Interim summary report

EURL-*Salmonella* Proficiency Test Cluster Analysis 2021

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

25 May 2022
Z&O letter report 2022-0046

1. Introduction

This document provides an overview of the results as produced by the participants in the EURL-*Salmonella* Proficiency Test (PT) Typing 2021, concerning the optional part on Cluster Analysis (CA).
A total of 19 NRLs participated in the cluster analysis; all 19 performed WGS analysis and 5 participants also performed MLVA analysis.
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 2021.

2. *Salmonella* strains for cluster analysis

A total of 10 *Salmonella* strains (21SCA01 – 21SCA10) were sent to the participants in the EURL-*Salmonella* PT Typing 2021, part CA. Background information on the strains is given in Table 1.

Table 1. Background information on the Salmonella *strains used for cluster analysis in 2021*

Strain code	Serovar	ST	MLVA-profile	Origin
21SCA01	Enteritidis	11	2-9-9-4-2	Human
21SCA02	Enteritidis	183	2-11-9-3-1	Human
21SCA03	Enteritidis	183	2-11-9-3-1	Human
21SCA04	Enteritidis	11	3-10-4-4-1	Human
21SCA05	Enteritidis	1925	3-10-5-4-1	Human
21SCA06 [a)]	Enteritidis	11	3-10-4-4-1	Human
21SCA07	Enteritidis	3406	2-14-NA-7-NA	Human
21SCA08	Enteritidis	11	3-10-4-4-1	Human
21SCA09 [a)]	Enteritidis	11	3-10-4-4-1	Human
21SCA10	Enteritidis	11	1-10-7-3-2	Human

[a)] 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 15 human surveillance strains collected and sequenced in 2019 were re-cultured from storage and submitted for MLVA and WGS analysis both directly and after sub-

culturing for 10 times. Subsequently, 9 strains were selected for inclusion in the PT (see also Figure 1). The 10th strain was to be a technical duplicate. Strain 21SCA06 and strain 21SCA09 shipment tubes were both prepared from the same blood-agar plate containing strain 21SCA06.

Cluster analysis was performed up to the choice of the participant by MLVA and/or WGS, and using their own routine method(s).
The pilot PT Cluster Analysis 2021 was mimicking an outbreak situation, with a *Salmonella* Enteritidis ST11, MLVA type 3-10-4-4-1 as the reference strain (21SCA-REF). Raw WGS data of this strain (21SCA-REF_R1.fq.gz and 21SCA-REF_R2.fq.gz) were made available through a secure ftp server.
For this particular PT2021 situation, the cluster definition was set at maximum 7 allelic differences from the reference sequence. 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 whether 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 specific performance criteria were set for this PT on cluster analysis. As a minimum, it was expected that participants would report the technical duplicate strains 21SCA06 and 21SCA09 to be (part of) one cluster.

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

Five participants (Laboratory codes 6, 7, 11, 23, 35) submitted cluster analysis results based on MLVA data.
The allelic profiles as submitted by the participants are given in Annex 1. Laboratory 7 reported a deviating result for strain 21SCA10.

Participants were asked to report per strain (Table 2) if a clustering match was found with the reference outbreak strain (21SCA-REF) in the EURL-*Salmonella* PT Typing 2021: *Salmonella* Enteritidis ST11, MLVA type 3-10-4-4-1.
The MLVA cluster definition for the PT Typing 2021 was set at no loci with a different number of repeats. Based on this cluster definition, MLVA-based results were expected to indicate strains 21SCA04, 21SCA06 (reference strain), 21SCA08 and 21SCA09 (technical duplicate of the reference strain) to be a clustering match with the reference outbreak strain as detailed in the PT Typing 2021.
All 5 participants reported the MLVA-based cluster analysis results completely as expected.

