JRP24-FBZSH9-BEONE D2.2

Workpackage 2

Responsible Partner: RIVM

Contributing partners: RIVM (30), INSA (36), SSI (13), PHE (22), NVI (33),RKI (11), APHA (21), PIWET (34)

GENERAL INFORMATION

European Joint Programme full title	**Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards**
European Joint Programme acronym	**One Health EJP – BeONE - Building integrative tools for OneHealth Surveillance**
Funding	**This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.**
Grant Agreement	**Grant agreement n° 773830**
Start Date	**01/01/2018**
Duration	**36 Months**

DOCUMENT MANAGEMENT

Project deliverable	D-BeONE.2.2 Final Outbreak detection algorithm
Project Acronym	BeONE
Author	Claudia Coipan
Other contributors	Friesema I., Franz E.
Due month of the report	M50
Actual submission month	M53
Type *R: Document, report* *DEC: Websites, patent filings, videos, etc.; OTHER*	R **Save date**: 22-mar-22
Dissemination level *PU: Public (default)* *CO: confidential, only for members of the consortium (including the Commission Services)*	PU **This is the default setting.** If this project deliverable should be confidential, please add justification here (may be assessed by PMT): ....
Dissemination *Author's suggestion to inform the following possible interested parties.*	OHEJP WP 1 ☐ OHEJP WP 2 ☐ OHEJP WP 3 ☐ OHEJP WP 4 ☐ OHEJP WP 5 ☐ OHEJP WP 6 ☐ OHEJP WP 7 ☐ Project Management Team ☐ Communication Team ☐ Scientific Steering Board ☐ National Stakeholders/Program Owners Committee ☐ EFSA ☐ ECDC ☐ EEA ☐ EMA ☐ FAO ☐ WHO ☐ OIE ☐ Other international stakeholder(s): ………………………………………………………………… Social Media: .... **<u>Other recipient(s)</u>:** ....

MULTITIER APPROACH ON CLUSTER DETECTION

Summary

Surveillance of foodborne pathogens allows identification and control of outbreaks. Collection of good epidemiological information is essential for outbreak investigation and source finding, and the most detailed information is collected from the questionnaires the infected subjects are required to fill in. However, at the moment of the initiation of the outbreak investigations, this information is likely to be missing. It is in only exceptional instances of very well organized surveillance systems and mandatory notifiable diseases, that subjects are required to fill in a questionnaire at the moment of sampling and diagnostics. Furthermore, the differences in digital competences between age groups and/or social categories may limit the use of digital questionnaires. This implies extra time and qualified personnel to convert the analogue questionnaires into electronic data that can be readily analysed. Yet another obstacle in prompt usage of epidemiological data is the progressive nature of outbreaks, with slowly accumulating number of cases in the beginning, and therefore limited use of statistical tests for identifying common exposure variables.

Whole-genome-sequencing (WGS) is being increasingly used as a routine typing tool for foodborne pathogens and it has revolutionized the field of molecular epidemiology. The benefits of WGS have been highlighted in many occasions (1-4) and include a much greater strain discrimination for bacterial typing, and the possibility to infer also phenotypic information (such as serovar, serotype, and antibiotic resistance patterns). This allowed an enhanced cluster resolution and an equally enhanced ability to detect clusters and to monitor disease trends.

Cluster detection based on WGS often assumes that the various isolates come from a common source and that the source is clonal in nature i.e. the population of the bacterial pathogen at the source is genetically homogeneous. Thus, all the isolates pertaining to a cluster can be traced back to a most recent common ancestor, which, in epidemiological terms, means that patients that share highly similar strains are likely to have contracted these from a common source. However, it should be emphasized that this assumption of a common source is a hypothesis that should be falsified with additional epidemiological investigation. Usually this involve a case-control study in order to identify this common source. The availability of matching food or environmental hypothesis significantly adds weight of evidence to the assumed common source of infection. A cautionary note must be struck here, as the fact that an isolate belongs to a cluster is indicative of a higher probability that it is related to the other isolates in the same cluster, but it cannot exclude the possibility of convergent evolution or random identity. Challenges still remain regarding the epidemiological interpretation of WGS. A major point of uncertainty is the definition of clusters.

