Retrieving Good-Quality *Salmonella* Genomes From the GenBank Database Using a Python Tool, SalmoDEST

Emeline Cherchame¹,², Guy Ilango¹,³ and Sabrina Cadel-Six¹

¹Salmonella and Listeria Unit, Laboratory for Food Safety, ANSES, Maisons-Alfort, France. ²Paris Brain Institute (ICM), Paris, France. ³Research Center for Respiratory Diseases, INSERM UMR 1100, Tours, France.

**ABSTRACT:** With the advent of next-generation whole-genome sequencing (WGS), the need for good-quality and well-characterised *Salmonella* genomes has increased over the past years. Good-quality complete genomes are often required for assembly reference mapping or phylogenetic single nucleotide polymorphism (SNP) analysis. Complete genomes or contigs from specific sources or serovars are also searched for clustering analysis or source attribution studies. Therefore, new bioinformatics tools are needed for the extraction of good-quality and well-characterised genomes from public databases. Here, we developed SalmoDEST, an open-source Python tool capable of extracting *Salmonella* genomes with a coverage higher than 50x and genome length over 4Mb from the GenBank database in the form of complete genomes or contigs, with verification of the serovar to which they belong and identification of the corresponding multi locus sequence type (MLST) profile. To validate the ability to SalmoDEST to screen for and retrieve genomes of good quality, we compared our results for S. Typhi complete genome with those available in the literature and extracted *Salmonella* genomes from bovine sources strains isolated worldwide. Finally, we provide in this study a list of 239 complete genomes for 123 serovars of *Salmonella* of high quality. SalmoDEST is a handy and easy-to-use open-source python tool to extract complete genomes or contigs that can be routinely used in public health, food safety and research laboratories. SalmoDEST (SALMONella Download gEnome Serotype sT) is available at https://github.com/I-Guy/SalmoDEST.

**KEYWORDS:** Salmonella, SalmoDEST open-source Python tool, good-quality genomes, complete reference genomes, serovar prediction, MLST profile determination

**RECEIVED:** November 16, 2021. **ACCEPTED:** January 26, 2022.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by funding by the French Ministry of Agriculture, Food and Forestry, by the Salmonella Network, part of the ANSES-Laboratory for Food Safety (France).

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Sabrina Cadel-Six, Salmonella and Listeria Unit, Laboratory for Food Safety, ANSES, 94701 Maisons-Alfort, France. Email: sabrina.cadel-six@anses.fr

**Introduction**

The investigation of genetic markers or genome relationships between different pathogens and microorganisms requires good-quality genomes. A large panel of good-quality genomes makes it possible to study chromosome rearrangements in more detail, identify sequences of interest and improve the identification of genetic clustering. Among the most frequently consulted sequence databases for collecting genomes is the open-access GenBank database, housed by the National Centre for Biotechnology Information (NCBI). GenBank annotates a collection of all publicly available nucleotide sequences generated by laboratories throughout the world from more than 100,000 distinct organisms. Release 242.0, produced in February 2021, contained over 12 trillion nucleotide bases in more than 2 billion sequences.¹ To facilitate the retrieval of genomes of interest from the GenBank database, we designed a workflow (called SalmoDEST) to search and download genomes with a coverage greater than 50x. The options of this tool make it possible to download either complete genomes or contigs. It is possible to choose to download protein fasta files, if desired, and an output directory where all the selected fasta files are kept. The SalmoDEST tool was developed for *Salmonella*, a well-known and widely distributed foodborne pathogen. *Salmonella enterica* is regulated in the European Union (EU) and monitored in the United States (US) and many other countries. In the US, the economic burden due to salmonellosis is estimated to be US$3.66 billion per year. In 2016, the incidences of culture-confirmed cases of salmonellosis were 14.51 and 20.4 cases per 100,000 population in the US and the EU, respectively.²,³ The economic, social and public health importance of diseases caused by *Salmonella* has brought many developing and developed countries to implement their monitoring systems with whole-genome sequencing (WGS) of the isolated strains, clustering by single nucleotide polymorphism (SNP) core-genome analysis for outbreaks and source attribution investigations. For countries that can carry out WGS, it is necessary to have access to *Salmonella* genomes from different regions of the world and for which the serovar has been verified and the multi-locus sequence type (MLST) profile identified. For countries in which WGS is still not readily available, carrying out studies based on good-quality and well-identified open-access *Salmonella* genomes can prove to be an essential asset.

