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#]

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

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

[c]Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

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

[#]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. A set of 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 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 which 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/).

37 **IMPORTANCE**

38 Toxigenic *Vibrio cholerae* of the O1 and O139 serogroups are the causative agent of cholera, an
39 acute diarrheal disease that plagued the world for centuries, if not millennia. Here, we introduce
40 a core genome multilocus sequence typing (cgMLST) scheme for *V. cholerae*. Using cgMLST,
41 we established an outbreak threshold that can efficiently identify outbreak related strains and
42 potential sources of introduction. We also defined a sublineage threshold that is similar to
43 traditional MLST sequence type which will provide context to this new typing method by
44 relating it to previous MLST results. cgMLST outperforms all existing methods in terms of
45 resolution, standardization, and ease-of-use, making this scheme the most suitable method for *V.*
46 *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 (10).
61 Being an indicator of healthcare and socio-economic disparities (11, 12), this disease is often
62 under-reported due to its negative influence on tourism as it implies poor water quality (13).
63 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 cholera 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 the 7th pandemic El Tor, El Tor sister, El Tor progenitor,

74 Classical, and Classical sister clades (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, due to their rapid mutation rate, VNTR regions 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 within *V. cholerae* (43–45) and between the species and its close relatives (46, 47),
124 SNP-based methods, although suitable in individual epidemiological studies, will be difficult to
125 serve as a universal classification method for *V. 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 (48–54). This method has the added
136 advantage of backwards compatibility with all MLST schemes. This means that it is possible to
137 determine MLST profiles of any isolates based on their cgMLST profiles, since cgMLST would
138 include all housekeeping genes by definition. This allows for a 1:1 mapping of any previously
139 established MLST scheme to the cgMLST scheme, helping consolidate already existing
140 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 (55) or major outbreak strains (56). However, cgMLST
147 offers even greater flexibility than MLST in this regard given the number of loci considered.
148 With large clustering thresholds, it is possible to identify lineage- or even sublineage-level
149 differences to study large scale patterns and answer broader ecological questions. Furthermore,
150 with smaller clustering thresholds where groups are created based on the sharing of a larger
151 number of alleles, it is possible to identify very closely related strains useful in epidemiological
152 studies. The benefits of defining clustering thresholds through cgMLST have already been
153 demonstrated in other human pathogens, such as *Brucella melitensis* (52), *Campylobacter jejuni*
154 (51) , *Clostridium difficile* (53), *Enterococcus faecium* (50), and *Listeria monocytogenes* (49).
155 In this study, we introduce a cgMLST scheme for the genome-wide typing of *V. cholerae*
156 and demonstrate its universality and efficacy by applying it to known cholera outbreaks around
157 the world. The advantages of cgMLST are presented by comparing the scheme with previously
158 established classification methods. Additionally, we have produced a 1:1 mapping of the
159 cgMLST scheme against two previously established MLST schemes for *V. cholerae* (20, 21),
160 allowing for the consolidation of existing classification information. The cgMLST scheme,
161 genome sequences used in this study, and relevant epidemiological information are publicly
162 available on PubMLST (https://pubmlst.org/vcholerae/), which allows for the automatic
163 annotation and subsequent analyses of hundreds of newly uploaded *V. cholerae* genomes in a
164 global context. This increase in efficiency, standardizability, and resolution compared to current
165 methods make cgMLST the most suitable classification scheme for large scale *V. cholerae*

166 surveillance. By applying this scheme to our collection of over 1,200 isolates collected around
167 the world, it was possible to establish outbreak and sublineage thresholds which not only allowed
168 us to validate the South Asian origin of many modern epidemics as proposed in previous studies
169 (5, 57, 58) but also identify clustering affinity among environmental strains, where isolates from
170 the same sublineage are also likely from the same geographic region. This pattern is not seen in
171 clinical isolates, as human hosts readily carry them over large geographical distances.
172

173 **RESULTS AND DISCUSSION**

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

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

189 resolve between closely related strains in the 7th pandemic El Tor lineage makes cgMLST more
190 suitable in outbreak surveillance than traditional MLST.
191
192 **Backwards compatibility with previous subspecies classification methods**
193 Much like how cgSTs are important in studying closely related strains typical in outbreaks, it is
194 also important in establishing a standardized nomenclature at a higher level to answer broader
195 ecological questions. Here, we propose a sublineage definition for *V. cholerae* based on cgMLST.
196 Pairwise allelic differences calculated between all isolates show three major peaks (Fig.
197 1A). The first peak ends at 40 allelic differences, and the second peak ends at 133 allelic
198 differences (Fig. 1B). The last peak begins at 2,200 allelic differences (Fig. 1A), which is
199 expected due to mutational saturation (i.e., every single allele in the scheme is different between
200 the two distantly related strains being compared). Both breaks (i.e., 40 or 133 allelic differences)
201 could represent a potential sublineage delineation. To choose between the two thresholds, the
202 clustering efficiency is measured by calculating the Dunn Index (DI) (60) (see Materials and
203 Methods). Since cluster distances are measured by allelic differences, the network with the best
204 clustering efficiency (i.e., the highest DI) will also produce clusters that best represent biological
205 relationships, as isolates are more closely related to themselves than to isolates from other
206 clusters. A DI was calculated for each clustering threshold in the range of 1 to 1,000 allelic
207 differences with 100 bootstrap replicates (Fig. 2). As the clustering threshold defines the
208 maximum number of allelic differences within a cluster, the smaller the threshold, the more
209 closely related the isolates are within a cluster. It is clear that DIs in the range of 0 to 50 allelic
210 differences are significantly lower than the DIs in the range of 100 to 350 allelic differences,
211 with 133 being a clear local maximum. Since 133 allelic differences has the best clustering

212 efficiency and also represents a natural break where most isolate pairs have either less than or
213 much greater number of allelic differences (Fig. 1B), it was therefore chosen as the sublineage
214 threshold.

215 Because cgMLST includes all housekeeping genes, information from the two MLST
216 schemes previously developed for *V. cholerae* (20, 21) can now be consolidated with the
217 cgMLST scheme by creating a 1:1 cgMLST to MLST map. To evaluate the similarities between
218 the sublineage threshold and the MLST schemes, we created a minimum spanning tree (MST)
219 for all Bangladesh isolates (n = 255) showing only edges with 133 allelic differences or fewer
220 (Fig. 3A and Fig. S2). Each cluster therefore represents a single sublineage. Bangladesh was
221 chosen to compare cgMLST and MLST as it is the most extensively sampled country both in
222 terms of clinical and environmental isolates in our dataset. Using this dataset, the chosen
223 sublineage threshold produces clusters that closely resemble ST clustering from traditional
224 MLST. Based on the 2013 MLST scheme (20), each sublineage corresponds to exactly one ST
225 (Fig. S2), whereas there is only one sublineage that contains two STs based on the 2016 MLST
226 scheme (21) (Fig. 3A). All but two isolates belong to ST1; N16961 and A19 belong to ST290,
227 which differs from ST1 at only one of seven MLST loci (Table S2). The reason these two
228 isolates are of a different MLST ST could only be partly explained; they were isolated at an
229 earlier time point (1970s near the start of the 7th pandemic (61)) than most of the remaining
230 isolates, which were isolated from 1991 onwards (Table S1).

231 It is impossible to visually evaluate similarities between two MSTs with over 1,200 nodes
232 each simply due to the sheer volume of data. The Adjusted Rand Index (ARI) was therefore used
233 as a metric to determine network similarities (62) (see Materials and Methods). In order to
234 determine whether the sublineage threshold (i.e., 133 allelic differences) is indeed the best match

235 to traditional MLST schemes, we chose 11 clustering thresholds distributed across the range of 1
236 to 1,000 allelic differences (Fig. 2) to compare with the MLST schemes. These additional
237 thresholds are chosen as they have a relatively high DI compared to their immediate neighbours.
238 More data points were chosen in the range of 105 to 330 allelic differences, as it was expected
239 that thresholds in this range will best match the traditional MLST schemes. Interestingly, all
240 thresholds in that range had comparable ARIs when compared to both the 2016 (21) and the
241 2013 MLST schemes (20) (Fig. 3B and Fig. S3), indicating that all of them, including the
242 sublineage threshold, produce clusters similar to the MLST scheme. This would suggest that
243 there can be a large range of diversity within a single MLST ST where isolates can have
244 anywhere from 0 (i.e., have the same cgST) to 330 allelic differences. Although clustering
245 thresholds between 105 to 330 allelic differences produce similar clusters to a traditional MLST
246 scheme, the sublineage threshold of 133 allelic differences was chosen as it has the best
247 clustering efficiency (Fig. 2) and it represents a natural breakpoint in the currently sampled
248 population (Fig. 1B).

249 A phylogenetic tree of 1,146 isolates was used to assess the phylogenetic support of the
250 sublineage threshold across different *V. cholerae* strains (Fig. 4). This tree includes all *V.*
251 *cholerae* isolates within our dataset with the exception of the recently published 116 clinical
252 isolates from the Yemen cholera outbreak (9), which all belongs to the 7th pandemic El Tor
253 lineage. The strains within the PG lineage are closely related with little genetic variation. These
254 lineages are therefore collapsed in the phylogenetic tree as the relationships between them are
255 not well resolved. All sublineages formed monophyletic clades, although in some cases the most
256 basal branch is of a different sublineage (e.g., *V. cholerae* strains T5 or 506315) creating
257 paraphyletic clades. Ideally, each sublineage would correspond to exactly one full monophyletic

258 clade. The reason this is not seen is likely the lack of sampling, leading to the grouping of
259 relatively distantly related isolates together in the same clade. Further sampling will likely
260 resolve these cases into two separate monophyletic clades. Out of 1,262 isolates, we identified
261 291 sublineages, 19 of which belong exclusively to the PG lineage and 223 are singletons. Much
262 like cgSTs, the rarefaction curve also indicates that the total sublineage diversity of *V. cholerae*
263 is far from being sampled (Fig. S1).

