1 **A *Vibrio Cholerae* Core Genome Multilocus Sequence Typing Scheme to Facilitate the**

2 **Epidemiological Study of Cholera**

3

4 Kevin Y. H. Liang,[a] Fabini D. Orata,[a] Mohammad Tarequl Islam,[a] Tania Nasreen,[a] Munirul

5 Alam,[b] Cheryl L. Tarr,[c] Yann F. Boucher[a#]

6

7 [a]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

8 [b]Infectious Diseases Division, International Centre for Diarrhoeal Disease Research,

9 Dhaka, Bangladesh

10 [c]Centers for Disease Control and Prevention, Atlanta, Georgia, USA

11

12 Running title: *V. cholerae* Core Genome Multilocus Sequence Typing

13

14 [#]Address correspondence to Yann F. Boucher, yboucher@ualberta.ca

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. 2,443 core genes ubiquitous in *V. cholerae* were used to analyze a comprehensive dataset of 1,262 clinical and environmental strains collected from 52 countries, including 70 genomes newly sequenced in this study. We established a sublineage threshold, based on 133 allelic differences, that creates clusters nearly identical to traditional MLST types providing context and backwards compatibility to new cgMLST classifications. We also defined an outbreak threshold, based on seven allelic differences, that is capable of identifying strains that are part of the same outbreak and closely related isolates which could give clues on its origin. Using this scheme, we confirmed the South Asian origin of modern epidemics and identified a strong geographic affinity among sublineages of environmental isolates. Advantages of cgMLST are highlighted by a 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/).

36 **IMPORTANCE**

37 Toxigenic *Vibrio cholerae* of the O1 and O139 serogroups are the causative agent of cholera, an
38 acute diarrheal disease that has plagued the world for centuries, if not millennia. Here, we
39 introduce a core-genome multilocus sequence typing (cgMLST) scheme for *V. cholerae*. Using
40 this scheme, we established an outbreak threshold that enabled us to efficiently identify outbreak
41 related strains and potential sources of introductions. We also defined a sublineage threshold that
42 is equivalent to traditional MLST sequence types, providing context to this new typing method.
43 This scheme is publicly available on PubMLST (https://pubmlst.org/vcholerae) with over 1,200
44 *V. cholerae* isolates collected from around the world. cgMLST outperforms all existing methods
45 in terms of resolution, standardization, and ease-of-use, making this scheme the most suitable
46 method for *V. cholerae* typing and surveillance worldwide.

47

48 **KEYWORDS**

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

51 **INTRODUCTION**

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

68 A typing method for use in global surveillance of pandemic causing pathogens such as *V.*
69 *cholerae* should be efficient and easy to use, with the potential to be applied to all *V. cholerae*
70 strains around the world. Therefore, it must have the capacity to analyze thousands of genomes
71 efficiently and new genomes should be easily typed as they get sequenced. As all cholera
72 outbreaks are caused by a single lineage of *V. cholerae*, the pandemic generating/phylocore
73 genome (PG) lineage, which includes 7th pandemic El Tor, El Tor sister, El Tor progenitor,

74 classical and classical sister clade (5, 18, 19), this method should also be able to differentiate
75 isolates at a fine scale and separate outbreaks caused by genetically similar strains. Such a
76 method will help create a comprehensive database with detailed epidemiological data that will
77 allow for the analysis of future outbreak strains in a global context and guide subsequent
78 epidemiological analyses. Different methods for subspecies level classification and outbreak
79 surveillance have been developed for *V. cholerae*. These methods include serotyping, multilocus
80 sequence typing (MLST) (20, 21), multilocus variable number of tandem repeats (VNTR)
81 analysis (MLVA) (22, 23), and single nucleotide polymorphism (SNP)-based approaches (14).
82 Despite the popularity of these methods, there are important limitations to each.
83 Serotyping based on the presence of cell surface O-antigens is one of the earliest attempts
84 at subspecies level classification of *V. cholerae*. There are now over 200 serogroups of *V.*
85 *cholerae* identified; however, only the toxigenic members of the O1 and O139 serogroups have
86 been found to be responsible for all major documented epidemics and pandemics (24, 25).
87 Serogroup O1 can be further divided into two biotypes (El Tor and Classical) and three serotypes
88 (Inaba, Hikojima, and Ogawa) (2). The lack of resolution within the epidemic strains and the
89 possibility of serogroup conversion (26) limits the use of serotyping in epidemiological studies.
90 MLST provides a standardized classification method that is based on a collection of six to
91 seven well-defined housekeeping genes (27). MLST was used to study a number of cholera
92 outbreaks and allowed the descriptions of its general population structure (28, 29). It is
93 reproducible and provides reliable results; however, it is unable to differentiate between closely
94 related strains which limits its use in outbreak surveillance (30, 31). In addition, direct
95 comparisons between different MLST schemes are difficult, as different schemes utilize different
96 housekeeping genes.

97 MLVA utilizes VNTR regions, which are under less selective pressure than housekeeping
98 genes. This method therefore provides greater resolution than MLST for some bacterial species
99 (32, 33). However, VNTR regions, due to their rapid mutation rate, are more affected by
100 homoplasy where two isolates may share the same MLVA profile due to convergent mutation
101 and not by vertical descent (34). As a result, MLVA may produce clusters that do not necessarily
102 reflect phylogenetic relationships (35). Two common PCR-based methods exist for the typing of
103 VNTR regions, but each have significant limitations (36). The first method is multiplex PCR
104 which can analyze all loci at once, but it is impossible to determine which bands correspond to
105 which loci; therefore, this method only produces a banding pattern for strain identification,
106 which makes it difficult to standardize and communicate results. The second method is the
107 separate amplification of VNTR regions but determining the number of repeats based on
108 amplicon size information alone is difficult if the difference in size is not large enough. In
109 addition, different types of mutations that do not necessarily change the number of repeats can
110 cause a change in amplicon size. Sequencing is needed to confirm MLVA profiles, but repeat
111 regions increase the chances of sequencing errors (37). Due to these limitations, stringent quality
112 control is required for reliable MLVA analysis (38).
113 SNP-based analysis is one of the most common whole-genome-based methods currently
114 being used and was applied to various outbreaks (14, 39, 40). It relies on the identification of
115 conserved SNPs in strains of interest using next-generation sequence reads or assembled
116 genomes. The number of SNPs can then be related to the evolutionary distance between isolates.
117 SNP-based analysis provides reliable results with sufficient resolution for epidemiological
118 studies, but it is sensitive to horizontal gene transfer and recombination events, as each event will
119 result in many SNPs being created. The number of SNPs between two strains, therefore, does not

120 necessarily reflect true phylogenetic relationship. SNPs found in recombinogenic regions should
121 therefore be removed which, depending on the organism of interest, can be anywhere from 30%
122 to 97% of SNPs identified (41, 42). Since recombination and horizontal gene transfer events are
123 common in *V. cholerae* (43–45), SNP-based methods, although suitable in individual
124 epidemiological studies, will be difficult to serve as a universal classification method for *V.*
125 *cholerae*.

126 Core genome MLST (cgMLST), also known as the gene-by-gene approach, overcomes
127 the various limitations of the previously mentioned subtyping methods and was established to
128 serve as a universally applicable standardized typing scheme. Similar to MLST, cgMLST relies
129 on individual gene sequences to differentiate between closely related strains; however, instead of
130 using only six to seven housekeeping genes, cgMLST utilizes hundreds to thousands of core
131 genes, which are commonly found in all strains of a species. By utilizing a much larger portion
132 of the genome, cgMLST provides superior resolution compared to traditional MLST schemes.
133 By combining the expandable and standardized classification method that made traditional
134 MLST favourable with the resolution of whole-genome-based methods, cgMLST is becoming
135 more popular in epidemiological and ecological studies (46–52). This method has the added
136 advantage of backwards compatibility with all MLST schemes, meaning it is possible to
137 determine MLST profiles of any isolates based on their cgMLST profiles, since cgMLST, by
138 definition, would include all housekeeping genes. This allows for a 1:1 mapping of any
139 previously established MLST scheme to the cgMLST scheme, helping consolidate already
140 existing classification information.

141 Another major benefit of cgMLST is that, much like traditional MLST methods, it is
142 possible to establish different clustering thresholds to define important groups. Clonal complexes

143 are examples of clustering thresholds established by MLST schemes, where each clonal complex
144 corresponds to a cluster of isolates that share at most one allelic difference across all seven genes
145 sequenced. Some important clonal complexes were shown to correspond to either groups
146 established by a previous typing method (53) or major outbreak strains (54). However, cgMLST
147 offers even greater flexibility than MLST in this regard, given the number of loci considered.
148 With small clustering thresholds where groups are created based on the sharing of a large
149 number of alleles, it is possible to identify closely related strains useful in epidemiological
150 studies. On the other hand, with larger clustering thresholds, it is possible to identify lineage- or
151 even sublineage-level differences to study large scale patterns and answer broader ecological
152 questions. The benefits of defining clustering thresholds with cgMLST schemes have already
153 been demonstrated in other human pathogens, such as *Brucella melitensis* (50), *Campylobacter*
154 *jejuni* (49) , *Clostridium difficile* (51), *Enterococcus faecium* (48), and *Listeria monocytogenes*
155 (47).

156 In this study, we introduce a cgMLST scheme for the genome-wide typing of *V. cholerae*
157 and demonstrate its universality and efficacy by applying it to known cholera outbreaks around
158 the world. The advantages of cgMLST are presented by comparing the scheme with previously
159 established classification methods. Furthermore, we have produced a 1:1 mapping of the
160 cgMLST scheme against two MLST schemes for *V. cholerae* (20, 21), allowing for the
161 consolidation of existing classification information. The cgMLST scheme, genome sequences
162 used in this study, and relevant epidemiological information are publicly available on PubMLST
163 (https://pubmlst.org/vcholerae/), which allows for the automatic annotation and subsequent
164 analyses of hundreds of newly uploaded *V. cholerae* genomes in a global context. This increase
165 in efficiency, standardizability, and resolution compared to current methods makes cgMLST the

166 most suitable classification scheme for large scale *V. cholerae* surveillance. By applying this
167 scheme to our collection of over 1,200 isolates collected around the world, it was possible to
168 establish outbreak and sublineage thresholds which not only allowed us to validate the South
169 Asian origin of many modern epidemics as proposed in previous studies (5, 55, 56) but also
170 identify a strong geographic signal among environmental strains, where isolates from the same
171 sublineage are also from the same geographic region; this is a pattern that is not seen in clinical
172 isolates.
173

174 **MATERIALS AND METHODS**

175 **Dataset description**

176 On November 6th 2018, 1,172 *V. cholerae* genomes consisting of 800 draft and complete
177 genomes and 372 sequence read archives (SRAs), available from both publicly available
178 database and private collection, were selected as our dataset. One hundred sixteen SRAs from a
179 recent study on the Yemen cholera outbreak (9) were subsequently added as an independent
180 evaluation of the cgMLST scheme (Table S1). The 488 SRAs were assembled using skesa (57)
181 or the CLC Genomics Workbench 7 (https://www.qiagenbioinformatics.com) using default
182 parameters. This total dataset of 1,288 included twenty-six genomes with less than 90% of the
183 core genes, which were identified using USearch (58) based on RAST (59) annotations. These
184 twenty-six genomes were removed from subsequent analyses resulting in a final dataset of 1,262
185 genomes collected from 52 countries and spanning 82 years from 1937 to 2018 (Table S2).
186 These include a historical collection from the 6th cholera pandemic, clinical isolates from
187 outbreaks in various countries (e.g., Bangladesh, India, Haiti, Yemen, the Democratic Republic

188 of Congo, and Russia), and environmental isolates from different parts of the world (e.g., USA,
189 Mexico, and Australia).
190
191 **Gene identification and allele assignments**
192 Instead of using the full dataset of 1,288 genomes, we selected a subset of high-quality genomes
193 because core gene identification is highly dependent on the initial dataset and the inclusion of
194 poorly assembled and/or sequenced data will reduce the number of core genes identified (47).
195 Firstly, 800 already assembled draft or complete genomes were selected for core gene
196 identification. Low-quality assemblies were then eliminated by removing genomes with less than
197 40× coverage and/or N50 values less than 40 kb. From a previously established cgMLST scheme
198 for *L. monocytogenes*, 40× coverage and 20 kb N50 value were used as cutoff thresholds, as
199 genomes that do not meet these criteria resulted in a low proportion of loci being called (47). The
200 40× coverage cutoff was adopted for this study; however, because the average *V. cholerae*
201 genome size (~ 4 Mb) is larger than the average *L. monocytogenes* genome (~ 3 Mb), 40 kb was
202 instead selected as the N50 cutoff. The use of these cutoffs resulted in the removal of 82
203 genomes.
204 The remaining 718 genomes were annotated using RAST (59) and USearch (58) and a
205 tentative set of core genes that were on average present in 99% of the genomes were selected. An
206 additional 13 genomes were removed, as they lacked more than 90% for the core genes, leaving
207 us with a dataset of 705 high-quality genomes. However, an additional 26 genomes were
208 subsequently removed for the core gene analysis as it has been previously suggested that they
209 form a highly divergent lineage within the *V. cholerae* (60–62), ensuring that the dataset used for
210 core gene identification consists only of unambiguously *V. cholerae* isolates (also as verified by

211 average nucleotide identity (63) and digital DNA–DNA hybridization (64) between genomes
212 (60–62)). Completeness and potential contamination of all remaining 679 genomes were also
213 independently evaluated by checkM, which estimates these values based on the presence and
214 number of copies of a set of pre-defined single copy marker genes (65) (Table S4). All genomes
215 were, according to the criteria established by checkM, nearly complete (≥97%) with medium to
216 low levels of contamination (<7%) (65).

