Bundesinstitut für Risikobewertung

DOI 10.17590/20191212-142336

Comparing the genetic material of pathogens to explain disease outbreaks

BfR Opinion No 047/2019 of 28 November 2019

Food can be contaminated with disease-causing pathogens such as bacteria, viruses or parasites. The worldwide trade of food products means that impurities can lead to disease outbreaks in locations that are far away from each other. The responsible regulatory authorities use molecular laboratory methods to promptly detect causes and potential links between incidences of disease. These are also continuously developed further through the application of them.

The German Federal Institute for Risk Assessment (BfR) has addressed the relevant questions for this topic and assessed to what extent the molecular methods of the Next Generation Sequencing procedure are suitable for detection of disease outbreaks and which data should also be included. This assessment primarily serves all authorities in the sector of public health, food safety and veterinary medicine for them to decide which method is suitable for detecting outbreaks.

“Next Generation Sequencing (NGS)” represents the second and third generations of sequencing of genetic material and offers the highest possible resolving to determine the nucleotide sequence of a DNA molecule or genome. The costs of the methods have significantly decreased in recent years.

Two methods are currently used to examine disease outbreaks:
1) For easy to cultivate pathogens available in pure form, whole genome sequencing (WGS) has been established worldwide. The genetic substance of the causative bacteria is isolated from the patient and compared to isolates of the same pathogen from food. This makes it possible to identify and evaluate the smallest differences in genetic material.
2) In contrast, in whole metagenome sequencing (WMS), genetic material is directly extracted from a food sample which often contains various bacteria. Microorganisms such as parasites or viruses which are difficult to cultivate or cannot be cultivated can be detected. Whole metagenome sequencing is suitable as a method for initial diagnosis if no specific pathogen is suspected.

To explain foodborne disease outbreaks, consulting epidemiological data on the examined pathogens may be necessary depending on the applied analytical method. With this data, it can be assessed as to whether a verified pathogen belongs to a particular outbreak. In ideal cases, explaining transmission routes and the source of the impurity is possible.

1 Subject of the assessment/Introduction

Microorganisms are subject to constant changing of their characteristics over generations. Changes may be quicker or slower depending on many extrinsic and intrinsic factors. If the change offers the microorganism a survival benefit (for example adaptation to environmental conditions), the probability that it will be passed on to subsequent descendants and be manifested in the population of the microorganism is very high. However the change may also be a disadvantage for the cells so that they cannot spread further. These new properties gained from the course of evolution are not always visible. However the microorganism's change is stored in the nucleotide sequence of the genome. Studies on the rate of the average change

of a bacteria genome, the mutation rate, assume around 1×10^{-3} mutations per genome per generation for *Escherichia coli*, for example (Lee et al., 2012). For the *Salmonella* serovar *S*. Choleraesuis var. Kunzendorf, the mutation rate was determined in relation to time with 1.02 bases per genome per year (Leekitcharoenphon et al. 2019). The rate for *S*. Enteritidis was determined in a similar way in one study with 1.01 bases per year (Deng et al., 2014).

Molecular typing methods are important tools to identify differences in the genetic material of an isolate by comparing additional isolates from the same species. Detecting such differences in a bacteria population which have occurred in brief periods (of a few years) therefore requires highly discriminatory typing methods. In the context of disease outbreaks, an outbreak strain usually spreads through direct contact with people or through intake of the bacteria via contaminated food. In unfavourable cases, the disease spreads over larger areas in the form of several clusters of outbreaks. Worldwide food trade encourages this sort of spreading. In the past, multilocus variable-number tandem repeat analysis (MLVA) and pulsed-field gel electrophoresis (PFGE) were used in outbreak investigations to compare suspected interrelated isolates. However these methods only have a limited resolving power so that a combination of typing methods is often used in outbreak investigations.

