frontiers in Microbiology

TECHNOLOGY REPORT
published: 03 August 2017
doi: 10.3389/fmicb.2017.01474

PathoBacTyper: A Web Server for Pathogenic Bacteria Identification and Molecular Genotyping

Ming-Hsin Tsai[1,2†], Yen-Yi Liu[1†] and Von-Wun Soo[2]*

[1] *Institute of Population Health Sciences, National Health Research Institutes, Miaoli County, Taiwan,* [2] *Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan*

With the decline in the cost of whole-genome sequencing because of the introduction of next-generation sequencing (NGS) techniques, many public health and clinical laboratories have started to use bacterial whole genomes for epidemiological surveillance and clinical investigation. For epidemiological and clinical purposes in this "NGS era," whole-genome-scale single nucleotide polymorphism (wgSNP) analysis for genotyping is considered suitable. In this paper, we present an online service, PathoBacTyper (http://halst.nhri.org.tw/PathoBacTyper/), for pathogenic bacteria identification and genotyping based on wgSNP analysis. More than 400 pathogenic bacteria can be identified and genotyped through this service. Four data sets containing 59 *Salmonella* Heidelberg isolates from three outbreaks with the same pulsed-field gel electrophoresis pattern, 34 *Salmonella* Typhimurium isolates from six outbreaks, 103 isolates of hospital-associated vancomycin-resistant *Enterococcus faecium* and 15 *Legionella pneumophila* isolates from clinical and environmental samples in Israel were used for demonstrating the operation and testing the performance of the PathoBacTyper service. The test results reveal the applicability of this service for epidemiological typing and clinical investigation.

Keywords: next-generation sequencing, pan-genome database, whole genome multilocus sequence typing, molecular typing, bacterial identification

OPEN ACCESS

Edited by:
Fumito Maruyama,
Kyoto University, Japan

Reviewed by:
Makoto Kuroda,
National Institute of Infectious Diseases, Japan
Tomotada Iwamoto,
Kobe Institute of Health, Japan

****Correspondence:***
Von-Wun Soo
soo@cs.nthu.edu.tw

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *01 May 2017*
Accepted: *20 July 2017*
Published: *03 August 2017*

Citation:
Tsai M-H, Liu Y-Y and Soo V-W (2017) PathoBacTyper: A Web Server for Pathogenic Bacteria Identification and Molecular Genotyping. Front. Microbiol. 8:1474. doi: 10.3389/fmicb.2017.01474

INTRODUCTION

The genomic DNA of organisms carries genetic information that is biologically functional. Decoding the entire genome sequence of an organism is a fundamental task in complex biological studies. Previously, the conventional Sanger sequencing method was used to decode the complete bacterial genome sequence; however, this method is very expensive and tedious. In recent years, considerable progress has been made in next-generation sequencing (NGS) technology. Currently, the NGS method can facilitate bacterial genome decoding within days and at a cost of less than US$100. Therefore, in the near future, whole-genome sequencing (WGS) is expected to be used in clinical and public health laboratories and to become a routine diagnostic and genotyping tool for disease surveillance, resistance prediction, cluster infection examination, and establishing evolutionary relationships among different strains (Parkhill et al., 2001; Merker et al., 2013; Struelens and Brisse, 2013; Franz et al., 2014; Gordon et al., 2014; Halachev et al., 2014; Joensen et al., 2014; Koser et al., 2014; Luo et al., 2014; Schmid et al., 2014). In addition to its clinical and public health applications, WGS is a very effective method for basic biomedical research such as studies on the pathogenesis of human diseases (Acke et al., 2014; Hoffmann et al., 2014; Meinel et al., 2014). However, the NGS platform typically generates millions of short sequences, and the analysis of such a large number of short WGS sequences to generate the required information, such as the genotype and resistance to different strains, is a challenge. Because most researchers in clinical and public health laboratories lack expertise in bioinformatics, developing a simple and easy-to-use analytical platform for automating the analysis of WGS primitive sequence fragments and for performing genotypic comparison of different strains in the laboratory is necessary.

The whole-genome-scale single nucleotide polymorphism (wgSNP) approach has been demonstrated to be suitable for bacterial strain genotyping (Achtman, 2008; Nielsen et al., 2011; Leekitcharoenphon et al., 2014). Many researchers have successfully applied wgSNP analysis for detecting outbreaks of different types of pathogenic bacteria (Holt et al., 2010; Bakker et al., 2011; Octavia et al., 2015; Taylor et al., 2015; Bekal et al., 2016). Several effective tools such as Lyve-SET (Katz et al., 2017), SNVPhyl (Petkau et al., 2016), and CFSAN SNP Pipeline (Davis et al., 2015) have been designed for manipulating WGS data to generate wgSNP metadata. However, these wgSNP tools are usually deployed as command-line programs, which is inconvenient and difficult for most wet-lab employees who must manipulate the WGS data in clinical and public health laboratories. Therefore, first-line lab personnel have a high demand for an easy-to-use tool that can help them generate wgSNP metadata from WGS data for further applications.

