Journal of Microbiology, Immunology and Infection 55 (2022) 102—106

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jmii.com

Original Article

cgMLST@Taiwan: A web service platform for *Vibrio cholerae* cgMLST profiling and global strain tracking

Yi-Syong Chen [1], Yueh-Hua Tu [1], Bo-Han Chen, Yen-Yi Liu, Yu-Ping Hong, Ru-Hsiou Teng, You-Wun Wang, Chien-Shun Chiou*

Center for Diagnostics and Vaccine Development, Centers for Disease Control, Ministry of Health and Welfare, Taiwan

Received 26 August 2020; received in revised form 28 December 2020; accepted 30 December 2020
Available online 15 January 2021

KEYWORDS
Vibrio cholerae;
Molecular typing;
Whole-genome sequencing;
Web service;
Core genome multilocus sequence typing (cgMLST);
Strain tracking

Abstract *Background:* Cholera, a rapidly dehydrating diarrheal disease caused by toxigenic *Vibrio cholerae*, is a leading cause of morbidity and mortality in some regions of the world. Core genome multilocus sequence typing (cgMLST) is a promising approach in generating genetic fingerprints from whole-genome sequencing (WGS) data for strain comparison among laboratories.
Methods: We constructed a *V. cholerae* core gene allele database using an in-house developed computational pipeline, a database with cgMLST profiles converted from genomic sequences from the National Center for Biotechnology Information, and built a REST-based web accessible via the Internet.
Results: We built a web service platform—cgMLST@Taiwan and installed a *V. cholerae* allele database, a cgMLST profile database, and computational tools for generating *V. cholerae* cgMLST profiles (based on 3,017 core genes), performing rapid global strain tracking, and clustering analysis of cgMLST profiles. This web-based platform provides services to researchers, public health microbiologists, and physicians who use WGS data for the investigation of cholera outbreaks and tracking of *V. cholerae* strain transmission across countries and geographic regions. The cgMLST@Taiwan is accessible at http://rdvd.cdc.gov.tw/cgMLST.

* Corresponding author. Central Region Laboratory, Center for Diagnostics and Vaccine Development, Centers for Disease Control, Ministry of Health and Welfare, No. 20 Wen-Sin South 3rd Road, Taichung, 40855, Taiwan. Fax: +886 4 24750474.
E-mail address: nipmcsc@cdc.gov.tw (C.-S. Chiou).
[1] YC Chen and YH Tu contributed equally to this work.

https://doi.org/10.1016/j.jmii.2020.12.007

Introduction

Vibrio cholerae are aquatic bacteria that live in brackish water and freshwater environments. Some strains of *V. cholerae* cause acute diarrheal disease, cholera, in humans. Only cholera toxin-producing strains of serogroups O1 and O139 have ever caused widespread epidemics of cholera.[1] In the past 200 years, seven cholera pandemics have occurred; the seventh pandemic originated in Indonesia in 1961 and is still ongoing.[2] Currently, cholera is still a severe public health threat in some regions of the world. In 2017, 34 countries, mostly in Africa, Asia, and the Americas, reported a total of 1,227,391 cases and 5,654 deaths to the World Health Organization.[3]

Many typing methods have been developed for assessing genetic relatedness among *V. cholerae* strains for an epidemiological study.[4] With the advancement of next-generation sequencing techniques, whole-genome sequencing (WGS) of bacterial isolates has become affordable and has been increasingly applied to characterize bacterial strains in evolutionary studies, disease outbreak investigations, and active disease surveillance as well as to track the global transmission of bacterial strains and clones.[5—12] Whole-genome single-nucleotide polymorphism (wgSNP) and core genome multilocus sequence typing (cgMLST) are the two most frequently used approaches to apply WGS data to generate genetic profiles for comparison of bacterial strains. cgMLST is a gene-by-gene comparison approach comprising the allelic profile of the core genes of a bacterial species.[13] cgMLST is easier than wgSNP to standardize for generating genetic profiles comparable among laboratories. Accordingly, cgMLST is considered an advisable approach for public health laboratories and academic institutes to use for analyzing WGS data for local, regional, and global disease surveillance networks and rapid strain tracking.[14] Recently, a *V. cholerae* cgMLST scheme has been developed and is publicly available on PubMLST (https://pubmlst.org/vcholerae/).[15]

