microorganisms

MDPI

Article

The Cano-eMLST Program: An Approach for the Calculation of Canonical Extended Multi-Locus Sequence Typing, Making Comparison of Genetic Differences Among Bunches of Bacterial Strains

Yen-Yi Liu [1], Ji-Wei Lin [2] and Chih-Chieh Chen [2,3,4,*]

1 Central Regional Laboratory, Center for Diagnostics and Vaccine Development, Centers for Disease Control, Taichung 40855, Taiwan; current788@gmail.com
2 Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung 80424, Taiwan; jwlin@imst.nsysu.edu.tw
3 Rapid Screening Research Center for Toxicology and Biomedicine, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
4 General Institute of Clinical Medicine, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
* Correspondence: chieh@imst.nsysu.edu.tw; Tel.: +886-7-5252000 (ext. 5791)

Received: 20 February 2019; Accepted: 1 April 2019; Published: 3 April 2019

Abstract: Extended multi-locus sequence typing (eMLST) methods have become popular in the field of genomic epidemiology. Before eMLST methods can be applied in epidemiological investigations, the selection of a suitable scheme is critical. The core genome scheme (cgMLST) has become the most popular eMLST approach for strain typing in the epidemiological domain. In addition to strain typing, many public health researchers and clinical microbiologists wish to investigate which genes cause genetic differences between compared strains. Therefore, a tool that can be used to extract canonical genes with an eMLST scheme would be particularly useful. In this study, we present cano-eMLST, a well-designed program that applies a feature-selection methodology to create a canonical locus combination with discriminatory power by traversing a genetic relatedness tree based on a user-selected scheme. The cano-eMLST program is provided mainly to help infectious disease laboratory researchers identify potential factors related to bacterial pathogenesis. The core program (tree-traversing approach) of cano-eMLST is implemented in Perl and Python. All the necessary dependencies and environmental settings are provided in the encapsulated version (VirtualBox or VMware) and self-installation version (all use source code and libraries).

Keywords: molecular typing; next-generation sequencing (NGS); core-genome multi-locus sequence typing (cgMLST); feature-selection

1. Introduction

Pulsed-field gel electrophoresis (PFGE) molecular typing technology has been successfully employed to investigate foodborne disease epidemics, particularly by the nationwide molecular classification of foodborne bacterial disease monitoring network (PulseNet), which was established in 1996 [1]. Currently, the disease surveillance network has expanded beyond the United States to a global foodborne disease surveillance network known as PulseNet International [2]. PFGE is a standard typing tool used in PulseNet laboratories. This typing method is highly reliable and reproducible, and the typing results can be easily interpreted. However, PFGE is laborious and time-consuming and must be undertaken by skilled technicians. Studies have demonstrated that PFGE cannot effectively type certain species such as *Shigella sonnei* and *Salmonella enterica* serovar Enteritidis [3,4]. Multisite

variability repeat sequence analysis was once a promising alternative to PFGE, but the method is highly species specific and not a common bacterial population tool. Furthermore, a comparison of its subgroup profiles is difficult across different laboratories [5,6]. Therefore, the optimal typing method for bacteria is whole-genome sequencing (WGS). With the advancement of next-generation sequencing (NGS) technology, WGS is likely to supersede PFGE as the PulseNet standard typing method in the near future.

However, the NGS platform typically generates millions of short sequences, and the analysis of numerous WGS sequences to generate the required information (e.g., regarding genotyping and resistance to different strains) is a challenge. Most employees in conventional laboratories lack expertise in bioinformatics, and therefore a simple and easy-to-use analytical platform is required for automating the analysis of WGS primitive sequence fragments and for performing genotypic comparisons of different strains in the laboratory. Whole-genome multi-locus sequence typing (wgMLST) is a suitable method for decoding genes (gene fingerprints) specific to a particular strain on the basis of the bacterial genome and NGS data. If wgMLST profiles can be generated from a common genome in the pan-genome database and compared between laboratories, they can be used as the standard data type for WGS and in infectious disease surveillance networks such as PulseNet.