217 Each orthologous gene was compared against the *V. cholerae* N16961 reference genome
218 using BLASTN (66) to determine gene function. Any gene family with no homolog in N16961
219 or classified as pseudogenes on the NCBI GenBank database were removed, meaning N16961
220 was 100% complete for the cgMLST scheme. Any gene that was present in more than one copy
221 in any of the initial 679 genomes was also removed, as they were considered paralogous. Thus,
222 in this context, core genes are defined as being present in at least 90% of the 679 high-quality
223 assembled genomes in a single copy. By choosing a relaxed cutoff of 90% completeness, we
224 accounted for missing genes due to sequencing, annotation, or assembly errors while ensuring
225 there is sufficient resolution to differentiate between closely related strains, with at least 2,199
226 loci remaining for classification purposes. The final cgMLST scheme utilizes a set of 2,443
227 single-copy core gene loci, which is 2,425,296 bp in size and covering approximately 61% of the
228 genome. The list of core genes is available on PubMLST (https://pubmlst.org/vcholerae/).

229 Automated scripts in BIGSdb (67) were used to perform allele calls and assignments for
230 all 1,262 isolates (Table S2). Allele calls were made only for complete coding sequences with a
231 minimum of 70% similarity and 70% length coverage at the nucleotide level, as previously
232 described (47). Default settings were used for all other parameters.

233

234 **Core-genome sequence type (cgST) assignment**

235 cgST, which was defined as a unique combination of alleles of all loci included in the scheme,
236 was assigned for all isolates, excluding those from the Yemen outbreak study (8), with an in-
237 house script, as previously described (68). Briefly, missing loci were replaced with the most
238 common allele when assigning cgSTs, allowing for a conservative estimate of diversity (68). The
239 116 isolates from the Yemen cholera outbreak study (9) were annotated automatically by
240 uploading them to PubMLST. PubMLST treated missing alleles as 'N'. cgSTs were assigned to
241 each allele profile, treating 'N' as a regular allele designation. However, different from typical
242 allele designations, 'N's can represent any allelic sequence; therefore, some isolates may contain
243 multiple cgST designations, all of which are possibly true cgSTs. For isolates with more than one
244 cgST suggested by PubMLST, postprocessing was done using an in-house script to identify the
245 most likely cgST, which was determined by assuming missing loci contained the most common
246 allele (Table S2). It is expected that as genome sequencing becomes more reliable, higher quality
247 genomes will be available and any missing data can be updated as needed.

248

249 **MLST scheme and sequence type (ST) assignments**

250 Two MLST schemes developed for *V. cholerae* were mapped to this cgMLST scheme. The first
251 MLST scheme developed in 2013 by Octavia *et al.* (20) was used to study global population
252 structure of non-O1/non-O139 *V. cholerae* and is currently hosted on PubMLST. All isolates
253 uploaded to PubMLST were automatically annotated with this scheme. Any missing data in this
254 scheme was ignored and no ST designation was assigned. The second MLST scheme developed
255 in 2016 by Kirchberger *et al.* (21) was used to study the population structure of environmental *V.*
256 *cholerae* in a region on the US East Coast. The second MLST scheme is not currently hosted on

257 PubMLST, but because the housekeeping genes in this scheme are also found in the cgMLST
258 scheme, a similar in-house script used in cgST assignments was used to assign ST designations.
259 Therefore, all isolates in this study were assigned three designations when possible – two ST
260 designations based on the two previous MLST schemes (20, 21) and one cgST designation based
261 on the cgMLST scheme from this study.
262
263 **Outbreak and sublineage clustering thresholds**
264 A clustering threshold was defined as the maximum number of allelic differences found within a
265 cluster. All clusters were produced based on the single-linkage clustering method, which meant
266 an isolate belonged to a cluster if it linked with any isolate within that cluster. Two metrics were
267 used as general guidelines for determining clustering thresholds. The first metric used was the
268 Dunn Index (DI), which measured clustering efficiencies (69). Briefly, the DI was highest for a
269 network (i.e., the network has the best clustering efficiency) when the intra-cluster distances
270 were minimized, and the inter-cluster distances were maximized. Since isolate distances were
271 measured based on allelic differences, a high DI resulted in clusters where isolates were more
272 closely related to those found within the same cluster than those found in a different cluster. The
273 DI was calculated using the R packages 'clvalid' and 'boot' with 100 bootstrap replicates for
274 each threshold and graphed using the R package 'ggplot2' (70–73).
275 The second metric used was the Adjusted Rand Index (ARI), which measured the level of
276 similarity between two networks when clustering the same set of isolates by measuring the
277 amount of agreements (i.e., the number of pairs that were grouped either as being in the same
278 cluster or different cluster in both networks) and disagreements (i.e., the number of pairs that
279 were grouped together in one network but grouped separately in another) (74). The values ranged

280 from -1 (i.e., two networks are exactly opposite) to 1 (i.e., two networks are identical). ARI was
281 used to determine the level of similarity between various clustering thresholds and the MLST
282 schemes. ARI was calculated using the R package 'clues' and graphed using 'ggplot2' (71, 73,
283 75).
284
285 **Minimum spanning tree (MST)**
286 All MSTs, unless otherwise specified, were constructed using GrapeTree MSTv2 (76). Loci with
287 missing data were included in the profile as "–". GrapeTree provided a novel algorithm that
288 accounted for missing data when constructing an MST, an important feature since missing data is
289 common in whole-genome- and core-genome-based analyses. GrapeTree is currently integrated
290 within PubMLST, which allows for quick visualization of the dataset with any provenance data.
291
292 **Phylogenetic analysis**
293 Parsnp v1.2 (77) was used to reconstruct the phylogenetic tree using *V. cholerae* N16961 as the
294 reference genome. The -x flag was used to enable filtering of SNPs in recombinogenic regions as
295 identified by PhiPack (78). Default settings were used for all other parameters. The phylogenetic
296 tree included 1,146 genomes (all genomes except for the 116 isolates from the recent Yemen
297 cholera outbreak study (9)). Since all isolates sequenced for the latter study belonged to the 7th
298 pandemic El Tor lineage, it would have had limited impact on the overall structure of the tree.
299 The phylogeny was visualized and annotated using iTOL (79).
300
301
302

303 **Biogeographical analysis of environmental isolates**

304 All isolates that were not part of the PG lineage (18, 19) were first clustered based on the
305 sublineage threshold using the python package 'networkX' (80). Missing alleles were replaced
306 with the most common allelic designation when calculating pairwise differences to establish a
307 more conservative estimate of diversity. The network was then visualized using Cytoscape (81).

308

309 **Data availability**

310 All previously sequenced *V. cholerae* genomes were obtained from the NCBI GenBank database.
311 65 additional genomes sequenced in this study were deposited on the same database. Table S6
312 lists all the accession numbers for all the genomes used in this study. In addition, all genome
313 sequences, allelic profiles, cgST designations, ST designations, and relevant epidemiological
314 data are publicly available on PubMLST (https://pubmlst.org/vcholerae/).

315

316 **RESULTS AND DISCUSSION**

317 **A high-resolution typing scheme for pandemic *V. cholerae***

318 The highest level of resolution of any cgMLST scheme is defined by cgSTs, where a unique
319 cgST represents a unique allelic profile. Isolates that belong to the same cgST are expected to be
320 phylogenetically very closely related, as although they may not have the exact genomic
321 sequence, they do have the same sequence for all 2,443 core gene loci. We identified a total of
322 1,026 cgSTs from 1,262 genomes collected from 52 countries. Even with our extensive dataset,
323 we have yet to sample anywhere close to the total predicted cgST diversity for the global *V.*
324 *cholerae* population (Fig. S1). All isolates were given at least one cgST designation and up to
325 two MLST ST designations based on two previously established MLST schemes (20, 21) (Table

326 S2). MLST STs are defined based on the unique combination of all loci of a particular MLST
327 scheme, which ideally uses six to seven well-defined housekeeping genes. Twelve STs are
328 exclusively present in the 7th pandemic El Tor lineage using the traditional MLST schemes (20,
329 21), while 560 cgSTs are uniquely present in this group based on the cgMLST scheme (Table
330 S2). As the El Tor lineage is responsible for most cholera outbreaks around the world since the
331 beginning of the 7th pandemic (82), this superior ability to resolve between closely related strains
332 in the 7th pandemic El Tor lineages makes cgMLST more suitable in outbreak surveillance than
333 traditional MLST schemes.
334
335 **Backwards compatibility with previous subspecies classification methods**
336 Much like how cgSTs are important in studying closely related strains typical in outbreaks, it is
337 also important in establishing a standardized nomenclature at a higher level to answer broader
338 ecological questions. Here, we propose a sublineage definition for *V. cholerae* based on our
339 cgMLST scheme.
340 Pairwise allelic differences calculated between all isolates shows three major peaks (Fig.
341 1A). The first peak ends at 40 allelic differences, and the second peak ends at 133 allelic
342 differences (Fig. 1B). The last peak begins at 2,200 allelic differences (Fig. 1A), which is
343 expected due to mutational saturation (i.e., every single allele in the scheme is different between
344 the two distantly related strains being compared). Both breaks (i.e., 40 or 133 allelic differences)
345 could represent a potential sublineage delineation. To choose between the two thresholds, the
346 clustering efficiency is measured by calculating the Dunn Index (DI) (69). Since cluster distances
347 are measured by allelic differences, the network with the best clustering efficiency (i.e., the
348 highest DI) will also produce clusters that best represent biological relationships, as isolates are

349 more closely related to themselves than to isolates from other clusters. A DI was calculated for
350 each clustering threshold in the range of 1 to 1,000 allelic differences with 100 bootstrap
351 replicates (Fig. 2). As the clustering threshold defines the maximum number of allelic
352 differences within a cluster, the smaller the threshold, the more closely related the isolates are
353 within a cluster. It is clear that DIs in the range of 0 to 50 allelic differences are significantly
354 lower than the DIs in the range of 100 to 350 allelic differences, with 133 being a clear local
355 maximum. Since 133 allelic differences has the best clustering efficiency and also represents a
356 natural break where most isolate pairs have either less than or much greater number of allelic
357 differences (Fig. 1B), it was therefore chosen as the sublineage threshold.

358 Because cgMLST includes all housekeeping genes, information from the two MLST
359 schemes previously developed for *V. cholerae* (20, 21) can now be consolidated with the
360 cgMLST scheme by creating a 1:1 cgMLST to MLST map. To evaluate the similarities between
361 the sublineage threshold and the MLST schemes, we created a minimum spanning tree (MST)
362 for all Bangladesh isolates (n = 255) showing only edges with 133 allelic differences or fewer
363 (Fig. 3A and Fig. S2). Each cluster therefore represents a single sublineage. Bangladesh was
364 chosen to compare cgMLST and MLST as it is the most extensively sampled country both in
365 terms of clinical and environmental isolates in our dataset. Using this dataset, the chosen
366 sublineage threshold produces clusters that closely resembles traditional MLST STs. Based on
367 the 2013 MLST scheme (20), each sublineage corresponds to exactly one ST (Fig. S2), whereas
368 there is only one sublineage that contains two STs based on the 2016 MLST scheme (21) (Fig.
369 3A). All but two isolates belong to ST1; N16961 and A19 belong to ST290, which differs from
370 ST1 at only one of seven MLST loci (Table S5). The reason these two isolates are of a different
371 MLST ST could only be partly explained; they were isolated at an earlier time point (1970s near

372 the start of the 7th pandemic (83)) than most of the remaining isolates, which were isolated from
373 1991 onwards (Table S2).

374 It is impossible to visually evaluate similarities between two MSTs with over 1,200 nodes
375 each simply due to the sheer volume of data. ARI was therefore used as a metric to determine
376 network similarities (74). In order to determine whether the sublineage threshold (i.e., 133 allelic
377 differences) is indeed the best match to traditional MLST schemes, we chose 11 clustering
378 thresholds distributed across the range of 1 to 1,000 allelic differences (Fig. 2) to compare with
379 the MLST schemes. These additional thresholds are chosen as they have a relatively high DI
380 compared to their immediate neighbours. More data points were chosen in the range of 105 to
381 330 allelic differences, as it was expected that thresholds in this range will best match the
382 traditional MLST schemes. Interestingly, all thresholds in that range had comparable ARIs (Fig.
383 3B), indicating that all of them, including the sublineage threshold, produce clusters similar to
384 the MLST scheme. This would suggest that there can be a large range of diversity within a single
385 MLST ST where isolates can have anywhere from 0 (i.e., have the same cgST) to 330 allelic
386 differences. Although clustering thresholds between 105 to 330 allelic differences produce
387 similar clusters to a traditional MLST scheme, 133 allelic differences was chosen as the
388 sublineage threshold as it has the best clustering efficiency (Fig. 2) and it represents a natural
389 breakpoint in the currently sampled population (Fig. 1B).

