Cluster-driven evolution and modularity uncover paths of cholera emergence

Mario López-Pérez[a,b,c], Trudy-Ann Grant[a,b#], Jose M. Haro-Moreno[c#], Asier Zaragoza-Solas[c], Salvador Almagro-Moreno[a,b*].

[a]Burnett School of Biomedical Sciences, College of Medicine, University of Central Florida, Orlando, Florida, USA; [b]National Center for Integrated Coastal Research, University of Central Florida, Orlando, Florida, USA; [c]Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Alicante, Spain.

*Address correspondence to: Salvador Almagro-Moreno, samoreno@ucf.edu

# Contributed equally

Running title: Cluster-driven evolution of *Vibrio cholerae*

Keywords:

Pathogen emergence / *Vibrio cholerae* / aquatic pathogen / virulence evolution / population genomics / cluster divergence

ABSTRACT

Pathogen emergence remains one of the most pressing public health concerns of our times. Here, using the agent of cholera, the *Vibrio cholerae* pandemic cholera group (PCG) as a model system, we investigate the evolutionary dynamics that lead to the emergence of human pathogens from environmental populations. Genomic comparison of over 1,100 *V. cholerae* genomes including novel isolates from this study, reveal a generalized cluster-driven phylogeny and evolution of the species. Emergence of PCG is largely based on modular acquisition of mobile genetic elements including small gene clusters and allelic variations, with distinct phylogenomic clusters acting as differential reservoirs of virulence factors. Surprisingly, PCG encodes few unique genes, which are mostly encoded within the super-integron, however, many genes within the group exhibit an aberrant degree of positive selection. Our analyses provide a compelling scenario for the emergence of pandemic clones in *V. cholerae* stablishing a blueprint for other bacterial pathogens.

INTRODUCTION

The emergence of human pathogens is one of the most pressing and persistent public health concerns of modern times, with new outbreaks of varying magnitude occurring regularly around the world[1–4]. Pathogen emergence is a complex and multifactorial phenomenon that culminates with a microorganism acquiring the ability to colonize and harm the human host, cause outbreaks and persist in the environment between them[5]. To date, the specific factors that allow certain strains within a species to emerge as pathogens remain mostly enigmatic[4]. Elucidating how the interplay between the numerous molecular and environmental drivers lead to this complex phenomenon is critical to develop surveillance platforms to predict emergent events and potential sources of outbreaks[4].

The Vibrionaceae is a highly diverse family of aquatic bacteria that encompass several species that represent distinct paradigms of how pathogens can emerge from environmental populations[6,7]. *Vibrio cholerae* is the most widely studied member of the Vibrionaceae and has been a model organism for countless molecular, ecological and epidemiological studies. *V. cholerae* is the causative agent of the severe diarrheal disease cholera, a scourge that affects millions of people each year and is responsible for over 100,000 deaths annually[8]. The disease typically affects nations with limited access to clean water and poor sanitation and can be lethal if not properly treated[8,9]. Cholera has existed in the Indian subcontinent for centuries, however, the disease started to spread worldwide after 1817 causing seven pandemics[8–10]. The first six pandemics are thought to have been caused by the Classical biotype *V. cholerae* O1, a strain that is now considered virtually extinct[8–11]. The seventh and current pandemic is caused by the El Tor biotype of *V. cholerae* O1 and started in Indonesia in 1961[12,13]. The recent cholera epidemics have increased in duration, intensity, and frequency warranting urgent attention to this infectious disease[8,9,14]. Furthermore, extended periods of warm weather, driven by climate change, have provided suitable conditions for the proliferation of pathogenic *Vibrio* spp.[14,15]. For instance, a recent multidecadal study demonstrated a

64 steady increase in the abundance of pathogenic Vibrios, including *V. cholerae*, over the
65 past half-century[16]. Overall, these concerning patterns only exacerbate the problem of
66 the emergence and reemergence of pathogenic Vibrios, the spread of virulence genes
67 and their proliferation[17].

68 Interestingly, only a limited number of *V. cholerae* strains can cause pandemic
69 cholera[18–20]. Specifically, only those confined to a phylogenetic group, the pandemic
70 cholera group (PCG), can cause the disease[19,21,22]. Even though the mechanisms of
71 cholera pathogenesis in humans have been extensively studied over the last several
72 decades, the evolutionary events and molecular constraints that explain why only
73 members from PCG can cause pandemic cholera remain unknown[23–28]. Some of the
74 major general events that are essential for the emergence of pandemic *V. cholerae* from
75 environmental populations include the acquisition of the lipopolysaccharide O1 antigen
76 cluster and the mobile genetic elements (MGEs) such as CTXΦ phage, Vibrio
77 pathogenicity Island-1 (VPI-1) and Vibrio pathogenicity Island-2 (VPI-2)[11]. Among
78 several other differences with Classical, El Tor strains have also acquired two unique
79 pathogenicity islands (PAIs): *Vibrio* Seventh Pandemic (VSP) 1 and 2. CTXΦ encodes
80 the genes for the cholera toxin, which is responsible for the severity of the profuse
81 diarrhea associated with cholera[29]. The toxin coregulated pilus (TCP), an essential
82 colonization factor, is encoded within VPI-1[30,31]. [35]This island also encodes several major
83 virulence regulators such as ToxT and TcpP[31,32]. VPI-2, among other, encodes the genes
84 for sialic acid utilization, which confers a competitive advantage in the gut[33]. VSP-I
85 encodes a regulator that plays a role in intestinal colonization and chemotaxis[34]. VSP-II
86 is associated with environmental survival and fitness of the bacterium[35–37]. It was recently
87 found that VSP-II and VPI-2 encode systems that prevent the uptake of foreign genetic
88 material[38]. These MGEs are always found in the PCG, however, they are not exclusive
89 to them as they are also encoded in some environmental strains of *V. cholerae*[39–41].
90 However, their abundance, specific distribution and dynamics within the *V. cholerae*
91 species remains poorly understood.

92 Horizontal acquisition of key virulence genes is a critical step in the emergence
93 of toxigenic clones of *V. cholerae*, however, it is not sufficient to explain why the ability
94 to cause cholera is limited to clones belonging to the PCG. This phylogenetically confined
95 distribution together with the presence of CTXΦ and the PAIs in some environmental
96 strains strongly indicate that other barriers and requirements exist that stringently limit
97 their emergence. To begin determining these barriers and the evolutionary origins of
98 pandemic clones of *V. cholerae*, we recently investigated the genomic prerequisites that
99 must be present in a population before a pathogenic clone can emerge from an
100 environmental gene pool[22]. We determined that the environmental ancestor of the
101 pandemic group had a particular genomic background containing a specific set of alleles
102 of core genes that conferred preadaptations to virulence and enhanced its potential to
103 give rise to pandemic disease, what we term virulence adaptive polymorphisms
104 (VAPs)[22]. Interestingly, our data shows that VAPs are present in environmental
105 populations and confer virulence-associated traits to strains encoding them. Overall, our
106 results indicate that there are core genomic factors beyond the acquisition of MGEs that
107 are critical for the emergence of pandemic *V. cholerae*. To date, the unique core genes
108 and allelic variations in the genomic background of pandemic *V. cholerae* that explain
109 their narrow distribution remain enigmatic.

110 In this study, we investigated the evolutionary paths that led with the emergence
111 of pandemic *V. cholerae* from environmental populations. To do this, first, we isolated
112 environmental *V. cholerae* strains from the Indian River Lagoon (IRL), a highly diverse
113 and expansive estuary located in Eastern Florida. Genome comparisons of these strains
114 against *V. cholerae* PCG unveil a phylogeny in which strains form cluster-like groups
115 following a similar pattern as what we recently found with *Vibrio vulnificus*[42,43]. We
116 extended our analyses to over 1,100 publicly available genomes and found that *V.*
117 *cholerae* strains are divided into four major clusters and numerous other minor clades.
118 We found that PCG belong to one large cluster (C24) and is located within one lineage
119 shared with among other environmental strains, the pandemic cholera lineage (PCL).

120 This topology suggests a scenario where clusters may differ in their virulence gene
121 content. In order to test this, first, we first investigated the evolutionary history,
122 environmental reservoirs and abundance of the O1 capsule and MGEs associated with
123 cholera emergence. We found widespread modularity of these elements, with distinct
124 phylogenomic clusters acting as differential reservoirs of virulence factors. However,
125 certain allelic variations and scarcely distributed islands act as bottlenecks drastically
126 narrowing the distribution of PCG-like gene clusters. Furthermore, PCG encodes very
127 few unique genes, which are mostly encoded within the super-integron, however, many
128 genes within the group exhibit an aberrant degree of positive selection, which may further
129 explain the confined distribution of these isolates. Our analyses provide a compelling
130 scenario for the emergence of pandemic clones in *V. cholerae* stablishing a blueprint for
131 other bacterial pathogens.
132

133 **RESULTS AND DISCUSSION**

134 **Isolation of *Vibrio cholerae* in the Indian River Lagoon.** *V. cholerae* is a natural
135 inhabitant of aquatic ecosystems where it can be found as free-living or attached to
136 abiotic or biotic surfaces[44,45]. We recently stablished two sampling sites along the eastern
137 coast of Florida where we isolated pathogenic strains of *Vibrio vulnificus*[43]. We sampled
138 those two sites, Feller's House (FH) and Shepards Park (SP), to investigate the potential
139 presence of *V. cholerae* strains (**Fig. 1A**). Using the isolation, enrichment and selection
140 methods described below, we identified a total of 178 potential *V. cholerae* isolates from
141 the different fractions at both locations during our sampling events (42 FH and 136 SP).
142 In order to confirm those isolates, we performed a PCR screening for the species-specific
143 gene *ompW*[46]. Interestingly, only 65 isolates were confirmed as bona-fide *V. cholerae* all
144 of them isolated from SP (**Fig. 1A**). *V. cholerae* isolates from SP were collected from all
145 the fractions that we examined (sediment, cyanobacteria, particle-associated and water)
146 except for oysters. Subsequently, we performed PCR-based fingerprinting of the strains
147 to reduce redundancies and minimize clonality. We used a combination of PCR-based

148 fingerprinting analyses, VCR-PCR and ERIC-PCR[47]. Based on fingerprinting patterns
149 and cluster analysis performed using BIONUMERICS v8.0 we generated a dendrogram
150 to group these observed patterns. Using this resulting dataset, we selected 35 isolates
151 with unique fingerprinting patterns for whole genome sequencing (**Fig. 1B and S1**).
152
153 **Phylogenomic and clustering of *V. cholerae* strains.** The 35 *V. cholerae* strains that
154 we sequenced were isolated from a diverse set of fractions: 6 from sediment, 10 from
155 cyanobacteria, 9 from the free-living water fraction and 10 from the particle-associated
156 one (**Table S1**). We examined the relationship of these strains with *V. cholerae* PCG by
157 analyzing the average nucleotide identity (ANI) in a pairwise comparison against *V.*
158 *cholerae* N16961, an O1 strain from the El Tor biotype (**Fig. S2**). The minimum values
159 obtained were ca. 97.5, always within the established cut-off for strains of the same
160 species. However, up to nine environmental strains showed a greater divergence with
161 the PCG (ANI > 98.2%) than with the rest of the strains (**Fig. S2**). Interestingly,
162 phylogenomic tree analysis based on whole-genome single nucleotide polymorphism
163 (SNP) (230,824 SNPs identified in 3,19 Mb of core genome) show that the majority of
164 the IRL strains form two major clusters one of them including the toxigenic strains of *V.*
165 *cholerae* (**Fig. 1B**). This evolutionary relationship shares similarity with that of *V.*
166 *vulnificus*, in which the pathogenic strains capable of causing septicemia in humans
167 belong to one clearly defined cluster[42]. This prompted us to characterize the *V. cholerae*
168 species to investigate whether the evolution of the pandemic strains follow a similar
169 cluster-based pattern as this would shed important light in the emergence and evolution
170 of the pathogen.
171 First, we collected the *V. cholerae* genomes available in public databases (based
172 on NCBI classification accessed in July 2020; see Materials and Methods) together with
173 our environmental isolates. Naturally, there is a significant bias in the databases towards
174 *V. cholerae* O1 and O139 strains, which are mostly clonal. Furthermore, it is likely that
175 several *V. cholerae* environmental strains that were uploaded in the databases are also

176 from clonal origin due to enrichment during the early steps of isolation. Therefore, in
177 order to obtain an unbiased phylogeny, we dereplicated all these genomes to an identity
178 of ANI>99%. From a total of 1,127 initial genomes, we obtained 227 non-clonal reference
179 *V. cholerae* clades (**Table S2**). To maximize the consistency of our analyses, we built
180 both a SNP-based phylogeny and genome clustering using ANI to determine the groups
181 that make up the *V. cholerae* species. Clustering using a threshold of 98% of ANI shows
182 that the species encompasses 34 clusters that include *V. cholerae* strains (**Table S3**).
183 The application of an ANI>97% results in a single cluster indicating that boundary of
184 intra-population sequence diversity in *V. cholerae* species is 97%. Next, we used a total
185 of 458,798 SNPs from a core genome size of 2.46 Mb to build the species phylogeny
186 using non-clonal reference clades and IRL strains (**Fig. 1C**). The species phylogeny
187 topology based on SNPs or ANI clustering are virtually identical, with some minor
188 exceptions that might be due to certain strains having a convergent history. Interestingly,
189 both approximations reveal that four major clusters account for 69.2% of the non-clonal
190 clades (157 out of a total of 227) including 31 of the 35 IRL isolates from this study (**Table**
191 **S3**). All IRL isolates are grouped together with previously isolated strains except for
192 strains IRLE0050 and IRLE0074 which formed a separate cluster (**Table S3**). Cluster 24
193 (C24) is the largest of the four main clusters containing 68 non-clonal clades including
194 PCG. On the other hand, most IRL strains belong to C5 (22 of the 35) (**Fig. 1C**). We did
195 not find a pattern that associate the phylogeny of the IRL isolates and their fraction of
196 isolation. The C24 phylogeny also reveals that PCG groups with nine other non-clonal
197 clades, including one from this study IRLE0178, into a separate branch in the
198 phylogenomic tree, what we have termed the Pandemic Cholera Lineage (PCL) (**Fig.**
199 **1C**). This provides us with a novel hierarchical approach based on three levels to
200 decipher the evolutionary history of pandemic strains of *V. cholerae*: (from highest to
201 lowest) a) the C24 within all *V. cholerae* strains, b) the PCL within C24 non-clonal clades
202 and c) the clonal *V. cholerae* pandemic strains (PCG) within the PCL lineage.
203

