Benchmarking of genotypic *Salmonella* serotype prediction (general)

Report number	#2
Responsible	Anthony Underwood (PHE)
Other partners/institutions involved	Lauren Cowley (PHE), Rolf Sommer Kaas (DTU), Pimlapas Leekitcharoenphon (DTU), Rob Davies (APHA), Mirko Rossi (University of Helsinki/ INNUENDO), Kathie Grant (PHE), Liljana Petrovska (APHA), Rene S. Hendriksen (DTU), Susanne Karlsmose Pedersen (DTU)
Launch date	Nov 2016
Deliverable date	Dec 2016

Purpose of the benchmarking exercise

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools for predicting the *Salmonella* serotype. Some EC regulations require the use of conventional serotyping methods. This could influence the need and velocity in the implementation of NGS for animal and food surveillance.

Tools benchmarked

Benchmarking by determining serotype genotypical using the following tools with default parameters:
1) MOST (PHE tool) run by DTU
2) SalmonellaTypeFinder 1.4 run by PHE
3) SeqSERO 1.2 stand-alone tool run by APHA and by PHE (as part of SalmonellaTypeFinder)
4) SISTR v1.0.1 run by INNUENDO (https://lfz.corefacility.ca/sistr-app/)

Species/genomes included

Three datasets have been collected for this study (See below Table 1 of "Tested serotypes", Annex A, Annex C). Strain selection was based on inclusion of a wide variation of serovars including commonly isolated serovars and rare serovars, seldom found (Table 2).

The Animal and Plant Health Agency (APHA) collected 78 serotyped *Salmonella* isolates. The dataset included 78 serotypes of which all were rare serovars. Bacterial DNA was extracted using the MagNA Pure LC DNA Isolation Kit III (Roche) according to manufacturer's instructions and sequencing libraries were prepared using the NexteraXT sample preparation method for sequenced on the Illumina HiSeq platform with paired-end 2x125bp reads (http://www.illumina.com).

The National Food Institute at DTU collected 208 serotyped *Salmonella* isolates (these dataset were not sequenced under ENGAGE project). The dataset included 208 isolates from 87 serotypes, received from the project '100K Salmonella project' (external subproject lead by this affiliated partner, data not included in this report). Genomic DNA was using an Invitrogen Easy-DNA™ Kit (Invitrogen, Carlsbad, CA, USA) and DNA concentrations were determined using the Qubit dsDNA BR assay kit (Invitrogen). The genomic DNA was prepared for Illumina pair-end sequencing using the Illumina (Illumina, Inc., San Diego, CA) NexteraXT® Guide 150319425031942 following the protocol revision C (http://support.illumina.com/downloads/nextera_xt_sample_preparation_guide_15031942.html). A sample of the pooled NexteraXT Libraries was loaded onto a Illumina HiSeq reagent cartridge using HiSeq Reagent Kit v2. The

libraries were sequenced using an Illumina Hiseq platform.

Public Health England (PHE) collected 500 serotyped *Salmonella* isolates. The dataset was selected to represent the serotypes that PHE receives routinely as a public health agency in the UK. It included 500 isolates from the PHE collection and representing 104 serotypes included in the PHE collection. DNA extraction of *Salmonella* isolates begins with a manual lysis using ATL buffer, Proteinase K and RNAase A (Qiagen, Hilden, Germany) (220µl, 20µl and 4µl respectively) before loading onto a Qiagen Qiasymphony SP for purification.
DNA quantification was performed using the Promega GloMax with the Invitrogen Quant-iT dsDNA Assay Kit (Broad range) (ThermoFisher Scientific, Waltham, Massachusetts, United States) according to the manufacturer's instructions. Genomic DNA was then processed using the NexteraXT® sample preparation method and sequenced with a standard 2x101 base protocol on a HiSeq 2500 Instrument in fast mode (Illumina, San Diego, CA, USA).

All selected isolates were serotyped phenotypically according to the WKLM scheme.

Table 2. Strain providers, number of isolates and serotypes.

Dataset	Isolates	Serotypes
APHA series	78	78
DTU series	208	87
PHE series	500	104
Total	786	196*

*The number of serotypes is unique serotypes across the total dataset and therefore not a sum of the serotypes within each dataset.

