science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://lio1.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

RECSALMO: Rapid typing and characterization tool for whole genome sequencing data of *Salmonella*

Kritchai Pooncharean[1], *Ekachai Chukeatirote*[2], *and Nuttachat Wisittipanit*[3*]

[1] Division of Biochemistry, School of Medical Science, University of Phayao, Phayao 56000, Thailand
[2] Department of Bioscience, School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand
[3] Department of Materials Engineering, School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand

***Corresponding author**: Nuttachat Wisittipanit
nuttachat.wis@mfu.ac.th

Received: 14 March 2024
Revised: 19 April 2024
Accepted: 19 April 2024
Published: 9 December 2024

Citation: Pooncharean, K., Chukeatirote, E., and Wisittipanit, N. (2024). RECSALMO: Rapid typing and characterization tool for whole genome sequencing data of Salmonella. Science, Engineering and Health Studies, 18, 24030004.

ABSTRACT

Salmonella is one of the most prevalent foodborne pathogens. Endemic events involving *Salmonella* now occur more frequently and must be handled efficiently and quickly. Whole genome sequencing (WGS) cost and speed have improved significantly during the past decades, with the number of bioinformatics tools based on WGS data also increasing. This research presented the "RECSALMO" tool, designed specifically to analyze *Salmonella* genome assemblies. This tool can quickly type and characterize a collection of *Salmonella* genomes, providing extensive information and assist related healthcare personnel to effectively and timely manage a *Salmonella* epidemic. The essential outputs of the tool include sequence-based typing comprising serotyping, MLST and cgMLST, antibiotic resistance genes with their associated classes/subclasses of antibiotic drugs, *Salmonella* pathogenicity islands (SPIs) from SPI-1 to SPI-17, and spacer profiles (CRISPR locus 1 and 2). Moreover, the tool generates SNP-based trees, CRISPR-based dendrograms and pie charts of serotypes and ST indexes. Due to the superior resolution, quick operational time, and low processing cost of WGS, RECSALMO is considered a practical bioinformatics tool for comprehensive and rapid examination of *Salmonella* genomes.

Keywords: bioinformatics tool; whole genome sequence; *Salmonella*; serotyping; pathogenicity islands; CRISPR

1. INTRODUCTION

Endemic or epidemic events related to *Salmonella* have been continuously increasing, together with the antibiotic resisting abilities of the pathogenic strains of *Salmonella* (Alenazy, 2022), making efficient epidemiological surveillance and prevention systems vital for public health. *Salmonella*, methodically classified as a genus, refers to a group of Gram-negative bacteria that comprises two species: *S. enterica* and *S. bongori*. *S. enterica* contains six subspecies and more than 2,600 serovars (Gal-Mor et al., 2014). *Salmonella* bacteria are well-known foodborne pathogens, which have long been a major concern for public health. They cause gastroenteritis, bacteremia, and enteric fever in infected patients, which can quickly lead to large-scale outbreaks if appropriate actions are not taken. Most of the outbreaks are triggered by the *Salmonella enterica* (*S. enterica*). Thus, accurate *Salmonella* typing and information on antibiotic-resistance properties of the infecting *Salmonella* strains are crucial to administer the correct antibiotic treatments to infected patients in a timely manner. The sources of the outbreaks must also be identified quickly so that resources can be deployed fast to resolve the situation. Bioinformatics analytical tools based on whole genome sequencing (WGS) are a viable and pragmatic option to effectively achieve these objectives.

WGS is now widespread in biotechnology research and clinical laboratories thanks to the technological advancement and relatively low cost of high-throughput sequencing technology. WGS is generally employed in research areas related to diagnostic microbiology for strain identification, profiling of antimicrobial resistance (AMR), and investigation of infection origins, providing a far greater resolution compared to standard molecular assays such as pulsed-field gel electrophoresis (PFGE). However, despite the huge potential and attainments of WGS, major adoption and transformation to this sequence-based analysis in actual clinical practices have been slow and problematic because most medical institutions lack sufficient resources and experts trained in bioinformatics-related fields. If WGS analysis becomes faster and accepted by mainstream in medical communities, health experts will have the most comprehensive analytical tool available, while specialized bioinformatics tools are also required. Currently, most WGS-related tools are fragmented, with only a few designed specifically to analyze and annotate bacterial genomes such as RAST (Aziz et al., 2008) and MicroScope (Vallenet et al., 2019). However, these tools cover a wide range of bacterial genera and/or species and might not provide accurate data for specific bacterial organisms. *Salmonella enterica* contains thousands of serotypes. Specific genomic data are essential in epidemiological investigations and surveillance such as serotypes, multi-locus sequence typing (MLST), AMR properties, *Salmonella* pathogenicity islands (SPIs), and clustered regularly interspaced short palindromic repeat (CRISPR) information but are rarely reported in the existing genomic annotation tools. Some *Salmonella*-specific tools might be able to provide insightful information but they are tedious to execute when dealing with a large set of genomes.

RECSALMO, a Linux-based Python program, is a rapid genome analysis tool designed specifically for genome assemblies of *Salmonella* strains. This particular program, a command-line tool, takes only a set of genome assemblies of *Salmonella* bacteria as its input (contained in a single folder) and performs a fast pre-screen of genomes before the full analysis (filtering out non-*Salmonella* genomes). It can process 20 genome assemblies in a relatively short time (approximately 80 seconds per genome). RECSALMO is essentially an automated pipe-line for *Salmonella* genome analysis based on WGS data, in which the overall operation includes *Salmonella* typing, AMR profiling, phylogenetic tree construction, pathogenicity island determination, and discovery of CRISPR spacers. The *Salmonella* typing process contains sequence-based serotyping, multilocus sequence typing (MLST) and core genome MLST (cgMLST). The AMR profiling procedure searches for the genes responsible for bacterial antimicrobial properties and classes of antibiotics associated with genetic components in such a way that their products (proteins) involve the biochemical pathways resisting these antibiotic classes, as essential information when treating infected patients. The construction of a phylogenetic tree is an imperative tool for an outbreak investigation. This employs the core genome alignment with all the *Salmonella* genome assemblies as the input and the LT2 *Salmonella* strain as the reference genome sequence. The detection of SPIs involves an in-house developed tool exploring a genome of *Salmonella* isolate for SPI-1 to SPI-23 regions based on the BLAST operations. Lastly, the finding of CRISPR spacers encompasses the search of predefined spacer sequences (~7k sequences, each with length ~32 bps) in a *Salmonella* genome. The outputs of all these processes include one summary file in both "csv" and "xlsx" format, one generated phylogenetic file, and native result files for each genome assembly (obtained from the typing and AMR profiling procedures).

