JRP24-FBZSH9-BEONE D1.3
Workpackage 1

Responsible Partner: INSA (36)

Contributing partners: All partners

GENERAL INFORMATION

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

DOCUMENT MANAGEMENT

Deliverable	**D-BeONE.1.3 Clustering congruence and thresholds**
WP and Task	JRP24-WP1-T3 Clustering congruence and thresholds
Leader	Vítor Borges
Other contributors	Leader partner: INSA (Verónica Mixão, Miguel Pinto, João Paulo Gomes, Vítor Borges); Contributor partners: All partners (https://doi.org/10.5281/zenodo.7445805)
Due month of the deliverable	**60**
Actual finalization month	**60**
Type ***R: Document, report*** ***DEC: Websites, patent fillings, videos, etc.*** ***OTHER***	R
Dissemination level ***PU: Public (default)*** ***CO: confidential, only for members of the consortium (including the Commission Services).***	PU
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 ☒ Other international stakeholder(s): …… Social Media: "One Health EJP" twitter **Other recipient(s):**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

CLUSTER CONGRUENCE AND THRESHOLDS

Introduction

Foodborne pathogens pose an important and serious threat to human and animal health. Therefore, genomics-based surveillance systems able to track the circulation of these pathogens and monitor their clinical and epidemiological relevant features have been implemented at the reference laboratories of multiple countries and sectors. However, there is a lack of harmonization between them that challenges the integration and comparison of data at international and intersectoral levels, and, ultimately, the establishment of fully integrative and efficient One Health genomic surveillance frameworks. In the BeONE project, we aimed to perform a massive comparison between different bioinformatics pipelines used for genomic surveillance of foodborne bacterial pathogens, as a way to assess the congruence and comparability of their clustering results. Specifically, BeONE WP1-T3 had as main objectives the:

- Evaluation of the behavior of different surveillance-oriented genomics pipelines in the definition of genetic clusters, through the identification of threshold ranges with high resolution for outbreak detection or stable threshold ranges for nomenclature design (useful for longitudinal surveillance);
- Determination of the congruence with traditional typing nomenclatures, such as Sequence Type (ST) or serovar;
- Identification of congruent resolution levels between pipelines, as a way to facilitate international data sharing and cooperation;
- Assessment of the similarities and discrepancies between pipelines in the identification of outbreak-related isolates.

Selection of WGS-based typing methods to be evaluated

This cluster congruence analysis counted with the participation of all BeONE partners, which allowed the inclusion of a wide variety of pipelines, from multiple countries and sectors (Figure 1). Based on the provided details for each pipeline, we established a strategy to avoid pipeline redundancy and increase the analysis efficiency (e.g., splitting the work between partners with matching pipelines). With the full support of WP1 team, each partner was responsible to install and run its own pipeline on the datasets generated in WP1-T2 for each of the four target species: *Listeria monocytogenes*, *Salmonella enterica*, *Escherichia coli* and *Campylobacter jejuni* (BeONE dataset complemented with extra data as detailed at https://doi.org/10.5281/zenodo.5808832) [1]. Assembly-based pipelines took as input all the assemblies available in these datasets, while read-based pipelines started from the QC-passed sequencing reads for the most represented STs/serovars in each dataset, which were then aligned to respective reference genome sequences (Table 1). Consequently, assembly-based pipelines provided clustering information considering the whole dataset, while all read-based pipelines provided clustering information per ST/serotype (one read-based pipeline also ran a subset with the main STs together). Clustering results were then shared with WP1 team for congruence analysis.

