





A *Vibrio cholerae* Core Genome Multilocus Sequence Typing Scheme To Facilitate the Epidemiological Study of Cholera

Kevin Y. H. Liang,^a  Fabini D. Orata,^a Mohammad Tarequl Islam,^a Tania Nasreen,^a Munirul Alam,^b Cheryl L. Tarr,^c  Yann F. Boucher^{a,d,e}

^aDepartment of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

^bInfectious Diseases Division, International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh

^cEnteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

^dSaw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

^eSingapore Center for Environmental Life Sciences Engineering, National University of Singapore, Singapore, Singapore

ABSTRACT Core genome multilocus sequence typing (cgMLST) has gained popularity in recent years in epidemiological research and subspecies-level classification. cgMLST retains the intuitive nature of traditional MLST but offers much greater resolution by utilizing significantly larger portions of the genome. Here, we introduce a cgMLST scheme for *Vibrio cholerae*, a bacterium abundant in marine and freshwater environments and the etiologic agent of cholera. A set of 2,443 core genes ubiquitous in *V. cholerae* were used to analyze a comprehensive data set of 1,262 clinical and environmental strains collected from 52 countries, including 65 newly sequenced genomes in this study. We established a sublineage threshold based on 133 allelic differences that creates clusters nearly identical to traditional MLST types, providing backwards compatibility to new cgMLST classifications. We also defined an outbreak threshold based on seven allelic differences that is capable of identifying strains from the same outbreak and closely related isolates that could give clues on outbreak origin. Using cgMLST, we confirmed the South Asian origin of modern epidemics and identified clustering affinity among sublineages of environmental isolates from the same geographic origin. Advantages of this method are highlighted by direct comparison with existing classification methods, such as MLST and single-nucleotide polymorphism-based methods. cgMLST outperforms all existing methods in terms of resolution, standardization, and ease of use. We anticipate this scheme will serve as a basis for a universally applicable and standardized classification system for *V. cholerae* research and epidemiological surveillance in the future. This cgMLST scheme is publicly available on PubMLST (<https://pubmlst.org/vcholerae/>).

IMPORTANCE Toxigenic *Vibrio cholerae* isolates of the O1 and O139 serogroups are the causative agents of cholera, an acute diarrheal disease that plagued the world for centuries, if not millennia. Here, we introduce a core genome multilocus sequence typing scheme for *V. cholerae*. Using this scheme, we have standardized the definition for subspecies-level classification, facilitating global collaboration in the surveillance of *V. cholerae*. In addition, this typing scheme allows for quick identification of outbreak-related isolates that can guide subsequent analyses, serving as an important first step in epidemiological research. This scheme is also easily scalable to analyze thousands of isolates at various levels of resolution, making it an invaluable tool for large-scale ecological and evolutionary analyses.

KEYWORDS *Vibrio cholerae*, whole-genome sequencing, cgMLST, cholera, core genome, epidemiological surveillance, gene-by-gene approach, multilocus sequence typing

Citation Liang KYH, Orata FD, Islam MT, Nasreen T, Alam M, Tarr CL, Boucher YF. 2020. A *Vibrio cholerae* core genome multilocus sequence typing scheme to facilitate the epidemiological study of cholera. *J Bacteriol* 202:e00086-20. <https://doi.org/10.1128/JB.00086-20>.

Editor Yves V. Brun, University of Montreal

Copyright © 2020 American Society for Microbiology. All Rights Reserved.

Address correspondence to Yann F. Boucher, yboucher@ualberta.ca.

Received 18 February 2020

Accepted 7 June 2020

Accepted manuscript posted online 15 June 2020

Published 19 November 2020

Cholera is transmitted in a fecal-oral route mostly by contaminated food or water (1, 2). The case fatality rate (CFR) of this disease can be up to 50% without treatment, but with proper medical care, the CFR is usually less than 1% (2, 3). In developed countries with proper water treatment facilities, cholera is practically nonexistent aside from imported cases. Unfortunately, this cannot be said for many developing countries lacking this infrastructure, where cholera has been endemic for centuries, such as in parts of South Asia (4). As it is also difficult to eradicate cholera (5), this disease often becomes endemic in regions where it has been introduced, for example, in Latin America in 1991 (6, 7), Haiti in 2010 (8), and Yemen in 2016 (9). It is estimated that there are over a million cholera cases each year, resulting in tens of thousands of deaths worldwide (10). Being an indicator of health care and socioeconomic disparities (11, 12), this disease is often underreported due to its negative influence on tourism, as it implies poor water quality (13). Together with the lack of a universally applicable and standardized classification method, outbreak surveillance and source attribution is often challenging (1, 8). The Haiti outbreak, for example, due to these limitations, required extensive genomic and epidemiological research from the beginning of the outbreak to determine the source of introduction, which was not confirmed until August 2011, even though cholera broke out in July 2010 (8, 14–17).

A typing method for use in global surveillance of pandemic-causing pathogens, such as *V. cholerae*, should be efficient and easy to use, with the potential to be applied to all *V. cholerae* strains around the world. Therefore, it must have the capacity to analyze thousands of genomes efficiently, and new genomes should be easily typed as they get sequenced. As all cholera outbreaks are caused by a single lineage of *V. cholerae*, the pandemic-generating/phylocore genome (PG) lineage, which includes the 7th pandemic El Tor, El Tor sister, El Tor progenitor, classical, and classical sister clades (5, 18, 19), this method should also be able to differentiate isolates at a fine scale and separate outbreaks caused by genetically similar strains. Such a method will help create a comprehensive database with detailed epidemiological data that will allow for the analysis of future outbreak strains in a global context and guide subsequent epidemiological analyses. Different methods for subspecies-level classification and outbreak surveillance have been developed for *V. cholerae*. These methods include serotyping, multilocus sequence typing (MLST) (20, 21), multilocus variable number of tandem repeats (VNTR) analysis (MLVA) (22, 23), and single-nucleotide polymorphism (SNP)-based approaches (14). Despite the popularity of these methods, there are important limitations to each.

Serotyping based on the presence of cell surface O antigens is one of the earliest attempts at subspecies-level classification of *V. cholerae*. There are now over 200 serogroups of *V. cholerae* identified; however, only the toxigenic members of the O1 and O139 serogroups have been found to be responsible for all major documented epidemics and pandemics (24, 25). Serogroup O1 can be further divided into two biotypes (El Tor and classical) and three serotypes (Inaba, Hikojima, and Ogawa) (2). The lack of resolution within the epidemic strains and the possibility of serogroup conversion (26) limit the use of serotyping in epidemiological studies.

MLST provides a standardized classification method that is usually based on a collection of six or seven well-defined housekeeping genes (27). MLST was used to study a number of cholera outbreaks and allowed the descriptions of general population structure (28, 29). It is reproducible and provides reliable results; however, it is unable to differentiate between closely related strains, which limits its use in outbreak surveillance (30, 31). In addition, direct comparisons between different MLST schemes are difficult, as different schemes utilize different housekeeping genes.

MLVA utilizes VNTR regions, which are under less selective pressure than housekeeping genes. Therefore, this method provides greater resolution than MLST for some bacterial species (32, 33). However, due to their rapid mutation rate, VNTR regions are more affected by homoplasmy, where two isolates may share the same MLVA profile due to convergent mutation and not by vertical descent (34). As a result, MLVA may produce clusters that do not necessarily reflect phylogenetic relationships (35). Two

common PCR-based methods exist for the typing of VNTR regions, but each has significant limitations (36). The first method is multiplex PCR, which can analyze all loci at once, but it is impossible to determine which bands correspond to which loci; therefore, this method only produces a banding pattern for strain identification, which makes it difficult to standardize and communicate results. The second method is the separate amplification of VNTR regions, but determining the number of repeats based on amplicon size information alone is difficult if the difference in size is not large enough. In addition, different types of mutations that do not necessarily change the number of repeats can cause a change in amplicon size. Sequencing is needed to confirm MLVA profiles, but repeat regions increase the chances of sequencing errors (37). Due to these limitations, stringent quality control is required for reliable MLVA (38).

SNP-based analysis is one of the most common whole-genome-based methods currently being used and was applied to various outbreaks (14, 39, 40). It relies on the identification of conserved SNPs in strains of interest using next-generation sequence reads or assembled genomes. The number of SNPs can then be related to the evolutionary distance between isolates. SNP-based analysis provides reliable results with sufficient resolution for epidemiological studies, but it is sensitive to horizontal gene transfer and recombination events, as each event will result in many SNPs being created. The number of SNPs between two strains, therefore, does not necessarily reflect the true phylogenetic relationship. Therefore, SNPs found in recombinogenic regions should be removed, which, depending on the organism of interest, can be anywhere from 30% to 97% of SNPs identified (41, 42). Since recombination and horizontal gene transfer events are common within *V. cholerae* (43–45) and between the species and its close relatives (46, 47), SNP-based methods, although suitable in individual epidemiological studies, will have difficulty serving as a universal classification method for *V. cholerae*.

Core genome MLST (cgMLST), also known as the gene-by-gene approach, overcomes the various limitations of the previously mentioned subtyping methods and was established to serve as a universally applicable standardized typing scheme. Similar to MLST, cgMLST relies on individual gene sequences to differentiate between closely related strains; however, instead of using only six or seven housekeeping genes, cgMLST utilizes hundreds to thousands of core genes, which are genes commonly found in all strains of a species. By utilizing a much larger portion of the genome, cgMLST provides superior resolution compared to traditional MLST schemes. By combining the expandable and standardized classification method that made traditional MLST favorable with the resolution of whole-genome-based methods, cgMLST is becoming more popular in epidemiological and ecological studies (48–54). This method has the added advantage of backwards compatibility with all MLST schemes. This means that it is possible to determine MLST profiles of any isolates based on their cgMLST profiles, since cgMLST would include all housekeeping genes by definition. This allows for a 1:1 mapping of any previously established MLST scheme to the cgMLST scheme, which helps consolidate existing classification information.

Another major benefit of cgMLST is that, much like traditional MLST methods, it is possible to establish different clustering thresholds to define important groups. Clonal complexes are examples of clustering thresholds established by MLST schemes, where each clonal complex corresponds to a cluster of isolates that share, at most, one allelic difference across all seven genes sequenced. Some important clonal complexes were shown to correspond to either groups established by a previous typing method (55) or major outbreak strains (56). However, cgMLST offers even greater flexibility than MLST in this regard, given the number of loci considered. With large clustering thresholds, it is possible to identify lineage level differences to study large-scale patterns and answer broader ecological questions (e.g., examining the entire PG lineage). Furthermore, with smaller clustering thresholds where groups are created based on the sharing of a larger number of alleles, it is possible to identify very closely related strains, which is useful in epidemiological studies. The benefits of defining clustering thresholds through cgMLST have already been demonstrated in other human pathogens, such as *Brucella melitensis*

(52), *Campylobacter jejuni* (51), *Clostridium difficile* (53), *Enterococcus faecium* (50), and *Listeria monocytogenes* (49).

In this study, we introduce a cgMLST scheme for the genome-wide typing of *V. cholerae* and demonstrate its universality and efficacy by applying it to known cholera outbreaks around the world. The advantages of cgMLST are presented by comparing the scheme with previously established classification methods. Additionally, we have produced a 1:1 mapping of the cgMLST scheme against two previously established MLST schemes for *V. cholerae* (20, 21), allowing for the consolidation of existing classification information. The cgMLST scheme, genome sequences used in this study, and relevant epidemiological information are publicly available on PubMLST (<https://pubmlst.org/vcholerae/>), which allows for the automatic annotation and subsequent analyses of hundreds of newly uploaded *V. cholerae* genomes in a global context. This increase in efficiency, standardizability, and resolution compared to current methods make cgMLST the most suitable classification scheme for large-scale *V. cholerae* surveillance. By applying this scheme to our collection of over 1,200 isolates collected around the world, it was possible to establish outbreak and sublineage thresholds that allowed us not only to validate the South Asian origin of many modern epidemics, as proposed in previous studies (5, 57, 58), but also to identify clustering affinity among environmental strains, where isolates from the same sublineage are also likely from the same geographic region. This pattern is not seen in clinical isolates, as human hosts readily carry them over large geographical distances.