All datasets were sequenced on an Illumina HiSeq.

Method
Four tools have been benchmarked in this study: Metric-Oriented Sequence Typer (MOST), SeqSero [1], SalmonellaTypeFinder, The *Salmonella In Silico* Typing Resource (SISTR).

Availability of tools:

MOST: https://github.com/phe-bioinformatics/MOST

SeqSero: https://github.com/denglab/SeqSero

SeqSero web tool: http://www.denglab.info/SeqSero

SalmonellaTypeFinder: https://cge.cbs.dtu.dk/services/SalmonellaTypeFinder/

SISTR: https://lfz.corefacility.ca/sistr-app/.

Briefly, about the three tools:

MOST is based on the first version of the tool "Short Read Sequence Typing" (SRST) [2]. MOST maps read data to MLST genes and infers an MLST type. The MLST type is subsequently looked up in a local MOST database that contains information on which serotypes that have been registered for the MLST type in question. The local database has been divided into two parts, one database containing only information from PHE and another with information collected from EnteroBase (http://enterobase.warwick.ac.uk/)

SeqSero is doing *in silico* molecular serotyping. In the sense that it maps read data to a local database of the genes that causes the phenotype of the specific serotypes. SeqSero thereby infers the phases and translates the

phase profile into a serotype.

SalmonellaTypeFinder is an attempt to merge the methods from the above tools. SalmonellaTypeFinder runs SeqSero and infers an MLST type using SRST2 [3]. The MLST type is subsequently looked up in a local database created from information in EnteroBase (this includes information from PHE) to determine which serotypes that have been registered for the particular MLST type. A serotype is then inferred from the MLST type based on the criteria that at least 3 registered isolates of the same serotype has been found with the MLST type in question, and at least 75% of the serotypes registered to the MLST type is identical. The final serotype is then found by comparing the serotype inferred by SeqSero and the serotype inferred by MLST. The serotype from SeqSero always takes precedence over the serotype inferred by MLST. If SeqSero reports several serotypes, the serotype (if any) agreeing with the MLST serotype is chosen.

SISTR is a bioinformatics platform for rapidly performing simultaneous *in silico* analyses for several leading subtyping methods on draft *Salmonella* genome assemblies. The serovar prediction module in the SISTR server utilizes O (somatic) and H (flagellar) antigen and/or serogroup-specific probes previously designed for our *Salmonella* Genoserotyping Array (SGSA), which provides serovar identification for 90% (n = 2,190) of serovars.

All isolates were trimmed using bbduk2 (part of the suit bbtools version 36.49) and *de novo* assembled using SPAdes. All isolates were analysed using all four tools. PHE ran the tools SalmonellaTypeFinder and SeqSero. SeqSero was run as a part of SalmonellaTypeFinder. DTU ran the tool MOST. The output from most is an array of all the serotypes registered to a specific MLST type, along with information on which were registered by PHE and which were not in EnteroBase. The authors of MOST (PHE) decided to predict a serotype by selecting the most commonly registered, by PHE, serotype for an MLST type. INNUENDO coordinator (University of Helsinki) ran SISTR.

Overall results

The results were divided into the serotype predictions that correlate with the expected serotype (Table 3, Figure 1) and those that does not correlate. The results that does not correlate has been further divided in to the predictions that give a different serotype than the expected (miscorrelation, Figure 2), the predictions that yields no result (no prediction, Figure 3), and the predictions that yield several possible serotypes, were the expected serotype is found among those (ambiguous, Figure 4). For more detail of all the results, see Supplementary Table 3 (Annex C).

Table 3. Serotype prediction results:

	MOST	SeqSero	SalmonellaTypeFinder	SISTR
Correlation	668 (85%)	508 (65%)	669 (85%)	694 (88%)
No Correlation	118 (15%)	278 (35%)	117 (15%)	92 (12%)
- Miscorrelation	33 (4%)	22 (3%)	26 (3%)	65 (8%)
- No prediction	85 (11%)	34 (4%)	34 (4%)	8 (1%)
- Ambiguous	0 (0%)	222 (28%)	57 (7%)	19 (2%)

7 miscorrelations (0.9% of all isolates) were identical across all four tools, meaning that all the tools agreed upon the predicted serotype. The relatively low correlation for APHA dataset was due to the fact that APHA dataset consist of rare serotypes (Table 4).

