Benchmarking of genotypic *Salmonella* serotype prediction complying to the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain – Method validation – Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures)

Report number	#3
Responsible	Eelco Franz (RIVM) and Pimlapas Leekitcharoenphon (DTU)
Other partners/institutions involved	Pimlapas Leekitcharoenphon (DTU), Liljana Petrovska (APHA), Kirsten Mooijman (RIVM), Eelco Franz (RIVM), Susanne Karlsmose Pedersen (DTU), Rene S. Hendriksen (DTU), Angela van Hoek (RIVM), Indra Bergval (RIVM), Rolf Sommer Kaas (DTU)
Benchmarking launched (date)	April 2017
Deliverable due (date)	June 2017

Purpose

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools for the *in silico* prediction of *Salmonella* serovars from raw whole genome sequencing data. The set-up of the interlaboratory (benchmarking) study complied with the Draft International Standard ISO 16140-6 (ISO/DIS 16140-6:2017 Microbiology of the food chain – Method validation – Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures).

Participants

Participants in this benchmarking exercise were institutions from the ENGAGE network, including also participation from EFSA representatives, and from RIVM.

Thirteen sets of results were submitted from the following institutions:

APHA (United Kingdom), BfR (Germany), DTU (Denmark), EFSA (2 sets of results), IZSLT (Italy), IZSVe (Italy), NIPH-NIH (Poland), NVRI (Poland), PHE (United Kingdom), RIVM (the Netherlands) (3 sets of results).

Participating institutes are identified by codes (1-13, see below) and each code is known only by the corresponding laboratory. The full list of laboratory codes is known only by the organizers (DTU).

Tools benchmarked

Benchmarking exercise component to determine species using the following tools and setup (each number refers to the corresponding participant and the (combination of) tools they used):

1. KmerFinder 2.1 (through Batch upload https://cge.cbs.dtu.dk/services/cge/index.php)
2. KmerFinder 2.0 (unix command line version, integrated in RIVM pipeline)
3. kmerid (PHE tool) tag version 2-1
4. CGE Tools (SPAdes 3.9, Assembler 1.2; SpeciesFinder 1.2; KmerFinder 2.0) (https://cge.cbs.dtu.dk/services/)
5. CGE KmerFinder 2.0. Scored method: winner takes it all. CGE SPADES 3.9 assembled sequences
6. SISTR (https://lfz.corefacility.ca/sistr-app/) 1.0.1; CGE KmerFinder 2.0 scored method: winner takes it all
7. CLC Genomics Workbench 9 & Species Finder 1.2 (https://cge.cbs.dtu.dk/services/SpeciesFinder/)
8. CLC Genomics Workbench 9 and SISTR app
9. CGE, blastn

10. Kraken version 0.10.5-beta, with MiniKraken DB. Options used: --fastq-input --gzip-compressed --quick --preload --paired
11. Blastn 2.3.0 -evalue 0.001 -outfmt "6 qseqid qlen sseqid sacc slen qstart qend sstart send evalue bitscore length pident mismatch gaps staxid sscinames" -perc_identity 95 -max_target_seqs 2 -qcov_hsp_perc 80 -db NT; for species confirmation: KmerFinder 2.0 (default options)
12. KmerFinder 2.0, scoring: Winner takes all, db: bacteria
13. KmerFinder 2.1 (BatchUpload of assembled data)

Benchmarking determining the *Salmonella* serovar genotypically using the following tools and setup:

1. SeqSero 1.0 (http://www.denglab.info/SeqSero) Reads paired-end
2. SISTR_cmd 0.3.6 (unix command line tool based on SISTR, integrated in our pipeline)
3. MOST (PHE tool) tag version 2-8, SeqSero (for antigenic formula) [-m 2 -b mem]
4. CGE Tools (SeqSero 1.2)
5. CGE SeqSero 1.2. We submitted raw sequences
6. SISTR (https://lfz.corefacility.ca/sistr-app/) v1.0.1
7. Seqsero 1.0 Genome Assembly and Species Finder 1.2
8. SISTR app
9. CGE (SeqSero, mlst)
10. SeqSero 1.0. Options used: -m2 (for pair-end reads), -b sam (for bwa samse/sampe)
11. SeqSero 1.0 -m2; for serotype confirmation: SISTR v0.3.4, --qc --no-cgmlst -f tab -o sistr-output.tab
12. SalmonellaTypeFinder 1.3
13. SeqSero 1.2 (paired end reads)

For further information on the serotyping tools, please see Appendix F – Benchmarking of genotypic *Salmonella* serotype prediction (general).