264 The sublineage concept has been applied to numerous pathogens and as such were each
265 defined differently depending on the pathogen in question. Some have defined sublineages based
266 on natural breaks in genetic similarities (49), while others may use sublineage to refer
267 specifically to traditional MLST STs (63) or even finer level of resolution below the MLST ST
268 level based on whole genome analyses (64). There is, however, one unifying feature of all
269 sublineage definitions – that they all refer to monophyletic clades. Sublineages are defined in this
270 study based on natural breaks in allelic differences calculated from cgMLST profiles and were
271 put into context by comparing with two traditional MLST schemes. We have shown that our
272 definition of sublineage results in monophyletic clades but also corresponds to any traditional
273 MLST ST designation (Fig. 4). This sublineage definition will therefore play a crucial role in
274 consolidating information from all previous MLST analyses.

275

276 **A universal South Asian origin for modern cholera outbreaks**

277 With the continual improvement of next-generation sequencing techniques, whole genome
278 sequencing is expected to become a standard practice or even the first identification tool used in
279 clinical and epidemiological studies. It is therefore critical to develop a rapid typing scheme for
280 genome sequence data that has the power to inform us about the relationship of a novel isolate

281 with known strains. This is done here by defining what we term an 'outbreak threshold' based on
282 cgMLST, which can identify outbreak related strains and potential sources of introduction. The
283 outbreak threshold is expected to be less than 40 allelic differences as isolates from the same
284 outbreak are very closely related (9, 14). There is a minor discontinuity at seven allelic
285 differences where most isolate pairs have either less or more than this number of allelic
286 differences (Fig. 1B). Looking at the DI, the local maximum in the range of 0 to 50 occurs at
287 seven allelic differences as well (Fig. 2), making this cutoff a likely candidate for an outbreak
288 threshold. When applying the outbreak threshold to a full dataset containing all sequenced *V.*
289 *cholerae* genomes meeting the minimum quality threshold, major clusters were examined to
290 evaluate the ability of cgMLST to identify strains that are part of the same outbreak.
291 One of the major outbreak clusters identified, with no prior information required,
292 contains the Haiti and the Yemen outbreaks, which are the two best documented cholera
293 outbreaks in modern history (8, 9, 14, 16, 65). Isolates collected from these outbreaks form a
294 single cluster with the Dominican Republic, Eurasian (India, Russia, Nepal, and Ukraine), and
295 African (Tanzania, Kenya, and Somalia) isolates (Fig. 5A). The Dominican Republic isolates are
296 closely related to the Haiti outbreak strains. Given the close proximity of the two countries, co-
297 located on the island of Hispaniola, it was expected that isolates from Haiti would eventually
298 spread to the Dominican Republic (14). The 7th pandemic El Tor lineage spread across the world
299 from South Asia in three separate waves (61). The third wave, being the most recent distribution
300 event, has been claimed to be responsible for the outbreaks in Haiti and Yemen (9). It is
301 therefore not surprising to see Haiti and Yemen form a single cluster with India (i.e., South Asia)
302 at its center. Nepal is the known source of introduction for the Haiti outbreak in 2010 (16), and

303 comparison with over 1,200 *V. cholerae* isolates from all over the world still shows the Nepalese
304 isolates as the closest relatives to the Haitian isolates (Fig. 5A).
305 Cholera is still endemic in Africa (10) and caused several major reported outbreaks in
306 different countries over last few decades (66) including Mozambique (23, 67) and Zimbabwe
307 (68). Another major outbreak cluster groups most of the Mozambique isolates together with two
308 Zimbabwe isolates (strains CP1038(11) and 2011EL-1137) and one USA isolate (2009V-1116)
309 (Fig. 5B). Based on cgMLST, it is evident that the two Zimbabwe isolates are closely related to
310 the Mozambique isolates, differing at only four or less alleles. The close proximity of the two
311 countries suggests that these are likely travel-associated cases. Although outbreaks involving the
312 Mozambique isolates (23) and the Zimbabwe isolates (58, 69) have been independently studied,
313 the link between these isolates have not been shown before. Global cgMLST analysis is therefore
314 an invaluable tool as it allows for the identification of links between independent studies.
315 However, with only two Zimbabwe isolates in the dataset, additional sampling in this region is
316 required to understand the epidemiology of this outbreak. According to the NCBI BioSample
317 database, strain 2009V-1116 was collected by the Centers for Disease Control and Prevention in
318 2009 and is associated with travel to Pakistan. Since the 7th pandemic El Tor lineage has been
319 circulating in Asian and Middle Eastern countries for a long time (70), it is possible that, at least
320 within our dataset, the Mozambique isolates are the closest relative to this specific Pakistan
321 strain.
322
323 **Confirmation of an African connection for the Yemen outbreak**
324 The Yemen cholera outbreak began in October 2016 with eleven confirmed cases
325 (http://www.emro.who.int/pandemic-epidemic-diseases/cholera/cholera-cases-in-yemen.html).

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

349 Unlike the limited samples available from African cholera outbreaks, the Haiti and
350 Yemen outbreaks are significant cases for epidemiological investigations because *V. cholerae*
351 has been heavily sampled from these countries as well as surrounding regions. Two major
352 limitations in genomic epidemiology have been the lack of a universal classification scheme and
353 a comprehensive database; however, this is no longer the case in the genomic era as sequencing
354 technology is becoming increasingly more accessible (8). A genomic approach, as shown here, is
355 able to produce accurate predictions of potential origins of outbreaks and provides us with
356 sufficient resolution to accurately track the spread of the disease. Therefore, genomic analysis
357 should be the first step in any epidemiological study as not only will it help guide subsequent
358 analyses and investigations, but consistently sequencing new genomes will also help expand and
359 refine the current global *V. cholerae* genome database.
360

361 **Increased resolution for the history of cholera in Mozambique: comparing cgMLST to**
362 **MLVA**

363 The 7th pandemic reached Africa in 1970 and cholera appeared in Mozambique at roughly the
364 same time (57). Since its introduction, cholera has been endemic in that country and has
365 continued to cause multiple outbreaks (23). A popular tool for outbreak investigation is MLVA
366 (32, 38), which was recently used to study *V. cholerae* strains collected in Mozambique over
367 multiple years (23). MLVA is a subspecies typing method similar to MLST in concept; however,
368 it utilizes VNTRs instead of using gene sequences. As VNTR mutates at a faster rate than
369 conserved genes, it has been shown that MLVA provides greater resolution than MLST for some
370 species (32, 33). To establish a direct comparison between our cgMLST scheme and this MLVA
371 scheme, we examined the MSTs created by both methods focusing only on shared isolates (Fig.

372 7A and 7B). The MLVA identified 26 profiles forming two clonal complexes (CCs) and four
373 singletons (Fig. 7A) (23). A similar population structure is seen with the cgMLST analysis (Fig.
374 7B), including the four singletons identified in the MLVA. The central node in the cgMLST
375 MST consists mostly of isolates with MLVA profile '8,4,6,18,21' similar to the central node in
376 the MLVA MST (23). The two CCs identified in the MLVA MST are also identified in the
377 cgMLST MST with the smaller CC2 being at least four allelic differences away from the larger
378 CC1.

379 Although there are a few MLVA types that were grouped into a single cgST, such as
380 cgST1 and cgST114, indicating that cgMLST was unable to resolve the differences in these
381 MLVA types, there are many MLVA types such as profile '2,4,6,18,21', profile '7,4,6,16,22',
382 profile '9,4,6,18,24', and profile '8,4,6,18,22' that were split into multiple cgSTs. Overall there
383 are 48 cgSTs as opposed to only 26 MLVA types, showing that cgMLST provides better
384 resolution overall than MLVA. The cgMLST analysis overlaid with isolation dates shows that
385 the Mozambique *V. cholerae* strains are highly clonal and strains from the same cgST can cause
386 outbreaks over multiple years (e.g., cgST114 and cgST94) (Fig. 7C), which corroborates the
387 claim made in the initial MLVA study that the same MLVA type can be seen over multiple years
388 (23). In addition to increased resolution, cgMLST also produces more reliable and reproducible
389 results than MLVA, as it eliminates errors associated with the detection of VNTR regions using
390 PCR or sequencing-based methods. For the same reason that MLST is less affected by
391 convergent evolution compared to MLVA (35), cgMLST is also less affected by convergent
392 evolution.
393

394 **Standardizing the genotypes responsible for the Haiti 2010 cholera outbreak: comparing**
395 **cgMLST and SNP-based analyses**

396 One of the largest cholera outbreaks in modern history occurred in Haiti following the
397 devastating earthquake in 2010 (8, 71). Prior to this outbreak, there were no documented cholera
398 cases in Haiti (14, 18). Since the initial introduction, *V. cholerae* now remains endemic in Haiti
399 and is responsible for thousands of cholera cases annually (71). Multiple studies have strongly
400 suggested that the Haitian strains were in fact imported from Nepal (by the UN Nepalese troops)
401 and the outbreak occurred as a result of both inappropriate sanitary practice and the lack of
402 screening of the UN troops upon their arrival in Haiti (8, 15, 16, 71).

403 A SNP-based approach was used to study the evolutionary dynamics of *V. cholerae* in
404 Haiti (14). This technique relies on the identification of SNPs in draft or closed genomes. The
405 primary benefit of this method is that assembly and annotation are not required. It is also capable
406 of resolving closely related strains using whole genome data. However, SNP-based methods are
407 highly influenced by recombination events (43, 44) and quality filter parameters chosen (72).