**Materials and Methods**

**Workflow description**

SalmoDEST is implemented as an open-source Python tool (https://github.com/I-Guy/SalmoDEST). It is based on a succession of two Python scripts and a Bash process (Figure 1).
SalmoDEST is a workflow designed to search and download Salmonella genomes from the NCBI GenBank database using either the ncbi-acc-download tool for complete genomes or ncbi-genome-download for contigs. Using these tools, the first Python script 'Get_HQ_Genome_1.py' in SalmoDEST automatically downloads the genome fasta files of the strains for which accession numbers are present in the input text file. Then, the serovar and MLST profile predictions of the downloaded genomes is carried out with a Bash process using SeqSero, MLSTseeman tool and Quast, respectively. The second Python script 'Get_HQ_Genome2.py' renames the downloaded fasta files, adding the accession number, the serovar and the MLST profile predictions as follows: antigenic formula or serovar name_ST_ID_Accession number (eg, Montevideo_81_42N_CP037893.1). The Python script 'Get_HQ_Genome2.py' also downloads the gff and gbk files and checks the quality of each genome. It retains only those with coverage greater than 50x and a genome length longer than 4 Mb, and removes the others. Finally, this Python script compresses (zips) all files.

Optionally, it is possible to choose to download fasta protein files, if desired, and, in addition, choose an output directory in which all the selected fasta files are stored.

Get_HQ_Genome_1.py script. The input file of SalmoDEST and the 'Get_HQ_Genome1.py' script is a text file, obtained from an NCBI Nucleotide database query (https://www.ncbi.nlm.nih.gov/nuccore) or compiled by the user, listing the accession numbers of the complete genomes or contigs to download.

If an NCBI Nucleotide database query is used, the ‘Complete Record’ must be exported into a destination ‘File’ in the ‘Accession List’ format sorted by ‘Default order’.

In the ‘Get_HQ_Genome_1.py’ script, the function named ‘getFastaFromNuccore’ downloads fasta files and transcribes the accession number of the downloaded fasta files in a tsv file. The function named ‘Renamer’ renames every fasta file as "ID_Accession.fasta" and creates a folder with the same name to which it moves the fasta files. The function named “Filter1Genome” works only if the user chooses the “complete
Selection of complete genomes from a public database. Complete reference genomes are often required for assembly reference mapping or phylogenetic SNP analysis for the mapping step and the calculation of pairwise distance between genomes. Nevertheless, for a single laboratory it may be difficult to have a complete set of reference genomes, particularly considering that the genus *Salmonella* is separated into six subspecies and over 2000 serovars. The SalmoDEST tool was tested to search, download and select all complete *Salmonella* reference genomes available in the GenBank database. SalmoDEST applies a coverage filter set to a minimum of 50x. A second manual filter is based on serovar identification. SalmoDEST was used to compare the listed serovars with the serovars predicted by Seqsero2 in the TableMergeFilter2 tsv file. In this study, SalmoDEST was tested using the list of accession numbers obtained using the NCBI ‘All Databases’ query: ‘*Salmonella*[title] AND *Genome*[title] AND *Salmonella enterica*[title] AND *Genome Assembly and Annotation report*[title]’ (https://www.ncbi.nlm.nih.gov/genome/browse/#/prokaryotes/152/) with the filter ‘Complete’ (on 24 June 2021). A list with 1648 accession numbers was retrieved, and after eliminating duplicates, 1048 unique accession numbers were found (Supplementary Table S1). The SalmoDEST option for complete genome mode ‘-m g’ was used. Finally, after serovar prediction and genome length verifications, 1040 genomes were retained and downloaded. Four tsv output files were produced, including the final TableMergeFilter2 tsv file (Supplementary Table S2).

Workflow application

In this study, we report two application examples for SalmoDEST. In the first example, we evaluate the ability of SalmoDEST tool to download complete *Salmonella* genomes from the NCBI GenBank database and, in the second, its ability to download *Salmonella* genome contigs for strains isolated from bovine sources.