Presented below is the correlation to each of the three datasets.

Table 4. Correlation result for different series of data:

	MOST	SeqSero	SalmonellaTypeFinder	SISTR
APHA	17 (22%)	44 (56%)	46 (59%)	35 (45%)
DTU	169 (81%)	155 (75%)	191 (92%)	192 (92%)
PHE	482 (97%)	309 (62%)	432 (86%)	467 (93%)

Conclusion

The results of this benchmarking study clearly demonstrate that serotyping using NGS data is a very feasible option. The tool with highest correlation, SISTR, gets 88% correlation with the conventional serotyping (Figure 5), and this is a conservative number, considered none of the isolates have be retested, to ensure correct serotyping.

The miscorrelation rate, cases were the tools predicted a different serotype than the expected, were 3-8% in this study. Additionally, at least half of these miscorrelations are heavily suspected to be mistakes in the conventional serotyping. Interestingly, the tool with highest correlation also seems to have the highest miscorrelation. It is not possible from this study to conclude why these miscorrelations happened, but the tools are under constant development and the errors made by the tools are decreasing with each new release.

Such a low miscorrelation rate would probably be hard to achieve for most labs that does conventional serotyping.

Three of the four tools archive similar scores with a correlation rate between 85-88% and “no correlation” rates between 12-15%. It is important to note that the lowest scoring tool “SeqSero” is an essential part of the higher scoring tool “SalmonellaTypeFinder”.

Additional notes

It is recommended to serotype the isolates were the predictions from the tools disagree with the expected serotype. This is especially important for the isolates where all tools have identical miscorrelations. Additionally, the different sequence quality and sequencing processing may have an effect on the results.

References

1. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015 May;53(5):1685-92.
2. Inouye M, Conway TC, Zobel J, Holt KE. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics. 2012 Jul 24;13:338
3. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014 Nov 20;6(11):90.

Table 1: Tested serotypes

Serotype	*count*	*Serotype*	*count*	*Serotype*	*count*
35:z10:-	1	Butantan	1	Hadar	7
38:k:-	1	Canastel	1	Haifa	8
4,12:d:-	1	Cerro	8	Havana	10
9,46:z45:-	1	Chandans	1	Heidelberg	7
Aberdeen	4	Chester	6	Hiduddify	1
Abony	5	Chicago	1	Hvittingfoss	5
Adelaide	5	Claibornei	1	Ibadan	5
Agama	5	Coeln	6	II 1,4,12:z29:e,n,x	1
Agbeni	5	Coleypark	1	II 16:m,t:[z42]	1
Ago	5	Colindale	7	II 21:z10:z6	1
Agona	6	Concord	8	II 55:k,z39:1	1
Agoueve	1	Corvallis	9	II 58:d:z6	1
Ajiobo	5	Derby	10	IIIa 41:z4,z23:-	1
Alachua	5	Dublin	8	IIIa 47:z4,z23:-	1
Albany	7	Dugbe	1	IIIa 51:g,z51:-	1
Altona	4	Durham	6	IIIa 56:z4,z23:z32	1
Amager	5	Ealing	3	IIIb 61:1,5,7:-	1
Amsterdam	2	Eastbourne	5	IIIb 65:z10:e,n,x,z15	1
Anatum	8	Elisabethville	1	Indiana	6
Anfo	1	Emek	6	Infantis	10
Ank	2	Enteritidis	24	Isangi	2
Apapa	2	Falkensee	1	Istanbul	1
Augustenborg	1	Fischerkietz	1	Itami	1
Bardo	1	Florida	1	Ituri	1
Bareilly	8	Fluntern	4	IV 48:g,z51:-	1
Bergen	1	Freetown	1	IV 50:g,z51:-	5
Bispebjerg	4	Fresno	1	Jangwani	5
Blockley	6	Friedenau	1	Javiana	7
Bonariensis	1	Gaminara	5	Jukestown	1
Bonn	3	Georgia	1	Kambole	1
Bovismorbificans	7	Give	8	Karachi	1
Braenderup	7	Glostrup	2	Kedougou	12
Brandenburg	7	Godesberg	1	Kentucky	12
Bredeney	7	Goldcoast	5	Kenya	5
Khami	3	Muenster	6	Stockholm	1
Kibi	1	Nagoya	1	Takoradi	6