Genomes of bacterial species and *Salmonella* serovars

According to ISO/DIS 16140-6, the following number and type of strains have to be tested, per laboratory, in an interlaboratory study (ILS) when validating an alternative serotyping method (ISO/DIS 16140-6 includes protocols for validation of alternative confirmation and typing procedures i.e. including also serotyping) for *Salmonella*: 16 different strains from target serovars, 4 strains from non-target serovars within target subspecies and 4 strains from non-target genus. In this ILS, 27 genomes without any pre-assembly or trimming of the following strains were tested (the strains in this study were not part of ENGAGE project):

- 18 isolates of 6 target *Salmonella* serovars:
    - Enteritidis (n=3), Hadar (n=3), Infantis (n=3), monophasic Typhimurium (n=3), Typhimurium (n=3), Virchow (n=3).
- 5 non-target *Salmonella* serovars:
    - Derby, Dublin, Kentucky, Mbandaka, Stanley.
- 4 strains from the same family (Enterobacteriaceae) but non-target genus:
    - *Citrobacter freundii*, *Escherichia coli*, *Klebsiella pneumoniae*, *Shigella flexneri.*

The genomic quality based on number of reads, N50, number of contigs and total base pairs of each strain was assessed (Table 6 – List of selected genomes) and they all were of good quality (genomic quality data can be found in the Supplementary Table 4 (Annex D)).

The National Institute for Public Health and the Environment (RIVM, the Netherlands), Centre for Zoonoses and Environmental Microbiology provided the genomes of six *Salmonella* genomes, serotyped by conventional methods, and one *E. coli* genome.

The Animal and Plant Health Agency (APHA, United Kingdom) provided seven *Salmonella* genomes, serotyped by conventional methods, one *Citrobacter freundii* genome and one *Klebsiella pneumoniae* genome, respectively.

The National Food Institute (DTU Food, Denmark) provided the 10 serotyped *Salmonella* genomes, serotyped by conventional methods and one *Shigella flexneri* genome.

All genomes were sequenced on either an Illumina MiSeq or Illumina HiSeq.

Overall results

The results were divided into the species and serovar predictions and correlated with the expected species and serovar (Tables 1 and 2 and Figures 1 and 2). The results that did not correlate with the expected result were further divided into predictions that give a different species and serovar than the expected (miscorrelation, Figures 1 and 2), predictions that yield no result (no prediction, Figures 1 and 2), and predictions that yield several possible serovars (ambiguous, Figures 1 and 2). Results are described into more detail in the Supplementary Table 4 (Annex D)).

Table 1. Correlation of in silico species prediction with conventional methods. Numbers represent the number of isolates. Total number of isolates is 27. Numbers in the header of the columns correspond to the listed participants for species prediction.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Correlation	25	25	26	24	24	25	25	26	26	25	26	25	25
No Correlation													
- Miscorrelation	2	2	1	2	1	1	2	1	1	2	1	1	2
- No prediction	0	0	0	1	2	1	0	0	0	0	0	1	0
- Ambiguous	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2. Correlation of in silico serovar prediction with conventional serotyping methods. Numbers represent the number of isolates. Total number of Salmonella *isolates is 23. Numbers in the header of the columns correspond to the listed participants for serovar prediction.*

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Correlation	20	22	22	17	19	22	16	22	20	20	22	22	20
No Correlation													
- Miscorrelation	1	1	1	0	1	1	1	1	1	1	1	1	1
- No prediction	0	0	0	4	1	0	6	0	0	0	0	0	0
- Ambiguous	2	0	0	2	2	0	0	0	2	2	0	0	2