Selection of complete genomes from public database. Microbiologists need to access to *Salmonella* serovar genomes from specific sources for many types of analyses such as clustering analyses, source attribution studies or when screening for molecular markers. Obtaining genomes from laboratories around the world is therefore a major advantage. Here, we tested the ability of the SalmoDEST tool to obtain *Salmonella* genomes from strains isolated from bovine sources worldwide. The SalmoDEST tool was tested using the list of assembly accession numbers obtained using the NCBI ‘All Databases’ query: ‘*Salmonella*[title] AND *Genome*[title] AND *Salmonella enterica*[title] AND *Genome Assembly and Annotation report*[title]’ (https://www.ncbi.nlm.nih.gov/genome/browse/#/prokaryotes/152/) with the following filters: ‘Contig’ AND ‘Bovine’ AND ‘bovine’ (on 24 June 2021), 89 unique accession numbers were included in the input txt file and downloaded. Four tsv output files were created, including the final TableMergeFilter2 tsv file (Supplementary Table S2).

Results and Discussion

The NCBI Nucleotide query carried out on 7 June 2021 resulted in 1648 accessions. After deduplication, 1048 unique accessions were included in the input txt file and downloaded by the SalmoDEST tool that we developed here. All these complete genomes were checked for 50x coverage, genome length and predicted serovar matching. Finally, 1040 complete genomes with good quality were downloaded and the MLST profile was determined. From the initial list of 1048 complete
Bioinformatics and Biology Insights

In the input txt file, SalmoDEST excluded one genome (CP060132.1) for incorrect serovar prediction and seven others (OU015718.1, OU015719.1, OU015720.1, OU015717.1, LR792437.1, LR792391.1 and LN868943.1) due to low genome length (genome lengths of < 4 Mb, comprised between 277 503 and 3 746 274 bases). We obtained 16 genomes of S. enterica subsp. salamae, 10 S. enterica subsp. arizonae, 13 S. enterica subsp. diarizonae, 10 S. enterica subsp. boute nae and 991 S. enterica subsp. enterica, representing 135 serovars with different antigenic formulas. No S. enterica subsp. indica genomes with a coverage higher than 50x were found. Four serovars were overrepresented (ie, more than 50 complete genomes) in the GenBank database and in our results: S. Typhi (ie, responsible for human typhoid fever with 124 genomes/1040), S. Enteritidis, S. Typhimurium and S. 4,[5],12: i:-, with 114/1040, 141/1040 and 56/1040 genomes, respectively. These latter three serovars are the non-typhoid Salmonella serovars the most frequently isolated worldwide. These serovars were followed by S. Heidelberg (40/1040), S. Newport (38/1040), S. Anatum (32/1040), S. Bareilly (30/1040), S. Indiana (22/1040), S. Montevideo (21/1040) and S. Senftenberg (20/1040) (Figure 2). Our results are consistent with CDC and EFSA reports. Since 2016, these 11 serovars have belonged to the top 30 most frequently isolated serovars in the EU and the US.

To validate the ability to SalmoDEST to screen for and retrieve complete genomes of good quality, we compared our results for S. Typhi with those available in the literature. As expected, in accordance with the study published by Yap and Thong in 2017, SalmoDEST was able to recover 124 S. Typhi. The SalmoDEST tool developed in this study succeeded in screening for and downloading good-quality reference genomes for S. Typhi, confirming its ability to make good-quality genomes available quickly.

Finally, due to the need for complete genomes for sequence assembly and for SNP phylogenetic analyses (ie, for mapping analyses and to calculate the pairwise distance between genomes), we constituted a panel of complete reference genomes for Salmonella from the SalmoDEST output obtained in this study. We selected 239 complete genomes from the initial 1040 genomes, with 10 S. enterica subsp. salamae, 8 S. enterica subsp. arizonae, 7 S. enterica subsp. diarizonae, 8 S. enterica subsp. boute nae and 206 S. enterica subsp. enterica, representing 123 serovars and 185 MLST profiles (Table 1 and Supplementary Table S5). When possible, the sequencing technology used for complete genome assembly (ie, both short and long reads) and coverage were taken in account for the selection of the final panel. This panel of complete genomes can be used by microbiologists in food poisoning and typhoid investigations involving Salmonella spp.
Table 1. List of good-quality complete *Salmonella* genomes (ID, serovar and MLST profile predictions) downloaded from the NCBI GenBank database on 28 June 2021.

