Supplementary Data 1

Details on the study design

Table S1.1. Datasets and sub-datasets used for the cluster congruence analysis, with indication of the species, Zenodo repositories, number of isolates, type of pipeline in which they were used (allele-based or SNP-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	Allele	-
		ST 1	343	SNP	F2365
		ST 5	269	SNP	CP006592
		ST 6	501	SNP	CP006046
		ST 8	247	SNP	CP006862
		ST 121	240	SNP	HG813249
		ST combined	1,600	SNP	CP006046
S. enterica	10.5281/zenodo.7267785/ 10.5281/zenodo.7119735	All isolates	2,974	Allele	-
		Enteritidis	725	SNP	AM933172
		Typhimurium	880	SNP	AE006468
		Infantis	120	SNP	H15092067901-2
		Serotype combined	1,725	SNP	AE006468
E. coli	10.5281/zenodo.7267844/ 10.5281/zenodo.7120057	All isolates	2,307	Allele	-
		O157:H7	863	SNP	Sakai (NC_002695)
C. jejuni	10.5281/zenodo.7267879/ 10.5281/zenodo.7120166	All isolates	3,686	Allele	-
		ST 21	361	SNP	NCTC_11168
		ST 45	165	SNP	HG428754
		ST 48	173	SNP	CP006006
		ST 50	218	SNP	NCTC_11168
		ST 257	110	SNP	H055140547
		ST combined	1,027	SNP	NCTC_11168

*Retrieved from the list of SnapperDB references

Study design and strategy for congruence analysis

Multiple European laboratories, from different countries and sectors, employed, whenever possible, their different bioinformatics pipelines for genomic surveillance of foodborne bacterial pathogens (Figure 1) on four WGS datasets of *L. monocytogenes* (n = 3300 isolates [50,51]), *S. enterica* (n = 2974 isolates [52,53]), *E. coli* (n = 2307 isolates [54,55]), and *C. jejuni* (n = 3686 isolates [56,57]) (details in the Methods section). After Quality Control (QC), allele-based pipelines ran with the whole datasets, while SNP-based independently ran with sub-datasets of the top represented Sequence Types (STs) or serotypes of each species (Table 1) to better mimic their common application in the discrimination of closely related strains (e.g. same ST or outbreak-related strains). This collaborative effort covered a broad variety of "pipelines" (hereinafter used as a proxy for the combination of software pipeline and schema/reference), including the most commonly used cg/wgMLST schemas, allele/SNP-callers and clustering methods (Table 1).

In order to assess pipeline clustering congruence and comparability, it was essential to obtain clustering information at all possible distance thresholds for each pipeline (both allele- and SNP-based pipelines). Given the heterogeneity of "end-point" pipeline outputs, this harmonization was achieved by ReporTree [58], which also allowed the application of the clustering methods used by the original laboratory, either single-linkage hierarchical clustering (HC) or Minimum-Spanning Tree (MST) generation through MSTreeV2 model of GrapeTree (GT) [59] (Table 1, details in the Methods section). In addition, whenever an allele matrix was provided, clustering was performed with both algorithms, reinforcing the power and magnitude of the comparison (Table 1). We took advantage of this vast amount of data to conduct four relevant complementary analyses:

i. **Evaluation of allele-based clustering and comparison of stability regions.** This preliminary evaluation targeted the pipelines running the whole dataset (i.e. allele-based pipelines) and aimed at providing a global perspective on the resolution power of each pipeline and identifying ranges of allelic distance (AD) thresholds associated with cluster stability, i.e., subsequent thresholds in which clustering remains similar (potentially relevant for nomenclature design). For this, we assessed the number and composition of clusters obtained at progressively increasing distance thresholds and determined the neighborhood Adjusted Wallace coefficient (nAWC), which varies from 0 (no congruence) to 1 (absolute congruence), at consecutive thresholds ($n + 1 \rightarrow n$), based on a previously described approach [86-88].

ii. **Evaluation of allele-based clustering congruence with traditional typing data.** This assessment targeted the pipelines running the whole dataset (i.e. allele-based pipelines) and aimed at identifying the threshold levels (and stability regions) with the highest congruence with traditional typing (ST, Clonal Complex [CC] or serotype) and, when possible, with WGS-derived pathogen main lineages (here inferred with PopPUNK [72], whenever possible). To this end, for each comparison, we calculated the AWC between each threshold of the allele-based typing and these classifications, as a measure of the probability that two samples that cluster together using one method (at a given threshold level) also belong to the same lineage, ST, CC or serotype [88]. This was conducted between all possible threshold levels in both directions (method A → method B and method B → method A). In addition, for each

comparison, we also calculated the Adjusted Rand (AR) coefficient as a measure of the overall agreement between the typing methods [88]. The three values calculated for each comparison were then combined into a "Congruence Score" (CS) ($CS = AWC_{A \rightarrow B} + AWC_{B \rightarrow A} + AR$), which varies from 0 (no congruence) to 3 (absolute congruence). By providing a more intuitive interpretation of WGS-based typing outputs, this evaluation is expected to show how the cg/wgMLST clustering relates with historical typing data, thus facilitating the adoption of WGS-based surveillance by laboratories starting this technological transition.

iii. **Evaluation of cluster congruence between different pipelines at all threshold levels.** This evaluation was conducted for the whole dataset (involving the pairwise comparison of all allele-based pipelines) and for the sub-datasets of the top represented STs or serotypes of each species (involving the pairwise comparison of both allele- and SNP-based pipelines). We aimed to determine: i) the threshold levels with the highest concordance between the different pipelines (i.e., threshold levels/regions with similar clustering); ii) whether the same clustering results are obtained at similar threshold levels for both pipelines; and, iii) the existence of regions of stability common to all pipelines. For each pairwise pipeline comparison, we assessed the above-mentioned coefficients and CS through the comparison of all thresholds from the lowest to the highest resolution, and identified the thresholds with highest concordance between two pipelines ("corresponding points"). These comparisons provide key information to facilitate international data sharing and cooperation, and to smooth future workflow upgrades (either to accommodate technological advances or upcoming international recommendations).

iv. **Evaluation of pipeline performance in identifying potential outbreak-related clusters.** This analysis aimed at assessing congruence and threshold variability to detect the same potential outbreak-related clusters. For this, we identified the genetic clusters at potential outbreak-level using each allele-based pipeline and, subsequently, assessed the thresholds at which the same cluster composition was (if it was) found by the other ones. By targeting the main focus of routine surveillance (detection and survey of clusters of potential public health interest), this output and associated tools are expected to facilitate and increase confidence in the direct communication between laboratories during a real-life outbreak scenario.