204 **Lipopolysaccharide (LPS) cluster**. The LPS plays numerous roles in the pathogenesis
205 of pandemic *V. cholerae* (e.g., endotoxic activities, immunological responses, etc.[48–50])
206 and is critical to determine the antigenic properties and classification of the bacterium[50].
207 Specifically, given that only O1 strains, together with the extinct O139, can cause
208 pandemic cholera we examined its LPS cluster to determine its distribution and
209 evolutionary history within the *V. cholerae* species (**Fig. 2**). The O1 gene cluster is
210 29.7Kb long and encompasses approximately 26 ORFs with potential or demonstrated
211 function (some are truncated preventing an exact number). The *rfa* gene cluster (LPS
212 core synthesis) is encoded at the 5' end of the island (**Fig. 2A** blue arrows). The *rfa*
213 cluster is widely distributed across all clades, with the notable exception of RfaF (ADP-
214 heptose--LPS heptosyltransferase 2, VC_0235) and RfaL (O-antigen ligase, VC_0237)
215 present in ca. 35 non-clonal clades. Nonetheless, the genes in the *rfa* cluster exhibit the
216 highest divergence within the LPS cluster (median identity 86-88%) and accumulate
217 numerous synonymous polymorphisms (**Fig. 2A**). Indicating that, like *V. vulnificus*, they
218 appear to be a preferred site to act as hot spots for the recombination of this island due
219 to the conservation and synteny of these genes[42]. The distribution of the central region
220 of the LPS cluster is much more restricted and is only present in 17 *V. cholerae* non-
221 clonal clades, however, those that encode it share 100% identity with the region of PCG.
222 This region contains the *rfb* cluster which is associated with antigen-O synthesis[51] (**Fig.
223 2A,** green arrows). Interestingly, only 5 out the 17 clades that encode this region belong
224 to the PCL, indicating that there is no direct correlation between the phylogeny and its
225 presence in PCG. The LPS cluster of PCG encodes four transposases, which likely
226 increase the variability within the island (**Fig. 2A and 2B**). These IS elements are widely
227 distributed among *V. cholerae* strains providing insight into how small gene clusters can
228 be introduced in the island. For instance, even though strains VC22 (C24) and 2479-86
229 (C27) encode the same version of the PCG LPS cluster (identity >99%), they do not
230 encode two of these transposases (**Fig. 2B**). Interestingly, these 2 transposases are
231 inserted within *rfbT* in PCG, leaving a truncated version of the gene (**Fig. 2B and 2C**).

232 *rfbT* encodes a methyltransferase that is responsible for seroconversion between *V.*
233 *cholerae* O1 Ogawa and Inaba serotypes, suggesting a major role for these IS
234 sequences in the emergence of PCG[52]. Furthermore, several strains from C23 encode
235 the same version of the hypervariable region as VC22 and 2479-86 (identity 99%) but
236 the *rfa* cluster is different (**Fig. 2B**). The first three genes of the island encode different
237 alleles of *rfaF*, *rfaG* and *rfaL*, and the identity of the remaining three is 95% (**Fig. 2B**).
238 The *rfa* cluster pattern of these strains is similar to that of strain A325 (C30) (**Fig. 2B**). In
239 addition, in this strain the two transposases that truncate the *rfb*T gene in the PCG are
240 inverted maintaining an intact version of the gene (**Fig. 2C**). Overall, we found a
241 remarkable degree of conservation of the LPS cluster, which is present in almost twenty
242 clades encoding a PCG-like version. This is particularly striking as capsular
243 hypervariability is a universal bacterial characteristic, suggesting some strong
244 evolutionary pressure maintaining this version of the island widespread in the population.
245 It appears that the island is a) exchanged by homologous recombination leading to the
246 whole gene cluster being conserved and b) the presence of IS elements increases the
247 variability by truncating genes such as *rfbT*.
248
249 **Abundance and distribution of virulence-associated mobile genetic elements in *V.***
250 ***cholerae* species.** There are several horizontally acquired MGEs that are directly
251 associated with the emergence of pandemic clones of *V. cholerae*. After we stablished
252 the phylogeny and clustering of the *V. cholerae* species, we investigated the distribution,
253 evolutionary dynamics and composition of the different virulence-associated MGEs.
254
255 *1. CTXΦ phage*. The genes for cholera toxin (CT) are encoded within the CTXΦ
256 phage[29]. The production of CT in the intestine is responsible for the severity of the
257 profuse diarrhea associated with cholera. CT leads to the secretion of electrolytes and
258 nutrients such as fatty acids into the lumen of the small intestine[8,53,54]. Due to the
259 lysogenic nature of CTXΦ, the insertion and deletion dynamics of this phage enables

260 recombination of genes, which leads to diversity within the pandemic strains ultimately
261 serving as an opportunity to increase the pathogenic potential of pandemic strains[29,55,56].
262 Inquiry into the distribution and composition of this element within *V. cholerae* shows that
263 there is no environmental strain that recapitulates all the phage genes of PCG. At most,
264 five clades and two clusters encode 10 of the 14 genes that make up the CTXΦ phage
265 of *V. cholerae* N16961: three from C24 (represented by AAS91, PivertUAT4Aug and
266 V060002) and two from C17 (3541-04 and 3528-08) (**Fig. 3A**). Four of these strains
267 encode the two subunits of the cholera toxin (CtxAB). However, genes coding for the
268 duplicated transcriptional regulator RstR (VC_1455 and VC_1464) are only found in one
269 strain: AAS91 (**Fig. 3A**, brown shade). This strain belongs to the PCL and encodes the
270 complete version of CTXΦ except for VC_1460. This gene encodes a minor coat protein
271 pIII related to host interaction as part of the infection mechanism (**Fig. 3B**)[57]. Instead, a
272 transposase has been inserted in this locus in strain AAS91, followed by a APH(3')-II
273 family aminoglycoside O-phosphotransferase (**Fig. 3B**). In addition, AAS91 has a third
274 copy of the triplet (RstA-RstB-RstR) instead of the two copies present in PCG strains
275 (**Fig. 3B**). The second most complete version of CTXΦ in *V. cholerae* belongs to strains
276 V060002 from PCL and 3528-08 (C17). Both strains have lost the gene coding for RstC
277 and encode a different allele of the transcriptional regulators RstR (**Fig. 3B**).
278 Genomic comparison of *V. cholerae* strains reveal that the phage variation
279 dynamics are of the "additive" kind, where small gene cassettes are integrated into the
280 3' part of the gene coding for RtxA[58]. All the CTXΦ variants have conserved intergenic
281 regions that separate and mark the insertion of these gene cassettes (*att* sites, marked
282 as black arrows). These types of insertions have been reported previously in *V. cholerae*
283 O139[59]. Interestingly, some strains such as 3528-08 (C17) can have up to five insertions
284 (**Fig. 3C**). The first two of the five cassettes are identical, and each cassette has a copy
285 of the two subunits of CT. The other three have lost the toxin genes and those coding for
286 RstB, RstR and RstA, and acquired a replication endonuclease (**Fig. 3C**). The

287 presence/absence of these cassettes is not correlated with the phylogeny and, when
288 present, they have a high identity indicating a fast turnover of these elements (**Fig. 3B**).
289 Furthermore, these cassettes can be integrated into other genome locations, as in the
290 case of the strain PivertUAT4Aug (**Fig. 3B**). In addition, using the sequence of *att* sites
291 as reference we also found a collection of cassettes in the same order (including the one
292 containing the two subunits of CT) in the close relative species *Vibrio mimicus*,
293 suggesting another potential donor of CTXΦ to non-toxigenic *V. cholerae* (**Fig S2**).
294
295 *2. Vibrio Pathogenicity Island-1 (VPI-1).* The toxin coregulated pilus (TCP), an
296 essential colonization factor, is encoded within VPI-1[30,31]. TCP mediates microcolony
297 formation, which is crucial for intestinal colonization, and acts as the receptor of CTXΦ[32].
298 VPI-1 also encodes two critical virulence regulators: ToxT and TcpP[31,32]. VPI-1 can be
299 transferred, via generalized transduction, between strains of *V. cholerae* and can also
300 form circular intermediates[60,61]. The variable region of the island is located between the
301 CDS VC_0821 (hypothetical protein) and VC_0846 (integrase) (**Fig. 4A**). The most
302 conserved version of this region can be found in nine clades, four of them within C24 (2
303 from PCL) and five from C17, including strain IRLE0081 isolated in this study (**Fig. 1C**).
304 The average identity of the proteins was highly conserved compared to PCG (*ca*. >97%),
305 however, we found some notable exceptions (**Fig. 4A**). First, among the proteins that
306 are part of the TCP operon, we found high divergence rates in the master regulator ToxT
307 (VC_0838) (**Fig. 4C**) and the toxin-coregulated pilus major pilin, TcpA (VC_0828) (**Fig.
308 4D**) with a median protein identity *ca*. 81% (**Fig. 4**). In addition, we found that the PCG
309 allele of the colonization factor TcpF (VC_0837) (**Fig. 4E**) was only present in three
310 strains (one PCL and two C17), in the rest, despite the conservation of synteny and
311 similarity of the whole TCP operon, the identity of this protein was less than 40%, likely
312 encoding a different allele of this poorly understood secreted protein[62–64] (**Fig. 4B**).
313 Finally, our genomic comparisons reveal that the gene coding for ToxR-activated gene
314 A, TagA (VC_0820) was absent from the C17 genomes, while being retained in three of

315 the four C24 genomes (**Fig. 4B**). Overall, despite widespread conservation in the synteny
316 and size of the clusters of genes that comprise VPI-1 in the *V. cholerae* species, the
317 existence of unique major allelic variations in a limited number of genes act as a stringent
318 bottleneck in the emergence of PCG.
319
320 *3. Vibrio Pathogenicity Island-2 (VPI-2).* VPI-2 is a large PAI that encodes the
321 genes for sialic acid utilization[33]. The capacity to utilize sialic acid as a carbon and energy
322 source confers *V. cholerae* a competitive advantage in the mucus-rich environment of
323 the gut, where sialic acid availability is extensive[65]. Sialic acid catabolism also mediates
324 a chemotactic response towards mucin and several environmental reservoirs of *V.*
325 *cholerae*[66]. VPI-2 is excisable and form a circular intermediate[61,67]. Interestingly, during
326 the excision event, there is cross-talk between VPI-2 and VPI-1[68]. Unlike other MGEs,
327 the differential abundance of the gene clusters within VPI-2 suggests that the island is
328 made up of different modules. Also, all but one of the IS elements of the island are
329 widespread among several clades in diverse locations, which favours the transfer of the
330 different modules. The canonical VPI-2 from O1 strains encodes three major modules,
331 from 5' to 3': a) a restriction modification system (RM), b) the Nan-Nag cluster, and c) a
332 Mu-phage like region (**Fig. 4F**). At the 3' end of VPI-2 there is also a small group of genes
333 encoding hypothetical proteins. The Nan-Nag cluster is present in 88 clades, whereas
334 the distribution of the RM system is more limited, being encoded by only 7 clades (**Fig.**
335 **4F**). The Mu-phage like module is present in nine clades, however, there is no
336 relationship between specific modules and species phylogeny. Like VPI-1, no clade other
337 than PCG encode a PCG-like version of VPI-2. Furthermore, strains that contain the
338 three individual modules sometimes encoded them in different locations within the
339 genome. For instance, strain PivertUAT4Aug (C24), the only one with the three major
340 modules, encode the first two modules (Nan-Nag and RM) within a version of VPI-2,
341 whereas the Mu-phage region is encoded in a different location in the genome (**Fig. 4G**).
342 Besides several genes missing, PivertUAT4Aug encodes a different allele of *hsdS*

343 (VC_1768) (**Fig. 4G**). Another instance of this modular distribution is exemplified by
344 strain N2824 (C16), which possesses the first two modules but not the Mu-phage.
345 Furthermore, the strain has lost the genes encoding the HsdS and HsdM proteins from
346 the RM cluster involved in DNA sequence recognition and methylation, respectively (**Fig.**
347 **4G**). Strain 3541-04 (C17) encodes a different cluster of genes at the insertion site of the
348 Nan-Nag cluster and RM system, with the Nan-Nag cluster being in another
349 chromosomal region (**Fig. 4G**). Although this strain is the only one that contains a small
350 cluster of four genes just like PCG and in the same position (from VC_1801 to VC_1804).
351 Besides the gene mentioned above that codes for HsdS (VC_1768), there are three
352 others that code for hypothetical proteins (VC_1787, VC_1788 and VC_1791) that are
353 only present in the PCG (highlighted in **Fig. 4F**). Genomic comparisons with
354 PivertUAT4Aug (C24) show that, similar to *rfb*T in the LPS gene cluster, VC_1788 and
355 VC_1791 are the two parts of a gene coding for the tape measure protein of phage Mu
356 that have been truncated by the insertion of two transposases in the PCG, annotated as
357 VC_1789 and VC_1790 (**Fig. 4G**). Overall, VPI-2 in PCG must be understood based on
358 its unique modular arrangement not by the individual units that constitute it. Its four
359 modules are associated with IS elements that are widely distributed among *V. cholerae*
360 environmental strains and scattered throughout the genome. This leads to the
361 convergence of all modules in the same genome being extremely low which contributes
362 to the exceptionality of PCG.
363
364 *4. Vibrio Seventh Pandemic Island I (VSP-I).* The role of VSP-I in *V. cholerae*
365 pathogenesis is not as clearly defined as other MGEs. VSP-I has been suggested to play
366 an environmental role related to chemotaxis[69] and encodes a regulator involved in
367 intestinal colonization (VC_0177)[34]. We found homologs of the eleven genes that
368 comprise this island in ten *V. cholerae* clades (**Fig. 5A**). Although four of them are found
369 in strains isolated from this study, IRLE0062, IRLE0049, IRLE0042 and IRLE0077, that
370 belonged to the same clade. The greatest sequence divergence corresponded to an