WGS has revolutionized the way epidemiological surveillance and investigation outbreaks are conducted by providing optimal resolutions of genomic-scale DNA sequences. The cost of WGS has now reduced to around US$1,000 per genome (Mardis, 2006), with studies of WGS-related data and analysis significantly increasing.

The WGS-based subtyping methods for *Salmonella* are similar to other prokaryotes, and include single nucleotide polymorphism (SNP)-based cluster analysis, MLST, cgMLST, whole genome MLST (wgMLST), and AMR profiling. SNP-based subtyping demonstrated high discriminatory power in the clustering process of *Salmonella* isolates by maximum likelihood phylogeny analysis (Chattaway et al., 2019) and also in separating outbreak *Salmonella* strains from non-outbreak *Salmonella* strains (Taylor et al., 2015; Mohammed and Thapa, 2020).

The MLST procedure utilizes seven housekeeping genes of *Salmonella* (genes required for fundamental cellular functions) for indexing gene sequence distinction (Maiden et al., 2013). A combination of these unique alleles called "allelic profile" defines the sequence type (ST) of each *Salmonella* isolate (Urwin and Maiden, 2003). These MLST characteristics enable rapid *Salmonella* subtyping that requires only a few genetic loci, while cgMLST expands the concept of MLST by using a much larger set of gene loci that could be in the order of hundreds or thousands depending on the number of 'core' genes as genes existing in the majority of *Salmonella* serotypes (often set at 95%) (Uelze et al., 2020).

The pathogenic competencies of *Salmonella*, which can cause fever, gastroenteritis, and stomach cramps most likely originate from SPIs that contain multiple genetic sections (called genomic islands) encoding virulence-related factors that harmfully interact with the hosts (Marcus et al., 2000). Each SPI discovered is defined by the symbol "SPI-X", where X is an integer number successively assigned in order of discovery. SPI-1 to SPI-5 are present in all serotypes of *S. enterica* while other SPIs only exist in certain *Salmonella* isolates (Espinoza et al., 2017). For instance, SPI-14 is unique to *S. typhimurium* (Wang et al., 2020) and *S. derby*, discovered in the United Kingdom, contains SPI-23 (Sévellec et al., 2018). SPI-1 and SPI-2 contain genes that encode the proteins needed to form type III secret system (T3SS) or the "needle complex". T3SS is the key ingredient of *Salmonella* that invades host epithelial cells (Lou et al., 2019; Hegazy et al., 2012). SPI-3, SPI-4, and SPI-5 were reported to have no impact on the virulence of *S. enteritidis* infecting chickens, while both SPI-1 and SPI-2 were found to be the major virulence-related genetic areas of *Salmonella* (Rychlik et al., 2009).

CRISPR is essentially part of the immune defense system in bacteria or archaea (prokaryotic organisms) against bacteriophages (viruses that specifically attack bacteria). CRISPR is a distinctive DNA region, first identified in *Escherichia coli*, which comprises numerous repeat and palindromic sequences (Fabre et al., 2012). Between these repeats are unique "spacer" sequences, which are analogous to bacteriophage DNA. The bacteria

utilize these spacer sequences as a "defensive memory" and deploy a Cas protein to cleave (or kill) any foreign phage DNA similar to such spacers. A Cas protein is encoded by a Cas gene associated with a CRISPR locus. There are several types of Cas protein in the CRISPR system. One of the most prominent Cas proteins is Cas9 which, when working alongside a guide RNA (gRNA) having a pre-defined target sequence, can cut a DNA strand at a precise location (Guerrero-Araya et al., 2021). Using this incredible property of the CRISPR system, the target sequence belonging to a guide RNA can be designed such that the Cas9 protein can cut DNA at any desired position in a genome.

2. MATERIALS AND METHODS

2.1 Tested dataset: Collection of *Salmonella* isolates

The dataset for the experimental run of the RECSALMO tool comprised 31 *Salmonella* genomes, including 12 *Salmonella* isolates extracted from gastroenteritis patients in Northern Thailand, 18 *Salmonella* strains from multiple countries obtained from EnteroBase (Zhou et al., 2020; Achtman et al., 2020), and 1 *Salmonella* reference genome i.e., *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2 acting as a control strain.

2.2 Genome-based serotyping

Distinct characteristics of antigens "O" and "H" on the outer wall of a *Salmonella* cell were used to identify its serotype according to the Kauffmann-White-Le Minor Scheme; thus, DNA regions encoding those antigens, if known, could help to accurately determine the serotype by comparing coding sequences to the existing serotype database. This study utilized SeqSero2 (Zhang et al., 2019), a k-mer-based serotype identification tool, to rapidly perform such sequence-based serotyping to obtain a predicted serotype for each genome assembly of a *Salmonella* isolate (the tool could also take raw sequencing reads as inputs).

2.3 MLST and cgMLST

One of the weaknesses of the serotyping method based on the surface antigens is that evolutionary characteristics are not taken into account. This can lead to inaccuracy in epidemiology investigation and surveillance. MLST (Achtman et al., 2012), a sequence-based typing method, resolves such concerns by employing seven housekeeping gene loci in the *Salmonella* genome: "aroC", "dnaN", "hemD", "hisD", "purE", "sucA", and "thrA". MLST designates an allele index for each gene loci using the order of novel allele discovery and then determines a unique MLST number called "sequence type" or ST based on the combination of these allele indexes. This study used FastMLST (Guerrero-Araya et al., 2021), a high performance MLST determining tool created in Python programming language to assign an ST number for each *Salmonella* genome assembly.

Another typing method that expands on the model of MLST and covers the core genes in a genome is cgMLST (Yan et al., 2021). In the case of *Salmonella*, the number of "core" genes is approximately 3,500. However, the cgMLST model does not have a fixed number of gene loci; therefore, some cgMLST schemes employ only a handful of core genes because they have rigorous requirements in gene selection. For instance, the qualified genes must have full coverage, with no indel positions allowed when aligned with the existing sequences in a database. This study used SISTR (Yoshida et al., 2016), a bioinformatics tool for rapid in silico *Salmonella* subtyping process using genome assemblies to perform cgMLST.

2.4 Determination of AMR information

Salmonella, like most bacteria, possesses self-protective AMR properties to withstand certain antibiotic drugs. Such properties most likely originate from the genetic determinants encoded in the *Salmonella* genome. The AMR properties could also stem from the genetic elements in *Salmonella* plasmids as well as through the conjugation process (horizontal gene transfer) (Holmes and Jobling, 1996). Some *Salmonella* strains have multiple antibiotic resistance genes (ARGs) that enable them to resist several antibiotics at the same time. The AMR attributes of bacteria, especially a pathogenic strain like *Salmonella*, cause serious concerns for public healthcare around the globe, and cases of *Salmonella* defying numerous antibiotic classes are on the rise. Therefore, precise data describing which antibiotics a *Salmonella* strain can resist are crucial to performing proper medication and treatment on infected patients.