| PREDICTED_SEROVAR | MLST PROFILE | ACCESSION NUMBER | PREDICTED_SEROVAR | MLST PROFILE | ACCESSION NUMBER | PREDICTED_SEROVAR | MLST PROFILE | ACCESSION NUMBER |
|-------------------|--------------|-------------------|-------------------|--------------|-------------------|-------------------|--------------|-------------------|
| 1,3,19:g, s,t:-   | 217          | CP038604.1        | II 56: b:z6       | 5324         | CP029995.1        | Oranienburg       | 3613         | CP033344.1        |
| Abaatetuba        | 2041         | CP007532.1        | II 56: z10:e, n,x,z15 | 2403 | CP029992.1        | Orion             | 684          | CP030235.1        |
| Aberdeen          | 426          | LS483453.1        | II 58: d:z6       | 3379         | CP070222.1        | Oslo              | 1370         | CP030231.1        |
| Abony             | 1483         | CP007534.1        | II 58: l, z13,z28-| 1141         | LS483477.1        | Ouakam            | 1610         | CP022116.1        |
| Adjame            | 3292         | CP049881.1        | IIIa -: z4, z23-  | 106          | CP053584.1        | Panama            | 48           | CP012346.1        |
| Adjame            | 4023         | CP054827.1        | IIIa 40: z4, z23- | 6216         | CP041011.1        | Paratyphi A       | 85           | CP000026.1        |
| Agona             | 13           | CP025452.1        | IIIa 41: z4, z23- | 2131         | CP000880.1        | Paratyphi A       | 129          | CP009049.1        |
| Albany or Duesseldorf | 292    | CP019177.1        | IIIa 48: z36-     | 3711         | LR134150.1        | Paratyphi B       | 28           | CP020492.1        |
| Albert            | 19           | CP044188.1        | IIIa 53: z4, z23,z32- | 2127       | CP022504.1        | Paratyphi B var. L(+) tartarate + | 307          | CP008861.1        |
| Anatum            | 64           | CP029800.1        | IIIa 53: z4, z23- | 874          | LR133910.1        | Paratyphi C or Choleraesuis or Typhisuis | 68          | CP007639.1        |
| Anatum            | 2167         | CP014620.1        | IIIa 62: z36-     | 2402         | CP006693.1        | Paratyphi C or Choleraesuis or Typhisuis | 68          | CP007639.1        |
| Antsalova         | 4407         | CP019116.1        | IIIa 63: g, z51-  | 1425         | CP029991.1        | Paratyphi C or Choleraesuis or Typhisuis | 90          | CP043773.1        |
| Apapa             | 19403.1      | IIIb 47: k:z35    | IIIb 47: k:z35    | 1195         | CP053583.1        | Paratyphi C or Choleraesuis or Typhisuis | 114         | CP000857.1        |
| Bareilly          | 203          | CP063684.2        | IIIb 48: i:z      | 574          | CP029989.1        | Paratyphi C or Choleraesuis or Typhisuis | 139         | CP012344.2        |
| Bareilly          | 909          | CP060503.1        | IIIb 50: k:z      | 430          | CP059886.1        | Paratyphi C or Choleraesuis or Typhisuis | 145         | CP051366.1        |
| Bareilly          | 5146         | CP034721.1        | IIIb 60: r:z      | 3457         | CP011289.1        | Pomona            | 451          | CP019186.1        |
| Bergen            | 1356         | CP019405.1        | IIIb 60: z52:z53  | 2830         | CP030180.1        | Poona             | 308          | CP046279.1        |
| Berta             | 435          | CP030005.1        | IIIb 61: i:z      | 57           | LS483474.1        | Poona             | 447          | CP037891.1        |
| Birkenhead        | 424          | CP045958.1        | IIIb 65: c:z      | 1260         | CP022135.1        | Poona             | 812          | LS483489.1        |
| Bispebjerg        | 251          | CP043027.1        | Indiana           | 17           | CP028131.1        | Poona             | 964          | CP019189.1        |
| Blockley          | 52           | CP043662.1        | Infantis          | 32           | CP047881.1        | Quebec            | 4409         | CP022019.1        |
| Blukwa            | 367          | LR134148.1        | Inverness         | 1384         | CP019181.1        | Reading           | 1628         | CP051307.1        |
| Bovismorbicans    | 142          | CP060517.1        | Irumu             | LR134144.1   | Rissen            | 469               | CP030190.1     |
| Bovismorbicans    | 1499         | CP069297.1        | Irumu             | LR134144.1   | Rissen            | 469               | CP030190.1     |
| Braenderup        | 22           | CP022490.1        | IV -: z4, z23-    | 963          | LS483478.1        | Saintpaul         | 27           | CP017723.1        |
| Brancaster        | 2133         | CP036166.1        | IV -: z4, z23-    | 3942         | CP051368.1        | Saintpaul         | 49           | CP053055.1        |
| Brandenburg       | 65           | CP02580.1         | IV [1],40:g, z51- | 2265         | CP053582.1        | Saintpaul         | 50           | CP045954.1        |
| Bredeney          | 241          | CP043222.1        | IV 16: z4, z32-   | 596          | CP045761.1        | Saintpaul         | 95           | CP023512.1        |
| Bredeney          | 897          | CP07533.1         | IV 41: z52-       | 3924         | CP054715.1        | Saintpaul         | 680          | CP022491.1        |
| Butantan          | 600          | CP046278.1        | IV 45: g, z51-    | 107          | CP030194.1        | Saintpaul         | 3602         | CP023166.1        |