In the present study, we built a web service platform and installed a *V. cholerae* allele database, a cgMLST profile database, and computational tools that allow users to upload whole-genomic sequences via the Internet to obtain cgMLST profiles and to conduct rapid genetic profile comparisons with the strains in the National Center for Biotechnology Information (NCBI).

Methods

Construction of web: cgMLST@Taiwan

The web was developed using the Django RESTful framework for its back end and React.js for its front end. All the functionality of the website is provided with representational state transfer (RESTful) API. The website provides several tools, including cgMLST profiling, strain tracking against a cgMLST profile database, and clustering analysis of cgMLST profiles.

Identification of core genes

We firstly used 1,056 genomes of O1 and non-O1 *V. cholerae* obtained from the NCBI database to construct a *V. cholerae* pan-genome allele database using the PGAdb-builder.[16] The genomes were first confirmed to be *V. cholerae* using the KmerFinder 3.1 (https://cge.cbs.dtu.dk/services/KmerFinder/). Loci of the pan-genome allele database with a frequency of ≧95% (present in ≧95% of the 1,056 genomes) were designated to be the core genes for *V. cholerae*. The most abundant amino acid sequences (the mode) for each of the core genes are selected as the reference sequence set for cgMLST profiling.

cgMLST profiling

We installed an in-house developed BENGA cgMLST profiling tool on the web. The tool accepts assembled sequences (contigs) files in a fasta format to perform cgMLST profiling. We suggest that users assemble reads using the latest version of SPAdes assembler.[17] Briefly, to profiling, each assembled genomic sequence is annotated using the Prodigal program Version 2.6.3 to identify open reading frames (ORFs),[18] and the corresponding nucleotide sequences of the ORFs are converted into SHA256 codes.[19] Codes are compared with those in the allele database, and the matched codes are assigned for the corresponding genes. For the remaining unmatched codes, the corresponding amino acid sequences of the codes are blasted with the reference sequence set of the core genes. As a query sequence shares ≧95% amino acid sequence identity and ≧75% coverage with the reference sequence of a locus, the corresponding code is added as a new allele to the locus. New codes are simultaneously added to the cgMLST profile of the query genome. A cgMLST profile is, therefore, output in a tsv file that comprises an array of SHA256 codes.

cgMLST database and strain tracking

We downloaded *V. cholerae* genomes from the Assembly and SRA databases of the NCBI. The genomes, which were first confirmed to be *V. cholerae* using the KmerFinder 3.1 (https://cge.cbs.dtu.dk/services/KmerFinder/) and had total reads ≧30X coverage, were selected for further analysis. Reads were assembled using the SPAdes assembler version 3.12.0, and cgMLST profiles were generated from assembled contigs using the BENGA cgMLST profiling tool. Plasmid types (Inc types), ST types, and antimicrobial resistance genes were identified from genomic sequences using the PlasmidFinder 2.1 (https://cge.cbs.dtu.dk/services/PlasmidFinder/), the MLST2.0 (https://cge.cbs.dtu.dk/services/MLST/), and the ResFinder 4.0 (https://cge.cbs.dtu.dk/services/ResFinder/). We provided the information, when available, including the source of genomic sequences (SRA or Assembly accession number), identifier (strain ID/alias), year of isolation, source of country, serogroup and serotype, ST type, plasmid types, resistance genes, and the number of void core genes, for each entry of the database. For strain tracking, the web allows users to