RESULTS AND DISCUSSION

A high-resolution typing scheme for pandemic *V. cholerae*. The highest level of resolution of any cgMLST scheme is defined by core genome sequence types (cgSTs), where a unique cgST represents a unique allelic profile. Isolates that belong to the same cgST are expected to be phylogenetically very closely related, as although they may not have the exact genomic sequence, they do have the same sequence for all 2,443 core gene loci used in cgMLST. We identified a total of 1,026 cgSTs from 1,262 genomes collected from 52 countries. Even with our extensive data set, we have yet to sample anywhere close to the total predicted cgST diversity for the global *V. cholerae* population (see Fig. S1 in the supplemental material). All isolates were given at least one cgST designation and up to two MLST sequence type (ST) designations based on two previously established MLST schemes (20, 21) (Table S1). MLST STs are defined based on the unique combination of all loci of a particular MLST scheme, which typically uses six or seven well-defined housekeeping genes. Only 12 STs are exclusively present in the 7th pandemic El Tor lineage identified using traditional MLST (20, 21), whereas 560 cgSTs are uniquely present in this group based on cgMLST (Table S1). As the El Tor lineage is responsible for most cholera outbreaks around the world since the beginning of the 7th pandemic (59), this superior ability to resolve between closely related strains in the 7th pandemic El Tor lineage makes cgMLST more suitable in outbreak surveillance than traditional MLST.

Backwards compatibility with previous subspecies classification methods. Much as cgSTs are important in studying closely related strains typical in outbreaks, they are also important in establishing a standardized nomenclature at a higher level to answer broader ecological questions. As the term “lineage” is commonly used to refer to the PG lineage or the major pandemic-causing clades, such as the El Tor and classical clades in *V. cholerae*, we propose the term “sublineage” to define subspecies of *V. cholerae* using cgMLST.

Pairwise allelic differences calculated between all isolates show three major peaks (Fig. 1A). The first peak ends at 40 allelic differences, and the second peak ends at 133 allelic differences (Fig. 1B). The last peak begins at 2,200 allelic differences (Fig. 1A), which is expected due to mutational saturation (i.e., every single allele in the scheme is different between the two distantly related strains being compared). Both breaks (i.e., 40 and 133 allelic differences) could represent a sublineage delineation. To choose between the two thresholds, the clustering efficiency is measured by calculating the

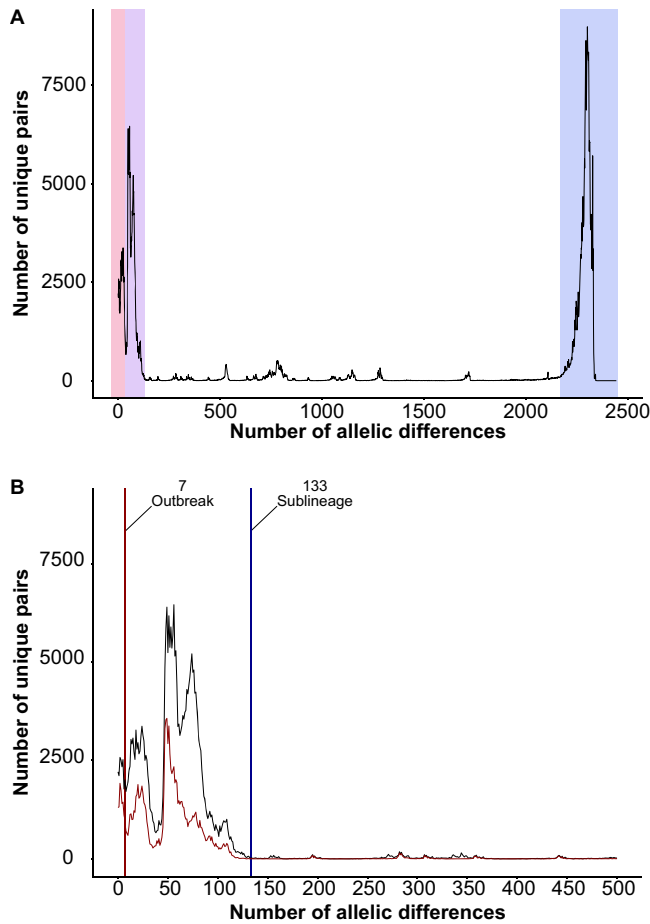


FIG 1 Pairwise allelic differences for all isolates used in this study. Both plots show the frequency of allelic mismatches in pairwise comparisons. (A) Pairwise comparisons of up to 2,443 allelic differences are shown. Major peaks are shaded. (B) Comparisons with up to 500 allelic differences are shown. Pairwise comparisons of only clinical isolates are shown in red. Vertical lines indicate the outbreak threshold (red) and sublineage threshold (blue).

Dunn index (DI) (60) (see Materials and Methods). Since cluster distances are measured by allelic differences, the network with the best clustering efficiency (i.e., the highest DI) will also produce clusters that best represent biological relationships, as isolates are more closely related to themselves than to isolates from other clusters. A DI was calculated for each clustering threshold in the range of 1 to 1,000 allelic differences with 100 bootstrap replicates (Fig. 2). As the clustering threshold defines the maximum number of allelic differences within a cluster, the smaller the threshold, the more closely related the isolates are within a cluster. It is clear that DIs in the range of 0 to 50 allelic differences are significantly lower than the DIs in the range of 100 to 350 allelic differences, with 133 being a clear local maximum. Since 133 allelic differences have the best clustering efficiency and also represents a natural break where most isolate pairs have either a lower or much greater number of allelic differences (Fig. 1B), it was chosen as the sublineage threshold.

Because cgMLST includes all housekeeping genes, information from the two MLST schemes previously developed for *V. cholerae* (20, 21) can now be consolidated with the cgMLST scheme by creating a 1:1 cgMLST to MLST map. To evaluate the similarities between the sublineage threshold and the MLST schemes, we created a minimum spanning tree (MST) for all Bangladesh isolates ($n = 255$), showing only edges with 133 allelic differences or fewer (Fig. 3A and Fig. S2). Therefore, each cluster represents a single sublineage. Bangladesh was chosen to compare cgMLST and MLST, as it is the

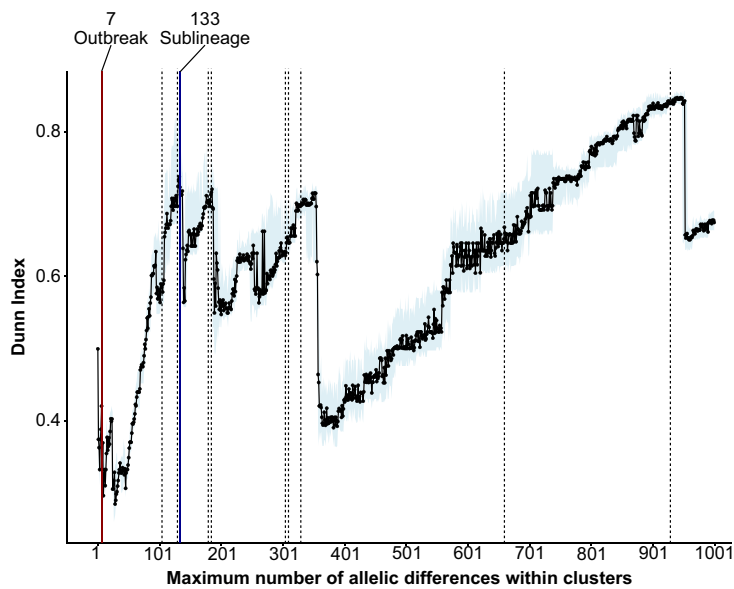


FIG 2 Plot showing the Dunn index for clustering thresholds, ranging from 1 to 1,000 allelic differences. Each clustering threshold is bootstrapped 100 times. The median, plotted with the light blue shade, indicates the 25th to 75th percentile range. Red and blue lines indicate the outbreak and sublineage thresholds, respectively. The dotted lines represent other clustering thresholds used in the adjusted Rand index calculations (Fig. 3B and Fig. S3).

most extensively sampled country in terms of both clinical and environmental isolates in our data set. Using this data set, the chosen sublineage threshold produces clusters that closely resemble ST clustering from traditional MLST. Based on the 2013 MLST scheme (20), each sublineage corresponds to exactly one ST (Fig. S2), whereas there is only one sublineage that contains two STs based on the 2016 MLST scheme (21) (Fig. 3A). All but two isolates belong to ST1; N16961 and A19 belong to ST290, which differs from ST1 at only one of seven MLST loci (Table S2). The reason these two isolates are of a different MLST ST could only be partly explained; they were isolated at an earlier time point (1970s, near the start of the 7th pandemic [61]) than most of the remaining isolates, which were isolated from 1991 onwards (Table S1).

It is impossible to visually evaluate similarities between two MSTs with over 1,200 nodes each simply due to the sheer volume of data. Therefore, the adjusted Rand index (ARI) was used as a metric to determine network similarities (62) (see Materials and Methods). To determine whether the sublineage threshold (i.e., 133 allelic differences) is indeed the best match to traditional MLST schemes, we chose 11 clustering thresholds distributed across the range of 1 to 1,000 allelic differences (Fig. 2) to compare with the MLST schemes. These additional thresholds are chosen because they have a relatively high DI compared to their immediate neighbors. More data points were chosen in the range of 105 to 330 allelic differences, as it was expected that thresholds in this range will best match the traditional MLST schemes. Interestingly, all thresholds in that range had ARIs comparable to those of both the 2016 (21) and the 2013 MLST schemes (20) (Fig. 3B and Fig. S3), indicating that all of them, including the sublineage threshold, produce clusters similar to the MLST scheme. This suggests that there can be a large range of diversity within a single MLST ST, where isolates can have anywhere from 0 (i.e., have the same cgST) to 330 allelic differences. Although clustering thresholds between 105 to 330 allelic differences produce clusters similar to those of a traditional MLST scheme, the sublineage threshold of 133 allelic differences was chosen, as it has the best clustering efficiency (Fig. 2) and represents a natural breakpoint in the currently sampled population (Fig. 1B).

A phylogenetic tree of 1,146 isolates was used to assess the phylogenetic support of the sublineage threshold across different *V. cholerae* strains (Fig. 4). This tree includes

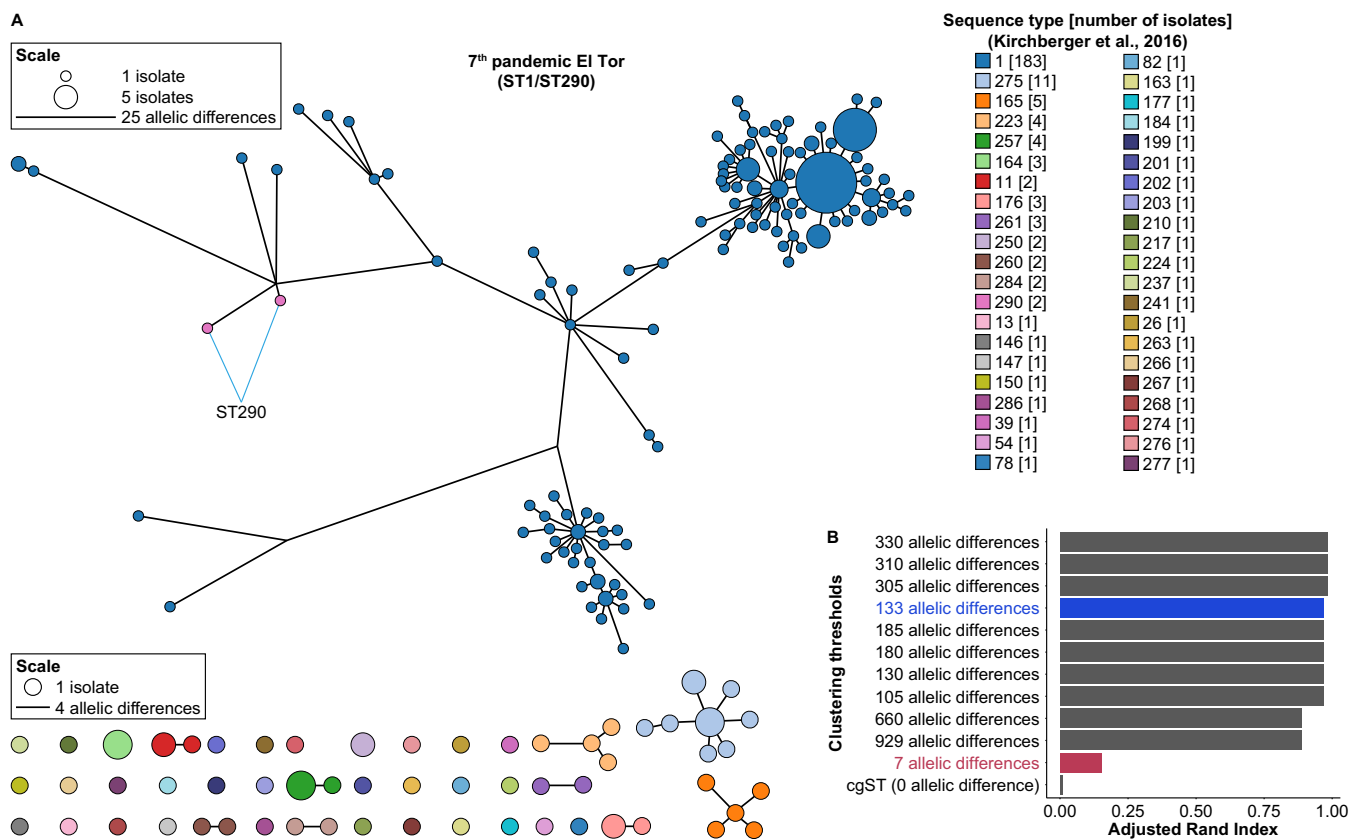


FIG 3 Evaluation of network similarities between the cgMLST sublineage threshold and MLST ST. (A) Networks of all sublineages identified using only *V. cholerae* isolates from Bangladesh ($n = 255$). Each cluster represents a sublineage and includes isolates with less than or equal to 133 allelic differences with each other. Each node represents a cgST and is colored by ST based on the 2016 MLST scheme (21). Sizes of the nodes are proportional to the number of isolates. The length of the connecting lines within a cluster is proportional to the number of allelic differences. (B) Adjusted Rand index for individual pairwise comparisons between predefined clustering thresholds (Fig. 2) and the 2016 MLST scheme (21). The sublineage clustering threshold (i.e., 133 allelic differences) and outbreak threshold (i.e., 7 allelic differences) are indicated in blue and red bars, respectively.