Table 2. Expected cluster analysis results and the cluster analysis results as reported by the 5 MLVA participants

Labcode	21 SCA01	21 SCA02	21 SCA03	21 SCA04	21 SCA05	21 SCA06[a)]	21 SCA07	21 SCA08	21 SCA09[a)]	21 SCA10
Expected	**No**	**No**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**Yes**	**No**
6	No	No	No	Yes	No	Yes	No	Yes	Yes	No
7	No	No	No	Yes	No	Yes	No	Yes	Yes	No
11	No	No	No	Yes	No	Yes	No	Yes	Yes	No
23	No	No	No	Yes	No	Yes	No	Yes	Yes	No
35	No	No	No	Yes	No	Yes	No	Yes	Yes	No

[a)] Technical duplicates.

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

Nineteen participants (Table 3) submitted cluster analysis results based on WGS data; 2 participants submitted both cgMLST-based and reference-based SNP data, and 1 participant submitted cgMLST-based as well as both assembly- and reference-based SNP data. Some details on the sequencing as performed by the participants are given in Annex 2.

WGS-based pre-test results as well as the PT 2021 results from the EURL-*Salmonella* are shown in Figure 1. Sequencing was performed in-house, on an Illumina NextSeq platform. Raw data were processed via an in-house developed Juno-assembly pipeline (https://rivm-bioinformatics.github.io/ids_bacteriology_man/juno-assembly.html), which includes the SPAdes 3.15.3 assembler. Cluster analysis was done in Ridom SeqSphere⁺, using the cgMLST Enterobase v2.0 scheme and visualised in a minimum spanning tree (MST, Figure 1).
The original WGS data from the human surveillance strains in 2019 are indicated with ELt0, the WGS data for initial testing are indicated with ELt1, and the WGS data after 10 times sub-culturing are indicated with ELt2. Strains with a stable and consistent cgMLST analysis result were selected to be included for the PT Cluster Analysis 2021. The PT 2021 set of strains were additionally tested both at the start of the PT (November 2021: ELt3) and at the end of the data submission period (February 2022: ELt4).