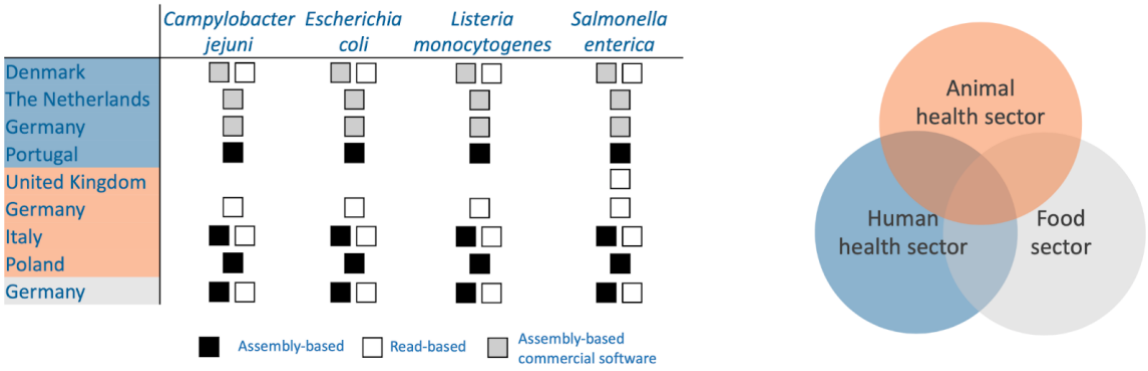

Figure 1. Summary of the different countries and sectors involved in the BeONE assessment of cluster congruence with indication of the pipeline diversity per country/sector and species.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

Table 1. Details on the datasets and sub-datasets used for the cluster congruence analysis, with indication of the species, Zenodo repositories with the BeONE and the extra datasets, the number of isolates, the type of pipeline in which they were used (assembly-based or read-based) and the name of the reference genome used for read mapping.

Species	Dataset source (BeONE / Extra)	Sub-dataset (ST or Serotype)	Number of isolates	Type of pipeline	Reference genome*
L. monocytogenes	10.5281/zenodo.7267486 / 10.5281/zenodo.7116878	All isolates	3,300	Assembly	-
		ST 1	343	Read	F2365
		ST 5	269	Read	CP006592
		ST 6	501	Read	CP006046
		ST 8	247	Read	CP006862
		ST 121	240	Read	HG813249
S. enterica	10.5281/zenodo.7267785 / 10.5281/zenodo.7119735	All isolates	2,974	Assembly	-
		Enteritidis	725	Read	AM933172
		Typhimurium	880	Read	AE006468
		Infantis	120	Read	H15092067901-2
E. coli	10.5281/zenodo.7267844 / 10.5281/zenodo.7120057	All isolates	2,307	Assembly	-
		O157:H7	863	Read	Sakai (NC_002695)
C. jejuni	10.5281/zenodo.7267879 / 10.5281/zenodo.7120166	All isolates	3,686	Assembly	-
		ST 21	361	Read	NCTC_11168
		ST 45	165	Read	HG428754
		ST 48	173	Read	CP006006
		ST 50	218	Read	NCTC_11168
		ST 257	110	Read	H055140547

*Retrieved from the list of SnapperDB references [2]

Obtaining clustering information at all distance thresholds – the development of ReporTree

As shown in Table 2, a high variety of pipelines was assessed in this study, covering the most commonly used schemas for allele-calling, allele/SNP-callers and clustering methods. For the purpose of cluster congruence analysis and pipeline comparison, it was essential to generate clustering information at all possible distance thresholds for each pipeline (both assembly- and read-based pipelines). Indeed, some pipelines do not reach this level of resolution, instead ending-up in intermediate outputs (e.g., allele/SNP or distance matrices) that are explored with non-automated approaches for cluster identification, such as the visual exploration of (large) dendrograms resulting from hierarchical clustering (HC) or Minimum-Spanning Trees (MST) provided by GrapeTree [3].

Table 2. Pipelines used in the BeONE cluster congruence analysis. HC – Hierarchical clustering; GT – GrapeTree.