The RECSALMO tool specifically employs AMRFinderPlus (Feldgarden et al., 2021), a bioinformatics tool from the National Center for Biotechnology Information (NCBI), to search for ARGs in a *Salmonella* genome. This tool reports the antibiotic classes/subclasses that the *Salmonella* strain is capable of resisting, with the found ARGs possibly responsible for this capability. A class of antibiotics is simply a group of antibiotic drugs that performs similar mechanisms in killing off bacteria. For instance, the aminoglycoside class inhibits the synthesis of bacteria proteins and sulfonamides that inhibit folate synthesis.

2.5 Phylogenetic tree construction

Phylogenetic trees present evolutionary associations among species or biological-related entities that can help to uncover their common ancestors. The trees are generally constructed based on genetic distances derived from multiple sequence alignments. For application in the epidemiological surveillance and investigation of *Salmonella* outbreaks, phylogenetic tree can be a great tool to accurately pinpoint the outbreak clusters. This study used ParSNP (Treangen et al., 2014) to build a phylogenetic tree using the entire dataset of the *Salmonella* genome assemblies as the inputs, with *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2 as the reference genome (NCBI Genome ID = SO4698-09 and accession = LN999997.1). The ParSNP tool effectively conducted the whole genome alignments, which took SNP into account for the similarity distance calculations.

2.6 Determination of SPIs

Salmonella, a pathogenic microorganism, contains regions of DNA sequences that encode numerous virulence factors. The bacteria use these to invade and attack the host. These regions in the *Salmonella* genome, called SPIs, are responsible for causing gastroenteritis, nausea, fever, and headache in infected patients. SPIs have a naming scheme as "SPI-X", in which X is a positive integer assigned by the successive discovery in a *Salmonella* strain.

Table 1. Details of SPIs where the RECSALMO tool is capable of searching in a *Salmonella* genome

SPI	Host *Salmonella* strain	Size
SPI-1	*Salmonella* enterica subsp. enterica serovar Typhi CT18	41.9 kb
SPI-2	*Salmonella* enterica subsp. enterica serovar Typhi CT18	41.6 kb
SPI-3	*Salmonella* enterica subsp. enterica serovar Typhi CT18	16.9 kb
SPI-4	*Salmonella* enterica subsp. enterica serovar Typhi CT18	23.4 kb
SPI-5	*Salmonella* enterica subsp. enterica serovar Typhi CT18	7.5 kb
SPI-6	*Salmonella* enterica subsp. enterica serovar Typhi CT18	58.7 kb
SPI-7	*Salmonella* enterica subsp. enterica serovar Typhi CT18	133.6 kb
SPI-8	*Salmonella* enterica subsp. enterica serovar Typhi CT18	6.9 kb
SPI-9	*Salmonella* enterica subsp. enterica serovar Typhi CT18	15.7 kb
SPI-10	*Salmonella* enterica subsp. enterica serovar Typhi CT18	32.9 kb
SPI-11	*Salmonella* enterica subsp. enterica serovar Choleraesuis str. SC-B67	15.7 kb
SPI-12	*Salmonella* enterica subsp. enterica serovar Choleraesuis str. SC-B67	11.1 kb
SPI-13	*Salmonella* enterica subsp. enterica serovar Gallinarum str. 287/91	23.9 kb
SPI-14	*Salmonella* enterica subsp. enterica serovar Typhimurium LT2	6.7 kb
SPI-15	*Salmonella* enterica subsp. enterica serovar Typhi CT18	6.4 kb
SPI-16	*Salmonella* enterica subsp. enterica serovar Typhi CT18	4.5 kb
SPI-17	*Salmonella* enterica subsp. enterica serovar Typhi CT18	5.1 kb

A tool for searching SPIs in a *Salmonella* genome was developed by the authors based on the SPI data in PAIDB v2.0 (Yoon et al., 2015) - a comprehensive database of pathogenicity islands and antimicrobial resistance islands (REIs) belonging to various organisms, which also covers the sequence data of SPIs from SPI-1 to SPI-14. This particular tool extends the SPI data coverage of PAIDB v2.0 further and stores sequence data of extra SPIs from SPI-15 to SPI-17, referred to in several literatures (Vernikos and Parkhill, 2006; Wang et al., 2020). This tool has the capability to discover SPI-1 to SPI-17 in a *Salmonella* genome. Table 1 shows details of each SPI that the tool discovered, including its name, size, and host *Salmonella* strain. For each target *Salmonella* genome, the developed SPI searching tool tries to find an SPI in the genome using the BLAST tool from the Biopython package (Cock et al., 2009). The essential parameters are percent identity and sequence coverage. If percent identity and sequence coverage both exceed certain thresholds set at 90% and 60%, respectively, then the query SPI sequence is considered to reside in the target genome.

2.7 Determination of spacers and CRISPR-based typing

CRISPRs refer to a specialized cluster of DNA sequences in bacteria/archaea that contains several identical and palindromic repeats of short-segmented DNA sequences called “DR” or direct repeats, interposed with dissimilar DNA sequences called “spacer” (Liu et al., 2011). In a *Salmonella* genome, there are usually two CRISPR loci namely CRISPR locus 1 (CRISPR 1) and CRISPR locus 2 (CRISPR 2) (Li et al., 2021). Each CRISPR locus possesses a unique DR sequence as the number of repeats and composition of spacers.

The RECSALMO tool determines the spacer content of CRISPR 1 and CRISPR 2 in each *Salmonella* genome assembly derived from CRISPR loci found in *Salmonella* genomes according to the study by Fabre et al (Fabre et al., 2012). Specifically, the tool utilizes the CRISPR Recognition Tool (CRT) (Bland et al., 2007) to detect CRISPR loci, repeats and spacer content in a genome. The CRT does not straightforwardly provide CRISPR 1 and CRISPR 2 loci, giving only the raw findings of DR and spacer sequences where continuous sequences of DR/spacer are assigned as a single locus. Therefore, the number of CRISPR loci found by the tool could exceed two loci, with some adjacent loci combined into a single locus such that the final number of CRISPR loci is always two. These CRISPR loci might also be found on different contigs of a *Salmonella* genome assembly where each contig can match either a plus or minus strand of a *Salmonella* reference genome. Thus, the combination of CRISPR loci (into two) must be carefully observed using their positions on the reference genome. The RECSALMO tool utilized *Salmonella* Typhimurium str. LT2 as the reference genome to keep the positions of CRISPR loci in check for the correct loci combination.