The wgMLST approach is becoming popular, and therefore many extended multi-locus sequence typing (eMLST) schemes, which involve different locus combinations within the whole-genome scheme, continue to be developed and evaluated by numerous public health research groups [7–14]. To use an eMLST method in bacterial molecular typing, the selection of a suitable typing scheme is critical. However, creating a scheme appropriate for use in the epidemiological typing of a specific bacterial species is challenging. In addition to the use of eMLST for typing, researchers usually wish to know the critical differences between the genomes they are comparing for antimicrobial and pathogenesis studies. Therefore, a user-friendly hands-on tool that helps to create a canonical typing scheme which is reduced to a human-readable quantity is crucial, particularly for infectious disease researchers seeking to identify potential factors related to bacterial pathogenesis. In this study, we present the program cano-eMLST, which uses a tree-traversing feature-selection approach to achieve this objective.

2. Materials and Methods

The genomes of bacterial strains are usually composed of core genomes in various proportions (genes present in all strains) and an accessory genome (genes not present in all strains) [15]. The cano-eMLST program (https://sourceforge.net/projects/cano-emlst/files/) operates according to user-selected bacterial genomes, through which it creates a database that recruits the core genes and all accessory genes of the included bacterial strains (pan-genome scheme). In the database created, each collected gene is assigned to a corresponding locus (gene group) as an allele according to the sequence identity. Next, the nonredundant alleles for each locus are serially numbered. The user-selected genomes are then compared with the created allele database using BLASTN [16], and the serial allele numbers (allelic profile) are taken from the schematic loci. The converted allelic profiles (one profile for one genome) can then be compared to generate a Hamming distance matrix (HDM). An unweighted pair group method with arithmetic mean (UPGMA) tree can then be constructed using the HDM. Our tree-traversing and feature-selection approaches can then be employed to walk through all the nodes of the dendrogram and select the most important loci that are distinguishable for each split simultaneously. After the tree traversal is complete, the final selected loci are defined for the highly discriminatory scheme.

To evaluate the schemes selected using the cano-eMLST approach, two empirical datasets consisting of 31 *Listeria monocytogenes* isolates and 10 *Escherichia coli* isolates were used [17]. The cano-eMLST program comprises five functional modules, namely the Contig Annotator, the pan-genome allele database (PGAdb) Builder, the pgMLST Profiler, the Dendro Plotter, and the Loci Extractor, which perform all the aforementioned steps. The Contig Annotator is used for annotating genome contigs, the PGAdb Builder is used for creating the pan-genome allele database (PGAdb)

on the basis of user-selected genomes, the pgMLST Profiler is used for generating allelic profiles by comparing user-selected genomes against the adjustable scheme selected from the constructed PGAdb, the Dendro Plotter is utilized for depicting UPGMA cladograms by using the allelic profiles generated from the same PGAdb, and the Loci Extractor is used for extracting the most important loci. The detailed descriptions of these five modules of the cano-eMLST program are provided in the following sections.

2.1. Contig Annotator

In the Contig Annotator step (Figure 1A), users import bacterial genome contigs (at least five genomes are required) in the FASTA format. The expensive computation required for contig annotation means that a computer with multicore processors for parallel processing is suggested for setting up the platform. In the platform, Prokka-v1.13 [18] is used for genome annotation.