all *V. cholerae* isolates within our data set, with the exception of the recently published 116 clinical isolates from the Yemen cholera outbreak (9), which all belong to the 7th pandemic El Tor lineage. The strains within the PG lineage are closely related with little genetic variation. These lineages are, therefore, collapsed in the phylogenetic tree, as the relationships between them are not well resolved. All sublineages formed monophyletic clades, although in some cases the most basal branch is of a different sublineage (e.g., *V. cholerae* strain T5 or 506315), creating paraphyletic clades. Ideally, each sublineage would correspond to exactly one full monophyletic clade. The reason this is not seen is likely the lack of sampling, leading to the grouping of relatively distantly related isolates together in the same clade. Further sampling will likely resolve these cases into two separate monophyletic clades. Out of 1,262 isolates, we identified 291 sublineages, 19 of which belong exclusively to the PG lineage and 223 are singletons. Much like cgSTs, the rarefaction curve indicates that the total sublineage diversity of *V. cholerae* is far from being sampled (Fig. S1).

The sublineage concepts have been applied to numerous pathogens and, as such, were defined differently depending on the pathogen in question. Some have defined sublineages based on natural breaks in genetic similarities (49), while others may use sublineage to refer specifically to traditional MLST STs (63) or an even finer level of resolution, below the MLST ST level, based on whole-genome analyses (64). There is, however, one unifying feature of all sublineage definitions, i.e., they all refer to monophyletic clades. Sublineages are defined in this study based on natural breaks in allelic differences calculated from cgMLST profiles and were put into context by comparison with two traditional MLST schemes. We have shown that our definition of

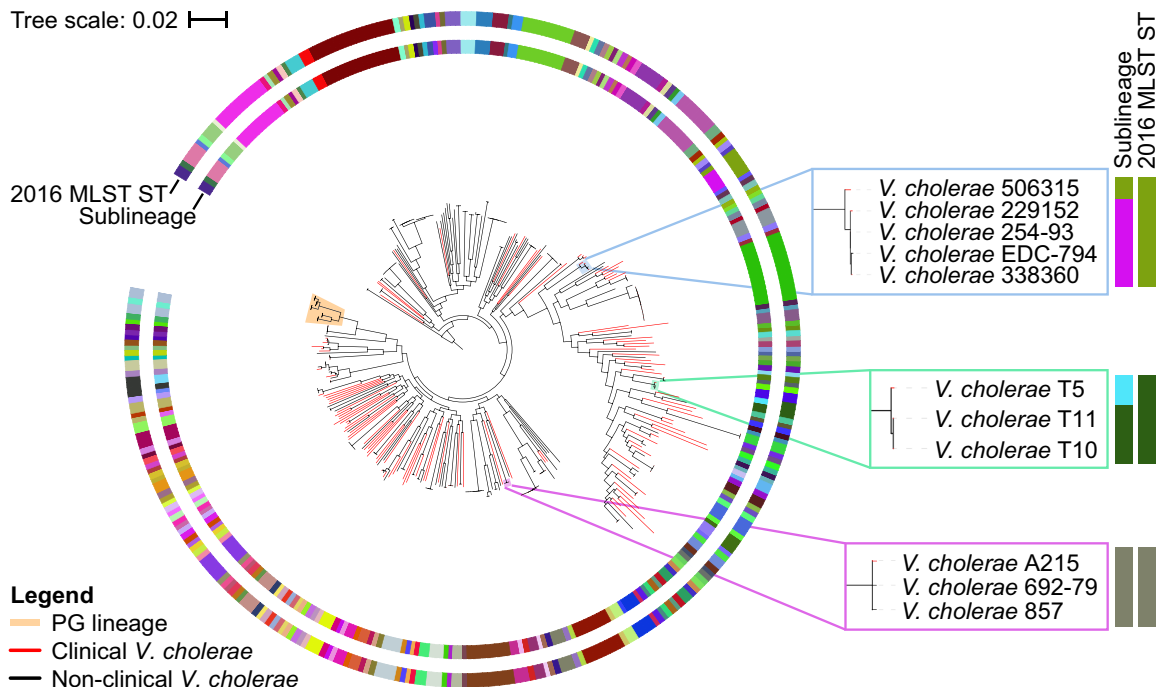


FIG 4 Phylogenetic tree of 1,146 *V. cholerae* isolates (excluding the 116 isolates from the Yemen cholera outbreak) reconstructed using Parsnp v1.2 (97). All groups inside the PG lineage (7th pandemic El Tor, El Tor progenitor, El Tor sister, classical, and classical sister) are collapsed. Outer rings represent clustering by sequence type based on the 2016 MLST scheme by Kirchberger and colleagues (21), whereas the inner ring represents clustering based on the sublineage threshold (i.e., 133 allelic differences). Branches of clinical strains are colored in red. The phylogenetic tree is rooted with a basal lineage to *V. cholerae* (collapsed) (79, 86).

sublineage results in monophyletic clades but also corresponds to any traditional MLST ST designation (Fig. 4). As MLST is one of the primary tools for studying *V. cholerae* and specific MLST STs were previously used to define specific cholera outbreaks (65), the ability to put MLST STs in the context of cgMLST through the use of sublineages will allow for an easier transition and will play a crucial role in consolidating information from all previous MLST analyses.

A universal South Asian origin for modern cholera outbreaks. With the continual improvement of next-generation sequencing techniques, whole-genome sequencing is expected to become a standard practice or even the first identification tool used in clinical and epidemiological studies. Therefore, it is critical to develop a rapid typing scheme for genome sequence data that has the power to inform us about the relationship of a novel isolate with known strains. This is done here by defining what we term an “outbreak threshold” based on cgMLST, which can identify outbreak-related strains and potential sources of introduction. The outbreak threshold is expected to be less than 40 allelic differences, as isolates from the same outbreak are very closely related (9, 14). There is a minor discontinuity at seven allelic differences, where most isolate pairs have either less or more than this number of allelic differences (Fig. 1B). Looking at the DI, the local maximum in the range of 0 to 50 occurs at seven allelic differences as well (Fig. 2), making this cutoff a likely candidate for an outbreak threshold. When applying the outbreak threshold to the full data set, containing all sequenced *V. cholerae* genomes meeting the minimum quality threshold, major clusters were examined to evaluate the ability of cgMLST to identify strains that are part of the same outbreak.

One of the major outbreak clusters identified, with no prior information required, contains the Haiti and the Yemen outbreaks, which are the two best-documented cholera outbreaks in modern history (8, 9, 14, 16, 66). Isolates collected from these outbreaks form a single cluster with the Dominican Republic, Eurasian (India, Russia, Nepal, and Ukraine), and African (Tanzania, Kenya, and Somalia) isolates (Fig. 5A). The

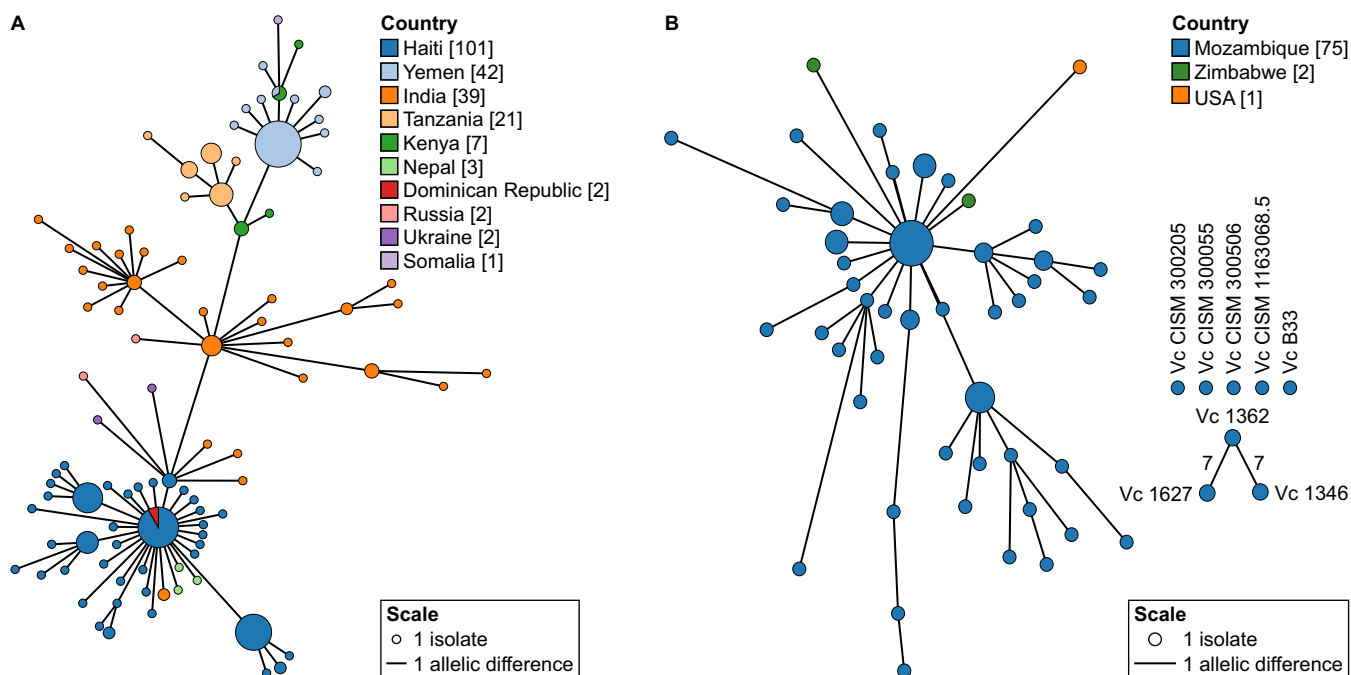


FIG 5 Minimum spanning trees isolated when the outbreak threshold was applied to the complete data set of 1,262 isolates. (A) All isolates that clustered together with the Haiti and Yemen isolates based on the clustering threshold of seven allelic differences. (B) All isolates that clustered with the Mozambique isolates based on the clustering threshold of seven allelic differences. Additional Mozambique isolates that are not part of the same outbreak cluster are also shown. Three isolates, two from Zimbabwe and one from the United States, are connected, as they share seven or fewer allelic differences with the Mozambique isolates. In both panels, the size of the nodes is proportional to the number of isolates. The length of the lines is proportional to the number of allelic differences, and all connections have fewer than or equal to seven allelic differences.