To demonstrate the process of loci combination in detail, suppose that the CRT tool detects three CRISPR loci in a *Salmonella* genome assembly, namely C1g, C2g, and C3g. C1g and C2g both match with contig A, while C3g matches with contig B. Moreover, contig A matches with the minus strand of the reference genome, while contig B matches with the plus strand. The positions of these found CRISPR loci and contigs on the reference genome shown in Figure 1 are only based on the plus strand. The end position of C2g (3078155) and the start position of C1g (3078180) differ by 25 bps (base pairs), while the end position of C1g (3079340) and the start position of C3g (3444722) differ by ~365K bps. Thus, C2g and C1g are combined into CRISPR locus 1 since the distance between the end and start positions of the two discovered loci is quite small (usually not exceeding 1000 bps). Then, C3g becomes CRISPR locus 2 since the distance between the end and start positions (C1g and C3g) is quite large (usually exceeding 10K bps).

Once CRISPR 1 and CRISPR 2 are detected, a sequence of each detected spacer (of each CRISPR locus) is compared to sequences in the tool’s local database. Its name is then assigned to the name of the spacer in which it matches perfectly (100% identity); however, if it does not match perfectly, the spacer is given a name according to the locus and genome assembly file name that it belongs to. Thus, for each CRISPR locus of a genome, a profile of spacers is constructed (list of spacers in the order that they appear), and then a pairwise profile alignment is performed between the two spacer profiles (from two *Salmonella*

genomes) under the same CRISPR locus. The alignment algorithm for the pairwise profile alignment is similar to the Needleman-Wunch algorithm, which performs global sequence alignment; however, this alignment does not employ a substitution matrix such as BLOSUM (BLOcks Substitution Matrix) but uses a simple scoring scheme where the match, mismatch, and gap scores are +1, 0, and −1, respectively. Figure 2 shows an example of the pairwise profile alignment result between two spacer profiles.

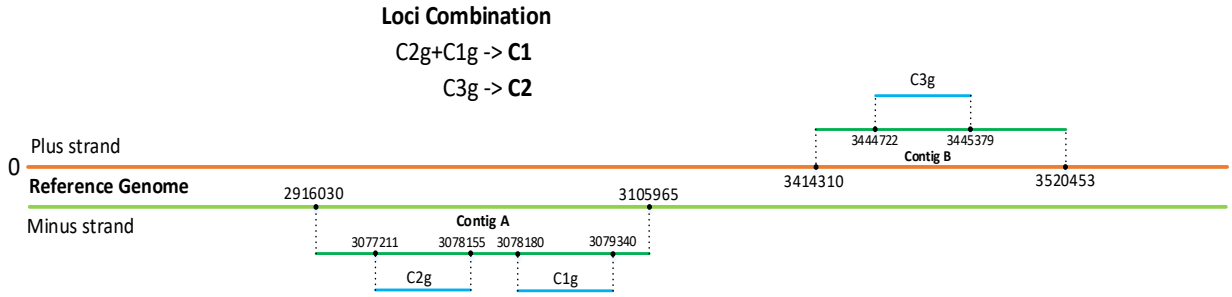


Figure 1. Positions of determined CRISPR loci and contigs on the reference genome

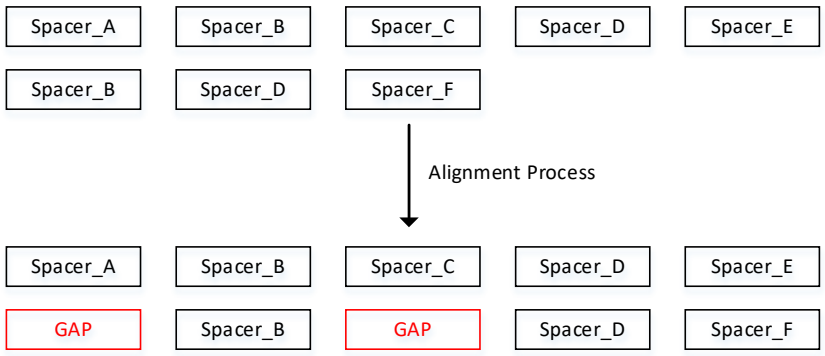


Figure 2. Example of pairwise profile alignment for CRISPR spacers of a single locus

According to Figure 1, the score for the alignment can be calculated as shown in Equation 1,

$$Score = gap + match + gap + match + mismatch = -1 + 1 - 1 + 1 + 0 = 0. \quad (1)$$

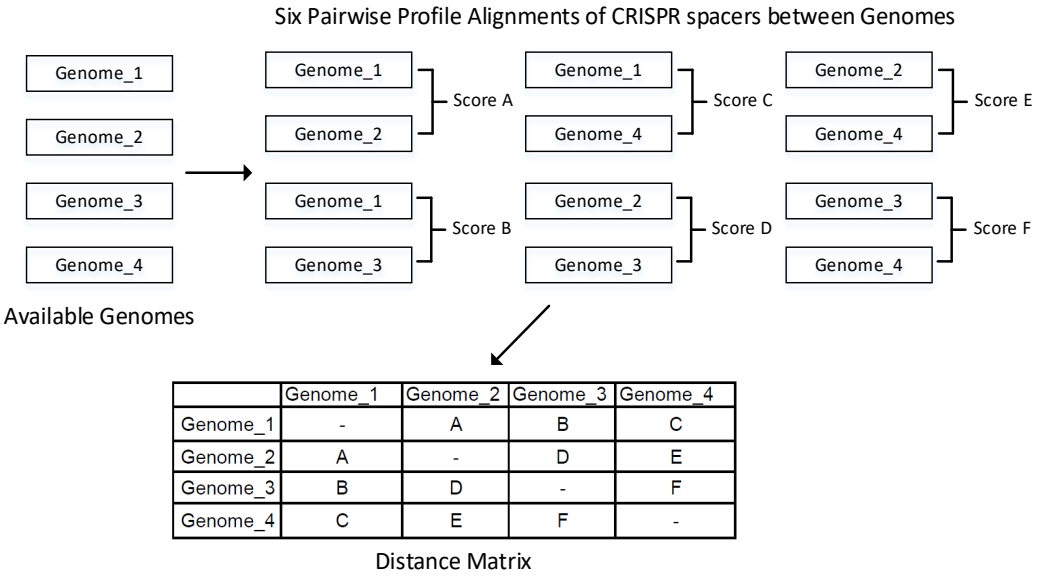
This tool only considers CRISPR 1 and CRISPR 2 for pairwise profile alignments. Therefore, the score for each profile alignment between two *Salmonella* genomes can be calculated using Equation 1. The total score from alignments between profiles of CRISPR 1 and CRISPR 2 is the summation of the two profile alignment scores. Then, the total scores between all pairs of *Salmonella* genomes are used to construct a distance matrix, which is employed to build a phylogenetic tree using the UPGMA algorithm.