390 A phylogenetic tree of 1,146 isolates was used to assess the phylogenetic support of the
391 sublineage threshold across different *V. cholerae* strains (Fig. 4). This tree includes all *V.*
392 *cholerae* isolates within our dataset with the exception of the 116 clinical isolates published
393 recently from the Yemen cholera outbreak study (9), which all belongs to the 7th pandemic El
394 Tor lineage. The strains within the PG lineage, which include the 7th pandemic El Tor, El Tor

395 progenitor, El Tor sister, Classical, and Classical sister groups (18, 19), are closely related with
396 little genetic variation. These lineages are therefore collapsed in the phylogenetic tree as the
397 relationships between them are not well resolved. All sublineages formed monophyletic clades,
398 although in some cases the most basal branch is of a different sublineage (e.g., *V. cholerae*
399 strains T5 or 506315) creating paraphyletic clades. Ideally, each sublineage would correspond to
400 exactly one full monophyletic clade. The reason this is not seen is likely the lack of sampling,
401 leading to the grouping of relatively distantly related isolates together in the same clade. Further
402 sampling will likely resolve these cases into two separate monophyletic clades. Out of 1,262
403 isolates, we identified 291 sublineages, 19 of which belong exclusively to the PG lineage. Of the
404 291 sublineages, 223 are singletons. Based on the rarefaction curve, much like cgSTs, the total
405 sublineage diversity of *V. cholerae* is far from being sampled (Fig. S1).
406 The concept of sublineage has been applied to numerous pathogens and as such were
407 each defined differently depending on the pathogen in question. Some have defined sublineages
408 based on natural breaks in genetic similarities (47), while others may use sublineage to refer
409 specifically to traditional MLST STs (84) or even finer level of resolution below the MLST ST
410 level based on whole genome analyses (85). There is, however, one unifying feature of all
411 sublineage definitions – that they all refer to monophyletic clades. Sublineages are defined in this
412 study based on natural breaks in allelic differences calculated from cgMLST profiles and were
413 put into context by comparing with two traditional MLST schemes. We have shown that our
414 definition of sublineage results in monophyletic clades (Fig. 4), but also corresponds to any
415 traditional MLST ST designation. This sublineage definition will therefore play a crucial role in
416 consolidating information from all previous MLST analyses.
417

418 **A universal south Asian origin for modern cholera outbreaks**

419 With the continual improvements of next-generation sequencing techniques, whole genome
420 sequencing is expected to become a standard practice or even the first identification tool used in
421 clinical and epidemiological studies. It is therefore critical to develop rapid typing scheme for
422 genome sequence data that has the power to inform us about the relationship of a novel isolate
423 with known strains. This is done here by defining what we term an 'outbreak threshold' based on
424 the cgMLST scheme, which can identify outbreak related strains and potential sources of
425 introduction. The outbreak threshold is expected to be less than 40 allelic differences as isolates
426 from the same outbreak are very closely related (9, 14). There is a minor discontinuity at seven
427 allelic differences where most isolate pairs have either less than or more than this number of
428 allelic differences (Fig. 1B). Looking at the DI, the local maximum in the range of 0 to 50 occurs
429 at seven allelic differences as well (Fig. 2), making this cutoff an interesting candidate for an
430 outbreak threshold. When applying the outbreak threshold to a full dataset containing all
431 sequenced *V. cholerae* genomes meeting the minimum quality threshold, major clusters were
432 examined to evaluate the ability of cgMLST to identify strains that are part of the same outbreak.

433 One of the major outbreak clusters identified, with no prior information required,
434 contains the Haiti and the Yemen outbreaks, which are the two best documented cholera
435 outbreaks in modern history (8, 9, 14, 16, 86). Isolates collected from these outbreaks form a
436 single cluster with the Dominican Republic, Eurasian (India, Russia, Nepal, and Ukraine), and
437 African (Tanzania, Kenya, and Somalia) isolates (Fig. 5A). The Dominican Republic isolates are
438 closely related to the Haiti outbreak strains. Given the close proximity of the two countries, co-
439 located on the island of Hispaniola, it was expected that isolates from Haiti would eventually
440 spread to the Dominican Republic (14). The 7th pandemic El Tor lineage spread across the world

441 from South Asia in three separate waves (83). The third wave, being the most recent distribution
442 event, has been claimed to be responsible for the outbreaks in Haiti and Yemen (9). It is
443 therefore not surprising to see Haiti and Yemen form a single cluster with India at its center.
444 Nepal is the known source of introduction for the Haiti outbreak in 2010 (16), and comparisons
445 with over 1,200 *V. cholerae* isolates from all over the world still show the Nepalese isolates are
446 indeed the closest relatives to the Haitian isolates (Fig. 5A).
447 Cholera is still endemic in Africa (10) and caused several major reported outbreaks in
448 different countries over last few decades (87) including Mozambique (23, 88) and Zimbabwe
449 (89). Another major outbreak cluster contains most of the Mozambique isolates in our dataset
450 together with two Zimbabwe isolates (strains CP1038(11) and 2011EL-1137) and one USA
451 isolate (2009V-1116) (Fig. 5B). Based on cgMLST analysis, it is evident that these two
452 Zimbabwe isolates are closely related to the Mozambique isolates differing at four or less alleles.
453 The close proximity of the two countries suggests these are likely travel-associated cases.
454 Although outbreaks involving the Mozambique isolates (23) and the Zimbabwe isolates (56, 90)
455 have been independently studied, the link between these isolates have not been shown before.
456 Global cgMLST analysis is therefore an invaluable tool as it allows for the identification of links
457 between independent studies. However, with only two Zimbabwe isolates in the dataset,
458 additional sampling in this region is required to understand the epidemiology of this outbreak.
459 According to the NCBI BioSample database, strain 2009V-1116 was collected by the Centers for
460 Disease Control and Prevention in 2009 and is associated with travel to Pakistan. Since the 7th
461 pandemic El Tor lineage has been circulating in Asian and Middle Eastern countries for a long
462 time (91), it is possible that, at least within our dataset, the Mozambique isolates are the closest
463 relative to this specific Pakistan strain.

464 **Confirmation of an African connection for the Yemen outbreak**

465 The Yemen cholera outbreak began in October 2016 with eleven confirmed cases
466 (http://www.emro.who.int/pandemic-epidemic-diseases/cholera/cholera-cases-in-yemen.html).
467 By January 2017, there has already been over 10,000 cholera cases with 99 associated deaths
468 (http://www.emro.who.int/pandemic-epidemic-diseases/cholera/weekly-update-cholera-cases-in-
469 yemen-15-jan-2017.html), and by the end of that year, there were over 900,000 cholera cases
470 (http://www.emro.who.int/pandemic-epidemic-diseases/cholera/outbreak-update-cholera-in-
471 yemen-19-december-2017.html). The Yemen cholera outbreak continues on today as the largest
472 cholera outbreak in modern history. As isolates from this outbreak were only recently made
473 available (9), they were not part of the initial dataset for the cgMLST scheme development.
474 These isolates were added and analyzed on PubMLST after the scheme had been established.
475 This set of isolates therefore serves as an independent test of the universality and applicability of
476 the cgMLST scheme. To determine the potential origin of the Yemen outbreak and its
477 phylogenetic relationships with existing *V. cholerae* strains, the Yemen isolates were compared
478 with other 7th pandemic El Tor isolates from Asian and African countries (Table S1). All allele
479 designations and cgST assignments were done automatically on PubMLST. MST was built using
480 these isolates and all connections with seven and fewer allelic differences are represented as
481 solid lines (Fig. 6). Isolates connected by solid lines therefore belong in the same outbreak
482 cluster as defined by the outbreak threshold of seven allelic differences. Isolates from Yemen,
483 Kenya, and Haiti all cluster with the central Indian isolates with seven or fewer allelic
484 differences; however, the closest relatives to the Yemen isolates are those from Kenya with four
485 or fewer allelic differences (Fig. 6). The Indian isolates are the next closest connection but there
486 is no direct linkage between these and the Yemen isolate. This pattern is consistent with the work

487 of Weill *et al.* (9), where they suggested that the Yemen outbreak strains may have come from
488 East Africa which itself came from South Asia based on SNP-base phylogenetic analysis and
489 Bayesian evolutionary analysis (9).

490 Unlike the limited samples available from African cholera outbreaks, the Haiti and
491 Yemen outbreaks are significant cases for epidemiological investigations because *V. cholerae*
492 has been heavily sampled from these countries as well as surrounding regions. Two major
493 limitations in genomic epidemiology have been the lack of a universal classification scheme and
494 a comprehensive database; however, this is no longer the case in the genomic era as sequencing
495 technology is becoming increasingly more accessible (8). A genomic approach, as shown here, is
496 able to produce accurate predictions of potential origins of outbreaks and provides us with
497 sufficient resolution to accurately track the spread of the disease. Therefore, genomic analysis
498 should be the first step in any epidemiological study as not only will it help guide subsequent
499 analyses and investigations, but consistently sequencing new genomes will also help expand and
500 refine the current global *V. cholerae* genome database.

501

502 **Increased resolution for the history of cholera in Mozambique: comparing cgMLST to**
503 **MLVA**

504 The 7th pandemic reached Africa in 1970 and cholera appeared in Mozambique at roughly the
505 same time (55). Since its introduction, cholera has been endemic in that country and has
506 continued to cause multiple outbreaks (23). A popular tool for outbreak investigation is MLVA
507 (32, 38), which was recently used to study *V. cholerae* strains collected in Mozambique over
508 multiple years (23). MLVA is a subspecies typing method similar to MLST in concept; however,
509 instead of using gene sequences, it utilizes VNTRs. As VNTR mutates at a faster rate than

510 conserved genes, it has been shown that MLVA provides greater resolution than MLST for some
511 species (32, 33). To establish a direct comparison between our cgMLST scheme and this MLVA
512 scheme, we examined the MSTs created by both methods focusing on only shared isolates (Fig.
513 7A and 7B). The MLVA identified 26 profiles forming two clonal complexes and four singletons
514 (Fig. 7A) (23). A similar population structure is seen with the cgMLST analysis (Fig. 7B),
515 including the four singletons identified in the MLVA. The central node in the cgMLST MST
516 consists mostly of isolates with MLVA profile 8,4,6,18,21, similar to the central node in the
517 MLVA MST (23). The two clonal complexes (CC) identified in the MLVA MST are also
518 identified in the cgMLST MST with the smaller CC2 being at least four allelic differences away
519 from the larger CC1.

520 Although there are a few MLVA types that were grouped into a single cgST, such as
521 cgST1 and cgST114, indicating cgMLST was unable to resolve the differences in these MLVA
522 types, there are many MLVA types such as profile 2,4,6,18,21, profile 7,4,6,16,22, profile
523 9,4,6,18,24, and profile 8,4,6,18,22 that were split into multiple cgSTs. Overall there are 48
524 cgSTs as opposed to only 26 MLVA types, showing that cgMLST provides better resolution
525 overall than MLVA. The cgMLST analysis overlaid with isolation dates shows that the
526 Mozambique *V. cholerae* strains are highly clonal and strains from the same cgST can cause
527 outbreaks over multiple years (e.g., cgST114 and cgST94) (Fig. 7C), which corroborates the
528 claim made in the initial MLVA study that the same MLVA type can be seen over multiple years
529 (23). In addition to increased resolution, cgMLST also produces more reliable and reproducible
530 results than MLVA, as it eliminates errors associated with the detection of VNTR regions using
531 PCR or sequencing-based methods. For the same reason that MLST is less affected by

532 convergent evolution compared to MLVA (35), cgMLST is also less affected by convergent
533 evolution.
534
535 **Standardizing the genotypes responsible for the Haiti 2010 cholera outbreak: comparing**
536 **cgMLST and SNP-based analyses**
537 One of the largest cholera outbreaks in modern history occurred in Haiti following the
538 devastating earthquake in 2010 (8, 92). Prior to this outbreak, there were no documented cholera
539 cases in Haiti (14, 18). Since the initial introduction, *V. cholerae* now remains endemic in Haiti
540 and is responsible for thousands of cholera cases annually (92). Multiple studies have strongly
541 suggested that the Haitian strains were in fact imported from Nepal (by the UN Nepalese troops)
542 and the outbreak occurred as a result of both inappropriate sanitary practice and the lack of
543 screening of the UN troops upon their arrival in Haiti (8, 15, 16, 92).
544 A SNP-based approach was used to study the evolutionary dynamics of *V. cholerae* in
545 Haiti (14). This technique relies on the identification of SNPs in draft or closed genomes. The
546 primary benefit of this method is that assembly and annotation are not required. It is also capable
547 of resolving closely related strains using whole-genome data. However, SNP-based methods are
548 highly influenced by recombination events (43, 44) and quality filter parameters chosen (93).
549 To establish a direct comparison between the cgMLST scheme and SNP-based analysis,
550 we focused on MSTs of only Haitian outbreak isolates (Fig. 8). All Haitian isolates are closely
551 related according to the cgMLST scheme, sharing at most four allelic differences with each other
552 (Fig. 8A). The Haitian and Nepalese isolates, therefore, also belong to the same sublineage (SL6)
553 which is consistent with the fact that these isolates belong to the same MLST ST (either ST1 or
554 ST69 based on the 2016 or 2013 MLST scheme, respectively (20, 21)) (Table S2). The overall

555 population structure is similar between the two methods where we have SNP ST1 as the center of
556 the MST and ST2 and ST3 extending from that likely ancestral genotype (Fig. 8). SNP ST1,
557 ST2, and ST3 can be split into 11, 2, and 3 different cgSTs, respectively (Fig. 8A). There is only
558 one case, cgST66, where it contains isolates from SNP ST1 and ST3. Overall, cgMLST was able
559 to differentiate 39% of the isolates while the SNP-based analysis differentiated 35%, showing
560 comparable level of resolution. As expected, both the cgMLST and the SNP-based analyses
561 showed that the Haiti outbreak is highly clonal where most isolates belong to the same cgST or
562 SNP ST (14). However, one important advantage of cgMLST over SNP-based analysis is that the
563 former can be easily standardized because it relies on a predefined set of core genes. Based on
564 these standardized genes, we can establish a systematic nomenclature system. This makes
565 cgMLST more suitable than SNP-based method as a universally applicable classification system
566 for epidemiological studies and research worldwide.
567

568 **Environmental isolates differ from clinical strains by their diversity and their associations**
569 **with specific geographical locations**

570 To look at the geographic signal of *V. cholerae*, we eliminated all clinical isolates and those that
571 belong to the PG lineage (18, 19). This is because the geographic signal of clinical strains can be
572 skewed, as pathogenic strains can travel long distances in a short period of time through
573 association with human hosts. The geographical analysis was therefore performed only with
574 environmental isolates.