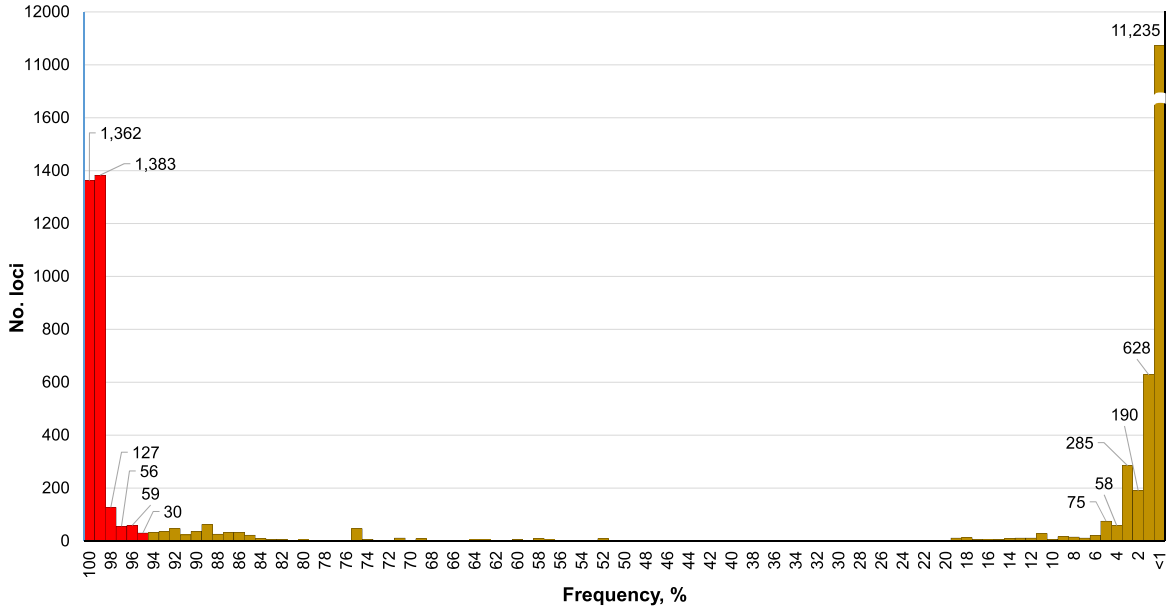

Figure 1. Frequency of loci (genes) over 1,056 *V. cholerae* genomes.

compare a cgMLST file per job and displays the 100 best-matches.

Clustering tool

We installed a tool for the clustering of cgMLST profiles, which are generated using the BENGA profiling tool and made up with SHA256 codes. Single linkage and unweighted pair group method with arithmetic mean (UPGMA) algorithms were provided for constructing a cgMLST tree. The web allows users to upload up to 500 cgMLST profiles per job for clustering analysis. A cgMLST tree (dendrogram) is output in png, pdf, svg, and Newick formats by option.

Results and discussion

The web

cgMLST@Taiwan is a user-friendly and straightforward browser-based platform for cgMLST profiling and strain tracking. Besides, it provides a tool for clustering analysis of cgMLST profiles that are generated using the BENGA profiling tool. The website takes a single page application (SPA) design as well as fool-proof design to avoid misleading and abuse. The website allows users to upload a maximum of 100 assembled genomic sequences in fasta format per job for cgMLST profiling, to upload a cgMLST profile for strain tracking, and to upload up to 500 cgMLST profiles for the construction of a cgMLST tree. The service is available at http://rdvd.cdc.gov.tw/cgMLST. The web can be browsed by only using Google Chrome™ and Firefox® browser.