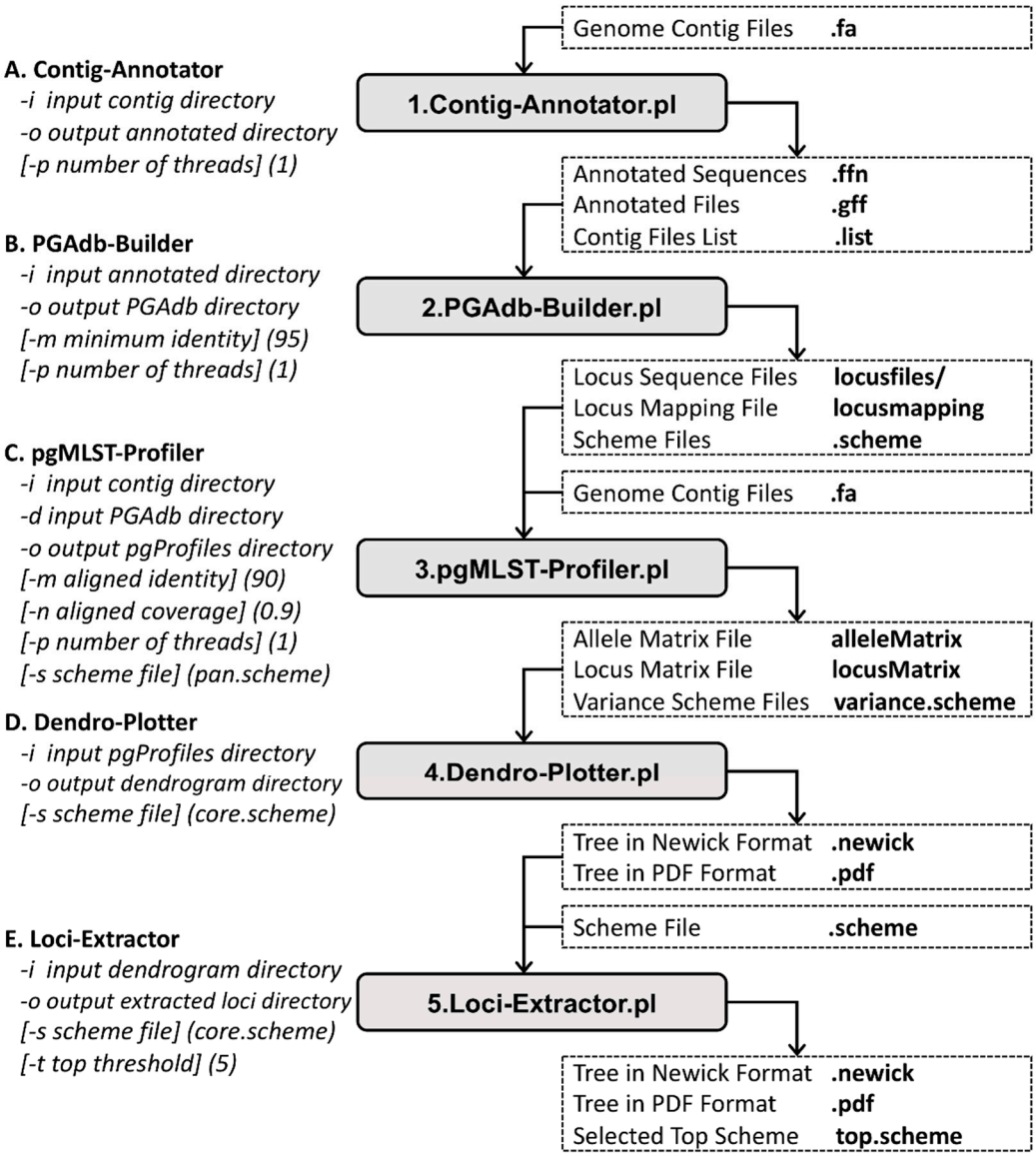

Figure 1. Schematic of cano-eMLST.

2.2. PGAdb Builder

In the PGAdb Builder step (Figure 1B), users import annotated bacterial genome files (at least five genomes are required) in the .ffn and .gff formats. In the platform, Roary-v3.12.0 [19] is used for

PGAdb construction. The running messages of the program appear in the terminal, and a log file is saved simultaneously.