A major point of uncertainty is the definition of clusters, as exemplified below for the main food-borne pathogens: Campylobacter jejuni/coli, Salmonella enterica, Escherichia coli, and Listeria monocytogenes (5). There are currently multiple working definitions for cluster delineation that include either a cut-off value for genetic dissimilarities either also a time component as the time interval that the isolates are being collected. Some of the common definitions for a cluster are:

- for C. jejuni/coli a genetic distance cut-off value of maximum 14 wg/cgMLST alleles, or 15 SNPs (6, 7),
- for E. coli a genetic distance cut-off value of maximum 10 for both alleles and SNPs (8, 9), but outbreaks with as many as 15 SNPs distance among the isolates in the cluster have been identified (10).
- for L. monocytogenes the most common definition is a maximum of 7 cgMLST alleles , or 3 SNPs (11-14). Additionally, some definitions include a maximum of two year interval between the isolates (11, 12). However, research of Food and Drugs Association (FDA), USA has identified outbreaks with as many as 23 SNPs distance among the isolates in the cluster (2, 15).
- for various serovars of S. enterica, a maximum of 10 alleles , where any two isolates are at most 5 alleles distant (1), and a maximum number of SNPs ranging from 2 for S. enterica Typhimurium (16), to 5 for S. enterica Enteritidis (17), and up to 13 for S. enterica Dublin (18) are being used. Additionally, the Center for Disease Control, USA (CDC) employs also a time interval of maximum 60 days (1). Here there are also exceptions to the rule, as S. Typhimurium

outbreaks have been identified with as many as 12 SNPs distance among the isolates in the cluster (19).

In order to provide a basis for a better informed integration of WGS in epidemiological surveillance we use an evolutionary view to the diversification of the bacterial isolates coming from a common shared source. We propose a multitier algorithm that incorporates concepts of evolutionary biology, and that makes use of not only cgMLST profiles but also cgSNPs.

1. For compatibility with the currently used molecular typing system – cgMLST, Hamming distance, and single-linkage hierarchical clustering (sl) – the first step uses these methods.

What is a cluster with a threshold distance?

One of the most often used distances in clustering genetic information is Hamming, and it can take values between 0 and the maximum number of loci used in the typing procedure.

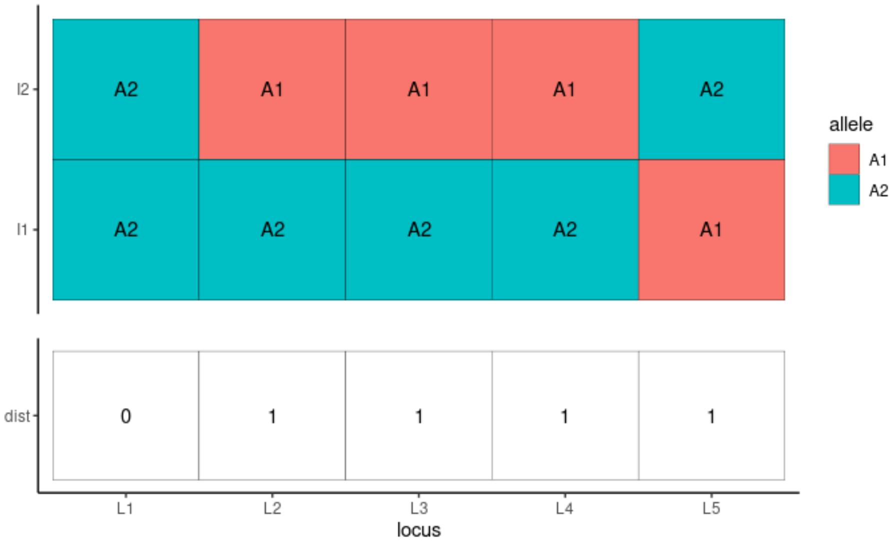
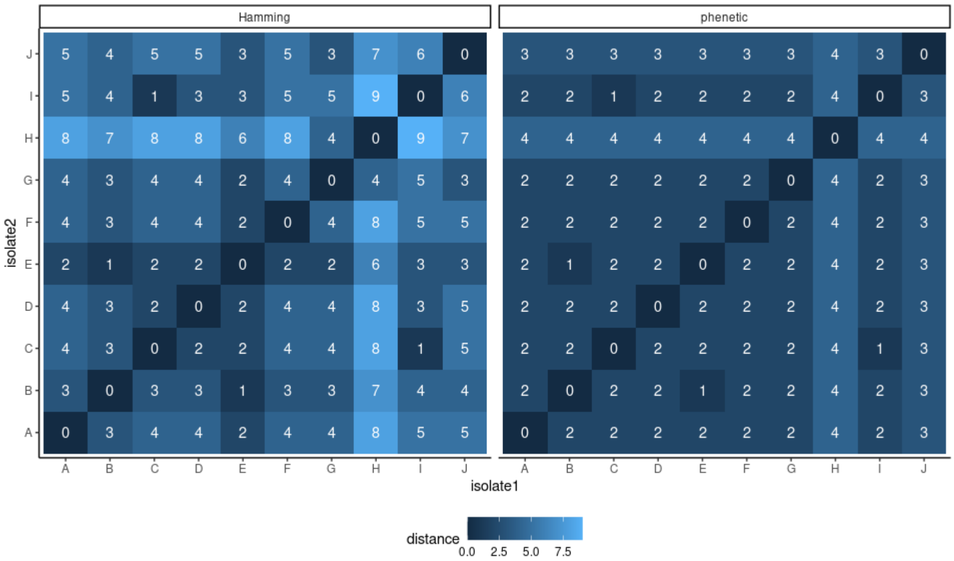


Figure 1. Distance between two isolates with five loci, and two alleles in their allelic profile. The total distance between two isolates will be the sum of distances across all loci

One can chose to use the Hamming distances as such, to infer nearest neighbours.