Pipeline	Type of pipeline	*Schema (if assembly) / Type of dataset (if read)*				*Output for clustering*	*Clustering method*
		Lm	*Se*	*Ec*	*Cj*		
chewieSnake	Assembly	Ruppitsch	Enterobase	Enterobase	PubMLST	Distance matrix	HC and GT
chewBBACA	Assembly	Pasteur	INNUENDO/ Enterobase	INNUENDO/ Enterobase	INNUENDO/ PubMLST	Allele matrix	HC and GT
Ridom SeqSphere	Assembly*	Ruppitsch	Enterobase	Enterobase	PubMLST	Distance matrix	HC
Bionumerics	Assembly*	Pasteur	Enterobase	Enterobase	Oxford	Distance matrix	HC
MentaLiST	Assembly	Pasteur	INNUENDO	INNUENDO	INNUENDO	Allele matrix	HC and GT
SnippySnake	Read	ST	Serotype	Serotype	ST	Distance matrix	HC
CSI Phylogeny	Read	ST/All	Serotype/All	Serotype/All	ST/All	Distance matrix	HC
WGSBAC	Read	ST	Serotype	Serotype	ST	Distance matrix	HC
SnapperDB	Read	ST	Serotype	Serotype	ST	Distance matrix	HC
kSNP3	Read	ST	Serotype	Serotype	ST	Distance matrix	HC

*Commercial software

To facilitate, speed-up and automate this workflow, we developed ReporTree (https://www.researchsquare.com/article/rs-1404655/v2) [4], a flexible tool that allows the rapid identification of genetic clusters at any (or all) distance threshold(s) and to generate surveillance-oriented reports based on the available metadata. Besides its direct contribution to enhance current genomics-based surveillance workflows, ReporTree was tailored to answer specific needs of WP1-T3:

- *Deal with output heterogeneity*
  ReporTree deals with multiple file formats (SNP/allele or distance matrices, trees/dendrograms, multiple-sequence alignments or VCFs), being able to process the outputs provided by the different pipelines used in this study. Noteworthy, ReporTree integrates the tool vcf2mst [5] to deal with VCF files as result of the collaboration with COVRIN project.

- *Obtain clustering at all possible thresholds*
  ReporTree automatically determines genetic clusters at all possible distance thresholds using several clustering methods, without the need of dendrogram/tree visualization. For instance, we shaped GrapeTree [3] open-source code so that ReporTree can obtain cluster information directly from MSTs (modified version of this tool is available at https://github.com/insapathogenomics/GrapeTree).

- *Determine cluster congruence and stability regions*
  ReporTree assesses several metrics that can be used to compare clustering information at consecutive distance thresholds of a pipeline (determining regions of stability) or at all thresholds of two methods (determining pipeline comparability). With this purpose, we modified the code of the Comparing Partitions tool (modified version of this tool is available at https://github.com/insapathogenomics/ComparingPartitions).

ReporTree benchmarking with the four "public datasets" compiled in WP1-T2 [1] showed that it can be smoothly implemented in routine surveillance workflows, with negligible computational and time costs. Complying with the FAIR principles, ReporTree is freely available in GitHub (https://github.com/insapathogenomics/ReporTree) under a GPL-3.0 license and at Docker hub (https://hub.docker.com/r/insapathogenomics/reportree). Besides its incorporation in the BeONE datahub (WP4), ReporTree is also integrated in the COHESIVE system.

Assessing clustering congruence between multiple methods at different resolution levels and linkage to traditional typing (preliminary results)

After all BeONE partners ran their pipelines on the provided datasets for each of the four species (Tables 1 and 2), we used ReporTree to obtain clustering information at all possible distance thresholds (partitions) of each pipeline. Taking advantage of ReporTree flexibility, this was conducted with two different clustering methods (HC single-linkage and GrapeTree MSTreeV2), reinforcing the power and magnitude of this study. ReporTree also allowed calculating the necessary metrics (e.g., Adjusted Wallace and Adjusted Rand coefficients) and scores to determine threshold ranges/points with similar cluster composition between: i) subsequent partitions of the same pipeline (as a mean to identify regions of cluster stability for each method), and, ii) all partitions of different pipelines (as a mean to identify thresholds with highest congruence between methods). These outputs constituted the basis to achieve the above-mentioned objectives.