In this paper, we present a web-service tool, PathoBacTyper, for pathogenic bacterial identification and molecular genotyping of more than 400 pathogenic bacteria. In this study, we demonstrated the operation of PathoBacTyper by identifying and genotyping four data sets, such as 59 *Salmonella* Heidelberg WGS raw reads from three outbreaks previously sequenced by Bekal et al. (2016), 34 *Salmonella* Typhimurium isolates (Leekitcharoenphon et al., 2014), 103 isolates of hospital-associated vancomycin-resistant *Enterococcus faecium* (De Been et al., 2015) and 15 *Legionella pneumophila* isolates from clinical and environmental samples in Israel (Moran-Gilad et al., 2015). The 34 *Salmonella* Typhimurium isolates are consisted of 18 strains from six previously studied outbreaks (Torpdahl et al., 2007; Petersen et al., 2011) and 16 unrelated strains. The 103 *E. faecium* strains are comprised of 46 strains isolated from German hospital during 2003 to 2006, 37 isolates from Danish hospitals during 2012–2013, and 20 isolates from Dutch hospitals during 2012 to 2013. The test results reveal the applicability of this service for epidemiological typing and clinical investigation.

METHODS AND IMPLEMENTATION

PathoBacTyper provides two functions: species identification and bacterial strain typing. In species identification, user-uploaded WGS reads are distributed to a prebuilt reference data set comprising 478 pathogenic bacterial genomes; possible species can then be identified according to the quantity of the mapped reads. After the species of the user-uploaded isolate whole genome sequence is recognized, the corresponding reference sequence of the species is automatically selected for the following strain typing process. **Figure 1** presents the overall workflow of PathoBacTyper. A detailed description of the methodologies used in PathoBacTyper for species identification and bacterial strain typing follows.

Species Identification

In the species identification process, the user-uploaded bacterial isolate WGS reads are mapped to a species genome reference database (SGRdb) comprising 478 pathogenic bacteria, which were selected from the list provided by Scholz et al. (2016). The isolates included in the SGRdb are presented in Supplementary Table S1. For mapping the reads, we applied the hash-based method (Alkan et al., 2009; Weese et al., 2009), which transforms tens of thousands of read sequences into a k-mer patterned hash table, and then the best match is rapidly obtained through comparison with the hash tables created from the reference genomes (i.e., genomes from the SGRdb in our case). In this study, a 30-mer hash table for the SGRdb (SGRdbht) was constructed using a sliding window with a 30-mer step size. The system creates hash keys by using the uploaded raw reads to query the SGRdbht for the species identification process. A coverage ratio is defined to evaluate the support level of the candidate sequence from the SGRdb. The coverage ratio is defined as $C = R/M$, where C is the coverage ratio, R denotes a number of mapping raw reads hash keys onto the positions of candidate sequence, and M denotes a total number of hash positions on the candidate sequence.

Bacterial Strain Typing

In the bacterial strain typing process, the user-uploaded WGS raw reads and the reference genome, which is selected from the SGRdb on the basis of the identification results from the species identification process or is uploaded by the user, are used as the input. We selected the Lyve-SET (Katz et al., 2017) method for calling wgSNPs; this method creates high-quality wgSNPs from WGS raw reads. The distance matrix computed using Lyve-SET is used for depicting a multidimensional scaling (MDS) plot, which can provide a good presentation of outbreak clusters. If input files are assembly contigs, WgSim is used to simulate raw reads in Lyve-SET. To omit the extra reads simulation process, the Parsnp (Treangen et al., 2014) method replaces Lyve-SET when assembly contig files are uploaded. In the process of calculating SNPs with assemblies, Parsnp is more accurate and faster than Lyve-SET. The FastTree (Price et al., 2010) is used to calculate the alignment files, which are outputted from both Lyve-SET and Parsnp to build phylogenetic trees. Moreover, we constructed a maximum-likelihood phylogenetic tree with confidence values labeled on the branches by using the ETE toolkit (Huerta-Cepas et al., 2016).