408 To establish a direct comparison between the cgMLST scheme and SNP-based analysis,
409 we focused on MSTs of only the Haitian outbreak isolates (Fig. 8). All Haitian isolates are
410 closely related according to the cgMLST scheme, sharing at most four allelic differences with
411 each other (Fig. 8A). The Haitian and Nepalese isolates, therefore, also belong to the same
412 sublineage (SL6) which is consistent with the fact that these isolates belong to the same MLST
413 ST (either ST1 or ST69 based on the 2016 or 2013 MLST scheme, respectively (20, 21)) (Table
414 S1). The overall population structure is similar between the two methods where we have SNP
415 ST1 as the center of the MST with ST2 and ST3 extending from that likely ancestral genotype
416 (Fig. 8). SNP ST1, ST2, and ST3 can be split into 11, 2, and 3 different cgSTs, respectively (Fig.

417 8A). There is only one case, that of cgST66, where it contains isolates from both SNP ST1 and
418 ST3. Overall, cgMLST was able to differentiate 39% of the isolates while the SNP-based
419 analysis differentiated 35%, showing comparable level of resolution. As expected, both the
420 cgMLST and the SNP-based analyses showed that the Haiti outbreak is highly clonal where most
421 isolates belong to the same cgST or SNP ST (14). However, one important advantage of
422 cgMLST over SNP-based analysis is that the former can be easily standardized because it relies
423 on a predefined set of core genes. Based on these standardized genes, a systematic nomenclature
424 system can be established. This makes cgMLST more suitable than the SNP-based method as a
425 universally applicable classification system for epidemiological studies and research worldwide.
426

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

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

434 Along with all the publicly available environmental strains that are not part of the PG
435 lineage, there are a total of 195 isolates spanning 9 countries. After grouping the isolates at the
436 sublineage level (i.e., each cluster have at most 133 allelic differences), it could be noted that all
437 isolates from the same sublineage also shared a country of origin, with the exception of strains
438 692-79 and 857 (Fig. 9), which are from the USA and Bangladesh, respectively. Phylogenetic
439 analysis shows these isolates to be closely related to strain A215, a clinical isolate from the USA

440 (Fig. 4). All three strains contain the *toxR* gene, a toxin transcriptional regulator common in
441 pathogenic *V. cholerae* (73), as well as genes encoding for the Mannose-sensitive hemagglutinin
442 pilus, the RTX toxin, and hemolysin (*hlyA*), all of which are putative virulence factors for this
443 species. In addition, strains A215 and 857 also harbor the zona occludens toxin gene. Similar
444 toxin gene contents among these three isolates and close phylogenetic relationships suggest that
445 strains 692-79 and 857 may also be potentially pathogenic and capable of surviving inside a
446 human host. This provides evidence that although clinical isolates can spread across the world
447 rapidly and closely related isolates can be from very different parts of the world, environmental
448 isolates from the same geographic origin share an affinity among each other at least at the
449 sublineage level. It is important to note that our current dataset contains a relatively small
450 number of environmental isolates that are not part of the PG lineage. Therefore, this distinct
451 distribution pattern based on geographic origin may be a result of currently insufficient sampling
452 of environmental *V. cholerae* worldwide. With large-scale environmental sampling, it will be
453 possible to determine with greater accuracy the evolutionary rate and distribution pattern of *V.*
454 *cholerae* in the environment using cgMLST. In addition, this method will become an invaluable
455 tool in dealing with these large datasets, as it provides an efficient and standardized method of
456 classification.
457
458 **Conclusion**
459 With an extensive collection of over 1,200 *V. cholerae* isolates, we developed a cgMLST scheme
460 based on 2,443 core genes. We established a sublineage-level definition based on 133 allelic
461 differences as part of our standardized classification scheme. It was determined by comparisons
462 with previous MLST schemes that the cgMLST sublineage classification can be used as a proxy

463 for traditional MLST. Additionally, the universality and applicability of the scheme have been
464 tested by looking at various cholera outbreak cases. We determined an outbreak threshold based
465 on seven allelic differences that groups isolates from the same outbreak together with strains
466 from the potential source of introduction. This threshold creates clusters that are consistent with
467 known epidemiological data when applied to the Haiti and Yemen cholera outbreaks, two of the
468 best-documented cholera outbreaks in modern history. Also, we were able to confirm the South
469 Asian origin of modern cholera outbreaks. Furthermore, although current sampling is limited, a
470 geographic signal at the sublineage level not seen in clinical strains could be identified among
471 environmental isolates that are not part of the PG lineage (18, 19). Lastly, this scheme is fully
472 implemented on PubMLST (https://pubmlst.org/vcholerae/) for public access. All newly
473 available genomes uploaded to PubMLST will be annotated automatically and a cgST
474 designation will be assigned to isolates with less than 100 missing loci. Relevant epidemiological
475 data and the variety of analytical and visualization tools are all integrated on PubMLST,
476 allowing for a quick analysis of any newly sequenced genome in a global context. This scheme
477 will be an important tool for future large-scale epidemiological and biogeographical research.
478

479 **MATERIALS AND METHODS**

480 **Dataset description**

481 On November 6th 2018, 1,172 *V. cholerae* genomes consisting of 800 draft and complete
482 genomes and 372 sequence read archives (SRAs), available from both publicly available
483 database and private collection, were selected as our dataset. One hundred sixteen SRAs from a
484 recent study on the Yemen cholera outbreak (9) were subsequently added as an independent
485 evaluation of the cgMLST scheme (Table S3). The 488 SRAs were assembled using skesa (74)

486 or the CLC Genomics Workbench 7 (QIAGEN) using default parameters. This total dataset of
487 1,288 included twenty-six genomes with less than 90% of the core genes, which were identified
488 using USearch (75) based on RAST (76) annotations. These 26 genomes were removed from
489 subsequent analyses resulting in a final dataset of 1,262 genomes collected from 52 countries and
490 spanning 82 years from 1937 to 2018 (Table S1). These include a historical collection from the
491 6th cholera pandemic, clinical isolates from outbreaks in various countries (e.g., Bangladesh,
492 India, Haiti, Yemen, the Democratic Republic of Congo, and Russia), and environmental isolates
493 from different parts of the world (e.g., Bangladesh, Haiti, USA, Mexico, Brazil, etc.).
494
495 **Gene identification and allele assignments**
496 Instead of using the full dataset of 1,288 genomes, we selected a subset of high-quality genomes
497 because core gene identification is highly dependent on the initial dataset and the inclusion of
498 poorly assembled and/or sequenced data will reduce the number of core genes identified (49).
499 Firstly, 800 already assembled draft or complete genomes were selected for core gene
500 identification. Low-quality assemblies were then eliminated by removing genomes with less than
501 40× coverage and/or N50 values less than 40 kb. From a previously established cgMLST scheme
502 for *L. monocytogenes*, 40× coverage and 20 kb N50 value were used as cutoff thresholds, as
503 genomes that do not meet these criteria resulted in a low proportion of loci being called (49). The
504 40× coverage cutoff was adopted for this study; however, because the average *V. cholerae*
505 genome size (~ 4 Mb) is larger than the average *L. monocytogenes* genome (~ 3 Mb), 40 kb was
506 instead selected as the N50 cutoff. The use of these cutoffs resulted in the removal of 82
507 genomes.

508 The remaining 718 genomes were annotated using RAST (76) and USearch (75), and a
509 tentative set of core genes that were on average present in 99% of the genomes were selected. An
510 additional 13 genomes were removed, as they lacked more than 90% for the core genes, leaving
511 us with a dataset of 705 high-quality genomes (Table S4). However, an additional 26 genomes
512 were subsequently removed for the core gene analysis as it has been previously suggested that
513 they form a highly divergent lineage within the *V. cholerae* (77–79), ensuring that the dataset
514 used for core gene identification consists only of unambiguously *V. cholerae* isolates (also as
515 verified by average nucleotide identity (80) and digital DNA–DNA hybridization (81) between
516 genomes (77–79)). Completeness and potential contamination of all remaining 679 genomes
517 were also independently evaluated by checkM, which estimates these values based on the
518 presence and number of copies of a set of pre-defined single copy marker genes (82) (Table S5).
519 All genomes were, according to the criteria established by checkM, nearly complete (≥97%) with
520 medium to low levels of contamination (<7%) (82).
521 Each orthologous gene was compared against the *V. cholerae* N16961 reference genome
522 using BLASTN (83) to determine gene function. Any gene family with no homolog in N16961
523 or classified as pseudogenes on the NCBI GenBank database were removed, meaning N16961
524 was 100% complete for the cgMLST scheme. Any gene that was present in more than one copy
525 in any of the initial 679 genomes was also removed, as they were considered paralogous. Thus,
526 in this context, core genes are defined as being present in at least 90% of the 679 high-quality
527 assembled genomes in a single copy. By choosing a relaxed cutoff of 90% completeness, we
528 accounted for missing genes due to sequencing, annotation, or assembly errors while ensuring
529 there is sufficient resolution to differentiate between closely related strains, with at least 2,199
530 loci remaining for classification purposes. The final cgMLST scheme utilizes a set of 2,443