575 Along with all the publicly available environmental strains that are not part of the PG
576 lineage, there are a total of 195 isolates spanning 9 countries. After grouping the isolates at the
577 sublineage level (i.e., each cluster have at most 133 allelic differences), it could be noted that all

578 isolates from the same sublineage also shared a country of origin, with the exception of strains
579 692-79 and 857 (Fig. 9), which are from the USA and Bangladesh, respectively. Phylogenetic
580 analysis shows these isolates to be closely related to strain A215, a clinical isolate from the USA
581 (Fig. 4). All three strains contain the *toxR* gene, a toxin transcriptional regulator common in
582 pathogenic *V. cholerae* (94), as well as genes encoding for the Mannose-sensitive hemagglutinin
583 pilus, the RTX toxin, and hemolysin (*hlyA*), all of which are putative virulence factors for this
584 species. In addition, strains A215 and 857 also harbor the zona occludens toxin gene. Similar
585 toxin gene contents among these three isolates and close phylogenetic relationships suggest that
586 strains 692-79 and 857 may also be potentially pathogenic and capable of surviving inside a
587 human host. This provides evidence that although clinical isolates can spread across the world
588 rapidly and closely related isolates can be from very different parts of the world, environmental
589 isolates show a strong geographic signal at least at the sublineage level. Currently, our dataset
590 contains a relatively small number of environmental isolates that are not part of the PG lineage.
591 With large-scale environmental sampling, it will be possible to determine with greater accuracy
592 the evolutionary rate and distribution pattern of *V. cholerae* in the environment. cgMLST will
593 become an invaluable tool in dealing with these big datasets, as it provides an efficient and
594 standardized method of classification.
595

596 **Conclusion**

597 With an extensive collection of over 1,200 *V. cholerae* isolates, we developed a cgMLST scheme
598 based on 2,443 core genes. The universality and applicability of the scheme have been tested by
599 looking at well-documented outbreaks around the world. We also established a sublineage-level
600 definition based on 133 allelic differences as part of our standardized classification scheme. It

601 was determined by comparisons with previous MLST schemes that the sublineage classification
602 can be used as a proxy for MLST. In addition to the sublineage definition, we determined an
603 outbreak threshold based on seven allelic differences that groups isolates from the same outbreak
604 together with strains from the potential source of introduction. This threshold creates clusters that
605 are consistent with known epidemiological data when applied to the Haiti and Yemen cholera
606 outbreaks, two of the best-documented cholera outbreaks in modern history. Based on this
607 scheme and utilizing the established sublineage and outbreak thresholds, we were also able to
608 confirm the South Asian origin of modern cholera outbreaks. Furthermore, focusing on only
609 environmental isolates that are not part of the PG lineage (18, 19), a strong geographic signal not
610 seen in clinical strains was identified. Lastly, this scheme is fully implemented on PubMLST
611 (https://pubmlst.org/vcholerae/) for public access. All newly available genomes uploaded to
612 PubMLST will be annotated automatically and a cgST designation will be assigned to isolates
613 with less than 100 missing loci. Relevant epidemiological data is already included, and the
614 variety of analytical and visualization tools integrated on PubMLST allows for quick analysis of
615 the dataset in a global context. This scheme will be an important tool for future large-scale
616 epidemiological and biogeographical research.
617

618 **ACKNOWLEDGMENTS**

619 KYHL, FDO, and YFB conceived the experiments. KYHL, FDO, and MTI performed all data
620 collection, bioinformatics and other analyses. FDO and TN performed genome sequencing and
621 analyses of Bangladesh isolates. TN, MA, and CLT provided strains used in this study.
622 We acknowledge the helpful feedback provided by Dr. Keith Jolley (University of
623 Oxford) regarding the development of the cgMLST scheme, as well as the implementation of

624 this scheme on PubMLST. We thank Monica Im (Centers for Disease Control and Prevention)
625 for assistance with obtaining some whole-genome sequences. This work was supported by the
626 Natural Sciences and Engineering Research Council (NSERC) of Canada (to YFB); the
627 Integrated Microbial Biodiversity program of the Canadian Institute for Advanced Research (to
628 YFB); federal appropriations to the Centers for Disease Control and Prevention through the
629 Advanced Molecular Detection Initiative (to CLT); graduate student scholarships from Alberta
630 Innovates – Technology Futures (to KYHL, FDO, and TN), NSERC (to KYHL and TN), and the
631 Bank of Montréal Financial Group (to FDO). MA of icddr,b acknowledges the governments of
632 Bangladesh, Canada, Sweden, and the United Kingdom for providing core/unrestricted support.
633 The funders had no role in study design, data collection and interpretation, or the decision to
634 submit the work for publication. The findings and conclusions in this report are those of the
635 authors and do not necessarily represent the official position of the Centers for Disease Control
636 and Prevention.

637 **REFERENCES**

638 1. Jahan S. 2016. Cholera - epidemiology, prevention and control, p. 145–157. *In* Makun,
639 HA (ed.), Significance, Prevention and Control of Food Related Diseases. Croatia:
640 InTech. InTechOpen, Rijeka, Croatia.

641 2. Momba M, Azab El-Liethy M. 2017. *Vibrio cholerae* and Cholera biotypes, p. online. *In*
642 Pruden, A, Ashbolt, N, Miller, J (eds.), Global Water Pathogen Project. Michigan State
643 University, Michigan.

644 3. Clemens JD, Nair GB, Ahmed T, Qadri F, Holmgren J. 2017. Cholera. Lancet 390:1539–
645 1549.

646 4. Kaper JB, Morris JG, Levine MM. 1995. Cholera. Clin Microbiol Rev 8:48–86.

647 5. Islam MT, Alam M, Boucher Y. 2017. Emergence, ecology and dispersal of the pandemic
648 generating Vibrio cholerae lineage. Int Microbiol 20:106–115.

649 6. Choi SY, Rashed SM, Hasan NA, Alam M, Islam T, Sadique A, Johura F-T, Eppinger M,
650 Ravel J, Huq A, Cravioto A, Colwell RR. 2016. Phylogenetic Diversity of *Vibrio cholerae*
651 Associated with Endemic Cholera in Mexico from 1991 to 2008. MBio 7:e02160-15.

652 7. Dalsgaard A, Skov MN, Serichantalergs O, Echeverria P, Meza R, Taylor DN. 1997.
653 Molecular evolution of *Vibrio cholerae* O1 strains isolated in Lima, Peru, from 1991 to
654 1995. J Clin Microbiol 35:1151–1156.

655 8. Orata FD, Keim PS, Boucher Y. 2014. The 2010 Cholera Outbreak in Haiti: How Science
656 Solved a Controversy. PLoS Pathog 10:e1003967.

657 9. Weill F-X, Domman D, Njamkepo E, Almesbahi AA, Naji M, Nasher SS, Rakesh A,
658 Assiri AM, Sharma NC, Kariuki S, Pourshafie MR, Rauzier J, Abubakar A, Carter JY,
659 Wamala JF, Seguin C, Bouchier C, Malliavin T, Bakhshi B, Abulmaali HH., Kumar D,
660 Njoroge SM, Malik MR, Kiiru J, Luquero FJ, Azman AS, Ramamurthy T, Thomson NR,
661 Quilici M-L. 2018. Genomic insights into the 2016–2017 cholera epidemic in Yemen.
662 Nature 565:230–233.

663 10. Ali M, Nelson AR, Lopez AL, Sack DA. 2015. Updated global burden of cholera in
664 endemic countries. PLoS Negl Trop Dis 9:e0003832.

665 11. Legros D. 2018. Global cholera epidemiology: Opportunities to reduce the burden of
666 cholera by 2030. J Infect Dis 218:S137–S140.

667 12. Mintz E, Omolo J, Ope M, Gathigi L, Thuranira M, Loharikar A, Abade A, Ayers T, De
668 Cock KM, Makayotto L, Oundo J, Ismail AM, Breiman RF, Langat D, Briere E, Amwayi
669 S, O'Reilly CE, Njeru I. 2013. A National Cholera Epidemic With High Case Fatality
670 Rates-Kenya 2009. J Infect Dis 208:S69–S77.

671 13. Sack DA, Sack RB, Chaignat C-L. 2006. Getting Serious about Cholera. N Engl J Med

672 355:649–651.

673 14. Katz LSS, Petkau A, Beaulaurier J, Tyler S, Antonova ESS, Turnsek MAA, Guo Y, Wang
674 S, Paxinos EEE, Orata F, Gladney LMM, Stroika S, Folster JPP, Rowe L, Freeman
675 MMM, Knox N, Frace M, Boncy J, Graham M, Hammer BKK, Boucher Y, Bashir A,
676 Hanage WPP, Domselaar GV Van, Tarr L, Van Domselaar G, Tarr CLL, Domselaar GV
677 Van. 2013. Evolutionary Dynamics of *Vibrio cholerae* O1 following a Single-Source
678 Introduction to Haiti. MBio 4:e00398-13.

679 15. Frerichs RR. 2016. Deadly river: Cholera and cover-up in post-earthquake Haiti. Cornell
680 University Press, Ithaca.

681 16. Frerichs RR, Keim PS, Barrais R, Piarroux R. 2012. Nepalese origin of cholera epidemic
682 in Haiti. Clin Microbiol Infect 18:E158–E163.

683 17. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V,
684 Pearson T, Waters AE, Prasad Upadhyay B, Devi Shrestha S, Adhikari S, Shakya G, Keim
685 PS, Aarestrup FM. 2011. Population Genetics of *Vibrio cholerae* from Nepal in 2010:
686 Evidence on the Origin of the Haitian Outbreak. MBio 2:e00157-11.

687 18. Boucher Y. 2016. Sustained Local Diversity of *Vibrio cholerae* O1 Biotypes in a
688 Previously Cholera-Free Country. MBio 7:e00570-16.

689 19. Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, Taviani E, Jeon Y-S, Kim DW,
690 Lee J-H, Brettin TS, Bruce DC, Challacombe JF, Detter JC, Han CS, Munk AC, Chertkov
691 O, Meincke L, Saunders E, Walters RA, Huq A, Nair GB, Colwell RR. 2009.
692 Comparative genomics reveals mechanism for short-term and long-term clonal transitions
693 in pandemic *Vibrio cholerae*. Proc Natl Acad Sci 106:15442–15447.

694 20. Octavia S, Salim A, Kurniawan J, Lam C, Leung Q, Ahsan S, Reeves PR, Nair GB, Lan
695 R. 2013. Population Structure and Evolution of Non-O1/Non-O139 *Vibrio cholerae* by
696 Multilocus Sequence Typing. PLoS One 8:e65342.

697 21. Kirchberger PC, Orata FD, Barlow EJ, Kauffman KM, Case RJ, Polz MF, Boucher Y.
698 2016. A small number of phylogenetically distinct clonal complexes dominate a coastal
699 *Vibrio cholerae* population. Appl Environ Microbiol 82:5576–5586.

700 22. Bwire G, Sack DA, Almeida M, Li S, Voeglein JB, Debes AK, Kagirita A, Buyinza AW,
701 Orach CG, Stine OC. 2018. Molecular characterization of *Vibrio cholerae* responsible for
702 cholera epidemics in Uganda by PCR, MLVA and WGS. PLoS Negl Trop Dis
703 12:e0006492.

704 23. Garrine M, Mandomando I, Vubil D, Nhampossa T, Acacio S, Li S, Paulson JN, Almeida
705 M, Domman D, Thomson NR, Alonso P, Stine OC. 2017. Minimal genetic change in
706 *Vibrio cholerae* in Mozambique over time: Multilocus variable number tandem repeat
707 analysis and whole genome sequencing. PLoS Negl Trop Dis 11:e0005671.

708 24. Safa A, Nair GB, Kong RYC. 2010. Evolution of new variants of *Vibrio cholerae* O1.

709 Trends Microbiol 18:46–54.

710 25. Boucher Y, Orata FD, Alam M. 2015. The out-of-the-delta hypothesis: Dense human
711 populations in low-lying river deltas served as agents for the evolution of a deadly
712 pathogen. Front Microbiol 6:1120.

713 26. Mandal S, Mandal MD, Pal NK. 2011. Cholera: A great global concern. Asian Pac J Trop
714 Med 4:573–580.

715 27. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J,
716 Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence
717 typing: A portable approach to the identification of clones within populations of
718 pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145.

719 28. Horwood P, Collins D, Jonduo M, Rosewell A, Dutta S, Dagina R, Ropa B, Siba P,
720 Greenhill A. 2011. Clonal Origins of *Vibrio cholerae* O1 El Tor Strains, Papua New
721 Guinea, 2009–2011. Emerg Infect Dis 17:2063.

722 29. Luo Y, Ye J, Jin D, Ding G, Zhang Z, Mei L, Octavia S, Lan R. 2013. Molecular analysis
723 of non-O1/non-O139 *Vibrio cholerae* isolated from hospitalised patients in China. BMC
724 Microbiol 13:52.

725 30. Maiden MCJ, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND.
726 2013. MLST revisited: The gene-by-gene approach to bacterial genomics. Nat Rev
727 Microbiol 11:728–736.

728 31. Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo T R, Jaykus L-A, DePaola A.
729 2008. Determination of Molecular Phylogenetics of *Vibrio parahaemolyticus* Strains by
730 Multilocus Sequence Typing. J Bacteriol 190:2831–2840.

731 32. Lam C, Octavia S, Reeves PR, Lan R. 2012. Multi-locus variable number tandem repeat
732 analysis of 7th pandemic *Vibrio cholerae*. BMC Microbiol 12:82.