Implementation

The PathoBacTyper service was built by integrating the species identification and bacterial strain typing functional modules in

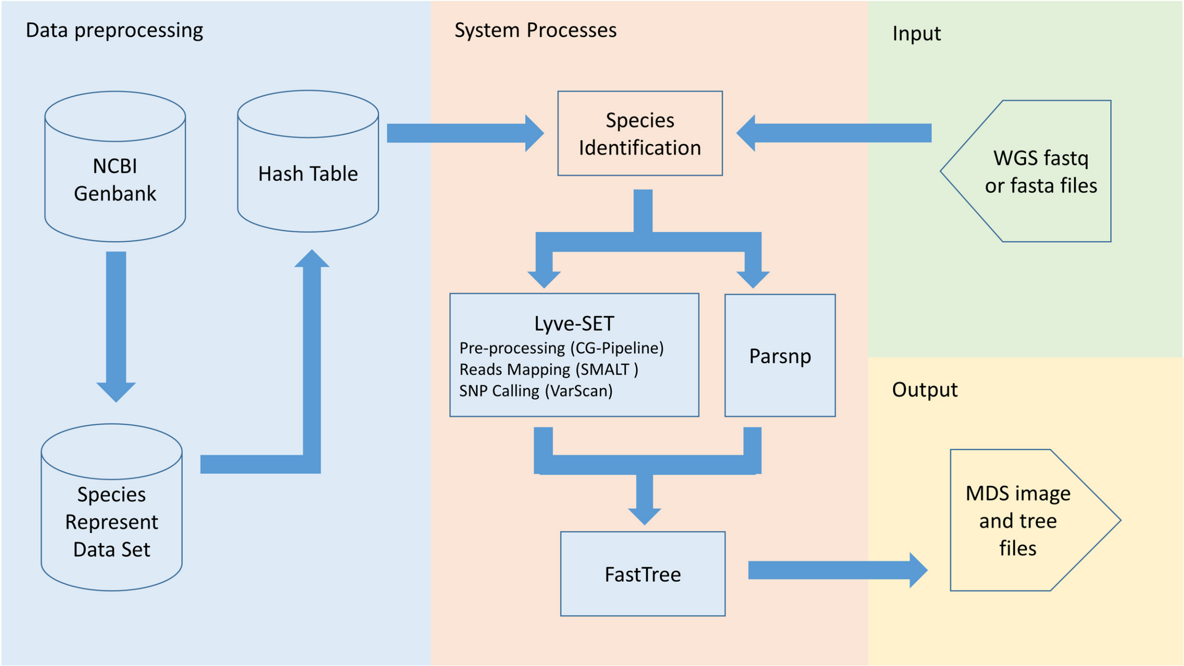


FIGURE 1 | The schematic work flow of PathoBacTyper.

Java programs. The web page was constructed using HTML, JavaScript, and JSP. The service runs on a Linux server with two 2.40-GHz Intel Xeon processors comprising 16 cores.

WEB SERVER

Input Format

PathoBacTyper accepts two genomic sequence formats, "fastq" and "fasta." The fastq data can be in the gzip format (.gz). All upload files are deleted after the analysis is completed, and the size of every uploaded file cannot exceed 1 GB. PathoBacTyper typically takes 587 and 101 min on average to finish all the processes for a data set comprising 59 *S.* Heidelberg genomes in the fastq and fasta formats, respectively. Users are encouraged to provide e-mail addresses for obtaining notification of the results when their jobs are finished. The home page for the user to upload WGS data is shown in **Figure 2A**.