Figure 3 shows the steps required to build a distance matrix. Starting from four available genomes, six pairwise profile alignments between genomes are performed to obtain six total scores (Score A to Score F). These scores are then used to construct a 4x4 distance matrix which is then employed to build the phylogenetic tree using the UPGMA algorithm.

2.8 Diagram and availability of the RECSALMO tool

The RECSALMO tool uses a batch of *Salmonella* genome assembly files (in FASTA format) as the input. Then, for each genome assembly, it performs sequence-based typing (serotype, MLST and cgMLST), determination of AMR genes/related antibiotics, determination of SPIs and determination of CRISPR locus 1 and 2. After all genome assemblies are analyzed, it constructs an SNP-based phylogenetic tree, a CRISPR-based phylogenetic tree, and pie charts of serotypes and MLST. Lastly, it creates an output file, which is essentially a summary file containing detailed results of all the performed analyses. The workflow diagram of the RECSALMO tool is shown in Figure 4.

The RECSALMO tool was entirely implemented in Python programming language version 3.8.6, using PyCharm IDE (PyCharm, the Python IDE for Professional Developers. Available online: http://jetbrains.com/pycharm) version 2022.3.1 under the Ubuntu OS version 20.04.5 LTS. All source codes and install/usage instructions of the RECSALMO tool are available at https://github.com/aongithub172/recsalmo under the GNU license. The program is open-source and only works in distributions of the Linux operating system.



	Genome_1	Genome_2	Genome_3	Genome_4
Genome_1	-	A	B	C
Genome_2	A	-	D	E
Genome_3	B	D	-	F
Genome_4	C	E	F	-

Figure 3. Steps required to build a distance matrix from available genomes. The example starts with four genomes which lead to six profile alignment scores used to construct the 4x4 distance matrix

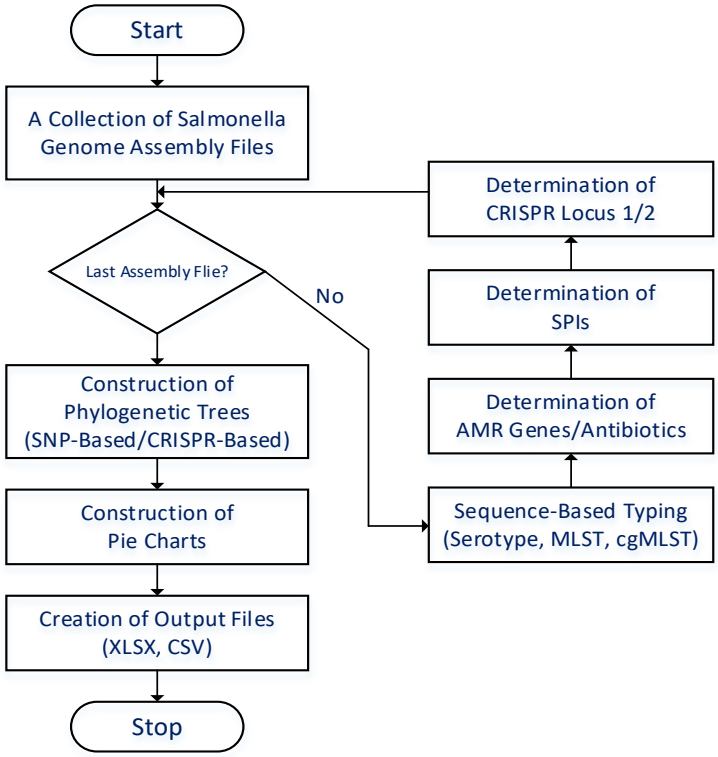


Figure 4. Workflow diagram of the RECSALMO tool

3. RESULTS AND DISCUSSION

As previously mentioned, the RECSALMO tool was experimentally run on 31 *Salmonella* genomes to obtain all genome annotation results, including *Salmonella* sequence-based typing, AMR information, SNP-based phylogenetic tree, list of SPIs, and CRISPR-based typing. All the essential outputs of the tool were aggregated in the file "summary.xlsx" (supplementary file). Pie charts of serotypes and ST indexes were also available in PNG image format, with annotation results of each genome in a separate folder, all in the TXT format. The program was run on the platform Intel® CoreTM i7-10750H CPU @2.60 GHz with 32 GB of RAM and 42 minutes were required to finish the experimental run of 31 genomes (an average of 81 seconds per *Salmonella* genome).

3.1 *Salmonella* sequence-based typing

Sequence-based typing of 31 *Salmonella* genomes included serotyping, MLST, and cgMLST performed by the SecSero2 tool, FastMLST tool, and SISTR tool, respectively. The sequence-based serotyping result indicated that the group of genomes was composed mostly of *S.* 1,4,[5],12:i:- as the monophasic variants of *Salmonella* Typhimurium, predominantly found in swine, while the second most found serotype was *Salmonella Poona*. The MLST result indicated that the group comprised mostly of ST34 strains, closely related to the monophasic *Salmonella* Typhimurium serotype, with the second most found sequence type ST447. Results of the sequence-based typing methods, including serotyping, MLST, and cgMLST are presented in the supplementary file: "summary.xlsx", while the supplementary file "comparison.xlsx" shows the comparison between serovars identified by a laboratory technique and those identified by RECSALMO. The 12 *Salmonella* strains isolated from gastroenteritis patients were identified by MALDI-TOF (matrix-assisted laser desorption ionization-time of flight) mass spectrometry. Serovar information of the 18 strains from the EnteroBase database was already available. The RECSALMO tool constructed pie charts for all the genome assemblies based on serotyping and MLST results. The pie charts gave a quick assessment of the group components by different typing methods.

3.2 AMR

The AMR analysis of 31 *Salmonella* genomes included the determination of ARGs contained in a *Salmonella* genome, along with antibiotic classes/subclasses. The AMR analysis results are presented in the supplementary file ("summary.xlsx": column = "AMR Gene/Antibiotic"). The ARGs and antibiotic classes/subclasses were re-formatted in the form of "A:U--V|…" where "A" is an ARG, "U" is an antibiotic class, and "V" is an antibiotic subclass. Each *Salmonella* genome showed detailed outputs of AMR analysis (direct results from the AMRFinderPlus tool) contained in its separate folder. If the tool cannot find any ARG in a genome, the result of this analysis will be empty for that particular genome. The presence of ARGs in genome does not necessarily mean that such *Salmonella* strains would be resistant to the associated antibiotic classes/subclasses because the tool cannot predict phenotypic resistance. The ARGs must be expressed in cell for the strain to resist certain antibiotics.

