frontiers in Microbiology

METHODS
published: 24 July 2019
doi: 10.3389/fmicb.2019.01687

cano-wgMLST_BacCompare: A Bacterial Genome Analysis Platform for Epidemiological Investigation and Comparative Genomic Analysis

*Yen-Yi Liu[1]†, Ji-Wei Lin[2]† and Chih-Chieh Chen[2,3]**

[1] Central Regional Laboratory, Center for Diagnostics and Vaccine Development, Centers for Disease Control, Taichung, Taiwan, [2] Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung, Taiwan, [3] Rapid Screening Research Center for Toxicology and Biomedicine, National Sun Yat-sen University, Kaohsiung, Taiwan

With the decreasing cost of next-generation sequencing, whole-genome sequence-based bacterial genome comparisons are expected to become a mainstream process in the public health domain. Extended multilocus sequence typing (MLST) methods are becoming increasingly popular for use in comparing bacterial genetic relatedness in epidemiological investigations. Several extended MLST schemes based on biological signatures have been reported. Among them, whole-genome MLST (wgMLST) has gradually become one of the most widely used approaches for bacterial strain typing. In addition to using bacterial typing, many researchers aim to identify differences in the genes of compared strains. Because these differences might provide insights into critical bacterial functions, such as virulence and pathogenicity, researchers usually study these genes that differ between strains. Hence, we designed a web service tool based on wgMLST-constructed tree topology coupled with the feature selection method to create the "canonical wgMLST (cano-wgMLST) tree." The genes that differ between strains are shown at each split of the tree, thereby directly providing information for performing a comparative genomic analysis for each strain pair. We demonstrated that this web service tool could be operated efficiently on two datasets consisting of 22 *Campylobacter jejuni* isolates and 59 *S.* Heidelberg isolates, respectively. We imposed this tool on a designated web server, cano-wgMLST_BacCompare, to enable users to create a wgMLST tree and canonical wgMLST tree automatically from their uploaded bacterial genomes for not only epidemiological investigation but also comparative genomic analysis. Additionally, detailed information on how to use this service is provided. The cano-wgMLST_BacCompare is available at http://baccompare.imst.nsysu.edu.tw.

Keywords: molecular typing, next generation sequencing, whole-genome multilocus sequence typing, feature selection, epidemiological investigation, comparative genomic analysis

OPEN ACCESS

Edited by:
David W. Ussery,
University of Arkansas for Medical Sciences, United States

Reviewed by:
Marja-Liisa Hänninen,
University of Helsinki, Finland
Ole Lund,
Technical University of Denmark, Denmark

Correspondence:
Chih-Chieh Chen
chieh@imst.nsysu.edu.tw

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *18 April 2019*
Accepted: *09 July 2019*
Published: *24 July 2019*

Citation:
Liu Y-Y, Lin J-W and Chen C-C (2019) cano-wgMLST_BacCompare: A Bacterial Genome Analysis Platform for Epidemiological Investigation and Comparative Genomic Analysis. Front. Microbiol. 10:1687. doi: 10.3389/fmicb.2019.01687

INTRODUCTION

The origin of the multilocus sequence typing (MLST) approach can be traced back to 1998, when it was proposed by Maiden et al. (1998) as a method for overcoming the typing data exchange problem. By constructing a reference database containing alleles for seven housekeeping gene loci, researchers can easily exchange information regarding isolated strain types by generating a standardized allele profile (sequence type), which consists of a serial number obtained from comparisons made using the reference database. Although *Neisseria meningitidis* was the first species used to test the MLST approach, the approach has been proven to be applicable to many bacterial species, and the information of at least 100 species has been included in MLST allele databases (Jolley and Maiden, 2010; Jolley et al., 2018). After its development, the MLST approach rapidly gained popularity within the public health community. Several studies have demonstrated the successful detection of epidemiological isolates by using the MLST approach (Mellmann et al., 2011; Perez-Losada et al., 2013; Ludeke et al., 2015).

Because schemes selected by the MLST approach contain only few housekeeping-related gene loci, their discriminatory power may not be sufficient for distinguishing closely related isolates. Therefore, various schemes that recruit more gene loci, such as ribosomal MLST (Jolley et al., 2012) and plasmid MLST (Garcia-Fernandez et al., 2008; Carattoli et al., 2014), have been developed. Although MLST-based approaches are useful to public health researchers, the costs of MLST experiments were high in the Sanger sequencing era. With the reduction in cost of next-generation sequencing, many studies have attempted to evaluate whole-genome MLST schemes for comparing the genomes of several common pathogenic bacteria, such as *Salmonella* Enteritidis (Pearce et al., 2018), *Listeria monocytogenes* (Chen et al., 2016), *Vibrio parahaemolyticus* (Gonzalez-Escalona et al., 2017), *Enterococcus faecium* (de Been et al., 2015), and *Campylobacter jejuni* (Cody et al., 2017). These studies have tended to extend the MLST scheme from housekeeping genes to whole-genome genes (i.e., whole-genome MLST and wgMLST) and demonstrated that wgMLST has favorable discriminatory power for distinguishing highly closely related strains. The difficulty in comparing strain types across different laboratories is a limitation of the SNP-based approach used for molecular typing. However, this limitation can be overcome by using the wgMLST approach for bacterial strain typing. Beyond typing, many public health researchers have aimed to identify genes that exhibit differences among compared strains, because these differences may provide insights into critical bacterial functions, such as pathogenicity and virulence. Therefore, a tool that can be used for investigating genomic differences might be helpful for researchers.