This massive comparison is currently being conducted for the four target species. Preliminary results are shown for *L. monocytogenes*. In general, we observe the existence of low stability (i.e., low congruence in cluster composition between subsequent distance thresholds) in the highest resolution region (spanning the "outbreak level") (Figure 2), followed by "plateau" regions of high stability (i.e., yielding similar cluster number and composition at distant thresholds), likely reflecting the pathogen

population structure and dataset diversity. For example, there is a good concordance between the beginning of the first large "plateau" of stability and traditional MLST nomenclature (Figure 2D). This stability landscape was parallel across all pipelines, which opened the possibility to compare their overall resolution power and identify threshold ranges with high congruence (e.g., that may be suitable for hierarchical nomenclature definition valid across pipelines). Indeed, an in-depth analysis revealed that inter-pipeline corresponding points, i.e., threshold points with highest concordance between two pipelines, follow a linear tendency across all partitions (Figure 3). This information can be key to facilitate the interpretation and comparison of clustering information resulting from different pipelines.

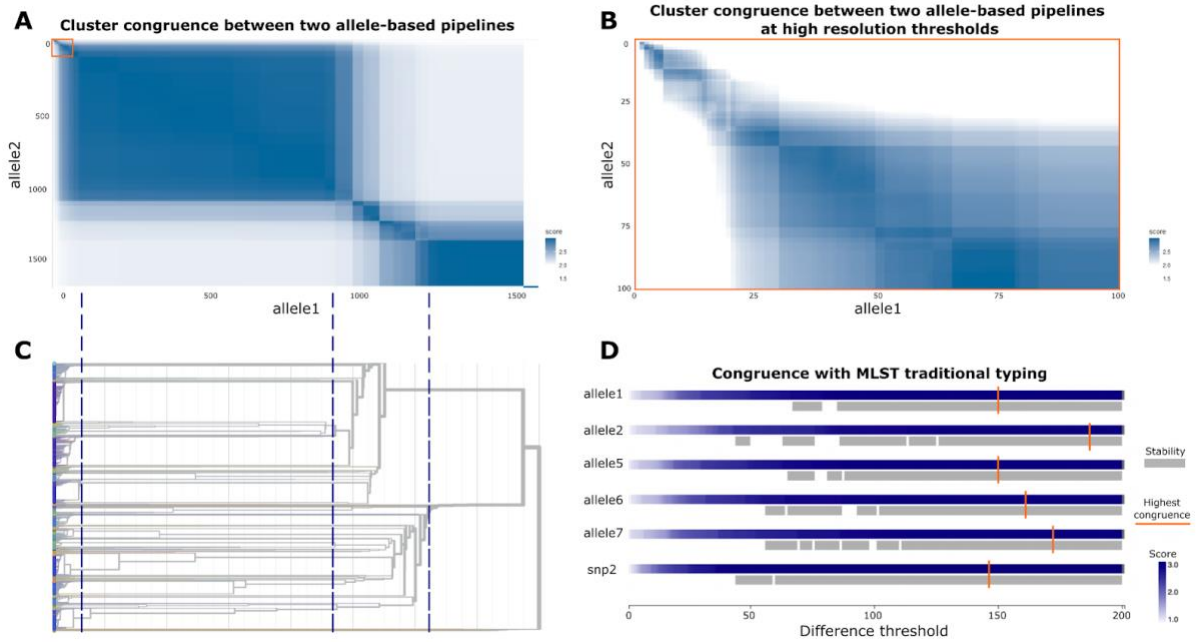

Figure 2. Example of the comparability of two pipelines using the L. monocytogenes dataset. A) Cluster congruence between two allele-based pipelines, with outbreak-resolution level highlighted in orange; B) Zoom-in in the outbreak-resolution level highlighted in A; C) Dendrogram of allele1 pipeline, with dashed lines connecting the distance to the stability regions observed in A; D) Cluster congruence in the first 200 distance thresholds of six pipelines, with stability regions marked in gray and the point of highest congruence with ST marked in orange.