Core genes

We identified 16,146 genes from 1,056 genomes. Of the genes, 76.7% existed in ≦5% of the genomes, and 18.7% in ≧95% of the genomes. A total of 3,017 loci with a frequency of ≧95% over the 1,056 genomes were designated as the core genes for *V. cholerae* (Fig. 1).

cgMLST profiling

The web allows users to upload WGS data in fasta format for cgMLST profiling and outputs a compressed file with an array of SHA256 codes for each of the uploaded genomes. When running profiling, the BENGA profiling program simultaneously adds new codes (alleles) to the corresponding loci and uses the new alleles for the cgMLST profile of the genome. Because new alleles (codes) are simultaneously added into the allele database and used in the progressing profiling process, an identical cgMLST profile can be generated for a genome at any time. And, because a nucleotide sequence is converted to a one-and-only SHA256 code, an identical cgMLST profile can be generated for a genome in different laboratories in which a common set of reference sequences for the core genes and the BENGA profiling tool are adopted. Therefore, cgMLST profiles generated in different laboratories can be entirely comparable. Our profiling system is superior to the system that uses the same allele calling step as that has been used for the 7-genes MLST.[20,21] Using the allele calling step, a common allele database with a large number of alleles has to be used to yield a comparable cgMLST profile. Since the new allele can not be simultaneously added in the allele database and used in the progressing profiling, profiles