Serotype	*count*	*Serotype*	*count*	*Serotype*	*count*
Kimuenza	1	Napoli	5	Tees	1
Kingston	5	Newport	11	Telelkebir	6
Kisangani	3	Nima	5	Teltow	1
Kisarawe	1	Nottingham	3	Tennessee	5
Kokomlemle	1	Offa	1	Thompson	7
Kottbus	6	Ohio	5	Toricada	1
Kuessel	1	Omifisan	1	Typhi	5
Landwasser	1	Onireke	0	Typhimurium	28
Lexington	1	Oranienburg	7	Uganda	3
Lille	1	Oritamerin	1	Umbilo	5
Litchfield	7	Oslo	5	Vejle	1
Liverpool	1	Panama	8	Vinohrady	1
Livingstone	6	Paratyphi A	5	Virchow	12
London	7	Paratyphi B	4	Virginia	1
Madjorio	1	Paratyphi B var Java	7	Vitkin	4
Malstatt	1	Pomona	1	Vogan	1
Manchester	3	Poona	7	Wangata	1
Manhattan	2	Potsdam	8	Waycross	1
Matopeni	1	Putten	1	Weltevreden	8
Mbandaka	11	Reading	2	Westhampton	1
Meleagridis	5	Richmond	5	Widemarsh	1
Mgulani	1	Rissen	9	Wilhelmsburg	1
Mikawasima	5	Rubislaw	5	Wippra	1
Minnesota	6	Saintpaul	8	Worthington	2
Mishmarhaemek	1	Sandiego	6		
Mississippi	5	Schwarzengrund	8		
Moero	1	Senftenberg	8		
Monschaui	7	Shipley	1		
Montevideo	8	Singapore	2		
Morningside	1	Solt	1		
Mpouto	1	Stanley	10		
Muenchen	8	Stanleyville	5		

Note: Information on the isolates included in this benchmarking analysis is available in Annex C.

Figure 1

Correlations. Y-axis represents number of isolates that serotype predictions correlate with the expected serotype.