531 single-copy core gene loci, which is 2,425,296 bp in size and covering approximately 61% of the
532 genome. The list of core genes is available on PubMLST (https://pubmlst.org/vcholerae/).
533 Automated scripts in BIGSdb (84) were used to perform allele calls and assignments for
534 all 1,262 isolates (Table S1). Allele calls were made only for complete coding sequences with a
535 minimum of 70% similarity and 70% length coverage at the nucleotide level, as previously
536 described (49). Default settings were used for all other parameters.
537
538 **Core genome sequence type (cgST) assignment**
539 cgST, which was defined as a unique combination of alleles of all loci included in the scheme,
540 was assigned for all isolates, excluding those from the Yemen outbreak study (8), with an in-
541 house script, as previously described (85). Briefly, missing loci were replaced with the most
542 common allele when assigning cgSTs, allowing for a conservative estimate of diversity (85). The
543 116 isolates from the Yemen cholera outbreak study (9) were annotated automatically by
544 uploading them to PubMLST. PubMLST treated missing alleles as 'N'. cgSTs were assigned to
545 each allele profile, treating 'N' as a regular allele designation. However, different from typical
546 allele designations, 'N's can represent any allelic sequence; therefore, some isolates may contain
547 multiple cgST designations, all of which are possibly true cgSTs. For isolates with more than one
548 cgST suggested by PubMLST, postprocessing was done using an in-house script to identify the
549 most likely cgST, which was determined by assuming missing loci contained the most common
550 allele (Table S2). It is expected that as genome sequencing becomes more reliable, higher quality
551 genomes will be available and any missing data can be updated as needed.
552

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

554 Two MLST schemes developed for *V. cholerae* were mapped to this cgMLST scheme. The first
555 MLST scheme developed in 2013 by Octavia and colleagues (20) was used to study the global
556 population structure of non-O1/non-O139 *V. cholerae* and is currently hosted on PubMLST. All
557 isolates uploaded to PubMLST were automatically annotated with this scheme. Any missing data
558 in this scheme was ignored and no ST designation was assigned. The second MLST scheme
559 developed in 2016 by Kirchberger and colleagues (21) was used to study the population structure
560 of environmental *V. cholerae* in a region on the US East Coast. The second MLST scheme is not
561 currently hosted on PubMLST, but because the housekeeping genes in this scheme are also
562 found in the cgMLST scheme, a similar in-house script used in cgST assignments was used to
563 assign ST designations. Therefore, all isolates in this study were assigned three designations
564 when possible – two ST designations based on the two previously established MLST schemes
565 (20, 21) and one cgST designation based on the cgMLST scheme from this study.

566

567 **Outbreak and sublineage clustering thresholds**

568 A clustering threshold was defined as the maximum number of allelic differences found within a
569 cluster. All clusters were produced based on the single-linkage clustering method, which meant
570 an isolate belonged to a cluster if it linked with any isolate within that cluster. Two metrics were
571 used as general guidelines for determining clustering thresholds. The first metric used was the
572 Dunn Index (DI), which measured clustering efficiencies (60). Briefly, the DI was highest for a
573 network (i.e., the network has the best clustering efficiency) when the intra-cluster distances
574 were minimized, and the inter-cluster distances were maximized. Since isolate distances were
575 measured based on allelic differences, a high DI resulted in clusters where isolates were more

576 closely related to those found within the same cluster than those found in a different cluster. The
577 DI was calculated using the R packages 'clvalid' and 'boot' with 100 bootstrap replicates for
578 each threshold and graphed using the R package 'ggplot2' (86–89).
579 The second metric used was the Adjusted Rand Index (ARI), which measured the level of
580 similarity between two networks when clustering the same set of isolates by measuring the
581 amount of agreements (i.e., the number of pairs that were grouped either as being in the same
582 cluster or different cluster in both networks) and disagreements (i.e., the number of pairs that
583 were grouped together in one network but grouped separately in another) (62). The values ranged
584 from -1 (i.e., two networks are exactly opposite) to 1 (i.e., two networks are identical). ARI was
585 used to determine the level of similarity between various clustering thresholds and the MLST
586 schemes. ARI was calculated using the R package 'clues' and graphed using 'ggplot2' (87, 89,
587 90).
588
589 **Minimum spanning tree (MST)**
590 All MSTs, unless otherwise specified, were constructed using GrapeTree MSTv2 (91). Loci with
591 missing data were included in the profile as "–". GrapeTree provided a novel algorithm that
592 accounted for missing data when constructing an MST, an important feature since missing data is
593 common in whole and core genome-based analyses. GrapeTree is currently integrated within
594 PubMLST, which allows for quick visualization of the dataset with any provenance data.
595
596 **Phylogenetic analysis**
597 Parsnp v1.2 (92) was used to reconstruct the phylogenetic tree using *V. cholerae* N16961 as the
598 reference genome. The -x flag was used to enable filtering of SNPs in recombinogenic regions as

599 identified by PhiPack (93). Default settings were used for all other parameters. The phylogenetic
600 tree included 1,146 genomes (all genomes except for the 116 isolates from the recent Yemen
601 cholera outbreak study (9)). Since all isolates sequenced for the latter study belonged to the 7th
602 pandemic El Tor lineage, it would have had limited impact on the overall structure of the tree.
603 The phylogeny was visualized and annotated using iTOL (94).
604
605 **Biogeographical analysis of environmental isolates**
606 All environmental isolates that were not part of the PG lineage (18, 19) were first clustered based
607 on the sublineage threshold using the python package 'networkX' (95). Missing alleles were
608 replaced with the most common allelic designation when calculating pairwise differences to
609 establish a more conservative estimate of diversity. The network was then visualized using
610 Cytoscape (96).
611
612 **Data availability**
613 All previously sequenced *V. cholerae* genomes and the additional 65 genomes sequenced in this
614 study are available on the NCBI GenBank database. Table S6 lists all the accession numbers for
615 all the genomes used in this study. In addition, all genome sequences, allelic profiles, cgST
616 designations, ST designations, and relevant epidemiological data are publicly available on
617 PubMLST (https://pubmlst.org/vcholerae/).
618

ACKNOWLEDGMENTS

KYHL, FDO, and YFB conceived the experiments. KYHL, FDO, and MTI performed all data collection and analyses. FDO and TN performed genome sequencing of Bangladesh isolates. MA and CLT provided isolates used in this study.

We thank Dr. 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 YFB); the Integrated Microbial Biodiversity program of the Canadian Institute for Advanced Research (to YFB); federal appropriations to the Centers for Disease Control and Prevention through the Advanced Molecular Detection Initiative (to CLT); and graduate student scholarships from Alberta Innovates – Technology Futures (to KYHL, FDO, MTI, and TN), NSERC (to KYHL and TN), the University of Alberta Faculty of Graduate Studies and Research (Queen Elizabeth II Graduate Scholarship to KYHL and TN), and the Bank of Montréal Financial Group (to FDO). MA of icddr,b 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.

640 **REFERENCES**

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

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

647 3. Clemens JD, Nair GB, Ahmed T, Qadri F, Holmgren J. 2017. Cholera. Lancet 390:1539–
648 1549.

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

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

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

655 7. Dalsgaard A, Skov MN, Serichantalergs O, Echeverria P, Meza R, Taylor DN. 1997.
656 Molecular evolution of *Vibrio cholerae* O1 strains isolated in Lima, Peru, from 1991 to
657 1995. J Clin Microbiol 35:1151–1156.

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

660 9. Weill F-X, Domman D, Njamkepo E, Almesbahi AA, Naji M, Nasher SS, Rakesh A, Assiri
661 AM, Sharma NC, Kariuki S, Pourshafie MR, Rauzier J, Abubakar A, Carter JY, Wamala
662 JF, Seguin C, Bouchier C, Malliavin T, Bakhshi B, Abulmaali HH., Kumar D, Njoroge SM,
663 Malik MR, Kiiru J, Luquero FJ, Azman AS, Ramamurthy T, Thomson NR, Quilici M-L.
664 2018. Genomic insights into the 2016–2017 cholera epidemic in Yemen. Nature 565:230–
665 233.

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

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

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

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

675 355:649–651.

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

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

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

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

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

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

697 20. Octavia S, Salim A, Kurniawan J, Lam C, Leung Q, Ahsan S, Reeves PR, Nair GB, Lan R.
698 2013. Population Structure and Evolution of Non-O1/Non-O139 *Vibrio cholerae* by
699 Multilocus Sequence Typing. PLoS One 8:e65342.

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

703 22. Bwire G, Sack DA, Almeida M, Li S, Voeglein JB, Debes AK, Kagirita A, Buyinza AW,
704 Orach CG, Stine OC. 2018. Molecular characterization of *Vibrio cholerae* responsible for
705 cholera epidemics in Uganda by PCR, MLVA and WGS. PLoS Negl Trop Dis
706 12:e0006492.

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

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

712 Trends Microbiol 18:46–54.

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

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

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

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

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

728 30. Maiden MCJ, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND.
729 2013. MLST revisited: The gene-by-gene approach to bacterial genomics. Nat Rev
730 Microbiol 11:728–736.

731 31. Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo T R, Jaykus L-A, DePaola A.
732 2008. Determination of Molecular Phylogenetics of *Vibrio parahaemolyticus* Strains by
733 Multilocus Sequence Typing. J Bacteriol 190:2831–2840.

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

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

741 34. Vogler AJ, Birdsell DN, Lee J, Vaissaire J, Doujet CL, Lapalus M, Wagner DM, Keim P.
742 2011. Phylogeography of *Francisella tularensis* ssp. holarctica in France. Lett Appl
743 Microbiol 52:177–180.

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

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

748 typing methods for outbreak detection and epidemiological surveillance. Eurosurveillance
749 18:20380.

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

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

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

759 40. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. 2014. Evaluation of
760 Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. PLoS One
761 9:e87991.

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

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

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

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

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

774 46. Orata FD, Kirchberger PC, Méheust R, Barlow EJ, Tarr CL, Boucher Y. 2015. The
775 Dynamics of Genetic Interactions between *Vibrio metoecus* and *Vibrio cholerae* , Two
776 Close Relatives Co-Occurring in the Environment. Genome Biol Evol 7:2941–2954.