Output Format

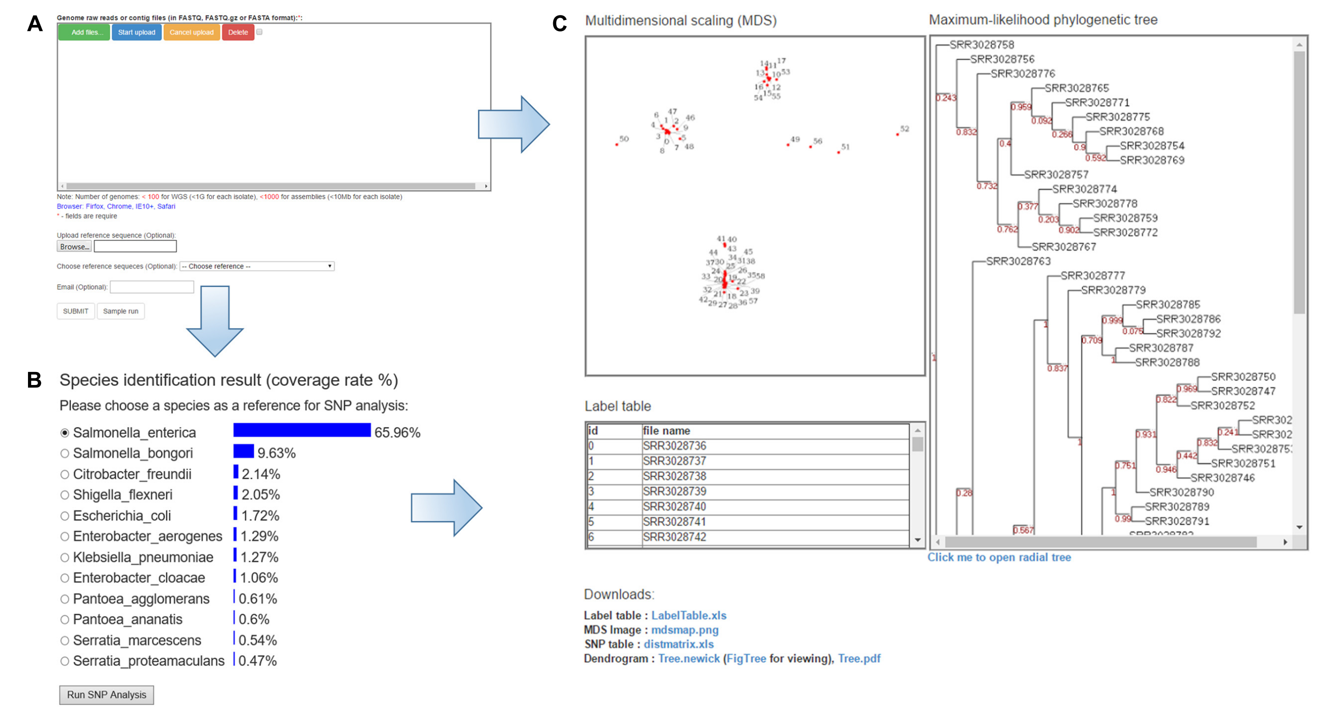
The PathoBacTyper output is composed of two sections, species identification (**Figure 2B**) and SNP analysis (**Figure 2C**). In the species identification part, the system lists the top 10 candidate species, ranked by confidence values. For the SNP analysis part, the result page includes (A) an MDS image file; (B) a label table, which provides a link from MDS labels to the upload file name; (C) a maximum-likelihood phylogenetic tree (in the pdf and png image file formats); and (D) a phylogenetic tree text file in the Newick format. The Newick tree file can be reused to draw the phylogenetic tree in other software. If users enter their email addresses on the job submission page, the system sends them a link to the URL of the result page when the job is finished.

EXAMPLE ANALYSIS

Four data sets are used to evaluate our implementation, such as 59 *S.* Heidelberg isolate genomes from three outbreaks with an identical pulsed-field gel electrophoresis type (Bekal et al., 2016), 34 *S.* Typhimurium isolate genomes from six outbreaks (Leekitcharoenphon et al., 2014), 103 isolates of hospital-associated vancomycin-resistant *E. faecium* (De Been et al., 2015), and 15 *L. pneumophila* isolates (Moran-Gilad et al., 2015). The strain name and ENA accession number of 103 *E. faecium* isolates and 15 *L. pneumophila* isolates are listed in the Supplementary Table S4, S5 separately. In *E. faecium* analysis, The PathoBacTyper successfully identified each subtype, as shown in Supplementary Figure S1. The Supplementary Figure S2 shows the result of *L. Pneumophila* analysis. The strains labeled by Lp-001, Lp-2002694p7, and Lp-012 are far from the major cluster, that is similar to the spanning tree in the original paper. Other strains are clustered that is consistent with literature data. Other details are demonstrated in following sections.

Demonstration of the Species Identification and the Molecular Typing Functionality for *S.* Heidelberg

We tested the operation of PathoBacTyper by using a data set comprising 59 *S.* Heidelberg isolate genomes from three outbreaks with an identical pulsed-field gel electrophoresis



Genome raw reads or contig files (in FASTQ, FASTQ.gz or FASTA format):*:

Note: Number of genomes: < 100 for WGS (<1G for each isolate), <1000 for assemblies (<10Mb for each isolate)
Browser: Firfox, Chrome, IE10+, Safari
* - fields are require

Upload reference sequence (Optional):

Choose reference sequeces (Optional): -- Choose reference --

Email (Optional):

Species identification result (coverage rate %)

Please choose a species as a reference for SNP analysis:

● Salmonella_enterica	65.96%
○ Salmonella_bongori	9.63%
○ Citrobacter_freundii	2.14%
○ Shigella_flexneri	2.05%
○ Escherichia_coli	1.72%
○ Enterobacter_aerogenes	1.29%
○ Klebsiella_pneumoniae	1.27%
○ Enterobacter_cloacae	1.06%
○ Pantoea_agglomerans	0.61%
○ Pantoea_ananatis	0.6%
○ Serratia_marcescens	0.54%
○ Serratia_proteamaculans	0.47%