In this study, by using the feature selection approach, we developed a user-friendly web server named cano-wgMLST_BacCompare to assist users in building a genetic relatedness tree based on the extracted canonical wgMLST gene set for not only epidemiological investigations but also functional investigations. We designed a two-layered feature selection approach for filtering loci. In the first layer, the whole-genome scheme is extracted for user-uploaded genome sequences. In the second layer, the "feature importance algorithm" is applied for selecting the most critical loci that possess a definite distinguishing ability. The cano-wgMLST_BacCompare web server is a user-friendly platform that can be used for performing comparative analyses. This web server mainly integrates the extraction of the whole genomes and identification of the most discriminatory loci. In addition, the genetic relatedness tree and heatmap profile indicating different genes for each split on the basis of the final created scheme are displayed on the result page.

METHODS AND IMPLEMENTATION

The cano-wgMLST_BacCompare server employs two major processes, namely whole-genome scheme extraction (GSE) and discriminatory loci refinement (DLR), that use the "feature importance" algorithm (Geurts et al., 2006). In the GSE process, a pipeline containing the "contig annotation" and "pan-genome allele database (PGAdb) creation" is used. In the DLR process, "feature importance levels with forests of trees" combined with a binary tree-traversal process were used to identify highly discriminatory loci for each selected split. The workflow for the cano-wgMLST_BacCompare server is displayed in **Figure 1**, and the details of methodologies used in this server are provided in the following sections.

GSE

Contig Annotation

We utilized a Prokka v1.11 (Seemann, 2014) pipeline to complete the annotation of contigs uploaded by users. The Prokka pipeline is a tool used for rapid prokaryotic genome annotation. The output file format of the annotation process is a "gff" file containing sequences and annotations to be used in the next step.

PGAdb Creation

A Roary v3.10.2 pipeline (Page et al., 2015) was used to process a large set of genomes. In this procedure, a "gff" file was used to arrange proteins into orthologous clusters. Paralogous genes were not included in the pan-genome data set. All orthologous clusters included a protein family with a 95% sequence identity. One locus (gene) represented one protein family. The "ffn" format files created in the previous step were used to convert proteins in each cluster into nucleotide sequences for establishing a pan-genome allele data set. In the case that one or more nucleotides

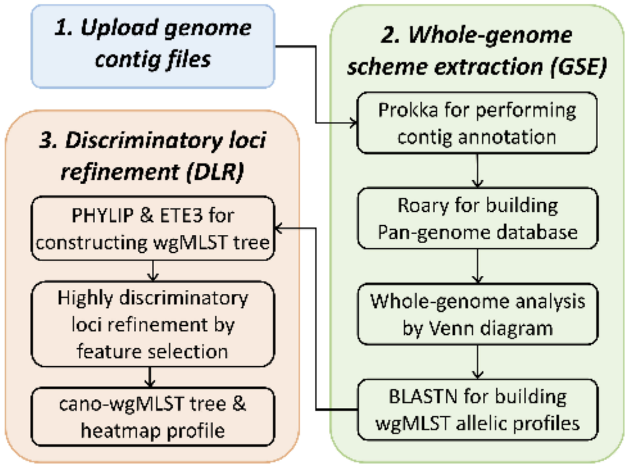

1. Upload genome contig files
2. Whole-genome scheme extraction (GSE)
Prokka for performing contig annotation
Roary for building Pan-genome database
Whole-genome analysis by Venn diagram
BLASTN for building wgMLST allelic profiles
3. Discriminatory loci refinement (DLR)
PHYLIP & ETE3 for constructing wgMLST tree
Highly discriminatory loci refinement by feature selection
cano-wgMLST tree & heatmap profile

FIGURE 1 | Schematic work flow of cano-wgMLST–BacCompare.

failed to provide a suitable match, sequences in a locus were defined as different alleles. The pan-genome allele data set was presented using a matrix table. Loci were encoded using the beginning three letters of genomes followed by a seven-digit number (e.g., SAL0000001or SAL0000002), where integers from one to *n* indicated the number of alleles in each locus.

DLR

wgMLST and cano-wgMLST Profiling

The uploaded genome contigs were compared with the constructed PGAdb by using BLASTN v2.2.30+ (Altschul et al., 1990) program with a minimum identity of 90% and coverage greater than 90% for the locus assignment and a exactly match for the allele assignment. The output indicated whether an allele was present in a locus. If alleles were absent, then "0" was assigned; if an allele was present, then the pre-established allele number was assigned. An "allelic sequence" was constructed after the comparison process.

Building a Genetic Relatedness Tree

A genetic relatedness tree was constructed from the previously established allelic sequence by using the clustering algorithm of the unweighted pair-group method with arithmetic averages in the PHYLIP v3.6 program (Felsenstein, 1981). Bootstrap values in a dendrogram were calculated using the ETE3 toolkit (Huerta-Cepas et al., 2016).

Feature Selection by "Feature Importance" Algorithm

Feature importance (Geurts et al., 2006) was applied in our study to determine discriminatory loci for each split. To include the features of the highest importance in the data sets, it was necessary to perform artificial classification. To this end, we used the ETE3 toolkit by traversing the Newick format tree produced from the user-selected scheme. The tree-traversal procedure was completed by using our in-house script. After the process, we could acquire a classified data set for performing the feature importance analysis (**Figure 2**). The feature importance method was based on Extra-Trees algorithms (Geurts et al., 2006), commonly referred to as decision trees (Quinlan, 1986). The classification process used artificial categorizations as targets to be matched using an Extra-Trees module. The feature importance analysis was completed using scikit-learn (Pedregosa et al., 2011), which is a machine learning module in Python. Results included the feature importance rank and magnitude. The schematic work flow of the tree-traversal and feature-extraction procedures are presented in **Figure 2**.