DNA sequencing means determining the nucleotide sequence in a DNA molecule. The first method which could be used to determine such a sequence was developed by Sanger at the end of the 1970s. A limited number of short DNA sections could be sequenced with this. However this method is not suitable for fully and cost efficiently sequencing genomes in a short period of time. With the increasing importance of sequencing in research and diagnostics, the next generation (second generation) of sequencing technology hit the market in the middle of the 2000s. With the devices developed for this, large-scale parallel sequencing of short nucleotide sequences of a DNA molecule could be performed. Using bioinformatic approaches makes it possible to determine the nucleotide sequence of the whole genome (for example *E. coli* genome of 5 megabases). There is now another sequencing technology available (the third generation) which can sequence individual molecules. This means that the polymerase chain reaction step which was still used in the second generation is no longer required. Third generation sequencers make it possible to sequence very long DNA molecules (ca. 10-50 kilobases) at once (Ronholm et al., 2016).

Next Generation Sequencing (NGS), the second and third generation of sequencing, offers the best possible resolving power for determining the nucleotide sequence of a DNA molecule or genome. The costs of NGS have reduced so much in recent years that the use of the technology for authorities in the area of public health, veterinary medicine and food safety has become affordable. This led to the use of NGS worldwide to determine the relationships between isolates as part of investigations into disease outbreaks or the monitoring of contamination in food production. The combination of this sequence comparison with relevant epidemiological information, and ideally with the evaluation of the product chain information, makes it possible to confidently explain the transmission routes and source. The success of an outbreak investigation therefore requires the continuous cooperation of microbiologists and epidemiologists to bring together the assessment of genomic and epidemiological proof for a certain result (World Health Organisation, 2018).

The German Federal Institute for Risk Assessment (BfR) has addressed the following questions on this topic:

1. NGS methods

a.) Are all NGS methods similarly suitable for examination as part of disease outbreaks?

b.) Which advantages and disadvantages of individual technical variants are important for investigating sample material as part of disease outbreaks?
c.) Which quality assurance measures are subject to applied NGS methods (for example as part of ring trials)?

2. Analytical results
a.) Which degree of similarity is required for results of investigations with NGS methods to assign them to specific disease outbreaks, for example clusters?
b.) How high is the explanatory power of the various techniques of the NGS procedure?
c.) Which information is required to consistently characterise a cluster/an identical molecular biological subtype?

3. Data on the examination material
a.) Which operating or product-related information is required or gathered in conjunction with sent food isolates and typing results (from sample material from the German federal states ["Laender"] responsible for food monitoring), for which purpose and how are these assigned?
b.) How and by which body should this information be used in the event of an outbreak?

The questions are answered as follows:

1. NGS method:
a.) Are all NGS methods similarly suitable for examination as part of disease outbreaks?

NGS methods permit large-scale determination of nucleotide sequences from genetic material in a sample. They usually combine (i) extraction of a nucleic acid from a sample, (ii) creation of libraries, (iii) large-scale sequencing of nucleic acid and (iv) bioinformatic evaluation of the raw sequencing parts. In principle, a distinction is made between DNA or RNA, which has been extracted from a pure culture (for example isolate) or from a complex compilation of various organisms (for example a food sample).

Two methods are currently possible for examining disease outbreaks, however their use depends on the initial situation of the occurrence. The method applied for easy-to-cultivate pathogens is whole genome sequencing (WGS). For this, the genome that comes from a bacteria isolate present in pure form is sequenced and then compared to additional genomes of suspected isolates sequenced in the same way. This approach means that individual nucleotide differences between genomes can be determined across larger sequence segments. Due to its precision and high resolving power, WGS has established itself worldwide for investigating relevant and easy-to-cultivate pathogens as part of the explanation of disease outbreaks.