371 integrase (VC_0183) and an XRE family transcriptional regulator (VC_0176) (**Fig. 5A**).
372 Nonetheless, only two strains encode a PCG-like VSP-I, V060002, which is
373 phylogenetically closer to PCG (C24), and strain VcCHNf4 from C5. V060002 has two
374 identical copies of the island in the genome that are in close proximity of each other in
375 the genome with the LPS cluster located between them (**Fig. 5B**). Another strain
376 belonging to C24 (2000), has the first gene at the 5' end replaced with a cluster of four
377 genes (**Fig. 5B**). The landscape of evolutionary possibilities for VSP-I is much more
378 limited than for the other MGEs due to its small size, nonetheless, and despite of this, it
379 is the rarest of all the islands. Suggesting this is a critical bottleneck in the emergence of
380 the current El Tor strains.
381
382 *5. Vibrio Seventh Pandemic Island II (VSP-II).* The presence of VSP-II in clinical
383 and environmental strains might be associated with environmental survival and fitness
384 of the bacterium[35–37]. A group of *V. cholerae* strains that caused an outbreak in Florida
385 associated with oyster consumption, encoded a novel bacteriocin and a pyocin protein
386 in the VSP-II element, which are typically associated with bacterial competition[37]. Our
387 analyses indicate that, similar to VPI-2, VSP-II is highly modular with several clusters of
388 genes that have a tRNA-Met insertion point. The direct repetition of the end of the tRNA
389 (highlighted as purple triangles) marks the insertion points of each module, which in most
390 cases match the presence of IS elements (**Fig. 5C and 5D**). Overall, this island serves
391 as an example of an "additive" island with three small modules. Module 1 comprises
392 genes VC_0490 to VC_0498, five of which are of viral origin but lack essential proteins
393 for packaging, indicating it is a defective prophage. Module 2 encompasses from
394 VC_0502 to VC_0510, however, the lack of possible annotation of its genes precludes
395 from assigning a specific function to this module. Nonetheless, VC_0502 and VC_0503
396 from this module are only present in strains from C24 (**Fig. 5C)**. Module 3 is comprised
397 of four genes (VC_0512 to VC_0515): VC_0512 and VC_0514 encode methyl-accepting
398 chemotaxis proteins, VC_0513 a transcriptional regulator and VC_0515 a predicted

399 signal transduction protein. These four genes are only present in C24 strains and
400 according to their annotation could be associated with environmental adaptations (**Fig.**
401 **5D**). Genomic comparison with strain MZO-3 (C24) reveal the presence of a novel CDS
402 in the PCG between VC_0512 and VC_513 that codes for a TnsD family Tn7-like
403 transposition protein (**Fig. 5D**). Overall, there are no identical versions of VSP-II outside
404 of PCG. The closest versions are encoded by strains N2723, MZO-3, and V060002
405 (C24), which are the only ones that possess a similar PCG-like modularity and synteny
406 of VSP-II (**Fig. 5D**). Relative to PCG, the changes in VSP-II from strain MZO-3 involve
407 only IS elements, the loss of a transposase (VC_0501) in Module 1 and the gain of
408 another in Module 3 (**Fig. 5D**). Whereas strain N2723 has lost genes coding for two
409 hypothetical proteins (VC_0496 and VC_0509) and a transcriptional regulator
410 (VC_0497) (**Fig. 5B**). In the rest of *V. cholerae* strains, VSP-II only maintains some of
411 these modules or clusters of genes, with the addition of other types of clusters different
412 from that of the PCG being common.

413 Overall, our analyses of the virulence-associated MGEs indicate that **a)** strains
414 from *V. cholerae* PCG have acquired the gene clusters encoding the major virulence
415 genes in a modular fashion, **b)** those MGEs are more widespread than previously
416 thought within the *V. cholerae* species, and **c)** a select set of genes dispersed in the
417 phylogeny of the species have a very limited distribution and **d)** unique allelic variations
418 in PCG act as stringent bottlenecks for emergence.

419

420 **Pangenome analysis.** Besides the modular acquisition of MGEs, are there other genes
421 associated with the emergence of PCG? In order to start addressing this question, we
422 investigated the unique genes encoded by PCG and those that the group inherited from
423 its closer ancestors. We considered the three major hierarchical levels in which the
424 pandemic strains of *V. cholerae* are included: cluster (C24), pandemic cholera lineage
425 (PCL), and the pandemic cholera group (PCG) (**Fig. 1C**). We analyzed the sequences

426 defining each of these groups at the pangenome scale, using the remaining set of clades
427 as outgroups to remove the background of other environmental *V. cholerae* sequences.
428 The Venn diagram in **Figure 6A** shows the total unique and shared gene family
429 clusters among the pangenomes of the different groups analyzed. The total number of
430 shared clusters of the *V. cholerae* species is 3,278 (core genome). The PCG encodes
431 29 unique gene families, 62% of which (18/29) are located in the integron of chromosome
432 II (**Fig. 6A**). Most of the genes could not be assigned a function or annotation,
433 nonetheless, among these unique genes families we found a transposase, a pullulanase,
434 a lysozyme inhibitor LprI family protein and several acetyltransferases (**Table S4**).
435 Among the remaining specific clusters that are encoded within the integron, most of them
436 are localized in MGEs. PCG inherited ten genes from the PCL. Half of those are also
437 concentrated in the integron and three are present in PAIs: the two homologs of the
438 transcriptional repressor RstR of CTXΦ (VC_1455 and VC_1464) and the ToxR-
439 activated gene A from VPI-1 (VC_0820) (**Table S4**). Finally, PCG inherited 17 unique
440 gene families from C24 with more than 80% of them being encoded within the integron
441 and the MGEs (**Fig. 6A**). Among these 17 genes we found two small clusters located in
442 VSP-II (VC_0502-VC_0503 and VC_0512-VC_0515) mentioned above in the island
443 description, as well as a cluster of genes inserted downstream of the 5' of CTXΦ (**Table**
444 **S4**). This cluster encodes a helicase, a transcriptional regulator, and a phage replication
445 protein, along with a Xer recombination activation factor involved in fragment integration
446 into the chromosome. Overall, our pangenome analyses reveal that, besides the
447 acquisition and modularity of MGEs, the integron plays a most critical role in the
448 emergence and evolution of the PCG significantly greater than the absence or presence
449 of other genes in the core genome.
450
451 **Evolution of the Pandemic Cholera Lineage.** Next, in order to understand at
452 fine resolution the evolution of PCG from the root of the PCL, we paired a) the history of
453 MGE acquisition, b) module exchange and c) gene gain and loss of the entire lineage.

454 Our analyses indicate that the capsule is the only element that is not modular in the PCL,
455 which contains 10 clades including the PCG, and appears to have evolved early in the
456 branching of PCL. It is present in 6 of the 9 non-PCG strains and is directly associated
457 with branching event 3 (**Fig. 6B**). VPI-1 and CTXΦ also appear to have been acquired
458 during branching event 3 and is present in the 2 clades that branch with PCG. On the
459 other hand, VSP-I and VSP-II are only found in 1 strain each, none of them with a direct
460 branching pattern related to PCG, which suggests independent acquisition events of the
461 islands by PCG (**Fig. 6B**). Only the closest relative to PCG encodes VPI-2 suggesting a
462 late acquisition of the island, likely during branching event 5. Overall, our results
463 demonstrate the mosaic-like relationship between mobile element acquisition and the
464 emergence of pandemic cholera strains with some (e.g. capsule) appearing early in the
465 phylogeny and others (e.g. VPI-2) being late acquisitions.
466
467 Given the importance of branching event 3 in the evolution of PCG, we analysed the
468 gene gain and loss of the different branches during the evolutionary process that led to
469 the PCG group excluding those genes related to PAIs. Our analyses highlight the
470 acquisition of small gene clusters by PCG during the evolution of its lineage after the split
471 in node 3 (**Table S5**). Several of these genes colocalize consecutively in the integron
472 suggesting a common evolutionary history. On chromosome I, we found a cluster of
473 genes related to sugar transfer, such as a glycosyltransferase, possibly involved in cell
474 wall decoration (**Table S5**). Two clusters were associated with transporters such as ion
475 channels and type IV toxin/antitoxin systems. On chromosome II, apart from the integron,
476 we found gene clusters associated with a transposase related to regulatory proteins and
477 one linked to a peptidase. On the other hand, the genes lost in relation to the other PCL
478 branches by PCG are related to membrane proteins such as one of the OmpA family, as
479 well as several secretion systems and transcriptional regulators (**Table S5**). Interestingly,
480 we found that PCG had lost a chitinase, which is a critical adaptation to the environment,
481 the reason behind this remains to be elucidated.

482

483 **Genetic variations associated with pathogen emergence (GVPEs)**. From our
484 analyses it appears that differences in overall gene content minimally account for the
485 confined distribution of pandemic cholera clones other than genes within the well-known
486 MGEs and virulence-associated clusters. In light of this, we investigated whether allelic
487 variations of core genes could help explain this confined distribution. We examined the
488 core genome microdiversity of PCG by estimating the ratio of nonsynonymous (d*N*) to
489 synonymous (d*S*) substitutions of each gene of the cholera group against the genomes
490 of every clade in Fig. 1C. We found a total of 248 genes on chromosome I and 153 on
491 chromosome II that showed signs of positive selection (d*N*/d*S*>1) (**Fig. 6C**). We
492 reasoned that those genes could possibly serve as genetic variations associated with
493 pathogen emergence (GVPEs)[4]. This recently coined term encompasses environmental
494 preadaptations such as virulence adaptive polymorphisms and host-associated
495 pathoadaptive mutations[4,22,43]. Interestingly, our analyses did not retrieve a single gene
496 that had a d*N*/d*S*>1 when compared to every other clade. Nonetheless, we identified a
497 total of eight genes in PCG (four from each chromosome) that exhibit positive selection
498 against more than 60 clades (**Fig. 6C**). Six of these genes (VC_0714, VC_1587,
499 VC_A0224, VC_A0233, VC_A0609 and VC_A0973) had no match in the databases and
500 therefore no function could be inferred, however, five of them encode transmembrane
501 domains or signal peptides indicating a possible functional relationship with
502 environmental sensing or interaction, either as secreted molecules, as receptors,
503 channels, or membrane regulators. The remaining two genes encode a 2-succinyl-6-
504 hydroxy-2,4-cyclohexadiene-1- carboxylate synthase (VC_1974), which is a hydrolytic
505 enzyme involved in vitamin K synthesis, as well as, a NifB/NifX family molybdenum-iron
506 cluster-binding protein (VC_1323), a cofactor of the enzyme dinitrogenase involved in
507 the reduction of nitrogen to ammonium (**Fig. 6C**). There is no clear pattern regarding the
508 functions associated with these genes, however, it is remarkable that, out of the over
509 200 clades analyzed, the PCG strongly exhibit signatures of selection for these 8. We

510 hypothesized that many of them could have arisen as GVPEs such as pathoadaptive
511 mutations in the human host after the emergence of toxigenic strains from PCG.
512
513 **CONCLUSIONS**
514 In this study, we took a systematic approach towards unraveling the evolutionary history
515 of the enigmatic Pandemic cholera group (PCG), which, to date, remains the only clade
516 within the *V. cholerae* species capable of causing the severe diarrheal disease cholera.
517 Phylogenomic analyses of the species reveal a cluster-based evolution of *V. cholerae*
518 with several hierarchical levels leading to the emergence of the PCG, which is contained
519 within a separate lineage (PCL) embedded within cluster 24. Our investigations on the
520 abundance and distribution of the virulence-associated genes and MGEs suggest the
521 existence of several layers in the evolutionary dynamics and emergence of *V. cholerae*
522 (**Fig. 7**).
523 We determined that even the hypervariable region of the LPS is found in a large
524 number of clusters. The high prevalence of the LPS cluster may be due to the different
525 exchange mechanism of this island compared to the acquired MGEs. The LPS locus
526 exhibit characteristics of a "replacement island", as it is transferred as a unit by
527 homologous recombination and not in small units like the other MGEs (**Fig. 7A**). On the
528 other hand, complete islands are composed of several modules generally bordered by
529 IS elements that contain small groups of genes resembling "additive islands" (**Fig. 7B**).
530 Furthermore, the small gene groups within these additive islands also demonstrate allelic
531 exchange within them and some can be fully conserved in synteny and similarity,
532 whereas a very limited number of them are exclusive to PCG (**Fig. 7C**). Our results
533 suggest that in order to analyze the evolutionary history of these elements, rather than
534 investigating presence or absence of the island, it is necessary to examine the specific
535 modules and individual alleles of some genes within them (**Fig. 7**). The evolution of the
536 PCL shows that we can only trace the phylogeny of the LPS cluster, the only island that
537 is exchanged by homologous recombination and which appears to have slower

538 replacement dynamics than the MGEs. However, both the example of the PCL and the
539 phylogeny of the different modules in the whole species suggest that the rest of the
540 islands have a rapid turnover and remain associated with each clade for a short time.
541 Overall, this fractal-like degree of molecular sophistication explains the absence of a
542 complete set of PCG-like elements in any other *V. cholerae* clade.

