population of interest has been exposed to (Pires et al., 2009; Hald & Pires, 2011). However, due to lack of such representative data some studies have used surrogate data, including data from different geographic regions and/or time periods, whereas other studies have simply not considered the relative occurrence of specific subtypes or sources that are otherwise known to play an important role (e.g. Mather et al., 2013). This may seriously bias the attribution results (Smid et al., 2013). So besides appreciating the population diversity and structure of the pathogen in question, the data and results should always be interpreted in the correct epidemiological context. This means that additional information relating to the data, such as time of sampling and origin of the sample is of paramount importance in order to draw conclusions and interpret the attribution results (EFSA Biohaz Panel, 2013).

7.2 Types of models

The most commonly used source attribution models employing subtyping data can be divided into frequency-matching models and population genetic models.

7.2.1 Frequency-matching models

Frequency-matching models compare distributions of pathogen subtypes between human and source isolates through mathematical models that can infer probabilistically the most likely sources of the human infections, assuming a unidirectional transmission pathway from sources to humans (Mughini-Gras et al., 2017). These models typically also incorporate epidemiological data on the human cases (e.g. travel history), and data on prevalence, food consumption, import-export flows, etc. as to better inform the attributions (de Knegt et al., 2015).

Based on years of surveillance data, it is clear, that within the same pathogen species, subtypes differ in their ability to cause disease in humans, often leading to different levels of severity (Pires and Hald, 2010; EFSA Panel on Biological Hazards, 2012). In addition, the various sources may have different impact in the human population making it difficult to identify a linear relationship between the occurrence (prevalence) of a particular subtype in various sources and the occurrence of reported human cases (Hald et al., 2004). Models based on frequency matching should therefore be able to account for variations between pathogen subtypes and sources.

Using data from the integrated Danish *Salmonella* surveillance program, a stochastic Bayesian model was developed to quantify the contribution of each of the major food animal sources to human *Salmonella* infections

(Hald et al., 2004; Hald et al., 2007; Pires and Hald, 2010; de Knegt et al., 2016). This model attributes domestically acquired laboratory-confirmed human *Salmonella* infections caused by different *Salmonella* subtypes as a function of the prevalence of these subtypes in animal and food sources and the amount of each food source available for consumption. The model compares impact across subtypes and is, through the Bayesian approach, able to estimate the relative impact of *Salmonella* subtypes and the included sources (Pires & Hald, 2010). However, these 'impact parameters' are difficult to interpret and can best be described as multiplication factors that help the model to arrive at the most likely solution given the observed data (Hald et al., 2004; David et al., 2012; de Knegt et al., 2016).

The Danish model has been adapted to attribute human salmonellosis in other EU countries (Pires et al., 2008; Wahlström et al., 2010; Valkenburgh et al., 2007; David et al., 2013), in EU as a whole (de Knegt et al., 2015), in the United States (Guo et al., 2011), New Zealand (Mullner et al., 2009), and Japan (Toyofuku et al., 2011), as well as for attribution of other food-borne pathogens e.g. *L. monocytogenes* (Little et al., 2010) and *Campylobacter* (Mullner et al., 2009; Boysen et al., 2013).

Frequency-matched models can employ both phenotypic and genotypic data. In fact, subtypes can be defined through any combination of phenotypic and/or genotypic data. Models developed for *Salmonella* have primarily used phenotypic data (serotyping, phage typing and antimicrobial susceptibility testing), but also the usefulness of MLVA-typing has been investigated. In a recent paper by de Knegt et al. (2016), adjustment of the discriminatory level of the MLVA typing for *Salmonella* source attribution was applied and discussed. Molecular methods have also been applied for *Campylobacter* (MLST) (Mullner et al., 2009; Boysen et al., 2013), whereas for Listeria monocytogenes a combination of phenotyping (serotyping) and molecular typing (AFLP) was used (Little et al., 2010). Overall, however, the value of applying WGS data in frequency-based source attribution models still needs to be explored.