An additional approach for investigation as part of disease outbreaks is whole metagenome sequencing (WMS). This is a method in which the DNA from a sample, such as for example from food, is extracted and this is subsequently analysed using NGS. The isolation and cultivation of the pathogen, which are often time-consuming, are avoided in this method. In addition to the time benefit, this also permits detection of microorganisms such as parasites or viruses which are difficult to cultivate or for which cultivation is not possible. However since pathogens are not cultivated before they are sequenced for WMS, no conclusions can be drawn on the viability and proliferation ability of the pathogen. What's more, the procedure also requires laborious bioinformatic analyses since complex sequence mixes need to be assigned to the individual pathogens. Furthermore, low coverage of the sequence can make comparison between samples difficult. The procedure is therefore suitable as a universal and

undirected method in the absence of a specific suspected case for preliminary diagnosis, and only under certain conditions for investigating disease outbreaks.

b.) Which advantages and disadvantages of individual technical variants are important for investigating sample material as part of disease outbreaks?

As described above, both second and third generation technologies fall under NGS technologies. For isolate sequencing of outbreak-relevant samples, the second generation of sequencing is used in particular, since this has an extremely high throughput as well as extremely low sequencing error rates. Sequence differences between various isolates are identified to detect outbreak clusters. It is therefore crucial that potential sequencing errors caused by technology can be differentiated from real biological signals. This is only possible if sequencing errors occur by chance and without an identifiable pattern since they can be relatively reliably rectified using statistical methods. The only technology which does not generate systematic errors is currently the *sequencing-by-synthesis* method (Cao et al., 2017). An additional sequencing technology, the *semi-conductor sequencing* method generates systematic error profiles which makes direct comparison with the results from sequencing from the *sequencing-by-synthesis* method difficult. Since the *sequencing-by-synthesis* method is now common in whole genome sequencing worldwide, sequence results for the investigation of outbreaks can be reliably compared to this procedure.

The “NGS bacteria characterisation” working group of the § 64 German Food and Feed Code (LFGB) was recently established. It is coordinated by the German Federal Office for Consumer Protection and Food Safety (BVL). The aim is to provide authorities entrusted with food monitoring with validated, effective and standardised methods of the NGS procedure for bacterial pathogens from foodborne outbreak investigations. One activity of this working group will be the comparability of different second generation sequencing methods when holding outbreak investigations.

In relation to the speed of the sequencing, both technologies do not take much. Essentially, the duration of the process depends on the length of the DNA molecules to be sequenced. The generated sequence lengths are, depending on the kit used, between 2x 75 bp and 2x 300 bp with the *sequencing-by-synthesis* method and between 200 and 400 bp with the *semi-conductor* method. However the additional preparation steps of library creation mentioned above must also be included in the process. The experience of the BfR with the *sequencing-by-synthesis* methods shows that the processing duration for a sequencing of short sequence parts (2x 75 bp) takes ca. 2 days from starting with the available DNA until the result to be evaluated is present. Creating longer sequences which may have an effect on the number of contigs (number of sequence parts compiled using bioinformatics) requires around 4 days of processing time. Cultivation of the pathogen for DNA isolation is not included in this calculation and may require a further 2 to 4 days depending on the pathogen.

The third generation of sequencing is able to generate fewer but extremely long sequences and can be used to create reference genomes for individual outbreaks. The error rates of the third generation technologies are so comparably high that these are not yet suitable for isolate sequence comparisons as part of investigations for disease outbreaks. However they are very useful for creating closed reference isolate sequences which can be combined with second generation sequence data and result in a very specific genome sequence. Reference sequences of isolates are required for bioinformatic evaluation. Numerous good reference sequences are therefore available for most relevant pathogens.

c.) Which quality assurance measures are subject to applied NGS methods (for example as part of ring trials)?