Label table

id	file name
0	SRR3028736
1	SRR3028737
2	SRR3028738
3	SRR3028739
4	SRR3028740
5	SRR3028741
6	SRR3028742

Downloads:

Label table : LabelTable.xls
MDS Image : mdsmap.png
SNP table : distmatrix.xls
Dendrogram : Tree.newick (FigTree for viewing), Tree.pdf

FIGURE 2 | The features of the PathoBacTyper server. **(A)** Input page. **(B)** Result page of the *Species identification*. **(C)** Output page of the results, such as phylogenetic tree MDS and label table.

type (Bekal et al., 2016). The WGS raw reads of these 59 *S.* Heidelberg isolates (Supplementary Table S2) were downloaded from the NCBI SRA database[1]. The SRA Toolkit[2] was used to convert the downloaded raw reads in the sra format to the fastq format. Moreover, we *de novo* assembled all 59 *S.* Heidelberg genomes with CLC v9.5.2. For processing the example data, approximately 9.5 h were required on a Linux server with two 2.40-GHz Intel Xeon processors comprising 16 cores to finish all the processes for outputting the MDS plot and the phylogenetic tree. We spent 19 min and approximately 1 min, depending on network traffic, uploading all 59 *S.* Heidelberg genomes in the fastq (raw reads) and fasta (assemblies) formats, respectively. The process times of the test data set were 587 and 101 min for raw reads and assemblies, respectively. The results of raw reads reveal that three outbreaks can be identified as distinct clusters in both the MDS plot (**Figure 3A**) and phylogenetic tree (**Figure 3B**). As shown in **Figure 3B**, the isolates differed at 10 SNPs or less within the same outbreak and over 50 SNPs among distinct outbreaks. The genetic relationships among the 59 isolates were highly concordant with the epidemiological definitions in a previous study (Bekal et al., 2016). We compared the results between the two different inputs, raw reads and assemblies (Supplementary Figure S3). Although both tree topologies were not identical, the three outbreaks can be clearly distinguished.

Demonstration of the Species Identification and the Molecular Typing Functionality for *S.* Typhimurium

The 34 isolates of *S.* Typhimurium are consisted of 18 strains from six outbreaks and 16 unrelated strains (Leekitcharoenphon et al., 2014). The WGS raw reads of these 34 *S.* Typhimurium isolates (Supplementary Table S3) were downloaded from the NCBI SRA database. Same as above section, the raw reads are converted to fastq format by the SRA Toolkit. The results illustrate that six outbreaks can be identified as distinct clusters in both the MDS plot (**Figure 4A**) and phylogenetic tree (**Figure 4B**). We tested the operation of PathoBacTyper with four data sets to confirm our implement is stable enough, and the output is consistent with literature.

DISCUSSION

The main purpose of our study is to implement a convenient platform for researchers who are interested in epidemiological typing and clinical investigation. The PathoBacTyper accepts two types of sequences, raw reads and assembly contigs, and no more parameters are required. This website provides a visualization result page, MDS plot and phylogenetic tree, which lets users easier to understand relationship among outbreaks. Furthermore, we provide several download links for users to validate the analysis results or redraw the tree in their favorite style. However, the computing power seems not sufficient to