Dominican Republic isolates are closely related to the Haiti outbreak strains. Given the close proximity of the two countries, colocated on the island of Hispaniola, it was expected that isolates from Haiti would eventually spread to the Dominican Republic (14). The 7th pandemic El Tor lineage spread across the world from South Asia in three separate waves (61). The third wave, being the most recent distribution event, has been claimed to be responsible for the outbreaks in Haiti and Yemen (9). Therefore, it is not surprising to see Haiti and Yemen form a single cluster with India (i.e., South Asia) at its center. Nepal is the known source of introduction for the Haiti outbreak in 2010 (16), and comparison with over 1,200 *V. cholerae* isolates from all over the world still shows the Nepalese isolates as the closest relatives to the Haitian isolates (Fig. 5A).

Cholera is still endemic to Africa (10) and caused several major reported outbreaks in different countries over the last few decades (67), including Mozambique (23, 68) and Zimbabwe (69). Another major outbreak cluster groups most of the Mozambique isolates together with two Zimbabwe isolates [strains CP1038(11) and 2011EL-1137] and one U.S. isolate (2009V-1116) (Fig. 5B). Based on cgMLST, it is evident that the two Zimbabwe isolates are closely related to the Mozambique isolates, differing at only four or fewer alleles. The close proximity of the two countries suggests that these are travel-associated cases. Although outbreaks involving the Mozambique isolates (23) and the Zimbabwe isolates (58, 70) have been independently studied, the links between these isolates have not been shown before. Global cgMLST analysis is an invaluable tool, as it allows for the identification of links between independent studies. However, with only two Zimbabwe isolates in the data set, additional sampling in this region is required to understand the epidemiology of this outbreak. According to the NCBI BioSample database, strain 2009V-1116 was collected by the Centers for Disease Control and Prevention in 2009 and is associated with travel to Pakistan. Since the 7th pandemic El Tor lineage has been circulating in Asian and Middle Eastern countries for a long time (71), it is possible that, at least within our data set, the Mozambique isolates are the closest relative to this specific Pakistan strain.

Confirmation of an African connection for the Yemen outbreak. The Yemen cholera outbreak began in October 2016 with eleven confirmed cases (<http://www.emro.who.int/pandemic-epidemic-diseases/cholera/cholera-cases-in-yemen.html>). By January 2017, there were already over 10,000 cholera cases with 99 associated deaths (<http://www.emro.who.int/pandemic-epidemic-diseases/cholera/weekly-update-cholera-cases-in-yemen-15-jan-2017.html>). By the end of that year, there were over 900,000 cholera cases (<http://www.emro.who.int/pandemic-epidemic-diseases/cholera/outbreak-update-cholera-in-yemen-19-december-2017.html>). The outbreak continues today as the largest cholera outbreak in modern history. As isolates from this outbreak were only recently made available (9), they were not part of the initial data set for cgMLST scheme development. These isolates were added and analyzed on PubMLST after the scheme had been established. This set of isolates serves as an independent test of the universality and applicability of the cgMLST scheme. To determine the potential origin of the Yemen outbreak and its phylogenetic relationships with existing *V. cholerae* strains, the Yemen isolates were compared with other 7th pandemic El Tor isolates from Asian and African countries (Table S3). All allele designations and cgST assignments were done automatically on PubMLST. MST was built using these isolates, and all connections with seven and fewer allelic differences are represented as solid lines (Fig. 6). Isolates connected by solid lines belong in the same outbreak cluster, as defined by the outbreak threshold of seven allelic differences. Isolates from Yemen, Kenya, and Haiti all cluster with the central Indian isolates, with seven or fewer allelic differences; however, the closest relatives to the Yemen isolates are those from Kenya, with four or fewer allelic differences (Fig. 6). The Indian isolates are the next closest connection, but there is no direct link between these and the Yemen isolate. This pattern is consistent with the work of Weill and colleagues (9), where they suggested that the Yemen outbreak strains came from East Africa, which themselves came from South Asia, based on SNP-based phylogenetic analysis and Bayesian evolutionary analysis.

Unlike the limited samples available from African cholera outbreaks, the Haiti and Yemen outbreaks are significant cases for epidemiological investigations, because *V. cholerae* has been heavily sampled from these countries as well as surrounding regions. Two major limitations in genomic epidemiology have been the lack of a universal classification scheme and a comprehensive database; however, this is no longer the case in the genomic era, as sequencing technology is becoming increasingly more accessible (8). A genomic approach, as shown here, is able to produce accurate predictions of potential origins of outbreaks and provides us with sufficient resolution to accurately track the spread of the disease. Therefore, genomic analysis should be the first step in any epidemiological study, as it will help guide subsequent analyses and investigations. Consistently sequencing new genomes will also help expand and refine the current global *V. cholerae* genome database.

Increased resolution for the history of cholera in Mozambique: comparing cgMLST to MLVA. The 7th pandemic reached Africa in 1970, and cholera appeared in Mozambique at roughly the same time (57). Since its introduction, cholera has been endemic to that country and has continued to cause multiple outbreaks (23). A popular tool for outbreak investigation is MLVA (32, 38), which was recently used to study *V. cholerae* strains collected in Mozambique over multiple years (23). MLVA is a subspecies typing method similar to MLST in concept; however, it utilizes VNTRs instead of using gene sequences. As a VNTR mutates at a higher rate than conserved genes, it has been shown that MLVA provides greater resolution than MLST for some species (32, 33). To establish a direct comparison between our cgMLST scheme and this MLVA scheme, we examined the MSTs created by both methods, focusing only on shared isolates (Fig. 7A and B). The MLVA identified 26 profiles forming two clonal complexes (CCs) and four singletons (Fig. 7A) (23). A similar population structure is seen with the cgMLST analysis (Fig. 7B), including the four singletons identified in the MLVA. The central node in the cgMLST MST consists mostly of isolates with MLVA profile 8, 4, 6, 18, 21, similar to the

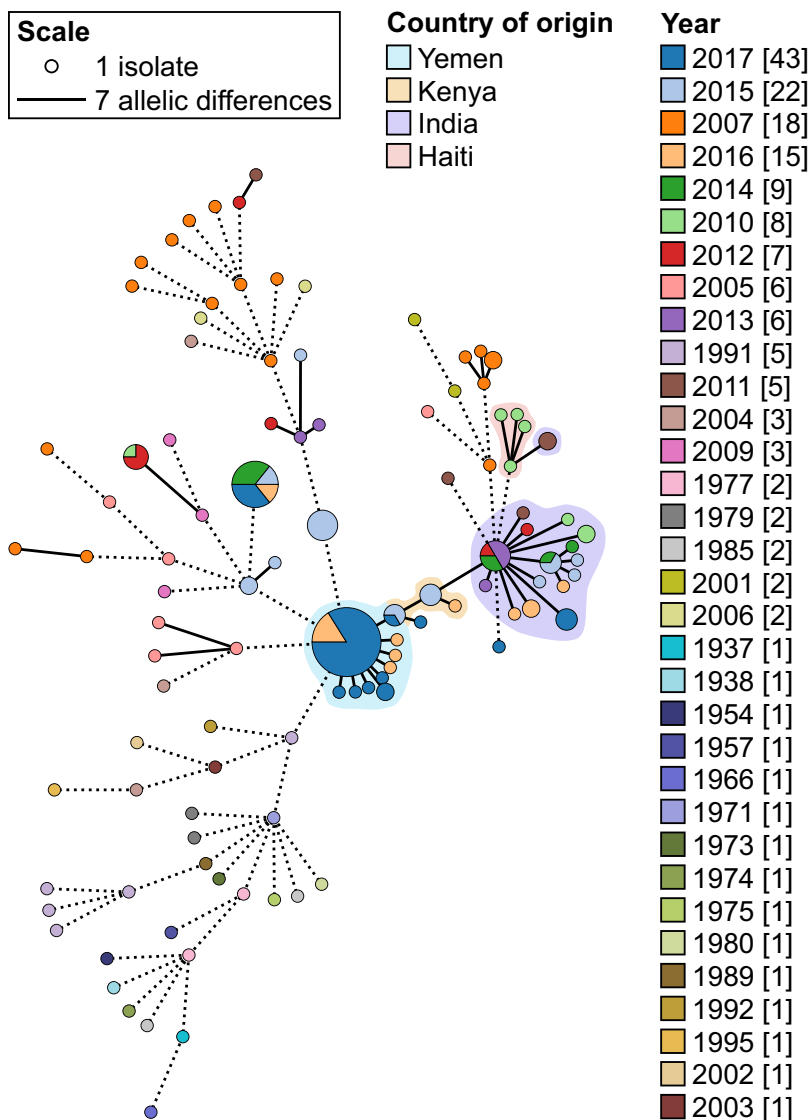


FIG 6 cgMLST MST of all Yemen isolates and representative 7th pandemic El Tor strains. All isolates connected by dotted lines share eight or more allelic differences (not drawn to scale). All isolates connected with solid lines share seven or fewer allelic differences (i.e., they belong to the same outbreak cluster; drawn to scale). Each node represents a cgST that is colored by year of collection. The outbreak clusters are shaded by country.

central node in the MLVA MST (23). The two CCs identified in the MLVA MST are also identified in the cgMLST MST, with the smaller CC2 being at least four allelic differences away from the larger CC1.

Although there are a few MLVA types that were grouped into a single cgST, such as cgST1 and cgST114, indicating that cgMLST was unable to resolve the differences in these MLVA types, there are many MLVA types, such as profile 2, 4, 6, 18, 21, profile 7, 4, 6, 16, 22, profile 9, 4, 6, 18, 24, and profile 8, 4, 6, 18, 22, that were split into multiple cgSTs. Overall, there are 48 cgSTs as opposed to only 26 MLVA types, showing that cgMLST provides better resolution than MLVA. The cgMLST analysis overlaid with isolation dates shows that the Mozambique *V. cholerae* strains are highly clonal, and strains from the same cgST can cause outbreaks over multiple years (e.g., cgST114 and cgST94) (Fig. 7C), which corroborates the claim made in the initial MLVA study that the same MLVA type can be seen over multiple years (23). In addition to increased resolution, cgMLST also produces more reliable and reproducible results than MLVA, as

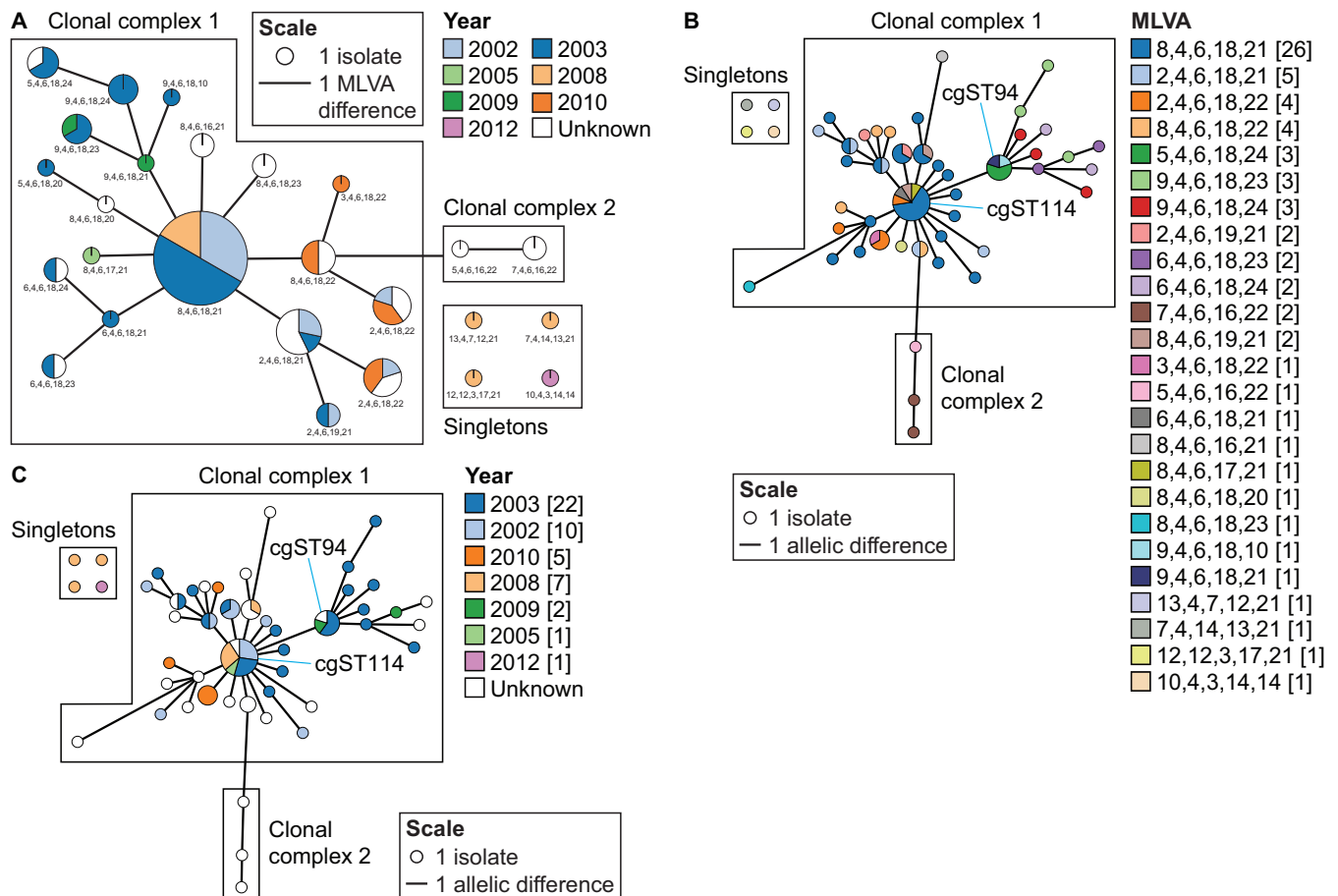


FIG 7 Comparison between cgMLST and MLVA with a focus on the Mozambique isolates. (A) Population structure of pandemic *V. cholerae* in Mozambique based on MLVA profiles by Garrine and colleagues (23). MST of the Mozambique isolates is based on the cgMLST scheme colored based on MLVA profiles (B) and year of isolation (C). All isolates in panels B and C connected with lines share seven or fewer allelic differences. For all panels, the size of the nodes is proportional to the number of isolates. The length of the lines is proportional to the number of allelic differences.