543 Genetic variations associated with pathogen emergence (GVPEs) in the form of
544 environmental preadaptations or intra-host pathoadaptive mutations are becoming
545 increasingly appreciated as major agents in the transition that allow bacteria to emerge
546 as human pathogens. We examined the core genome microdiversity of *V. cholerae* and
547 determined the large extent of allelic variations of genes that uniquely exhibit strong
548 signatures of positive selection in PCG. This finding strongly suggests that, besides the
549 intricate mosaic associated with MGEs modularity, unique allelic variations in PCG could
550 help explain this confined distribution (**Fig. 7D**). Future work will determine whether these
551 GVPEs stem from host adaptations or were present during the early evolution of the
552 PCG. Overall, our investigations provide unique insights on the paths that led to the
553 emergence of a major human pathogen and can extrapolated to other bacterial threats.
554

555 **FIGURE LEGENDS**

556 **Fig. 1. Phylogenomic and population structure of *Vibrio cholerae*. A**. Map of Florida
557 indicating the locations where the environmental *V. cholerae* strains were isolated
558 (Shepard Park and Feller´s House). **B**. Whole-genome SNP-based phylogenetic tree
559 showing the relationships among the *V. cholerae* genomes obtained in this study and a
560 representative strain from the Pandemic Cholera Group (*V. cholerae* O1 biovar El Tor
561 N16961; red branch) **C**. Phylogenomic tree of dereplicated *V. cholerae* strains available
562 in public databases and strains from this study. The tree was reconstructed from single
563 nucleotide polymorphisms of the core genome. Branches of members belonging to the
564 four main clusters are colored in green (C5), violet (C11), blue (C16) and red (C24). The

565 colored circles represent the source of isolation of the environmental strains collected in
566 this study. Grey circles located at the branch nodes represent the bootstrap values.
567
568 **Fig. 2. Distribution and evolutionary dynamics of the O1 Lipopolysaccharide (LPS)**
569 **cluster in the *V. cholerae* species. A**. The area outlined in gray represents the number
570 of homologous proteins found in the dereplicated *V. cholerae* genomes. The white circles
571 with dotted lines indicate the median protein identity. **B**. Genomic comparison of PCG-
572 like LPS clusters identified in *V. cholerae* environmental genomes. **C**. Schematic
573 representation of the variable zone of the LPS containing the *rfb*T gene and the IS
574 elements inserted in the PCG allele. IS elements are highlighted in yellow.
575
576 **Fig. 3. Integron-like dynamics govern genetic variation of CTXΦ phage. A**.
577 Distribution and structure of the CTXΦ phage in the *V. cholerae* species. The area
578 outlined in gray represents the number of homologous proteins found in the dereplicated
579 *V. cholerae* genomes. The white circles with dotted lines indicate the median protein
580 identity. Red rectangles on the genes indicate their exclusive presence in the PCL. The
581 black arrows indicate the genomic position where the cassettes have been inserted. **B**.
582 Genomic comparison of similar CTXΦ phage found in dereplicated *V. cholerae* genomes.
583 **C**. Schematic representation CTXΦ phage found in *V. cholerae* 3528-08 both the
584 numbers and the rectangles mark the different clusters acquired. Clusters of genes with
585 a known function follow the same color pattern. IS elements are highlighted in yellow.
586
587 **Fig. 4. Modularity and allelic variations shape the evolutionary dynamics of Vibrio**
588 **pathogenicity islands 1 and 2.** Structures of **(A)** VPI-1 and **(F)** VPI-2. The area outlined
589 in gray represents the number of homologous proteins found in the dereplicated *V.*
590 *cholerae* genomes. The white circles with dotted lines indicate the median protein
591 identity. **B**. Genomic comparison of homologous VPI-1 in dereplicated *V. cholerae*

592 genomes denote a reduced number of allelic variations in some PCG genes. Sequence
593 alignment of the protein sequences of **(C)** ToxT **(D)** TcpA and **(E)** TcpF from PCG and
594 IRLE0081 (C17). **G**. Genomic comparison of VPI-2 variants encoded in dereplicated *V.*
595 *cholerae* genomes highlight pervasive modularity of the island. Clusters of genes with a
596 known function follow the same color pattern. IS elements are highlighted in yellow.
597
598 **Fig. 5. Distribution and evolutionary dynamics of the Vibrio seventh pandemic**
599 **islands I and II.** Structures of **(A)** VSP-I and **(C)** VSP-II. The gray area represents the
600 number of homologous proteins found in the dereplicated *V. cholerae* genomes. The
601 white circles with dotted lines indicate the median protein identity. Red rectangles on the
602 genes indicate their exclusive presence in the PCL. Pink triangles indicate the presence
603 of tRNA-direct repeat genomic islands. **B**. Genomic comparison of homologous VSP-I
604 across *V. cholerae* genomes. **D**. Genomic comparison of VSP-II variants encoded in
605 dereplicated *V. cholerae* genomes highlight pervasive modularity of the island. IS
606 elements are highlighted in yellow.
607
608 **Fig. 6. Evolutionary trajectories and genome dynamics of *V. cholerae* PCG. A**.
609 *Pangenome analysis*. The Venn diagram shows the numbers of unique and shared gene
610 families between the pandemic cholera group (purple), pandemic cholera lineage (blue),
611 Cluster 24 (light blue) and the rest of clades (greyQ) **B**. *PCL evolution*. The tree
612 designates the evolutionary relationships of the different clades within the PCL
613 highlighting the loss and gain of genes through the lineage. Black numbers indicate an
614 overall gain and grey an overall loss. The colored circles represent the specific element
615 as well as the percentage of completeness in each of the clades in reference to PCG.
616 LPS, O1 lipopolysaccharide cluster; VPI-1, Vibrio pathogenicity island-1; VPI-2, Vibrio
617 pathogenicity island-2; VSP-1, Vibrio seventh pandemic island-1; VSP-2, Vibrio seventh
618 pandemic island-2. **C.** *Microdiversity analysis*. PCG genes under positive selection

619 compared to the rest of the clades. X-axis indicates the number of clades in which
620 d*N*/d*S*>1 and the Y-axis shows the average of the d*N*/d*S* value
621
622 **Fig. 7. Hierarchical model of the emergence of the Pandemic cholera group in *V.***
623 ***cholerae*.** The evolutionary history of PCG suggests that the emergence of this group
624 has not been a linear process. Our analyses indicate that this process has unfolded at
625 four levels: A) The first one involving only the LPS where the island is replaced as a unit
626 by homologous recombination ("replacement island"). This replacement is slow and
627 allows tracing its phylogeny. B) The second level is related to the MGEs, which are
628 islands of the additive type composed of several modules bordered by IS elements that
629 have a very rapid exchange dynamics. In addition, these modules have the potential to
630 be inserted into other regions of the genome. This rapid exchange precludes from
631 associating these modules to specific clades. This is exemplified by the MGEs such as
632 VPI-2 and VSP-II. C) Furthermore, a third level in the process of PCG emergence is
633 associated with allelic variations within these modules, such as those we identified in the
634 genes encoding TcpF or ToxT in VPI-1. D) Finally, the fourth level is composed of genetic
635 variations associated with pathogen emergence (GVPEs), which are mutations in core
636 genes associated with the emergence of pathogenic strains (e.g. VC1323 or VC0714)
637 and exhibit strong positive selection in the PCG.
638
639 **SUPPLEMENTARY MATERIAL**
640 **Table S1.** List of *Vibrio cholerae* collected in this study, together with genomic features
641 and metadata
642 **Table S2.** List of dereplicated (ANI>99%) *Vibrio cholerae* reference strains from the
643 NCBI
644 **Table S3.** List of dereplicated (ANI>98%) *Vibrio cholerae* reference including IRL strains
645 **Table S4**. Shared family gene clusters in the pangenomes of the three hierarchical levels

646 **Table S5**. Summary of the gene gain/loss of the node 3 from the pandemic cholera
647 lineage.
648 **Figure S1**. **Dendrogram representing the relatedness of *V. cholerae* isolates**. The
649 dendrogram is based on PCR fingerprints generated by VCR-PCR. Groups of similarity
650 were established using the UPGMA method in BioNumerics v8.0. The red circles indicate
651 clonal isolates that were selected for whole genome sequencing.
652 **Figure S2. Pairwise comparison among the *V. cholerae* genomes.** The pairwise
653 comparison was obtained using the average nucleotide identity (ANI) from strains
654 isolated in this study and the Pandemic Cholera Group (*V. cholerae* O1 biovar El Tor).
655 **Figure S3**. **Genomic comparison of CTXΦ phage**. Comparison between the CTXΦ
656 phage genomic island encoded by Pandemic Cholera Group strains (*V. cholerae* O1
657 biovar El Tor N16961) and *Vibrio mimicus*.
658
659 **MATERIAL AND METHODS**
660 **Sampling sites**. We collected samples at two environmentally distinctive locations along
661 the IRL (Easter Florida, USA) in three sampling events. For each sampling campaign we
662 obtained samples from several fractions: three associated with the water bodies (Water
663 filtered through 20µm, 5µm, and 0.22µm) and one from the sediment. One of the
664 sampling locations, Fellers House Field Station (N28˚54'25.315"; W80˚49'15.017"), is
665 located within the federally-protected Canaveral National Seashore. The second
666 sampling site, Shepard Park, is located in the Port St. Lucie area (N27˚11'48.864";
667 W80˚15'33.172"), due to urbanization and agricultural expansion, experiences nutrient
668 over-enrichment leading to excessive macroalgal bloom (Fig. 1A)[43].
669
670 **Isolation of *V. cholerae.*** *Water.* Isolation of *V. cholerae* from water samples collected
671 at the aforementioned sampling locations was performed using a modified protocol from
672 Huq *et al* [71]. Briefly, water samples (~3L) were collected in triplicate from each site and
673 stored cold until arrival at the laboratory. To investigate the prevalence of *V. cholerae* in

674 various fractions of the aquatic column and to separate particle-associated and free-
675 living bacteria, 500 mL of collected water was filtered successively through 20µm, 5µm
676 and 0.2µm filters (Sterlitech) using a vacuum filtration system. The respective membrane
677 filters harboring potential *vibrio* spp. were then suspended in 25ml phosphate buffered
678 saline (PBS) and vortexed vigorously to release the contents. Each resuspended sample
679 was then inoculated (1:10 v/v) in alkaline peptone water (APW) and incubated at 37°C
680 for 12-14 hrs. This was repeated for each replicate. Enriched cultures were subjected to
681 growth on a series of selective agar specific for the isolation of *V. cholerae.* First,
682 enriched cultures were serially diluted 10-fold and plated on CHROMagar Vibrio (CaV;
683 CHROMagar), a chromagenic media on which *V.* cholerae forms turquoise blue colonies.
684 Secondly, turquoise blue colonies from CaV were transferred onto Thiosulfate Citrate
685 Bile Salts Sucrose (TCBS; Sigma) agar plates. Colonies that appeared turquoise blue
686 on CaV followed by yellow on TCBS were presumed to be *V. cholerae*. *Sediment.*
687 Isolation of *V. cholerae* was done using previous methods [43]. Briefly, sediment samples
688 were collected in triplicates from each sampling site using universal corer. 25g of the
689 sediment samples from each replicate was weighed and suspended in 25 ml of PBS
690 (Gibco). Samples were homogenized, enriched, and studied for the presence of *V.*
691 *cholerae* as described above. *Cyanobacteria.* Cyanobacteria floccules were pelleted,
692 supernatant removed and fresh APW added to homogenize the sample. Samples were
693 enriched and studied for the presence of *V. cholerae* as described above.
694
695 **Strains and growth conditions.** *V. cholerae* strains were routinely grown on Luria
696 Bertani (LB) agar at 37 ºC for approximately 16hrs, unless otherwise stated. For liquid
697 cultures, isolated colonies selected from agar plates were grown aerobically at 37ºC in
698 LB broth for ~16hrs.
699
700 **Clonal analysis of *V. cholerae* isolates.** To confirm isolates of *V. cholerae,* presence
701 of the species-specific gene *ompW* was determined by PCR[46]. Genomic DNA from the