(Continued)
| PREDICTED_SEROVAR | MLST PROFILE | ACCESSION NUMBER | PREDICTED_SEROVAR | MLST PROFILE | ACCESSION NUMBER | PREDICTED_SEROVAR | MLST PROFILE | ACCESSION NUMBER |
|-------------------|--------------|------------------|-------------------|--------------|------------------|-------------------|--------------|------------------|
| Carmel 2123       | LS483455.1   | 2882             | Sanjuan 785       | LR134159.1   | 785              |                   |              |
| Cerro 367         | CP008925.1   | 2053             | Schoeneberg 96    | CP045447.1   |                  |                   |              |
| Chester 24        | CP019178.1   | 152              | Senftenberg 14    | CP038591.1   |                  |                   |              |
| Coeln 534         | LR134190.1   | 471              | Senftenberg 185   | CP016837.1   |                  |                   |              |
| Concord 599       | CP028196.1   | 198              | Senftenberg 210   | AP020332.1   |                  |                   |              |
| Corvallis 1541     | CP027677.1   | 906              | Senftenberg       | CP034233.1   |                  |                   |              |
| Cubana 286        | CP006055.1   | 212              | Senftenberg 3179  | CP012349.1   |                  |                   |              |
| Dakar 5734        | CP046280.1   | 808              | Solden 319        | CP014994.1   |                  |                   |              |
| Dayton 40         | CP028900.1   | 214              | Stanley 1986      | CP034716.1   |                  |                   |              |
| Derby 71          | CP026609.1   | 491              | Stanleyville 97   | CP017727.1   |                  |                   |              |
| Derby 72          | CP022494.1   | 2247             | Stanleyville      | CP034701.1   |                  |                   |              |
| Djakarta 11       | CP019409.1   | 413              | Telekebir 450     | CP030217.1   |                  |                   |              |
| Dublin 10         | CP032393.1   | 155              | Sundsvall 5323    | LS483457.1   |                  |                   |              |
| Dublin 4406       | CP019179.1   | 504              | Taksny 2204       | LR134146.1   |                  |                   |              |
| Enteritidis 11    | CP063700.1   | 413              | Telekebir        | CP030217.1   |                  |                   |              |
| Enteritidis 3175  | CP008928.1   | 4976             | Tennessee 319     | CP014994.1   |                  |                   |              |
| Florida 931       | LS483454.1   | 18               | Thompson 26       | CP032141.1   |                  |                   |              |
| Fresno 649        | CP032444.1   | 413              | Typhi 1           | CP003278.1   |                  |                   |              |
| Gallinarum or Enteritidis 78 | CP019035.1 | 3016             | Typhi 2          | AL513382.1   |                  |                   |              |
| Gallinarum or Enteritidis 92 | CP022963.1 | 155              | Typhi 8          | LT904887.1   |                  |                   |              |
| Gallinarum or Enteritidis 136 | CP018633.1 | 85                | Typhi 2138       | LT905088.1   |                  |                   |              |
| Gallinarum or Enteritidis 331 | AM933173.1 | 129              | Typhi 2209       | CP029918.1   |                  |                   |              |
| Gallinarum or Enteritidis 1972 | CP045955.1 | 140              | Typhium 19        | AE006468.2   |                  |                   |              |
| Gallinarum or Enteritidis 3304 | CP045956.1 | 5372             | Typhium 34        | CP045952.1   |                  |                   |              |
| Gaminara 2439     | CP024165.1   | 1245             | Typhium 36        | CP036168.1   |                  |                   |              |
| Gaminara 2440     | CP030288.1   | 548              | Typhium 99        | CP020922.1   |                  |                   |              |
| Gateshead 6131    | CP046291.1   | 4                | Typhium 128       | HG326213.1   |                  |                   |              |
| PREDICTED SEROVAR | MLST PROFILE | ACCESSION NUMBER | PREDICTED SEROVAR | MLST PROFILE | ACCESSION NUMBER | PREDICTED SEROVAR | MLST PROFILE | ACCESSION NUMBER |
|-------------------|--------------|------------------|-------------------|--------------|------------------|-------------------|--------------|------------------|
| Give              | 516          | CP046277.1       | Montevideo        | 81           | CP037893.1       | Typhimurium        | 213          | CP035547.1       |
| Goldcoast or Brikama | 358       | CP062223.1       | Montevideo        | 138          | CP040380.1       | Typhimurium        | 302          | CP014356.1       |
| Goldcoast or Brikama | 2529      | LR134158.1       | Muenchen          | 83           | CP016014.1       | Typhimurium        | 328          | CP025736.1       |
| Grumpensis        | 751          | CP030223.1       | Muenchen          | 112          | CP045056.1       | Typhimurium        | 568          | CP064919.1       |
| Hadar             | 33           | CP022069.2       | Muenchen          | 112          | CP045063.1       | Typhimurium        | 568          | LR862421.1       |
| Havana            | 1237         | LR134187.1       | Muenster          | 321          | CP019198.1       | Typhimurium        | 2066         | CP009102.1       |
| Heidelberg        | 15           | CP005995.1       | Muenster          | 2210         | CP040562.1       | Typhimurium        | 2110         | CP040562.1       |
| Hidalgo or Cocody | 195          | CP022663.1       | Napoli            | 2095         | CP063140.1       | Typhimurium        | 3631         | CP039854.1       |
| Hillingdon        | 205          | CP019410.1       | Newport           | 5            | CP015923.1       | Typhimurium        | 5036         | CP029840.1       |
| Hvittingfoss      | 434          | CP045831.1       | Newport           | 31           | CP007559.2       | Typhimurium        | 5401         | CP033226.2       |
| Hvittingfoss      | 446          | CP022503.1       | Newport           | 45           | CP012598.1       | Uganda             | 684          | CP051398.1       |
| I 4: [5], t2: b-  | 2379         | CP039610.1       | Newport           | 118          | CP015924.1       | Virchow            | 16           | CP045945.1       |
| I 9: [g, m, q]-   | 2912         | CP019406.1       | Newport           | 132          | CP025323.1       | Wandsworth         | 1498         | CP019417.1       |
| I 9: [g, p, q]-   | 10           | CP030207.1       | Newport           | 166          | CP012144.1       | Waycross           | 2460         | CP034707.1       |
| II: z39: k, x15   | 3706         | LS483496.1       | Newport           | 350          | CP016010.1       | Weltevreden        | 365          | CP014996.1       |
| II: z4: z24: z39  | 4415         | LS483456.1       | Newport           | 4157         | CP039436.1       | Weltevreden        | 2384         | LN800524.1       |
| II: x: z39: k     | 3910         | CP034717.1       | Newport           | 4166         | CP039437.1       | Weslaco            | 1088         | LR131413.1       |
| II: z39: k        | 1211         | CP022139.1       | Oranienburg       | 23           | CP019197.1       | Yoruba             | 1316         | CP030209.1       |
| II: z39: k        | 1211         | CP022139.1       | Oranienburg       | 23           | CP019197.1       | Yoruba             | 1316         | CP030209.1       |
Salmonella contig genomes from bovine sources