777 47. Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL,
778 Polz MF. 2011. Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create
779 Endemicity within Global *Vibrio cholerae* Populations. MBio 2.

780 48. Neumann B, Prior K, Bender JK, Harmsen D, Klare I, Fuchs S, Bethe A, Zühlke D, Göhler
781 A, Schwarz S, Schaffer K, Riedel K, Wieler LH, Werner G. 2019. A Core Genome
782 Multilocus Sequence Typing Scheme for *Enterococcus faecalis*. J Clin Microbiol
783 57:e01686-18.

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

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

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

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

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

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

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

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

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

820 58. Reimer A, Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L,
821 Olsen-Rasmussen M, Frace M, Sammons S, Dahourou G, Boncy J, Smith A, Mabon P,
822 Petkau A, Graham M, Gilmour M, Gerner-Smidt P. 2011. Comparative Genomics of *Vibrio*

823 *cholerae* from Haiti, Asia, and Africa. Emerg Infect Dis 17:2113.

824 59. Salim A, Lan R, Reeves PR. 2005. *Vibrio cholerae* pathogenic clones. Emerg Infect Dis
825 11:1758–1760.

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

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

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

832 63. Lucidarme J, Hill DMC, Bratcher HB, Gray SJ, du Plessis M, Tsang RSW, Vazquez JA,
833 Taha M-K, Ceyhan M, Efron AM, Gorla MC, Findlow J, Jolley KA, Maiden MCJ, Borrow
834 R. 2015. Genomic resolution of an aggressive, widespread, diverse and expanding
835 meningococcal serogroup B, C and W lineage. J Infect 71:544–552.

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

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

845 66. World Health Organization. 2017. Weekly epidemiological record.

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

851 68. Sardar T, Mukhopadhyay S, Bhowmick AR, Chattopadhyay J. 2013. An optimal cost
852 effectiveness study on Zimbabwe cholera seasonal data from 2008-2011. PLoS One
853 8:e81231.

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

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

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

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

865 73. Childers BM, Klose KE. 2007. Regulation of virulence in Vibrio cholerae: The ToxR
866 regulon. Future Microbiol 2:335–344.

867 74. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for
868 scrupulous assemblies. Genome Biol 19:153.

869 75. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
870 Bioinformatics 26:2460–2461.

871 76. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S,
872 Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil
873 LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O,
874 Vonstein V, Wilke A, Zagnitko O, Formsma K, Kubal M, Vonstein V, Stevens R, McNeil
875 LK, Edwards RA, Pusch GD, Reich C, Glass EM, Olsen GJ, Paczian T, Overbeek RA,
876 Meyer F, Vassieva O, DeJongh M, Osterman AL, Disz T, Best AA, Gerdes S, Parrello B,
877 Bartels D, Olson R, Paarmann D. 2008. The RAST server: rapid annotations using
878 subsystems technology. BMC Genomics 9:75.

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

882 78. Liang K, Islam MT, Hussain N, Winkjer NS, Im MS, Rowe LA, Tarr CL, Boucher Y. 2019.
883 Draft Genome Sequences of Eight Vibrio sp. Clinical Isolates from across the United States
884 That Form a Basal Sister Clade to *Vibrio cholerae*. Microbiol Resour Announc 8:e01473-
885 18.

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

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

892 81. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species
893 delimitation with confidence intervals and improved distance functions. BMC
894 Bioinformatics 14:60.

895 82. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM :

896 assessing the quality of microbial genomes recovered from isolates, single cells, and
897 metagenomes. Genome Res 25:1043–1055.

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

900 84. Jolley KA, Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation at
901 the population level. BMC Bioinformatics 11:595.

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

905 86. Brock G, Pihur V, Datta S, Datta S. 2008. clValid: An R Package for Cluster Validation. J
906 Stat Softw 25:1–22.

907 87. R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna,
908 Austria.

909 88. Canty A, Ripley BD, others. 2017. boot: Bootstrap R (S-Plus) Functions. R Packag version
910 1.

911 89. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New
912 York.

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

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

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

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

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

925 95. Hagberg A, Swart P, S Chult D. 2008. Exploring network structure, dynamics, and function
926 using NetworkX, p. 11–16. *In* Varoquaux G, Vaught T, Millman J (ed), Proceedings of the
927 7th Python in Science Conference (SciPy2008).

928 96. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B,
929 Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of

930 Biomolecular Interaction Networks. Genome Res 13:2498–2504.

931

932 **FIGURE LEGENDS**

933 **Fig. 1.** Pairwise allelic differences for all isolates used in this study. Both plots show the
934 frequency of allelic mismatches in pairwise comparisons. A) Pairwise comparisons of up to
935 2,443 allelic differences are shown. Major peaks are shaded. B) Comparisons with up to 500
936 allelic differences are shown. Pairwise comparisons of only clinical isolates are shown in red.
937 Vertical lines indicate the outbreak threshold (red) and sublineage threshold (blue).

938

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

944

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

955 **Fig. 4.** Phylogenetic tree of 1,146 *V. cholerae* isolates (excluding the 116 isolates from the
956 Yemen cholera outbreak) reconstructed using Parsnp v1.2 (92). All group inside the PG lineage
957 (7th pandemic El Tor, El Tor progenitor, El Tor sister, Classical, and Classical sister) are
958 collapsed. Outer rings represents clustering by sequence type based on the 2016 MLST scheme
959 by Kirchberger and colleagues (21), whereas the inner ring represents clustering based on the
960 sublineage threshold (i.e., 133 allelic differences). Branches of clinical strains are colored in red.
961 The phylogenetic tree is rooted with a basal lineage to *V. cholerae* (collapsed) (77, 78).
962
963 **Fig. 5.** Minimum spanning trees isolated when the outbreak threshold was applied to the
964 complete dataset of 1,262 isolates. A) All isolates that clustered together with the Haiti (dark
965 blue) and Yemen (light blue) isolates based on the clustering threshold of seven allelic
966 differences. B) All isolates that clustered with the Mozambique isolates (dark blue) based on the
967 clustering threshold of seven allelic differences. Additional Mozambique isolates that are not
968 part of the same outbreak cluster are also shown. Three isolates, two from Zimbabwe (green) and
969 one from the USA (orange), are connected as they share seven or fewer allelic differences with
970 the Mozambique isolates. In both panels, the size of the nodes is proportional to the number of
971 isolates. Length of the lines is proportional to the number of allelic differences and all
972 connections have less than or equal to seven allelic differences.
973
974 **Fig. 6.** cgMLST MST of all Yemen isolates and representative 7th pandemic El Tor strains. All
975 isolates connected by dotted lines share eight or more allelic differences (not drawn to scale). All
976 isolates connected with solid lines share seven or fewer allelic differences (i.e., they belong to
977 the same outbreak cluster; drawn to scale). Each node represents a cgST that is colored by year

978 of collection. The outbreak clusters are shaded by country.

979

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

987

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

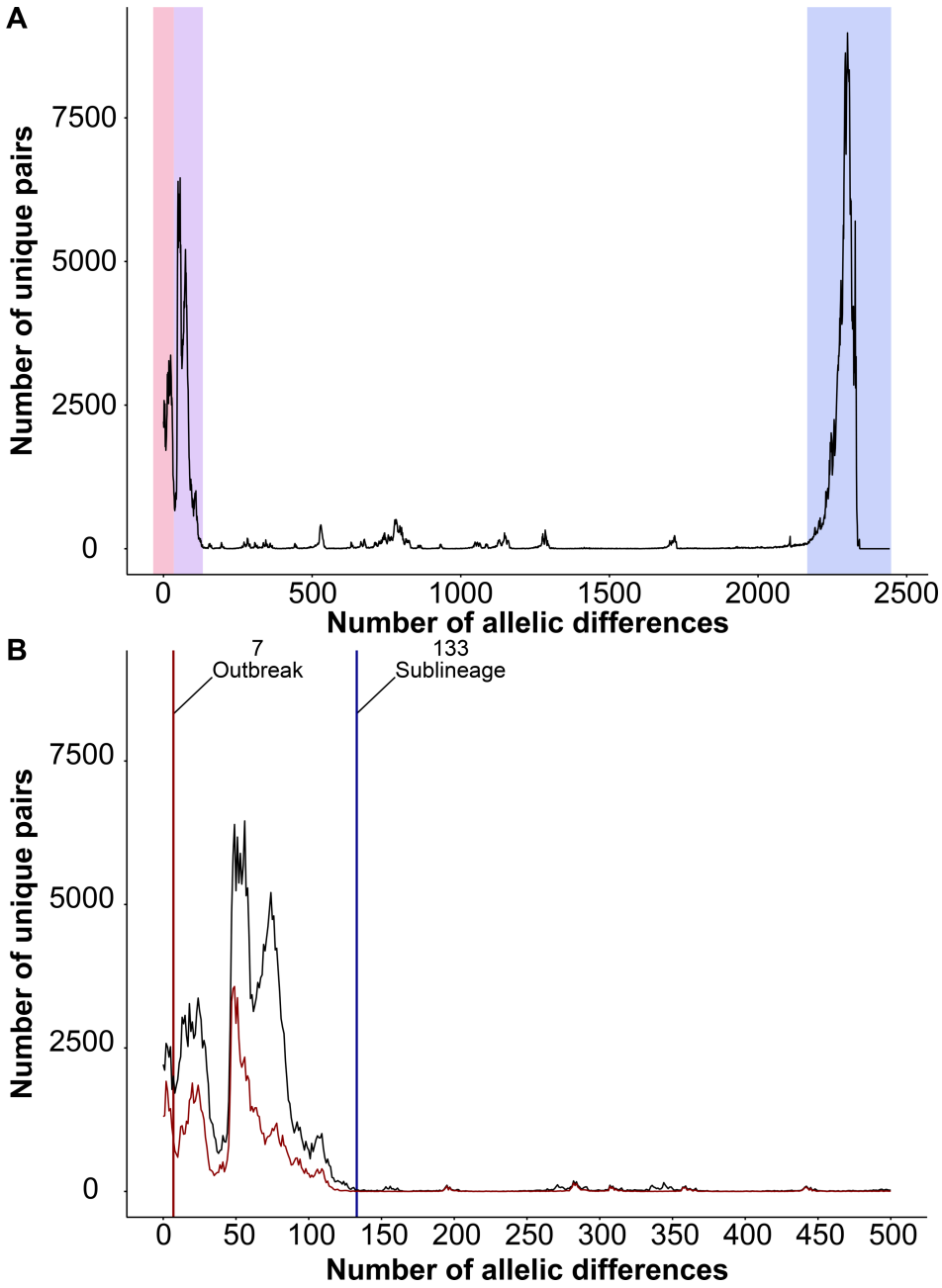
997

998 **Fig. 9.** Sublineage clusters of non-clinical environmental isolates that are not part of the PG
999 lineage. Clusters are constructed using networkX (95) and visualized with Cytoscape (96).
1000 Missing loci were assumed to contain the most common allele when calculating allelic

1001 differences. Isolates are connected only if they share 133 allelic differences or fewer with each

1002 other. Each node represents an isolate and is colored by the country of origin.