NGS methods are not currently accredited at the BfR due to their complexity. Various quality-assurance measures are implemented during library preparation, the sequencing process on the device and the bioinformatic evaluation.
Quality assurance measures during library preparation include both the concentration examination of the DNA and a determination of the DNA fragment lengths. This data is an important prerequisite for sequencing to be performed in a high-quality way on the device. In the subsequent bioinformatic evaluation, the raw data created for each sequenced isolate is trimmed (bases with poor quality are filtered out) and assembled (numerous shorter sequences are put together into longer sequence parts using bioinformatics). The quality inspection is performed using an automatically created Assembly Report, which lists the significant quality parameters which are typically used to be able to estimate the quality of sequence data. The following are used for inspection: the data of the predicted species, the data amounts or coverage of the assemblies (this provides information on how frequently, on average, each nucleotide of the genome was sequenced), the number of sequence parts (contigs) as well as the length of the whole assemblies, which should correspond to the expected length of the genome (this provides information on any contamination). The limit values applied are used to determine whether the quality of sequence data is sufficient. They are based firstly on specific past experience and secondly on exchange of knowledge with scientists from the National Food Institute of the Technical University of Denmark (DTU-Food), from the U. S. Food and Drug Administration (FDA), and information from the literature.

Each larger (for example changing the sequencing kit) or smaller (for example shortening the number of sequencing cycles) adjustment in the flow chart is validated and verified in the laboratory and using bioinformatics. Control isolates are used as part of this to estimate whether key bioinformatic values (coverage, number of contigs, assembly, length) are changing beyond the acceptable degree or whether sequence differences are observed with the originally applied protocol.

The BfR regularly participates in WGS ring trials to further assure the quality of the practical laboratory work and the bioinformatic analyses too. This includes the following performance comparison tests:

- Global Microbial Identifier Initiative (Participation in 2015, 2016 and 2017)
- Various ring trials as part of the ENGAGE (Establishing Next Generation sequencing Ability for Genomic analysis in Europe) and COMPARE (Collaborative management platform for detection and analyses of (re-) emerging and foodborne outbreaks in Europe) projects
- Sequencing comparison attempts of defined isolates with the RKI
- UNSGM bioinformatic ring trial (Practical Exercise in Support of the United Nations Secretary-General's Mechanism for Investigation of Alleged Use of Biological Weapons, with Special Consideration of the Functional Subunit Approach) for detecting pathogenicity factors in sequence data

The BfR is seeking the accreditation of the NGS method for the sequencing of pathogen isolates as part of disease outbreaks. Preparatory measures for this have been started. The "NGS bacteria characterisation" working group of the § 64 German Food and Feed Code

(LFGB) was recently established by the German Federal Office for Consumer Protection and Food Safety (BVL). The BfR has taken over the chair function and is actively incorporating its expertise into the standardisation process. For this, ring trials will be organised and implemented with the members of the working groups. § 64 LFGB working groups are established by the BVL and used to provide the authorities entrusted with food monitoring with validated, efficient and standardised methods as part of the "Statutory Compilation of Methods for Sampling and Investigating Food" (ASU).

2. Analytical results:
When answering questions on various bioinformatic analysis options, the BfR refers to those used in outbreak investigations of isolate sequences using WGS. The answer to question 2b comes first to aid understanding.

b.) How high is the explanatory power of the various techniques of the NGS method?

Several approaches are currently used for analysing WGS results in outbreak investigations: (i) core genome MLST (cgMLST), considering thousands of genes which are available in most isolates of a species or genus, (ii) whole genome MLST (wgMLST) considering all genes including the variable additional genes of a species and (iii) the reference mapping of high-quality Single-Nucleotide Polymorphism (SNP) based Clustering Pipelines (Kovac et al., 2017). CgMLST/wgMLST are based on the MLST concept. In this, DNA sequence variations are determined in a set of fixed genes. These are distributed throughout the genome and show specific variations due to mutations or recombinations despite preservation. The typing of the isolate is done using the specific allele profile of the genes, which is then reflected in its corresponding sequence type (ST). While MLST is used with an analysis based on *de novo* assembly (gene-by-gene approximation), the SNP analysis is based on a comparison of individual point mutations of the isolate sequence for a pre-set, closely-related reference genome. This type of analysis is called "reference-based mapping". No genes to be analysed are set in this procedure. SNPs can therefore also be taken into account outside of gene-coded sequences.