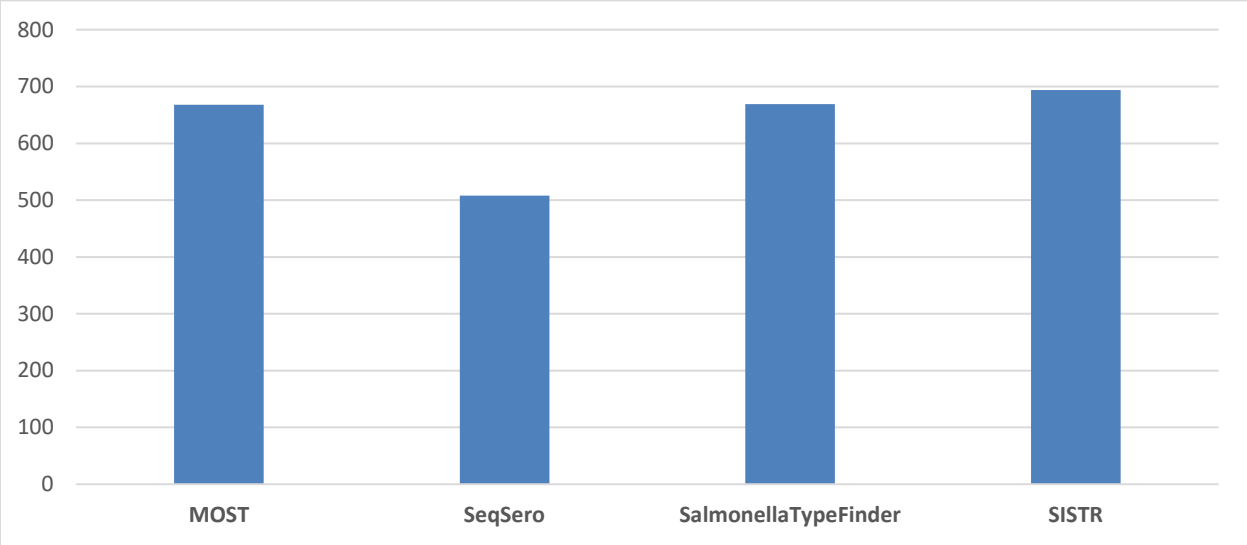


Figure 2

Miscorrelation. Y-axis represents number of isolates that serotype predictions give a different serotype than the expected.

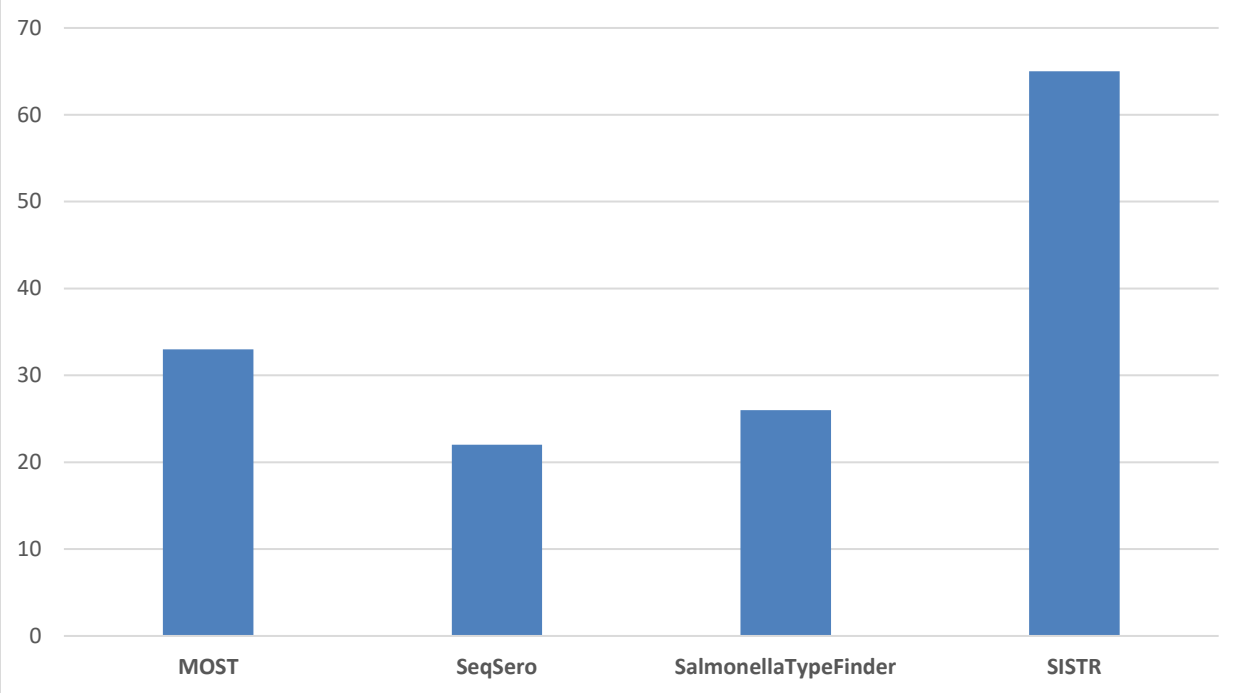


Figure 3
No prediction. Y-axis represents number of isolates that serotype predictions yield no result.