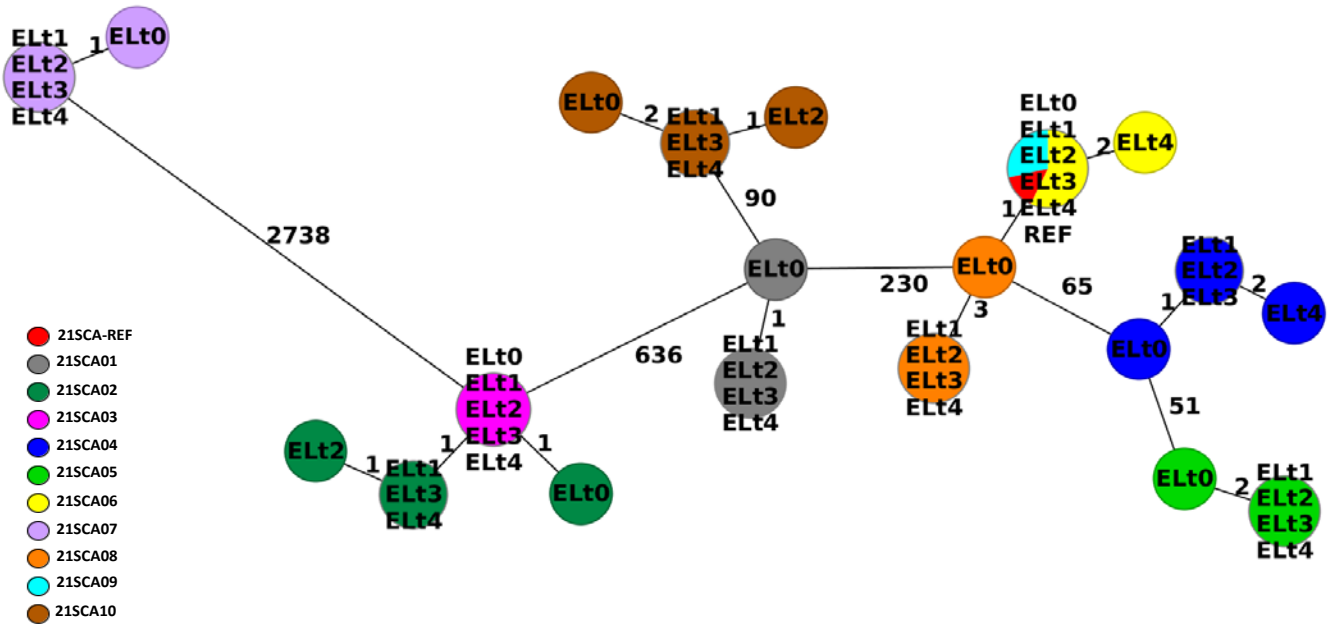


*Figure 1. MST of the EURL-*Salmonella *pre-test and PT 2021 results, (RidomSeqSphere⁺, cgMLST (3002), pairwise ignoring missing values).*

All participants' raw data (compressed fastq files) were successfully processed through the Juno-assembly pipeline as mentioned. The *de novo* assembled genomes (fasta files) were analysed in Ridom SeqSphere⁺, using the cgMLST Enterobase v2.0 and visualised in a MST (Figure 2). Data per strain are given in Annex 3. Data for strain 21SCA09-Lab21 were excluded due to a 85,2% of Good Targets, which is below the quality threshold.