Robinson–Foulds Metric

The Robinson–Foulds (RF) metric (Robinson and Foulds, 1981) is the most widely used measure of phylogenetic tree similarity. Given two phylogenetic trees, the RF metric counts the number of splits or clades induced by one of the trees but not the other. In this study, the RF metric was used to measure the distance between genetic relatedness trees built on the basis of the whole-genome scheme (wgMLST tree) and highly discriminatory loci (cano-wgMLST tree).

Implementation

The cano-wgMLST_BacCompare server integrates five functional modules: "contig annotation," "PGAdb creation," "wgMLST profiling" (containing wgMLST and cano-wgMLST), "genetic relatedness tree construction," and "highly DLR." All of these modules were implemented in Perl. The computational task of "feature importance" analysis was completed using scikit-learn (Pedregosa et al., 2011), which is a machine learning module in Python. The webpage was constructed using HTML, JavaScript, PHP, jQuery, and jvenn (Bardou et al., 2014) JavaScript libraries.

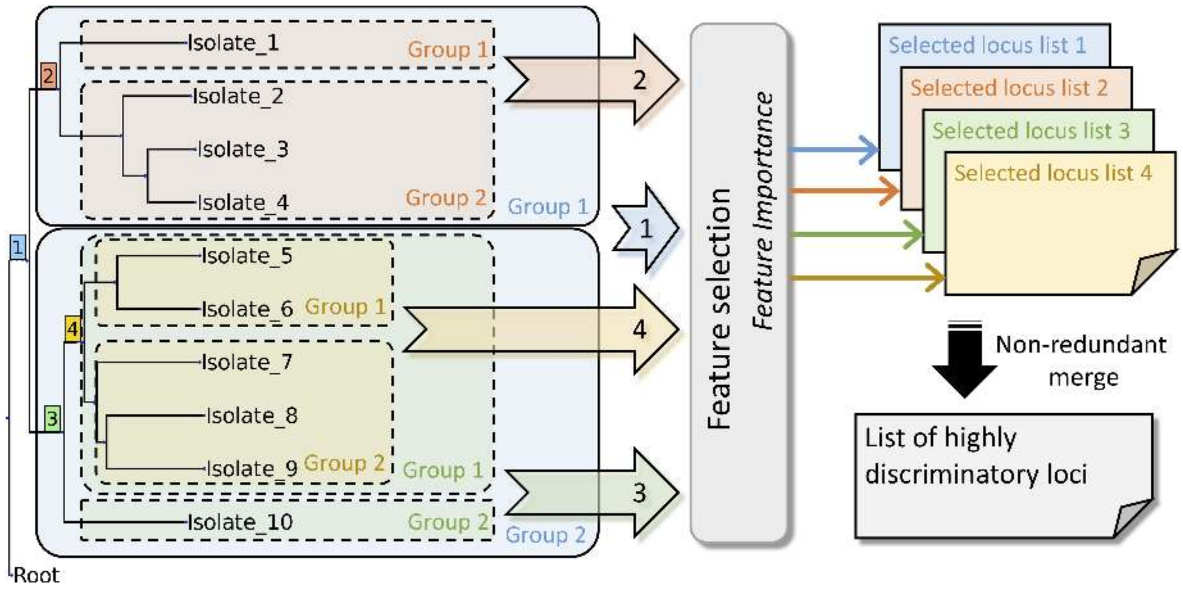


FIGURE 2 | Schematic work flow for identifying highly discriminatory loci.

The server runs on a 24-core Linux cluster with 2.40 GHz Intel Xeon processors.