it eliminates errors associated with the detection of VNTR regions using PCR- or sequencing-based methods. For the same reason that MLST is less affected by convergent evolution than MLVA (35), cgMLST is also less affected by convergent evolution.

Standardizing the genotypes responsible for the Haiti 2010 cholera outbreak: comparing cgMLST and SNP-based analyses. One of the largest cholera outbreaks in modern history occurred in Haiti following the devastating earthquake in 2010 (8, 72). Prior to this outbreak, there were no documented cholera cases in Haiti (14, 18). Since the initial introduction, *V. cholerae* now remains endemic in Haiti and is responsible for thousands of cholera cases annually (72). Multiple studies have strongly suggested that the Haitian strains were in fact imported from Nepal (by the United Nations [UN] Nepalese troops), and the outbreak occurred as a result of both inappropriate sanitary practice and the lack of screening of the UN troops upon their arrival in Haiti (8, 15, 16, 72).

A SNP-based approach was used to study the evolutionary dynamics of *V. cholerae* in Haiti (14). This technique relies on the identification of SNPs in draft or closed genomes. The primary benefit of this method is that assembly and annotation are not required. It is also capable of resolving closely related strains using whole-genome data. However, SNP-based methods are highly influenced by recombination events (73), which occur regularly among *V. cholerae* isolates (43, 44), and quality filter parameters chosen (74).

To establish a direct comparison between the cgMLST scheme and SNP-based analysis, we focused on MSTs of only the Haitian outbreak isolates (Fig. 8). All Haitian

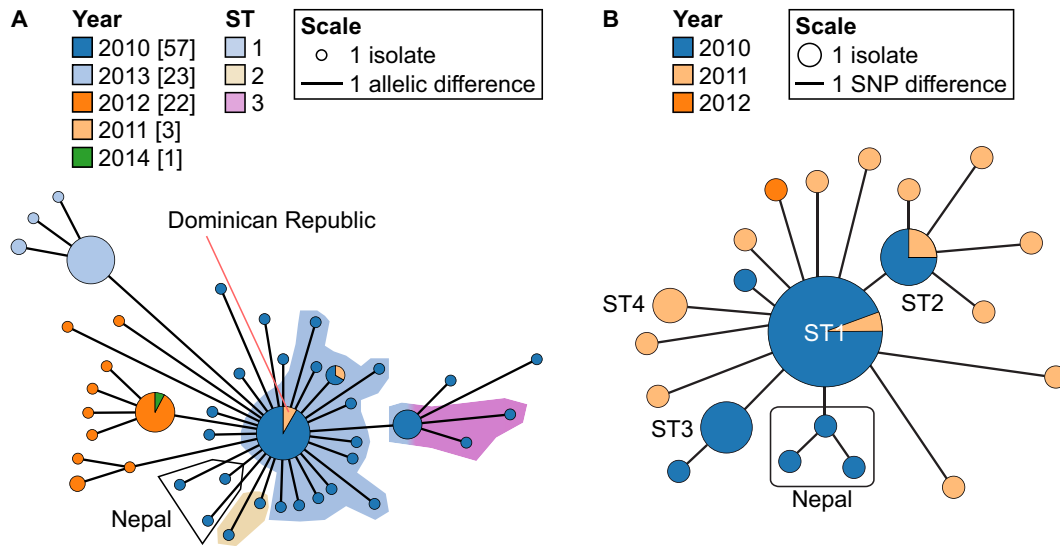


FIG 8 Comparison between cgMLST and SNP-based analysis with a focus on the Haiti outbreak and related isolates. (A) MST of isolates from the 2010 cholera outbreak in Haiti. All lines indicate connections of four or fewer allelic differences. Each node represents a cgST, which is colored by year of isolation. Background shading represents ST designations based on 45 high-quality SNPs by Katz and colleagues (14). Note that cgST66 contains a mix of colors, as it contains both ST1 and ST3. Any isolate from countries other than Haiti is indicated. The length of the lines is proportional to the number of allelic differences. (B) MST constructed from whole-genome SNP data (14). The length of the lines indicates the number of nucleotide substitutions. The size of the nodes is proportional to the number of isolates.

isolates are closely related according to the cgMLST scheme, sharing at most four allelic differences with each other (Fig. 8A). The Haitian and Nepalese isolates, therefore, also belong to the same sublineage (SL6), which is consistent with the fact that these isolates belong to the same MLST ST (either ST1 or ST69 based on the 2016 or 2013 MLST scheme, respectively [20, 21]) (Table S1). The overall population structure is similar between the two methods (Fig. 8), where we have SNP ST1 as the center of the MST with ST2 and ST3 extending from that likely ancestral genotype. SNP ST1, ST2, and ST3 can be split into 14, 2, and 3 different cgSTs, respectively (Fig. 8A). There is only one case, that of cgST66, where it contains isolates from both SNP ST1 and ST3. Overall, cgMLST was able to differentiate 39% of the isolates, while the SNP-based analysis differentiated 35%, showing comparable levels of resolution. As expected, both the cgMLST and the SNP-based analyses showed that the Haiti outbreak is highly clonal, where most isolates belong to the same cgST or SNP ST (14). However, cgMLST is easily standardized across various sequencing tools through the use of a predefined set of core genes. In addition, unlike SNP-based analyses, a systematic nomenclature system based on these standardized genes can be established. Furthermore, cgMLST analyses are also less affected by recombination events (30). Analyses of various cgMLST clusters, cgSTs, and the construction of MST based on the cgMLST scheme is also automated on PubMLST (75). This makes cgMLST more suitable than the SNP-based method as a universally applicable classification system for epidemiological studies and research worldwide.

Environmental isolates differ from clinical strains by their diversity and their associations with specific geographical locations. To look at the geographic signal of *V. cholerae*, we eliminated all clinical isolates and those that belong to the PG lineage (18, 19). This is because the geographic signal of clinical strains can be skewed, as pathogenic strains can travel long distances in a short period of time through association with human hosts. Therefore, the geographical analysis was performed only with environmental isolates.

Along with all of the publicly available environmental strains that are not part of the PG lineage, there are a total of 195 isolates spanning 9 countries. By using a larger

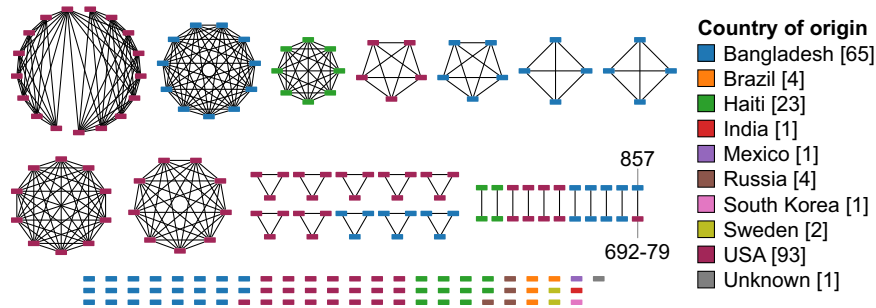


FIG 9 Sublineage clusters of nonclinical environmental isolates that are not part of the PG lineage. Clusters are constructed using NetworkX (100) and visualized with Cytoscape (101). Missing loci were assumed to contain the most common allele when calculating allelic differences. Isolates are connected only if they share 133 allelic differences or fewer with each other. Each node represents an isolate and is colored by the country of origin.

clustering threshold that groups together more distantly related isolates, it is possible to examine broader ecological patterns that otherwise would have been hidden if all analyses were done at the highest level of resolution (i.e., using cgSTs). Therefore, we chose to group isolates at the sublineage (i.e., each cluster has at most 133 allelic differences) rather than the cgST level. It could be noted that all isolates from the same sublineage also shared a country of origin, with the exception of strains 692-79 and 857 (Fig. 9), which are from the United States and Bangladesh, respectively. Phylogenetic analysis shows these isolates to be closely related to strain A215, a clinical isolate from the United States (Fig. 4). All three strains contain the *toxR* gene, a toxin transcriptional regulator common in pathogenic *V. cholerae* (76), as well as genes encoding the mannose-sensitive hemagglutinin pilus, the RTX toxin, and hemolysin (*hlyA*), all of which are putative virulence factors for this species. In addition, strains A215 and 857 also harbor the zona occludens toxin gene. Similar toxin gene contents among these three isolates and close phylogenetic relationships suggest that strains 692-79 and 857 also are pathogenic and capable of surviving inside a human host. This provides evidence that although clinical isolates can spread across the world rapidly and closely related isolates can be from very different parts of the world, environmental isolates from the same geographic origin share an affinity to each other, at least at the sublineage level. It is important to note that our current data set contains a relatively small number of environmental isolates that are not part of the PG lineage. Therefore, this distinct distribution pattern based on geographic origin may be a result of currently insufficient sampling of environmental *V. cholerae* worldwide. Consistent with our finding, however, a broader study that examined the metagenomes from 79 sewage sites from around the world also observed a more pronounced geographic clustering between environmental species and less clustering for pathogens (77). Nonetheless, large-scale environmental sampling of *V. cholerae* will make it possible to determine with greater accuracy the evolutionary rate and distribution pattern of the species in the environment using cgMLST. In addition, this method will become an invaluable tool in dealing with these large data sets, as it provides an efficient and standardized method of classification.

Conclusions. With an extensive collection of over 1,200 *V. cholerae* isolates, we developed a cgMLST scheme based on 2,443 core genes. We established a sublineage-level definition based on 133 allelic differences as part of our standardized classification scheme. It was determined by comparisons with previous MLST schemes that the cgMLST sublineage classification can be used as a proxy for traditional MLST. Additionally, the universality and applicability of the scheme have been tested by looking at various cholera outbreak cases. We determined an outbreak threshold based on seven allelic differences that groups isolates from the same outbreak together with strains from the potential source of introduction. This threshold creates clusters that are consistent with known epidemiological data when applied to the Haiti and Yemen

cholera outbreaks, two of the best-documented cholera outbreaks in modern history. We were also able to confirm the South Asian origin of modern cholera outbreaks. Furthermore, although current sampling is limited, a geographic signal at the sublineage level not seen in clinical strains could be identified among environmental isolates that are not part of the PG lineage (18, 19). Lastly, this scheme is fully implemented on PubMLST (<https://pubmlst.org/vcholerae/>) for public access. All newly available genomes uploaded to PubMLST will be annotated automatically, and a cgST designation will be assigned to isolates with fewer than 100 missing loci. Relevant epidemiological data and a variety of analytical and visualization tools are all integrated on PubMLST (75), allowing for a quick analysis of any newly sequenced genome in a global context. This scheme will be an important tool for future large-scale epidemiological and biogeographical research.