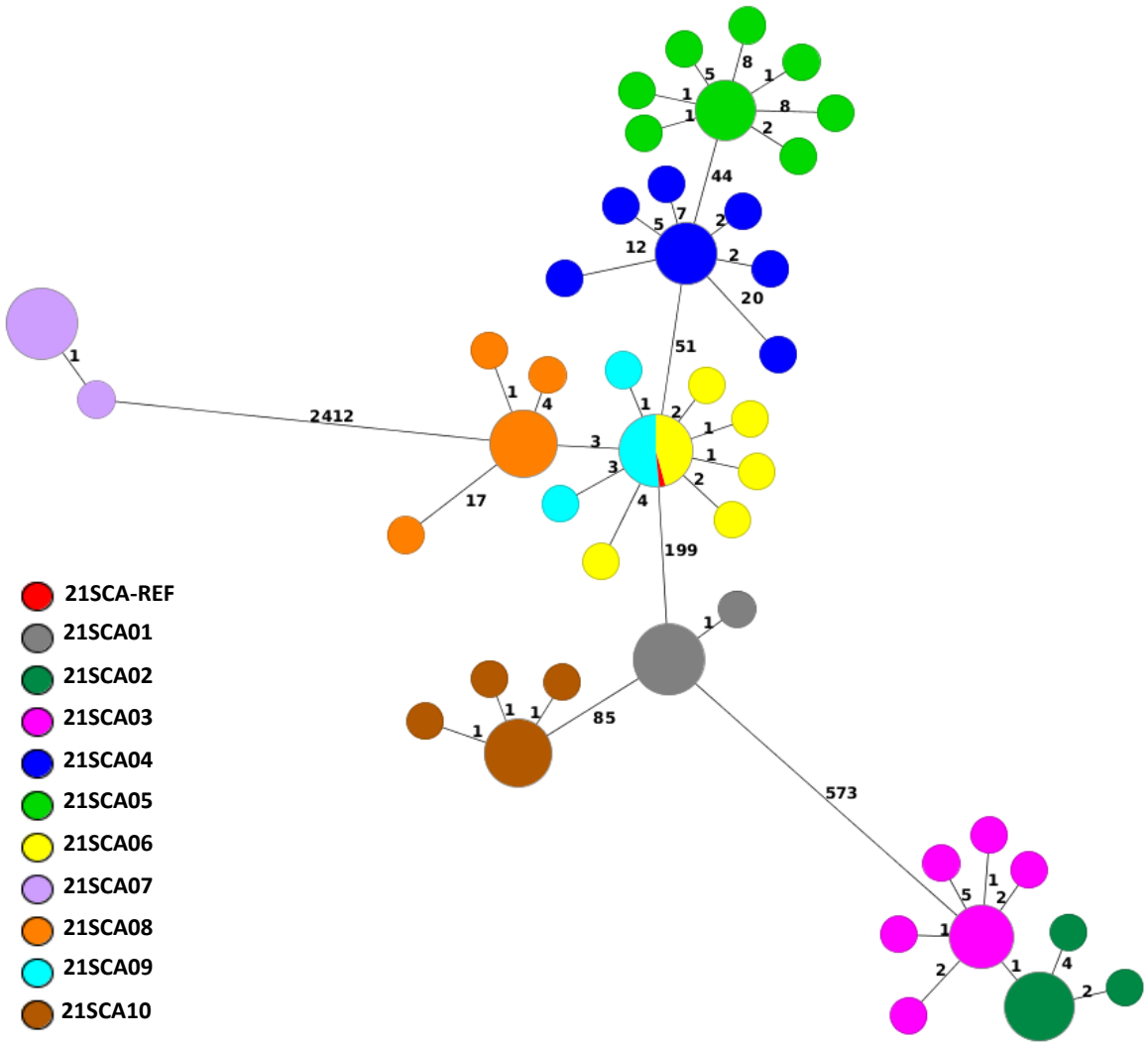


Figure 2. MST of the strains from the participants' processed raw data (Ridom SeqSphere⁺, cgMLST (3002), pairwise ignoring missing values)

Participants were asked to report per strain (Table 3) if a clustering match was found with the reference outbreak strain in the EURL-*Salmonella* PT Typing 2021: 21SCA-REF (*Salmonella* Enteritidis ST11, MLVA type 3-10-4-4-1).

The WGS cluster definition for the PT Typing 2021 was set at maximum 7 allelic differences from the reference. Based on this (cgMLST-)cluster definition, WGS-based results were expected to indicate strains 21SCA06 (reference strain), 21SCA08, and 21SCA09 (technical duplicate of the reference strain) to be a clustering match with the provided reference outbreak strain 21SCA-REF data as detailed in the PT Typing 2021 (also see Figure 1).

Fourteen of the 23 submissions (3 participants with multiple submissions) reported the WGS-based cluster analysis results completely as expected (Table 3).
Annex 4 shows per submission the participants' distance matrix data for their comparison of the 21SCA-REF with the 10 tested strains. Based on these distance matrix data, all but one (Lab 12) of the 13 cgMLST submissions were reported in accordance with the PT 2021 cluster definition of a maximum of 7 allelic differences, even though this subsequently ended up in a deviation from the expected result (Annex 4).

Because no cluster definition was specified for SNP-based analysis, the 10 SNP submissions were based on the participants' internal criteria. The apparent variety in these internal criteria may, partly, explain the differences in cluster analysis results reported for strain 21SCA08 (Annex 4).

Table 3. Expected cluster analysis results and the cluster analysis results as reported per data analysis method by the 19 WGS participants

Labcode-method	21 SCA01	21 SCA02	21 SCA03	21 SCA04	21 SCA05	21 SCA06[a)]	21 SCA07	21 SCA08	21 SCA09[a)]	21 SCA10
Expected	**No**	**No**	**No**	**No**	**No**	**Yes**	**No**	**Yes**	**Yes**	**No**
1-SNPa	No	No	No	No	No	Yes	No	No	Yes	No
2-SNPa	No	No	No	No	No	Yes	No	Yes	Yes	No
6-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
6-SNPa	No	No	No	No	No	Yes	No	Yes	Yes	No
6-SNPr	No	No	No	No	No	Yes	No	Yes	Yes	No
7-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
10-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
10-SNPr	No	No	No	No	No	Yes	No	Yes	Yes	No
11-SNPr	No	No	No	No	No	Yes	No	Yes	Yes	No
12-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
14-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
16-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
19-cgMLST	No	No	No	No	No	Yes	No	No	Yes	No
19-SNPr	No	No	No	No	No	Yes	No	No	Yes	No
21-SNPr	No	No	No	No	No	Yes	No	No	Yes	No
22-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
23-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
24-cgMLST	No	No	No	No	No	Yes	No	Yes	Yes	No
26-cgMLST	No	No	No	No	No	Yes	No	No	Yes	No
27-SNPr	No	No	No	No	No	Yes	No	No	Yes	No
30-SNPa	No	No	No	No	No	Yes	No	No	Yes	No
34-cgMLST	No	No	No	No	No	Yes	No	No	Yes	No
35-cgMLST	No	No	No	No	No	Yes	No	No	Yes	No

[a)] Technical duplicates. SNPa: assembly-based SNP data, SNPr: reference-based SNP data.