Correlation for species prediction of all participants was more than 88%. Most of the tools failed to predict *Shigella flexneri* but identified *Shigella sonnei* instead. Almost all the tools predicted all *Salmonella enterica* correctly. The exception was ENGAGE-BM-16 which participants 4 and 5 did not predict correctly. Correlation of serovar prediction was between 74% and 96% The tools that resulted in a 96% correlation were SISTR (v1.0.1 and v0.3.6), SeqSero 1.0 (command line version), SalmonellaTypeFinder 1.3 and MOST. Most tools predicted *S.* Hadar (ENGAGE-BM-14) as *S.* Eko, meaning that these tools agreed upon the predicted serovar. Either conventional serotyping misclassified this isolate or the incorrect fastq files were added to the test panel. Many tools predicted *S.* Hadar (ENGAGE-BM-11 and ENGAGE-BM-13) ambiguously as *S.* Hadar/*S.* Istanbul. Colony form variation (the variable expression of minor antigens by different single-colony picks from the same strain) may occur with the expression of the O:6_1 antigen by some serogroup C2 serovars (Hendriksen et al., 2009; Popoff, 2001). SeqSero was the most used tool, however the correlation between the different versions (web-based or command line)/modes of input data (raw reads or assembled genomes) varied from 74.1% to 96.3%. This variation might be due to the choice of assembly tools, different options/parameters in web-based and command line version and to the operator. The second most used tool was SISTR that resulted in a 96.3% correlation.

Additionally, the results were evaluated following the data analysis and interpretation described in ISO/DIS 16140-6:2017. For this evaluation, the reference and alternative methods were compared for the target strains as well as for the non-target strains (inclusivity and exclusivity study, see Table 3).

Table 3. Comparison and interpretation of results between the reference and alternative methods for the inclusivity study (target strains) and for the exclusivity study (non-target strains)

Result of the (reference or alternative) method per strain		***Interpretation***
Reference confirmation procedure	**Alternative confirmation method**	**Alternative confirmation method compared to reference confirmation procedure** [a]
+	+	PA
+	-	ND
-	+	PD
-	-	NA

[a] PA: Positive agreement; ND: Negative deviation; PD: Positive deviation; NA: Negative agreement

The results of the inclusivity and exclusivity analysis were compared to the acceptability limits (AL) indicated in ISO/DIS 16140-6:2017 (these acceptability limits are based on expert opinions), and are summarized in Table 4 (species level) and Table 5 (serovar level). For the evaluation at species level it was noticed that with two tools one *Salmonella* strain (BM-16) could not be identified and with three tools *Citrobacter freundii* (BM-06) was wrongly identified as *Salmonella* (Table 4). For the evaluation at serovar level, the outcome '*S.* Hadar/*S.* Istanbul', instead of '*S.* Hadar' was still considered correct for reasons as described above. Additionally *S.* Hadar (ENGAGE-BM-14) was excluded from further analysis, because of inconsistent results between conventional and WGS serotyping. It was noticed that with three tools some *Salmonella* serovars could not be identified. These concerned 7 strains and in total 9 incidences (Table 5).

Table 4. Outcome inclusivity/exclusivity analysis at species level

	N	**PA**	**ND**	**NA**	**PD**	**ND-PD**	**AL**	**ND+PD**	**AL**
Inclusivity	299	297	2	0	0	2	3	2	5
Exclusivity	52	0	0	49	3	Not Applicable	Not Applicable	3	3

PA: Positive agreement; ND: Negative deviation; PD: Positive deviation; NA: Negative agreement; AL: Acceptability limits (in the ISO WG working on ISO 16140-6 it was agreed for the Exclusivity not to set targets for ND-PD).