702 isolates was extracted using the Gentra Puregene Yeast/Bact. Kit (Qiagen) and used as
703 template for PCRs. To identify the clonal populations of *V.* cholerae in the confirmed
704 dataset, DNA fingerprinting patterns were examined using a combination of *Vibrio*
705 *cholerae* Repeats-PCR (VCR-PCR) and Enterobacterial Repetitive Intergenic
706 Consensus Sequence (ERIC)-PCR as described by Teh *et al* [47]. Genomic DNA was used
707 to perform VCR- and ERIC-PCR and resulting PCR products were electrophoresed and
708 imaged using UVP ChemStudio (AnalytikJena) to observe fingerprinting patterns. Cluster
709 analysis of DNA patterns was performed using BioNumerics v8.0 (Applied Maths, Inc.)
710 and dendrogram generated based on the Unweighted Pair Group Method with the
711 Arithmetic Mean (UPGMA) method[47,72,73]. Based on resulting DNA fingerprinting
712 patterns, a total of 35 independent clones were identified and selected for whole genome
713 sequencing (WGS). Genomic DNA of each clone was submitted to Microbial Genome
714 Sequencing Center (MiGS) for WGS using the Illumina NextSeq 2000 platform.
715
716 **Genome assembly, gene prediction, and annotation.** Reads were trimmed using
717 Trimmomatic v0.36[74] and assembled de novo with SPAdes version 3.11.1[75]. Coding DNA
718 sequences (CDS) from the assembled contigs were predicted using Prodigal version
719 2.6.3 using *-a output.proteins -d output.genes -c -p meta* parameters[76]. Then, tRNAs
720 were obtained using tRNAscan-SE version 1.4[77] together with ssu-align version 0.1.1,
721 and rRNA genes using meta-rna[78]. Predicted proteins were compared against the
722 National Center for Biotechnology Information nonredundant database (NCBI nr) using
723 DIAMOND[79]. In addition, for functional annotation we used HMMscan version 3.1b2[80] for
724 the comparison against COG v2003 (update 2014)[81] and TIGFRAM v15.0 (September
725 2014)[82] databases.
726
727 **Recovery of *V. cholerae* genomes.** A total of 1,127 *V. cholerae* genomes were
728 downloaded from RefSeq (accessed July 2021) and subjected to a dereplication step
729 with drep[83] at 99 % nucleotide identity. Briefly, we force dRep to sort and select as seeds

730 for the clustering genomes with better score. For instance, chromosomes in a single
731 contig, followed by higher N50 and L50 values, and lastly by the completeness and
732 contamination values calculated with checkm[84]. This resulted in 227 dereplicated
733 reference *V. cholerae* genomes.
734
735 **Phylogeny of *V. cholerae* dereplicated genomes.** Single nucleotide polymorphisms
736 (SNPs) among *V. cholerae* dereplicated genomes and the PCG reference genome *Vibrio*
737 *cholerae* El Tor N16961 (RefSeq accession number GCF_900205735.1) were calculated
738 with PhaME[85] with default options, resulting in a core genome length of 2,468,623 bp
739 and 458,798 polymorphic positions considered. Then, a maximum likelihood
740 phylogenetic tree of SNPs was constructed using iqtree v 1.6.12[86] using the ultrafast
741 bootstrap approach (5000 replicates) and the best fitted model GTR+F+R4.
742
743 **Gene cluster search.** To determine the distribution and degree of prevalence of proteins
744 within the studied *V. cholerae* genomes, we extracted the LPS cluster, CTXΦ and
745 pathogenicity islands from the reference PCG strain based on the manual annotation
746 found in the NCBI Genbank. Proteins were isolated from the clusters and analyzed using
747 Prodigal[76], then used as a reference in a BLASTP search against the proteins from the
748 genomes[87,88]. We kept all proteins that matched against the pathogenic protein database
749 with at least 70% identity and the alignment covered at least 70% of the both proteins.
750
751 **Pangenome analysis.** To evaluate the pangenome of *V. cholerae*, in a first approach
752 dereplicated sequences were grouped in clusters at 98% identity using cd-hit[89], resulting
753 in 34 clusters, four of them (C5, C11, C16, and C24) comprising 188 out of 262 genomes.
754 Then, pangenomes were calculated for each cluster individually using PPanGGOLiN[90]
755 with default parameters. A second pangenome step with the protein families of clusters
756 C5, C11 y C16 was performed using PPanGGOLiN to simplify the comparison among
757 the pathogenic reference genome el Tor, C24 and the new superpangenome C5-C11-

758 C16. Ortholog proteins among the three groups (el Tor, C24 and C5-C11-C16 were
759 calculated using orthoMCL[91] (> 30 % amino acid identity, > 70 % alignment). A Venn
760 diagram showing proteins shared and not shared among groups was performed in R
761 using the ggVennDiagram package (https://github.com/gaospecial/ggVennDiagram).
762
763 **Determination of Gain and Loss of genes in the pandemic cholera lineage.** In order
764 to understand the flow of genes among pandemic *V. cholerae* and the closest genomes
765 from cluster C24, those belonging to the pandemic cholera lineage, gene presence and
766 absence was tested using the gain loss mapping engine GLOOME[92]. The expectations
767 and probabilities of both gain and loss events were estimated using stochastic mapping.
768 The phyletic pattern of presence/absence of orthologous genes was then classified using
769 hmmscan[80] against the Pfam database v35.0[93], using an evalue of < $1e^{-5}$ as a threshold
770 for protein annotation.
771
772 **Microdiversity analysis.** To identify which genes in each genome were under positive
773 selection, we calculated the nonsynonymous (d*N*) and synonymous (d*S*) substitutions
774 between ortholog proteins found in both the studied *V. cholerae* genomes and the
775 reference PCG strain, we used the orthologr package in R[42], a pipeline that identifies
776 orthologous protein pairs, performs codon alignment and finally computes the final dN/dS
777 values. We considered a gene under positive selection if the dN/dS ratio > 1.
778
779 **ACKNOWLEDGEMENTS**
780 This work was supported by an NSF CAREER award (#2045671) to SAM. This work was
781 also supported by grant "FLEX3GEN" PID2020-118052GB-I00 (cofounded with FEDER
782 funds) from the Spanish Ministerio de Economía, Industria y Competitividad to MLP.
783
784 **COMPETING INTERESTS**
785 The authors declare no competing interests.

786

787 **AUTHOR CONTRIBUTIONS**

788 SAM conceived the study. MLP, AZS and JHM performed bioinformatic analyses. TAG
789 collected samples, isolated *V. cholerae* and sequenced genomes. The manuscript was
790 written by MLP and SAM. All authors read and approved the final version.

791

792 **REFERENCES**

793 1. Alexander, K. A. *et al.* What factors might have led to the emergence of Ebola in
794 West Africa? *PLOS Negl Trop Dis* 9, e0003652 (2015).

795 2. Faye, O. *et al.* Molecular Evolution of Zika Virus during Its Emergence in the 20th
796 Century. *PLOS Negl Trop Dis* 8, e2636 (2014).

797 3. Wellington, E. M. *et al.* The role of the natural environment in the emergence of
798 antibiotic resistance in Gram-negative bacteria. *The Lancet Infectious Diseases* 13,
799 155–165 (2013).

800 4. Balasubramanian, D., López-Pérez, M., Grant, T.-A., Ogbunugafor, C. B. & Almagro-
801 Moreno, S. Molecular mechanisms and drivers of pathogen emergence. *Trends*
802 *Microbiol* (2022) doi:10.1016/j.tim.2022.02.003.

803 5. Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature* 451, 990–
804 993 (2008).

805 6. Thompson, F. L., Iida, T. & Swings, J. Biodiversity of vibrios. *Microbiology and*
806 *molecular biology reviews : MMBR* 68, 403–431 (2004).

807 7. Reen, F. J., Almagro-Moreno, S., Ussery, D. & Boyd, E. F. The genomic code:
808 inferring Vibrionaceae niche specialization. *Nature reviews. Microbiology* 4, 697–704
809 (2006).

810 8. Kanungo, S., Azman, A. S., Ramamurthy, T., Deen, J. & Dutta, S. Cholera. *Lancet*
811 399, 1429–1440 (2022).

812 9. Clemens, J. D., Nair, G. B., Ahmed, T., Qadri, F. & Holmgren, J. Cholera. *The*
813 *Lancet* 390, 1539–1549 (2017).

814 10. Ali, M., Nelson, A. R., Lopez, A. L. & Sack, D. A. Updated Global Burden of Cholera
815 in Endemic Countries. *PLOS Negl Trop Dis* 9, e0003832 (2015).

816 11. Balasubramanian, D., Murcia, S., Ogbunugafor, C. B., Gavilan, R. & Almagro-
817 Moreno, S. Cholera dynamics: lessons from an epidemic. *Journal of medical*
818 *microbiology* 20, 106 (2021).

819 12. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh
820 cholera pandemic. *Nature* 477, 462–465 (2011).

821 13. Islam, M. T., Alam, M. & Boucher, Y. Emergence, ecology and dispersal of the
822 pandemic generating Vibrio cholerae lineage. *International microbiology : the official*
823 *journal of the Spanish Society for Microbiology* 20, 106–115 (2017).

824 14. Baker-Austin, C. *et al.* Vibrio spp. infections. *Nature Reviews Disease Primers* 4, 8
825 (2018).

826 15. Austin, C. B., Pruzzo, C., Oliver, J. D. & Garzón, D. D. Vibrios – from genes to
827 ecosystems. *Environmental microbiology* 22, 4093–4095 (2020).

828 16. Vezzulli, L. *et al.* Climate influence on Vibrio and associated human diseases
829 during the past half-century in the coastal North Atlantic. *Proceedings of the National*
830 *Academy of Sciences* 113, E5062-71 (2016).

831 17. Trinanes, J. & Martinez-Urtaza, J. Future scenarios of risk of Vibrio infections in a
832 warming planet: a global mapping study. *Lancet Planet Heal* 5, e426–e435 (2021).

833 18. Chatterjee, S. *et al.* Incidence, virulence factors, and clonality among clinical strains
834 of non-O1, non-O139 Vibrio cholerae isolates from hospitalized diarrheal patients in
835 Kolkata, India. *Journal of clinical microbiology* 47, 1087–1095 (2009).

836 19. Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-
837 term clonal transitions in pandemic Vibrio cholerae. *Proceedings of the National*
838 *Academy of Sciences* 106, 15442–15447 (2009).

839 20. Dziejman, M. *et al.* Comparative genomic analysis of Vibrio cholerae: genes that
840 correlate with cholera endemic and pandemic disease. *Proceedings of the National*
841 *Academy of Sciences of the United States of America* 99, 1556–1561 (2002).

842 21. Boucher, Y. Sustained local diversity of Vibrio cholerae O1 biotypes in a previously
843 cholera-Free country. *mBio* 7, e00570-16 (2016).

844 22. Shapiro, B. J., Levade, I., Kovacikova, G., Taylor, R. K. & Almagro-Moreno, S.
845 Origins of pandemic Vibrio cholerae from environmental gene pools. *Nature*
846 *Microbiology* 2, 16240 (2016).

847 23. DiRita, V. J. & Mekalanos, J. J. Periplasmic interaction between two membrane
848 regulatory proteins, ToxR and ToxS, results in signal transduction and transcriptional
849 activation. *Cell* 64, 29–37 (1991).

850 24. DiRita, V. J., Parsot, C., Jander, G. & Mekalanos, J. J. Regulatory cascade controls
851 virulence in Vibrio cholerae. *Proceedings of the National Academy of Sciences of the*
852 *United States of America* 88, 5403–5407 (1991).

853 25. Finkelstein, R. A. & LoSpalluto, J. J. Pathogenesis of experimental cholera.
854 Preparation and isolation of choleragen and choleragenoid. *The Journal of*
855 *experimental medicine* 130, 185–202 (1969).

856 26. Miller, V. L., Taylor, R. K. & Mekalanos, J. J. Cholera toxin transcriptional activator
857 toxR is a transmembrane DNA binding protein. *Cell* 48, 271–279 (1987).

858 27. Skorupski, K. & Taylor, R. K. A new level in the Vibrio cholerae ToxR virulence
859 cascade: AphA is required for transcriptional activation of the tcpPH operon. *Molecular*
860 *microbiology* 31, 763–771 (1999).

861 28. Taylor, R. K., Manoil, C. & Mekalanos, J. J. Broad-host-range vectors for delivery of
862 TnphoA: use in genetic analysis of secreted virulence determinants of Vibrio cholerae.
863 *The Journal of Bacteriology* 171, 1870–1878 (1989).

864 29. Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage
865 encoding cholera toxin. *Science* 272, 1910–1914 (1996).

866 30. Taylor, R. K., Miller, V. L., Furlong, D. B. & Mekalanos, J. J. Use of phoA gene
867 fusions to identify a pilus colonization factor coordinately regulated with cholera toxin.
868 *Proc National Acad Sci* 84, 2833–2837 (1987).

869 31. Karaolis, D. K. *et al.* A Vibrio cholerae pathogenicity island associated with
870 epidemic and pandemic strains. *Proceedings of the National Academy of Sciences of*
871 *the United States of America* 95, 3134–3139 (1998).

872 32. Faruque, S. M. & Mekalanos, J. J. Pathogenicity islands and phages in Vibrio
873 cholerae evolution. *Trends in microbiology* 11, 505–510 (2003).

874 33. Jermyn, W. S. & Boyd, E. F. Characterization of a novel Vibrio pathogenicity island
875 (VPI-2) encoding neuraminidase (nanH) among toxigenic Vibrio cholerae isolates.
876 *Microbiology* 148, 3681–3693 (2002).

877 34. Davies, B. W., Bogard, R. W., Young, T. S. & Mekalanos, J. J. Coordinated
878 regulation of accessory genetic elements produces cyclic di-nucleotides for V. cholerae
879 virulence. *Cell* 149, 358–370 (2012).

880 35. O'Shea, Y. A. *et al.* The Vibrio seventh pandemic island-II is a 26.9 kb genomic
881 island present in Vibrio cholerae El Tor and O139 serogroup isolates that shows
882 homology to a 43.4 kb genomic island in V. vulnificus. *Microbiology* 150, 4053–4063
883 (2004).

884 36. Taviani, E. *et al.* Discovery of novel Vibrio cholerae VSP-II genomic islands using
885 comparative genomic analysis. *FEMS Microbiology Letters* 308, 130–137 (2010).

886 37. Haley, B. J. *et al.* Genomic and Phenotypic Characterization of Vibrio cholerae
887 Non-O1 Isolates from a US Gulf Coast Cholera Outbreak. *PLoS ONE* 9, e86264
888 (2014).

889 38. Jaskólska, M., Adams, D. W. & Blokesch, M. Two defence systems eliminate
890 plasmids from seventh pandemic Vibrio cholerae. *Nature* 604, 323–329 (2022).