2.3. pgMLST Profiler

In the pgMLST Profiler step (Figure 1C), users enter (1) the directory of bacterial genomes, (2) the directory of the PGAdb, (3) the output pgProfiles directory, and (4) the threshold of sequence identity and aligned coverage which create allelic profiles through blasting against the allele sequences within the selected scheme. The pan-genome scheme is selected by default if users do not specify a user-defined scheme file. All running messages appear in the log console, and the allelic profiles are generated locally.

2.4. Dendro Plotter

In the Dendro Plotter step (Figure 1D), users enter (1) the result directory of the pgMLST Profiler, (2) the output directory of the Dendro Plotter, and (3) the selected scheme for plotting the dendrogram. The core genome scheme is selected by default if users do not specify a user-defined scheme file. We employ the ETE3 Toolkit [20] for tree visualization. The dendrogram is plotted automatically and can be saved as a .pdf or .newick file.

2.5. Loci Extractor

In the Loci Extractor step (Figure 1E), users enter (1) the result directory of the Dendro Plotter, (2) the output directory of the Loci Extractor, (3) the selected scheme, and (4) the threshold for the selection of the most important loci for plotting the dendrogram. The pseudocodes of the tree-traversal and feature-extraction procedures are presented in Table 1. We employ the ETE3 Toolkit [20] for tree visualization. The dendrogram is plotted automatically and can be saved as a .pdf or .newick file.

Table 1. The Loci Extractor pseudocode.

```
1   Import ETE toolkit and scikit-learn library
2   Read newick tree file to be traversal
3   for each node in tree:
4           if node is leaf:
5                   continue
6           (subtree_1, subtree_2) = get_children (node)
7           if (leaf_amount(subtree_1) < 3 and leaf_amount(subtree_2) < 3):
8                   continue
9           (class_1, class_2) = (subtree_1, subtree_2)
10          Informative loci selected by feature importance
11  Non-redundant merge of the informative loci
```

3. Results

We used two benchmark datasets comprising 28 empirical outbreak isolates and three outgroups belonging to *Listeria monocytogenes* and three empirical outbreak isolates and seven outgroups belonging to *Escherichia coli* [17] to demonstrate the use of the cano-eMLST program. These datasets were used to help measure the closeness of the predicted dendrogram (Figures 2B and 3B) to the original dendrogram (Figures 2A and 3A). Input files of genomic data used for cano-eMLST were required to be in the FASTA format. The time required by the Contig Annotator and PGAdb Builder steps to process data concerning the 31 isolates was approximately 1 h. The pgMLST Profiler and Dendro Plotter steps each took approximately 2 min to be completed for the data of the 31 isolates. The final step of the traversing-tree (31 isolates) required 5 min to complete. After cano-eMLST processing, the final selected loci were reduced to 59 for the *Listeria monocytogenes* set (Figure 2) and 25 for the *Escherichia coli* set (Figure 3). These results can be compared with the core genome scheme comprising 2828 and 4650 loci for the *Listeria monocytogenes* and *Escherichia coli* sets, respectively.

The resulting dendrogram of the example dataset exhibits high concordance (the measurements of Robinson–Foulds distances [21] were both equal to 0) between the test sets and our approach (Figures 2 and 3).