While all differences in a set which occur for the reference genome are taken into account in SNP-based clustering methods, only allele differences without a distinction in the number and type of mutations between the isolates to be compared are taken into account for cgMLST/wgMLST. This can lead to a higher resolving power being achieved with the SNP analysis than with the cgMLST/wgMLST analysis. The great benefit of cgMLST/wgMLST is the application of nomenclature schemes which make it possible to create clear identification for the sequence to be investigated. It is easy to communicate without needing to share raw sequences. CgMLST/wgMLST also require less effort in terms of computer power compared to SNP-based procedures. The various bioinformatic options which assess the quality of the SNPs and then make selections which are considered in an SNP analysis mean that the raw sequences of all isolates to be compared must always be present. Using SNP analyses is appropriate for relatively closely related genomes, i.e. within a serovar or MLST group. In principle, results are only comparable if they have been generated with the same or equivalent software programmes or identical software algorithms. This applies to both methods (SNP – cgMLST/wgMLST). What's more, in SNP analyses the exact same references need to be used because otherwise a result may be different to the SNPs found. In contrast, a uniform scheme and also a centrally managed nomenclature for new allele numbers need to be used for cgMLST analyses. One aim is to have the latter but it is not currently available centrally. Only different isolated solutions are available until now. A central data analysis which exchanges raw sequencing data is therefore offered for correct assessment.

The validity of both cgMLST/wgMLST and SNP-based analyses is fundamentally very well suited for determining differences between isolates as part of outbreak investigations and leads to similar results. Both permit classification of groups of genomes into sub-types (cluster types) and an assessment of whether two genomes are (almost) identical in terms of molecular biology. Potential thresholds for the allele distance or SNP distance are nevertheless different in principle. Applying cgMLST is more suitable if several users need to systematically analyse each new isolate which is added to a common database at the same time (for example in the event of an outbreak), and in particular if the sequence information is not publicly accessible. For investigation of phylogeny, using cgMLST or SNP procedures can provide more robust analyses than wgMLST, since the latter only covers regions of genomes in which all strains are available. However wgMLST can lead to a higher resolution due to the additional consideration of the variable gene of a species.
SNP and cgMLST/wgMLST approaches assess genetic variations in somewhat different ways and should be seen as complementary. In particular, if one method alone cannot provide a clear answer, both analytical methods should be performed to obtain a better assessment.

Independently of the analytical method, incorporating epidemiological data from the suspected isolates is required for an explanation. This is only way that conclusions can be drawn on whether this gene belongs to an outbreak or to a cluster. Isolates compiled in a cluster do not necessarily belong to an outbreak if epidemiological data on the sequence comparison data contradicts this.

a.) Which degree of similarity is required for results of investigations with NGS methods to assign them to specific disease outbreaks, for example clusters?

c.) Which information is required to consistently characterise a cluster/an identical molecular biological subtype?

Questions 2a and 2c are answered together since there is a big overlap for this issue.

In the WGS analysis, the number of SNP or allele differences is used to construct phylogenetic trees which provide information on the evolutionary history of the isolate. From a biological point of view, a high sequence similarity shows through the WGS analysis that isolates have a recent common ancestor. In contrast, a low similarity means that it comes from an older common ancestor in the best case (Plightling et al., 2018). One fundamental assumption of molecular epidemiology is that phylogeny reflects epidemiological relationship, i.e. clinical isolates or food or environmental isolates which are closely related in terms of clinic and phylogenetic aspects are probably epidemiologically or causally linked (Besser et al., 2018). However this assumption does not always apply since complex or indirect links may be involved which can occur at any point along the food chain. Drawing epidemiological and food safety conclusions together is therefore decisive to achieve a coherent interpretation of the WGS analysis. The WGS analysis provides reliable evidence that isolates are genetically related but this does not necessarily mean that a clinical case is directly associated with a specific food. It is therefore essential to have epidemiological information available to support the phylogenetic results (Jagadeesan et al, 2019).