891 39. Rivera, I. N., Chun, J., Huq, A., Sack, R. B. & Colwell, R. R. Genotypes associated
892 with virulence in environmental isolates of Vibrio cholerae. *Applied and environmental*
893 *microbiology* 67, 2421–2429 (2001).

894 40. Faruque, S. M. *et al.* Genetic diversity and virulence potential of environmental
895 Vibrio cholerae population in a cholera-endemic area. *Proceedings of the National*
896 *Academy of Sciences of the United States of America* 101, 2123–2128 (2004).

897 41. Gennari, M., Ghidini, V. & Lleo, M. M. Virulence genes and pathogenicity islands in
898 environmental Vibrio strains non-pathogenic to humans. *FEMS microbiology ecology*
899 82, 563–573 (2012).

900 42. López-Pérez, M. *et al.* Evolutionary model of cluster divergence of the emergent
901 marine pathogen Vibrio vulnificus: From genotype to ecotype. *mBio* 10, 697 (2019).

902 43. López-Pérez, M. *et al.* Ecological diversification reveals routes of pathogen
903 emergence in endemic Vibrio vulnificus populations. *Proc National Acad Sci* 118,
904 e2103470118 (2021).

905 44. Almagro-Moreno, S. & Taylor, R. K. Cholera: Environmental Reservoirs and Impact
906 on Disease Transmission. *Microbiology Spectrum* 1, 149–165 (2013).

907 45. Lutz, C., Erken, M., Noorian, P., Sun, S. & McDougald, D. Environmental reservoirs
908 and mechanisms of persistence of Vibrio cholerae. *Frontiers in Microbiology* 4, 375
909 (2013).

910 46. Nandi, B. *et al.* Rapid Method for Species-Specific Identification of Vibrio cholerae
911 Using Primers Targeted to the Gene of Outer Membrane Protein OmpW. *J Clin*
912 *Microbiol* 38, 4145–4151 (2000).

913 47. Teh, C. S. J., Thong, K. L., Osawa, R. & Chua, K. H. Comparative PCR-based
914 fingerprinting of Vibrio cholerae isolated in Malaysia. *J Gen Appl Microbiol* 57, 19–26
915 (2011).

916 48. Chatterjee, S. N. & Chaudhuri, K. Lipopolysaccharides of Vibrio cholerae I. Physical
917 and chemical characterization. *Biochimica Et Biophysica Acta Bba - Mol Basis Dis*
918 1639, 65–79 (2003).

919 49. Chatterjee, S. N. & Chaudhuri, K. Lipopolysaccharides of Vibrio cholerae II.
920 Genetics of biosynthesis. *Biochimica Et Biophysica Acta Bba - Mol Basis Dis* 1690, 93–
921 109 (2004).

922 50. Chatterjee, S. N. & Chaudhuri, K. Lipopolysaccharides of Vibrio cholerae: III.
923 Biological functions. *Biochimica Et Biophysica Acta Bba - Mol Basis Dis* 1762, 1–16
924 (2006).

925 51. Schnaitman, C. A. & Klena, J. D. Genetics of lipopolysaccharide biosynthesis in
926 enteric bacteria. *Microbiol Rev* 57, 655–682 (1993).

927 52. Liang, W. *et al.* Sequence polymorphisms of rfbT among the Vibrio cholerae O1
928 strains in the Ogawa and Inaba serotype shifts. *Bmc Microbiol* 13, 173–173 (2013).

929 53. Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T. & Calderwood, S. B. Cholera.
930 *Lancet* 379, 2466–2476 (2012).

931 54. Rivera-Chávez, F. & Mekalanos, J. J. Cholera toxin promotes pathogen acquisition
932 of host-derived nutrients. *Nature* 572, 244–248 (2019).

933 55. Faruque, S. M. & Mekalanos, J. J. Phage-bacterial interactions in the evolution of
934 toxigenic Vibrio cholerae. *Virulence* 3, 556–565 (2014).

935 56. Mukhopadhyay, A. K., Chakraborty, S., Takeda, Y., Nair, G. B. & Berg, D. E.
936 Characterization of VPI pathogenicity island and CTXphi prophage in environmental
937 strains of Vibrio cholerae. *The Journal of Bacteriology* 183, 4737–4746 (2001).

938 57. Houot, L. *et al.* Electrostatic interactions between the CTX phage minor coat protein
939 and the bacterial host receptor TolA drive the pathogenic conversion of Vibrio cholerae.
940 *J Biol Chem* 292, 13584–13598 (2017).

941 58. Rodriguez-Valera, F., Martin-Cuadrado, A.-B. & López-Pérez, M. Flexible genomic
942 islands as drivers of genome evolution. *Curr Opin Microbiol* 31, 154–160 (2016).

943 59. Dorman, M. J. *et al.* High quality reference genomes for toxigenic and non-toxigenic
944 Vibrio cholerae serogroup O139. *Sci Rep-uk* 9, 5865 (2019).

945 60. O'Shea, Y. A. & Boyd, E. F. Mobilization of the Vibrio pathogenicity island between
946 Vibrio cholerae isolates mediated by CP-T1 generalized transduction. *FEMS*
947 *Microbiology Letters* 214, 153–157 (2002).

948 61. Murphy, R. A. & Boyd, E. F. Three pathogenicity islands of Vibrio cholerae can
949 excise from the chromosome and form circular intermediates. *Journal of bacteriology*
950 190, 636–647 (2008).

951 62. Kirn, T. J., Bose, N. & Taylor, R. K. Secretion of a soluble colonization factor by the
952 TCP type 4 pilus biogenesis pathway in Vibrio cholerae. *Molecular microbiology* 49,
953 81–92 (2003).

954 63. Kirn, T. J. & Taylor, R. K. TcpF is a soluble colonization factor and protective
955 antigen secreted by El Tor and classical O1 and O139 Vibrio cholerae serogroups.
956 *Infection and immunity* 73, 4461–4470 (2005).

957 64. Megli, C. J. *et al.* Crystal structure of the Vibrio cholerae colonization factor TcpF
958 and identification of a functional immunogenic site. *Journal of molecular biology* 409,
959 146–158 (2011).

960 65. Almagro-Moreno, S. & Boyd, E. F. Sialic acid catabolism confers a competitive
961 advantage to pathogenic Vibrio cholerae in the mouse intestine. *Infection and immunity*
962 77, 3807–3816 (2009).

963 66. Reddi, G., Pruss, K., Taylor, R. K., Cottingham, K. L. & Almagro-Moreno, S.
964 Catabolism of mucus components influences motility of Vibrio cholerae in the presence
965 of environmental reservoirs. *PLoS ONE* 13, e0201383 (2018).

966 67. Almagro-Moreno, S., Napolitano, M. G. & Boyd, E. F. Excision dynamics of Vibrio
967 pathogenicity island-2 from Vibrio cholerae: role of a recombination directionality factor
968 VefA. *BMC microbiology* 10, 306 (2010).

969 68. Carpenter, M. R., Rozovsky, S., Boyd, E. F. & DiRita, V. J. Pathogenicity Island
970 Cross Talk Mediated by Recombination Directionality Factors Facilitates Excision from
971 the Chromosome. *The Journal of Bacteriology* 198, 766–776 (2016).

972 69. Grim, C. J. *et al.* Occurrence of the Vibrio cholerae Seventh Pandemic VSP-I Island
973 and a New Variant. *OMICS: A Journal of Integrative Biology* 14, 1–7 (2010).

974 70. Iqbal, N., Guérout, A.-M., Krin, E., Roux, F. L. & Mazel, D. Comprehensive
975 Functional Analysis of the 18 Vibrio cholerae N16961 Toxin-Antitoxin Systems
976 Substantiates Their Role in Stabilizing the Superintegron. *J Bacteriol* 197, 2150–2159
977 (2015).

978 71. Huq, A. *et al.* Detection, Isolation, and Identification of Vibrio cholerae from the
979 Environment. *Curr Protoc Microbiol* 26, 6A.5.1-6A.5.51 (2012).

980 72. Naha, A. *et al.* Development and Evaluation of a PCR Assay for Tracking the
981 Emergence and Dissemination of Haitian Variant ctxB in Vibrio cholerae O1 Strains
982 Isolated from Kolkata, India. *J Clin Microbiol* 50, 1733–1736 (2012).

983 73. Zou, W. *et al.* Evaluation of Pulsed-Field Gel Electrophoresis Profiles for
984 Identification of Salmonella Serotypes. *J Clin Microbiol* 48, 3122–3126 (2010).

985 74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
986 sequence data. *Bioinformatics* 30, 2114–2120 (2014).

987 75. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its
988 Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19, 455–477
989 (2012).

990 76. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
991 identification. *BMC Bioinformatics* 11, 1–11 (2010).

992 77. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of
993 transfer RNA genes in genomic sequence. *Nucleic acids research* 25, 955–964 (1997).

994 78. Huang, Y., Gilna, P. & Li, W. Identification of ribosomal RNA genes in metagenomic
995 fragments. *Bioinformatics* 25, 1338–1340 (2009).

996 79. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
997 DIAMOND. *Nature methods* 12, 59–60 (2015).

998 80. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* 7,
999 e1002195 (2011).

1000 81. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic
1001 classification of proteins from complete genomes. *Nucleic acids research* 29, 22–28
1002 (2001).

1003 82. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional
1004 identification of proteins. *Nucleic acids research* 29, 41–43 (2001).

1005 83. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and
1006 accurate genomic comparisons that enables improved genome recovery from
1007 metagenomes through de-replication. *Isme J* 11, 2864–2868 (2017).

1008 84. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.
1009 CheckM: assessing the quality of microbial genomes recovered from isolates, single
1010 cells, and metagenomes. *Genome Research* 25, 1043–1055 (2015).

1011 85. Shakya, M. *et al.* Standardized phylogenetic and molecular evolutionary analysis
1012 applied to species across the microbial tree of life. *Sci Rep-uk* 10, 1723 (2020).

1013 86. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and
1014 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol*
1015 *Biol Evol* 32, 268–274 (2015).

1016 87. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
1017 alignment search tool. *J Mol Biol* 215, 403–410 (1990).

1018 88. Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database
1019 search programs. *Nucleic acids research* 25, 3389–3402 (1997).

1020 89. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
1021 generation sequencing data. *Bioinformatics* 28, 3150–3152 (2012).

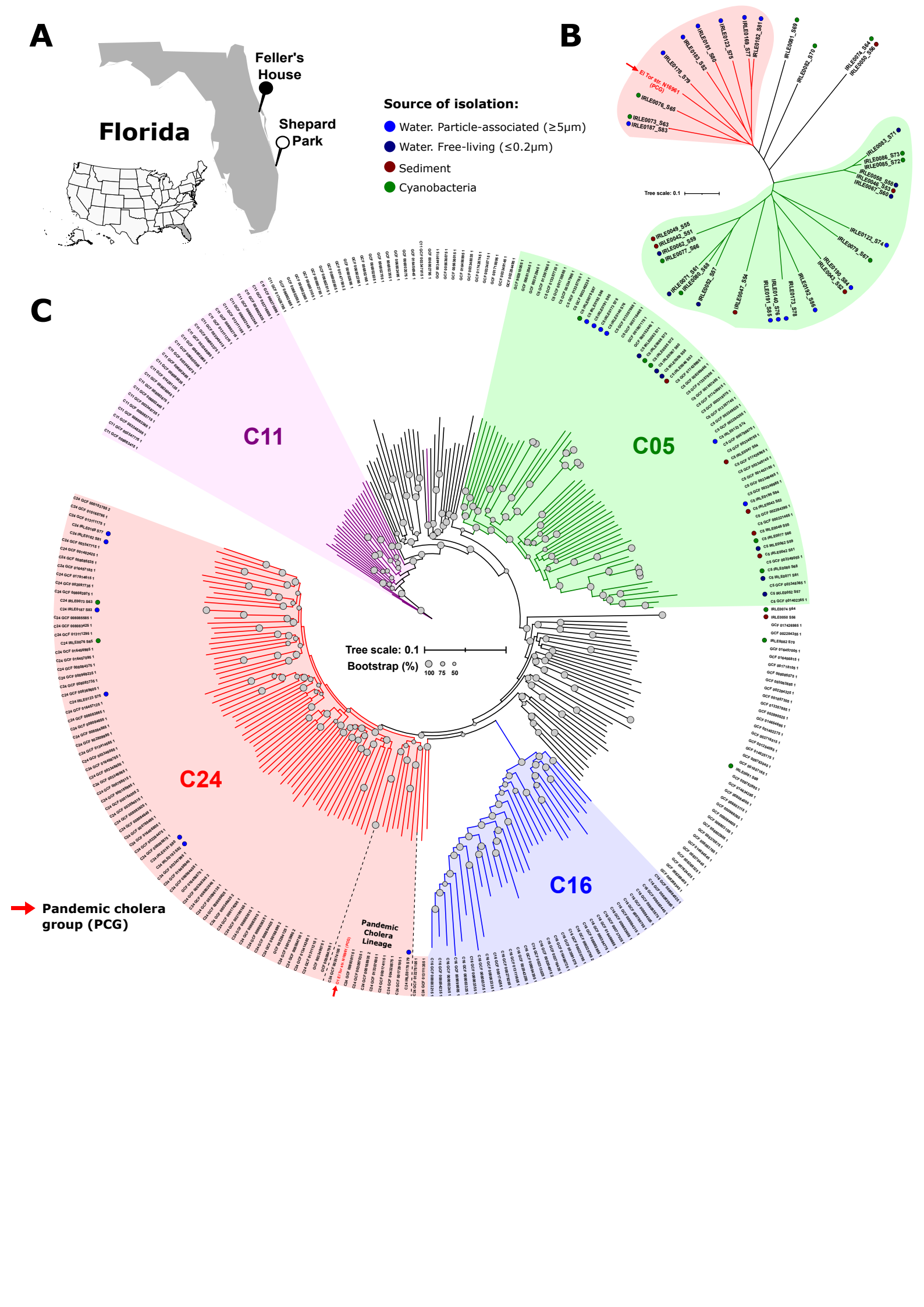
1022 90. Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned
1023 pangenome graph. *Plos Comput Biol* 16, e1007732 (2020).