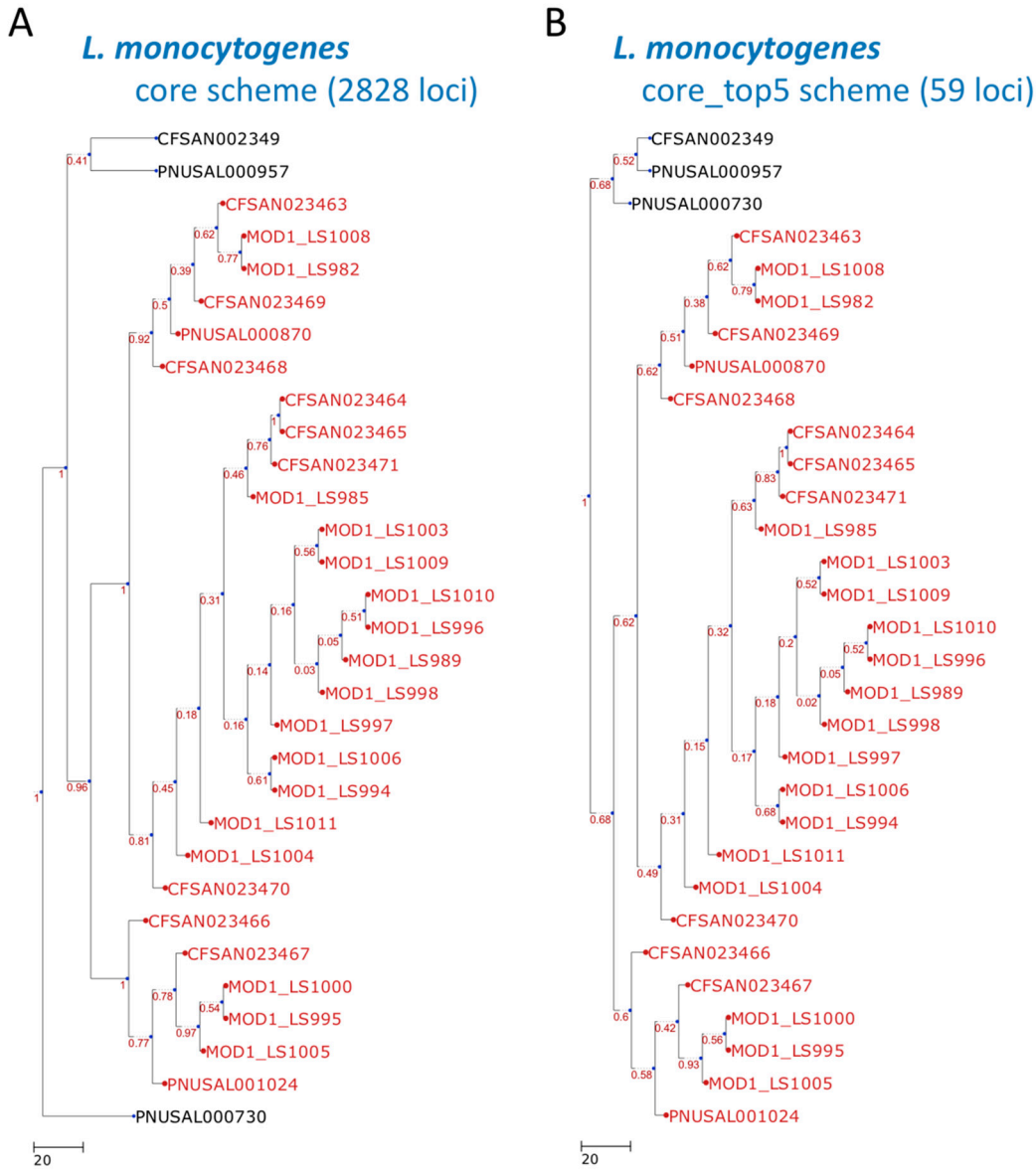

Figure 2. Dendrogram constructed with cgMLST profiles for 31 *Listeria monocytogenes* isolates. The cgMLST trees were generated according to the (**A**) core scheme (2828 loci) and (**B**) core_top5 scheme (59 loci). The outbreak or event-related taxa are colored red. The core scheme refers to the core genome scheme generated in step 2 by the PGAdb Builder. The core_top5 scheme is a subset of the core scheme that unites the five most discriminatory loci for each split. The scheme was generated in step five by the Loci Extractor.