7.2.2. Population genetic models

Driven by the latest years of development in high-throughput DNA sequencing techniques, a whole new set of tools have emerged. Common for them all, are that they compare pathogen strains from different sources by investigating how closely they are related and how they may have evolved from each other (EFSA Biohaz Panel, 2013). Some of these methods directly provide attribution estimates, where a number or proportion of human cases is attributed to specific source. Among these are the Bayesian clustering algorithm STRUCTURE (Pritchard et al., 2000) and the Asymmetric Island Model (AIM) (Wilson et al., 2008), both developed for MLST data. Other methods are based on clustering techniques that visualise the relatedness of bacterial strains using graphical representation e.g. the Minimum Spanning Trees (Feil et al., 2004; Spratt et al., 2004). Although visualisation tools do not provide attribution estimates, they do give an insight into the population structure of a pathogen and can support the conclusions drawn from the mathematical models.

Several studies show the value of source attribution based on molecular subtyping methods using probabilistic population genetics models such as STRUCTURE and the AIM. For *Campylobacter* for instance, it has been possible to identify some degree of host association between certain sequence types (ST) and a particular host reservoir despite the weakly clonal population structure of this pathogen (Dingle et al., 2001; McCarthy et al., 2007), resulting in attribution estimates (Wilson et al., 2008; Sheppard et al., 2009; Strachan et al., 2009; Mullner, 2009; Mughini-Gras et al., 2012; Boysen et al., 2013). The assumption is that the animal and environmental reservoirs of *Campylobacter* are separate populations within which the bacteria evolve through

mutation and horizontal gene transfer (recombination), and between which genes may flow (migrate). Based on the estimated amount of mutation, recombination and migration, each human case is assigned probabilistically to the source populations. From these individual probabilities, the total amount of human disease attributable to each source is estimated. Most of the studies so far have been on *Campylobacter*, but *Salmonella* data using MLVA typing have also recently been applied in the AIM (Mughini-Gras et al., 2014) and it is expected that the methods also will be applicable to other zoonotic pathogens such as *L. monocytogenes* and STEC as whole-genome sequencing become more widely used (EFSA Biohaz Panel, 2013).

7.3 Future attribution studies using WGS data and machine learning algorithms

With the high-throughput WGS techniques, a high level of standardisation and consequently meaningful comparison of results between technologies is expected. The most significant advantage of WGS compared to other methods is that the former results in a much higher data resolution. In fact, the application of WGS should in theory be able to provide us with all we need to know about a certain bacterial strain. However, the large number of variables in WGS datasets and the high data dimensionality challenges the currently available methods. Specifically for source attribution, research should focus on ways to define meaningful subtypes (or groups of features) that are able to distinguish between sources and can be used as input for the mathematical attribution models. Such subtypes could include one or more of the housekeeping genes currently used for MLST typing, but they could also be based on a whole new set of 'host-associated' genes/features identified from the WGS data analysed using for instance wgMLST or k-mer.

Recently, machine learning algorithms (e.g. random forest, neural networks, Adaboost) have shown huge potential for identifying relevant "features" in WGS data, thereby enabling the making of strong predictions (Davis et al., 2016). The hypothesis is that a machine learning model can recognize particular features/instances based on the strain sequences used as data inputs. For source attribution, it would be relevant to use supervised learning, where sequences from strains of known origin (e.g. animals or food) are annotated as such. The algorithm will then try to identify 'host-associated' features in order to predict the origin of the sequence. The models typically use part of the whole dataset as a training dataset and another part for testing model accuracy, as well as other performance measure such as positive and negative predictive power. This is followed by the use of the algorithm itself to predict features in an unknown sample based only on the sequence i.e. for source attribution to predict the source origin of strains isolated from humans. The random forest algorithm is considered particularly suitable for WGS datasets that may include hundreds or thousands of relevant features, but where each feature may contain only a small amount of information (Breiman, 2001; Davis et al., 2016).