But can also use single-linkage clustering. What this method entails is, as the name suggests, finding a link between all isolates in a dataset, by using an agglomerative process. The process starts with two points in the dataset between which the distance is minimal across all possible pairs of points in the dataset. These are then linked together to form a first cluster. A next point in the dataset will be searched where the distance to this initial cluster is minimal. However, since in the initial cluster there are two points, a choice has to be made as to which distance to be used. The single-linkage method implies that the minimal distance will be chosen (by contrast, in average-linkage, the average of the distances will be taken, and in complete linkage, the maximum of the two will be taken). The distances resulting from such an algorithm are called phenetic distances. The following figure illustrates the differences between the Hamming and the phenetic distances for a set of 10 isolates.



Hamming

isolate2	A	B	C	D	E	F	G	H	I	J
J	5	4	5	5	3	5	3	7	6	0
I	5	4	1	3	3	5	5	9	0	6
H	8	7	8	8	6	8	4	0	9	7
G	4	3	4	4	2	4	0	4	5	3
F	4	3	4	4	2	0	4	8	5	5
E	2	1	2	2	0	2	2	6	3	3
D	4	3	2	0	2	4	4	8	3	5
C	4	3	0	2	2	4	4	8	1	5
B	3	0	3	3	1	3	3	7	4	4
A	0	3	4	4	2	4	4	8	5	5

phenetic

isolate2	A	B	C	D	E	F	G	H	I	J
J	3	3	3	3	3	3	3	4	3	0
I	2	2	1	2	2	2	2	4	0	3
H	4	4	4	4	4	4	4	0	4	4
G	2	2	2	2	2	2	0	4	2	3
F	2	2	2	2	2	0	2	4	2	3
E	2	1	2	2	0	2	2	4	2	3
D	2	2	2	0	2	2	2	4	2	3
C	2	2	0	2	2	2	2	4	1	3
B	2	0	2	2	1	2	2	4	2	3
A	0	2	2	2	2	2	2	4	2	3

Figure 2. Hamming and single-linkage phenetic distances in a set of ten isolates

Based on the Hamming distances, the pairs of isolates E & B and I & C, both have the smallest distance (1). So the clustering will start with these pairs: cluster(B,E) and cluster(C,I). The next smallest distance in the remaining pairs of isolates is from A, to isolate E (2); however, because isolate E is already in cluster(B,E), the distance from A to E will apply to the whole cluster and its members, thus the distance to isolate B becoming also 2. Same holds true for isolates D, F and G and cluster(B,E), all having a distance of 2 to one of the isolates in (B,E), thus becoming distance 2 also to the other isolate. Simultaneously, D is the nearest isolate also for cluster(C,I), with a distance of 2. And isolates C and E have also a distance of 2. At this point we have cluster (A,D,F,G,(B,E),(C,I)), where all the distances have been reduced to 2, when the actual distances are as high as 5. The next closest isolate is J, which has only a distance of 3 from isolate E; thus the distance to the whole cluster is 3, and the new larger cluster becomes (J,(A,D,F,G,(B,E),(C,I))). Finally, isolate H has a distance of 4 to the existing cluster and will thus be incorporated last in the clustering process, showing a phenetic distance of 4 to all other isolates, when the Hamming distance can be as high as 9. The resulting dendrogram/tree is shown in the next figure.