Among the recognised pathogens causing human disease, almost 60% are of animal origin\(^2\) and cattle bred for meat and for milk are common reservoirs of Salmonella spp.\(^2\) Almost 40% of a herd can be infected, and the risk of infection increases with the size of the herd.\(^2\) Salmonellosis in cattle puts producers at risk for direct economic losses associated with mortality or body weight loss, and also indirect losses caused by reduced feed conversion or veterinary care costs.\(^2\) Genomes from strains isolated from cattle can be used in source attribution studies, as well as in searches for specific host marker sequences. Our test successfully downloaded Salmonella genomes of strains isolated from bovine animals. The SalmoDEST tool was able to download 88 contig genomes of Salmonella isolated from bovine sources with a coverage of \(>50x\), lengths of \(>4\) Mb and correct serovar prediction from the initial input list file of 89 genomes. One genome (GCA_004744895,1) was excluded due to a genome length of \(<4\) Mb (Supplementary Tables 3 and 4). Fifty-two entries in the TableMergeFilter2.tsv file showed missing information on coverage and sequence type in the gbk files of the corresponding genomes. Interestingly, among the 88 contig genomes downloaded, the most represented serovars were S. Typhimurium (28 contig genomes/88), S. Newport (14/88) and S. Dublin (11/88). These three serovars are well known for contaminating bovine animals in the EU and the US.\(^18,20,22\)

