



Interim summary report

EURL-*Salmonella* Proficiency Test Cluster Analysis 2020

Wilma Jacobs-Reitsma, RIVM, Bilthoven, the Netherlands
 Robin Diddens, RIVM, Bilthoven, the Netherlands
 Angela van Hoek, RIVM, Bilthoven, the Netherlands
 Kirsten Mooijman, RIVM, Bilthoven, the Netherlands

27 May 2021
 Z&O letter report 2021-0031

1. Introduction

This document provides an overview of the results as produced by the participants in the EURL-*Salmonella* Proficiency Test (PT) Typing 2020, concerning the optional part on the second pilot on Cluster Analysis (CA).

The evaluations of the individual laboratory results were sent to each of the participants separately. The full results will be reported in more detail in the final report on the EURL-*Salmonella* PT Typing 2020.

2. *Salmonella* strains for cluster analysis

A total of 10 *Salmonella* strains (shipped as SCA01–SCA10, but from now on indicated with 20SCA01 – 20SCA10) were sent to the participants in the EURL-*Salmonella* PT Typing 2020, part Cluster Analysis. Background information on the strains is given in Table 1.

Table 1. Background information on the *Salmonella* strains used for cluster analysis in 2020

Strain code	Serovar	ST	MLVA-profile	Origin
20SCA01 ^{a)} (=19SCA09)	4,[5],12:i:-	34	3-13-9-NA-211	Human
20SCA02 ^{a)}	4,5,12:i:-	34	3-14-9-NA-211	Human
20SCA03 ^{a)}	4,5,12:i:-	34	3-15-9-NA-211	Human
20SCA04 ^{a)}	4,12:i:-	34	3-14-13-NA-211	Human
20SCA05 ^{a)}	4,5,12:i:-	34	3-14-13-NA-211	Human
20SCA06 ^{a) b)}	4,5,12:i:-	34	3-14-13-NA-211	Human
20SCA07 (=19SCA07)	Typhimurium	19	5-9-14-9-211	Human
20SCA08 ^{a) b)}	4,5,12:i:-	34	3-14-13-NA-211	Human
20SCA09 ^{a)}	4,12:i:-	34	3-11-8-NA-211	Human
20SCA10 (=19SCA03)	Typhimurium	19	3-16-7-17-311	Human

^{a)} Typhimurium, monophasic variant as determined by PCR.

^{b)} Technical duplicates (in bold).



Strains were selected by the EURL-*Salmonella* to be suitable for analysis by using either MLVA or WGS. A set of 11 human surveillance strains collected and sequenced in 2020 and 4 strains from the PT Cluster Analysis 2019 were re-cultured from storage and submitted for MLVA and WGS analysis both directly and after re-culturing for 10 times. Subsequently, 9 strains

were selected for inclusion in the PT (also see Figure 1). The 10th strain was to be a technical duplicate; strain 20SCA06 and strain 20SCA08 shipment tubes were both prepared from the same blood-agar plate containing strain 20SCA06.

Cluster analysis was performed up to the choice of the participant by PFGE and/or MLVA and/or WGS (or any combination of these methods), and using their own routine method(s) of choice. However, the Protocol of the PT Typing 2020 already indicated that PFGE is no longer performed at the EURL-*Salmonella* and evaluation of PFGE results would only be based on comparing the results as sent in by PFGE participants.

The pilot PT Cluster Analysis 2020 was mimicking an outbreak situation, with a monophasic *Salmonella* Typhimurium ST34, MLVA type 3-14-13-NA-211 as the reference strain. Raw WGS data of this strain (fastq-files) were made available through a secure ftp server. For this particular PT2020 situation, the cluster definition was set at maximum 6 allelic differences from the reference sequence (REF). For MLVA, the cluster definition was set at no loci with a different number of repeats.

Participants were asked to analyse the 10 *Salmonella* strains and to report per strain if a clustering match with the reference strain was found or not.

Evaluation (per methodology) of the participants' cluster analysis results was done by comparing the participants' results to the expected results in the outbreak investigation setting, as pre-defined by the EURL-*Salmonella*.

No performance criteria were set for this second pilot PT on cluster analysis. As a minimum, it was expected that participants would report the technical duplicate strains 20SCA06 and 20SCA08 to be (part of) one cluster.

A total of 19 NRLs and 2 external partners participated in the cluster analysis, with 2 participants for PFGE analysis, 6 for MLVA analysis and 21 participants for WGS analysis (Table 2).