Deviation from the expected result

Annex 1 Expected and reported MLVA results for all 5 participants

Labcode	21SCA01	21SCA02	21SCA03	21SCA04	21SCA05
Expected	**2-9-9-4-2**	**2-11-9-3-1**	**2-11-9-3-1**	**3-10-4-4-1**	**3-10-5-4-1**
6	2-9-9-4-2	2-11-9-3-1	2-11-9-3-1	3-10-4-4-1	3-10-5-4-1
7	2-9-9-4-2	2-11-9-3-1	2-11-9-3-1	3-10-4-4-1	3-10-5-4-1
11	2-9-9-4-2	2-11-9-3-1	2-11-9-3-1	3-10-4-4-1	3-10-5-4-1
23	2-9-9-4-2	2-11-9-3-1	2-11-9-3-1	3-10-4-4-1	3-10-5-4-1
35	2-9-9-4-2	2-11-9-3-1	2-11-9-3-1	3-10-4-4-1	3-10-5-4-1

Labcode	21SCA06[a)]	21SCA07	21SCA08	21SCA09[a)]	21SCA10
Expected	**3-10-4-4-1**	**2-14-NA-7-NA**	**3-10-4-4-1**	**3-10-4-4-1**	**1-10-7-3-2**
6	3-10-4-4-1	2-14-NA-7-NA	3-10-4-4-1	3-10-4-4-1	1-10-7-3-2
7	3-10-4-4-1	2-14 -NA-7-NA	3-10-4-4-1	3-10-4-4-1	NA-10-7-3-2
11	3-10-4-4-1	2-14-NA-7-NA	3-10-4-4-1	3-10-4-4-1	1-10-7-3-2
23	3-10-4-4-1	2-14-0-7-0	3-10-4-4-1	3-10-4-4-1	1-10-7-3-2
35	3-10-4-4-1	2-14-NA-7-NA	3-10-4-4-1	3-10-4-4-1	1-10-7-3-2

[a)] Technical duplicates.

Loci reported in the order: SENTR7-SENTR5-SENTR6-SENTR4-SE-3.