Analysis of WGS data may also provide the information needed to quantify the difference between various subtypes/strains with regard to causing human illness and thereby assist with the characterisation of 'pathotypes' (Brul et al., 2012). This has been emphasised for Shigatoxigenic *E. Coli,* where MLST combined with the determination of virulence genes provides better insight than MLST alone into identifying source strains significant for human disease (Hauser et al., 2013; Ji et al., 2010). Virulence genes are therefore considered important to include in subtypes used for STEC risk assessment and source attribution. Hence, for both source attribution and risk assessment, novel approaches that are able to consider both genetic and functional relationship between strains would be very useful, especially if the functional traits relate to factors important for human infections such as virulence, antimicrobial resistance and survivability (e.g. acid tolerance) (EFSA Biohaz Panel, 2013). As part of COMPARE WP1, a random forest machine learning approach that uses WGS data from pathogens/potential pathogens in food and returns an estimate of the resulting risk/health burden at the

population level is currently being developed using *L. monocytogenes* as a case study (Njage et al., in prep.). If successful, this approach could be of great value for both future microbial risk assessment and source attribution.

References

Barco L, Barrucci F, Olsen JE and Ricci A, 2013. *Salmonella* source attribution based on microbial subtyping. Int J Food Microbiol, 163, 193-203.

Batz MB, Doyle MP, Morris G, Jr., Painter J, Singh R, Tauxe RV, Taylor MR and Lo Fo Wong DM, 2005. Attributing illness to food. Emerg Infect Dis, 11, 993-999.

Boysen L, Rosenquist H, Larsson JT, Nielsen EM, Sorensen G, Nordentoft S and Hald T, 2013. Source attribution of human campylobacteriosis in Denmark. Epidemiol Infect, 1-10.

Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

Brul, S., Bassett, J., Cook, P., Kathariou, S., McClure, P., Jasti, P. R., & Betts, R., 2012. 'Omics' technologies in quantitative microbial risk assessment. Trends in Food Science and Technology, 27(1), 12–24. doi:10.1016/j.tifs.2012.04.004

David JM, Guillemot D, Bemrah N, Thebault A, Brisabois A, Chemaly M, Weill FX, Sanders P and Watier L, 2012. The Bayesian microbial subtyping attribution model: robustness to prior information and a proposition. Risk Anal, 33, 397-408.

David JM, Sanders P, Bemrah N, Granier SA, Denis M, Weill FX, Guillemot D and Watier L, 2013. Attribution of the French human Salmonellosis cases to the main food-sources according to the type of surveillance data. Prev Vet Med, 110, 12-27.

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., … Stevens, R., 2016. Antimicrobial Resistance Prediction in PATRIC and RAST. Scientific Reports, 6(1), 27930. doi:10.1038/srep27930

de Knegt LV, Pires SM, Hald T, 2015. Attributing foodborne salmonellosis in humans to animal reservoirs in the European Union using a multi-country stochastic model. Epidemiol. Infect., 143, 1175–1186. doi:10.1017/S0950268814001903

de Knegt, L. V., Pires, S. M., Löfström, C., Sørensen, G., Pedersen, K., Torpdahl, M., Nielsen, E. M., Hald, T., 2016. Application of Molecular Typing Results in Source Attribution Models: The Case of Multiple Locus Variable Number Tandem Repeat Analysis (MLVA) of *Salmonella* Isolates Obtained from Integrated Surveillance in Denmark. Risk Analysis, 36(3), p.571-588. http://dx.doi.org/10.1111/risa.12483

Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJ, Urwin R and Maiden MC, 2001. Multilocus sequence typing system for Campylobacter jejuni. J Clin Microbiol, 39, 14-23.

Domingues AR, Pires SM, Halasa T, Hald T, 2012.Source attribution of human campylobacteriosis using a meta-analysis of case-control studies of sporadic infections. Epidemiology and Infection, 140(6), 970–981. doi:10.1017/S0950268811002676

EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), 2013. Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 1 (evaluation of methods and applications). EFSA Journal 2013;11(12):3502, 84 pp. doi:10.2903/j.efsa.2013.3502.

Feil EJ, Li BC, Aanensen DM, Hanage WP and Spratt BG, 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol, 186, 1518-1530.

Guo C, Hoekstra RM, Schroeder CM, Pires SM, Ong KL, Hartnett E, Naugle A, Harman J, Bennett P, Cieslak P, Scallan E, Rose B, Holt KG, Kissler B, Mbandi E, Roodsari R, Angulo FJ and Cole D, 2011. Application of Bayesian techniques to model the burden of human salmonellosis attributable to U.S. food commodities at the point of processing: adaptation of a Danish model. Foodborne Pathog Dis, 8, 509-516.