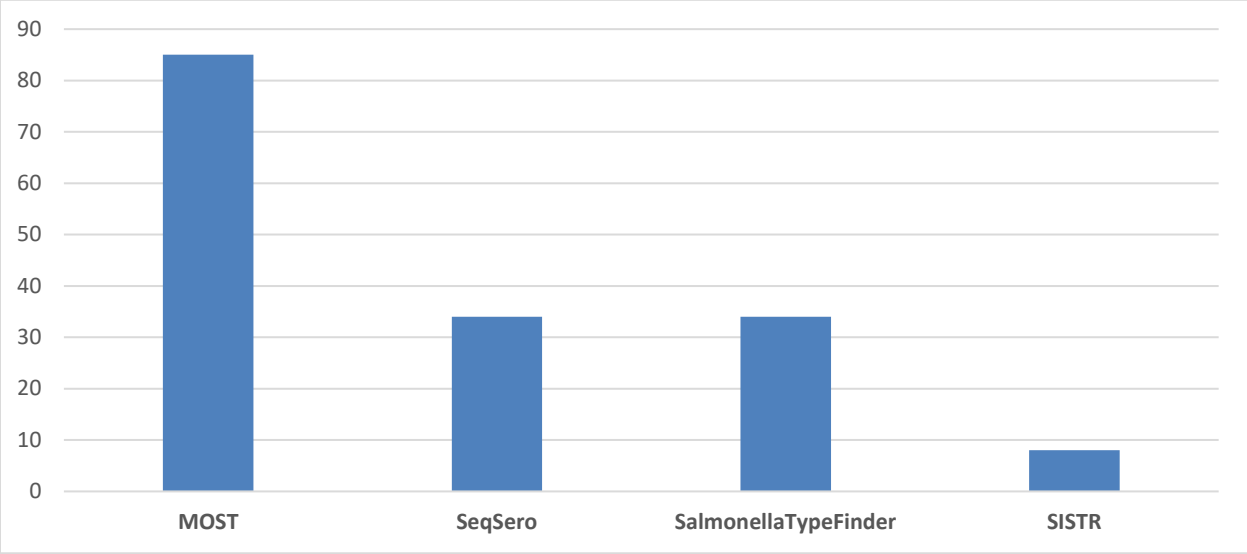


Figure 4
Ambiguous. Y-axis represents number of isolates that serotype predictions yield several possible serotypes, and the expected serotype is found among those.