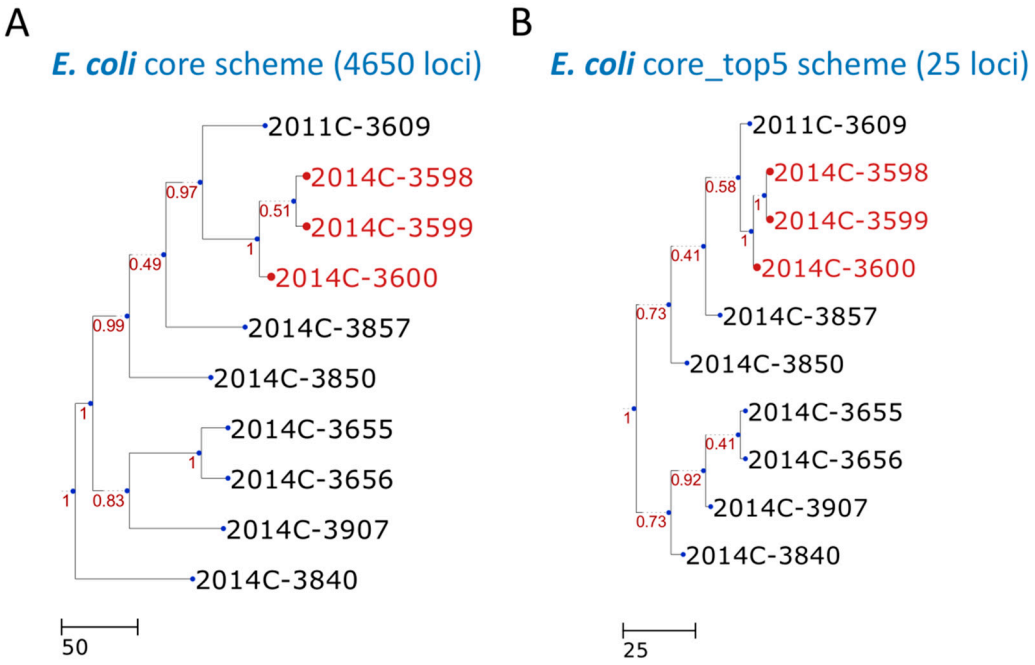

Figure 3. Dendrogram constructed with cgMLST profiles for ten *Escherichia coli* isolates. The cgMLST trees were generated according to the (**A**) core scheme (4650 loci) and (**B**) core_top5 scheme (25 loci). The outbreak or event-related taxa are colored red. The core scheme refers to the core genome scheme generated in step 2 by the PGAdb Builder. The core_top5 scheme is a subset of the core scheme that unites the five most discriminatory loci for each split. The scheme was generated in step five by the Loci Extractor.

Another real-case dataset that consisted of next-generation sequencing data for 34 *Salmonella* Typhimurium, sequenced by Leekitcharoenphon et al. [22], was also used to evaluate our approach. As illustrated in Supplementary Figure S1, the genetic relationships among the 34 isolates constructed using the cano-eMLST approach were highly concordant with the relationships of the isolates determined using the single nucleotide polymorphism (SNP)-based method (six foodborne disease outbreaks), as shown in a study by Leekitcharoenphon [22]. The detailed running times of jobs with different numbers of *Salmonella* isolates were compared with the default parameter setting on a desktop computer with two CPUs (6 cores, 12 threads, 2.10 GHz) and 64 GB RAM (Supplementary Figure S2).

4. Discussion

In our test, the schemes used for comparing the differences among isolates can dramatically reduce the number of loci from thousands to dozens. If the trees computed with the core genome scheme resemble those computed with the reduced core genome scheme, the reduced core genome loci are critically different from the compared isolates. Hence, the reduced core genome scheme may be used for typing, as well as in comparative genomic studies (e.g., for finding virulence factors). Currently, several popular genomic comparison tools, such as Mauve [23] and LAST [24], provide genome-to-genome comparison. The advantage of our tool is that it can be used to compare bunch isolates simultaneously and to extract the most different loci among the isolates compared for further studies.