3.3 SNP-based phylogenetic tree

A phylogenetic tree built from 31 *Salmonella* genomes was based on SNP with *Salmonella* Typhimurium str. LT2 as the reference genome. The reference strain was also included in the 31 genomes as the control isolate. There were 32 nodes in the tree, and the two reference strains (the exact same genomic sequences) should be close to each other in the tree. All the related files of the SNP-based phylogenetic tree were contained in the folder "treeoutput" (directly under the user-defined output folder). The generated tree was in "ggr" file format, which can be opened by the Gingr program. The phylogenetic tree is shown in Figure 5. The reference genome ("refgenome_so46998_09.fasta.ref" highlighted in blue) is adjacent to its counterpart genome ("REF_LT2.fasta") in the tree, confirming the accuracy of the tree, while the *Salmonella monophasic* strains (BK_SAL*X*) were all in the same group.

3.4 SPIs

SPI determination results from 31 *Salmonella* genomes are presented in the supplementary file ("summary.xlsx": column = "SPI"). The RECSALMO tool utilized its own FASTA-format file containing the sequences of SPIs from SPI-1 to SPI-17 to search for SPIs in a *Salmonella* strain. All the found SPIs were written in the format "SPI-X|SPI-Y|…" where X and Y are the discovered SPI indexes in a genome. Each *Salmonella* genome also had detailed outputs of SPI finder results in a separate folder, which described the SPI locations found in a strain, including percent identity and percent of sequence coverage.

3.5 In silico CRISPR-based typing of *Salmonella* spps

In silico CRISPR-based typing analysis of 31 *Salmonella* genomes determined the spacer content (profile) in each genome and created the UPGMA-based phylogenetic tree from the spacer profile alignments. The spacer content (from CRISPR locus 1 and locus 2) was kept in a separate folder describing the list of spacers in order of discovery and in accordance with the RECSALMO's own database of spacer names/sequences. An image of the phylogenetic tree ("crispr_dendrogram.png") is presented in the main output folder. Figure 6 shows the UPGMA-based phylogenetic tree from the spacer profile alignment of 31 *Salmonella* isolates.

4. CONCLUSION

RECSALMO is a bioinformatics tool that assists in the surveillance and epidemiological investigation of *Salmonella* bacteria. In essence, the tool combines already existing and effective tools, e.g., Secsero2, SISTR, FastMLST, and AMR Finder with the authors' own tools, e.g., SPI determiner and CRISPR-based typing tool to investigate the genomes of *Salmonella*. The RECSALMO tool can assist healthcare professionals and medical experts to rapidly pinpoint the origins of the *Salmonella* endemic, and determine suitable antibiotics to cure infected people. The tool also provides critical information about the whole group of *Salmonella* genomes. RECSALMO is simple to use as it needs only two input arguments: an input folder containing *Salmonella* genome assemblies, and an output folder containing all the analysis results for each genome and the entire assemblies.

The RECSALMO tool generates pie charts of serotypes/ST indexes, SNP-based trees, CRISPR-based dendrograms, and summary files that describe species identifications, serotypes, MLST (ST indexes), cgMLST, ARGs/associated antibiotics, SPIs, and spacer profiles (CRISPR locus 1 and 2). Each genome has its own output in a separate folder that contains detailed results from all the individual tools, such as "amrfinder_result.txt" describing an ARG and its associated class/subclass of antibiotic for each contig, and "crispr_result.txt" describing direct repeat and spacer sequences for all the CRISPR loci found. Future versions of the RECSALMO tool will extend the coverage of SPI determination to include SPIs from SPI-18 to SPI-23.

ACKNOWLEDGMENT

This research was supported by the NSRF (National Science, Research and Innovation Fund) [grant number: 652A01026].

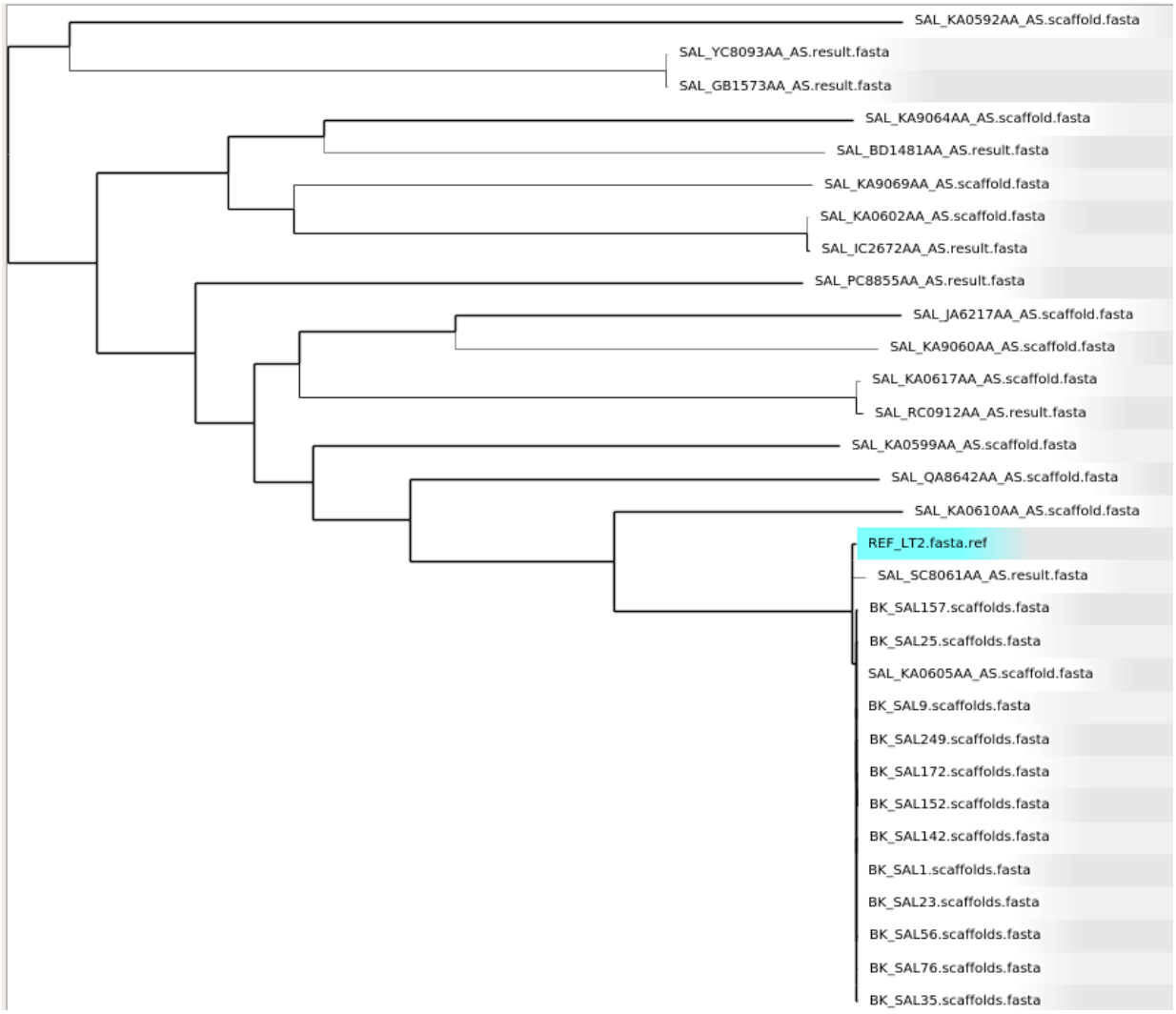


Figure 5. SNP-based phylogenetic tree built from 50 *Salmonella* isolates with *Salmonella* Typhimurium str. LT2 as the reference genome