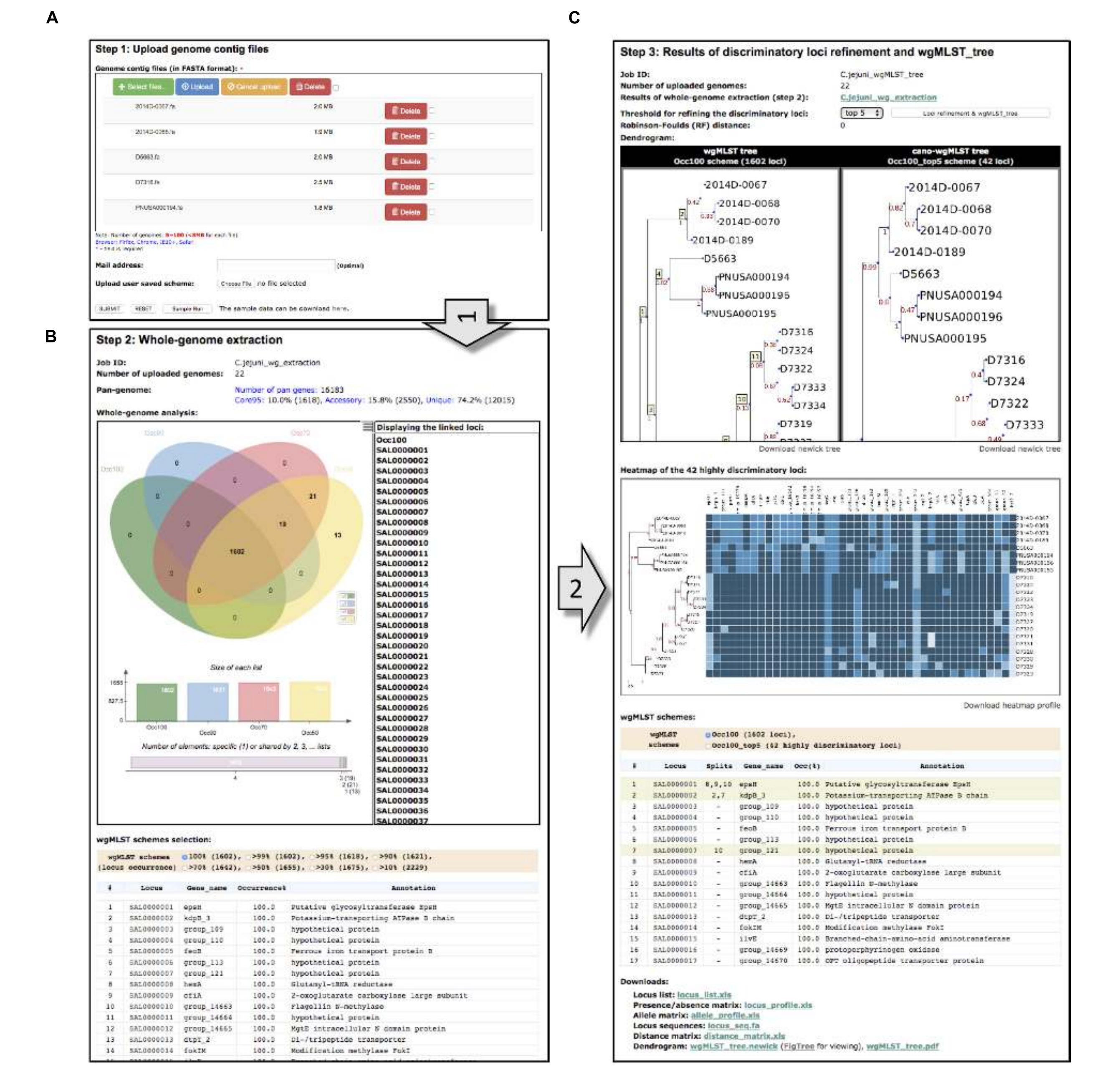
WEB SERVER

Input Format

The cano-wgMLST_BacCompare server is initiated by uploading a set of bacterial genome contigs (**Figure 3A**); then, standard processes, namely contig annotation, PGAdb creation, and wgMLST profiling, are run. The parameter of "protein sequence identity" for PGAdb creation is set at 95%, and parameters for "alignment coverage" and "alignment identity" for wgMLST profiling are both set at 90%. Users are encouraged to provide e-mail addresses to receive notifications when jobs are completed.

Output Format

The output of GSE comprises the following information. (A) A summary of settings; (B) an illustration of the probability



Step 1: Upload genome contig files

Genome contig files (in FASTA format): *

2014D-0067.fa	2.0 MB	Delete
2014D-0068.fa	1.9 MB	Delete
D5663.fa	2.0 MB	Delete
D7316.fa	2.5 MB	Delete
PNUSA000194.fa	1.8 MB	Delete

Mail address: (optional)

Upload user saved scheme: Choose File no file selected

SUBMIT RESET Sample Run The sample data can be download here.

Step 2: Whole-genome extraction

Job ID: C.jejuni_wg_extraction
Number of uploaded genomes: 22

Pan-genome: Number of pan genes: 16183
Core95: 10.0% (1618), Accessory: 15.8% (2550), Unique: 74.2% (12015)

Whole-genome analysis:

Displaying the linked loci:
Occ100
SAL0000001
SAL0000002
SAL0000003
SAL0000004
SAL0000005
SAL0000006
SAL0000007
SAL0000008
SAL0000009
SAL0000010
SAL0000011
SAL0000012
SAL0000013
SAL0000014
SAL0000015
SAL0000016
SAL0000017
SAL0000018
SAL0000019
SAL0000020
SAL0000021
SAL0000022
SAL0000023
SAL0000024
SAL0000025
SAL0000026
SAL0000027
SAL0000028
SAL0000029
SAL0000030
SAL0000031
SAL0000032
SAL0000033
SAL0000034
SAL0000035
SAL0000036
SAL0000037

wgMLST schemes selection:

wgMLST schemes (locus occurrence) ● 100% (1602), ○ >99% (1602), ○ >95% (1618), ○ >90% (1621), ○ >70% (1642), ○ >50% (1655), ○ >30% (1675), ○ >10% (2229)

#	Locus	Gene_name	Occurrence%	Annotation
1	SAL0000001	epsH	100.0	Putative glycosyltransferase EpsH
2	SAL0000002	kdpB_3	100.0	Potassium-transporting ATPase B chain
3	SAL0000003	group_109	100.0	hypothetical protein
4	SAL0000004	group_110	100.0	hypothetical protein
5	SAL0000005	feoB	100.0	Ferrous iron transport protein B
6	SAL0000006	group_113	100.0	hypothetical protein
7	SAL0000007	group_121	100.0	hypothetical protein
8	SAL0000008	hemA	100.0	Glutamyl-tRNA reductase
9	SAL0000009	cfiA	100.0	2-oxoglutarate carboxylase large subunit
10	SAL0000010	group_14663	100.0	Flagellin N-methylase
11	SAL0000011	group_14664	100.0	hypothetical protein
12	SAL0000012	group_14665	100.0	MgtE intracellular N domain protein
13	SAL0000013	dtpT_2	100.0	Di-/tripeptide transporter
14	SAL0000014	fokIM	100.0	Modification methylase FokI

Step 3: Results of discriminatory loci refinement and wgMLST_tree

Job ID: C.jejuni_wgMLST_tree
Number of uploaded genomes: 22
Results of whole-genome extraction (step 2): C.jejuni_wg_extraction
Threshold for refining the discriminatory loci: top 5 Loci refinement & wgMLST_tree
Robinson-Foulds (RF) distance: 0
Dendrogram:

wgMLST tree Occ100 scheme (1602 loci)	cano-wgMLST tree Occ100_top5 scheme (42 loci)
Download newick tree	Download newick tree