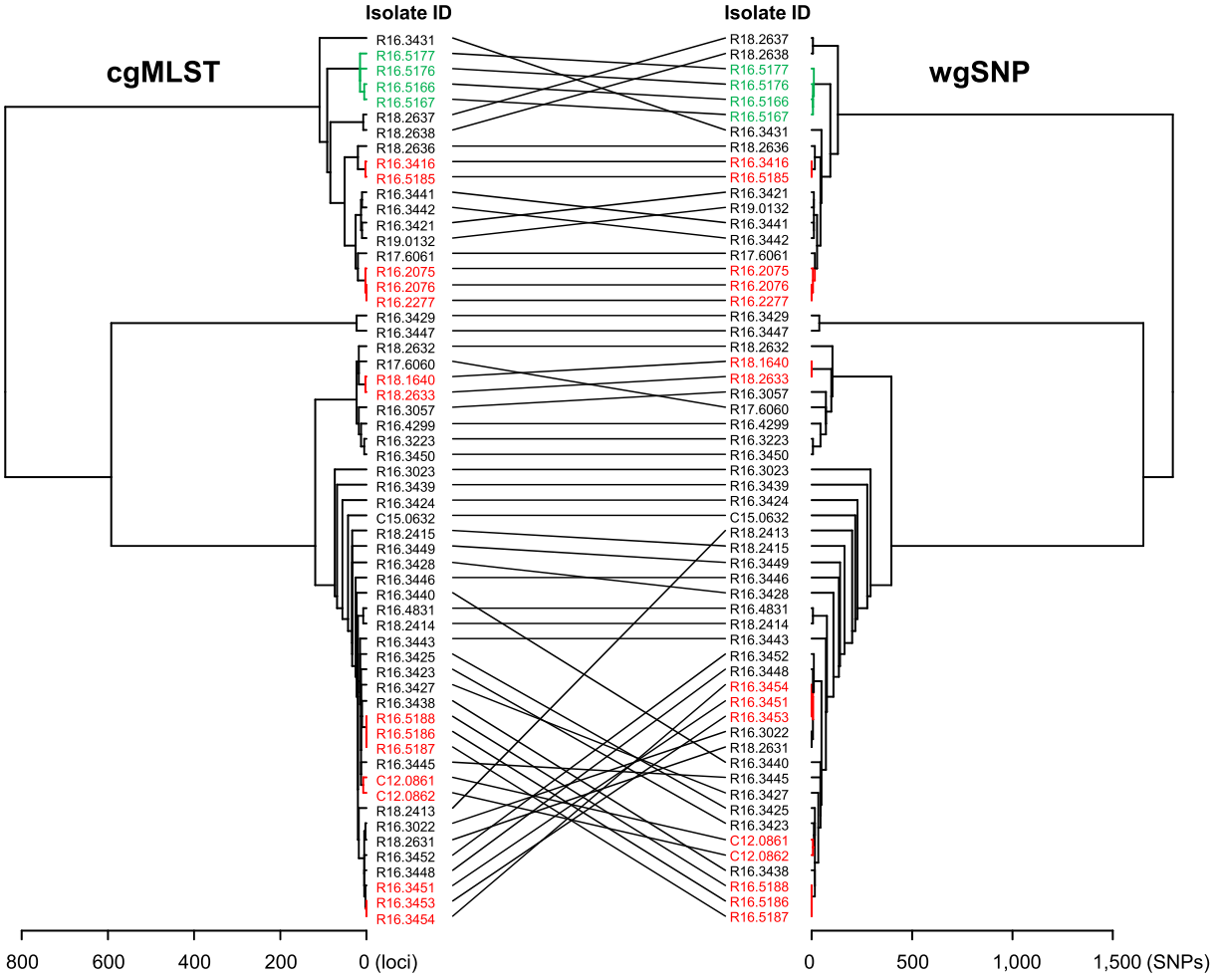

Figure 2. Comparison of phylogenetic trees constructed using the cgMLST profiles and wgSNP profiles for 58 *V. cholerae* isolates from cholera cases in Taiwan. cgMLST profiles were generated using the 3,017 core genes cgMLST scheme developed in this study and the wgSNP profiles were called using the option of strict SNP filtering tool provided in the BioNumerics version 7.6.3. The trees were constructed using the UPGMA algorithm. The 4 isolates from the 1962 cholera outbreak are marked in green; the isolates from each of 6 epidemiologically-linked clusters are marked in red.

generated for a genome using different versions of an allele database could be not completely identical. Using our BENGA profiling system, an identical profile can be generated for a genome in any laboratory at any time.

cgMLST database and strain tracking

The database installed on the web contains 6,270 cgMLST profiles, which were generated from the genomes deposited in the NCBI database by August 10, 2020. Strain tracking is performed by uploading a cgMLST profile to compare it with those in the database. Up to 100 best-matched profiles are displayed for each job, and the 100 profiles and the relevant information can be downloaded.

We constructed a UPGMA genetic tree with the 6,270 cgMLST profiles to show a global genetic structure of *V. cholerae* strains (the tree can be found on the web: About). The cgMLST tree shows that ST69, ST75, and ST73 are the most abundant O1 clones. Non-O1, O139 *V. cholerae* strains are highly diverse; they fall into many genetically distant ST clones. ST69 clone comprises serogroup O1 and O139 *V. cholerae* strains, supporting the conclusion of a study that serogroup O139 strains, which caused the first massive epidemic of cholera in India and Bangladesh in 1992, are derived from the seventh pandemic O1 clone.[22]

Discrimination of closely related strains

A cgMLST profile comprises an allele array of as many as 3,017 core genes; thus, this cgMLST typing scheme should possess an excellent discriminatory power in distinguishing among closely related strains. To demonstrate the usefulness of the cgMLST method in discerning closely related *V. cholerae* strains, we analyzed a panel of 58 *V. cholerae* isolates, of which 54 were collected from cholera patients in Taiwan from 2003 to 2008 and 4 from the 1962 cholera outbreak that occurred in Taiwan,[23,24] using our cgMLST and a wgSNP approaches. The genetic relatedness among the *V. cholerae* isolates established with cgMLST, and wgSNP profiles were highly concordant. Both approaches could clearly distinguish among isolates from 6 epidemiologically-related clusters (Fig. 2).

In conclusion, our web-based service platform allows users to generate cgMLST profiles and rapidly compare the genetic relatedness between a query strain and those in the NCBI database. Using our reference sequence set of core genes, the BENGA profiling tool can generate the same cgMLST profile for a genome in different laboratories at any time. Therefore, using this system, each laboratory can construct its allele database for cgMLST profiling, and the profiles are fully comparable among laboratories. With the BENGA profiling system, public health laboratories can easily standardize a cgMLST protocol and implement a WGS-based genotyping method as a standard molecular subtyping tool for cross-border disease surveillance. This web platform will benefit researchers, public health microbiologists, and physicians who seek to use WGS data for cholera outbreak investigations and global tracking of transmission of *V. cholerae* strains.