Deviation from the expected result

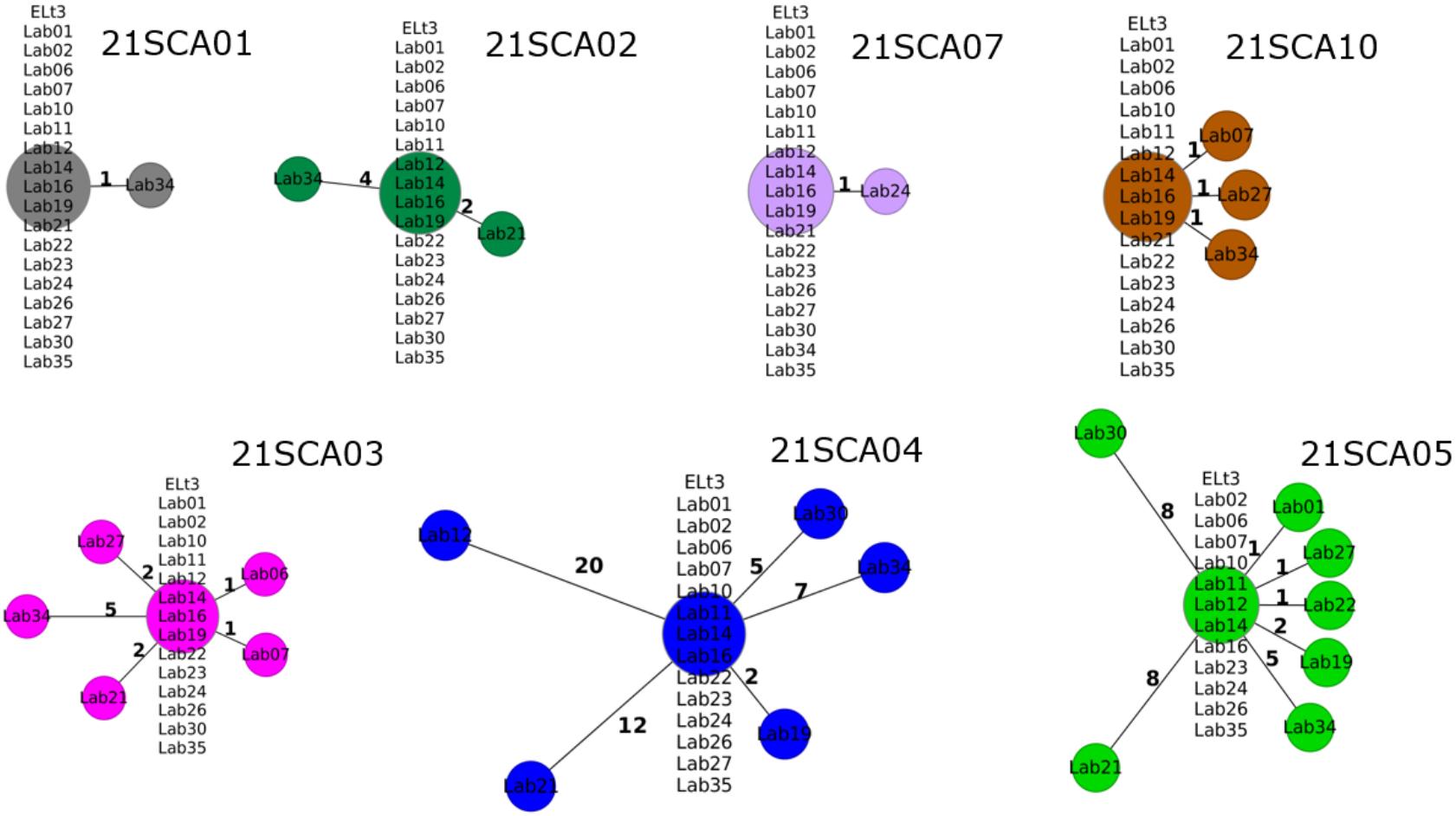
Annex 2 Sequencing details as reported by the 19 WGS participants

Labcode	Wet lab [a)]	WGS platform	Data analysis	Tool for analysis	Method for cluster analysis
1	In-Out-Out	Illumina NovaSeq	SNP-based - assembly-based	CSIPhylogeny 1.4	Maximum likelihood (ML)
2	In-In-In	Illumina MiSeq	SNP-based - assembly-based	In house pipeline[b)]	Maximum likelihood (ML)
6-cgMLST	In-Out-Out	Illumina MiSeq	cgMLST-based	PyMLST v1	hierarchical clustering
6-SNPa	In-Out-Out	Illumina MiSeq	SNP-based - assembly-based	Roary and Prank	Maximum likelihood (ML)
6-SNPr	In-Out-Out	Illumina MiSeq	SNP-based - reference-based	BWA, bcftools, RAxML	Maximum likelihood (ML)
7	In-In-In	Illumina MiSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
10-cgMLST	In-Out-Out	Illumina NovaSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
10-SNPr	In-Out-Out	Illumina NovaSeq	SNP-based - reference-based	in-house pipeline iVARCall2	Maximum likelihood (ML)
11	In-In-Out	Illumina NextSeq	SNP-based - reference-based	MINTyper 1.0	Minimum Spanning Tree (MST)
12	In-In-In	Illumina MiSeq	cgMLST-based	Linux cgmlst finder	Minimum Spanning Tree (MST)
14	In-In-In	Illumina MiniSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
16	In-In-In	Illumina MiSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
19-cgMLST	In-In-In	Illumina NextSeq	cgMLST-based	inhouse automated ChewieSnake Pipeline	single linkage hierarchical clustering
19-SNPr	In-In-In	Illumina NextSeq	SNP-based - reference-based	inhouse automated SnippySnake Pipeline	single linkage hierarchical clustering
21	In-In-In	Illumina MiSeq	SNP-based - reference-based	In-house pipeline	Minimum Spanning Tree (MST)
22	In-In-In	Illumina NextSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)
23	In-In-In	Illumina MiSeq	cgMLST-based	ChewBBaCa	Minimum Spanning Tree (MST)
24	In-In-In	Illumina MiSeq	cgMLST-based	BioNumerics	Minimum Spanning Tree (MST)
26	In-Out-Out	Illumina NovaSeq	cgMLST-based	chewbbaca	Neighbor joining (NJ)
27	In-In-In	Illumina NextSeq	SNP-based - reference-based	Snippy, Gubbins	Maximum likelihood (ML)
30	In-In-In	Illumina MiSeq	SNP-based - assembly-based	CGE CSIPhylogeny 1.4	ML included in CSYPhylogeny
34	In-In-In	Illumina NextSeq	cgMLST-based	EnterobaseChewBBACA	MSTreeV2
35	In-In-In	Illumina MiSeq	cgMLST-based	in-house Galaxy pipeline	Neighbor joining (NJ)
EURL-Salm	In-In-In	Illumina NextSeq	cgMLST-based	Ridom SeqSphere	Minimum Spanning Tree (MST)