Table 5. Outcome inclusivity/exclusivity analysis at serovar level

	N	**PA**	**ND**	**NA**	**PD**	**ND-PD**	**AL**	**ND+PD**	**AL**
Inclusivity	221	212	9	0	0	9	3	9	5
Exclusivity	117	0	0	114	3	Not Applicable	Not Applicable	3	3

PA: Positive agreement; ND: Negative deviation; PD: Positive deviation; NA: Negative agreement; AL: Acceptability limits

Conclusions

The results of this benchmarking study demonstrate that serotyping using WGS data is a promising option. The tools predicting the *Salmonella* serovars in the most optimal way, in the current study, were, SISTR, SeqSero, SalmonellaTypeFinder followed by MOST, resulting in a 96.3% correlation with the conventional serotyping. This value was observed for MOST (only 1 participant used it), SISTR (3 participants used it), SeqSero (2 participants out of 8 who used this tool), and SalmonellaTypeFinder (only 1 participant used). The most optimal tool in this study based on unequal numbers of participants that used the tools. This was a limitation to evaluate the best tool in this study.

When analysing the data in accordance with ISO/DIS 16140-6:2017, the evaluation of results at species level showed to be within the acceptability limits, but at serovar level they exceeded these limits. This latter was mainly caused by the fact that in 9 incidences the *Salmonella* serovar of the target strains could not be identified. Testing non-target strains additional to target strains in such a study showed to be important as with 3 tools *Citrobacter* was incorrectly identified as *Salmonella*. The quality of the sequences and the choice of assembly tools and/or different options/parameter settings still need some attention when using WGS for serotyping *Salmonella* as participants who used different settings or assembly tools (also same tool using different online platforms), they got different serotyping results.

Additional notes

It is recommended to re-serotype, using the conventional serotyping, the isolates where the predictions from the tools disagree with the expected serovar.

References

ISO/DIS 16140-6:2017 Microbiology of the food chain — Method validation —Part 6: Protocol for the validation of alternative (proprietary) methods for microbiological confirmation and typing procedures

Hendriksen et al., 2009. WHO Global Salm-Surv External Quality Assurance System for Serotyping of *Salmonella* Isolates from 2000 to 2007. J Clin Microbiol 47(9): 2729-2736.

Popoff, M. Y., 2001. Guidelines for the preparation of *Salmonella* antisera, 6th ed. WHO Collaborating Centre for Reference and Research on *Salmonella.* Institut Pasteur, Paris, France.