MATERIALS AND METHODS

Data set description. On 6 November 2018, 1,172 *V. cholerae* genomes consisting of 800 draft and complete genomes and 372 sequence read archives (SRAs), available from both a publicly available database and a private collection, were selected as our data set. Sixty-five of these genomes were sequenced in this study using the PacBio or Illumina sequencing platform, as previously described (78, 79). To confirm that the genomes indeed belong to the *V. cholerae* species, we calculated average nucleotide identity values against the *V. cholerae* N16961 reference genome (80) using JSpecies (81). All of the genomes exhibited at least 95% average nucleotide identity against N16961, which is the currently accepted standard for species delineation (82) and correctly verifying the identity of the genomes. One hundred sixteen SRAs from a recent study on the Yemen cholera outbreak (9) were subsequently added as an independent evaluation of the cgMLST scheme (see Table S3 in the supplemental material). The 488 SRAs were assembled using skesa (83) or CLC Genomics Workbench 7 (Qiagen) using default parameters. This total data set of 1,288 included twenty-six genomes with less than 90% of the core genes, which were identified using USearch (84) based on RAST annotations (85). These 26 genomes were removed from subsequent analyses, resulting in a final data set of 1,262 genomes collected from 52 countries and spanning 82 years, from 1937 to 2018 (Table S1). These include a historical collection from the 6th cholera pandemic, clinical isolates from outbreaks in various countries (e.g., Bangladesh, India, Haiti, Yemen, the Democratic Republic of Congo, Russia, etc.), and environmental isolates from different parts of the world (e.g., Bangladesh, Haiti, the United States, Mexico, Brazil, etc.).

Gene identification and allele assignments. Instead of using the full data set of 1,288 genomes, we selected a subset of high-quality genomes, because core gene identification is highly dependent on the initial data set, and the inclusion of poorly assembled and/or sequenced data will reduce the number of core genes identified (49). First, 800 already assembled draft or complete genomes were selected for core gene identification. Low-quality assemblies were then eliminated by removing genomes with less than 40 \times coverage and/or N_{50} values less than 40 kb. From a previously established cgMLST scheme for *L. monocytogenes*, 40 \times coverage and a 20-kb N_{50} value were used as cutoff thresholds, as genomes that do not meet these criteria resulted in a low proportion of loci being called (49). The 40 \times coverage cutoff was adopted for this study; however, because the average *V. cholerae* genome size (~4 Mb) is larger than the average *L. monocytogenes* genome (~3 Mb), 40 kb was instead selected as the N_{50} cutoff. The use of these cutoffs resulted in the removal of 82 genomes.

The remaining 718 genomes were annotated using RAST (85) and USearch (84), and a tentative set of core genes that were present in 99% of the genomes, on average, were selected. An additional 13 genomes were subsequently removed, as they lacked more than 90% of the core genes, leaving us with a data set of 705 high-quality genomes (Table S4). However, 26 genomes were further removed for the core gene analysis, as it was previously suggested that they form a highly divergent lineage within *V. cholerae* (78, 79, 86), ensuring that the data set used for core gene identification consists only of unambiguously *V. cholerae* isolates. The completeness and potential contamination of the remaining 679 genomes were also independently evaluated by checkM, which estimates these values based on the presence and number of copies of a set of predefined single-copy marker genes (87) (Table S5). All genomes were, according to the criteria established by checkM, nearly complete ($\geq 97\%$), with medium to low levels of contamination ($< 7\%$) (87).

Each orthologous gene was compared against the *V. cholerae* N16961 reference genome using BLASTN (88) to determine gene function. Any gene family with no homolog in N16961 or classified as pseudogenes in the NCBI GenBank database were removed, meaning N16961 was 100% complete for the cgMLST scheme. Any genes that were present in more than one copy in any of the initial 679 genomes were also removed, as they were considered paralogous. Thus, in this context, core genes are defined as being present in at least 90% of the 679 high-quality assembled genomes in a single copy. By choosing a relaxed cutoff of 90% completeness, we accounted for missing genes due to sequencing, annotation, or assembly errors while ensuring there is sufficient resolution to differentiate between closely related strains, with at least 2,199 loci remaining for classification purposes. The final cgMLST scheme utilizes a set of 2,443 single-copy core gene loci, which is 2,425,296 bp in size and covers approximately 61% of the genome. The list of core genes is available on PubMLST (<https://pubmlst.org/vcholerae/>).

Automated scripts in BIGSdb (89) were used to perform allele calls and assignments for all 1,262 isolates (Table S1). Allele calls were made only for complete coding sequences with a minimum of 70% similarity and 70% length coverage at the nucleotide level, as previously described (49). Default settings were used for all other parameters.

cgST assignment. cgST, which was defined as a unique combination of alleles of all loci included in the scheme, was assigned for all isolates, excluding those from the Yemen outbreak study (9), with an in-house script, as previously described (90). Briefly, missing loci were replaced with the most common allele when assigning cgSTs, allowing for a conservative estimate of diversity (90). The 116 isolates from the Yemen cholera outbreak study (9) were annotated automatically by uploading them to PubMLST. PubMLST treated missing alleles as "N." cgSTs were assigned to each allele profile, treating N as a regular allele designation. However, different from typical allele designations, Ns can represent any allelic sequence; therefore, some isolates may contain multiple cgST designations, all of which are possibly true cgSTs. For isolates with more than one cgST suggested by PubMLST, postprocessing was done using an in-house script to identify the most likely cgST, which was determined by assuming missing loci contained the most common allele (Table S2). It is expected that as genome sequencing becomes more reliable, higher-quality genomes will be available and any missing data can be updated as needed.

MLST scheme and ST assignments. Two MLST schemes developed for *V. cholerae* were mapped to this cgMLST scheme. The first MLST scheme, developed in 2013 by Octavia and colleagues (20), was used to study the global population structure of non-O1/non-O139 *V. cholerae* and is currently hosted on PubMLST. All isolates uploaded to PubMLST were automatically annotated with this scheme. Any missing data in this scheme were ignored and no ST designation was assigned. The second MLST scheme, developed in 2016 by Kirchberger and colleagues (21), was used to study the population structure of environmental *V. cholerae* in a region on the U.S. East Coast. The second MLST scheme is not currently hosted on PubMLST, but because the housekeeping genes in this scheme are also found in the cgMLST scheme, a similar in-house script used in cgST assignments was used to assign ST designations. Therefore, all isolates in this study were assigned three designations when possible, two ST designations based on the two previously established MLST schemes (20, 21) and one cgST designation based on the cgMLST scheme from this study.

Outbreak and sublineage clustering thresholds. A clustering threshold was defined as the maximum number of allelic differences found within a cluster. All clusters were produced based on the single-linkage clustering method, which meant an isolate belonged to a cluster if it linked with any isolate within that cluster. Two metrics were used as general guidelines for determining clustering thresholds. The first metric used was the Dunn index (DI), which measured clustering efficiencies (60). Briefly, the DI was highest for a network (i.e., the network has the best clustering efficiency) when the intracluster distances were minimized and the intercluster distances were maximized. Since isolate distances were measured based on allelic differences, a high DI resulted in clusters where isolates were more closely related to those found within the same cluster than those found in a different cluster. The DI was calculated using the R packages *clvalid* and *boot*, with 100 bootstrap replicates for each threshold, and graphed using the R package *ggplot2* (91–94).

The second metric used was the adjusted Rand index (ARI), which measured the level of similarity between two networks when clustering the same set of isolates by measuring the amount of agreement (i.e., the number of pairs that were grouped either as being in the same cluster or different clusters in both networks) and disagreements (i.e., the number of pairs that were grouped together in one network but grouped separately in another) (62). The values ranged from -1 (i.e., two networks are exactly opposite) to 1 (i.e., two networks are identical). ARI was used to determine the level of similarity between various clustering thresholds and the MLST schemes. ARI was calculated using the R package *clues* and graphed using *ggplot2* (92, 94, 95).

MST. All MSTs, unless otherwise specified, were constructed using GrapeTree MSTv2 (96). Loci with missing data were included in the profile as dashes. GrapeTree provided a novel algorithm that accounted for missing data when constructing an MST, an important feature, since missing data are common in whole- and core genome-based analyses. GrapeTree is currently integrated within PubMLST (75), which allows for quick visualization of the data set with any provenance data.

Phylogenetic analysis. Parsnp v1.2 (97) was used to reconstruct the phylogenetic tree using *V. cholerae* N16961 as the reference genome. The $-x$ flag was used to enable filtering of SNPs in recombinogenic regions as identified by PhiPack (98). Default settings were used for all other parameters. The phylogenetic tree included 1,146 genomes (all genomes except for the 116 isolates from the recent Yemen cholera outbreak study [9]). Since all isolates sequenced for the latter study belonged to the 7th pandemic El Tor lineage, it would have had limited impact on the overall structure of the tree. The phylogeny was visualized and annotated using iTOL (99).

Biogeographical analysis of environmental isolates. All environmental isolates that were not part of the PG lineage (18, 19) were first clustered based on the sublineage threshold using the python package *networkX* (100). Missing alleles were replaced with the most common allelic designation when calculating pairwise differences to establish a more conservative estimate of diversity. The network was then visualized using Cytoscape (101).

Data availability. All previously sequenced *V. cholerae* genomes and the additional 65 genomes sequenced in this study are available from the NCBI GenBank and the PubMLST databases. Table S6 lists all the accession numbers and PubMLST IDs for all the genomes used in this study. In addition, all genome sequences, allelic profiles, cgST designations, ST designations, and relevant epidemiological data are publicly available on PubMLST (<https://pubmlst.org/vcholerae/>).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.9 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.1 MB.

SUPPLEMENTAL FILE 3, XLSX file, 10.4 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 5, XLSX file, 0.03 MB.

SUPPLEMENTAL FILE 6, XLSX file, 0.03 MB.

SUPPLEMENTAL FILE 7, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

K.Y.H.L., F.D.O., and Y.F.B. conceived the experiments. K.Y.H.L., F.D.O., and M.T.I. performed all data collection and analyses. F.D.O. and T.N. performed genome sequencing of Bangladesh isolates. M.A. and C.L.T. provided isolates used in this study.

We thank Keith Jolley (University of Oxford) for providing valuable feedback regarding the development of the cgMLST scheme, as well as the implementation of this scheme on PubMLST. We also thank Monica Im (Centers for Disease Control and Prevention) for assistance with obtaining whole-genome sequences.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (to Y.F.B.); the Integrated Microbial Biodiversity program of the Canadian Institute for Advanced Research (to Y.F.B.); federal appropriations to the Centers for Disease Control and Prevention through the Advanced Molecular Detection Initiative (to C.L.T.); and graduate student scholarships from Alberta Innovates–Technology Futures (to K.Y.H.L., F.D.O., M.T.I., and T.N.), NSERC (to K.Y.H.L., and T.N.), the University of Alberta Faculty of Graduate Studies and Research (Queen Elizabeth II Graduate Scholarship to K.Y.H.L., and T.N.), and the Bank of Montréal Financial Group (to F.D.O.). M.A. acknowledges the governments of Bangladesh, Canada, Sweden, and the United Kingdom for providing core/unrestricted support.