[1] https://www.ncbi.nlm.nih.gov/sra
[2] https://github.com/ncbi/sra-tools

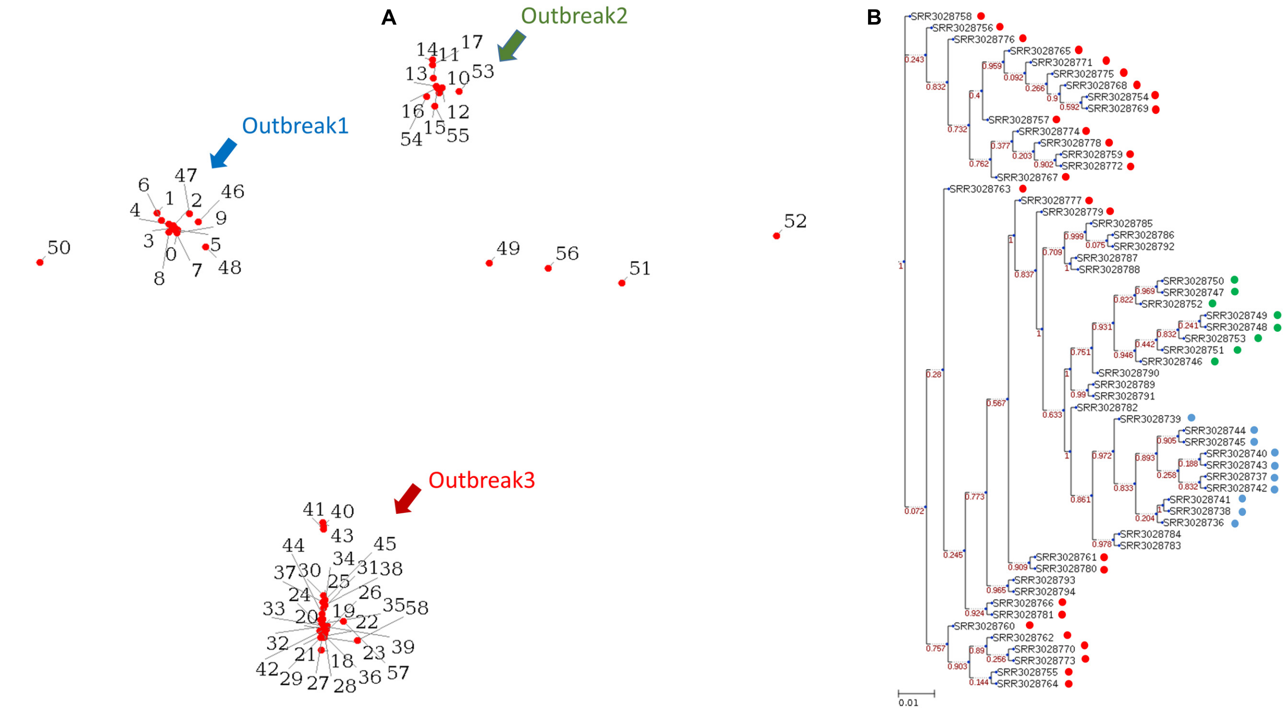


FIGURE 3 | Reveals the 59 *S.* Heidelberg isolates from three outbreaks, which are identified as distinct clusters in both the MDS plot **(A)** and phylogenetic tree **(B)**. The blue, green, and red circles represent the outbreaks outbreak 1, outbreak 2, and outbreak 3 separately in **(B)**.

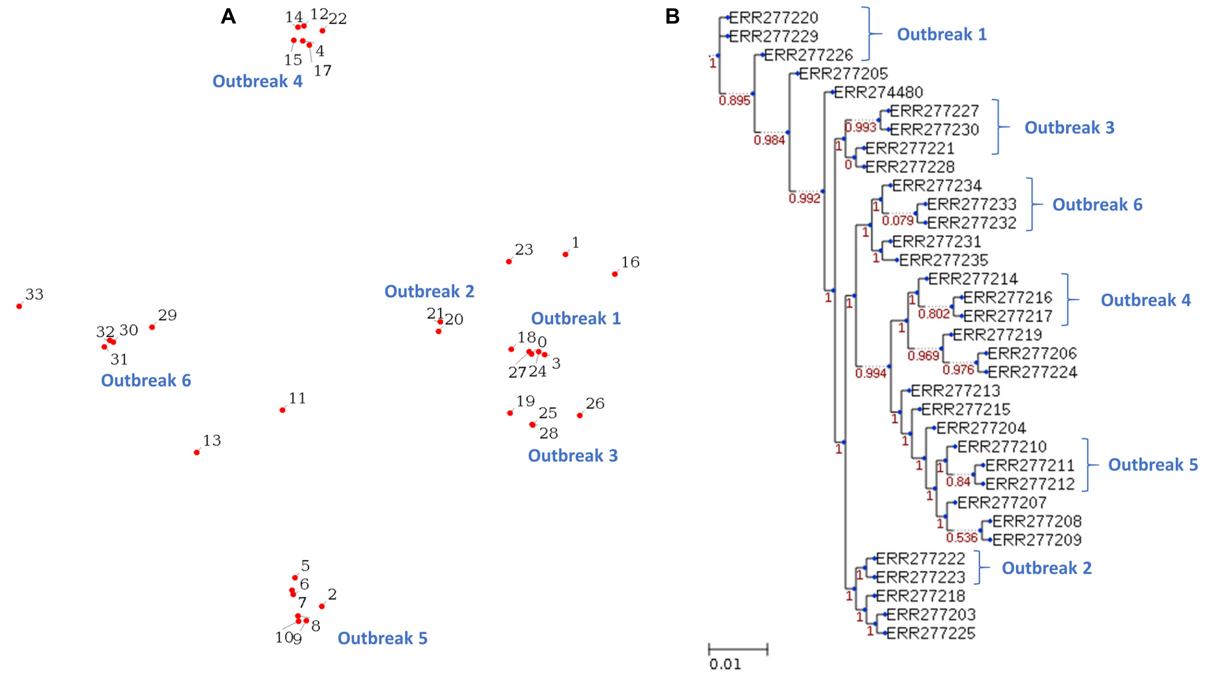


FIGURE 4 | Reveals the 34 *S.* Typhimurium isolates from six outbreaks, which are identified as distinct clusters in both the MDS plot **(A)** and phylogenetic tree **(B)**. The six outbreaks are labeled with outbreak 1–6 in **(A, B)**.