This cano-eMLST program is expected to be available to clinical laboratories after an environment is established using the virtual machine we provided. The cano-eMLST can be used for rapidly capturing molecular information on the strains, such as the genetic relationship among strains and pathogenicity. Therefore, the program is highly useful, not only for investigating and controlling nosocomial infection in hospitals, but also for disease monitoring and epidemiological investigation. Because the clinical laboratories of hospitals are the target users of this tool, we have provided

step-by-step documentation on the installation and use of this program (see https://sourceforge.net/projects/cano-emlst/files/).

The application of eMLST approaches for epidemiological surveillance is set to become a trend in the public health domain. Several schemes with various functions have been reported in the literature. In addition to typing, the analysis of the differences between compared strains may be crucial for identifying the pathogenesis or antimicrobial drug resistance factors. In this study, we describe a feature-selection approach that can select a small set of loci (usually dozens) and adequately reproduce the genetic relatedness construct typically based on a large set of loci (usually thousands). Moreover, this dramatic locus reduction and the simplicity of the tree-traversing approach could enable researchers to determine the differences between various isolates. Thus, the cano-eMLST approach may be a valuable method for undertaking comparative genomic studies.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-2607/7/4/98/s1.

Author Contributions: Conceived and designed the experiments: Y.-Y.L. and C.-C.C. Performed the experiments and analyzed the data: Y.-Y.L., J.-W.L. and C.-C.C. Built analysis tools: Y.-Y.L., J.-W.L. and C.-C.C. Wrote the paper: Y.-Y.L. and C.-C.C.

Funding: This study was mainly supported by the grant from the 'Academic Summit Program' of the Ministry of Science and Technology (MOST 107-2311-B-110-001), an NSYSU-KMU joint research project, (#NSYSUKMU 107-P031), and the 'Rapid Screening Research Center for Toxicology and Biomedicine of Higher Education Enhancement Project' of National Sun Yat-sen University and Ministry of Education, Taiwan.

Acknowledgments: We are grateful to the National Center for High-Performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) of Taiwan for providing computational resources and storage resources.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Swaminathan, B.; Barrett, T.J.; Hunter, S.B.; Tauxe, R.V. PulseNet: The molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* **2001**, *7*, 382–389. [CrossRef] [PubMed]
2. Swaminathan, B.; Gerner-Smidt, P.; Ng, L.K.; Lukinmaa, S.; Kam, K.M.; Rolando, S.; Gutierrez, E.P.; Binsztein, N. Building PulseNet International: An interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog. Dis.* **2006**, *3*, 36–50. [CrossRef] [PubMed]
3. Liang, S.Y.; Watanabe, H.; Terajima, J.; Li, C.C.; Liao, J.C.; Tung, S.K.; Chiou, C.S. Multilocus Variable-Number Tandem Repeat Analysis for Molecular Typing of *Shigella sonnei*. *J. Clin. Microbiol.* **2007**, *45*, 3574–3580. [CrossRef]
4. Boxrud, D.; Pederson-Gulrud, K.; Wotton, J.; Medus, C.; Lyszkowicz, E.; Besser, J.; Bartkus, J.M. Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of *Salmonella enterica* serotype Enteritidis. *J. Clin. Microbiol.* **2007**, *45*, 536–543. [CrossRef]
5. Chiou, C.S. Multilocus variable-number tandem repeat analysis as a molecular tool for subtyping and phylogenetic analysis of bacterial pathogens. *Expert Rev. Mol. Diagn.* **2010**, *10*, 5–7. [CrossRef]
6. Chiou, C.S.; Izumiya, H.; Thong, K.L.; Larsson, J.T.; Liang, S.Y.; Kim, J.; Koh, X.P. A simple approach to obtain comparable Shigella sonnei MLVA results across laboratories. *Int. J. Med. Microbiol.* **2013**, *303*, 678–684. [CrossRef]
7. Kluytmans-van den Bergh, M.F.; Rossen, J.W.; Bruijning-Verhagen, P.C.; Bonten, M.J.; Friedrich, A.W.; Vandenbroucke-Grauls, C.M.; Willems, R.J.; Kluytmans, J.A. Whole-Genome Multilocus Sequence Typing of Extended-Spectrum-Beta-Lactamase-Producing Enterobacteriaceae. *J. Clin. Microbiol.* **2016**, *54*, 2919–2927. [CrossRef]
8. Kingry, L.C.; Rowe, L.A.; Respicio-Kingry, L.B.; Beard, C.B.; Schriefer, M.E.; Petersen, J.M. Whole genome multilocus sequence typing as an epidemiologic tool for Yersinia pestis. *Diagn. Microbiol. Infect. Dis.* **2016**, *84*, 275–280. [CrossRef] [PubMed]