1003 **FIGURES AND FIGURE LEGENDS**



1004

1005 **Fig. 1.** Pairwise allelic differences for all isolates used in this study. Both plots show the

1006 frequency of allelic mismatches in pairwise comparisons. A) Pairwise comparisons of up to

1007 2,443 allelic differences are shown. Major peaks are shaded. B) Comparisons with up to 500

1008 allelic differences are shown. Pairwise comparisons of only clinical isolates are shown in red.

1009 Vertical lines indicate the outbreak threshold (red) and sublineage threshold (blue).

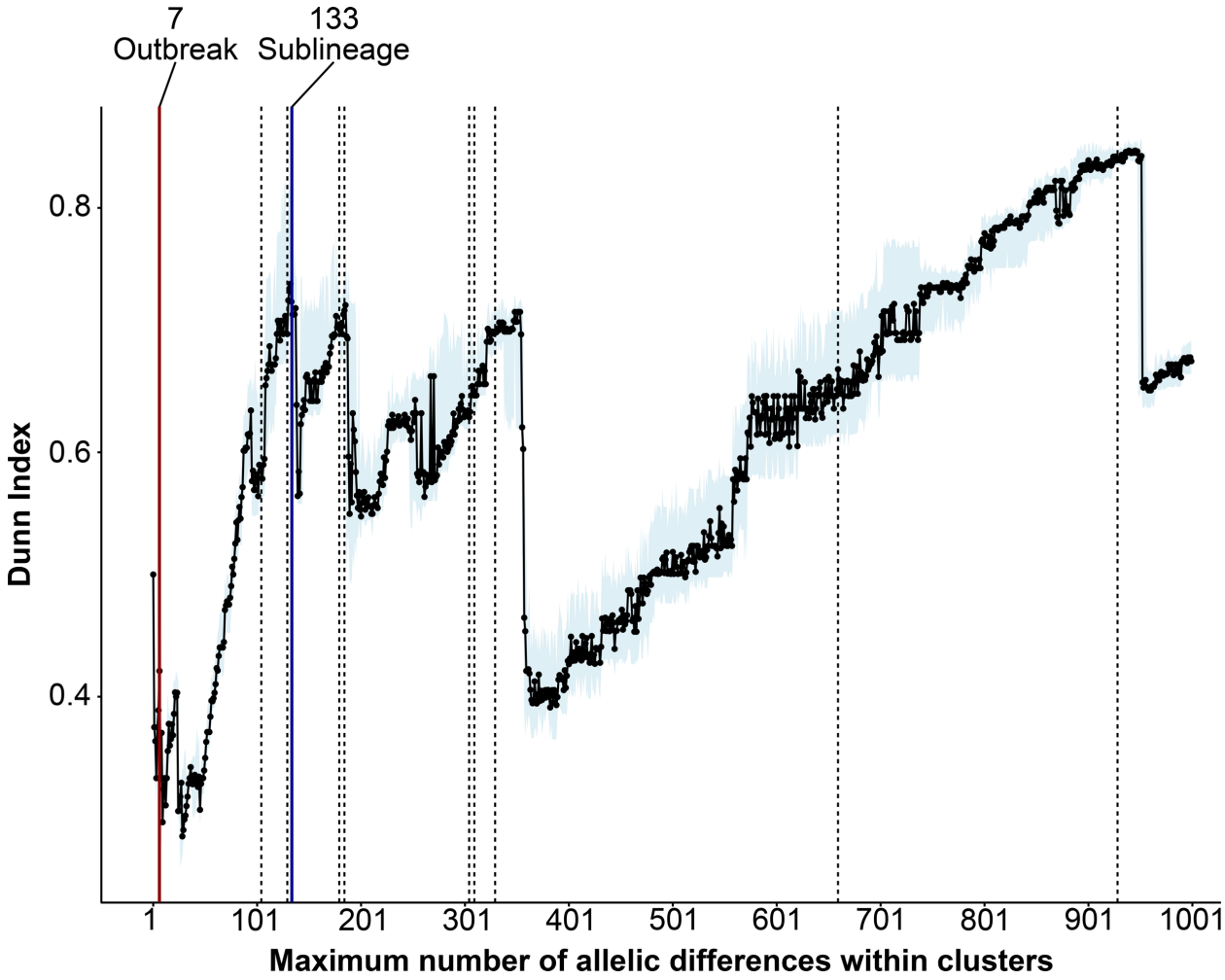


Fig. 2. Plot showing the Dunn Index for clustering thresholds ranging from 1 to 1,000 allelic differences. Each clustering threshold is bootstrapped 100 times. The median, plotted with the light blue shade, indicates the 25th to 75th percentile range. Red and blue vertical lines indicate the outbreak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

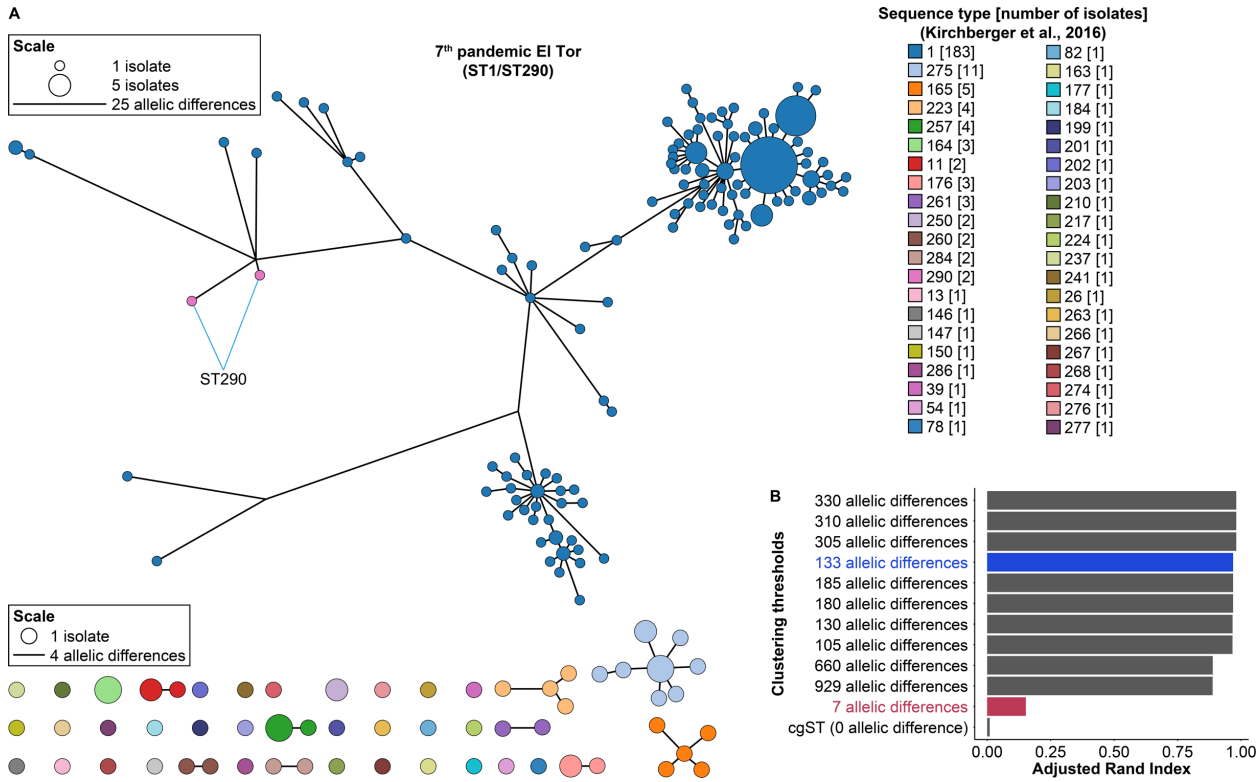


Fig. 3. Evaluation of network similarities between the cgMLST sublineage threshold and MLST ST. A) Networks of all sublineages identified using only *V. cholerae* isolates from Bangladesh (*n* = 255). Each cluster represents a sublineage and includes isolates with less than or equal to 133 allelic differences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