Due to the variety of types of bacteria, the different epidemiological contexts and different WGS analysis approaches, setting a threshold should be avoided in principle. For example, this applies for interpretation during an outbreak for both the cgMLST and the SNP analysis (Pightling et al., 2018; Schürch et al., 2018; Jagadeesan et al, 2019). Some species or serotypes are less diverse than others, for example *Salmonella* Enteritidis is relatively clonal (Al-

lard et al., 2013). Furthermore, the environment in which a species is located can apply evolutionary pressure which influences the mutation rate and generation time (Deatherage et al., 2017). The interpretation of the genetic relationships from strains which are based on SNP or allele differences must therefore be supplemented with expert knowledge on the respective pathogen, including understanding of its genetic diversity in the food chain and the representativeness of the investigated isolate (Besser et al, 2018; Schürch et al, 2018; Jagadeesan et al, 2019). The WGS analysis of each foodborne outbreak scenario must be assessed independently of the others, whereby epidemiological and investigations related to the food chain should be held to provide as much information as possible for interpretation.

For an initial rough estimation, it can be assumed that two pathogen isolates which have a difference of 0-20 SNP/Allele, for example, are considered closely related (Jagadeesan et al, 2019). They therefore probably have a recent common ancestor which comes from a common source. If such closely related isolates from different locations in a food production facility are isolated, the most likely scenario is that the same strain has spread within the production environment.

Various outbreaks have shown that isolates from ill people which fall under a threshold of a few SNPs in an SNP-based analysis exhibit a strong time-related correlation in respect to the date of occurrence of the symptoms (Deng et al, 2014). A study of seven different *S*. Enteritidis outbreaks showed a divergence of 3 SNPs within an outbreak and the nearest non-outbreak strains are differentiated by an average of 42.4 SNPs (Taylor et al, 2016). But the number of SNP differences between isolates may remain low over a longer outbreak period. In an outbreak study on *Listeria monocytogenes,* the variability of the isolate sequences within two outbreak clusters was 5 SNPs over a period of three years (Gillesberg Lassen, 2016). The data was confirmed in combination with epidemiological investigations in food companies and patient surveys.

If the sequences of two isolates are very different, for example > 50-100 SNPs or alleles, the isolates are generally not considered to be related. The probability that they do not come from the same source is therefore very high.

Isolates are not always within the thresholds mentioned above. For example, isolates can be differentiated by 30 SNPs/allele in a food processing facility but they belong to the same cluster compared to other isolates. This suggests that isolates have a common ancestor and have probably developed from a resilient strain which persists in the facility (Elson et al., 2019). This can occur if the quantity of microbial population experiences is greatly reduced, for example from disinfectants, since random mutations could lead to diversification of the predecessor strain (Jagadeesan et al, 2019).

However the thresholds mentioned above can also be exceeded by outbreaks associated with one source. In one case, three *Salmonella* serovars (*S*. Poona, *S*. Pomona and *S*. Sandiego) were simultaneously involved in exposure to salmonella through small turtles. The differences of the associated isolates were up to 17 SNPs for *S*. Poona and up to 30 SNPs for *S*. Pomona (https://www.cdc.gov/salmonella/small-turtles-03-12/epi.html) (Jagadeesan et al, 2019). In addition, 401 isolates associated with a multinational European outbreak of *S.* Enteritidis phage type 14b, with eggs as a source, showed a maximum difference of 23 SNP (Dallman et al., 2016).

3. Data on the investigation material:

a.) Which operating or product-related information is required or gathered in conjunction with submitted food isolates and typing results (from sample material from the German federal

states ["Laender"] responsible for food monitoring), for which purpose and how are these assigned?