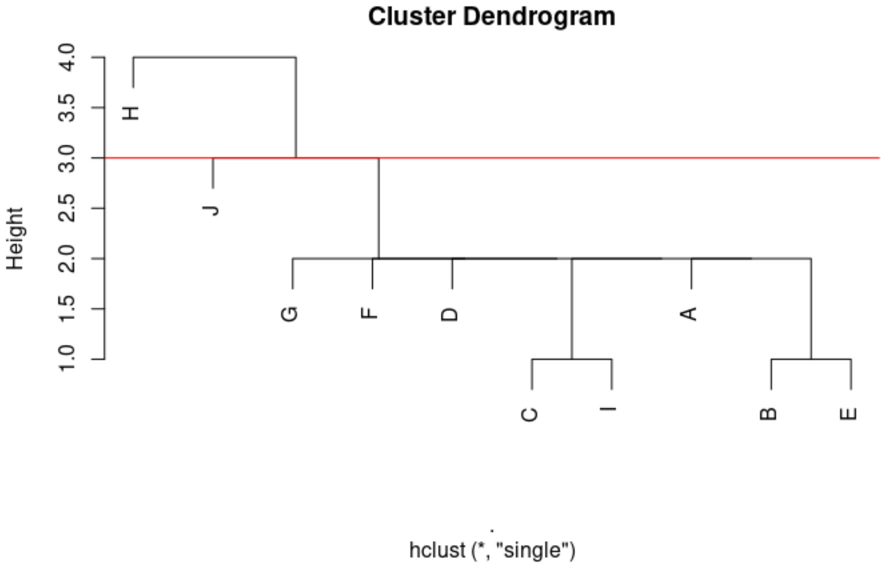


Figure 3. Dendrogram based on single-linkage clustering of ten isolates. Red horizontal line corresponds to a random chosen threshold

From the dendrogram it is clear that if one would want to see a cluster with a threshold distance of 3 (red line), all but one isolate will be included in this cluster, although their actual Hamming distances are sometimes much higher than that threshold. The phenetic distances are thus not identical to the actual Hamming distances.

The use of single linkage combined with a variable number of missing loci for each isolate allows, however, for allelic distances within a cluster much exceeding the predefined threshold of seven alleles. Thus, the shape of the distribution of the cgMLST Hamming distances within a cluster can be bimodal, indicating the existence of two subclusters.

Furthermore, since the gene is not the structural unit of DNA, a Hamming distance of one based on alleles might be an underestimation of the distances based on nucleotide polymorphisms.

The use of single linkage combined with a variable number of missing loci for each isolate allows, however, for allelic distances within a cluster much exceeding the predefined threshold of allelic dissimilarities. Furthermore, the structural unit of DNA is not the gene but the nucleotide, which means that allelic distances may be in fact an underestimation of the dissimilarity among isolates.