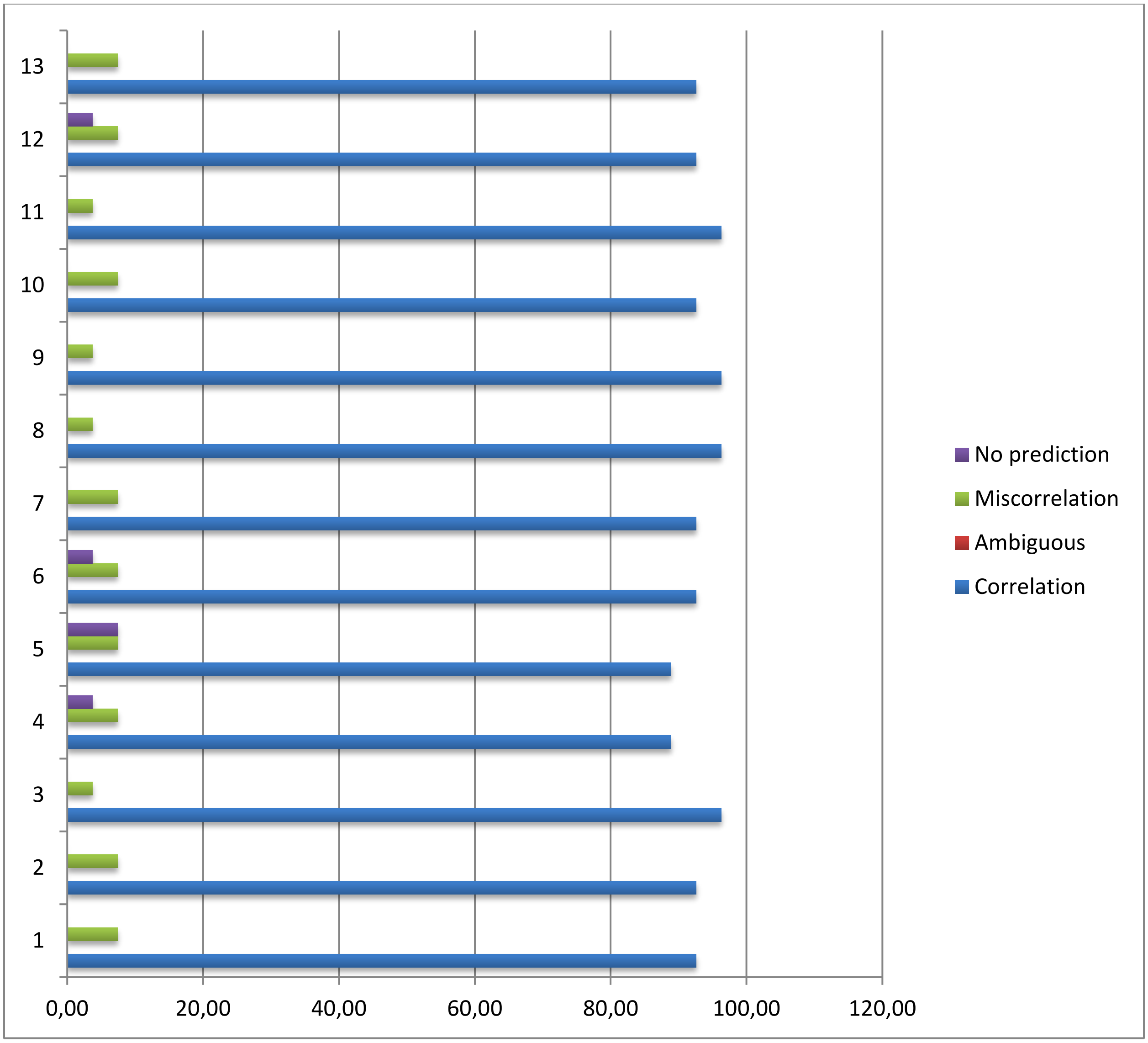

Figure 1: Species prediction

X-axis represents percentage of correlation, ambiguous, miscorrelation and no prediction of species prediction.

Y-axis corresponds to the list of benchmark tools and participants for species prediction.