To submit isolates from various matrices (food, feed, primary production, production environments and others, except for human isolates) for investigation in the German Microbiological National Reference Laboratories (NRL) and consiliary laboratories in the BfR, an electronic submission form in Excel file format is available on the homepage of the BfR. All senders are asked to complete the submission form as fully as possible and to send it to the BfR electronically by email and/or in paper form. The data is documented in the laboratory information system.

The following metadata is requested in the BfR submission form in text form and as ADV (the Working Committee of the Surveying Authorities of the States of the Federal Republic of Germany) codes: Sample number, German General Administrative Regulation (AVV) data number, preliminary pathogen finding, sampling date, isolation date, sampling location, matrix, processing condition, reason for sampling, operating type, German Livestock Movement Regulation number.

This information is only compulsory for samples from zoonosis monitoring. For all other samples, a minimum of the sample number, pathogen, sampling date and matrix must be sent to the BfR.

The metadata is gathered by the BfR for assignment of individual isolates of a pathogen as part of

- Outbreak investigations
- Epidemiological research
- Implementation of the directive 2003/99/EC on monitoring of zoonoses
- Zoonoses Control Regulation (EC) 2160/2003
- Collection of statistics in relation to the prevalence of pathogens in different matrices outside of the matters and programmes mentioned above.

The BfR only receives precise information on producers of samples which permit clean assignment of samples to individual companies in exceptional cases and on a voluntary basis. This important information for outbreak investigations usually remains with the German federal states ("Laender") and is only requested by BVL in the affected German federal states through the highest state authorities.

If the involved state authorities would like a goods flow analysis with traceability of suspected food as part of an outbreak investigation, the BfR also supports this if asked. It does this with data collection and analyses, visualisations and assessment of these using the FoodChain-Lab software developed for this purpose.

b.) How and by which body should this information be used in the event of an outbreak?

To hold an efficient investigation and to safely interpret the results of foodborne disease outbreaks, linking the available sequence data of suspected isolates to the epidemiological data is necessary. Sequencing and metadata should therefore be available in databases which are interlinked. Experts must perform the analysis of isolate sequences which determines the relationship of the isolates to be compared using bioinformatics. If isolates are classed as closely related, this forms the basis for the subsequent epidemiological investigations (for

example assignment of samples to individual companies, distributors, measures for control examinations in the companies, etc.).

The sequence comparison should be performed by an expert pathogen laboratory. This may be a laboratory which has obtained expertise for phylogenetic comparisons of the respective pathogen. A metadata set for an initial epidemiological assessment should be managed and run by an authorised body which has the necessary experience for this (transfer of data by the authorities). This initial assessment must not contain sensitive data on companies (for example the company name or location). Depending on the jurisdictions involves, these can remain in the respective German federal state ("Land") accordingly. They will only be incorporated into a more in-depth epidemiological investigation by the affected state if there is increased suspicion. The isolate source data should be provided in a public repository (for example ENA or NCBI) as soon as possible. This provides the basis for the exchange of sequence data and the possibility of performing additional local analyses under identical software conditions. This also secures the availability of sequences for investigations that are part of international outbreaks.

Further information on Next Generation Sequencing is available on the BfR website

Submission form for isolates and samples (Excel file format)
https://www.bfr.bund.de/de/einsendeformular_fuer_isolate_und_proben-9257.html

Research project: *Establishment Next Generation sequencing Ability for Genomic analysis of Bacterial Pathogens in Europe* (ENGAGE)
https://www.bfr.bund.de/en/new_approaches_in_identifying_and_characterising_microbiological_and_chemical_hazards__engage_-202739.html

External Next Generation Sequencing links

Global Microbial Identifier initiative to build a DNA genome database for identification and diagnosis of microbial pathogens, which the BfR also works with
https://www.globalmicrobialidentifier.org/

BfR "Opinions app"

2 References

Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW. 2013. On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. PLoS One 8(1), e55254.

Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. 2018. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect 24, 335–341.

Cao Y, Fanning S, Proos S, Jordan K, Srikumar S. 2017. A Review on the applications of Next Generation Sequencing technologies as applied to food-related microbiome studies Front Microbiol 8, 1829.