Table 2. Participation in Cluster Analysis in 2020, per method or combination of methods used

Participating in:			Number of participants	Laboratory codes
		WGS	15	1, 2, 3, 6, 12, 14, 18, 19, 21, 24, 25, 32, 34, 91, 96
	MLVA	WGS	4	8, 11, 28, 31
PFGE	MLVA	WGS	2	17, 33
Total PFGE:	Total MLVA:	Total WGS:	Total overall:	
2	6	21	21	

3. Evaluation of the cluster analysis results based on PFGE data

Only 2 participants (Laboratory codes 17 and 33) submitted results based on PFGE data and were using BioNumerics for the cluster analysis. The combined data sets in BioNumerics are shown in Figure 1. Similarity was calculated as recommended by EFSA (Jacobs et al., 2014) using the Dice coefficient, with both tolerance and optimization at 1,5%. As shown for both participants, this would lead to a clustering match of the REF strain 20SCA06 with strains 20SCA03, 20SCA04, 20SCA05, and 20SCA08 (the technical duplicate). This was correctly reported by participant 33 (Table 3), although this participant also remarked that adjusting the tolerance and optimization to 1%, strain 20SCA03 would not be included in the cluster anymore. By using these adjusted settings, the PFGE clustering would match with both the MLVA-based and the WGS-based clustering. Such a slight difference in settings may also explain the result as reported by participant 17 (Table 3).

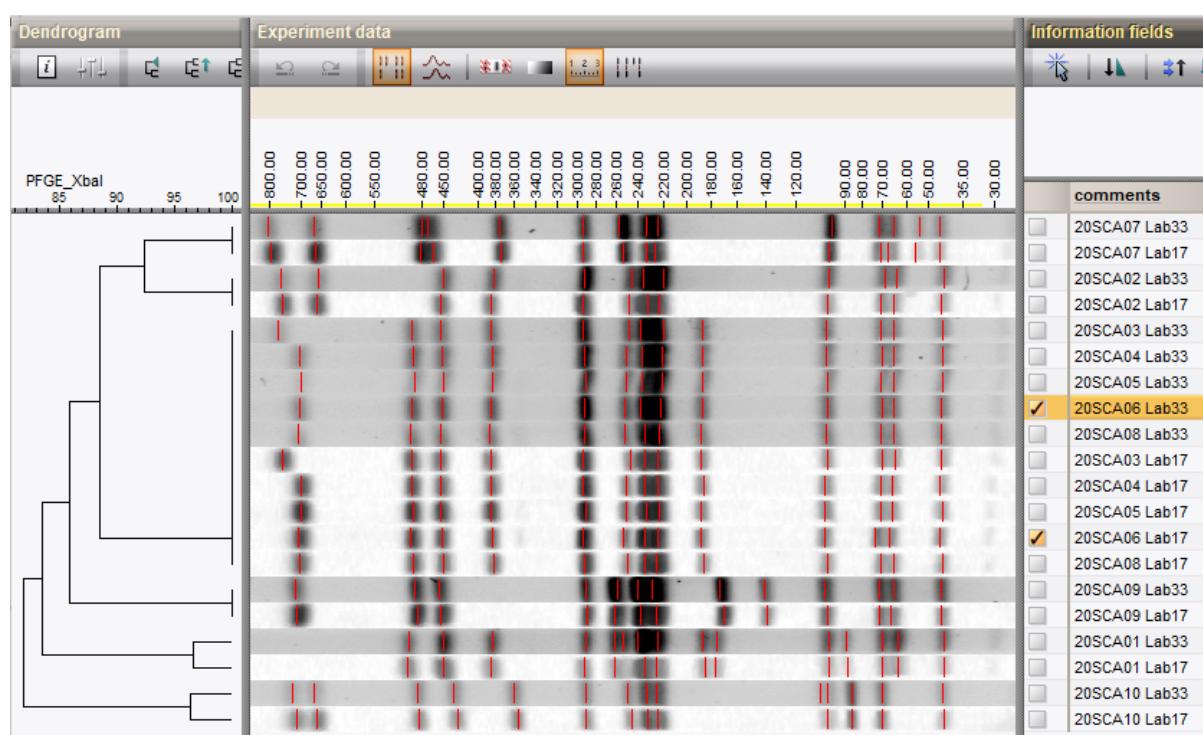


Figure 1. Cluster analysis based on PFGE data from Laboratory codes 17 and 33.