Heatmap of the 42 highly discriminatory loci:

Download heatmap profile

wgMLST schemes:

wgMLST schemes ● Occ100 (1602 loci), ○ Occ100_top5 (42 highly discriminatory loci)

#	Locus	Splits	Gene_name	Occ(%)	Annotation
1	SAL0000001	8,9,10	epsH	100.0	Putative glycosyltransferase EpsH
2	SAL0000002	2,7	kdpB_3	100.0	Potassium-transporting ATPase B chain
3	SAL0000003	-	group_109	100.0	hypothetical protein
4	SAL0000004	-	group_110	100.0	hypothetical protein
5	SAL0000005	-	feoB	100.0	Ferrous iron transport protein B
6	SAL0000006	-	group_113	100.0	hypothetical protein
7	SAL0000007	10	group_121	100.0	hypothetical protein
8	SAL0000008	-	hemA	100.0	Glutamyl-tRNA reductase
9	SAL0000009	-	cfiA	100.0	2-oxoglutarate carboxylase large subunit
10	SAL0000010	-	group_14663	100.0	Flagellin N-methylase
11	SAL0000011	-	group_14664	100.0	hypothetical protein
12	SAL0000012	-	group_14665	100.0	MgtE intracellular N domain protein
13	SAL0000013	-	dtpT_2	100.0	Di-/tripeptide transporter
14	SAL0000014	-	fokIM	100.0	Modification methylase FokI
15	SAL0000015	-	ilvE	100.0	Branched-chain-amino-acid aminotransferase
16	SAL0000016	-	group_14669	100.0	protoporphyrinogen oxidase
17	SAL0000017	-	group_14670	100.0	OPT oligopeptide transporter protein

Downloads:
Locus list: locus_list.xls
Presence/absence matrix: locus_profile.xls
Allele matrix: allele_profile.xls
Locus sequences: locus_seq.fa
Distance matrix: distance_matrix.xls
Dendrogram: wgMLST_tree.newick (FigTree for viewing), wgMLST_tree.pdf

FIGURE 3 | Features of the cano-wgMLST_BacCompare server. **(A)** Input page of the upload genome contig files; **(B)** output page of the whole-genome scheme extraction; **(C)** output page of the highly discriminatory loci refinement and wgMLST tree.