50x coverage

The value of 50x was chosen for Salmonella in the SalmoDEST tool following the recommendations of the European Centre for Disease Control and Prevention (ECDC).\(^24\) The amount of data generated per Salmonella isolate by a DNA sequencer is substantial (ie, megabytes) and a trade-off must be struck between genome coverage (ie, quality) and the size of the files generated. For example, although a coverage of 30x is typically sufficient for routine surveillance of foodborne pathogens, the appropriate coverage threshold is platform-dependent and may also vary by organism.\(^25\) ECDC has fixed a coverage of 50x for Salmonella, considering this value as reasonable for corresponding file size.\(^24\) Coverage is frequently considered as the main quality metric typically used in WGS. Furthermore, the quality of genome sequences also have an impact on successful in silico serovar prediction. Missing or incomplete MLST and cgMLST loci sequences largely contribute to errors in identification.\(^6\)\(^,26\) Similarly, partial or missing antigenic data in the \(rftp\) region (ie, the O-antigen flippase and polymerase genes) and the \(fljC\) and \(fljB\) genes influence in silico serovar prediction.\(^6\) Good coverage prevents poor MLST, cgMLST and, antigenic data and contributes to the correct listing of the serovar.\(^6\)\(^,26,27\)

Errors in serotyping

Salmonella genomes from GenBank have already revealed errors in the serovar listed in their metadata. In 2016, Yoshioka et al carried out in silico serovar prediction on over 4,291 genomes extracted from GenBank, and revealed that 3.5% gave incorrect serovar predictions and that 1.8% had missing or ambiguous metadata, making it impossible to ascertain the listed phenotypic serovar.\(^26\) For this reason, we integrated the Bash process in the SalmoDEST tool to query the SeqSero\(^2\) and MLSTseeman\(^7\) tools. SeqSero is a Web-based tool developed by the Centres for Disease Control and Prevention (CDC) in Atlanta, GA (US) for determining Salmonella serotypes using the \(rftp\) region and the \(fljC\) and \(fljB\) alleles.\(^6,28\) SeqSero2 was chosen because it is the only tool that relies on characterising genetic determinants of Salmonella serovars without consulting any markers, such as MLST types; it saves time because it predicts serovars directly from raw sequencing reads and not from assemblies, and finally it is able to detect inter-serovar contaminations.\(^6\) The MLSTseeman is a tool developed by Torsten Seemann in 1991\(^7\) that scans contig files against traditional PubMLST typing schemes conceived as part of the development of the first MLST scheme in 1998.\(^29\) making it possible to include all levels of sequence data, from single gene sequences up to and including complete, finished genomes.\(^30\)

Information on serovar and MLST type were integrated in SalmoDEST to enable genome verification and because they are integral to surveillance and outbreak investigations.

Conclusion

SalmoDEST is a handy and easy-to-use tool that can be routinely used in public health, food safety and research laboratories to extract complete Salmonella reference genomes of high quality from GenBank. It can also be used to download contig genomes from a list of assembly IDs. A coverage of 50x, as well as correct Salmonella genome size and serovar and MLST type prediction, are used as quality controls for both genome modes (ie, complete and contig genomes search and download). Moreover, SalmoDEST screens downloaded genomes for contamination by using the SeqSero2 tool for serovar prediction.

Acknowledgements

We thank Laurent Vigneron (ANSES) for providing high-performance computing resources.