Table 3. Number of clusters, and their identification as reported by the 2 PFGE participants

Laboratory code	# Clusters reported	Cluster 1
17	1	SCA04;SCA05;SCA06;SC08
33	1	SCA03-SCA04-SCA05-SCA06-SCA08

4. Evaluation of the clusters analysis results based on MLVA data

Six participants (Laboratory codes 8, 11, 17, 28, 31, and 33) submitted cluster analysis results based on MLVA data.

The allelic profiles as submitted by the participants are given in Annex 1.



Participants were asked to report per strain (Table 4) if a clustering match was found with the reference outbreak strain (REF) in the EURL-*Salmonella* PT Typing 2020: monophasic *Salmonella* Typhimurium, ST34, MLVA type 3-14-13-NA-211.

The MLVA cluster definition for the PT Typing 2020 was set at no loci with a different number of repeats. Based on this cluster definition, MLVA-based results were expected to indicate strains 20SCA04, 20SCA05, SCA06 (reference strain) and 20SCA08 (technical duplicate of the reference strain) to be a clustering match with the reference outbreak strain REF as detailed in the PT Typing 2020.

Five participants (Laboratory codes 8, 17, 28, 31, and 33) out of the 6 submissions reported the MLVA-based cluster analysis results completely as expected.

Laboratory 11 reported incorrect results for strains 20SCA05 and 20SCA07, most likely due to a swap between those 2 strains (also see Annex 1).

Table 4. Expected cluster analysis results and the cluster analysis results as reported by the 6 MLVA participants

Labcode	20SCA01	20SCA02	20SCA03	20SCA04	20SCA05	20SCA06	20SCA07	20SCA08	20SCA09	20SCA10
Expected	No	No	No	Yes	Yes	Yes	No	Yes	No	No
8	No	No	No	Yes	Yes	Yes	No	Yes	No	No
11	No	No	No	Yes	No	Yes	Yes	Yes	No	No
17	No	No	No	Yes	Yes	Yes	No	Yes	No	No
28	No	No	No	Yes	Yes	Yes	No	Yes	No	No
31	No	No	No	Yes	Yes	Yes	No	Yes	No	No
33	No	No	No	Yes	Yes	Yes	No	Yes	No	No

5. Evaluation of the cluster analysis results based on WGS data

Twenty-one participants (Table 5) submitted cluster analysis results based on WGS data; 2 participants submitted both cgMLST-based and SNP-based data. Some details on the sequencing as performed by the participants are given in Annex 2.

All WGS-based pre-test results as well as the PT 2020 results from the EURL-*Salmonella* are shown in Figure 2. Sequencing was performed externally, on an Illumina NovaSeq platform. Raw data were processed via an in-house developed pipeline (assembly_pipeline: <https://github.com/Papos92>), which includes the SPAdes assembler. Cluster analysis was done in Ridom SeqSphere⁺, using the cgMLST Enterobase v2.0 scheme and visualised in a minimum spanning tree (MST, Figure 2).

The original WGS data from the human surveillance strains in 2020 and the PT 2019 in November 2019 are indicated with 20SCA_0, the WGS data for initial testing are indicated with 20SCA_1, and the WGS data after 10 times sub-culturing are indicated with 20SCA_2. The PT 2020 data (November 2020) are indicated without underscore. Stable strains were selected to be included for the PT Cluster Analysis 2020. In addition, the variable strain 19SCA03 from the PT 2019 was included in the PT 2020 as strain 20SCA10, still showing this variability along its history (Figure 2 and Figure 3).