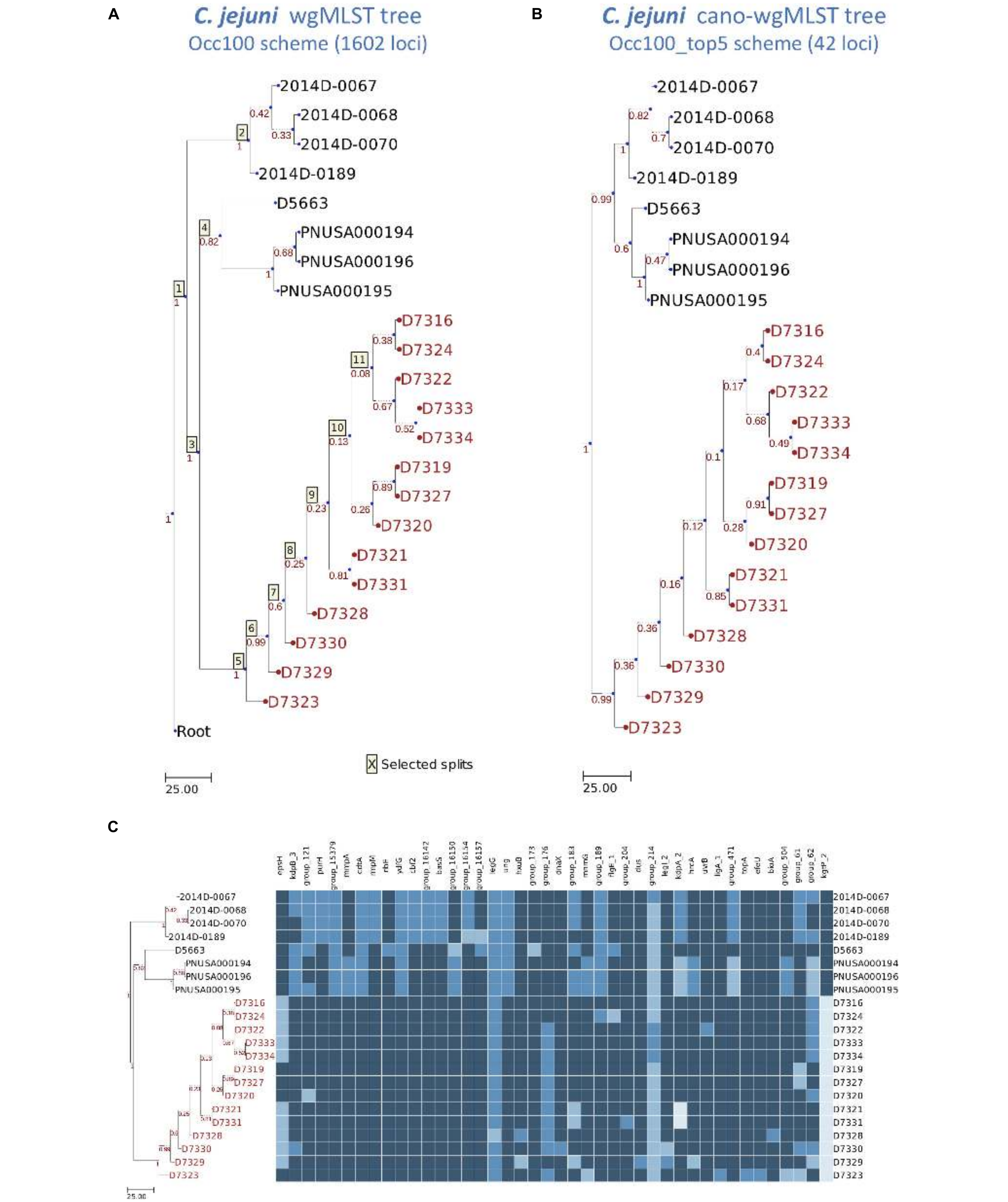


FIGURE 4 | Dendrogram and heatmap constructed using wgMLST profiles for 22 *Campylobacter jejuni* isolates. **(A)** The wgMLST tree generated on the basis of the Occ100 scheme (1602 loci). **(B)** The cano-wgMLST tree generated on the basis of the Occ100_top5 scheme (42 loci). The Occ100 scheme refers to the set of loci that are present in all isolates, and the Occ100_top5 scheme is a subset of the Occ100 scheme that unites the five most discriminatory loci for each split. Outbreak or event-related taxa are colored red. **(C)** The heatmap of 42 highly discriminatory loci. Different alleles in the same column are indicated by different colors.

ratio of a locus accounting for core95 genes (occurrence among isolates ≥95%), accessory genes (larger than one allele and <95%), and unique genes (one allele only) in the pan-genome database. The GSE output also includes (C) a four-part Venn diagram with ellipses that illustrates the composition of the whole-genome schemes (Occ100, Occ90, Occ70, and Occ50), marking the inclusion of loci when 100%, 90%, 70%, and 50% of isolates are presented, respectively; (D) a description table for the selected typing scheme; and (E) a button to perform the "loci refinement and wgMLST_tree" function. The output of DLR includes (A) a summary of settings; (B) a wgMLST tree constructed using the user-selected whole-genome scheme; (C) a cano-wgMLST tree constructed using the highly discriminatory loci; (D) Newick tree files to download; (E) a heatmap of the highly discriminatory loci; (F) a description table showing the

TABLE 1 | List of the 42 highly discriminatory loci selected by 22 *Campylobacter jejuni* isolates.

#	Locus	Splits[a]	Gene name	Annotation
1	SAL0000001	8,9,10	*epsH*	Putative glycosyltransferase EpsH
2	SAL0000002	2,7	*kdpB_3*	Potassium-transporting ATPase B chain
3	SAL0000007	10	*group_121*	Hypothetical protein
4	SAL0000258	1	*purH*	Bifunctional purine biosynthesis protein PurH
5	SAL0000726	3	*group_15379*	Hypothetical protein
6	SAL0000810	4	*mmpA*	Metalloprotease MmpA
7	SAL0000920	3	*cdtA*[b]	Cytolethal distending toxin subunit A precursor
8	SAL0001454	1	*rmpM*	Outer membrane protein class 4 precursor
9	SAL0001455	4	*ribE*	Riboflavin synthase
10	SAL0001471	3	*ydfG*[b]	NADP-dependent 3-hydroxy acid dehydrogenase YdfG
11	SAL0001477	1	*cbf2*	Putative peptidyl-prolyl *cis-trans* isomerase Cbf2 precursor
12	SAL0001479	1	*group_16142*	Hypothetical protein
13	SAL0001481	1	*basS*	Sensor protein BasS
14	SAL0001487	4	*group_16150*	Hypothetical protein
15	SAL0001491	2	*group_16154*	Hypothetical protein
16	SAL0001494	2	*group_16157*	Hypothetical protein
17	SAL0001508	8	*legG*	GDP/UDP-N,N'-diacetylbacillosamine 2-epimerase (hydrolyzing)
18	SAL0001515	3	*ung*[b]	Uracil-DNA glycosylase
19	SAL0001524	6,8	*hxuB*	Heme/hemopexin transporter protein HuxB precursor
20	SAL0001525	4	*group_173*	Hypothetical protein
21	SAL0001527	11	*group_176*	Putative type I restriction enzymeP M protein
22	SAL0001529	7	*dnaX*	DNA polymerase III subunit tau
23	SAL0001530	2,9	*group_183*	Hypothetical protein
24	SAL0001531	5	*mnmG*	tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG
25	SAL0001534	11	*group_189*	Hypothetical protein
26	SAL0001538	11	*flgE_1*	Flagellar hook protein FlgE
27	SAL0001539	9	*group_204*	Ribonuclease J 1
28	SAL0001540	6	*dus*	Putative tRNA-dihydrouridine synthase
29	SAL0001542	10,11	*group_214*	Hypothetical protein
30	SAL0001543	6,7	*legI_2*	N,N'-diacetyllegionaminic acid synthase
31	SAL0001544	9	*kdpA_2*	Potassium-transporting ATPase A chain
32	SAL0001548	6	*hrcA*	Heat-inducible transcription repressor HrcA
33	SAL0001557	11	*uvrB*	UvrABC system protein B
34	SAL0001560	5	*ligA_1*	DNA ligase
35	SAL0001565	4	*group_471*	Flagellar protein FlaG
36	SAL0001572	5	*topA*	DNA topoisomerase 1
37	SAL0001573	5	*efeU*	Ferrous iron permease EfeU precursor
38	SAL0001578	8	*bioA*	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
39	SAL0001595	5	*group_504*	Hypothetical protein
40	SAL0001599	7,10	*group_61*	Aminoglycoside 3-N-acetyltransferase
41	SAL0001600	2,6,7,8,9,10	*group_62*	Hypothetical protein
42	SAL0001601	3	*kgtP_2*[b]	Alpha-ketoglutarate permease

[a] *The split number can be referenced to the split marker that is labeled in the wgMLST tree (**Figure 4A**).* [b] *Selected genes that have the potential to be used for distinguishing between outbreak and non-outbreak isolates.*

selected whole-genome scheme and the highly discriminatory loci (labeled in lime); and (G) a summary of output files to download. Examples of GSE and DLR outputs are presented in **Figures 3B,C**, respectively.