733 33. Chenal-Francisque V, Passet V, Brisse S, Cantinelli T, Diancourt L, Pourcel C, Lecuit M,
734 Leclercq A, Tran-Hykes C, Bracq-Dieye H. 2013. Optimized Multilocus Variable-
735 Number Tandem-Repeat Analysis Assay and Its Complementarity with Pulsed-Field Gel
736 Electrophoresis and Multilocus Sequence Typing for *Listeria monocytogenes* Clone
737 Identification and Surveillance. J Clin Microbiol 51:1868–1880.

738 34. Vogler AJ, Birdsell DN, Lee J, Vaissaire J, Doujet CL, Lapalus M, Wagner DM, Keim P.
739 2011. Phylogeography of *Francisella tularensis* ssp. holarctica in France. Lett Appl
740 Microbiol 52:177–180.

741 35. Struelens MJ, Brisse S. 2013. From molecular to genomic epidemiology: Transforming
742 surveillance and control of infectious diseases. Eurosurveillance 18:20386.

743 36. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, Grundmann H,
744 Friedrich AW, on behalf of the ESCMID Study Group. 2013. Overview of molecular

745 typing methods for outbreak detection and epidemiological surveillance. Eurosurveillance
746 18:20380.

747 37. Klassen JL, Currie CR. 2012. Gene fragmentation in bacterial draft genomes: extent,
748 consequences and mitigation. BMC Genomics 13:14.

749 38. Danin-Poleg Y, Cohen LA, Gancz H, Broza YY, Goldshmidt H, Malul E, Valinsky L,
750 Lerner L, Broza M, Kashi Y. 2007. *Vibrio cholerae* Strain Typing and Phylogeny Study
751 Based on Simple Sequence Repeats. J Clin Microbiol 45:736–746.

752 39. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, Murphy N, Holliman R,
753 Sefton A, Millar M, Dyson ZA, Dougan G, Holt KE. 2016. An extended genotyping
754 framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. Nat
755 Commun 7:12827.

756 40. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. 2014. Evaluation of
757 Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. PLoS One
758 9:e87991.

759 41. Chen C, Zhang W, Zheng H, Lan R, Wang H, Du P, Bai X, Ji S, Meng Q, Jin D, Liu K,
760 Jing H, Ye C, Gao GF, Wang L, Gottschalk M, Xu J. 2013. Minimum core genome
761 sequence typing of bacterial pathogens: A unified approach for clinical and public health
762 microbiology. J Clin Microbiol 51:2582–2591.

763 42. Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R, Xu J. 2016. Population structure
764 and minimum core genome typing of Legionella pneumophila. Sci Rep 6:21356.

765 43. Wang R, Yu D, Yue J, Kan B. 2016. Variations in SXT elements in epidemic *Vibrio*
766 *cholerae* O1 El Tor strains in China. Sci Rep 6:22733.

767 44. Meibom KL, Blokesch M, Dolganov NA, Wu C-Y, Schoolnik GK. 2005. Chitin induces
768 natural competence in *Vibrio cholerae*. Science 310:1824–1827.

769 45. Borgeaud S, Metzger LC, Scrignari T, Blokesch M. 2015. The type VI secretion system of
770 *Vibrio cholerae* fosters horizontal gene transfer. Science 347:63–68.

771 46. Neumann B, Prior K, Bender JK, Harmsen D, Klare I, Fuchs S, Bethe A, Zühlke D,
772 Göhler A, Schwarz S, Schaffer K, Riedel K, Wieler LH, Werner G. 2019. A Core Genome
773 Multilocus Sequence Typing Scheme for *Enterococcus faecalis*. J Clin Microbiol
774 57:e01686-18.

775 47. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT,
776 Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS,
777 Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-
778 Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-
779 Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and
780 epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol 2:1–10.

781 48. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers
782 M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen D, Willems RJL. 2015. Core
783 Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of
784 *Enterococcus faecium*. J Clin Microbiol 53:3788–3797.

785 49. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ. 2017. Core Genome
786 Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter*
787 *jejuni* and *C. coli* Human Disease Isolates. J Clin Microbiol 55:2086–2097.

788 50. Janowicz A, De Massis F, Ancora M, Camma C, Patavino C, Battisti A, Prior K, Harmsen
789 D, Scholz H, Zilli K, Sacchini L, Di Giannatale E, Garofolo G. 2018. Core Genome
790 Multilocus Sequence Typing and Single Nucleotide Polymorphism Analysis in the
791 Epidemiology of *Brucella melitensis* Infections. J Clin Microbiol 56:e00517-18.

792 51. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. 2018. Defining and Evaluating a
793 Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of
794 *Clostridium difficile*. J Clin Microbiol 56:e01987-17.

795 52. Jones RC, Harris LG, Morgan S, Ruddy MC, Perry M, Williams R, Humphrey T, Temple
796 M, Davies AP. 2019. Phylogenetic Analysis of *Mycobacterium tuberculosis* Strains in
797 Wales by Use of Core Genome Multilocus Sequence Typing To Analyze Whole-Genome
798 Sequencing Data. J Clin Microbiol 57:e02025-18.

799 53. Sails AD, Swaminathan B, Fields PI. 2003. Clonal complexes of *Campylobacter jejuni*
800 identified by multilocus sequence typing correlate with strain associations identified by
801 multilocus enzyme electrophoresis. J Clin Microbiol 41:4058–4067.

802 54. Leavis HL, Bonten MJ, Willems RJ. 2006. Identification of high-risk enterococcal clonal
803 complexes: global dispersion and antibiotic resistance. Curr Opin Microbiol 9:454–460.

804 55. Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, Keddy KH, Salje H,
805 Moore S, Mukhopadhyay AK, Bercion R, Luquero FJ, Ngandjio A, Dosso M, Monakhova
806 E, Garin B, Bouchier C, Pazzani C, Mutreja A, Grunow R, Sidikou F, Bonte L, Breurec S,
807 Damian M, Njanpop-Lafourcade B-M, Sapriel G, Page A-L, Hamze M, Henkens M,
808 Chowdhury G, Mengel M, Koeck J-L, Fournier J-M, Dougan G, Grimont PAD, Parkhill J,
809 Holt KE, Piarroux R, Ramamurthy T, Quilici M-L, Thomson NR. 2017. Genomic history
810 of the seventh pandemic of cholera in Africa. Science 358:785–789.

811 56. Reimer A, Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L,
812 Olsen-Rasmussen M, Frace M, Sammons S, Dahourou G, Boncy J, Smith A, Mabon P,
813 Petkau A, Graham M, Gilmour M, Gerner-Smidt P. 2011. Comparative Genomics of
814 *Vibrio cholerae* from Haiti, Asia, and Africa. Emerg Infect Dis 17:2113.

815 57. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for
816 scrupulous assemblies. Genome Biol 19:153.

817 58. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
818 Bioinformatics 26:2460–2461.

819 59. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S,
820 Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil
821 LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O,
822 Vonstein V, Wilke A, Zagnitko O, Formsma K, Kubal M, Vonstein V, Stevens R, McNeil
823 LK, Edwards RA, Pusch GD, Reich C, Glass EM, Olsen GJ, Paczian T, Overbeek RA,
824 Meyer F, Vassieva O, DeJongh M, Osterman AL, Disz T, Best AA, Gerdes S, Parrello B,
825 Bartels D, Olson R, Paarmann D. 2008. The RAST server: rapid annotations using
826 subsystems technology. BMC Genomics 9:75.

827 60. Islam MT, Liang K, Im MS, Winkjer J, Busby S, Tarr CL, Boucher Y. 2018. Draft
828 Genome Sequences of Nine *Vibrio* sp. Isolates from across the United States Closely
829 Related to Vibrio cholerae. Microbiol Resour Announc 7:e00965-18.

830 61. Liang K, Islam MT, Hussain N, Winkjer NS, Im MS, Rowe LA, Tarr CL, Boucher Y.
831 2019. Draft Genome Sequences of Eight Vibrio sp. Clinical Isolates from across the
832 United States That Form a Basal Sister Clade to *Vibrio cholerae*. Microbiol Resour
833 Announc 8:e01473-18.

834 62. Liang K, Orata FD, Winkjer NS, Rowe LA, Tarr CL, Boucher Y. 2017. Complete
835 Genome Sequence of *Vibrio* sp. Strain 2521-89, a Close Relative of *Vibrio cholerae*
836 Isolated from Lake Water in New Mexico, USA. Genome Announc 5:e00905-17.

837 63. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007.
838 DNA–DNA hybridization values and their relationship to whole-genome sequence
839 similarities. Int J Syst Evol Microbiol 57:81–91.

840 64. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species
841 delimitation with confidence intervals and improved distance functions. BMC
842 Bioinformatics 14:60.

843 65. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM :
844 assessing the quality of microbial genomes recovered from isolates, single cells, and
845 metagenomes. Genome Res 25:1043–1055.

846 66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
847 search tool. J Mol Biol 215:403–410.

848 67. Jolley KA, Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation
849 at the population level. BMC Bioinformatics 11:595.

850 68. Garg P, Aydanian A, Smith D, Morris JGJG, Nair GB, Stine OC, Antonia A, Smith D,
851 Morris JGJG, Nair GB, Stine OC. 2003. Molecular Epidemiology of O139 *Vibrio*
852 *cholerae*: Mutation, Lateral Gene Transfer, and Founder Flush. Emerg Infect Dis 9:810–
853 814.

854 69. Dunn JC. 1974. Well-separated clusters and optimal fuzzy partitions. J Cybern 4:95–104.

855 70. Brock G, Pihur V, Datta S, Datta S. 2008. clValid: An R Package for Cluster Validation. J

856 Stat Softw 25:1–22.

857 71. R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna,
858 Austria.

859 72. Canty A, Ripley BD, others. 2017. boot: Bootstrap R (S-Plus) Functions. R Packag
860 version 1.

861 73. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New
862 York.

863 74. Hubert L, Arabie P. 1985. Comparing partitions. J Classif 2:193–218.

864 75. Chang F, Qiu W, Zamar RH, Lazarus R, Wang X. 2010. clues: An R Package for
865 Nonparametric Clustering Based on Local Shrinking. J Stat Softw 33:1–16.

866 76. Zhou Z, Alikhan N, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA,
867 Achtman M. 2018. GrapeTree: visualization of core genomic relationships among 100,000
868 bacterial pathogens. Genome Res 28:1395–1404.

869 77. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The harvest suite for rapid core-
870 genome alignment and visualization of thousands of intraspecific microbial genomes.
871 Genome Biol 15:524.

872 78. Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting
873 the presence of recombination. Genetics 172:2665–2681.

874 79. Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): An online tool for phylogenetic
875 tree display and annotation. Bioinformatics 23:127–128.

876 80. Hagberg A, Swart P, S Chult D. 2008. Exploring network structure, dynamics, and
877 function using NetworkX, p. 11–16. *In* Varoquaux G, Vaught T, Millman J (ed),
878 Proceedings of the 7th Python in Science Conference (SciPy2008).

879 81. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski
880 B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of
881 Biomolecular Interaction Networks. Genome Res 13:2498–2504.

882 82. Salim A, Lan R, Reeves PR. 2005. *Vibrio cholerae* pathogenic clones. Emerg Infect Dis
883 11:1758–1760.

884 83. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi
885 SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN,
886 Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G. 2011. Evidence
887 for several waves of global transmission in the seventh cholera pandemic. Nature
888 477:462–465.

889 84. Lucidarme J, Hill DMC, Bratcher HB, Gray SJ, du Plessis M, Tsang RSW, Vazquez JA,

890 Taha M-K, Ceyhan M, Efron AM, Gorla MC, Findlow J, Jolley KA, Maiden MCJ,
891 Borrow R. 2015. Genomic resolution of an aggressive, widespread, diverse and expanding
892 meningococcal serogroup B, C and W lineage. J Infect 71:544–552.

893 85. Royer G, Fourreau F, Boulanger B, Mercier-Darty M, Ducellier D, Cizeau F, Potron A,
894 Podglajen I, Mongardon N, Decousser J-W. 2019. Local outbreak of extended-spectrum
895 β-lactamase SHV2a-producing *Pseudomonas aeruginosa* reveals the emergence of a new
896 specific sub-lineage of the international ST235 high-risk clone. J Hosp Infect.

897 86. Eppinger M, Pearson T, Koenig SSK, Pearson O, Hicks N, Agrawal S, Sanjar F, Galens
898 K, Daugherty S, Crabtree J, Hendriksen RS, Price LB, Upadhyay BP, Shakya G, Fraser
899 CM, Ravel J, Keim PS. 2014. Genomic Epidemiology of the Haitian Cholera Outbreak: a
900 Single Introduction Followed by Rapid, Extensive, and Continued Spread Characterized
901 the Onset of the Epidemic. MBio 5:e01721-14.

902 87. World Health Organization. 2017. Weekly epidemiological record.

903 88. Cavailler P, Lucas M, Perroud V, McChesney M, Ampuero S, Guérin PJ, Legros D,
904 Nierle T, Mahoudeau C, Lab B, Kahozi P, Deen JL, von Seidlein L, Wang XY, Puri M,
905 Ali M, Clemens JD, Songane F, Baptista A, Ismael F, Barreto A, Chaignat CL. 2006.
906 Feasibility of a mass vaccination campaign using a two-dose oral cholera vaccine in an
907 urban cholera-endemic setting in Mozambique. Vaccine 24:4890–4895.

908 89. Sardar T, Mukhopadhyay S, Bhowmick AR, Chattopadhyay J. 2013. An optimal cost
909 effectiveness study on Zimbabwe cholera seasonal data from 2008-2011. PLoS One
910 8:e81231.

911 90. Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, Haley BJ, Taviani E,
912 Hine E, Su Q, others. 2012. Genomic diversity of 2010 Haitian cholera outbreak strains.
913 Proc Natl Acad Sci 109:E2010--E2017.