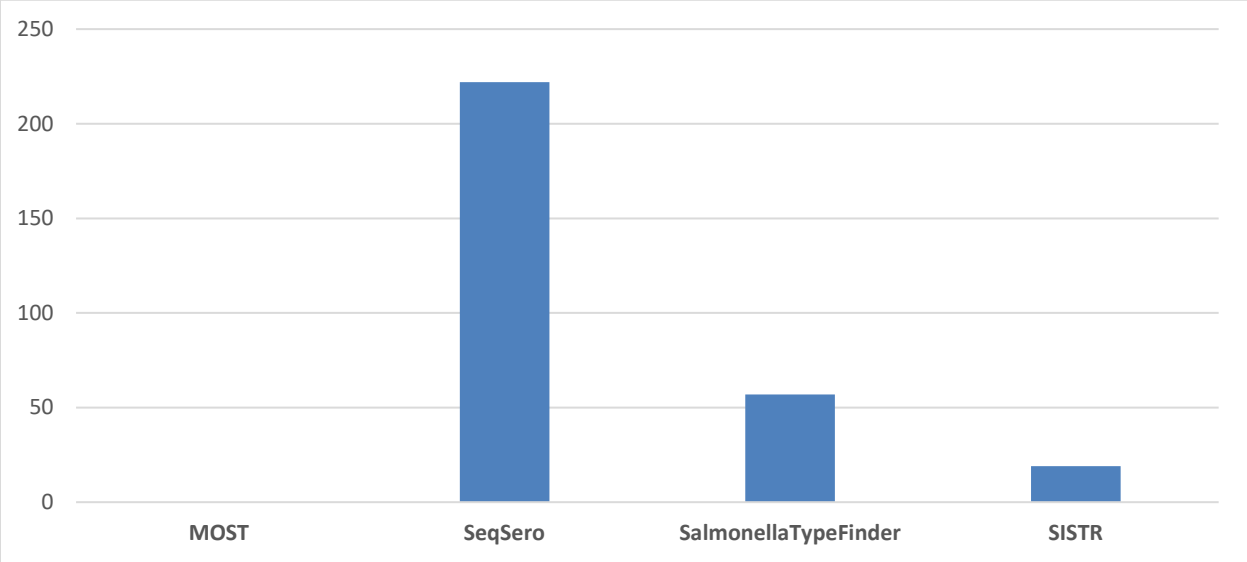
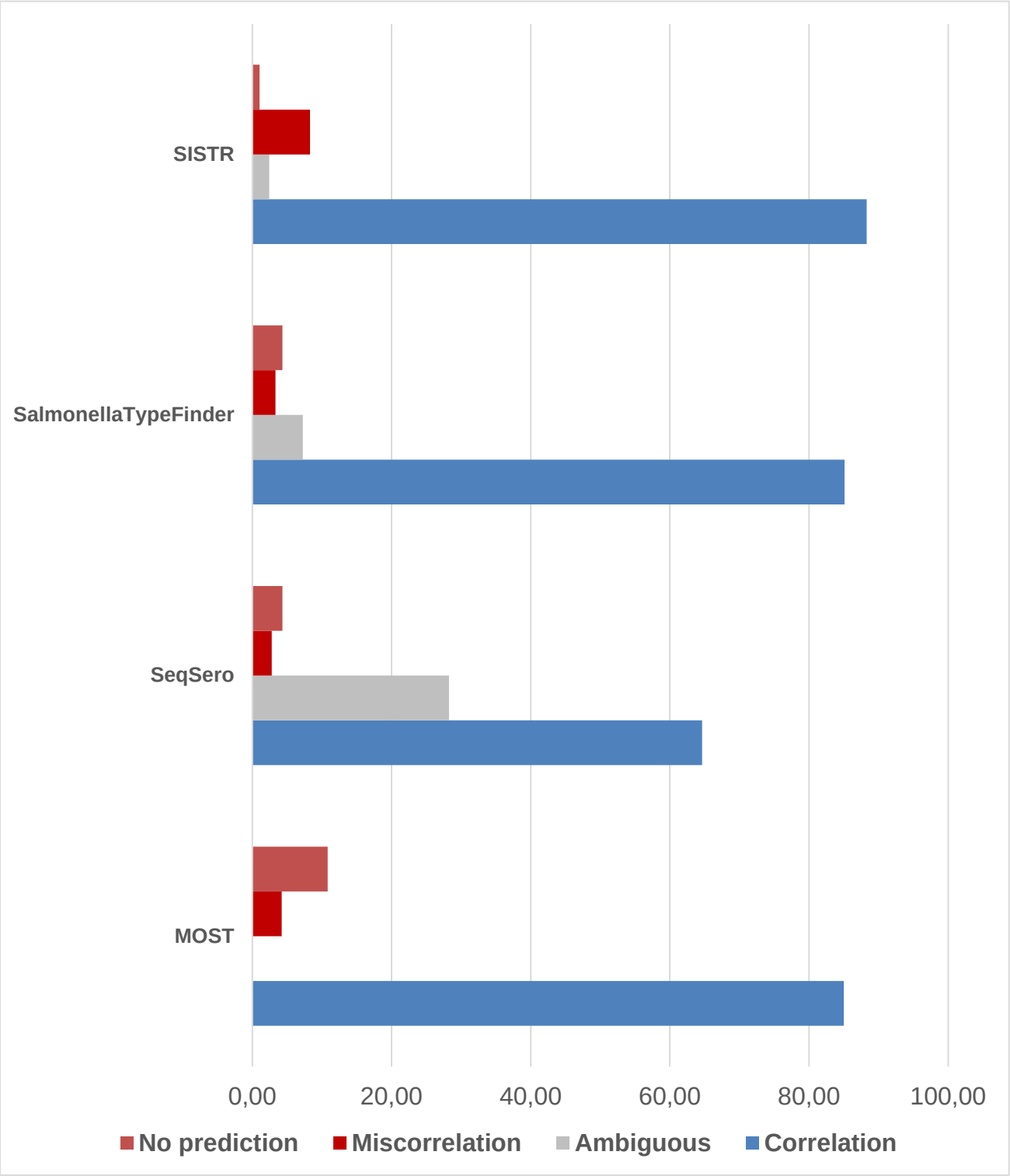


Figure 5
Summary of correlation, miscorrelation, no prediction and ambiguous. X-axis represents percentage.



--- --- ---