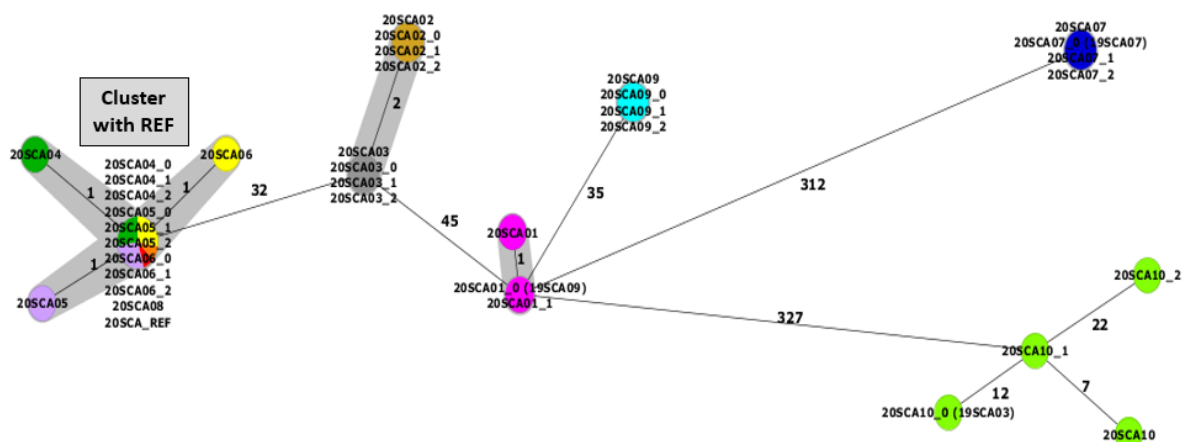


Figure 2. MST of the EURL-Salmonella pre-test and PT 2020 results, (RidomSeqSphere⁺, *S. enterica* MLST (7) and cgMLST (3002), pairwise ignoring missing values).

All but one participants' raw data (fastq files) were successfully processed through the assembly pipeline as mentioned. Raw data from participant 28 were processed using a Unicycler assembly pipeline ((Galaxy Version 0.4.8.0)), because this concerned single-end fastq files which cannot be analysed by our in-house assembly pipeline. All *de novo* assembled genomes (fasta files) were analysed in Ridom SeqSphere⁺, using the cgMLST Enterobase v2.0 and visualised in a MST (Figure 3). Data per strain are given in Annex 3.

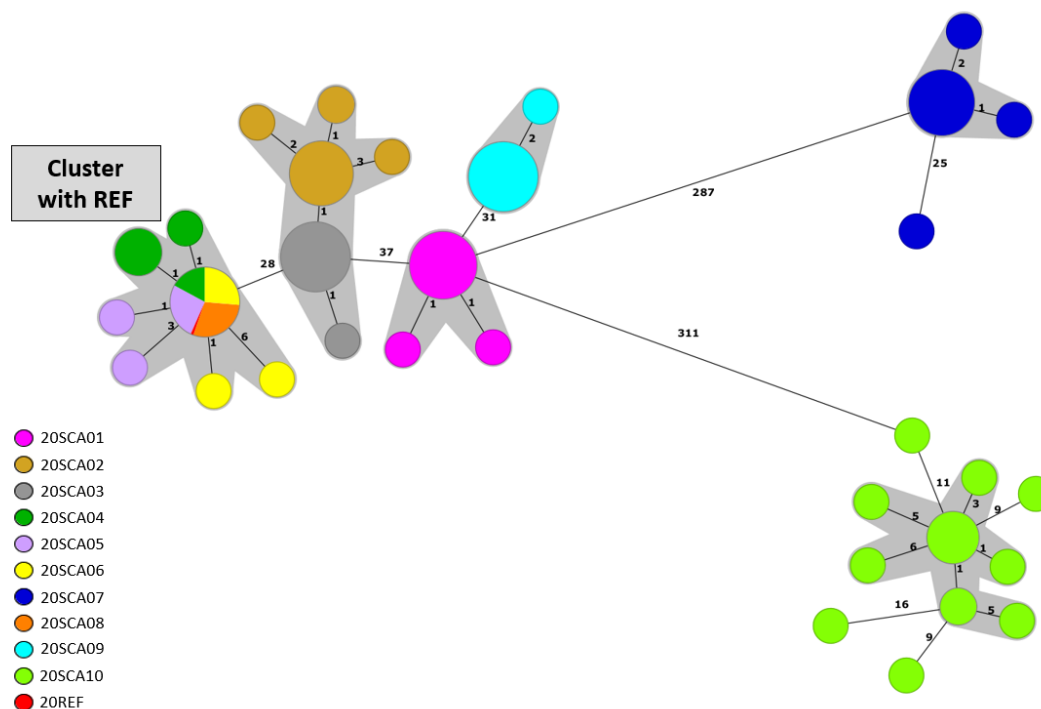


Figure 3. MST of all strains from all participants' processed raw data (Ridom SeqSphere⁺, *S. enterica* MLST (7) and cgMLST (3002), pairwise ignoring missing values)



Like in the PT Typing 2019 and in the pre-tests, strain 20SCA10 (originating from 19SCA03) showed quite some variation among the participants.

Participants were asked to report per strain (Table 5) if a clustering match was found with the reference outbreak strain (REF) in the EURL-Salmonella PT Typing 2020: 20SCA_REF_R1.fq.gz and 20SCA_REF_R2.fq.gz (monophasic *Salmonella* Typhimurium, ST34, MLVA type 3-14-13-NA-211).