914 91. Hu D, Liu B, Feng L, Ding P, Guo X, Wang M, Cao B, Reeves PR, Wang L. 2016.
915 Origins of the current seventh cholera pandemic. Proc Natl Acad Sci 113:E7730–E7739.

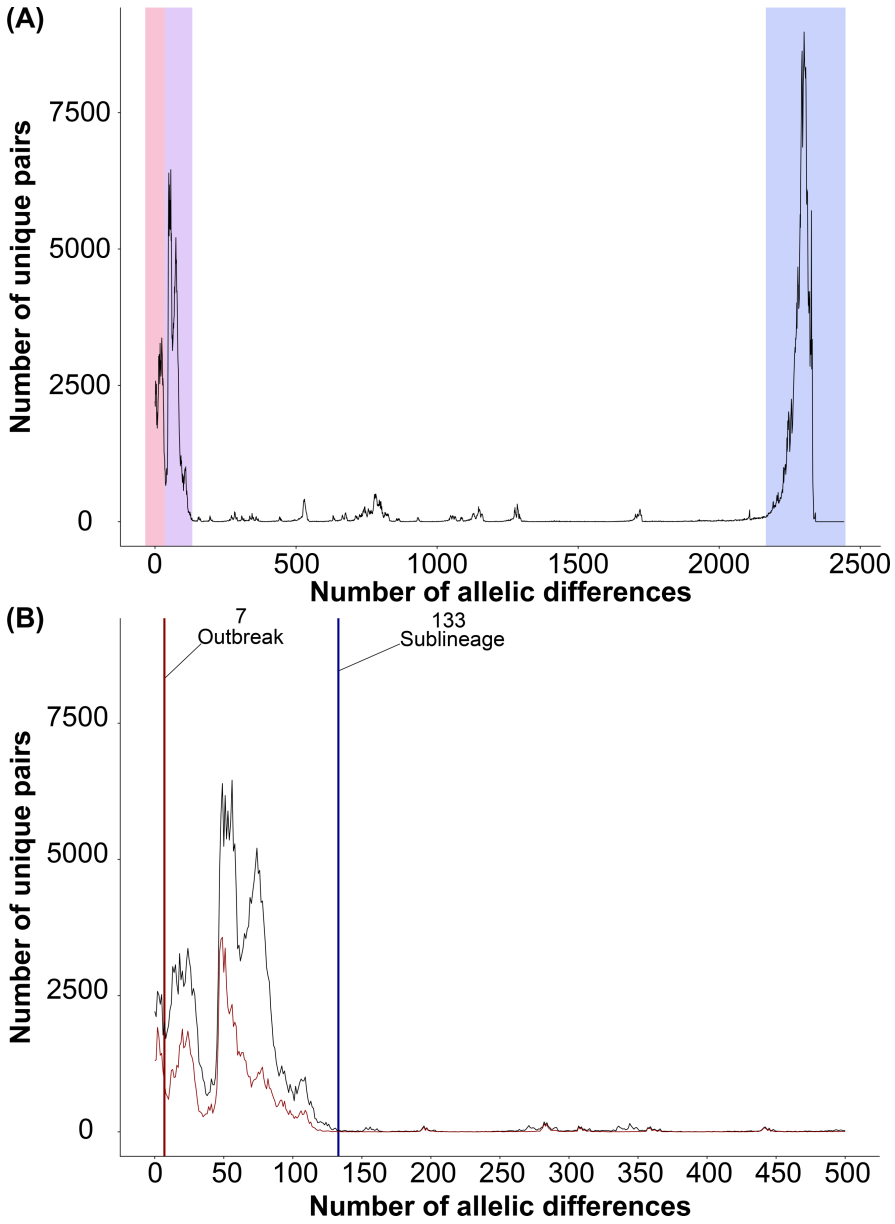
916 92. Guillaume Y, Ternier R, Vissieres K, Casseus A, Chery MJ, Ivers LC. 2018. Responding
917 to cholera in Haiti: Implications for the national plan to eliminate cholera by 2022. J Infect
918 Dis 218:S167–S170.

919 93. Pightling AW, Petronella N, Pagotto F. 2014. Choice of Reference Sequence and
920 Assembler for Alignment of *Listeria monocytogenes* Short-Read Sequence Data Greatly
921 Influences Rates of Error in SNP Analyses. PLoS One 9:e104579.

922 94. Childers BM, Klose KE. 2007. Regulation of virulence in Vibrio cholerae: The ToxR
923 regulon. Future Microbiol 2:335–344.

924

925 **FIGURES AND FIGURE LEGENDS**



926

927 **Fig. 1.** Pairwise allelic differences for all isolates used in this study. Both plots show the

928 frequency of allelic mismatches in pairwise comparisons. A) Pairwise comparisons of up to

929 2,443 allelic differences are shown. Major peaks are shaded. B) Comparisons with up to 500

930 allelic differences are shown. Vertical lines indicate the outbreak threshold (red) and sublineage

931 threshold (blue). Pairwise comparisons of only clinical isolates are shown in red.

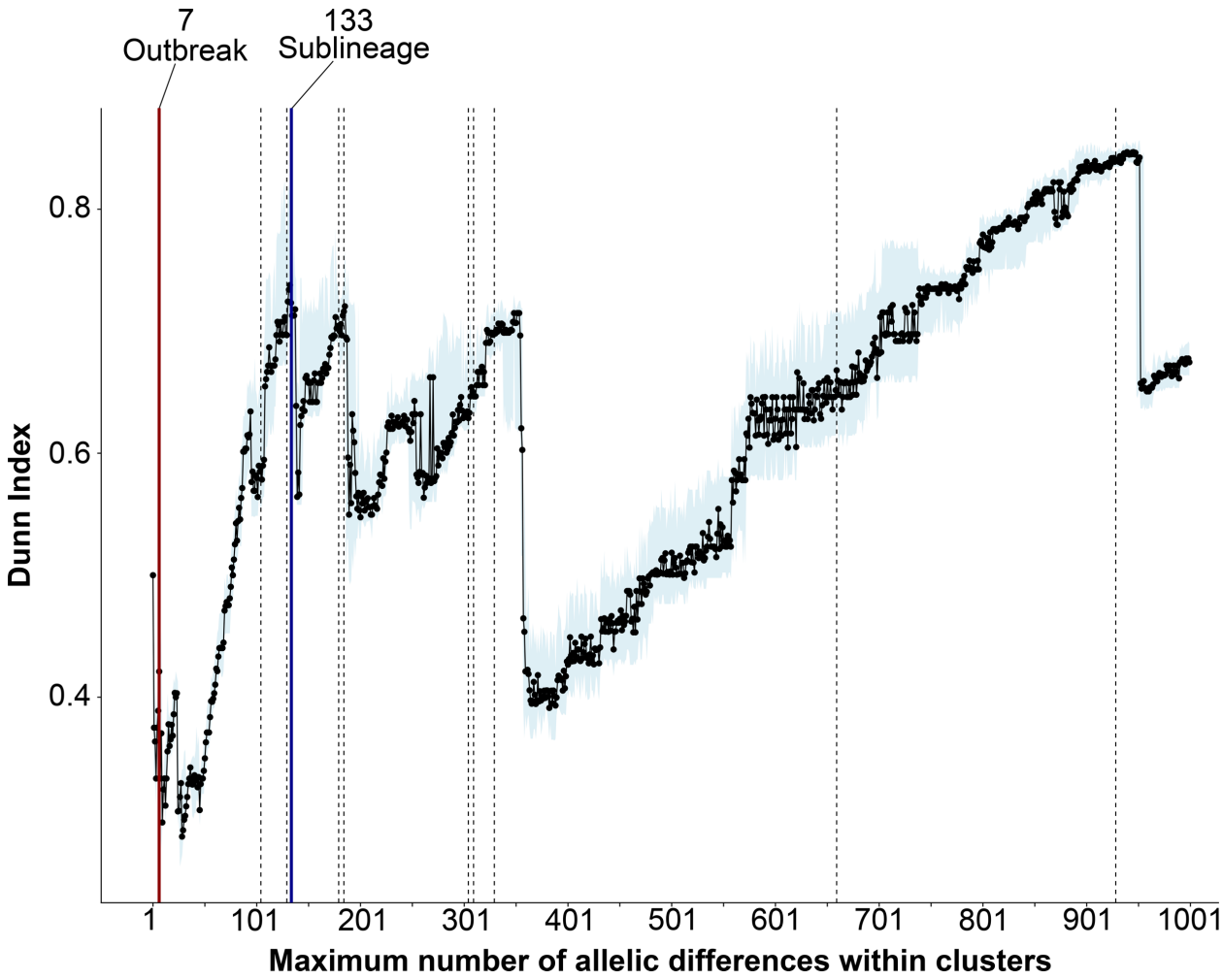


Fig. 2. Plot showing the Dunn Index for clustering thresholds ranging from 1 to 1,000 allelic differences. Each clustering threshold is bootstrapped 100 times. The median, plotted with the light blue shade, indicates the 25th to 75th percentile range. Dark blue and dark red vertical lines indicate the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

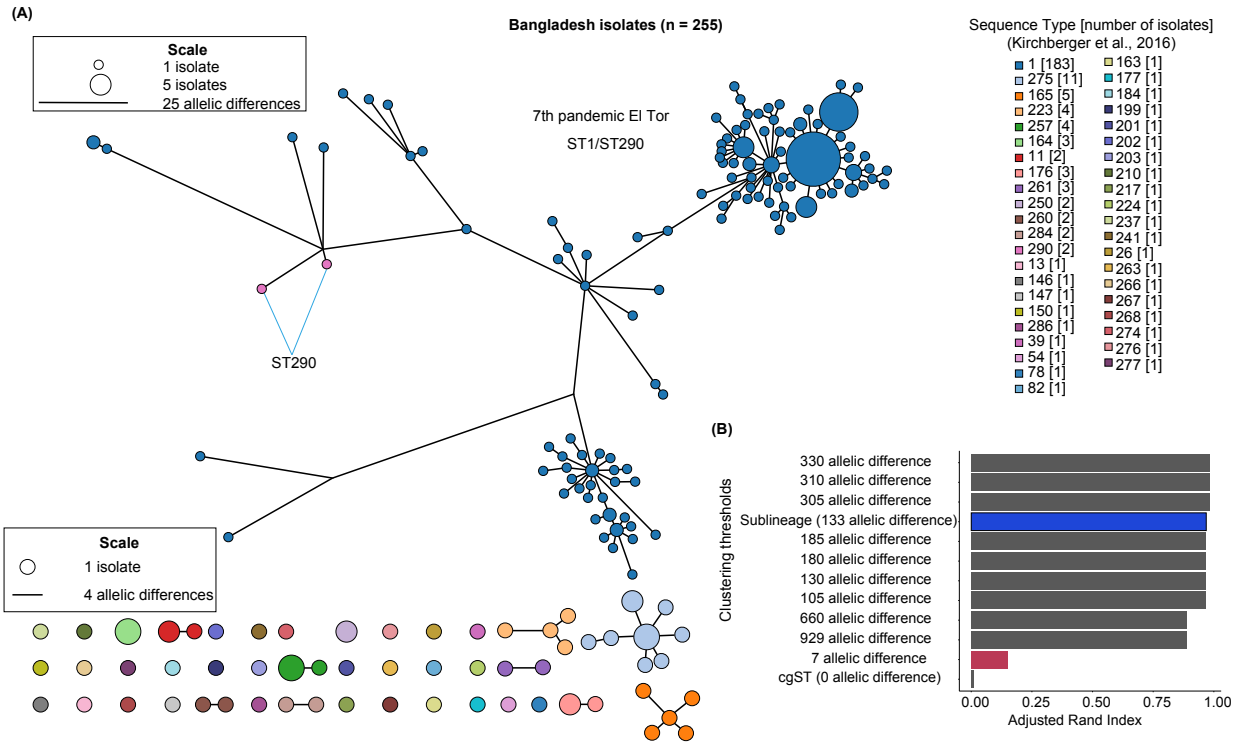
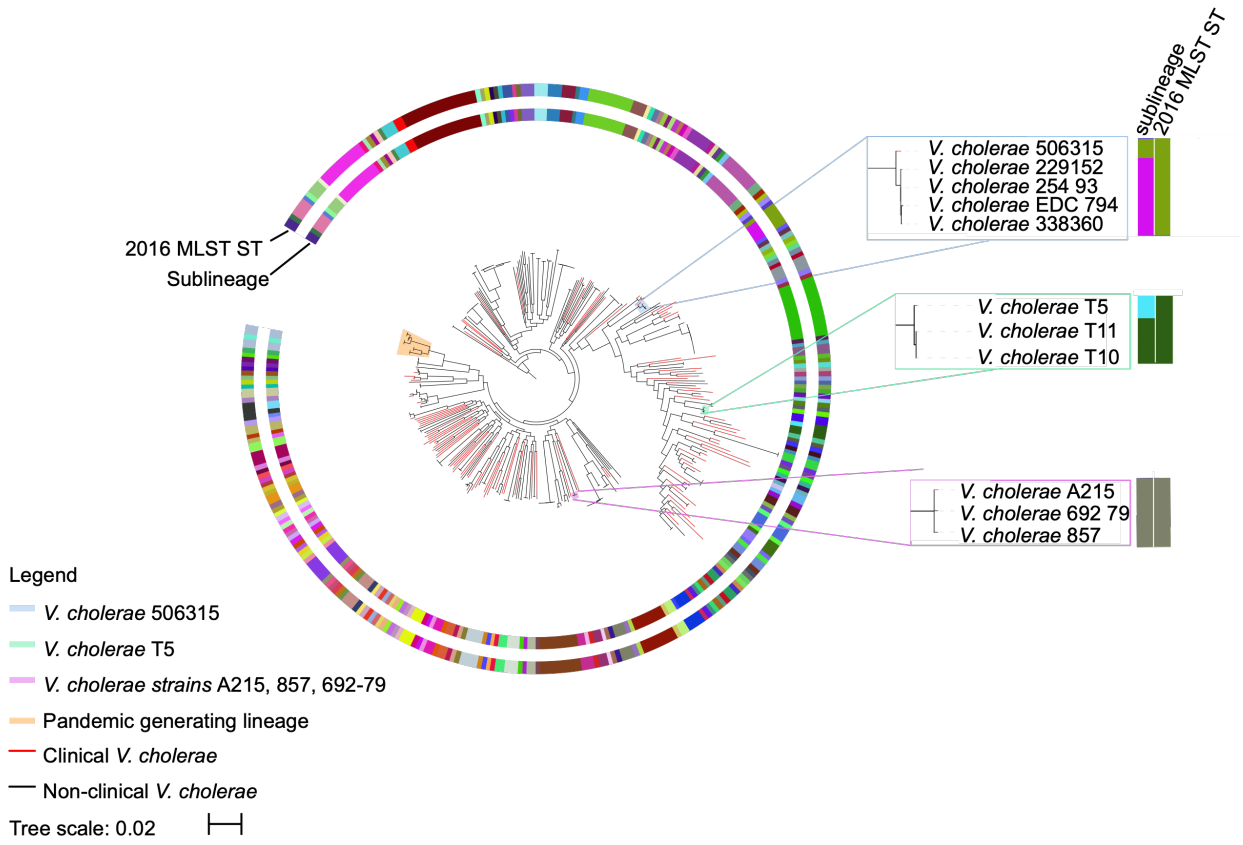