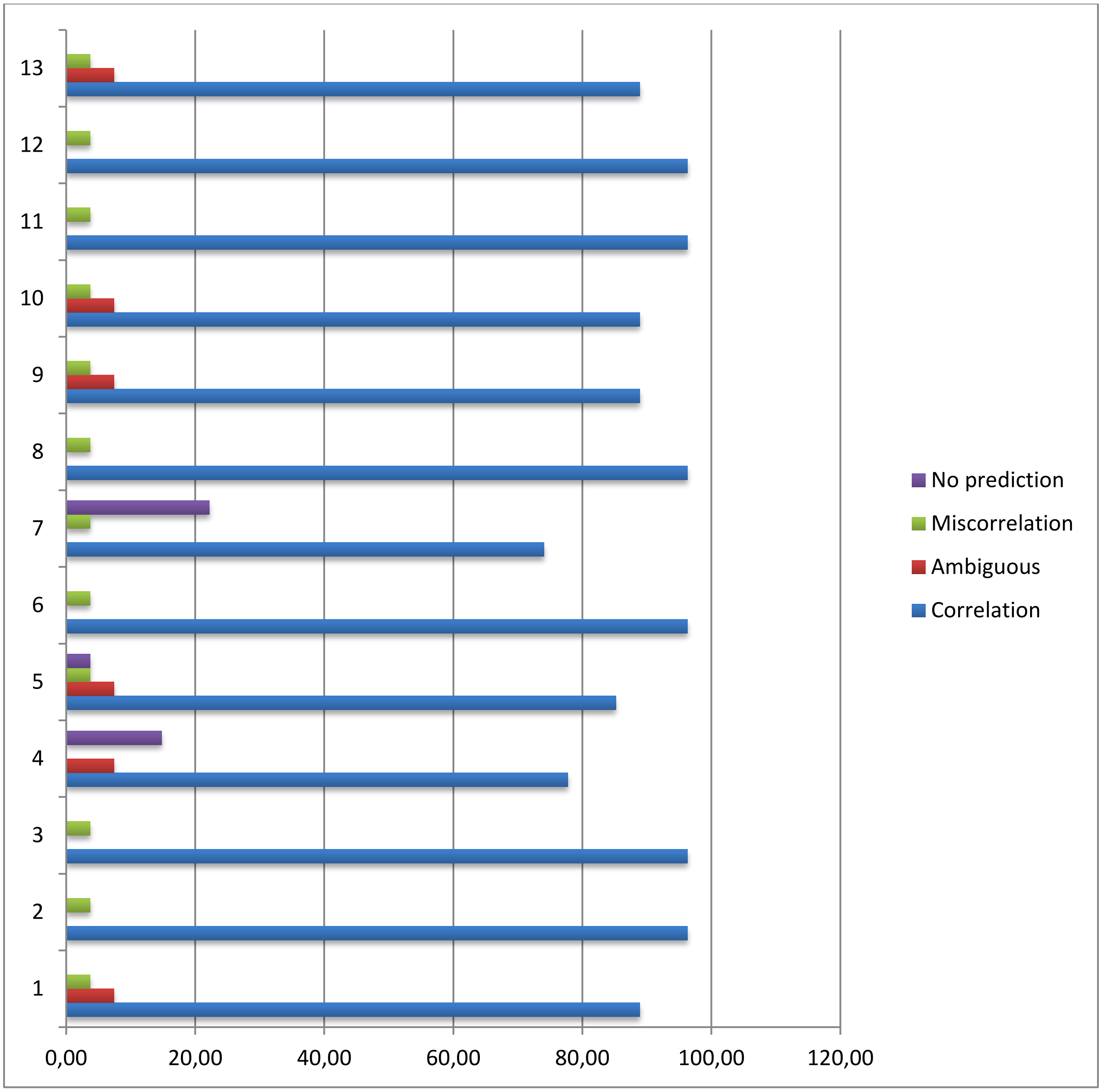

Figure 2: Serovar prediction
X-axis represents percentage of correlation, ambiguous, miscorrelation and no prediction of Salmonella serotype prediction.
Y-axis corresponds to the list of benchmark tools and participants for species prediction.

Table 6: List of selected genomes; selected (sequence data of) strains for interlaboratory study WGS serotyping *Salmonella*
Indicated are the strains selected per institute

Target strains	**RIVM**	**APHA**	**DTU**
Typhimurium	ENGAGE-BM-20	ENGAGE-BM-03	ENGAGE-BM-01
monophasic Typhimurium	ENGAGE-BM-04	ENGAGE-BM-10	ENGAGE-BM-19
Enteritidis	ENGAGE-BM-21	ENGAGE-BM-25	ENGAGE-BM-09
Hadar	ENGAGE-BM-11		ENGAGE-BM-13, ENGAGE-BM-14
Infantis	ENGAGE-BM-15	ENGAGE-BM-22	ENGAGE-BM-16
Virchow	ENGAGE-BM-18	ENGAGE-BM-24	ENGAGE-BM-07
Non-target *Salmonella* serovars			
Dublin		ENGAGE-BM-26	
Stanley		ENGAGE-BM-27	
Derby			ENGAGE-BM-05
Kentucky			ENGAGE-BM-23
Mbandaka			ENGAGE-BM-08
Same family, but non-target genus			
Citrobacter freundii		ENGAGE-BM-06	
Escherichia coli	ENGAGE-BM-17		
Klebsiella pneumoniae		ENGAGE-BM-12	
Shigella flexneri			ENGAGE-BM-02

--- --- ---