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

- Jahan S. 2016. Cholera—epidemiology, prevention and control, p 145–157. In Makun HA (ed), Significance, prevention and control of food related diseases. InTechOpen, Rijeka, Croatia.
- Momba M, Azab El-Liethy M. 2017. *Vibrio cholerae* and cholera biotypes. In Pruden A, Ashbolt N, Miller J (ed), Global water pathogen project. Michigan State University, Lansing, Michigan.
- Clemens JD, Nair GB, Ahmed T, Qadri F, Holmgren J. 2017. Cholera. *Lancet* 390:1539–1549. [https://doi.org/10.1016/S0140-6736\(17\)30559-7](https://doi.org/10.1016/S0140-6736(17)30559-7).
- Kaper JB, Morris JG, Levine MM. 1995. Cholera. *Clin Microbiol Rev* 8:48–86. <https://doi.org/10.1128/CMR.8.1.48>.
- Islam MT, Alam M, Boucher Y. 2017. Emergence, ecology and dispersal of the pandemic generating *Vibrio cholerae* lineage. *Int Microbiol* 20:106–115. <https://doi.org/10.2436/20.1501.01.291>.
- Choi SY, Rashed SM, Hasan NA, Alam M, Islam T, Sadique A, Johura F-T, Eppinger M, Ravel J, Huq A, Cravioto A, Colwell RR. 2016. Phylogenetic diversity of *Vibrio cholerae* associated with endemic cholera in Mexico from 1991 to 2008. *mBio* 7:e02160-15. <https://doi.org/10.1128/mBio.02160-15>.
- Dalsgaard A, Skov MN, Serichantalergs O, Echeverria P, Meza R, Taylor DN. 1997. Molecular evolution of *Vibrio cholerae* O1 strains isolated in Lima, Peru, from 1991 to 1995. *J Clin Microbiol* 35:1151–1156. <https://doi.org/10.1128/JCM.35.5.1151-1156.1997>.
- Orata FD, Keim PS, Boucher Y. 2014. The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathog* 10:e1003967. <https://doi.org/10.1371/journal.ppat.1003967>.
- Weill F-X, Domman D, Njamkepo E, Almesbahi AA, Naji M, Nasher SS, Rakesh A, Assiri AM, Sharma NC, Kariuki S, Pourshafie MR, Raugier J, Abubakar A, Carter JY, Wamala JF, Seguin C, Bouchier C, Malliavin T, Bakhshi B, Abulmaali HH, Kumar D, Njoroge SM, Malik MR, Kiiru J, Luquero FJ, Azman AS, Ramamurthy T, Thomson NR, Quilici M-L. 2019. Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* 565:230–233. <https://doi.org/10.1038/s41586-018-0818-3>.
- Ali M, Nelson AR, Lopez AL, Sack DA. 2015. Updated global burden of cholera in endemic countries. *PLoS Negl Trop Dis* 9:e0003832. <https://doi.org/10.1371/journal.pntd.0003832>.
- Legros D, Partners of the Global Task Force on Cholera Control. 2018. Global cholera epidemiology: opportunities to reduce the burden of cholera by 2030. *J Infect Dis* 218:S137–S140. <https://doi.org/10.1093/infdis/jiy486>.
- Loharikar A, Briere E, Ope M, Langat D, Njeru I, Gathigi L, Makayotto L, Ismail AM, Thurairam M, Abade A, Amwayi S, Omolo J, Oundo J, De Cock KM, Breiman RF, Ayers T, Mintz E, O'Reilly CE. 2013. A national cholera epidemic with high case fatality rates—Kenya 2009. *J Infect Dis* 208: S69–S77. <https://doi.org/10.1093/infdis/jit220>.
- Sack DA, Sack RB, Chaignat C-L. 2006. Getting serious about cholera. *N Engl J Med* 355:649–651. <https://doi.org/10.1056/NEJMp068144>.
- Katz LSS, Petkau A, Beaulaurier J, Tyler S, Antonova ESS, Turnsek MAA, Guo Y, Wang S, Paxinos EEE, Orata F, Gladney LMM, Stroika S, Folster JPP, Rowe L, Freeman MMM, Knox N, Frace M, Boncy J, Graham M, Hammer BKK, Boucher Y, Bashir A, Hanage WPP, Van Domselaar GV, Tarr L, Van Domselaar G, Tarr CLL, Van Domselaar GV. 2013. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *mBio* 4:e00398-13. <https://doi.org/10.1128/mBio.00398-13>.
- Frerichs RR. 2016. Deadly river: cholera and cover-up in post-earthquake Haiti. Cornell University Press, Ithaca, NY.
- Frerichs RR, Keim PS, Barrais R, Piarroux R. 2012. Nepalese origin of

- cholera epidemic in Haiti. *Clin Microbiol Infect* 18:E158–E163. <https://doi.org/10.1111/j.1469-0691.2012.03841.x>.
17. Hendriksen RS, Price LB, Schupp JM, Gillette JD, Kaas RS, Engelthaler DM, Bortolola V, Pearson T, Waters AE, Prasad Upadhyay B, Devi Shrestha S, Adhikari S, Shakya G, Keim PS, Aarestrup FM. 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *mBio* 2:e00157-11. <https://doi.org/10.1128/mBio.00157-11>.
 18. Boucher Y. 2016. Sustained local diversity of *Vibrio cholerae* O1 biotypes in a previously cholera-free country. *mBio* 7:e00570-16. <https://doi.org/10.1128/mBio.00570-16>.
 19. Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, Taviani E, Jeon Y-S, Kim DW, Lee J-H, Bretin TS, Bruce DC, Challacombe JF, Detter JC, Han CS, Munk AC, Chertkov O, Meincke L, Saunders E, Walters RA, Huq A, Nair GB, Colwell RR. 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci U S A* 106:15442–15447. <https://doi.org/10.1073/pnas.0907787106>.
 20. Octavia S, Salim A, Kurniawan J, Lam C, Leung Q, Ahsan S, Reeves PR, Nair GB, Lan R. 2013. Population structure and evolution of non-O1/non-O139 *Vibrio cholerae* by multilocus sequence typing. *PLoS One* 8:e65342. <https://doi.org/10.1371/journal.pone.0065342>.
 21. Kirchberger PC, Orata FD, Barlow EJ, Kauffman KM, Case RJ, Polz MF, Boucher Y. 2016. A small number of phylogenetically distinct clonal complexes dominate a coastal *Vibrio cholerae* population. *Appl Environ Microbiol* 82:5576–5586. <https://doi.org/10.1128/AEM.01177-16>.
 22. Bwire G, Sack DA, Almeida M, Li S, Voeglein JB, Debes AK, Kagirita A, Buyinza AW, Orach CG, Stine OC. 2018. Molecular characterization of *Vibrio cholerae* responsible for cholera epidemics in Uganda by PCR, MLVA and WGS. *PLoS Negl Trop Dis* 12:e0006492. <https://doi.org/10.1371/journal.pntd.0006492>.
 23. Garrine M, Mandomando I, Vubil D, Nhampossa T, Acacio S, Li S, Paulson JN, Almeida M, Domman D, Thomson NR, Alonso P, Stine OC. 2017. Minimal genetic change in *Vibrio cholerae* in Mozambique over time: multilocus variable number tandem repeat analysis and whole genome sequencing. *PLoS Negl Trop Dis* 11:e0005671. <https://doi.org/10.1371/journal.pntd.0005671>.
 24. Safa A, Nair GB, Kong R. 2010. Evolution of new variants of *Vibrio cholerae* O1. *Trends Microbiol* 18:46–54. <https://doi.org/10.1016/j.tim.2009.10.003>.
 25. Boucher Y, Orata FD, Alam M. 2015. The out-of-the-delta hypothesis: dense human populations in low-lying river deltas served as agents for the evolution of a deadly pathogen. *Front Microbiol* 6:1120. <https://doi.org/10.3389/fmicb.2015.01120>.
 26. Mandal S, Mandal MD, Pal NK. 2011. Cholera: a great global concern. *Asian Pac J Trop Med* 4:573–580. [https://doi.org/10.1016/S1995-7645\(11\)60149-1](https://doi.org/10.1016/S1995-7645(11)60149-1).
 27. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 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. <https://doi.org/10.1073/pnas.95.6.3140>.
 28. Horwood P, Collins D, Jonduo M, Rosewell A, Dutta S, Dagina R, Ropa B, Siba P, Greenhill A. 2011. Clonal origins of *Vibrio cholerae* O1 El Tor strains, Papua New Guinea, 2009–2011. *Emerg Infect Dis* 17:2063–2065. <https://doi.org/10.3201/eid1711.110782>.
 29. Luo Y, Ye J, Jin D, Ding G, Zhang Z, Mei L, Octavia S, Lan R. 2013. Molecular analysis of non-O1/non-O139 *Vibrio cholerae* isolated from hospitalised patients in China. *BMC Microbiol* 13:52. <https://doi.org/10.1186/1471-2180-13-52>.
 30. Maiden MCJ, Van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* 11:728–736. <https://doi.org/10.1038/nrmicro3093>.
 31. Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo TR, Jaykus L-A, DePaola A. 2008. Determination of molecular phylogenetics of *Vibrio parahaemolyticus* strains by multilocus sequence typing. *J Bacteriol* 190:2831–2840. <https://doi.org/10.1128/JB.01808-07>.
 32. Lam C, Octavia S, Reeves PR, Lan R. 2012. Multi-locus variable number tandem repeat analysis of 7th pandemic *Vibrio cholerae*. *BMC Microbiol* 12:82. <https://doi.org/10.1186/1471-2180-12-82>.
 33. Chenal-Francois V, Passet V, Brisse S, Cantinelli T, Diancourt L, Pourcel C, Lecuit M, Leclercq A, Tran-Hykes C, Bracq-Dieye H. 2013. Optimized multilocus variable-number tandem-repeat analysis and its complementarity with pulsed-field gel electrophoresis and multilocus sequence typing for *Listeria monocytogenes* clone identification and surveillance. *J Clin Microbiol* 51:1868–1880. <https://doi.org/10.1128/JCM.00606-13>.
 34. Vogler AJ, Birdsell DN, Lee J, Vaissaire J, Doujet CL, Lapalus M, Wagner DM, Keim P. 2011. Phylogeography of *Francisella tularensis* ssp. holarctica in France. *Lett Appl Microbiol* 52:177–180. <https://doi.org/10.1111/j.1472-765X.2010.02977.x>.
 35. Struelens MJ, Brisse S. 2013. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. *Euro Surveill* 18:20386. <https://doi.org/10.2807/ese.18.04.20386-en>.
 36. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijk JM, Laurent F, Grundmann H, Friedrich AW, ESCMID Study Group. 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill* 18:20380. <https://doi.org/10.2807/ese.18.04.20380-en>.
 37. Klassen JL, Currie CR. 2012. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 13:14. <https://doi.org/10.1186/1471-2164-13-14>.
 38. Danin-Poleg Y, Cohen LA, Gancz H, Broza YY, Goldshmidt H, Malul E, Valinsky L, Lerner L, Broza M, Kashi Y. 2007. *Vibrio cholerae* strain typing and phylogeny study based on simple sequence repeats. *J Clin Microbiol* 45:736–746. <https://doi.org/10.1128/JCM.01895-06>.
 39. Wong VK, International Typhoid Consortium, Baker S, Connor TR, Pickard D, Page AJ, Dave J, Murphy N, Holliman R, Sefton A, Millar M, Dyson ZA, Dougan G, Holt KE. 2016. An extended genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. *Nat Commun* 7:12827. <https://doi.org/10.1038/ncomms12827>.
 40. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. 2014. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLoS One* 9:e87991. <https://doi.org/10.1371/journal.pone.0087991>.
 41. Chen C, Zhang W, Zheng H, Lan R, Wang H, Du P, Bai X, Ji S, Meng Q, Jin D, Liu K, Jing H, Ye C, Gao GF, Wang L, Gottschalk M, Xu J. 2013. Minimum core genome sequence typing of bacterial pathogens: a unified approach for clinical and public health microbiology. *J Clin Microbiol* 51:2582–2591. <https://doi.org/10.1128/JCM.00535-13>.
 42. Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R, Xu J. 2016. Population structure and minimum core genome typing of *Legionella pneumophila*. *Sci Rep* 6:21356. <https://doi.org/10.1038/srep21356>.
 43. Wang R, Yu D, Yue J, Kan B. 2016. Variations in SXT elements in epidemic *Vibrio cholerae* O1 El Tor strains in China. *Sci Rep* 6:22733. <https://doi.org/10.1038/srep22733>.
 44. Meibom KL, Blokesch M, Dolganov NA, Wu C-Y, Schoolnik GK. 2005. Chitin induces natural competence in *Vibrio cholerae*. *Science* 310:1824–1827. <https://doi.org/10.1126/science.1120096>.
 45. Borgeaud S, Metzger LC, Scrinari T, Blokesch M. 2015. The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* 347:63–68. <https://doi.org/10.1126/science.1260064>.
 46. Orata FD, Kirchberger PC, Méheust R, Barlow EJ, Tarr CL, Boucher Y. 2015. The dynamics of genetic interactions between *Vibrio metoecus* and *Vibrio cholerae*, two close relatives co-occurring in the environment. *Genome Biol Evol* 7:2941–2954. <https://doi.org/10.1093/gbe/evw193>.
 47. Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Baptiste E, Lopez P, Tarr CL, Polz MF. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *mBio* 2:e00335-10. <https://doi.org/10.1128/mBio.00335-10>.
 48. Neumann B, Prior K, Bender JK, Harmsen D, Klare I, Fuchs S, Bethe A, Zühlke D, Göhler A, Schwarz S, Schaffer K, Riedel K, Wieler LH, Werner G. 2019. A core genome multilocus sequence typing scheme for *Enterococcus faecalis*. *J Clin Microbiol* 57:e01686-18. <https://doi.org/10.1128/JCM.01686-18>.
 49. Moura A, Criscuolo A, Pousee H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larssonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francois V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2017. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol* 2:1–10. <https://doi.org/10.1038/nmicrobiol.2016.185>.
 50. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen

- D, Willems R. 2015. Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. J Clin Microbiol 53:3788–3797. <https://doi.org/10.1128/JCM.01946-15>.
51. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden M. 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. <https://doi.org/10.1128/JCM.00080-17>.
 52. Janowicz A, De Massis F, Ancora M, Camma C, Patavino C, Battisti A, Prior K, Harmsen D, Scholz H, Zilli K, Sacchini L, Di Giannatale E, Garofolo G. 2018. Core genome multilocus sequence typing and single nucleotide polymorphism analysis in the epidemiology of *Brucella melitensis* infections. J Clin Microbiol 56:e00517-18. <https://doi.org/10.1128/JCM.00517-18>.
 53. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. 2018. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. J Clin Microbiol 56:e01987-17. <https://doi.org/10.1128/JCM.01987-17>.
 54. Jones RC, Harris LG, Morgan S, Ruddy MC, Perry M, Williams R, Humphrey T, Temple M, Davies AP. 2019. Phylogenetic analysis of *Mycobacterium tuberculosis* strains in Wales by use of core genome multilocus sequence typing to analyze whole-genome sequencing data. J Clin Microbiol 57:e02025-18. <https://doi.org/10.1128/JCM.02025-18>.
 55. Sails AD, Swaminathan B, Fields PI. 2003. Clonal complexes of *Campylobacter jejuni* identified by multilocus sequence typing correlate with strain associations identified by multilocus enzyme electrophoresis. J Clin Microbiol 41:4058–4067. <https://doi.org/10.1128/jcm.41.9.4058-4067.2003>.
 56. Leavis HL, Bonten MJ, Willems RJ. 2006. Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance. Curr Opin Microbiol 9:454–460. <https://doi.org/10.1016/j.mib.2006.07.001>.
 57. Weill F-X, Domman D, Njamkepo E, Tarr C, Raugier J, Fawal N, Keddy KH, Salje H, Moore S, Mukhopadhyay AK, Bercion R, Luquero FJ, Ngandjio A, Dosso M, Monakhova E, Garin B, Bouchier C, Pazzani C, Mutreja A, Grunow R, Sidikou F, Bonte L, Breurec S, Damian M, Njanpop-Lafourcade B-M, Sapriel G, Page A-L, Hamze M, Henkens M, Chowdhury G, Mengel M, Koeck J-L, Fournier J-M, Dougan G, Grimont PAD, Parkhill J, Holt KE, Piarroux R, Ramamurthy T, Quilici M-L, Thomson NR. 2017. Genomic history of the seventh pandemic of cholera in Africa. Science 358:785–789. <https://doi.org/10.1126/science.aad5901>.
 58. Reimer A, V. cholerae Outbreak Genomics Task Force, Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L, Olsen-Rasmussen M, Frace M, Sammons S, Dahourou G, Boncy J, Smith A, Mabon P, Petkau A, Graham M, Gilmour M, Gerner-Smidt P. 2011. Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. Emerg Infect Dis 17:2113–2121. <https://doi.org/10.3201/eid1711.110794>.
 59. Salim A, Lan R, Reeves PR. 2005. *Vibrio cholerae* pathogenic clones. Emerg Infect Dis 11:1758–1760. <https://doi.org/10.3201/eid1111.041170>.
 60. Dunn JC. 1974. Well-separated clusters and optimal fuzzy partitions. J Cybern 4:95–104. <https://doi.org/10.1080/01969727408546059>.
 61. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462–465. <https://doi.org/10.1038/nature10392>.
 62. Hubert L, Arabie P. 1985. Comparing partitions. J Classif 2:193–218. <https://doi.org/10.1007/BF01908075>.
 63. Lucidarme J, Hill DMC, Bratcher HB, Gray SJ, Du Plessis M, Tsang RSW, Vazquez JA, Taha M-K, Ceyhan M, Efron AM, Gorla MC, Findlow J, Jolley KA, Maiden MCJ, Borrow R. 2015. Genomic resolution of an aggressive, widespread, diverse and expanding meningococcal serogroup B, C and W lineage. J Infect 71:544–552. <https://doi.org/10.1016/j.jinf.2015.07.007>.
 64. Royer G, Fourreau F, Boulanger B, Mercier-Darty M, Ducellier D, Cizeau F, Patron A, Podglajen I, Mongardon N, Decusser J-W. 2019. Local outbreak of extended-spectrum β -lactamase SHV2a-producing *Pseudomonas aeruginosa* reveals the emergence of a new specific sub-lineage of the international ST235 high-risk clone. J Hosp Infect 104:33–39. <https://doi.org/10.1016/j.jhinf.2019.07.014>.
 65. Tu Y-H, Chen B-H, Hong Y-P, Liao Y-S, Chen Y-S, Liu Y-Y, Teng R-H, Wang Y-W, Chiou C-S. 2020. Emergence of *Vibrio cholerae* O1 sequence type 75 in Taiwan. Emerg Infect Dis 26:164–166. <https://doi.org/10.3201/eid2601.190934>.
 66. Eppinger M, Pearson T, Koenig SSK, Pearson O, Hicks N, Agrawal S, Sanjar F, Galens K, Daugherty S, Crabtree J, Hendriksen RS, Price LB, Upadhyay BP, Shakya G, Fraser CM, Ravel J, Keim PS. 2014. Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. mBio 5:e01721-14. <https://doi.org/10.1128/mBio.01721-14>.
 67. World Health Organization. 2017. Cholera. Wkly Epidemiol Rec 92: 521–536.
 68. Cavailler P, Lucas M, Perroud V, McChesney M, Ampuero S, Guérin PJ, Legros D, Nierle T, Mahoudeau C, Lab B, Kahozo P, Deen JL, von Seidlein L, Wang XY, Puri M, Ali M, Clemens JD, Songane F, Baptista A, Ismael F, Barreto A, Chagnat CL. 2006. Feasibility of a mass vaccination campaign using a two-dose oral cholera vaccine in an urban cholera-endemic setting in Mozambique. Vaccine 24:4890–4895. <https://doi.org/10.1016/j.vaccine.2005.10.006>.
 69. Sardar T, Mukhopadhyay S, Bhowmick AR, Chattopadhyay J. 2013. An optimal cost effectiveness study on Zimbabwe cholera seasonal data from 2008–2011. PLoS One 8:e81231. <https://doi.org/10.1371/journal.pone.0081231>.
 70. Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, Haley BJ, Taviani E, Hine E, Su Q, Tallon LJ, Prosper JB, Furth K, Hoq MM, Li H, Fraser-Liggett CM, Cravioto A, Huq A, Ravel J, Cebula TA, Colwell RR. 2012. Genomic diversity of 2010 Haitian cholera outbreak strains. Proc Natl Acad Sci U S A 109:E2010–E2017. <https://doi.org/10.1073/pnas.1207359109>.
 71. Hu D, Liu B, Feng L, Ding P, Guo X, Wang M, Cao B, Reeves PR, Wang L. 2016. Origins of the current seventh cholera pandemic. Proc Natl Acad Sci U S A 113:E7730–E7739. <https://doi.org/10.1073/pnas.1608732113>.
 72. Guillaume Y, Ternier R, Vissieres K, Casseus A, Chery MJ, Ivers LC. 2018. Responding to cholera in Haiti: implications for the national plan to eliminate cholera by 2022. J Infect Dis 218:S167–S170. <https://doi.org/10.1093/infdis/jiy491>.
 73. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoesser N, Peto TEA, Crook DW, Walker AS. 2020. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience 9:giaa007. <https://doi.org/10.1093/gigascience/giaa007>.
 74. Pightling AW, Petronella N, Pagotto F. 2014. Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* Short-read sequence data greatly influences rates of error in SNP analyses. PLoS One 9:e104579. <https://doi.org/10.1371/journal.pone.0104579>.
 75. Jolley KA, Bray JE, Maiden M. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. <https://doi.org/10.12688/wellcomeopenres.14826.1>.
 76. Childers BM, Klose KE. 2007. Regulation of virulence in *Vibrio cholerae*: the ToxR regulon. Future Microbiol 2:335–344. <https://doi.org/10.2217/17460913.2.3.335>.
 77. Ahrenfeldt J, Waisi M, Loft IC, Clausen P, Allesøe R, Szarvas J, Hendriksen RS, Aarestrup FM, Lund O. 2020. Metaphylogenetic analysis of global sewage reveals that bacterial strains associated with human disease show less degree of geographic clustering. Sci Rep 10:3033. <https://doi.org/10.1038/s41598-020-59292-w>.
 78. Liang K, Orata FD, Winkjer NS, Rowe LA, Tarr CL, Boucher Y. 2017. Complete genome sequence of *Vibrio* sp. strain 2521-89, a close relative of *Vibrio cholerae* isolated from lake water in New Mexico, USA. Genome Announc 5:e00905-17. <https://doi.org/10.1128/genomeA.00905-17>.
 79. Liang K, Islam MT, Hussain N, Winkjer NS, Im MS, Rowe LA, Tarr CL, Boucher Y. 2019. Draft genome sequences of eight *Vibrio* sp. clinical isolates from across the United States that form a basal sister clade to *Vibrio cholerae*. Microbiol Resour Announc 8:e01473-18. <https://doi.org/10.1128/MRA.01473-18>.
 80. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleischmann RD, Nierman WC, White O, Salzberg SL, Smith HO, Colwell RR, Mekalanos JJ, Venter JC, Fraser CM. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483. <https://doi.org/10.1038/35020000>.
 81. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard

- for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 106: 19126–19131. <https://doi.org/10.1073/pnas.0906412106>.
82. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57: 81–91. <https://doi.org/10.1099/ijs.0.64483-0>.
 83. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol* 19:153. <https://doi.org/10.1186/s13059-018-1540-z>.
 84. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>.
 85. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O, Formsma K, Kubal M, Vonstein V, Stevens R, McNeil LK, Edwards RA, Pusch GD, Reich C, Glass EM, Olsen GJ, Paczian T, Overbeek RA, Meyer F, Vassieva O, DeJongh M, Osterman AL, Disz T, Best AA, Gerdes S, Parrello B, Bartels D, Olson R, Paarmann D. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. <https://doi.org/10.1186/1471-2164-9-75>.
 86. Islam MT, Liang K, Im MS, Winkler J, Busby S, Tarr CL, Boucher Y. 2018. Draft genome sequences of nine *Vibrio* sp. isolates from across the United States closely related to *Vibrio cholerae*. *Microbiol Resour Anounc* 7:e00965-18. <https://doi.org/10.1128/MRA.00965-18>.
 87. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>.
 88. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
 89. Jolley KA, Maiden M. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11:595. <https://doi.org/10.1186/1471-2105-11-595>.
 90. Garg P, Aydanian A, Smith DW, Morris JG, Nair GB, Stine OC. 2003. Molecular epidemiology of O139 *Vibrio cholerae*: mutation, lateral gene transfer, and founder flush. *Emerg Infect Dis* 9:810–814. <https://doi.org/10.3201/eid0907.020760>.
 91. Brock G, Pihur V, Datta S, Datta S. 2008. cValid: an R package for cluster validation. *J Stat Softw* 25:1–22.
 92. R Core Team. 2017. R: a language and environment for statistical computing. R Core Development Team, Vienna, Austria.
 93. Canty A, Ripley BD. 2017. boot: bootstrap R (S-Plus) functions. R package, version 1. R Core Development Team, Vienna, Austria.
 94. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
 95. Chang F, Qiu W, Zamar RH, Lazarus R, Wang X. 2010. clues: an R package for nonparametric clustering based on local shrinking. *J Stat Softw* 33:1–16.
 96. Zhou Z, Alikhan N, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M. 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 28: 1395–1404. <https://doi.org/10.1101/gr.232397.117>.
 97. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 15:524. <https://doi.org/10.1186/s13059-014-0524-x>.
 98. Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681. <https://doi.org/10.1534/genetics.105.048975>.
 99. Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128. <https://doi.org/10.1093/bioinformatics/btl529>.
 100. Hagberg A, Swart P, S Chult D. 2008. Exploring network structure, dynamics, and function using NetworkX, p 11–16. *In* Varoquaux G, Vaught T, Millman J (ed), Proceedings of the 7th Python in science conference (SciPy2008). SciPy, Pasadena, CA.
 101. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. <https://doi.org/10.1101/gr.1239303>.