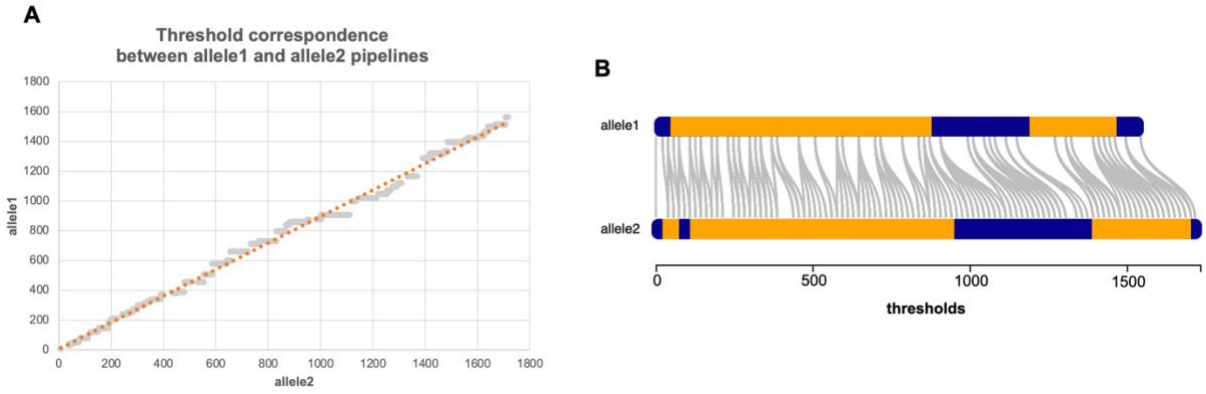

Figure 3. Clustering correspondence between two assembly-based pipelines. A) Threshold points with highest concordance between two pipelines marked in grey, with the respective linear tendency line shown in orange. B) Graphical representation of the threshold points of each pipeline, with their respective stability regions highlighted in orange. Grey lines connect inter-pipeline correspondence points.

Taking advantage of the overall good congruence between pipelines, we will also identify and compare the minimum threshold of each pipeline able to cluster the same set of highly close-related isolates (i.e., "outbreak"). By scaling-up this assessment to hundreds of "outbreak" clusters genetically inferred from the datasets (by applying thresholds commonly used as a proxy to detect "outbreaks"), as well as to epidemiologically verified outbreaks (provided by BeONE consortium as detailed in Deliverable 1.2), this study is thus expected to provide strong data on the performance and comparability of different pipelines for outbreak detection.

Concluding remarks

The joint collaboration between all BeONE partners resulted into this thorough analysis, where we could retrieve important information regarding the comparability between different genomics surveillance methods. Noteworthy, a manuscript where the extensive description of the methodology and the results of this analysis are provided is being prepared and is expected to be published during 2023. Based on these preliminary observations, we anticipate that this study will provide important information regarding the behavior of each pipeline and facilitate the direct comparability of their results. Moreover, as the code used to generate these outputs will be publicly available, it will open the possibility for other institutes to analyze their own methods and assess how they compare to others. This outcome potentiates the impact of this task beyond the BeONE partners, while providing valuable knowledge to facilitate and promote (international and intersectoral) data sharing, cooperation and communication during routine surveillance and outbreak investigation.

References

1. Mixão et al. (2021) D-BeONE.1.2 BeONE dataset (Version 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7335590
2. Dallman et al. (2018) SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics.* **34(17):** 3028-3029. https://doi.org/10.1093/bioinformatics/bty212
3. Zhou et al. (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res*. **28(9):** 1395–1404. https://doi.org/10.1101%2Fgr.232397.117
4. Mixão et al. (2022) ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data. *Research Square*. PPR552299. https://doi.org/10.21203/rs.3.rs-1404655/v2
5. Di Pasquale et al. (2021) SARS-CoV-2 surveillance in Italy through phylogenomic inferences based on Hamming distances derived from pan-SNPs, -MNPs and -InDels. *BMC Genomics*. **22:** 782. https://doi.org/10.1186%2Fs12864-021-08112-0