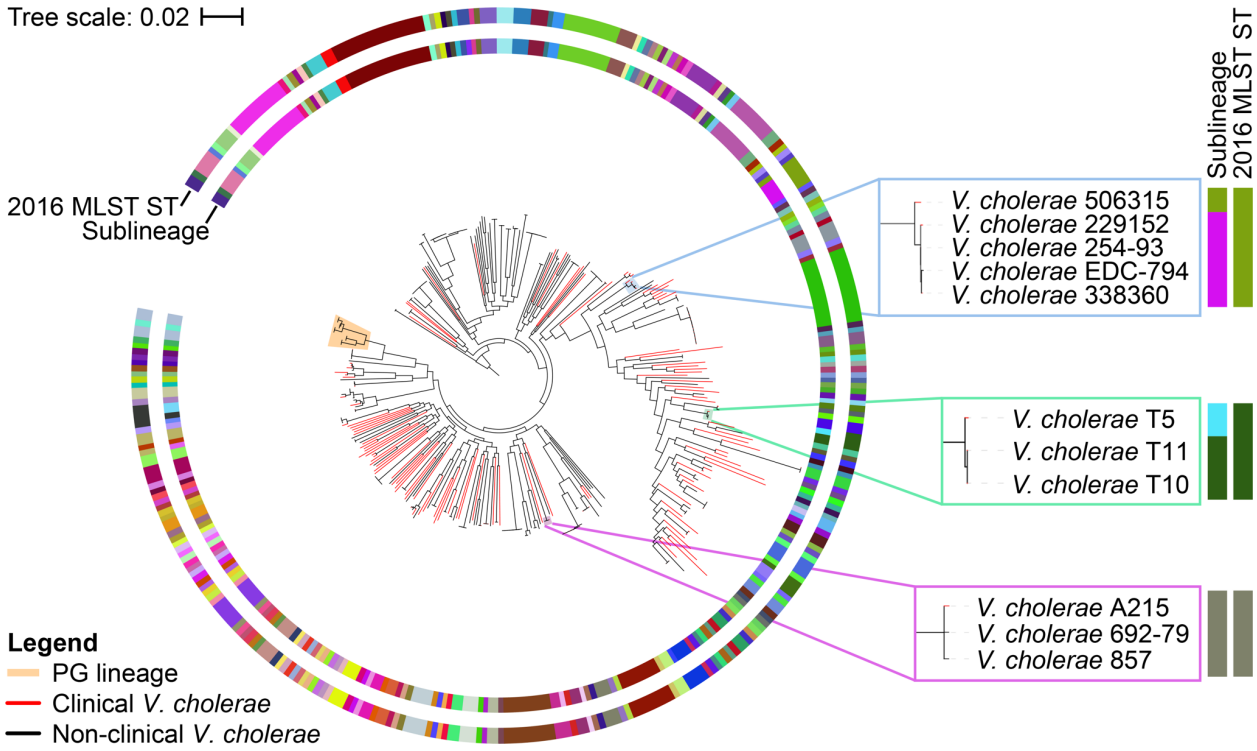


Fig. 4. Phylogenetic tree of 1,146 *V. cholerae* isolates (excluding the 116 isolates from the Yemen cholera outbreak) reconstructed using Parsnp v1.2 (92). All group inside the PG lineage (7th pandemic El Tor, El Tor progenitor, El Tor sister, Classical, and Classical sister) are collapsed. Outer rings represents 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) (77, 78).

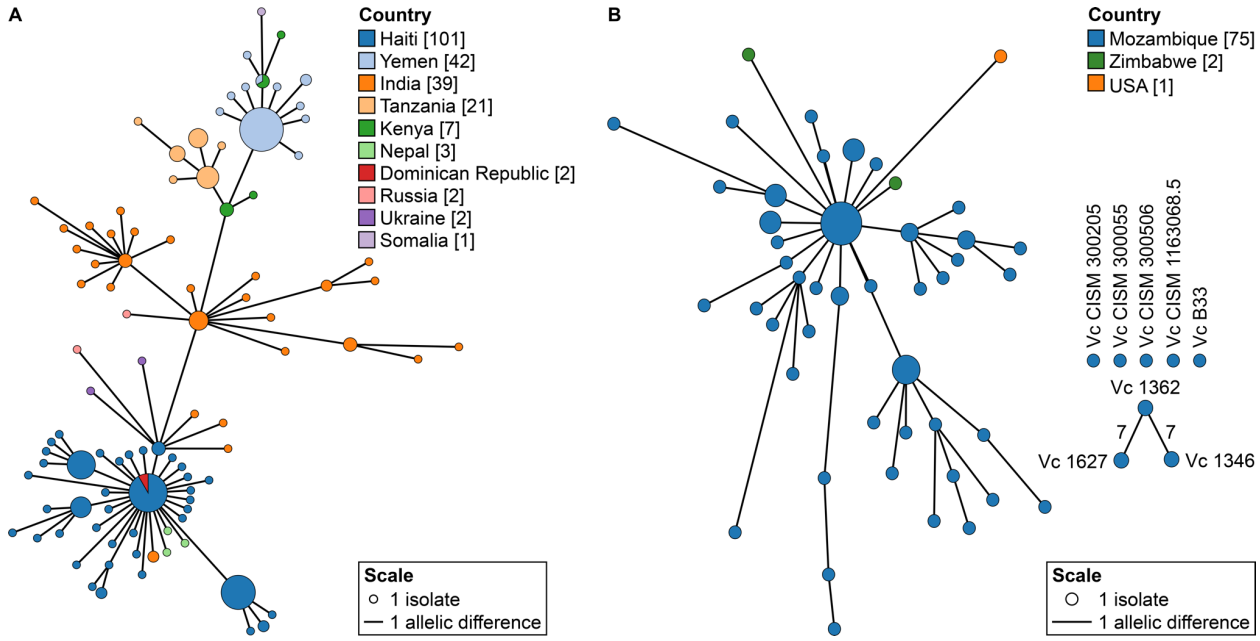


Fig. 5. Minimum spanning trees isolated when the outbreak threshold was applied to the complete dataset of 1,262 isolates. A) All isolates that clustered together with the Haiti (dark blue) and Yemen (light blue) isolates based on the clustering threshold of seven allelic differences. B) All isolates that clustered with the Mozambique isolates (dark blue) 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.

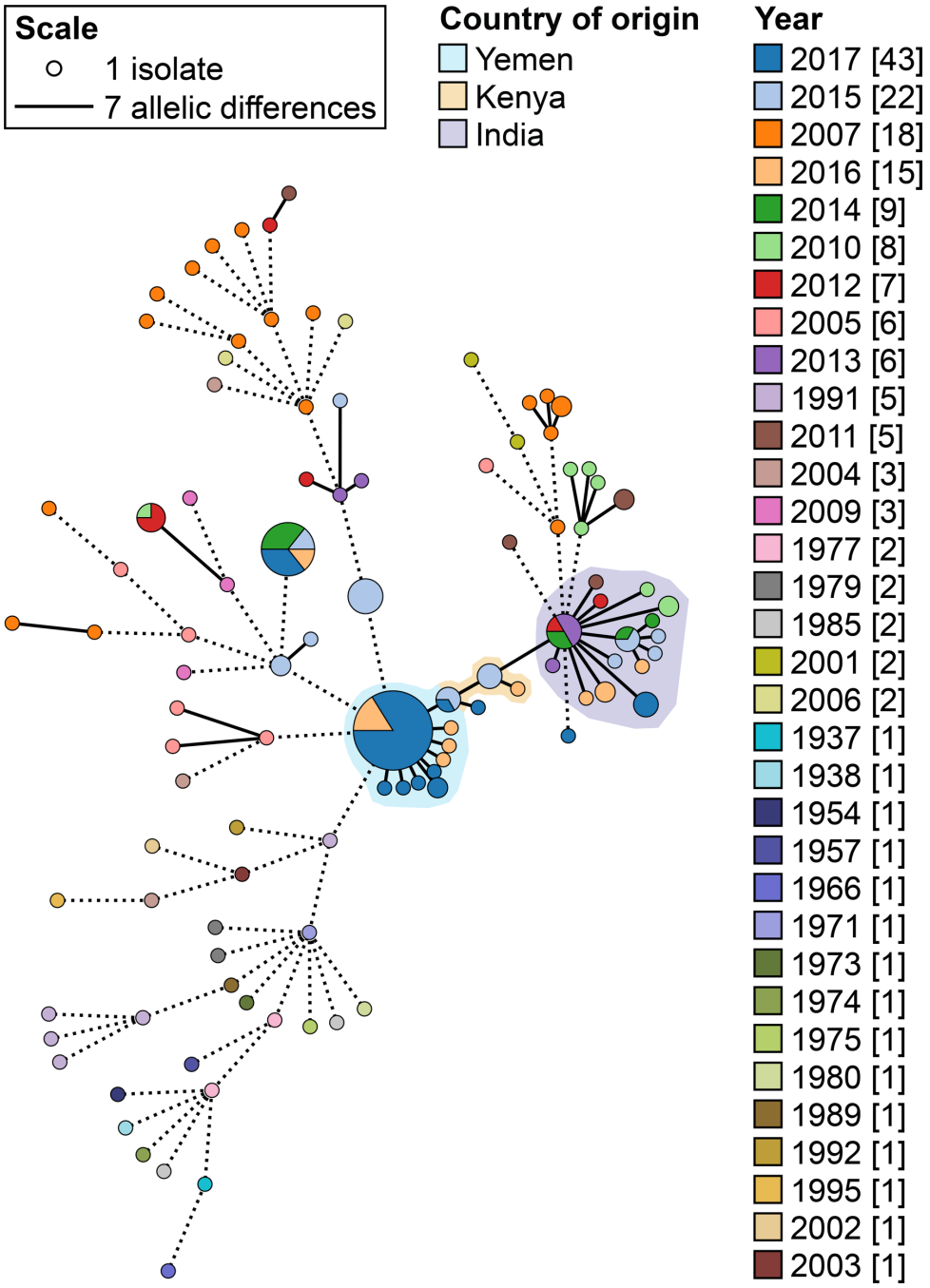


Fig. 6. cgMLST MST of all Yemen isolates and representative 7th pandemic El Tor strains. All isolates connected by dotted lines share eight or more allelic differences (not drawn to scale). All isolates connected with solid lines share seven or fewer allelic differences (i.e., they belong to the same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