EXAMPLES ANALYSIS

To evaluate the cano-wgMLST_BacCompare, we selected a data set comprising 22 isolates of *C. jejuni* from the benchmark for phylogenetic pipeline validation published by the Centers for Disease Control and Prevention (Timme et al., 2017). For this data set, the *C. jejuni* PGAdb contained 16,183 loci, of which 10.0% (1618 loci) belonged to the core95 genome, 15.8% (2550 loci) belonged to the accessory genome, and 74.2% (12,015 loci) belonged to unique genes (**Figure 3B**). In this step, we defined the core95 genome as that containing genes present in 95% of the tested genomes, an accessory genome as that containing genes present in two or more but less than 95% of the genomes, and unique genes as those present only in a single genome. The whole genome was further analyzed by constructing a Venn diagram. The PGAdb was then used to construct the wgMLST tree and cano-wgMLST tree and identify highly discriminatory loci for 22 *C. jejuni* isolates by using the DLR module (**Figure 3C**). In this step, the Occ100 scheme (genes present in all of the tested genomes) was used as the default scheme for constructing the wgMLST tree. The filtered Occ100_top5 scheme (a subset of the Occ100 scheme that unites the five most discriminatory loci for each split) was generated using tree-traversal and feature selection approaches, and the scheme was then used to construct the cano-wgMLST tree. Subsequently, the RF metric was used to measure the distance between the wgMLST tree and cano-wgMLST tree. Finally, the heatmap of highly discriminatory loci was also plotted to easily analyze allele distribution among uploaded isolates. The operation required approximately 1 h for the benchmark data set on a Linux server with 2.40 GHz Intel Xeon processors comprising 24 cores.

To analyze 22 *C. jejuni* isolates, 14 outbreak isolates could be clearly grouped, regardless of whether we used the Occ100 scheme (1602 loci) (**Figure 4A**) or the Occ100_top5 scheme (42 loci) (**Figure 4B**). The resulting dendrogram of the example data set exhibits high concordance (the measurements of the RF distance is equal to 0) between the wgMLST tree (**Figure 4A**) and cano-wgMLST tree (**Figure 4B**). We inferred that the Occ100_top5 scheme (the selected 42 highly discriminatory loci) can sufficiently distinguish between outbreak and sporadic isolates. The selected 42 loci were then used to generate a heatmap profile (**Figure 4C**). Detailed information of the 42 highly discriminatory loci is provided in **Table 1**. A heatmap can show how significantly the effect loci are distinguished for each split. By analyzing the heatmap profile, we observed that the genes *cdtA*, *ydfG*, *ung*, and *kgtP_2* had potential to be used for distinguishing between outbreak and non-outbreak isolates and that these genes were all selected from split 3 (**Table 1**). Among them, *cdtA* encoded the cytolethal distending toxin subunit A precursor, which was identified as a virulence-associated gene (Hickey et al., 2000); *ung* encoded uracil-DNA glycosylase, which is likely to be involved in the repair of uracil-containing DNA during base excision repair (Gaasbeek et al., 2009); and *kgtP* encoded α-ketoglutarate permease, which imports a tricarboxylic acid cycle intermediate required for the biosynthesis of glutamate, proline, arginine, and glutamine (Reid et al., 2008).

Another real case dataset consisting of next generation sequencing data for 59 *S.* Heidelberg, sequenced by Bekal et al. (2016), was also used to evaluate our approach. As illustrated in **Supplementary Figure S1**, the genetic relationships among the 59 isolates constructed using the cano-wgMLST approach were highly concordant with the relationships of the isolates determined using the high-quality core genome single-nucleotide variant (hqSNV) approach (three foodborne disease outbreaks), as shown in a Bekal's study (Bekal et al., 2016). The selected 125 loci were then used to generate a heatmap profile (**Supplementary Figure S2**). Detailed information of the 125 highly discriminatory loci is provided in **Supplementary Table S1**.

In summary, these examples demonstrate that our cano-wgMLST_BacCompare server not only correctly constructs genetic relatedness trees but also has the potential to select highly discriminatory loci from user-uploaded genome sequences.

DISCUSSION AND CONCLUSION

The proposed online tool, cano-wgMLST_BacCompare, which comprises two modules (i.e., GSE and DLR), was established to help users conduct epidemiological investigations and comparative genomic analyses involving bacterial whole-genome sequences. A strong advantage of the cano-wgMLST_BacCompare server is its incorporation of the feature selection method to filter the most important loci, which indicate the key genes contributing to the diversity between compared strains for each split along the wgMLST genetic relatedness tree. Because a tree can only be correctly interpreted from an evolutionary viewpoint, many horizontal gene transfer events (e.g., conjugation of plasmids) that might interfere with the correctness of a tree are usually excluded from calculations for the dendrogram. Therefore, it is reasonable to use conserved genetic markers, such as core genes, to construct the skeleton of the genetic relatedness tree; the further locus-reducing process is then applied on the basis of this tree topology. However, since many researchers may need to investigate differences in genes between strains for examining pathogenicity and antimicrobial resistance. Therefore, although our default tree building is based on the core scheme, we also provide different scheme options comprising gene loci with lower occurrences for users to choose from if their aim is mainly to investigate targets usually belonging to accessory genes. We believe that the cano-wgMLST_BacCompare can serve as a powerful online tool for not only epidemiological analysis but also comparative genomic analysis.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/WGS-standards-and-analysis/datasets.