The WGS cluster definition for the PT Typing 2020 was set at maximum 6 allelic differences from the reference (REF). Based on this cluster definition, WGS-based results were expected to indicate strains 20SCA04, 20SCA05, SCA06 (reference strain) and 20SCA08 (technical duplicate of the reference strain) to be a clustering match with the provided reference outbreak strain REF as detailed in the PT Typing 2020 (also see Figure 2).

All but 1 of the 23 submissions (2 participants with both a SNP-based and a cgMLST-based submission) reported the WGS-based cluster analysis results completely as expected (Table 5). Laboratory 32 reported strain 20SCA08 not to be clustering with the reference strain, but remarked that "I would from this analysis without any metadata also suggest strain SCA08 to possibly be part of the cluster due to 9 SNP differences". Notably, the cgMLST-based analysis on all participants' data showed no allelic differences for clustering strain 20SCA08 at all (Annex 2).

Two participants commented that strains 20SCA02 and 20SCA03 would fall into the definition of a second cluster (Figure 2 and Figure 3). Note that this was not the case when using the PFGE-based or MLVA-based cluster definitions (Figure 1 and Table 1).

Table 5. Expected cluster analysis results and the cluster analysis results as reported by the 21 WGS participants

Labcode	20SCA01	20SCA02	20SCA03	20SCA04	20SCA05	20SCA06	20SCA07	20SCA08	20SCA09	20SCA10
Expected	No	No	No	Yes	Yes	Yes	No	Yes	No	No
1	No	No	No	Yes	Yes	Yes	No	Yes	No	No
2-cgMLST	No	No	No	Yes	Yes	Yes	No	Yes	No	No
2-SNP	No	No	No	Yes	Yes	Yes	No	Yes	No	No
3	No	No	No	Yes	Yes	Yes	No	Yes	No	No
6-cgMLST	No	No	No	Yes	Yes	Yes	No	Yes	No	No
6-SNP	No	No	No	Yes	Yes	Yes	No	Yes	No	No
8	No	No	No	Yes	Yes	Yes	No	Yes	No	No
11	No	No	No	Yes	Yes	Yes	No	Yes	No	No
12	No	No	No	Yes	Yes	Yes	No	Yes	No	No
14	No	No	No	Yes	Yes	Yes	No	Yes	No	No
17	No	No	No	Yes	Yes	Yes	No	Yes	No	No
18	No	No	No	Yes	Yes	Yes	No	Yes	No	No
19	No	No	No	Yes	Yes	Yes	No	Yes	No	No
21	No	No	No	Yes	Yes	Yes	No	Yes	No	No
24	No	No	No	Yes	Yes	Yes	No	Yes	No	No
25	No	No	No	Yes	Yes	Yes	No	Yes	No	No
28	No	No	No	Yes	Yes	Yes	No	Yes	No	No
31	No	No	No	Yes	Yes	Yes	No	Yes	No	No
32	No	No	No	Yes	Yes	Yes	No	No	No	No
33	No	No	No	Yes	Yes	Yes	No	Yes	No	No
34	No	No	No	Yes	Yes	Yes	No	Yes	No	No
91	No	No	No	Yes	Yes	Yes	No	Yes	No	No
96	No	No	No	Yes	Yes	Yes	No	Yes	No	No



Annex 1 Expected and reported MLVA results for all 6 participants

Labcode	20SCA01	20SCA02	20SCA03	20SCA04	20SCA05	20SCA06 (REF)	20SCA07	20SCA08 (Ref)	20SCA09	20SCA10
Expected	3-13-9-NA-211	3-14-9-NA-211	3-15-9-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	5-9-14-9-211	3-14-13-NA-211	3-11-8-NA-211	3-16-7-17-311
8	3-13-9-NA-211	3-14-9-NA-211	3-15-9-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	5-9-14-9-211	3-14-13-NA-211	3-11-8-NA-211	3-16-7-17-311
11	3-13-9-NA-211	3-16-7-17-311	3-11-8-NA-211	3-14-13-NA-211	5-9-14-9-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-15-9-NA-211	3-14-9-NA-211
17	3-13-9-NA-211	3-14-9-NA-211	3-15-9-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	5-9-14-9-211	3-14-13-NA-211	3-11-8-NA-211	3-16-7-17-311
28	3-13-9-NA-211	3-14-9-NA-211	3-15-9-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	5-9-14-9-211	3-14-13-NA-211	3-11-8-NA-211	3-16-7-17-311
31	3-13-9-NA-211	3-14-9-NA-211	3-15-9-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-9-14-9-211	3-14-13-NA-211	3-11-8-NA-211	3-16-7-17-311
33	3-13-9-NA-211	3-14-9-NA-211	3-15-9-NA-211	3-14-13-NA-211	3-14-13-NA-211	3-14-13-NA-211	5-9-14-9-211	3-14-13-NA-211	3-11-8-NA-211	3-16-7-17-311

Loci reported in the order: STTR9, STTR5, STTR6, STTR10, STTR3.

