JRP24-FBZSH9-BEONE D1.2
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.1.2 Finalized BeONE dataset**
WP and Task	JRP24-WP1-T2 Dataset selection and curation
Leader	Vítor Borges
Other contributors	Verónica Mixão (INSA), Miguel Pinto (INSA), Holger Brendebach (BfR), Simon Tausch (BfR), Carlus Deneke (BfR) and Karin Lagesen (NVI)
Due month of the deliverable	**48**
Actual finalization month	**48**
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 **<u>Other recipient(s)</u>:**

This meeting is part of the European Joint Programme One Health EJP.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

FINALIZED BEONE DATASET

Introduction

The goal of the BeONE project is to develop integrative solutions for One Health surveillance, in which molecular and epidemiological data for foodborne pathogens, specifically *Listeria monocytogenes, Salmonella enterica, Escherichia coli* (STEC) and *Campylobacter jejuni*, can be analyzed. In this context, WP1-T2 aimed to compile a dataset (including sequencing reads and respective metadata) that captures the genomic diversity within the populations of each of the four target species. This output is expected to contribute to the accomplishment of the specific objectives of the different WPs, particularly:

- WP1: cluster congruence analysis between the different partner's pipelines
- WP2: testing novel algorithms (combining genomic and epidemiological data) for outbreak detection
- WP3: testing metadata structures and the compliance and viability of the metadata combination from different sectors and regions
- WP3/4: testing both the dashboard and the database
- WP4: testing the behavior of the system to handle epi-data from different countries in a user-oriented manner

The BeONE dataset

Although some of the above-mentioned WPs' specific objectives could be achieved by subsampling existing public data, the robustness of the results would be likely lower and the combination of metadata from different sources would be lacking for the epidemiological analyses in that case. Therefore, a collaborative effort was made by the BeONE partners (representing multiple European countries and/or sectors) to compile a controlled dataset (the BeONE dataset) with sequencing data and metadata of samples from human, food, or/and veterinary sources. To guarantee that this compilation and the posterior usage of the dataset would not only comply with the best scientific and ethical practices, but also be subject to rules agreed by all the BeONE partners, a Material Transfer Agreement (MTA) was circulated *a priori* for revision and final acceptance.

This task involved the set-up of the NVI's SAGA server for data upload, storage, and analysis. Data submission was performed following guidelines for data upload (account login, anonymization, and transfer of data), which also included specific instructions to fill a metadata submission form elaborated using EFSA Data Collection Framework for Controlled Vocabularies, in collaboration with BfR partner, leader of WP3. An additional step for read anonymization including the renaming of sequencing reads and shuffling their order in the file was performed. Quality control/filtering of the sequencing reads, confirmation of species identification (and contamination checking), genome assembly and *in silico* typing (i.e., the classification of microorganisms at intra-species level) was performed with the AQUAMIS pipeline (Deneke et al. 2021). The final BeONE dataset comprises a total of 3,884 isolates (all samples shared within the consortium that passed this dataset curation step), from which the anonymized sequencing reads were released in the European Nucleotide Archive (ENA) and the anonymized genome assemblies in the Zenodo repository [1,426 *L. monocytogenes* (accession: PRJEB57166 and 10.5281/zenodo.7267486); 1,540 *S. enterica* (accession: PRJEB57179 and 10.5281/zenodo.7267785); 308 *E. coli* (accession: PRJEB57098 and 10.5281/zenodo.7267844); 610 *C. jejuni* (accession: PRJEB57119 and 10.5281/zenodo.7267879)].

The public dataset - a complement that ensures wide genetic diversity

Contrary to initial expectations, the BeONE dataset did not capture a wide genetic diversity for each of the four target species, which is a requirement for other project tasks. Therefore, additional samples were carefully selected among the WGS data publicly available at the beginning of the analysis

(November 2021) in ENA or the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), in order to ensure the representativeness of the genomic diversity within public databases (assessed in terms of sequence type or serotype, depending on the species) in the final dataset. The sequencing reads of these samples were subjected to the same quality check, genome assembly and curation methodologies as described above for the samples of the BeONE dataset. In the end, a so-called "public dataset" with the 8,383 samples that passed the curation step was released in Zenodo repository [1,874 *L. monocytogenes* (accession: 10.5281/zenodo.7116878); 1,434 *S. enterica* (accession: 10.5281/zenodo.7119735), 1,999 *E. coli* (accession: 10.5281/zenodo.7120057); 3,076 *C. jejuni* (accession: 10.5281/zenodo.7120166)].

Conclusions

The joint collaboration between the different BeONE partners led to the compilation of a controlled dataset for each of the four foodborne pathogens of interest, including sequencing data and metadata of samples from human, food, or/and veterinary sources, as well as sets of epidemiologically verified outbreak isolates. Considering the unpublished nature of some of the data, this dataset was fully anonymized, in compliance with the signed MTA. To ensure a wide genetic diversity, an additional dataset ("public dataset") was compiled for each of the species to be used as a complement to the BeONE dataset. Together, the BeONE dataset and the "public dataset" represent a useful resource not only for the different WPs of the BeONE project, but also for the whole scientific community.