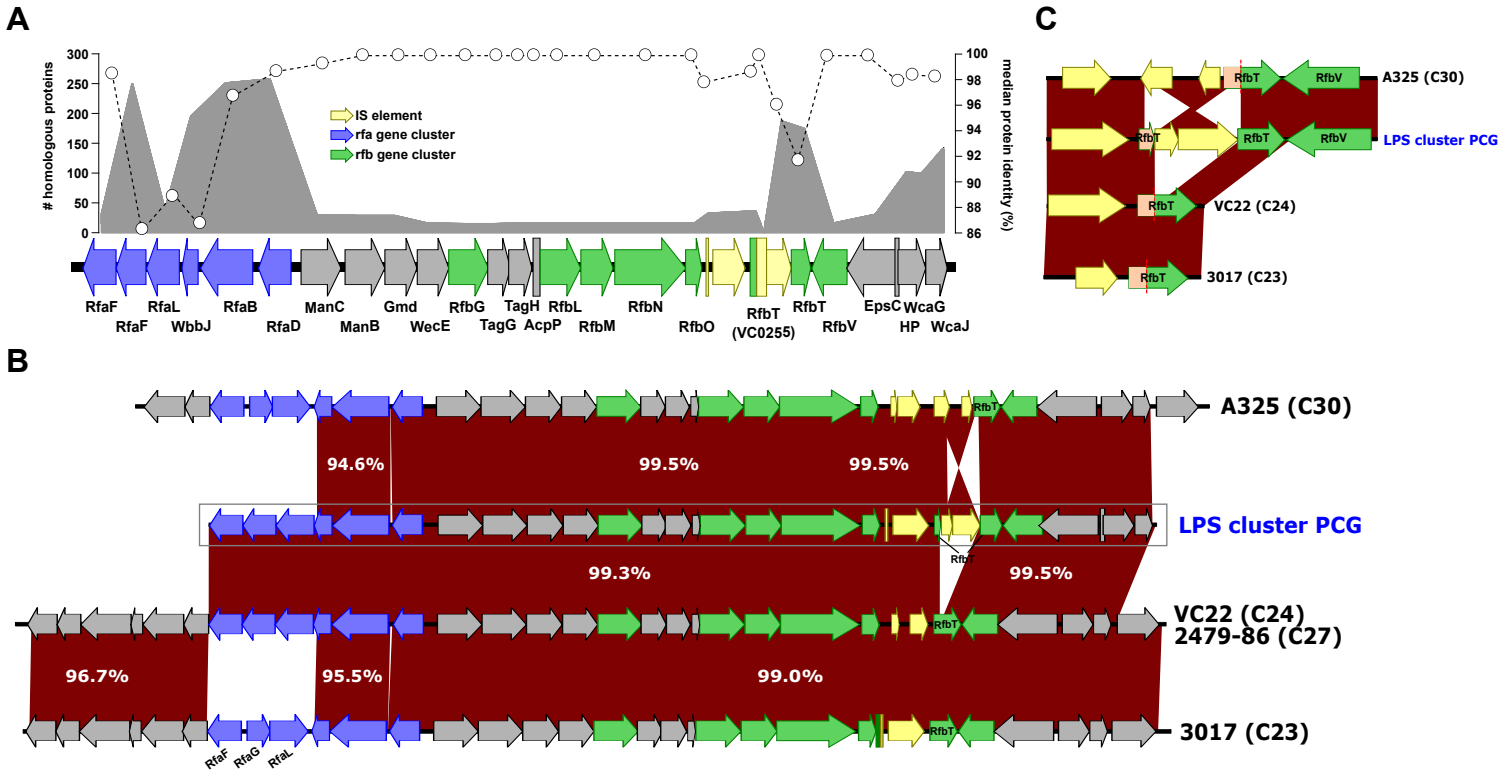
1024 91. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for
1025 eukaryotic genomes. *Genome Research* 13, 2178–2189 (2003).

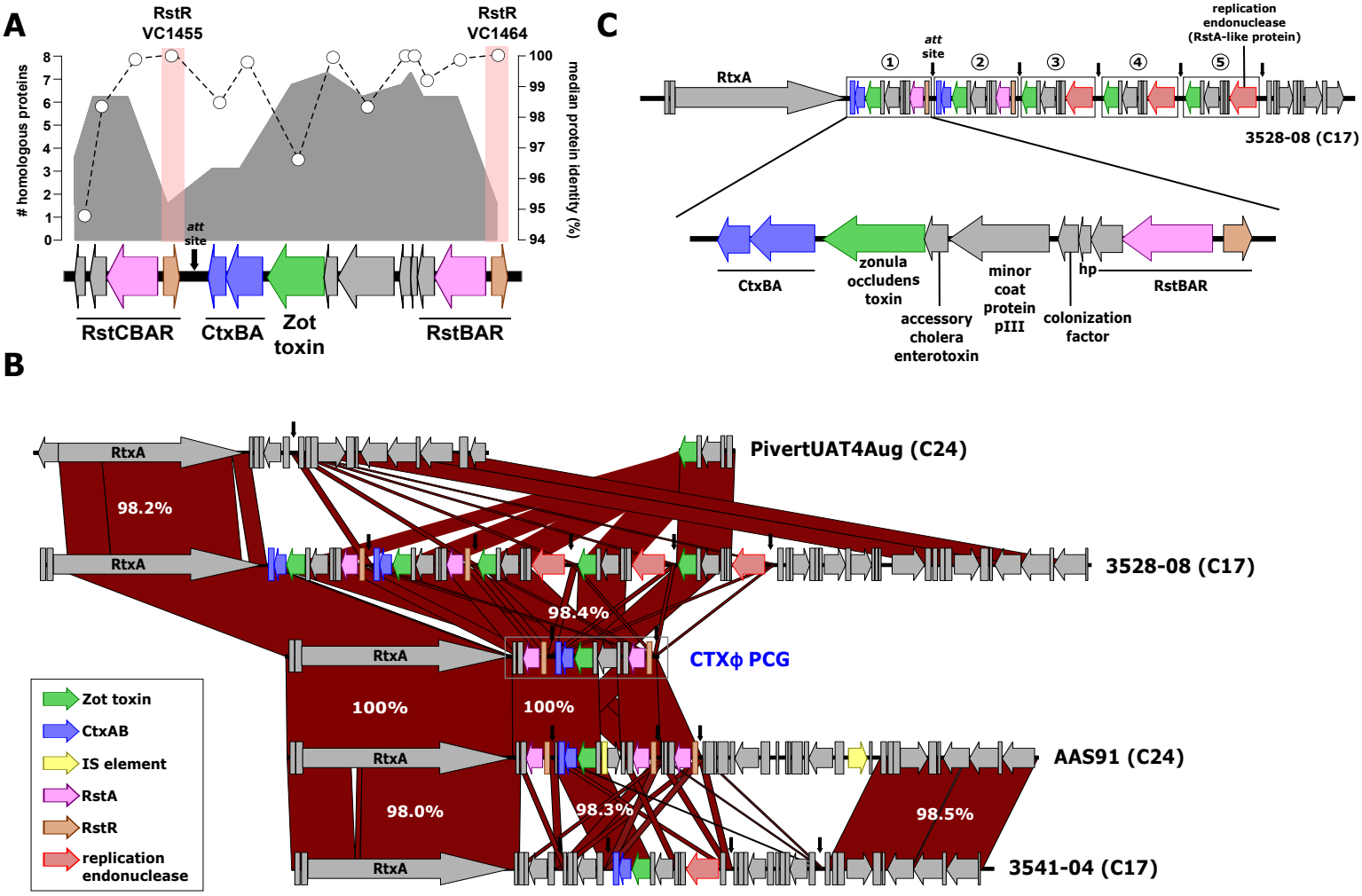
1026 92. Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D. & Pupko, T. GLOOME: gain
1027 loss mapping engine. *Bioinformatics* 26, 2914–2915 (2010).

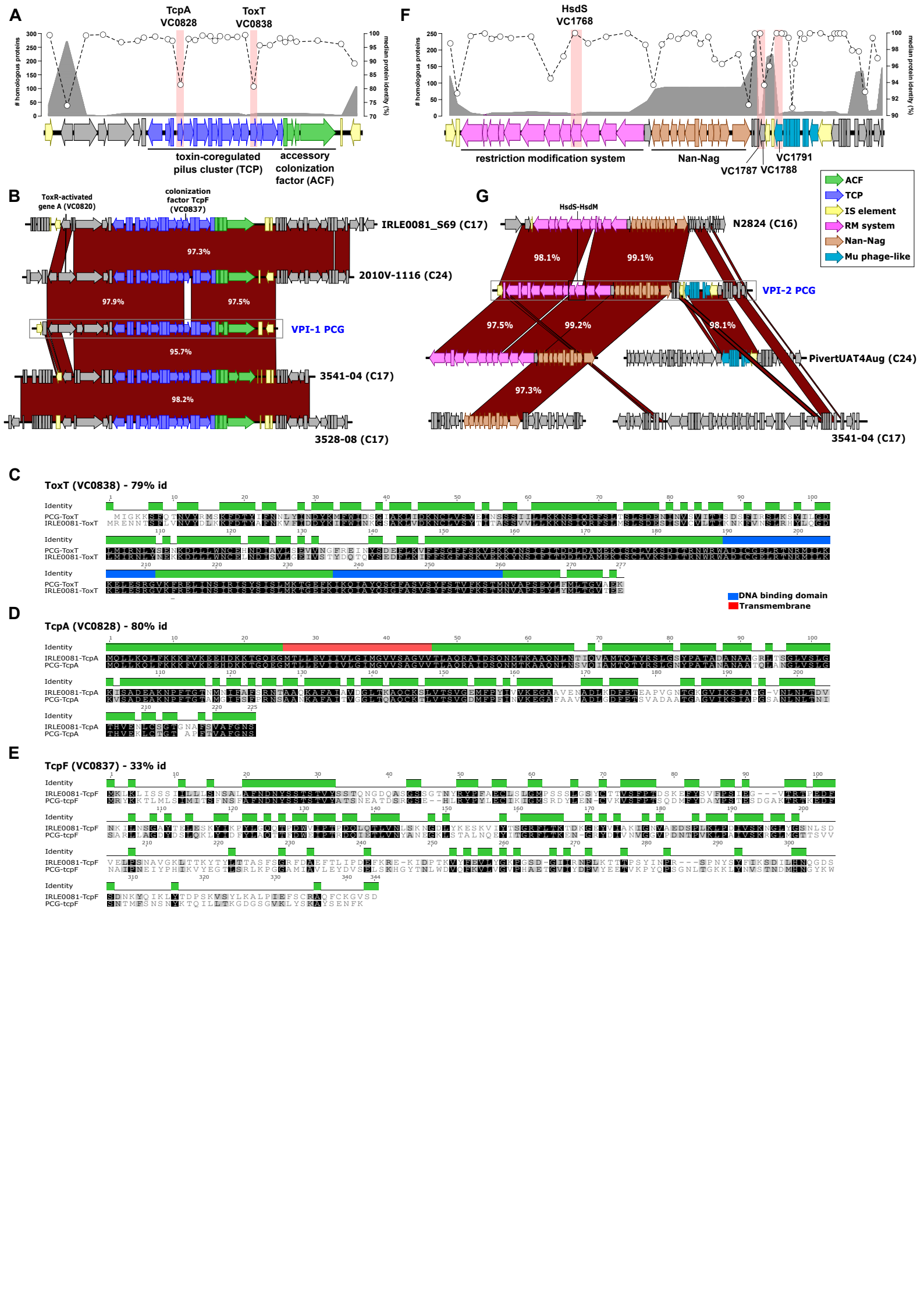
1028 93. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* 49,
1029 gkaa913- (2020).