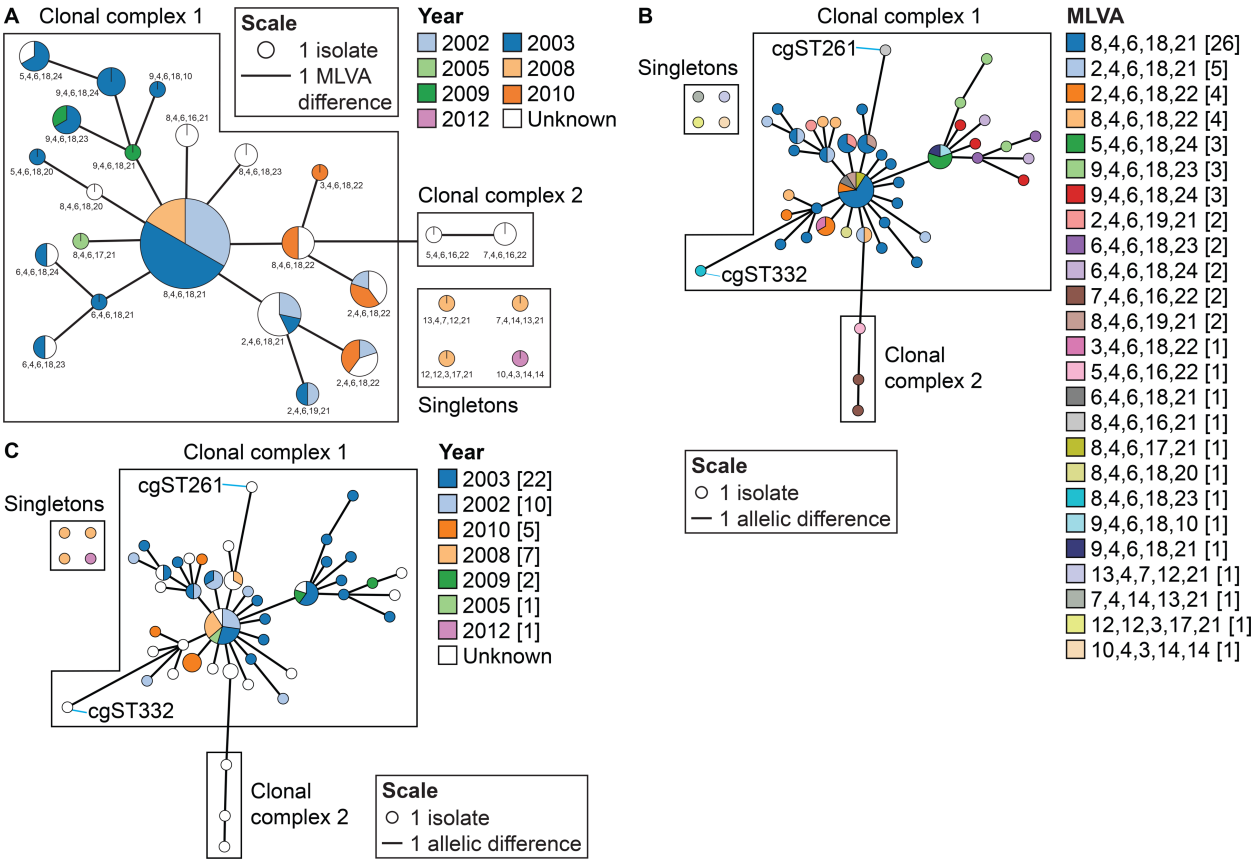
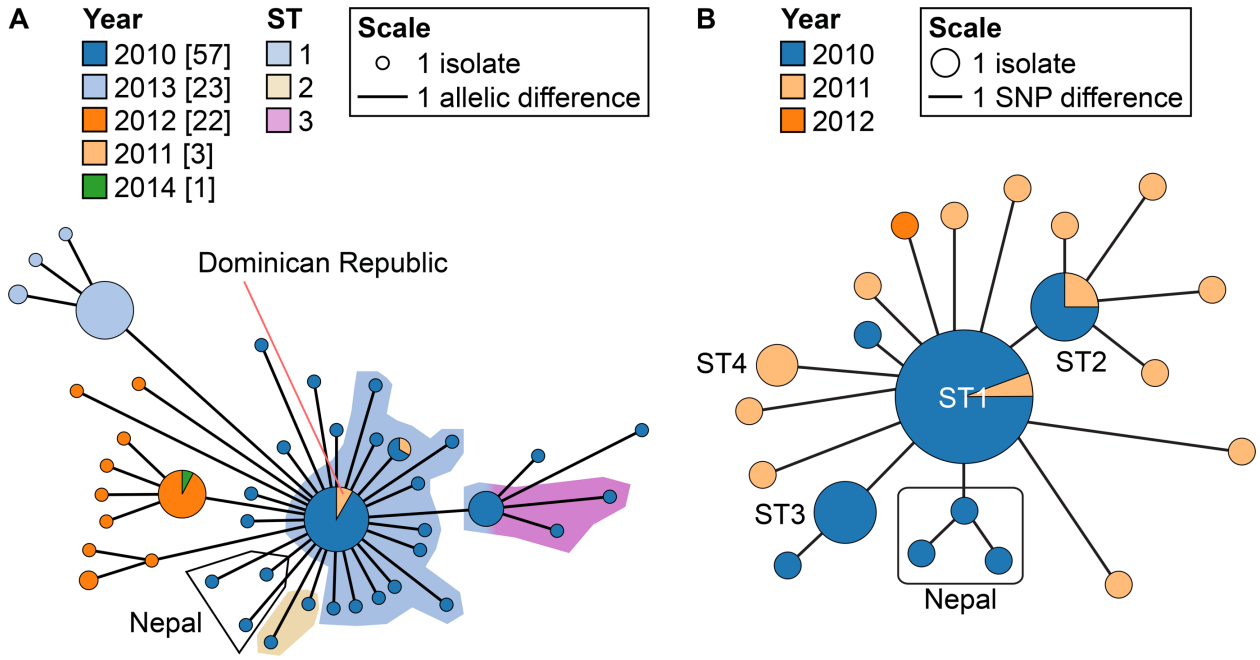


Fig. 7. Comparison between cgMLST and MLVA with a focus on the Mozambique isolates. A) Population structure of pandemic *V. cholerae* in Mozambique based on MLVA profiles by Garrine and colleagues (23). MST of the Mozambique isolates based on the cgMLST scheme colored based on B) MLVA profiles and C) 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.



1060

1061 **Fig. 8.** Comparison between cgMLST and SNP-based analysis with a focus on the Haiti outbreak

1062 and related isolates. A) MST of isolates from the 2010 cholera outbreak in Haiti. All lines

1063 indicate connections of four or fewer allelic differences. Each node represents a cgST which is

1064 colored by year of isolation. Background shading represent ST designations based on 45 high-

1065 quality SNPs by Katz and colleagues (14). Note that cgST66 contains a mix of color as it

1066 contains both ST1 and ST3. Any isolate from countries other than Haiti is indicated. The length

1067 of the lines is proportional to the number of allelic differences. B) MST constructed from whole

1068 genome SNP data (14). The length of the lines indicates the number of nucleotide substitutions.

1069 Size of the nodes is proportional to the number of isolates.

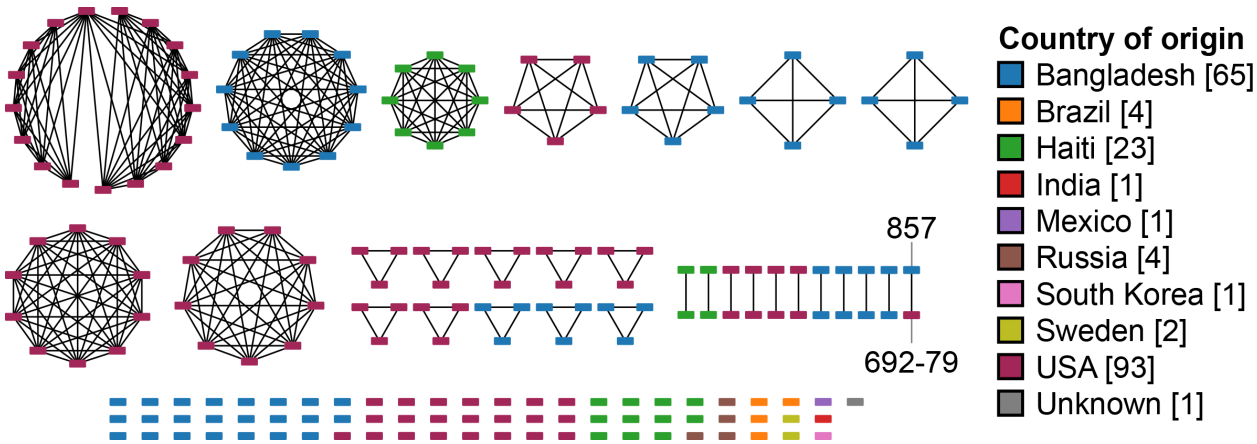


Fig. 9. Sublineage clusters of non-clinical environmental isolates that are not part of the PG lineage. Clusters are constructed using networkX (95) and visualized with Cytoscape (96). Missing loci were assumed to contain the most common allele when calculating allelic differences. Isolates are connected 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.