Fig. 3. Evaluation of network similarities between cgMLST sublineage threshold (133 allelic differences) and MLST ST. A) Networks of all sublineages identified using only *V. cholerae* isolates from Bangladesh (*n* = 255). Each cluster represents a sublineage and includes isolates with less than or equal to 133 allelic differences with each other. Each node represents a cgST and is colored by sequence type based on the 2016 MLST scheme (21). Size 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.



949

950 **Fig. 4.** Phylogenetic tree of 1,146 *V. cholerae* isolates (excluding the 116 isolates from the
951 Yemen outbreak study (9)) reconstructed using Parsnp v1.2 (77). All group inside the PG lineage
952 (7th pandemic El Tor, El Tor progenitor, El Tor sister, Classical, and Classical sister) are
953 collapsed. Outer rings represents clustering by sequence type based on the 2016 MLST scheme
954 by Kirchberger *et al.* (21), whereas the inner ring represents clustering based on the sublineage
955 threshold (i.e., 133 allelic differences). Branches of clinical strains are colored in red. The
956 phylogenetic tree is rooted with a basal lineage to *V. cholerae* (collapsed) (60, 61).

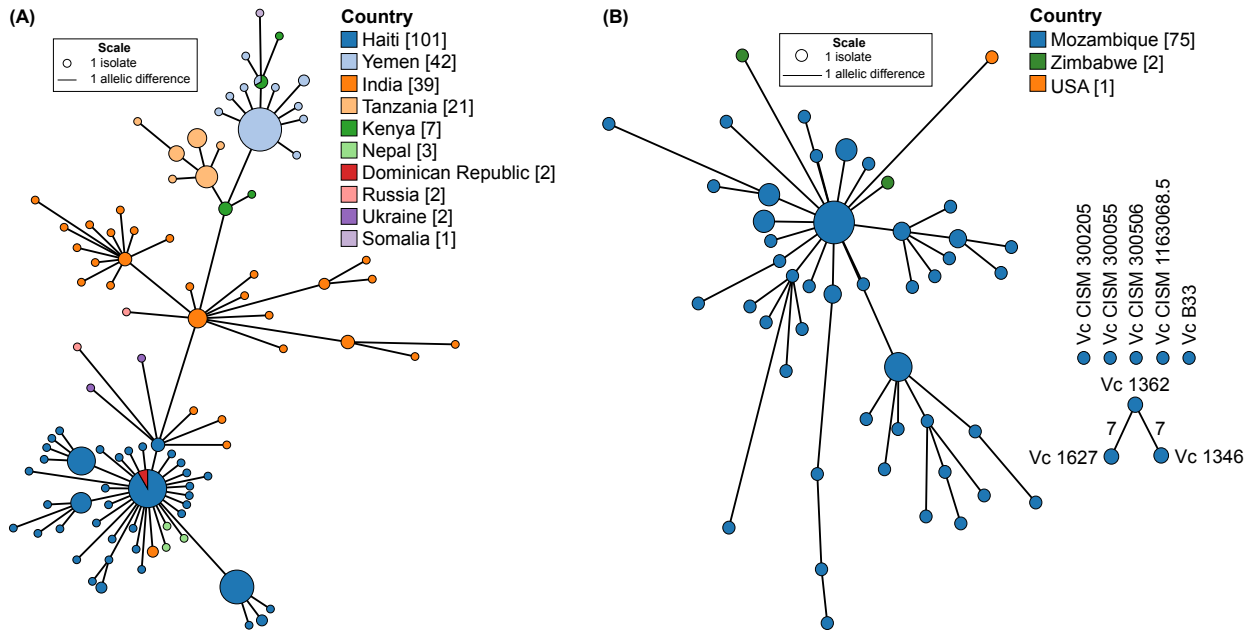


Fig. 5. Minimum spanning trees isolated when the outbreak threshold (7 allelic differences) was applied to the complete dataset of 1,264 isolates. A) All isolates which clustered together with the isolates from Haiti (blue) and Yemen (light blue) based on the clustering threshold of seven allelic differences. B) All isolates 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 (green) and one from the USA (orange), 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. Length of the lines is proportional to the number of allelic differences and all connections have less than or equal to seven allelic differences.

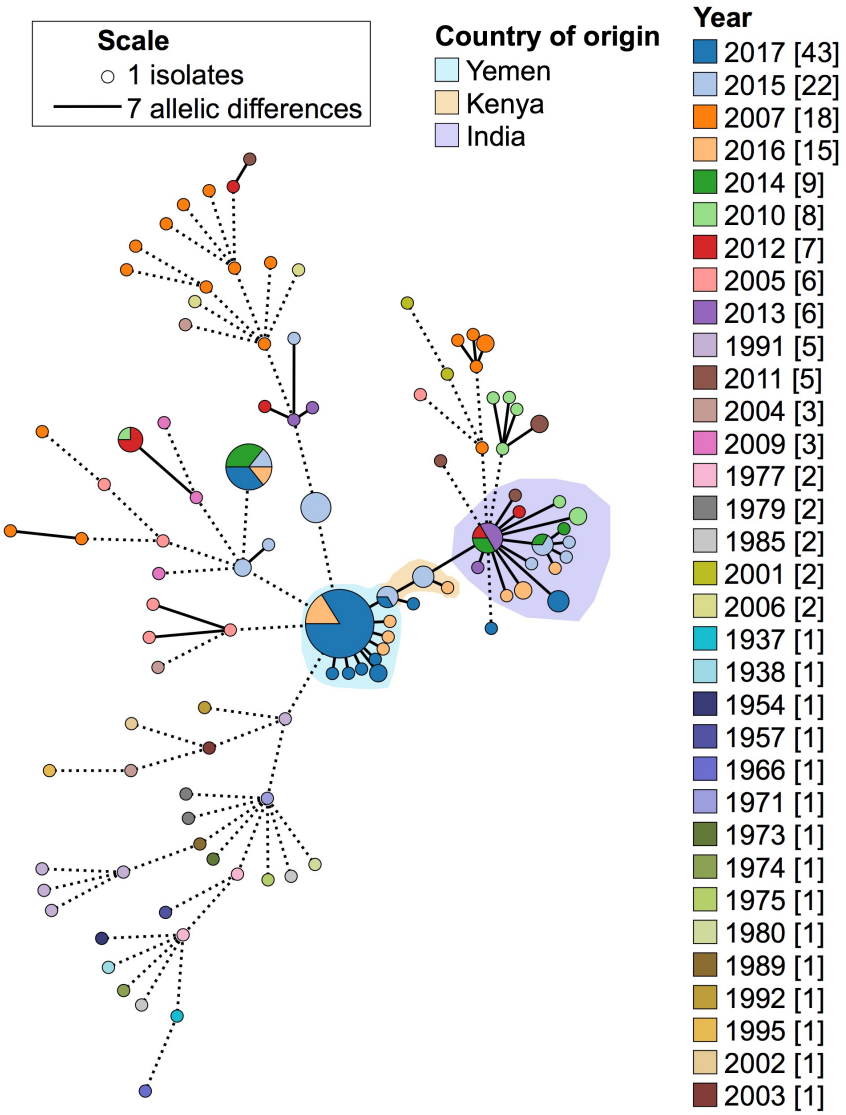


Fig. 6. cgMLST Minimum Spanning Tree of all Yemen isolates and representative 7th pandemic El Tor strains (Table S1). All isolates connected by dotted lines share eight or more allelic differences (not drawn to scale). All isolates connected with solid lines share seven or fewer allelic differences (i.e., 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.

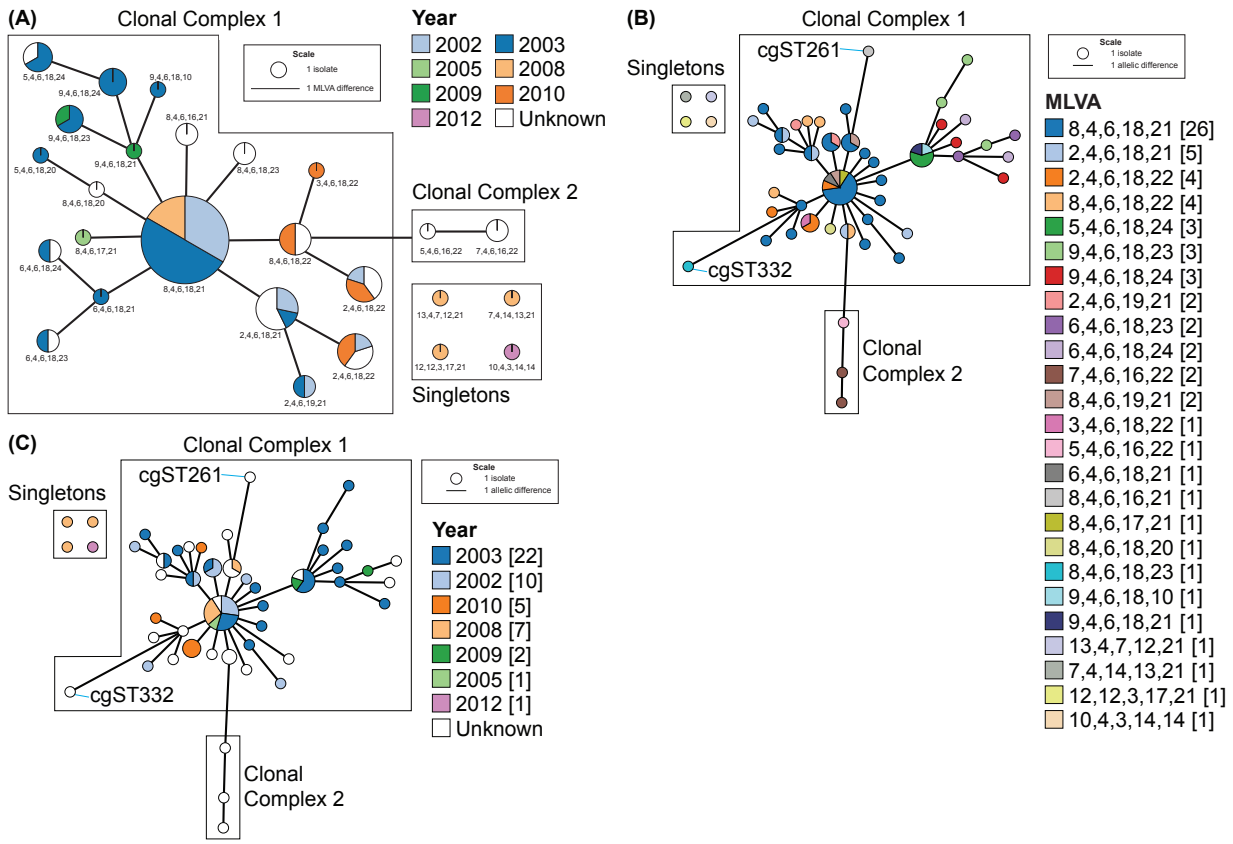


Fig. 7. Comparison between cgMLST and MLVA focusing on Mozambique isolates. A) Population structure of pandemic V. cholerae in Mozambique based on MLVA profiles from Garrine *et al.* (23). B) MST of the Mozambique isolates based on the cgMLST scheme colored based on MLVA profiles. C) Same MST of Mozambique isolates colored by year of isolation. All isolates in 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.