Hald T, D Vose, HC Wegener, T Koupeev, 2004. A Bayesian Approach to Quantify the Contribution of Animal-Food Sources to Human Salmonellosis. Risk Analysis Feb, 24(1): 251-65.

Hald T, Lo Fo Wong DMA, Aarestrup FM, 2007. The attribution of human infections with antimicrobial resistant *Salmonella* bacteria in Denmark to sources of animal origin. Foodborne Pathogens and Disease 4(3): 313-326.

Hald T and S M Pires, 2011. “Attributing the burden of foodborne disease to specific sources of infection” in Tracing pathogens in the food chain. Eds. S Brul, P M Fratamico and T A McMeekin. Woodhead Publishing Series in Food Science, Technology and Nutrition No. 196. ISBN 1 84569 496 1.

Hald T, Aspinall W, Devleesschauwer B, Cooke R, Corrigan T, Havelaar AH, et al., 2016. World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation. PLoS ONE 11(1): e0145839. doi:10.1371/ journal.pone.0145839.

Hauser E, Mellmann A, Semmler T, Stoeber H, Wieler LH, Karch H, Kuebler N, Fruth A, Harmsen D, Weniger T, Tietze E and Schmidt H, 2013. Phylogenetic and molecular analysis of food-borne shiga toxin-producing Escherichia coli. Appl Environ Microbiol, 79, 2731-2740.

Havelaar AH, Braunig J, Christiansen K, Cornu M, Hald T, Mangen MJ, Molbak K, Pielaat A, Snary E, Van Pelt W, Velthuis A and Wahlstrom H, 2007. Towards an integrated approach in supporting microbiological food safety decisions. Zoonoses Public Health, 54, 103-117.

Ji XW, Liao YL, Zhu YF, Wang HG, Gu L, Gu J, Dong C, Ding HL, Mao XH, Zhu FC and Zou QM, 2010. Multilocus sequence typing and virulence factors analysis of Escherichia coli O157 strains in China. J Microbiol, 48, 849-855.

Little CL, Pires SM, Gillespie IA, Grant K and Nichols GL, 2010. Attribution of human Listeria monocytogenes infections in England and Wales to ready-to-eat food sources placed on the market: adaptation of the Hald *Salmonella* source attribution model. Foodborne Pathog Dis, 7, 749-756.

McCarthy ND, Colles FM, Dingle KE, Bagnall MC, Manning G, Maiden MC and Falush D, 2007. Host-associated genetic import in Campylobacter jejuni. Emerg Infect Dis, 13, 267-272.

Mather AE, Reid SW, Maskell DJ, Parkhill J, Fookes MC, Harris SR, Brown DJ, Coia JE, Mulvey MR, Gilmour MW, Petrovska L, de Pinna E, Kuroda M, Akiba M, Izumiya H, Connor TR, Suchard MA, Lemey P, Mellor DJ, Haydon DT and Thomson NR, 2013. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. Science, 341, 1514-1517.

Mughini Gras L, Smid JH, Wagenaar JA, de Boer AG, Havelaar AH, Friesema IH, French NP, Busani L and van Pelt W, 2012. Risk factors for campylobacteriosis of chicken, ruminant, and environmental origin: a combined case-control and source attribution analysis. PLoS One, 7, e42599.

Mughini-Gras, L., Smid, J., Enserink, R., Franz, E., Schouls, L., Heck, M., & van Pelt, W., 2014. Tracing the sources of human salmonellosis: A multi-model comparison of phenotyping and genotyping methods. Infection Genetics and Evolution, 28, 251–260. doi:10.1016/j.meegid.2014.10.003

Mughini-Gras, L., Franz, E., & van Pelt, W., 2017. New paradigms for *Salmonella* source attribution based on microbial subtyping. Food Microbiology. doi:10.1016/j.fm.2017.03.002

Mullner P, Jones G, Noble A, Spencer SE, Hathaway S and French NP, 2009. Source attribution of food-borne zoonoses in New Zealand: a modified Hald model. Risk Anal, 29, 970-984.