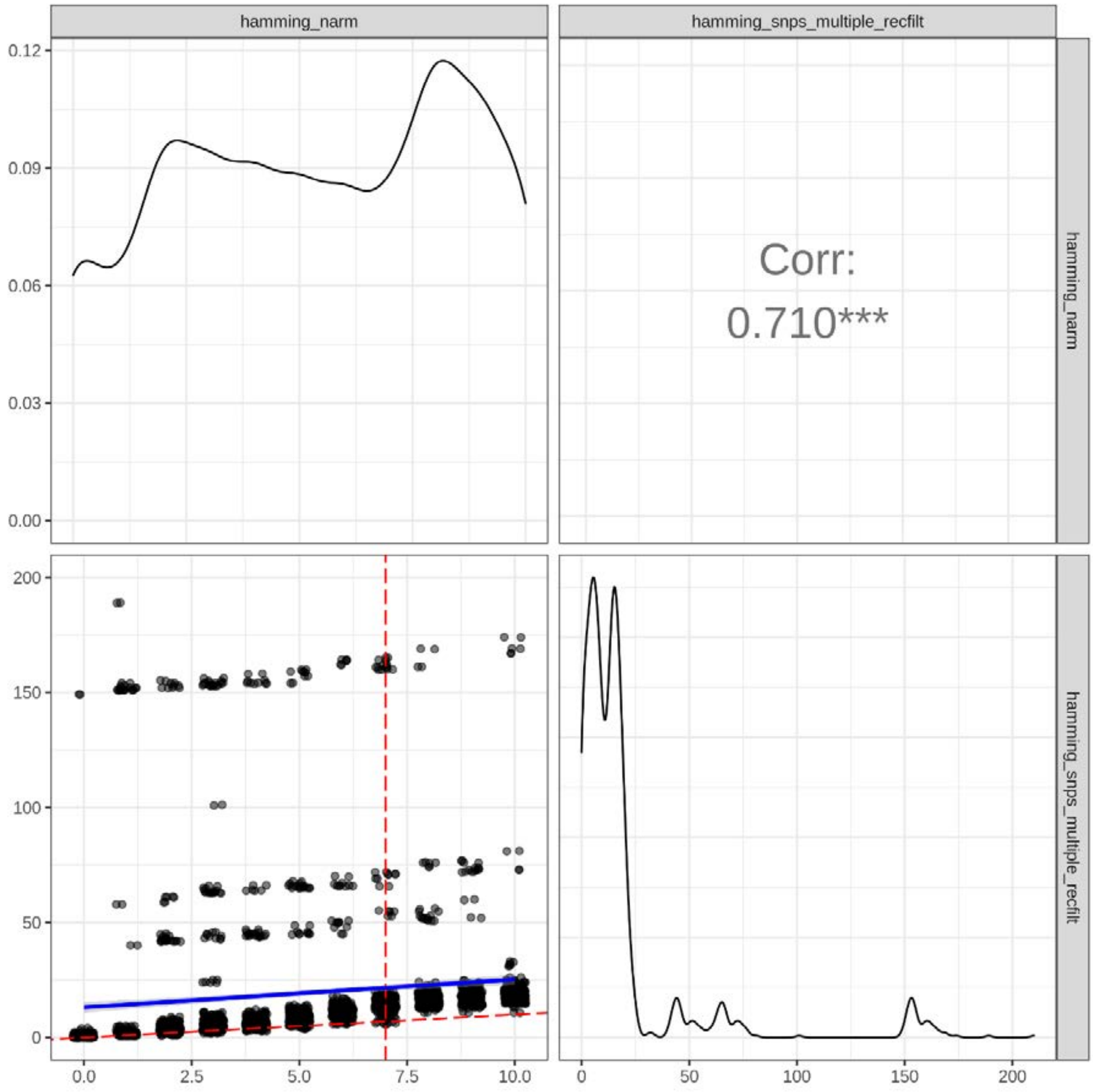
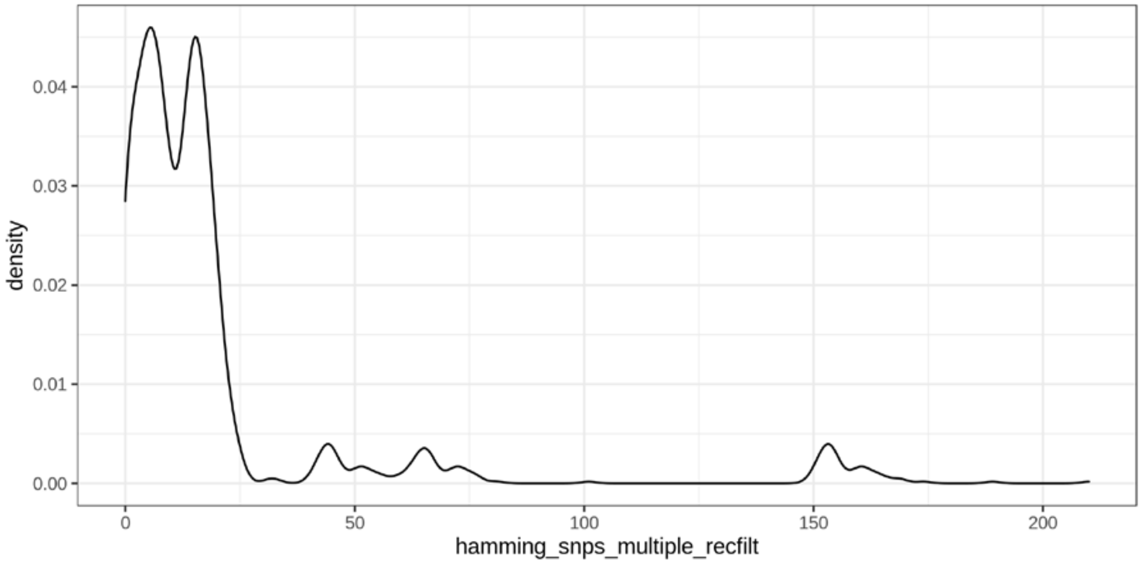


Figure 4. correlation of Hamming distances based on genes/alleles and SNPs within a cgMLST s7 defined cluster.

2. We thus use a second step involving SNP-calling, pairwise Hamming distances, and Gaussian mixture models to split potential multimodal distributions of pairwise dissimilarities, which are often indicative of multiple clusters.



B.