AUTHOR CONTRIBUTIONS

Y-YL and C-CC conceived and designed the experiments. All authors performed the experiments, analyzed the data, built analysis tools and webpages, and wrote the manuscript.

FUNDING

This study was mainly supported by the grant from Ministry of Science and Technology (MOST grant numbers 107-2311-B-110-001 and 108-2221-E-110-061-MY2), NSYSU-KMU joint research project (#NSYSUKMU 108-P003), and "Rapid Screening Research Center for Toxicology and Biomedicine of Higher Education Enhancement Project" of the National Sun Yat-sen University and Ministry of Education, Taiwan.

ACKNOWLEDGMENTS

We are grateful to the Institute of Medical Science and Technology of National Sun Yat-sen University of Taiwan and the National Center for High-Performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) of Taiwan for providing the hardware, software, and computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01687/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1006/jmbi.1990.9999

Bardou, P., Mariette, J., Escudie, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. *BMC Bioinform.* 15:293. doi: 10.1186/1471-2105-15-293

Bekal, S., Berry, C., Reimer, A. R., Van Domselaar, G., Beaudry, G., Fournier, E., et al. (2016). Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar heidelberg clone in the context of outbreak investigations. *J. Clin. Microbiol.* 54, 289–295. doi: 10.1128/JCM.02200-15

Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O., Villa, L., et al. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi: 10.1128/AAC.02412-14

Chen, Y., Gonzalez-Escalona, N., Hammack, T. S., Allard, M. W., Strain, E. A., and Brown, E. W. (2016). Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. *Appl. Environ. Microbiol.* 82, 6258–6272. doi: 10.1128/aem.01532-16

Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D., and Maiden, M. C. J. (2017). Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human Disease Isolates. *J. Clin. Microbiol.* 55, 2086–2097. doi: 10.1128/JCM.00080-17

de Been, M., Pinholt, M., Top, J., Bletz, S., Mellmann, A., van Schaik, W., et al. (2015). Core genome multilocus sequence typing scheme for high- resolution typing of *Enterococcus faecium*. *J. Clin. Microbiol.* 53, 3788–3797. doi: 10.1128/jcm.01946-15

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi: 10.1007/bf01734359

Gaasbeek, E. J., van der Wal, F. J., van Putten, J. P., de Boer, P., van der Graaf-van Bloois, L., de Boer, A. G., et al. (2009). Functional characterization of excision repair and RecA-dependent recombinational DNA repair in Campylobacter jejuni. *J. Bacteriol.* 191, 3785–3793. doi: 10.1128/JB.01817-08

Garcia-Fernandez, A., Chiaretto, G., Bertini, A., Villa, L., Fortini, D., Ricci, A., et al. (2008). Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum beta-lactamases in *Escherichia coli* and *Salmonella* of human and animal origin. *J. Antimicrob. Chemother.* 61, 1229–1233. doi: 10.1093/jac/dkn131

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Gonzalez-Escalona, N., Jolley, K. A., Reed, E., and Martinez-Urtaza, J. (2017). Defining a core genome multilocus sequence typing scheme for the global epidemiology of *Vibrio parahaemolyticus*. *J. Clin. Microbiol.* 55, 1682–1697. doi: 10.1128/JCM.00227-17

Hickey, T. E., McVeigh, A. L., Scott, D. A., Michielutti, R. E., Bixby, A., Carroll, S. A., et al. (2000). Campylobacter jejuni cytolethal distending toxin mediates release of interleukin-8 from intestinal epithelial cells. *Infect. Immun.* 68, 6535–6541. doi: 10.1128/iai.68.12.6535-6541.2000

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 158, 1005–1015. doi: 10.1099/mic.0.055459-0

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.*org website and their applications*. *Wellcome Open Res.* 3:124. doi: 10.12688/wellcomeopenres.14826.1

Jolley, K. A., and Maiden, M. C. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform.* 11:595. doi: 10.1186/1471-2105-11-595

Ludeke, C. H., Gonzalez-Escalona, N., Fischer, M., and Jones, J. L. (2015). Examination of clinical and environmental Vibrio parahaemolyticus isolates by multi-locus sequence typing (MLST) and multiple-locus variable-number tandem-repeat analysis (MLVA). *Front. Microbiol.* 6:564. doi: 10.3389/fmicb.2015.00564

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., et al. (2011). Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6:e22751. doi: 10.1371/journal.pone.0022751

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Pearce, M. E., Alikhan, N. F., Dallman, T. J., Zhou, Z., Grant, K., and Maiden, M. C. J. (2018). Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. *Int. J. Food Microbiol.* 274, 1–11. doi: 10.1016/j.ijfoodmicro.2018.02.023

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Perez-Losada, M., Cabezas, P., Castro-Nallar, E., and Crandall, K. A. (2013). Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect. Genet. Evol.* 16, 38–53. doi: 10.1016/j.meegid.2013.01.009

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106.

Reid, A. N., Pandey, R., Palyada, K., Whitworth, L., Doukhanine, E., and Stintzi, A. (2008). Identification of Campylobacter jejuni genes contributing to acid adaptation by transcriptional profiling and genome-wide mutagenesis. *Appl. Environ. Microbiol.* 74, 1598–1612. doi: 10.1128/AEM.01508-07

Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Timme, R. E., Rand, H., Shumway, M., Trees, E. K., Simmons, M., Agarwala, R., et al. (2017). Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* 5:e3893. doi: 10.7717/peerj.3893

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.