Njage, PMK., Henry, C., Leekitcharoenphon, P., Hendriksen, RS., Hald, T., in prep. Machine learning methods as a tool for predicting risk of illness applying next generation sequencing data.

Pintar, K.D., Thomas, K.M., Christidis, T., Otten, A., Nesbitt, A., Marshall, B., Pollari, F., Hurst, M., Ravel, A., 2016. A comparative exposure assessment of Campylobacter in Ontario, Canada. Risk Anal. http://dx.doi.org/10.1111/risa.12653.

Pires SM, Kaesboher A, Spitznagel H, Whalstrom H, Nichols G, David J, Van Pelt W, Baumann A and T H, 2008. *Salmonella* source attribution in different European countries. Proceedings in FoodMicro 2008, Aberdeen, Scotland. Available online at: http://aberdeen.conference-services.net/programme.asp?conferenceID=1143&action=prog_categories (last accessed on 12/12/2013).

Pires SM, Evers EG, van Pelt W, Ayers T, Scallan E, Angulo FJ, Havelaar A and Hald T, 2009. Attributing the human disease burden of foodborne infections to specific sources. Foodborne Pathog Dis, 6, 417-424.

Pires SM, Hald T, 2010. Assessing the Differences in Public Health Impact of *Salmonella* Subtypes Using a Bayesian Microbial Subtyping Approach for Source Attribution. Foodborne Pathog Dis. 7(2): 143-151.

Pires, SM, Vieira, A, Perez, E, Wong, DLF, Hald, T, 2012. Attributing human foodborne illness to food sources and water in Latin America and the Caribbean using data from outbreak investigations. International Journal of Food Microbiology, 152 (3): 129–138.

Pires, S. M., 2013. Assessing the Applicability of Currently Available Methods for Attributing Foodborne Disease to Sources, Including Food and Food Commodities. Foodborne Pathogens and Disease, 10(3), 206–213. doi:10.1089/fpd.2012.1134

Pritchard JK, Stephens M and Donnelly P, 2000. Inference of population structure using multilocus genotype data. Genetics, 155, 945-959.

Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID, Maiden MC and Forbes KJ, 2009. Campylobacter genotyping to determine the source of human infection. Clin Infect Dis, 48, 1072-1078.

Smid JH, Mughini Gras L, de Boer AG, French NP, Havelaar AH, Wagenaar JA and van Pelt W, 2013. Practicalities of using non-local or non-recent multilocus sequence typing data for source attribution in space and time of human campylobacteriosis. PLoS One, 8, e55029.

Son I, Zheng J, Keys CE, Zhao S, Meng J and Brown EW, 2013. Analysis of pulsed field gel electrophoresis profiles using multiple enzymes for predicting potential source reservoirs for strains of *Salmonella* Enteritidis and *Salmonella* Typhimurium isolated from humans. Infect Genet Evol, 16, 226-233.

Spratt BG, Hanage WP, Li B, Aanensen DM and Feil EJ, 2004. Displaying the relatedness among isolates of bacterial species -- the eBURST approach. FEMS Microbiol Lett, 241, 129-134.

Strachan NJ, Gormley FJ, Rotariu O, Ogden ID, Miller G, Dunn GM, Sheppard SK, Dallas JF, Reid TM, Howie H, Maiden MC and Forbes KJ, 2009. Attribution of Campylobacter infections in northeast Scotland to specific sources by use of multilocus sequence typing. J Infect Dis, 199, 1205-1208.

Toyofuku H, Pires SM and Hald T, 2011. *Salmonella* source attribution in Japan by a microbial subtyping approach. EcoHealth (ISSN: 1612-9202), 7 (Suppl.1): S22-S23.

Tustin, J., Laberge, K., Michel, P., Reiersen, J., Dadadottir, S., Briem, H., Hardardottir, H., Kristinsson, K., Gunnarsson, E., Fridriksdottir, V., Georgsson, F., 2011. A national epidemic of campylobacteriosis in Iceland, lessons learned. Zoonoses Public Health 58, 440e447.