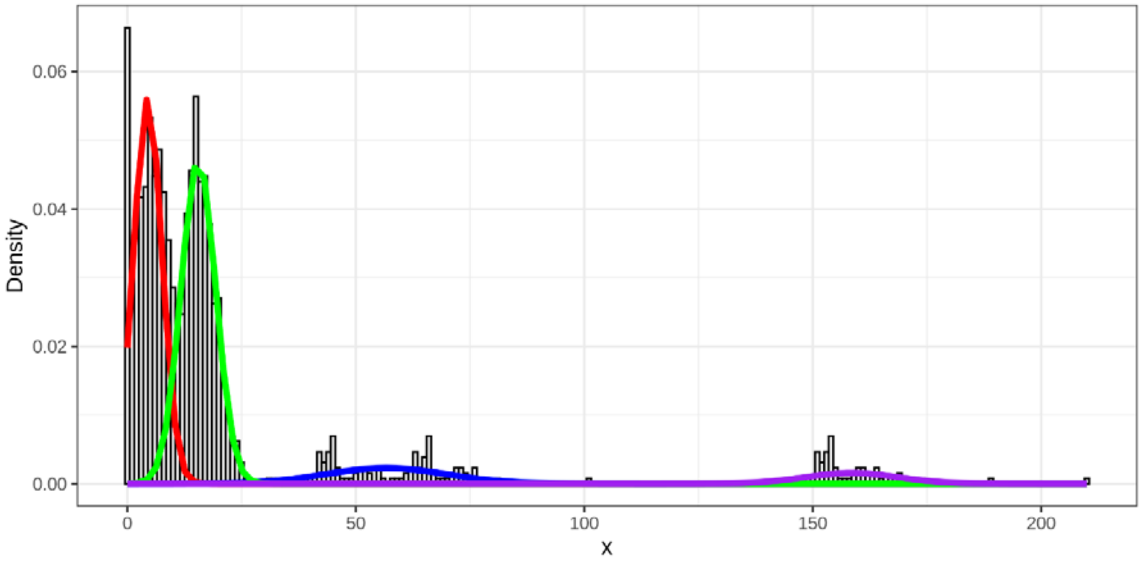


Figure5. Distribution of snp distances among the sequences of Lm included in a cluster as defined by single-linkage hierarchical clustering at threshold 10 A. density plot of the distances, B. split of the multimodal distribution into several unimodal distributions using Gaussian mixture models

Of the pairwise unimodal distributions identified by the Gaussian mixture models we retain the distribution with the smallest mode, and the sequences it comprises as a more probable cluster.

This approach can only be used when the number of sequences is high. However, time is critical for outbreak control and, therefore, timely identification of clusters, indicative of a potential outbreak, would require the use of a small number of sequences.

3. Therefore, in a third step we calculate the expected SNP distances between any two isolates based on Kingman's coalescent model of shared ancestry (20).

The SNP variation is suitable for inferring evolutionary relations among the bacterial isolates, as nucleotides are the direct object of the mutational process. The most common source of genetic variation is mutation. The mutation rate, defined as the number of mutations per generation time, can be used in terms of "molecular clock" in order to refine the previously defined clusters. Thus, the incorporation of the time span between any two isolates and the expected mutation rate might act as prior knowledge in calculating the maximum possible distance observed among the isolates in the hypothesis of common descent.

Some of the main factors impacting on the perceived genetic distances among the isolates from a common source are:

- mutation rate – number of mutations introduced in the genome per generation
- generation time (doubling time) – the time necessary for the bacterial cells to divide
- substitution rate – the product of mutation rate and generation time, but more often expressed as number of substitutions per real time unit (e.g. year)
- effective bacterial population size

The expected distance between any number of bacterial isolates presumed to have originated from the same sample is expressed as a distribution resulting from forward-time simulations of bacterial evolution. The main assumptions in the simulations are:

- a homogeneous initial bacterial population
- constant generation time
- non-overlapping generations
- finite-sites model
- equal transition rate
- constant population size
- no selection pressure and no differential fitness

For a hypothetical population with 1000 individuals, a genome of 4.5 Mb, a mutation rate of $7*10^{-10}$ per site/generation, the distribution of the segregating sites after 1000 generations and 1000 simulations, is depicted in Figure 6.

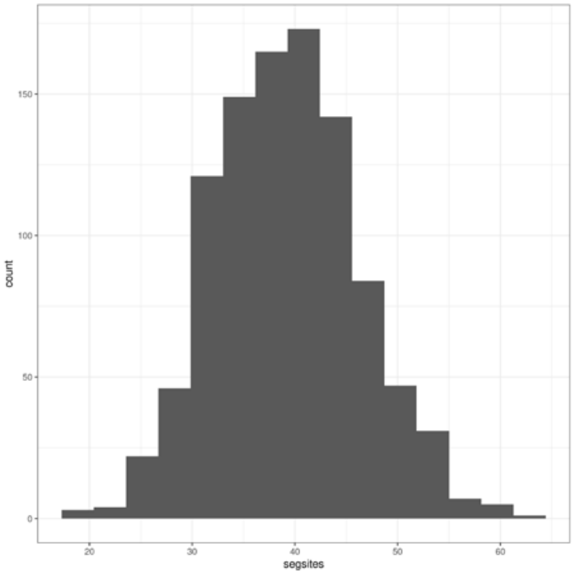


Figure 6. Variation of segregating sites from 1000 simulations.

Of each simulation we draw randomly 10000 samples of size equal to the number of individuals in the cluster identified in the previous steps, the effect of sampling being tested by comparing the distributions corresponding to different number of samples (Figure 7) by means of Kolmogorov-Smirnov test.