satisfy all requests from internet. The NGS sequences alignment essentially consume a lots of computing resource, therefore a standalone version of PathoBacTyper is needed. Additionally, the function of species identification is limited if user upload a species which is not included in the database SGRdb. To solve the problems mentioned above, we provide PathoBacTyper as a virtual machine image in the download page[3]. Users can download and run PathoBacTyper on their own server that allows new species addition and the SGAdb rebuilding. However, a virtual machine deployment and manual operation of species addition are inconvenient. Therefore, a standalone GUI-based tool, such as java application, with automatic new species addition and SGAdb rebuilding is needed and will be implemented in our next version of PathoBacTyper.

CONCLUSION

The PathoBacTyper web server comprising two functions, species identification and bacterial strain typing, was established for researchers to perform epidemiological typing and clinical investigation for more than 400 pathogenic bacterial organisms. With NGS becoming a routine approach in clinical and public health laboratories, a research use only analysis platform for handling hundreds of thousands of WGS data is crucial. Such molecular genotyping will provide an information about strain-relatedness among outbreaks to support the infection control by field study. Through the PathoBacTyper service, users can directly upload WGS raw reads or assemblies of bacterial isolates for performing species identification and strain typing without expertise in bioinformatics. We believe that PathoBacTyper is a very powerful online tool for disease outbreak investigation and surveillance.

AUTHOR CONTRIBUTIONS

M-HT and Y-YL conceived and designed the experiments. M-HT and Y-YL performed the experiments and analyzed the data. M-HT and Y-YL contributed materials/analysis tools. M-HT, Y-YL, and V-WS wrote the paper.

FUNDING

This study was supported by grant from Ministry of Science and Technology (MOST-103-2221-E-007-090-), Taiwan.

ACKNOWLEDGMENT

We are grateful to NHRI, for providing the hardware and software resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01474/full#supplementary-material

[3] http://halst.nhri.org.tw/PathoBacTyper/download.jsp

REFERENCES

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* 62, 53–70. doi: 10.1146/annurev.micro.62.081307.162832

Acke, F. R., Malfait, F., Vanakker, O. M., Steyaert, W., De Leeneer, K., Mortier, G., et al. (2014). Novel pathogenic COL11A1/COL11A2 variants in Stickler syndrome detected by targeted NGS and exome sequencing. *Mol. Genet. Metab.* 113, 230–235. doi: 10.1016/j.ymgme.2014.09.001

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* 41, 1061–1067. doi: 10.1038/ng.437

Bakker, H. C., Switt, A. I., Cummings, C. A., Hoelzer, K., Degoricija, L., Rodriguez-Rivera, L. D., et al. (2011). A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common *Salmonella enterica* subsp. *enterica* serovar montevideo pulsed-field gel electrophoresis type. *Appl. Environ. Microbiol.* 77, 8648–8655. doi: 10.1128/AEM.06538-11

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

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., et al. (2015). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput. Sci.* 1:e20. doi: 10.7717/peerj-cs.20

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

Franz, E., Delaquis, P., Morabito, S., Beutin, L., Gobius, K., Rasko, D. A., et al. (2014). Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. *Int. J. Food Microbiol.* 187, 57–72. doi: 10.1016/j.ijfoodmicro.2014.07.002

Gordon, N. C., Price, J. R., Cole, K., Everitt, R., Morgan, M., Finney, J., et al. (2014). Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *J. Clin. Microbiol.* 52, 1182–1191. doi: 10.1128/JCM.03117-13

Halachev, M. R., Chan, J. Z., Constantinidou, C. I., Cumley, N., Bradley, C., Smith-Banks, M., et al. (2014). Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant *Acinetobacter baumannii* in Birmingham, England. *Genome Med.* 6:70. doi: 10.1186/s13073-014-0070-x

Hoffmann, M., Zhao, S., Pettengill, J., Luo, Y., Monday, S. R., Abbott, J., et al. (2014). Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype heidelberg isolates from humans, retail meats, and animals. *Genome Biol. Evol.* 6, 1046–1068. doi: 10.1093/gbe/evu079