Deviation from the expected result

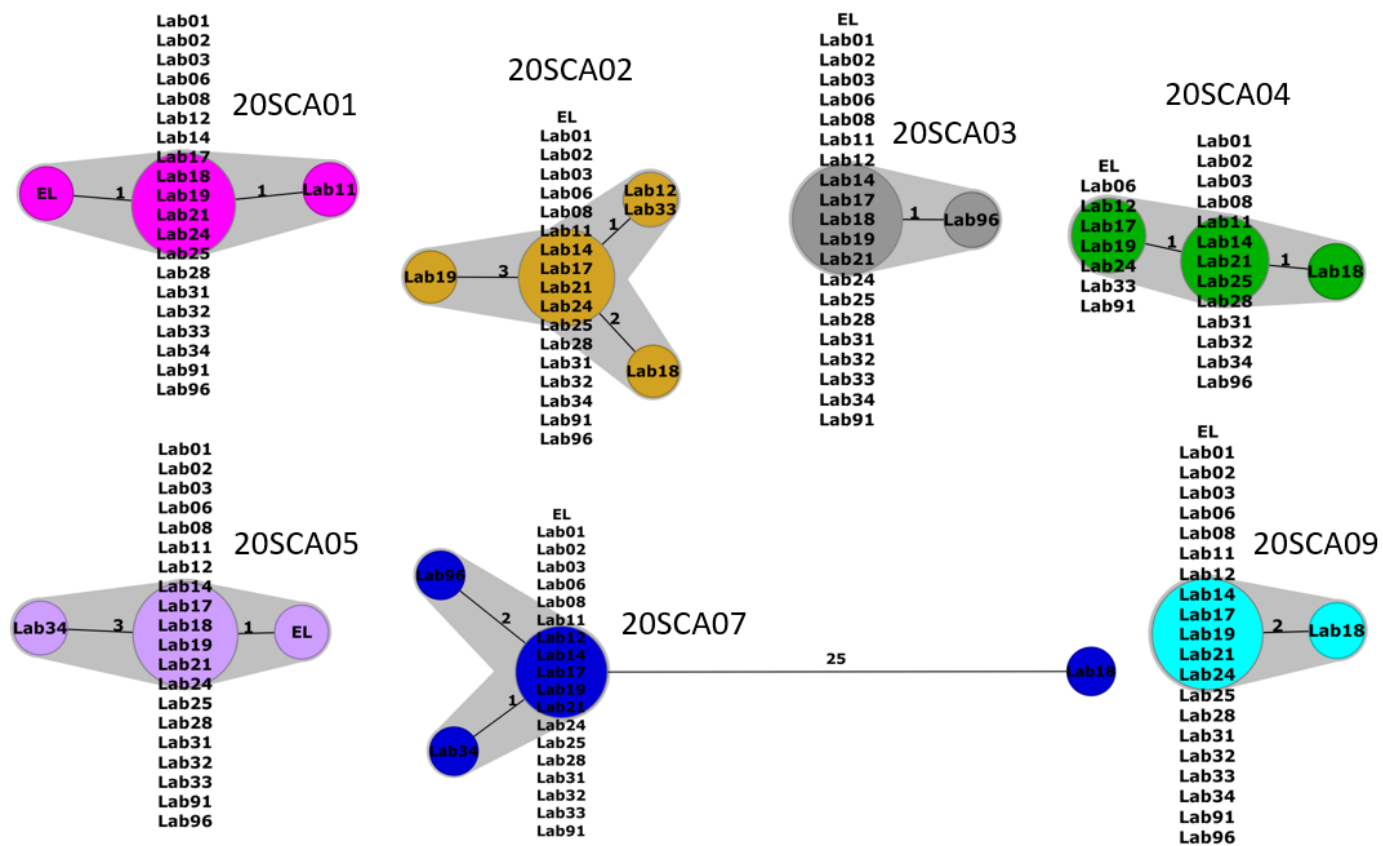


Annex 2 Sequencing details as reported by the 21 WGS participants

Labcode	DNA extraction, library preparation and sequencing performed	WGS platform used	Data analysis used	Tool used for analysis	Method used for cluster analysis
EURL-Salm	Outsourced	Illumina NovaSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
1	In-house	MiniSeq Illumina	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
2-cgMLST	DNA extraction : in-house ; Library and sequencing : outsourced	Illumina NovaSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
2-SNP	DNA extraction : in-house ; Library and sequencing : outsourced	Illumina NovaSeq	SNP-based - reference-based	in-house : iVarCall2	Maximum likelihood (ML)
3	In-house	Illumina NextSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
6-cgMLST	In-house	Illumina MiSeq	cgMLST-based	BioNumerics	Minimum Spanning Tree (MST)
6-SNP	In-house	Illumina MiSeq	SNP-based - reference-based	BioNumerics	Minimum Spanning Tree (MST)
8	In-house	Illumina MiSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
11	Outsourced	Illumina MiSeq	cgMLST-based	in house automated CHEWBACCA based pipeline	single linkage hierarchical clustering
12	In-house	Illumina NextSeq	cgMLST-based	inhouse automated CHEWBACCA based Pipeline	single linkage hierarchical clustering
14	DNA extraction in-house, WGS outsourcing	Illumina NovaSeq	cgMLST-based	BioNumerics	Minimum Spanning Tree (MST)
17	In-house	Illumina MiSeq	cgMLST-based	in-house Galaxy	Neighbor joining (NJ)