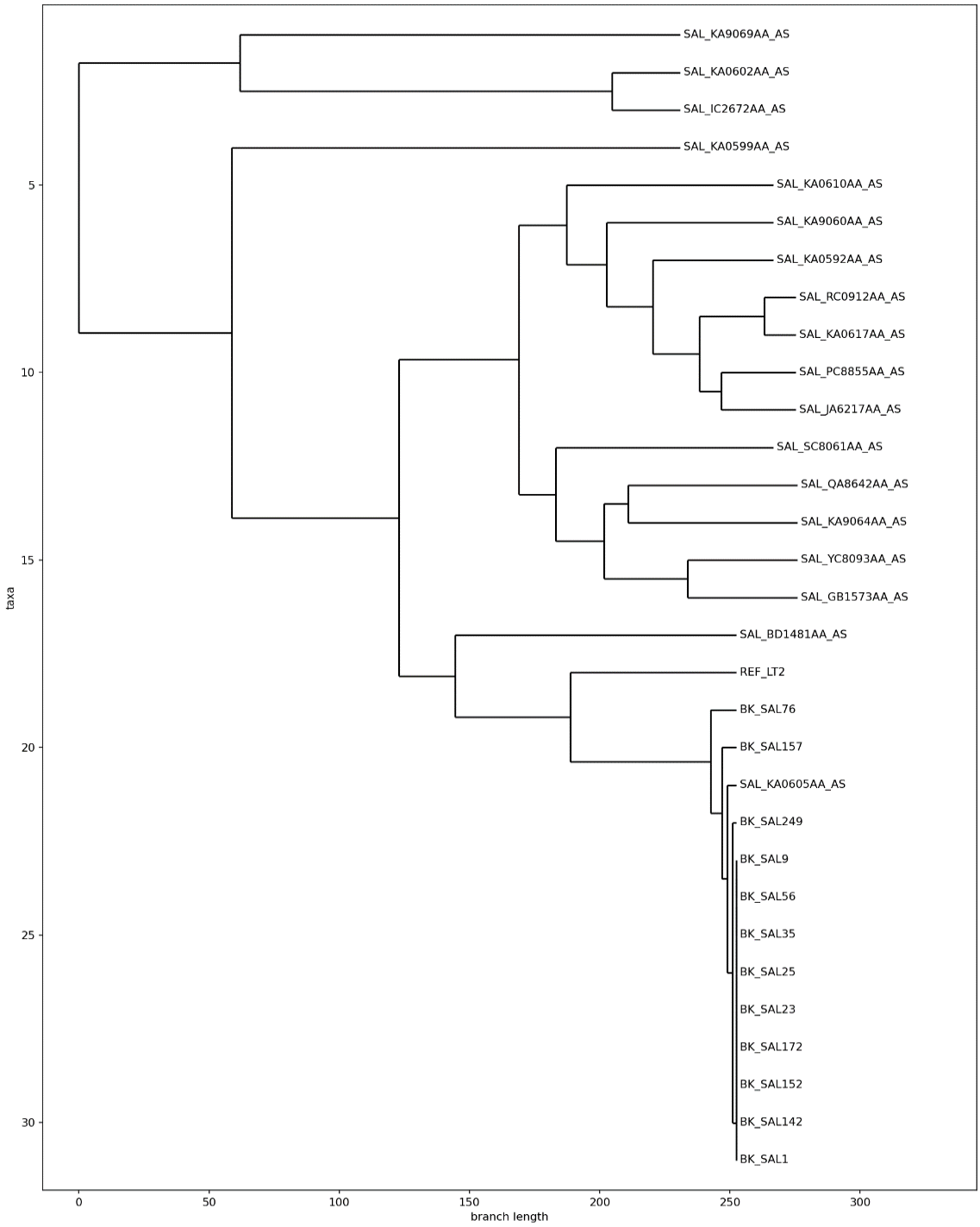


Figure 6. UPGMA-based phylogenetic tree from the spacer profile alignment of 31 *Salmonella* isolates

REFERENCES

Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M. G., Hale, J. L., Harbottle, H., Uesbeck, A., Dougan, G., Harrison, L. H., and Brisse, S. (2012). Multilocus sequence typing and a replacement for serotyping in *Salmonella enterica. PLoS Pathogens*, 8(6), e1002776.

Achtman, M., Zhou, Z., Alikhan, N.-F., Tyne, W., Parkhill, J., Cormican, M., Chiou, C. S., Torpdahl, M., Litrup, E., Prendergast, D. M., Moore, J. E., Strain, S., Kornschober, C., Meinersmann, R., Uesbeck, A., Weill, F.-X., Coffey, A., Andrews-Polymenis, H., Curtiss Rd, R., and Fanning, S. (2020). Genomic diversity of *Salmonella enterica* - The UoWUCC 10K genomes project. *Welcome Open Research*, 5, 223.

Alenazy, R. (2022). Antibiotic resistance in *Salmonella*: Targeting multidrug resistance by understanding efflux pumps, regulators and the inhibitors. *Journal of King Saud University - Science*, 34(7), 102275.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., … Zagnitko, O. (2008). The RAST server: Rapid annotations using subsystems technology. *BMC Genomics,* 9, 75.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, 8, 209.

Chattaway, M. A., Dallman, T. J., Larkin, L., Nair, S., McCormick, J., Mikhail, A., Hartman, H., Godbole, G., Powell, D., Day, M., Smith, R., and Grant, K. (2019). The transformation of reference microbiology methods and surveillance for *Salmonella* with the use of whole genome sequencing in England and Wales. *Frontiers in Public Health,* 7, 317.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25(11), 1422–1423.

Espinoza, R. A., Silva-Valenzuela, C. A., Amaya, F. A., Urrutia, Í. M., Contreras, I., and Santiviago, C. A. (2017). Differential roles for pathogenicity islands SPI-13 and SPI-8 in the interaction of *Salmonella* enteritidis and *Salmonella* typhi with murine and human macrophages. *Biological Research,* 50, 5.