[a)] Wet lab preparations: DNA extraction, Library preparation, sequencing. IN: In-house, Out: Outsourced.

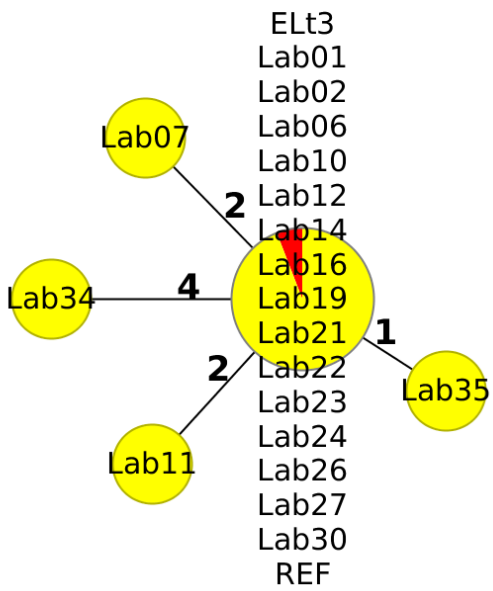
[b)] based on parSNP, Gubbins, creating a ML tree in IQTree, creating a SNP distance matrix with snp-dists (https://github.com/NorwegianVeterinaryInstitute/ALPPACA/wiki/Pipeline-and-program-descriptions).

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

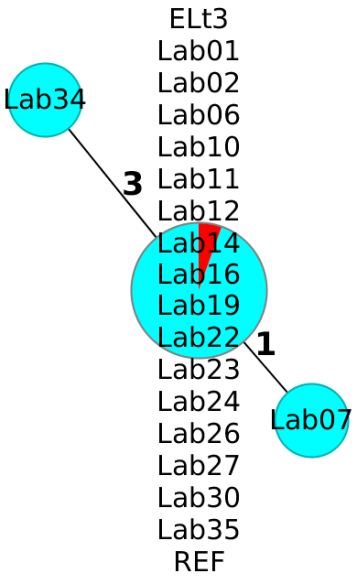


Annex 3 MSTs of each strain, using all participants' processed raw data (Ridom SeqSphere+, cgMLST (3002), pairwise ignoring missing values), continued.
Data for strain 21SCA09-Lab21 were excluded due to a 85,2% of Good Targets, which is below the quality threshold.