18	Outsourced	Illumina MiSeq	SNP-based - assembly-based		Neighbor joining (NJ)
19	DNA extraction: in-house, library prep and sequencing outsourced	NovaSeq6000	cgMLST-based	chewbbaca, used Salmonella.cgMLSTv2 from Enterobase	Neighbor joining (NJ)
21	In-house	Illumina MiSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
24	In-house	Illumina MiSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
25	In-house	Illumina MiSeq	cgMLST-based	chewBBACA, https://github.com/B-UMMI/chewBBACA	Minimum Spanning Tree (MST)
28	In-house	Illumina NextSeq	SNP-based - reference-based	CSI Phylogeny 1.4; https://cge.cbs.dtu.dk/services/CSIPhylogeny/	Maximum likelihood (ML)
31	In-house	Illumina MiSeq	SNP-based - reference-based	In-house pipeline	Minimum Spanning Tree (MST)
32	In-house	Illumina MiSeq	SNP-based - assembly-based	In house pipeline based on parSNP, Gubbins, creating a ML tree in IQTree, creating a SNP distance matrix with snp-dists (https://github.com/NorwegianVeterinaryInstitute/ALPPACA/wiki/Pipeline-and-program-descriptions)	Maximum likelihood (ML)
33	In-house	Illumina MiSeq	cgMLST-based	chewBBaca	Minimum Spanning Tree (MST)
34	In-house	Illumina MiSeq	SNP-based - reference-based	Snippy, Gubbins, Raxml, iTol	Maximum likelihood (ML)
91	In-house	Illumina HiSeq	SNP-based - reference-based	SNapper DB	Variant Call Format
96	Outsourced	Illumina NextSeq	cgMLST-based	https://chewbbaca.online/species/4/ ; https://github.com/B-UMMI/chewBBACA	MSTtree V2 GrapeTree https://github.com/achtman-lab/GrapeTree