Fabre, L., Zhang, J., Guigon, G., Hello, S. L., Guibert, V., Accou-Demartin, M., de Romans, S., Lim, C., Roux, C., Passet, V., Diancourt, L., Guibourdenche, M., Issenhuth-Jeanjean, S., Achtman, M., Brisse, S., Sola, C., and Weill, F.-X. (2012). CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS ONE,* 7(5), e36995.

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., and Klimke, W. (2021). AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance stress response, and virulence. *Scientific Reports,* 11, 12728.

Gal-Mor, O., Boyle, E. C., and Grassi, G. A. (2014). Same species, different diseases: How and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Frontiers in Microbiology*, 5, 391.

Guerrero-Araya, E., Muñoz, M., Rodríguez, C., and Paredes-Sabja, D. (2021). FastMLST: A multi-core tool for multilocus sequence typing of draft genome assemblies. *Bioinformatics and Biology Insights*, 15, 11779322211059238.

Hegazy, W. A. H., Xu, X., Metelitsa, L., and Hensel, M. (2012). Evaluation of *Salmonella enterica* type III secretion system effector proteins as carriers for heterologous vaccine antigens. *Infection and Immunity,* 80(3), 1193–1202.

Holmes, R. K., and Jobling, M. G. (1996). Chapter 5: Genetics. In *Medical Microbiology* (Baron, S., Ed.), 4th, Galveston, TX: University of Texas Medical Branch.

Ilyas, B., Tsai, C. N., and Coombes, B. K. (2017). Evolution of *Salmonella*-host cell interactions through a dynamic bacterial genome. *Frontiers in Cellular and Infection Microbiology*, 7, 428.

Li, C., Wang, Y., Gao, Y., Li, C., Ma, B., and Wang, H. (2021). Antimicrobial resistance and CRISPR typing among *Salmonella* isolates from poultry farms in China. *Frontiers in Microbiology*, 12, 730046.

Liu, F., Barrangou, R., Gerner-Smidt, P., Ribot, E. M., Knabel, S. J., and Dudley, E. G. (2011). Novel virulence gene and Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) multilocus sequence typing scheme for subtyping the major serovars of *Salmonella enterica* subsp. *enterica*. *Applied Environment Microbiology*, 77(6), 1946–1956.

Lou, L., Zhang, P., Piao, R., and Wang, Y. (2019). *Salmonella* Pathogenicity Island 1 (SPI-1) and its complex regulatory network. *Frontiers in Cellular and Infection Microbiology,* 9, 270.

Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., and McCarthy, N. D. (2013). MLST revisited: The gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology,* 11(10), 728–736.

Marcus, S. L., Brumell, J. H., Pfeifer, C. G., and Finlay, B. B. (2000). *Salmonella* pathogenicity islands: Big virulence in small packages. *Microbes and Infection*, 2(2), 145–156.

Mardis, E. R. (2006). Anticipating the $1,000 genome. *Genome Biology*, 7(7), 112.

Mohammed, M., and Thapa, S. (2020). Evaluation of WGS-subtyping methods for epidemiological surveillance of foodborne salmonellosis. *One Health Outlook,* 2, 13.

Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., Björkman, J. T., Dallman, T., Reimer, A., Enouf, V., Larsonneur, E., Carleton, H., Bracq-Dieye, H., Katz, L. S., Jones, L., Touchon, M., Tourdjman, M., Walker, M., Stroika, S., ... Brisse, S. (2016). Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nature Microbiology*, 2, 16185.

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 156(5), 935–949.

Rychlik, I., Karasova, D., Sebkova, A., Volf, J., Sisak, F., Havlickova, H., Kummer, V., Imre, A., Annamaria Szmolka, A., and Nagy, B. (2009). Virulence potential of five major pathogenicity islands (SPI-1 to SPI-5) of *Salmonella* enterica serovar enteritidis for chickens. *BMC Microbiology,* 9, 268.

Sévellec, Y., Vignaud, M.-L., Granier, S. A., Lailler, R., Feurer, C., Le Hello, S., Mistou, M.-Y., and Cadel-Six, S. (2018). Polyphyletic nature of *Salmonella enterica* serotype derby and lineage-specific host-association revealed by genome-wide analysis. *Frontiers in Microbiology,* 9, 891.

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15, 224.

Taylor, A. J., Lappi, V., Wolfgang, W. J., Lapierre, P., Palumbo, M. J., Medus, C., and Boxrud, D. (2015). Characterization of foodborne outbreaks of *Salmonella enterica* serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance

and outbreak detection. *Journal of Clinical Microbiology*, 53(10), 3334–3340.

Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., and Malorny, B. (2020). Typing methods based on whole genome sequencing data. *One Health Outlook,* 2, 3.

Urwin, R., and Maiden, M. C. J. (2003). Multi-locus sequence typing: A tool for global epidemiology. *Trends in Microbiology,* 11(10), 479–487.

Vallenet, D., Calteau, A., Dubois, M., Amours, P., Bazin, A., Beuvin, M., Burlot, L., Bussell, X., Fouteau, S., Gautreau, G., Lajus, A., Langlois, J., Planel, R., Roche, D., Rollin, J., Rouy, Z., Sabatet, V., and Médigue, C. (2019). MicroScope: An integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, 48(D1), D579–D589.

Vernikos, G. S., and Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, 22(18), 2196–2203.

Wang, M., Qazi, I. H., Wang, L., Zhou, G., and Han, H. (2020). *Salmonella* virulence and immune escape. *Microorganisms,* 8(3), 407.

Yan, S., Zhang, W., Li, C., Liu, X., Zhu, L., Chen, L., and Yang, B. (2021). Serotyping, MLST, and core genome MLST analysis of *Salmonella enterica* from different sources in China during 2004–2019. *Frontiers in Microbiology*, 12, 688614.

Yoon, S. H., Park, Y.-K., and Kim, J, F. (2015). PAIDB v2.0: Exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Research*, 43(D1), D624–D630.

Yoshida, C. E., Kruczkiewicz, P., Laing, C. R., Jingohr, E. J., Gannon, V. P. J., Nash, J. H. E., and Taboada, E. N. (2016). The *Salmonella in silico* typing resource (SISTR): An open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *Plos ONE*, 11(1), e0147101.

Zhang, S., den Bakker, H. C., Li, S., Chen, J., Dinsmore, B. A., Lane, C., Lauer, A. C., Fields, P. I., and Deng, X. (2019). SeqSero2: Rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. *Applied and Environmental Microbiology*, 85(23), e01746–19.

Zhou, Z., Alikhan, N.-F., Mohamed, K., and Achtman, M. (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny and *Escherichia* core genomic diversity. *Genome Research*, 30(1), 138–152.