Valkenburgh S, van Oosterom R, Stenvers O, Aalten M, Braks M, Schimmer B, Van De Giessen AW, Van Pelt W and Langelaar M, 2007. Zoonoses and zoonotic agents in humans, food, animals and feed in The Netherlands 2003-2006. Available online at: http://www.rivm.nl/bibliotheek/rapporten/330152001.pdf (last accessed on 13/12/2013).

Wahlstrom H, Andersson Y, Plym-Forshell L and Pires SM, 2010. Source attribution of human *Salmonella* cases in Sweden. Epidemiol Infect, 139, 1246-1253.

Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart CA and Diggle PJ, 2008. Tracing the source of campylobacteriosis. PLoS Genet, 4, e1000203.

8. Conclusion

This document has presented analysis algorithms, related to methods and analysis strategies, associated with WGS and relevant for disease surveillance and public health epidemiology. In many countries, routine microbiological surveillance of foodborne diseases is undergoing profound changes in recent years as WGS is being incorporated and therefore this whole area of work is developing rapidly. The use of WGS has already provided breakthroughs for outbreak detection and investigations. An example hereof is the application within listeria-epidemiology in Denmark (Chapter 5) which led to a number of outbreaks being found and solved in a manner not possible before the introduction of WGS. Apart from being a valuable tool for outbreak investigations, the public health application of WGS also promises to give a deeper insight into the epidemiology of the diseases, including sporadic disease, as seen with the attempts to use the methodology for source attribution (Chapter 7). Similarly, application within veterinary science for production-animal disease surveillance will provide a better understanding of factors affection colonization and transmission routes. These will all, hopefully, translate into development of areas of intervention or targeted programs, which will work to reduce the number of foodborne infections.

Surveillance of foodborne disease will encompass an analysis of the data provided by the WGS methods, but also of 'classical' epidemiological data, time-place-person data. These are sometimes in a WGS context referred to as Metadata. Analysis methods useful for public health surveillance will therefore need to be able to handle both types of data. Chapter 3 describes a series of relevant statistical cluster detection methods developed for use with pre-WGS diagnostics and typing. They are developed to find temporal and geographical clusters and generally rely on historical data from a stable surveillance system for comparison. It is clear that these methods are still relevant in a WGS world, the central question still being if clusters exist in time and space. It is also clear that the presented methods are challenged by the level of detail, which WGS provides, and in particular by the lack of a stable historical baseline. More work to develop and calibrate these methods is therefore needed. The integration of metadata and WGS data is also unresolved when it comes to sharing of data across borders. Whereas the Compare project envisions a future where WGS data flow freely, this may not be possible with metadata for reasons of patient confidentiality or economical concerns within the food and veterinary fields. Therefore, analysis methods also need to consider this aspect. A proposed soon-to-begin cooperation between Compare and ECDC concerning European WGS surveillance data of selected foodborne infections will help resolve these matters. Also, to work with data, there will be a requirement for visualization tools, on-line epidemiological 'dashboards' showing sequence information (possibly involving phylogenetic presentations) combined with time-place-person data in an epidemiological relevant format. This aspect has not been treated in this document, but is frequently mentioned as a need by epidemiologists.

In order to work with routinely generated WGS surveillance data in a public health context, the sequence data will in general need to be digested and interpreted to reach an 'entity', so that a strain can be compared to other strains (Chapters 2 and 4). A number of possibilities to do this exist (see example-boxes in Chapter 2) and the choice hereof will depend on the specific surveillance situation. For the foodborne bacterial infections, public health institutes are increasingly realising that a detailed comparison of sequence information of a large number of strains is impractical and the focus is turning towards a uniform use of cg- or wg-MLST protocols (see Nadon et al, 2017, referenced in Chapter 4). Among the benefits of such a strategy is that the method will be (at least partly) generic for different organisms and that the MLSTs will produce a strain 'name' (an ST). This can be easily communicated and it will be sufficient to do the (local/national) analysis using this name only. Moreover as the

nomenclature is hierarchical, analyses can be performed at different levels of sensitivity and there is less need for complex phylogenetic analyses.

List of Figures, Tables and Boxes