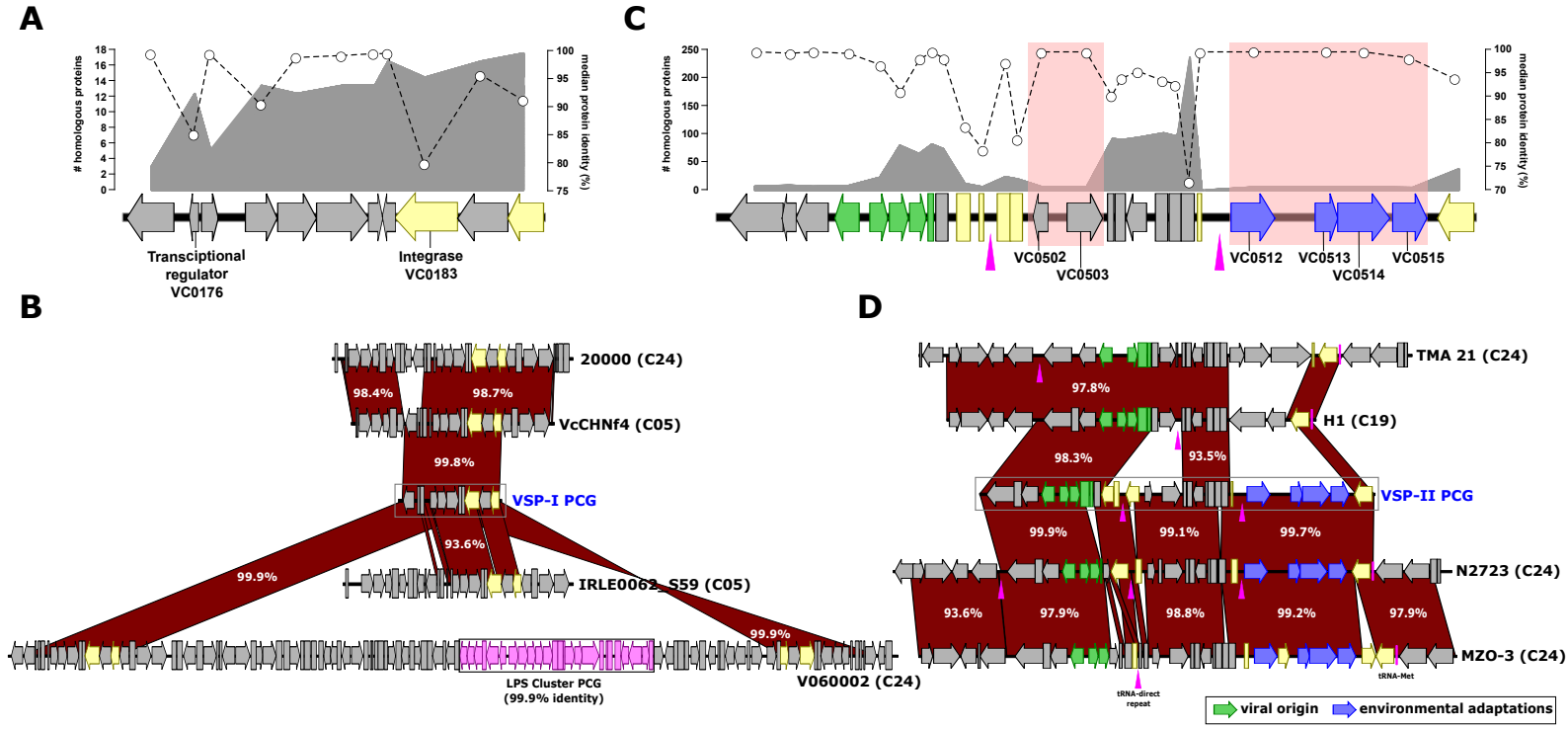
1030

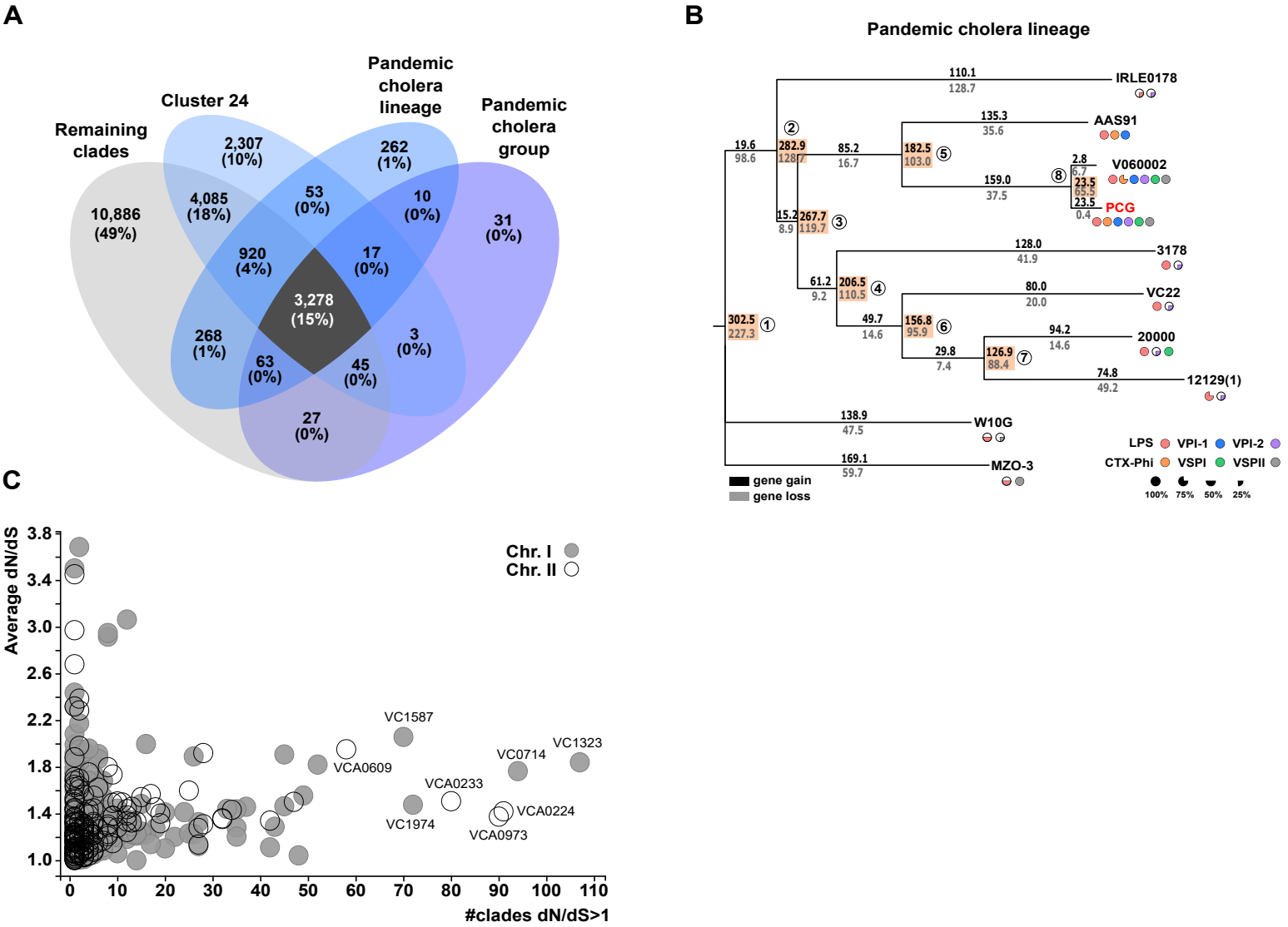

A
Florida
Feller's House
Shepard Park
Source of isolation:
Water. Particle-associated (≥5μm)
Water. Free-living (≤0.2μm)
Sediment
Cyanobacteria
B
El Tor str. N16961 (PCG)
Tree scale: 0.1
C
C11
C05
C24
C16
Tree scale: 0.1
Bootstrap (%)
100 75 50
Pandemic Cholera Lineage
Pandemic cholera group (PCG)

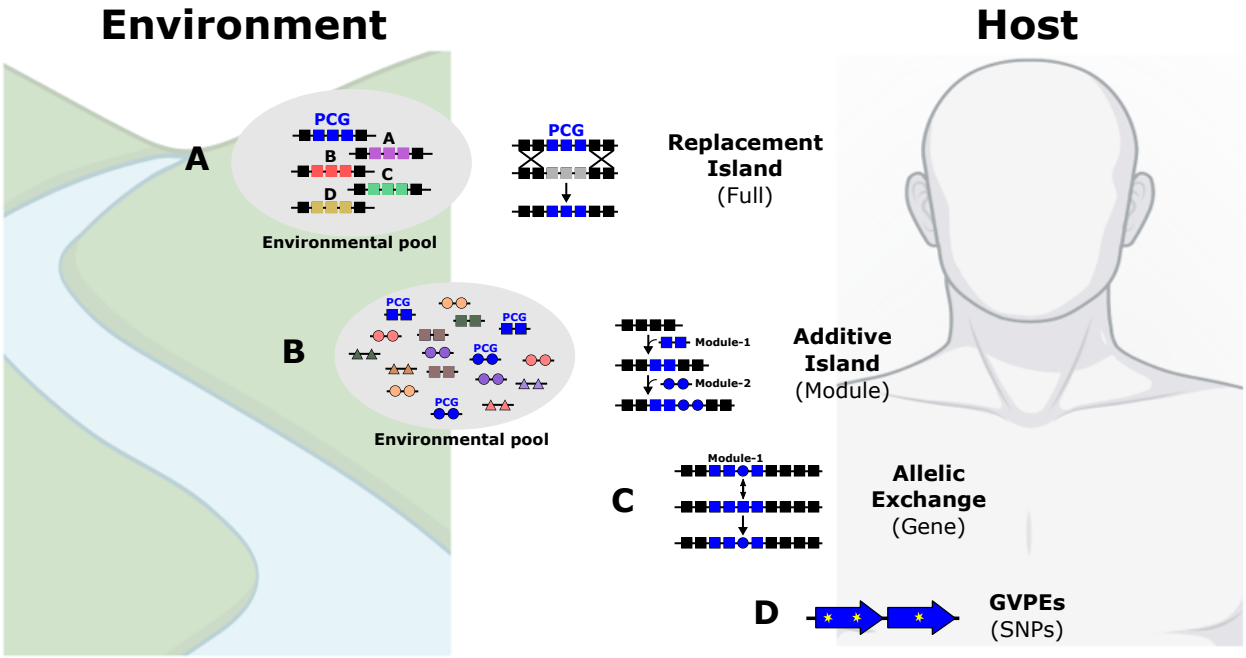

A
homologous proteins
median protein identity (%)
IS element
rfa gene cluster
rfb gene cluster
RfaF
RfaF
RfaL
WbbJ
RfaB
RfaD
ManC
ManB
Gmd
WecE
RfbG
TagG
TagH
AcpP
RfbL
RfbM
RfbN
RfbO
RfbT
(VC0255)
RfbT
RfbV
EpsC
HP
WcaG
WcaJ
C
A325 (C30)
LPS cluster PCG
VC22 (C24)
3017 (C23)
RfbT
RfbV
B
A325 (C30)
LPS cluster PCG
VC22 (C24)
2479-86 (C27)
3017 (C23)
94.6%
99.5%
99.5%
99.3%
99.5%
96.7%
95.5%
99.0%
RfaF
RfaG
RfaL


A
RstR
VC1455
RstR
VC1464
homologous proteins
median protein identity (%)
att
site
RstCBAR
CtxBA
Zot
toxin
RstBAR
C
RtxA
att
site
replication
endonuclease
(RstA-like protein)
3528-08 (C17)
CtxBA
zonula
occludens
toxin
accessory
cholera
enterotoxin
minor
coat
protein
pIII
hp
colonization
factor
RstBAR
B
RtxA
98.2%
PivertUAT4Aug (C24)
3528-08 (C17)
98.4%
CTXϕ PCG
100%
100%
AAS91 (C24)
98.0%
98.3%
98.5%
3541-04 (C17)
Zot toxin
CtxAB
IS element
RstA
RstR
replication
endonuclease


A
TcpA
VC0828
ToxT
VC0838
homologous proteins
median protein identity (%)
toxin-coregulated pilus cluster (TCP)
accessory colonization factor (ACF)
F
HsdS
VC1768
restriction modification system
Nan-Nag
VC1787
VC1788
VC1791
ACF
TCP
IS element
RM system
Nan-Nag
Mu phage-like
B
ToxR-activated gene A (VC0820)
colonization factor TcpF (VC0837)
IRLE0081_S69 (C17)
2010V-1116 (C24)
VPI-1 PCG
3541-04 (C17)
3528-08 (C17)
97.3%
97.9%
97.5%
95.7%
98.2%
G
HsdS-HsdM
N2824 (C16)
VPI-2 PCG
PivertUAT4Aug (C24)
3541-04 (C17)
98.1%
99.1%
97.5%
99.2%
98.1%
97.3%
C
ToxT (VC0838) - 79% id
Identity
PCG-ToxT
IRLE0081-ToxT
DNA binding domain
Transmembrane
D
TcpA (VC0828) - 80% id
IRLE0081-TcpA
PCG-TcpA
E
TcpF (VC0837) - 33% id
IRLE0081-TcpF
PCG-tcpF


A
homologous proteins
median protein identity (%)
Transcriptional regulator VC0176
Integrase VC0183
C
VC0502
VC0503
VC0512
VC0513
VC0514
VC0515
B
20000 (C24)
VcCHNf4 (C05)
VSP-I PCG
IRLE0062_S59 (C05)
V060002 (C24)
LPS Cluster PCG (99.9% identity)
98.4%
98.7%
99.8%
93.6%
99.9%
D
TMA 21 (C24)
H1 (C19)
VSP-II PCG
N2723 (C24)
MZO-3 (C24)
97.8%
98.3%
93.5%
99.9%
99.1%
99.7%
93.6%
97.9%
98.8%
99.2%
97.9%
tRNA-direct repeat
tRNA-Met
viral origin
environmental adaptations


A
Remaining clades
Cluster 24
Pandemic cholera lineage
Pandemic cholera group
10,886 (49%)
2,307 (10%)
262 (1%)
31 (0%)
4,085 (18%)
53 (0%)
10 (0%)
920 (4%)
17 (0%)
3,278 (15%)
268 (1%)
3 (0%)
63 (0%)
45 (0%)
27 (0%)
B
Pandemic cholera lineage
IRLE0178
AAS91
V060002
PCG
3178
VC22
20000
12129(1)
W10G
MZO-3
gene gain
gene loss
LPS
VPI-1
VPI-2
CTX-Phi
VSPI
VSPII
100%
75%
50%
25%
C
Average dN/dS
#clades dN/dS>1
Chr. I
Chr. II
VC1587
VC1323
VC0714
VCA0609
VCA0233
VCA0224
VC1974
VCA0973


Environment
Host
A
PCG
A
B
C
D
Environmental pool
PCG
Replacement Island
(Full)
B
PCG
PCG
PCG
PCG
Environmental pool
Module-1
Module-2
Additive Island
(Module)
C
Module-1
Allelic Exchange
(Gene)
D
GVPEs
(SNPs)