Author contributions

Y-HT and Y-SC contributed equally to the construction of the webserver and the development of tools. B-HC, Y-YL, Y-PH, R-HT, and Y-WW analyzed genomic sequences, tested and validated the webserver, and the tools. C-SC designed the study, analyzed and interpreted the data, and drafted the manuscript. All authors approved the final version.

Funding

This study was supported by the Ministry of Health and Welfare, Taiwan with Grant No. MOHW108-CDC-C-315-134516.

References

1. Centers for Disease Control and Prevention. *CDC Yellow Book 2020: health information for international travel*. New York: Oxford University Press; 2017.
2. Morris Jr JG. Cholera-modern pandemic disease of ancient lineage. *Emerg Infect Dis* 2011;**17**:2099—104.
3. World health organization. Cholera Annual Report 2017. *Wkly Epidemiol Rec* 2018;**93**:489—500.
4. Rahaman MH, Islam T, Colwell RR, Alam M. Molecular tools in understanding the evolution of *Vibrio cholerae*. *Front Microbiol* 2015;**6**:1040.
5. Weill FX, Domman D, Njamkepo E, Almesbahi AA, Naji M, Nasher SS, et al. Genomic insights into the 2016-2017 cholera epidemic in Yemen. *Nature* 2019;**565**:230—3.
6. Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, et al. Genomic history of the seventh pandemic of cholera in Africa. *Science* 2017;**358**:785.
7. Kuleshov KV, Vodop'ianov SO, Dedkov VG, Markelov ML, Deviatkin AA, Kruglikov VD, et al. Travel-associated *Vibrio cholerae* O1 El Tor, Russia. *Emerg Infect Dis* 2016;**22**:2006—8.
8. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, et al. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *mBio* 2011;**2**:e00157. 11.
9. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 2011;**477**:462—5.
10. Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, et al. 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* 2015;**53**:3334—40.
11. Liao YS, Liu YY, Lo YC, Chiou CS. Azithromycin-nonsusceptible *Shigella flexneri* 3a in men who have sex with men, Taiwan, 2015-2016. *Emerg Infect Dis* 2017;**23**:345—6.
12. Chiou CS, Izumiya H, Kawamura M, Liao YS, Su YS, Wu HH, et al. The worldwide spread of ciprofloxacin-resistant *Shigella sonnei* among HIV-infected men who have sex with men, Taiwan. *Clin Microbiol Infect* 2016;**22**:383 e11—6.
13. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* 2013;**11**:728—36.
14. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill* 2017;**22**.
15. Liang KYH, Orata FD, Islam MT, Nasreen T, Alam M, Tarr CL, et al. A *Vibrio cholerae* core genome multilocus sequence typing scheme to facilitate the epidemiological study of cholera. *J Bacteriol* 2020;**202**:e00086-20.
16. Liu YY, Chiou CS, Chen CC. PGAdb-builder: a web service tool for creating pan-genome allele database for molecular fine typing. *Sci Rep* 2016;**6**:36213.
17. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455—77.
18. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf* 2010;**11**:119.
19. Lamberger M, Mendel F. Higher-order differential attack on reduced SHA-256. *IACR Cryptol Print Arch* 2011;**2011**:37.
20. Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, et al. Typing methods based on whole genome sequencing data. *One Health Outlook* 2020;**2**:3.
21. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 1998;**95**:3140—5.
22. Faruque SM, Sack DA, Sack RB, Colwell RR, Takeda Y, Nair GB. Emergence and evolution of *Vibrio cholerae* O139. *Proc Natl Acad Sci U S A* 2003;**100**:1304—9.
23. Tu YH, Chen BH, Hong YP, Liao YS, Chen YS, Liu YY, et al. Emergence of *Vibrio cholerae* O1 sequence type 75 in Taiwan. *Emerg Infect Dis* 2020;**26**:164—6.
24. Yen CH. A recent study of cholera with reference to an outbreak in Taiwan in 1962. *Bull World Health Organ* 1964;**30**:811—25.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmii.2020.12.007.