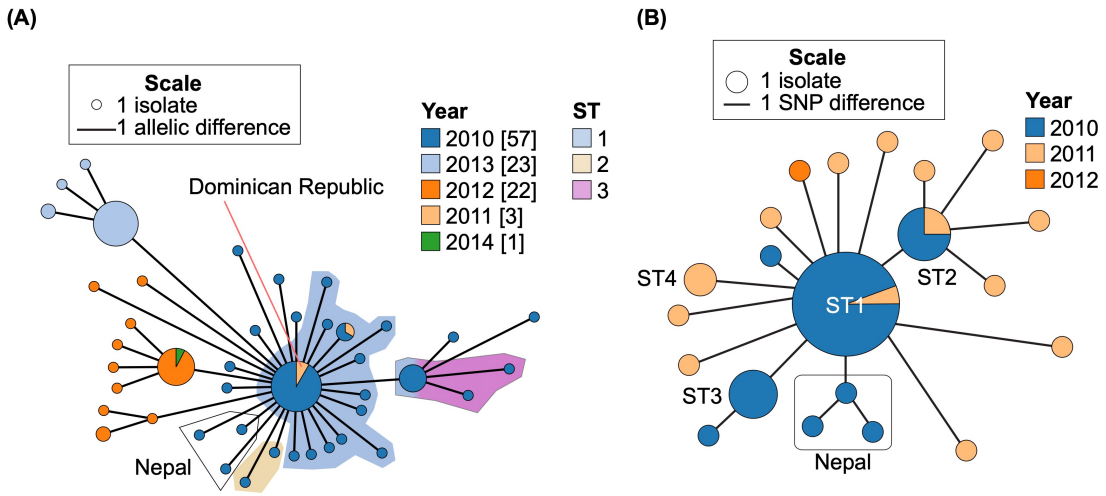


Fig. 8. Comparison between cgMLST and SNP based analysis focusing on Haiti outbreak related strains. A) Minimum spanning tree (MST) of isolates from the 2010 cholera outbreak in Haiti. All lines indicate connections of four or fewer allelic differences. Each node represents a cgST which is colored by year of isolation. Background shading represent ST designations based on 45 high-quality SNPs (14). cgST66 contains a mix of color as it contains both ST1 and ST3 isolates. Any isolate from countries other than Haiti is labeled. The length of the lines is proportional to the number of allelic differences. B) MST constructed from whole genome SNP data as in Katz *et al.* (14). The length of the lines indicates the number of nucleotide substitutions. Size of the nodes is proportional to the number of isolates.

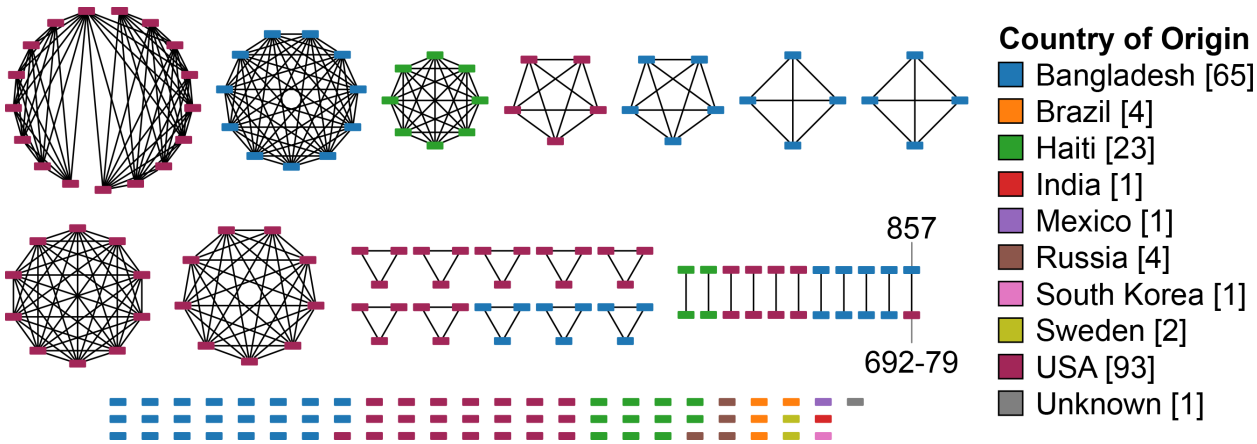


Fig. 9. Sublineage clusters constructed using networkX (80) and visualized on Cytoscape (81) showing non-clinical environmental isolates that are not part of the PG lineage. Missing loci were assumed to contain the most common allele when calculating allelic differences. Isolates are connected only if they share 133 allelic differences or fewer with each other. Each node represents an isolate and is colored by the country of isolation.

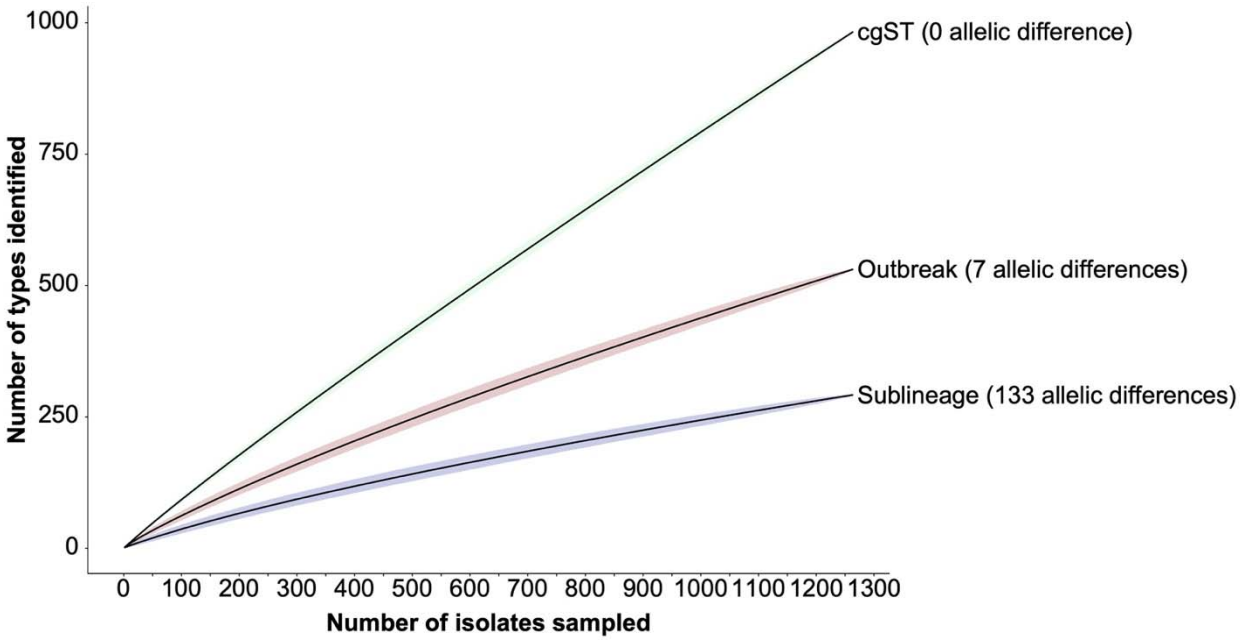


Fig. S1: Rarefaction curve for cgST, outbreak threshold (7 allelic difference) and the sublineage threshold (133 allelic difference) computed using mothur (1) with default parameters.

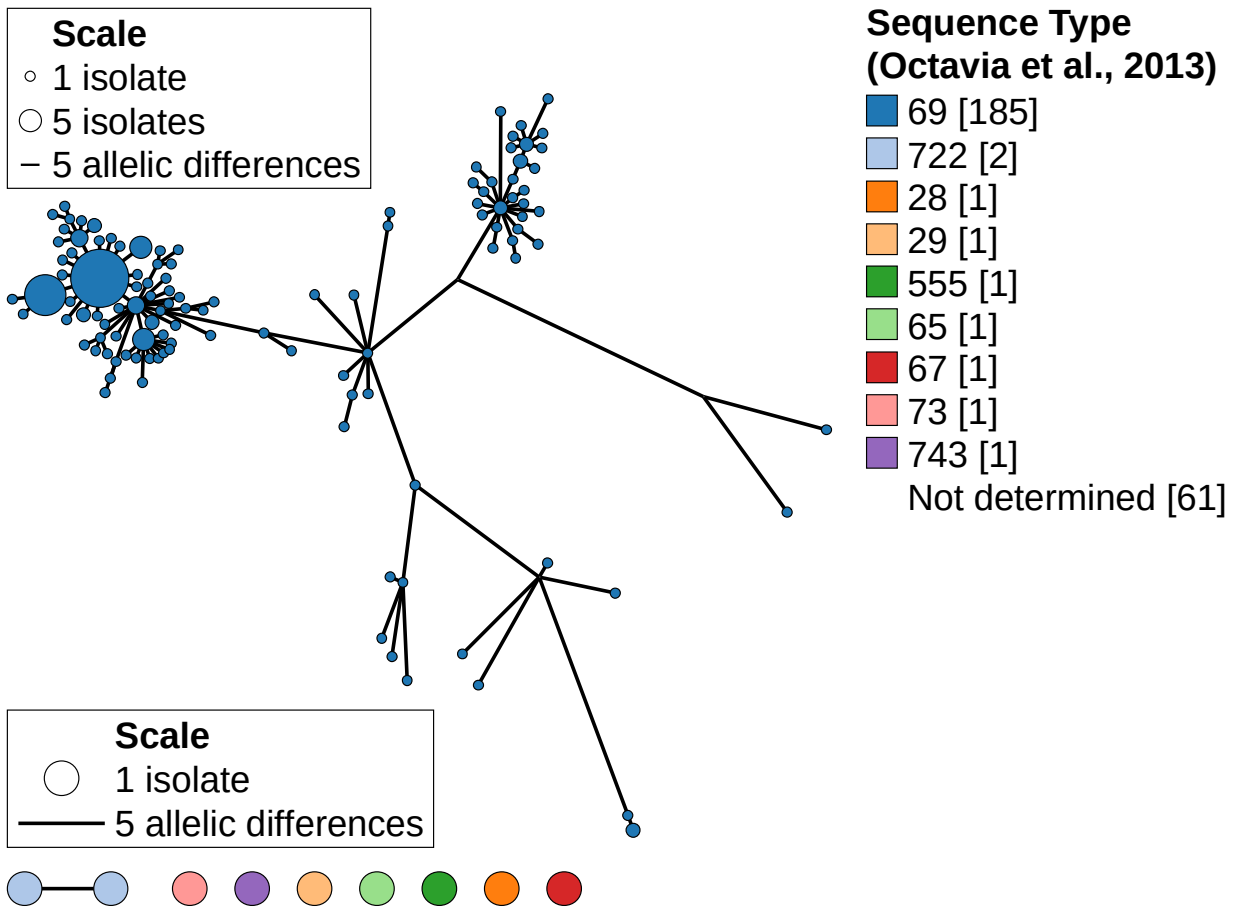


Fig. S2: Evaluation of network similarities between cgMLST sublineage threshold (133 allelic differences) and the 2013 MLST scheme (2). Networks of all sublineages identified using only *V. cholerae* isolates from Bangladesh (*n* = 255). Each cluster represents a sublineage and includes isolates with less than or equal to 133 allelic differences with each other. Each node represents a cgST and is colored by sequence type based on the 2013 MLST scheme (2). Size 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.

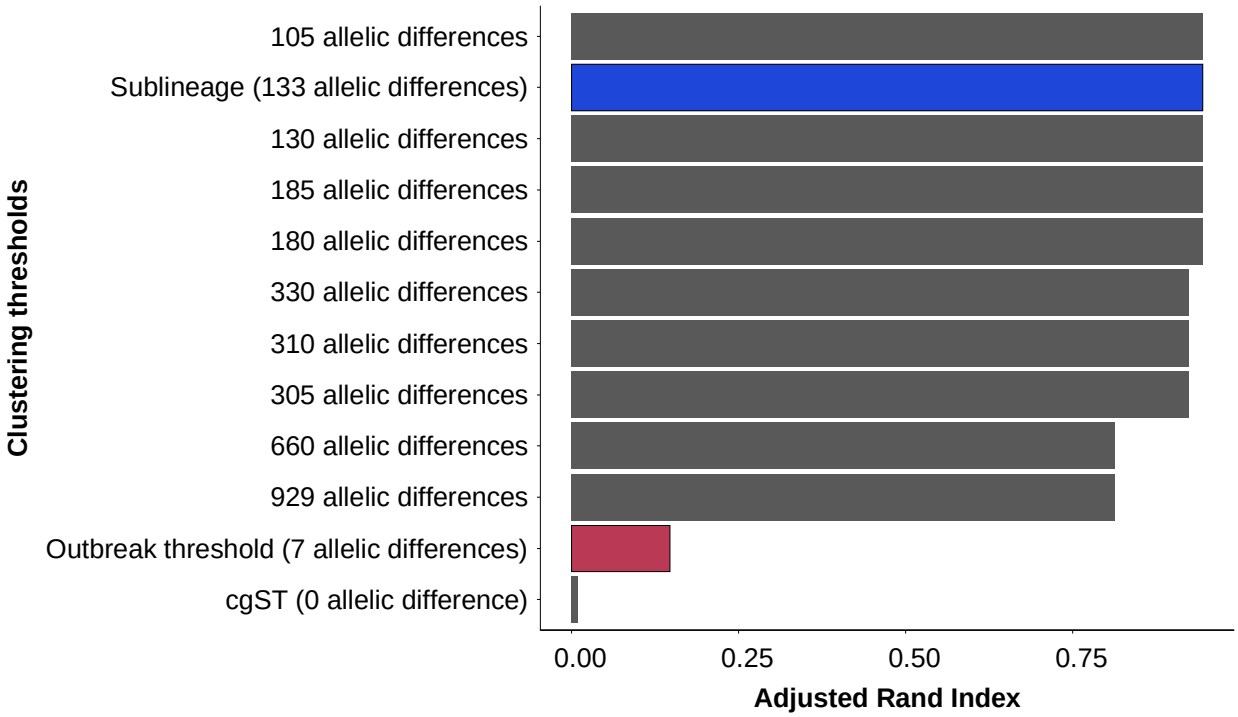


Fig. S3: Adjusted rand index calculated with the same method as before (Fig 3B), but compared with the 2013 MLST scheme (2). The same allelic thresholds were chosen and the outbreak and sublineage thresholds are highlighted in red and blue respectively.

References

1. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, C. F. Weber. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75:7537–7541.
2. 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.