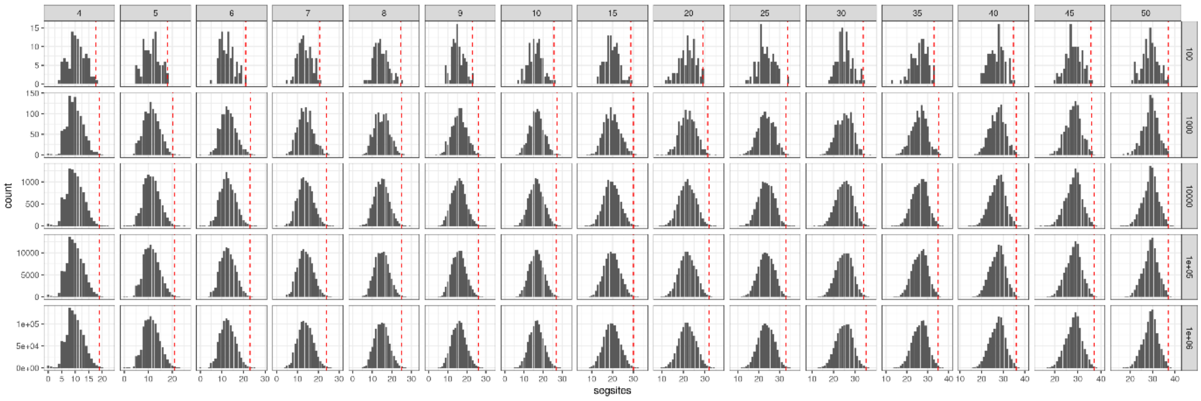


Figure 7. Impact of sampling on the expected number of segregating sites. There were no significant differences on the distributions based on 10^3 to 10^6 samples.

4. In a fourth step, the comparison of the distribution of observed dissimilarities with the expected one is expressed as the probability that the two or *n* selected isolates come from a common ancestor in a given timespan.
   The probability of observing *s* segregating sites is calculated as the cumulative probability to the right of the quantile corresponding to *s* on the precomputed distributions. For a population in the median range of diversification, the expected distribution of segregating sites is given in Figure 8. The red vertical line indicates the number of segregating sites observed empirically, and the blue area to the cumulative probability of values equal or higher to the observed segregating sites. The higher this probability is, the more likely the n sequences come from the same homogeneous population.

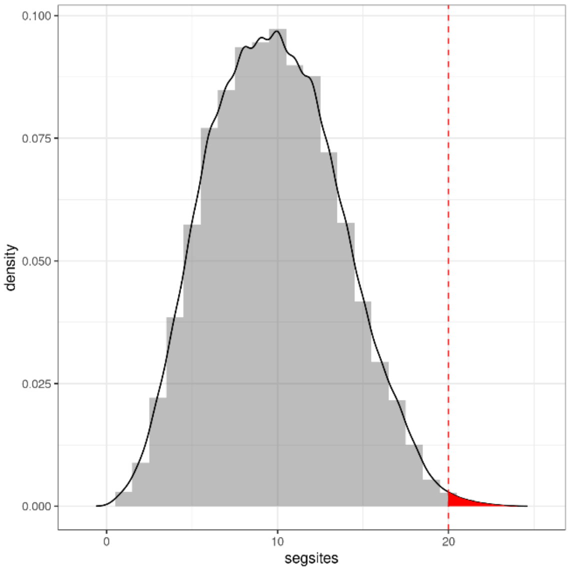


Figure 8. Distribution of segregating sites in a theoretical population. The observed number of segregating sites in the identified cluster is depicted as a vertical line and the probability of observing the respective number is the shaded blue area to its right.

Further testing, sensitivity analyses and validation on epidemiologically confirmed outbreaks are currently running. Our preliminary results indicate that in the context of effective disease surveillance a two-tier approach could be beneficial, using cgMLST for a first screening of the isolates, followed by SNP phylogenetic analyses to refine the clusters and identify relatedness among the isolates at a higher resolution. A potential rapid implementation might be based on precomputed distributions for a relatively broad range of

Limitations:

The Gaussian mixture models are applicable and reliable for relatively large sets of observations (sequences).

The coalescent model of evolution has numerous assumptions that may not (always) be met in real bacterial populations. It is however an approximation that is currently achievable with the population parameters described to date in the literature. It can be further improved and extended, as our knowledge of the bacterial population dynamics progresses.

The use of maximum values for the various simulation parameters provide a conservative frame for defining related bacterial sequences (isolates), allowing for false-positives, and reducing the probability of false-negatives.