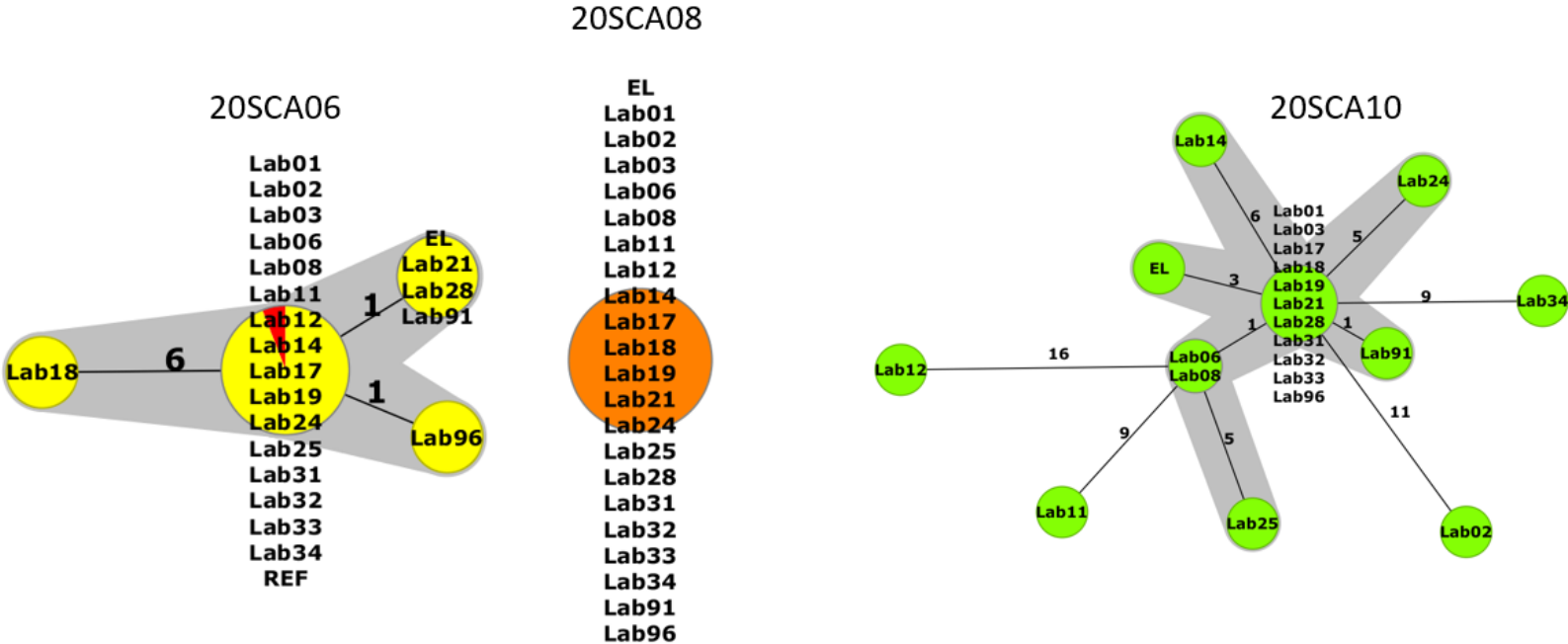


Annex 3 MSTs of each strain, using all participants' processed raw data (Ridom SeqSphere+, *S. enterica* MLST (7) and cgMLST (3002), pairwise ignoring missing values).





Annex 3 MSTs of each strain, using all participants' processed raw data (Ridom SeqSphere+, *S. enterica* MLST (7) and cgMLST (3002), pairwise ignoring missing values), continued.





References

ECDC (European Centre for Disease Prevention and Control), 2011. ECDC Technical Report, ECDC 2011-06, version 1.2. Laboratory standard operating procedure for MLVA of *Salmonella enterica* serotype Typhimurium. Available online:

https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/1109_SOP_Salmonella_Typhimurium_MLVA.pdf

Jacobs, W., S. Kuiling, K. van der Zwaluw, 2014. Molecular typing of *Salmonella* strains isolated from food, feed and animals: state of play and standard operating procedures for pulsed field gel electrophoresis (PFGE) and Multiple-Locus Variable number tandem repeat Analysis (MLVA) typing, profiles interpretation and curation. EFSA supporting publication 2014:EN-703, 74 pp. Available online:

http://www.efsa.europa.eu/sites/default/files/scientific_output/files/main_documents/703e.pdf

Wick, R.R., L.M. Judd, C.L. Gorrie, and K.E. Holt (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595. Available online:

<https://doi.org/10.1371/journal.pcbi.1005595>

List of abbreviations

cgMLST	core genome Multilocus Sequence Typing
EFSA	European Food Safety Authority
EL	EURL- <i>Salmonella</i> Laboratory
EU	European Union
EURL- <i>Salmonella</i>	European Union Reference Laboratory for <i>Salmonella</i>
MLVA	Multiple-Locus Variable number of tandem repeat Analysis
MST	Minimum Spanning Tree
NRLs- <i>Salmonella</i>	National Reference Laboratories for <i>Salmonella</i>
PFGE	Pulsed Field Gel Electrophoresis
REF	Reference
RIVM	National Institute for Public Health and the Environment
ST	Sequence Type
WGS	Whole Genome Sequencing

Contact for this PT

Wilma Jacobs-Reitsma: wilma.jacobs@rivm.nl

National Institute for Public Health and the Environment (RIVM)

Centre for Zoonosis and Environmental microbiology (Z&O/ internal mailbox 63)

Antonie van Leeuwenhoeklaan 9 P.O. Box 1, 3720 BA Bilthoven, The Netherlands

EURL-*Salmonella* website: www.eurlsalmonella.eu