Holt, K. E., Baker, S., Dongol, S., Basnyat, B., Adhikari, N., Thorson, S., et al. (2010). High-throughput bacterial SNP typing identifies distinct clusters of *Salmonella* Typhi causing typhoid in Nepalese children. *BMC Infect. Dis.* 10:144. doi: 10.1186/1471-2334-10-144

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

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., et al. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* 52, 1501–1510. doi: 10.1128/JCM.03617-13

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., et al. (2017). A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Front. Microbiol.* 8:375. doi: 10.3389/fmicb.2017.00375

Koser, C. U., Ellington, M. J., and Peacock, S. J. (2014). Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.* 30, 401–407. doi: 10.1016/j.tig.2014.07.003

Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O., and Aarestrup, F. M. (2014). Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLoS ONE* 9:e87991. doi: 10.1371/journal.pone.0087991

Luo, T., Yang, C., Peng, Y., Lu, L., Sun, G., Wu, J., et al. (2014). Whole-genome sequencing to detect recent transmission of *Mycobacterium tuberculosis* in settings with a high burden of tuberculosis. *Tuberculosis* 94, 434–440. doi: 10.1016/j.tube.2014.04.005

Meinel, D. M., Margos, G., Konrad, R., Krebs, S., Blum, H., and Sing, A. (2014). Next generation sequencing analysis of nine *Corynebacterium ulcerans* isolates reveals zoonotic transmission and a novel putative diphtheria toxin-encoding pathogenicity island. *Genome Med.* 6:113. doi: 10.1186/s13073-014-0113-3

Merker, M., Kohl, T. A., Roetzer, A., Truebe, L., Richter, E., Rusch-Gerdes, S., et al. (2013). Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. *PLoS ONE* 8:e82551. doi: 10.1371/journal.pone.0082551

Moran-Gilad, J., Prior, K., Yakunin, E., Harrison, T. G., Underwood, A., Lazarovitch, T., et al. (2015). Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Euro Surveill.* 20:21186. doi: 10.2807/1560-7917.ES2015.20.28.21186

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Octavia, S., Wang, Q., Tanaka, M. M., Kaur, S., Sintchenko, V., and Lan, R. (2015). Delineating community outbreaks of *Salmonella enterica* serovar Typhimurium by use of whole-genome sequencing: insights into genomic variability within an outbreak. *J. Clin. Microbiol.* 53, 1063–1071. doi: 10.1128/JCM.03235-14

Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., et al. (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* 413, 848–852. doi: 10.1038/35101607

Petersen, R. F., Litrup, E., Larsson, J. T., Torpdahl, M., Sørensen, G., Müller, L., et al. (2011). Molecular characterization of *Salmonella* Typhimurium highly successful outbreak strains. *Foodborne Pathog. Dis.* 8, 655–661. doi: 10.1089/fpd.2010.0683

Petkau, A., Mabon, P., Sieffert, C., Knox, N., Cabral, J., Iskander, M., et al. (2016). SNVPhyl: A Single Nucleotide Variant Phylogenomics Pipeline for Microbial Genomic Epidemiology. bioRxiv

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Schmid, D., Allerberger, F., Huhulescu, S., Pietzka, A., Amar, C., and Kleta, S. (2014). Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-2013. *Clin. Microbiol. Infect.* 20, 431–436. doi: 10.1111/1469-0691.12638

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13, 435–438. doi: 10.1038/nmeth.3802

Struelens, M. J., and Brisse, S. (2013). From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. *Euro. Surveill.* 18:20386.

Taylor, A. J., Lappi, V., Wolfgang, W. J., Lapierre, P., Palumbo, M. J., Medus, C., et al. (2015). Characterization of foodborne outbreaks of *Salmonella enterica* serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. *J. Clin. Microbiol.* 53, 3334–3340. doi: 10.1128/JCM.01280-15

Torpdahl, M., Sørensen, G., Lindstedt, B. A., and Nielsen, E. M. (2007). Tandem repeat analysis for surveillance of human *Salmonella* Typhimurium infections. *Emerg. Infect. Dis* 13, 388–395. doi: 10.3201/eid1303.060460

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 Biol.* 15:524. doi: 10.1186/s13059-014-0524-x

Weese, D., Emde, A. K., Rausch, T., Doring, A., and Reinert, K. (2009). RazerS–fast read mapping with sensitivity control. *Genome Res.* 19, 1646–1654. doi: 10.1101/gr.088823.108

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.

Copyright © 2017 Tsai, Liu and Soo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.