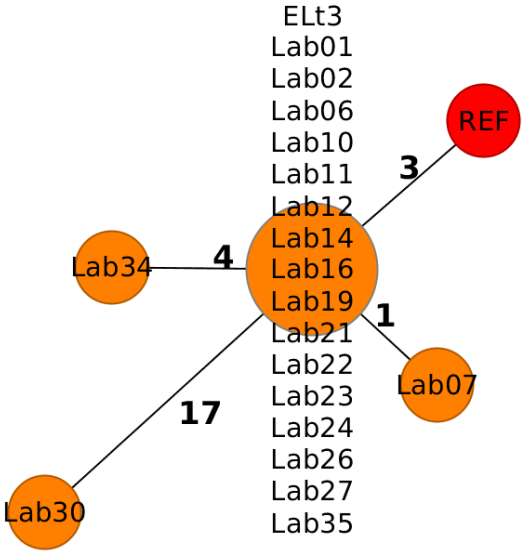
21SCA06 plus REF



21SCA09 plus REF



21SCA08 plus REF



Annex 4 Per submission, the participants' distance matrix data for their comparison of the 21SCA-REF with the 10 tested strains. Last column: reported cluster analysis result for strain 21SCA08 (also see Table 3).

Labcode-method	21SCA-REF	21SCA01	21SCA02	21SCA03	21SCA04	21SCA05	21SCA06	21SCA07	21SCA08	21SCA09	21SCA10		21SCA08
1-SNPa	0	427	1444	1444	136	152	5	29352	11	4	444		No
2-SNPa	0	427	1461	1461	133	149	1	40860	10	1	450		Yes
6-SNPa	0	288	974	977	97	106	0	29800	3	0	302		Yes
30-SNPa	0	420	1435	1436	142	166	0	29291	44	2	441		No
6-SNPr	0	453	1563	1561	140	157	0	40941	9	0	481		Yes
10-SNPr	0	455	1571	1567	136	155	2	45908	10	1	476		Yes
11-SNPr	0	383	1361	1360	129	147	0	29299	9	0	414		Yes
19-SNPr	0	431	1464	1461	133	146	0	44867	9	0	449		No
21-SNPr	0	376	1265	1262	105	126	0	37250	8	0	390		No
27-SNPr	0	401	1378	1378	126	141	1	39910	11	1	424		No
6-cgMLST	0	233	682	680	69	83	0	2745	6	0	234		Yes
7-cgMLST	0	232	681	681	67	82	0	2749	4	0	232		Yes
10-cgMLST	0	232	680	680	67	81	0	2746	4	0	232		Yes
12-cgMLST	0	243	707	708	76	89	2	2767	9	3	248		Yes
14-cgMLST	0	232	680	680	67	81	0	2745	4	0	232		Yes
16-cgMLST	0	232	681	680	67	81	0	2750	4	0	232		Yes
19-cgMLST	0	233	691	692	70	85	1	2716	9	1	238		No
22-cgMLST	0	232	680	681	67	82	0	2750	4	0	232		Yes
23-cgMLST	0	228	691	691	69	85	0	2703	6	0	232		Yes
24-cgMLST	0	200	200	200	69	86	0	200	5	0	200		Yes
26-cgMLST	0	314	928	926	109	127	4	3467	8	3	318		No
34-cgMLST	0	236	697	700	72	86	1	2737	10	2	241		No
35-cgMLST	0,0000	0,0825	0,2425	0,2415	0,0257	0,0307	0,0000	0,9289	0,0039	0,0000	0,0825		No
ELt3-cgMLST	0	232	681	681	67	81	0	2750	4	0	232		Yes

Deviation from the expected result

Deviation based on internal result

References

European Centre for Disease Prevention and Control. Laboratory standard operating procedure for multiple-locus variable-number tandem repeat analysis of *Salmonella enterica* serotype Enteritidis. Stockholm:ECDC; 2016. doi 10.2900/973540. Available online: https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/Salmonella-Enteritidis-Laboratory-standard-operating-procedure.pdf

Acknowledgements

We would like to thank the IDS-bioinformatics team for their valuable help with the Juno-assembly pipeline.

List of abbreviations

cgMLST	core genome Multilocus Sequence Typing
EFSA	European Food Safety Authority
EL	EURL-*Salmonella* Laboratory
EU	European Union
EURL-*Salmonella*	European Union Reference Laboratory for *Salmonella*
MLVA	Multiple-Locus Variable number of tandem repeat Analysis
MST	Minimum Spanning Tree
NRLs-*Salmonella*	National Reference Laboratories for *Salmonella*
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