Dallman T, Inns T, Jombart T, Ashton P, Loman N, Chatt C, Messelhaeusser U, Rabsch W, Simon S, Nikisins S, Bernard H, le Hello S, Jourdan da-Silva N, Kornschober C, Mossong J, Hawkey P, de Pinna E, Grant K, Cleary P. 2016. Phylogenetic structure of European *Salmonella* Enteritidis outbreak correlates with national and international egg distribution network. Microb. Genom 2(8), e000070.

Deatherage DE, Kepner JL, Bennett AF, Lenski RE, Barrick JE. 2017. Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. Proc Natl Acad Sci Unit States 114(10):E1904 LP-E1912.

Deng X, Desai, PT, den Bakker, HC., Mikoleit M, Tolar B, Trees E, Hendriksen RS, Frye JG, Porwollik S, Weimer BC, Wiedmann M, Weinstock GM, Fields PI, McClelland M. 2014. Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. Emerg Infect Dis 20, 1481–1489.

Elson R, Awofisayo-Okuyelu A, Greener T, Swift C, Painset A, Amar C, Newton A, Aird H, Swindlehurst M, Elviss N, Foster K, Dallman TJ, Ruggles R, Grant K. 2019. Utility of WGS to describe the persistence and evolution of *L. monocytogenes* strains within crabmeat processing environments linked to outbreaks. J Food Protect 82:30-38.

Gillesberg Lassen S, Ethelberg S, Björkman JT, Jensen T, Sørensen G, Kvistholm Jensen A, Sørensen G, Kvistholm Jensen A, Müller L, Nielsen EM, Mølbak K. 2016. Two listeria outbreaks caused by smoked fish consumption-using whole genome sequencing for outbreak investigations. Clin Microbiol Infect 22 (7), 620–624.

Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, McClure P, Kimura B, Ching Chai L, Chapman J, Grant K. 2019. The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiol 79:96-115.

Kovac J, den Bakker H, Carroll LM, Wiedmann M.. 2017. Precision food safety: A systems approach to food safety facilitated by genomics tools. Trends An Chem 96:52-61.

Lee H, Popodi E, Tang H., Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci Unit. States 109: E2774-E2783;

Leekitcharoenphon P, Sørensen G, Löfström C, Battisti A, Szabo I, Wasyl D, Slowey R, Zhao S, Brisabois A, Kornschober C, Kärssin A, Szilárd J, Černý T, Svendsen CA, Pedersen K, Aarestrup FM, Hendriksen RS. 2019. Cross-Border transmission of *Salmonella* Choleraesuis var. Kunzendorf in European pigs and wild boar: infection, genetics, and evolution. Front Microbiol. 10:179.

Pightling, AW, Pettengill, JB., Luo Y, Baugher JD, Hugh Rand H, Strain E. 2018. Interpreting Whole-Genome Sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. Front Microbiol 9:1482, doi: 10.3389/fmicb.2018.01482

Ronholm J, Nasheri JN, Petronella N, Pagotto F. 2016. Navigating microbiological food safety in the era of whole-genome sequencing. Clin Microbiol Rev 29 (4):837-57.

Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect 24 (4), 350–354.

Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, Boxrud D. 2015. Characterization of foodborne outbreaks of *Salmonella enterica* serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-based analysis for surveillance and outbreak detection. J Clin Microbiol 53(10):3334-40.

World Health Organization (WHO). 2018. Whole genome sequencing for foodborne disease surveillance Landscape paper. ISBN 978-92-4-151368-9.

About the BfR

The German Federal Institute for Risk Assessment (BfR) is a scientifically independent institution within the portfolio of the Federal Ministry of Food and Agriculture (BMEL) in Germany. It advises the German federal government and federal states on questions of food, chemical and product safety. The BfR conducts its own research on topics that are closely linked to its assessment tasks.

This text version is a translation of the original German text which is the only legally binding version.