9. Higgins, P.G.; Prior, K.; Harmsen, D.; Seifert, H. Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of Acinetobacter baumannii. *PLoS ONE* **2017**, *12*, e0179228. [CrossRef] [PubMed]
10. Bletz, S.; Janezic, S.; Harmsen, D.; Rupnik, M.; Mellmann, A. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of Clostridium difficile. *J. Clin. Microbiol.* **2018**, *56*, e01987-17. [CrossRef]
11. De Been, M.; Pinholt, M.; Top, J.; Bletz, S.; Mellmann, A.; van Schaik, W.; Brouwer, E.; Rogers, M.; Kraat, Y.; Bonten, M.; et al. Core Genome Multilocus Sequence Typing Scheme for High- Resolution Typing of Enterococcus faecium. *J. Clin. Microbiol.* **2015**, *53*, 3788–3797. [CrossRef]
12. Moran-Gilad, J.; Prior, K.; Yakunin, E.; Harrison, T.G.; Underwood, A.; Lazarovitch, T.; Valinsky, L.; Luck, C.; Krux, F.; Agmon, V.; et al. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Euro Surveill.* **2015**, *20*, 21186. [CrossRef] [PubMed]
13. Ruppitsch, W.; Pietzka, A.; Prior, K.; Bletz, S.; Fernandez, H.L.; Allerberger, F.; Harmsen, D.; Mellmann, A. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. *J. Clin. Microbiol.* **2015**, *53*, 2869–2876. [CrossRef] [PubMed]
14. Kohl, T.A.; Diel, R.; Harmsen, D.; Rothganger, J.; Walter, K.M.; Merker, M.; Weniger, T.; Niemann, S. Whole-genome-based Mycobacterium tuberculosis surveillance: A standardized, portable, and expandable approach. *J. Clin. Microbiol.* **2014**, *52*, 2479–2486. [CrossRef]
15. Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **2005**, *15*, 589–594. [CrossRef] [PubMed]
16. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
17. Timme, R.E.; Rand, H.; Shumway, M.; Trees, E.K.; Simmons, M.; Agarwala, R.; Davis, S.; Tillman, G.E.; Defibaugh-Chavez, S.; Carleton, H.A.; et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* **2017**, *5*, e3893. [CrossRef] [PubMed]
18. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef]
19. Page, A.J.; Cummins, C.A.; Hunt, M.; Wong, V.K.; Reuter, S.; Holden, M.T.; Fookes, M.; Falush, D.; Keane, J.A.; Parkhill, J. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **2015**, *31*, 3691–3693. [CrossRef]
20. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [CrossRef]
21. Robinson, D.F.; Foulds, L.R. Comparison of Phylogenetic Trees. *Math. Biosci.* **1981**, *53*, 131–147. [CrossRef]
22. Leekitcharoenphon, P.; Nielsen, E.M.; Kaas, R.S.; Lund, O.; Aarestrup, F.M. Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. *PLoS ONE* **2014**, *9*, e87991. [CrossRef] [PubMed]
23. Darling, A.C.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **2004**, *14*, 1394–1403. [CrossRef] [PubMed]
24. Kielbasa, S.M.; Wan, R.; Sato, K.; Horton, P.; Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **2011**, *21*, 487–493. [CrossRef] [PubMed]