Even a high probability of shared recent descent is not a guarantee for it. Other processes might influence the observed distribution of the segregating sites. Identical sequences may occur under natural settings either by chance or at the hand of convergent evolution mechanisms. Furthermore, one of the major assumptions is that the observed number of segregating sites is an accurate reflection of the real one. However, errors may be introduced at a number of steps, be it in vivo (in differential survival of the infectious bacteria), in vitro (in culturing the bacterial isolates), or in silico (in calling the nucleotide polymorphisms). Advances in methodology and understanding the biological processes can improve the models used here.

For foodborne pathogens, collaboration with the food and animal health sectors are, therefore, key in advancing our understanding of the bacterial population dynamics and, consequently, of the data generating processes underlying the outbreaks.

References

1. Besser JM, Carleton HA, Trees E, Stroika SG, Hise K, Wise M, et al. Interpretation of Whole-Genome Sequencing for Enteric Disease Surveillance and Outbreak Investigation. Foodborne pathogens and disease. 2019;16(7):504-12.
2. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. Front Microbiol. 2018;9:1482.
3. Prevention ECfD, Control. Expert opinion on whole genome sequencing for public health surveillance2016. Available from: https://www.ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/whole-genome-sequencing-for-public-health-surveillance.pdf.
4. WHO. Whole genome sequencing for foodborne disease surveillance: WHO; 2018.
5. Authority EFS, Prevention ECfD, Control. The European Union One Health 2018 Zoonoses Report. EFSA Journal. 2019;17(12):e05926.
6. Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, et al. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol. 2013;51(8):2526-34.
7. Llarena AK, Taboada E, Rossi M. Whole-Genome Sequencing in Epidemiology of Campylobacter jejuni Infections. J Clin Microbiol. 2017;55(5):1269-75.
8. Dekker JP, Frank KM. Next-Generation Epidemiology: Using Real-Time Core Genome Multilocus Sequence Typing To Support Infection Control Policy. J Clin Microbiol. 2016;54(12):2850-3.
9. Roer L, Hansen F, Thomsen MCF, Knudsen JD, Hansen DS, Wang M, et al. WGS-based surveillance of third-generation cephalosporin-resistant Escherichia coli from bloodstream infections in Denmark. J Antimicrob Chemother. 2017;72(7):1922-9.
10. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of Escherichia coli O157:H7 outbreaks. Proc Natl Acad Sci U S A. 2011;108(50):20142-7.
11. 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. Nature Microbiology. 2016;2(2):16185.

12. Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen N, et al. Real-Time Whole-Genome Sequencing for Surveillance of Listeria monocytogenes, France. Emerg Infect Dis. 2017;23(9):1462-70.
13. Kvistholm Jensen A, Nielsen EM, Björkman JT, Jensen T, Müller L, Persson S, et al. Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. Clin Infect Dis. 2016;63(1):64-70.
14. Van Walle I, Björkman JT, Cormican M, Dallman T, Mossong J, Moura A, et al. Retrospective validation of whole genome sequencing-enhanced surveillance of listeriosis in Europe, 2010 to 2015. Eurosurveillance. 2018;23(33):1700798.
15. Wang Q, Holmes N, Martinez E, Howard P, Hill-Cawthorne G, Sintchenko V. It Is Not All about Single Nucleotide Polymorphisms: Comparison of Mobile Genetic Elements and Deletions in Listeria monocytogenes Genomes Links Cases of Hospital-Acquired Listeriosis to the Environmental Source. J Clin Microbiol. 2015;53(11):3492-500.
16. Phillips A, Sotomayor C, Wang Q, Holmes N, Furlong C, Ward K, et al. Whole genome sequencing of Salmonella Typhimurium illuminates distinct outbreaks caused by an endemic multi-locus variable number tandem repeat analysis type in Australia, 2014. BMC Microbiology. 2016;16(1):211.
17. Pijnacker R, Dallman TJ, Tijsma ASL, Hawkins G, Larkin L, Kotila SM, et al. An international outbreak of Salmonella enterica serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. Lancet Infect Dis. 2019;19(7):778-86.
18. Ågren ECC, Wahlström H, Vesterlund-Carlson C, Lahti E, Melin L, Söderlund R. Comparison of whole genome sequencing typing results and epidemiological contact information from outbreaks of Salmonella Dublin in Swedish cattle herds. Infect Ecol Epidemiol. 2016;6:31782.
19. Octavia S, Wang Q, Tanaka MM, Kaur S, Sintchenko V, Lan R. Delineating community outbreaks of Salmonella enterica serovar Typhimurium by use of whole-genome sequencing: insights into genomic variability within an outbreak. Journal of clinical microbiology. 2015;53(4):1063-71.
20. Kingman, J. F. C. The coalescent. Stochastic processes and their applications. 1982;13(3): 235-248.