Author Contributions

SC-S, EC and GI conceived the study. EC and GI contributed equally to the design and analysis of data. GI conceptualised the algorithms. EC implemented scripts and executed commands. SC-S drafted the manuscript. EC reviewed the draft. All authors commented and approved the final manuscript, take public responsibility for appropriate portions of the content and agree to be accountable for all aspects of the work in terms of accuracy or integrity.
Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. GenBank release notes. NCBI. https://www.ncbi.nlm.nih.gov/genbank/release/.
2. Gal-Mor O. Persistent infection and long-term carriage of typhoidal and nontyphoidal salmonellae. Clin Microbiol Rev. 2019;32:e00088-18.
3. USDA. Economic service research cost of foodborne illness estimates for Salmonella (non-typhoidal). https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses.aspx. Updated 2018.
4. Blin K. ncbi-acc-download. https://github.com/kblin/ncbi-acc-download.
5. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072-1075.
6. Zhang S, den Bakker HC, Li S, et al. SeqSero2: rapid and improved salmonella serotype determination using whole-genome sequencing data. Appl Environ Microbiol. 2019;85:e01746-19.
7. Grimont P, Weill F-X. Antigenic Formulae of the Salmonella serovars. 9th ed. Paris: WHO Collaborating Center for Reference and Research on Salmonella Institut Pasteur; 2007.
8. Elnekave E, Hong SL, Lim S, et al. Transmission of multidrug-resistant Salmonella Typhi and Paratyphi A. Clin Microbiol Rev. 2014;26:3034-3038.
9. Seemann T. mlst. https://github.com/tseemann/mlst.
10. Elnekave E, Hong SL, Lim S, et al. Transmission of multidrug-resistant Salmonella enterica Subspecies enterica 4,[5],12:i:- sequence type 34 between Europe and the United States. Emerg Infect Dis. 2020;26:3034-3038.
11. Fabre L, Le Hello S, Roux C, Issenhuth-Jeanjean S, Weill FX. CRISPR is an optimal target for the design of specific PCR assays for salmonella enterica serotypes Typhi and Paratyphi. PLoS Negl Trop Dis. 2014;8:e2671.
12. Felten A, Guillier L, Radomski N, Mistou MY, Lailler R, Cadel-Six S. Genome target evaluator (GTEvaluator): a workflow exploiting genome dataset to measure the sensitivity and specificity of genetic markers. PLoS ONE. 2017;12:e0182082.
13. Gal-Mor O. Persistent infection and long-term carriage of typhoidal and nontyphoidal salmonellae. Clin Microbiol Rev. 2019;32:e00088-18.
14. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. EFSA J. 2018;16:e05500.
15. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union One Health 2019 Zoonoses report. EFSA J. 2019;17:e09526.
16. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union One Health 2018 Zoonoses report. EFSA J. 2019;17:e09526.
17. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union One Health 2019 Zoonoses report. EFSA J. 2021;19:e06406.
18. Yap KP, Ithong KL. Salmonella Typhi genomics: envisaging the future of typhoid eradication. Trop Med Int Health. 2017;22:918-925.
19. Oueslati W, Rjeibi MR, Mhadhbi M, Jbeli M, Zreili S, Ennouchi A. Prevalence, virulence and antibiotic susceptibility of Salmonella spp. strains, isolated from beef in Greater Tunis (Tunisia). Meat Sci. 2016;119:154-159.
20. Cummings KJ, Warnick LD, Alexander KA, et al. Duration of fecal Salmonella shedding following clinical disease among dairy cattle in the northeastern USA. Prev Vet Med. 2009;92:134-139.
21. Oueslati W, Rjeibi MR, Mhadhbi M, Jbeli M, Zreili S, Ennouchi A. Prevalence, virulence and antibiotic susceptibility of Salmonella spp. strains, isolated from beef in Greater Tunis (Tunisia). Meat Sci. 2016;119:154-159.
22. Cummings KJ, Warnick LD, Alexander KA, et al. Duration of fecal Salmonella shedding following clinical disease among dairy cattle in the northeastern USA. Prev Vet Med. 2009;92:134-139.
23. Hoelzer K, Moreno Swift AI, Wiedmann M. Animal contact as a source of human non-typhoidal salmonellosis. Veterinary Research. 2011;42:1-28.
24. EFSA. Expert Opinion on the Introduction of Next-generation Typing Methods for food- and Waterborne Diseases in the EU and EEA. Stockholm: ECDC; 2015.
25. Nadon C, Van Walle I, Gerner-Smidt P, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017;22:30544.
26. Yoshida CE, Kruzkiewicz P, Laing CR, et al. The Salmonella In Silico Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. PLoS ONE. 2016;11:e0147101.
27. Zhang S, Yin Y, Jones MB, et al. Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015;53:1685-1692.
28. Uelze L, Borowiak M, Denke C, et al. Performance and accuracy of four open-source tools for in silico serotyping of Salmonella spp. Complete genome short-read sequencing data. Appl Environ Microbiol. 2020;86:e02265-19.
29. Maiden M, Byravare J, Feil E, Morelli G, Russell J, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. PLoS One. 1998;95:3140-3145.
30. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.