Origins of pandemic clones from environmental gene pools

B. Jesse Shapiro[1], Inès Levade[1#], Gabriela Kovacikova[2#], Ronald K. Taylor[2], Salvador Almagro-Moreno[2,3]*

[1]Department of Biological Sciences, University of Montreal, Montreal, Quebec, Canada.

[2]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA. [3]Burnett School of Biomedical Sciences, College of Medicine, University of Central Florida, Orlando, Florida, USA.

*For correspondence: samoreno@ucf.edu

[#]These authors contributed equally

Abbreviations: PG, Pandemic genome; EG, Environmental genome; SNP, Single nucleotide polymorphism; VAP, Virulence adaptive polymorphism; NS, Nonsynonymous; S, Synonymous.

Abstract

Some microbes can transition from an environmental lifestyle to a pathogenic one[1-4]. This ecological switch typically occurs through the acquisition of horizontally acquired virulence genes[5-8]. However, the genomic features that must be present in a population prior to the acquisition of virulence genes and emergence of pathogenic clones remain unknown. We hypothesized that virulence adaptive polymorphisms (VAPs) circulate in environmental populations and are required for this transition. We developed a comparative genomic framework for identifying VAPs, using *Vibrio cholerae* as a model. We then characterized several environmental VAP alleles to show that, while some of them reduced the ability of clinical strains to colonize a mammalian host, other alleles conferred efficient host colonization. These results show that VAPs are present in environmental bacterial populations prior to the emergence of virulent clones. We propose a scenario in which VAPs circulate in the environment, they become selected and enriched under certain ecological conditions, and finally a genomic background containing several VAPs acquires virulence factors that allows for its emergence as a pathogenic clone.

30 **Main text**

31 Numerous bacterial pathogens have emerged from environmental populations[1-4,9]. These
32 virulent clones evolve through the acquisition of toxins and host colonization factors[5-8]. Given
33 that the genes encoding these factors can often spread widely by horizontal gene transfer, it is
34 surprising that only a limited number of pathogenic clones have emerged from any particular
35 bacterial species. As a model of how environmental gene pools give rise to pandemic clones,
36 we used *Vibrio cholerae*, a genetically diverse group of aquatic bacteria that include a confined
37 phylogenetic group, the "pandemic genome" group (PG), that can cause the severe diarrheal
38 disease cholera in humans[10-14]. Seven pandemics of cholera have been recorded to date, all
39 caused by the PG group. The first six pandemics were caused by strains of the classical
40 biotype[12,13]; the current and seventh pandemic is caused by strains of the El Tor biotype, and
41 has spread across the globe in several waves of transmission[15]. Virulence in *V. cholerae* PG is
42 mainly determined by two virulence factors: the cholera toxin (CT) and the toxin-coregulated
43 pilus (TCP), which are encoded within horizontally acquired genetic elements, the CTXΦ phage
44 and the Vibrio Pathogenicity Island-1 (VPI-1) respectively[16-20]. Both CTXΦ and VPI-1 are always
45 found in the PG group, however, they are also encoded in some environmental populations of
46 *V. cholerae*[10,11,21-28]. Furthermore, even though some non-PG strains can cause gastrointestinal
47 infections, only strains from the PG clade have ever emerged as source of pandemic
48 cholera[12,13,29-31].

49 To investigate the evolutionary origins of pandemic clones of *V. cholerae* and the
50 potential for their reemergence, we analyzed 43 *V. cholerae* genomes sequenced from clinical
51 and environmental samples (Methods; Supplementary Table 1). These genomes span the
52 known genetic diversity of *V. cholerae* (Supplementary Text), and were divided into a primary
53 dataset of 22 genomes and a replication dataset of 22 genomes (Supplementary Table 2), with
54 one reference genome in common, MJ-1236. In the primary dataset, we chose 7 PGs, including
55 both classical and El Tor, to represent the genetic diversity of the pandemic group. We

56 compared these with 15 non-clinical environmental genomes (EGs): 10 EGs from worldwide
57 samples to include global diversity, and five sympatric isolates from the Great Bay Estuary
58 (GBE) in New Hampshire, USA, a region with no recent history of cholera outbreaks[32].
59 Consistent with the results of previous studies[10,15,32,33], PGs form a distinct monophyletic
60 group compared to EGs, based on the aligned core genome (Fig. 1a). Other than the PG group,
61 there is little phylogenetic structure and the tree is star-like, with a similar pattern being
62 observed in both the primary and replication datasets (Supplementary Fig. 1). Reticulations in
63 the phylogenetic network indicate recombination or homoplasies (repeated mutations in
64 independent lineages at the same locus), consistent with a large, genetically diverse and
65 recombining *V. cholerae* population[33,34] (Supplementary Fig. 1). Given such a recombining
66 population with mobile virulence factors, it remains puzzling why the ability to cause pandemic
67 cholera is limited to the PG group.
68 To find unique features of PGs that could explain their pandemic success, we first
69 searched for genes present in PGs but absent in EGs. Although there are a few genes which
70 appear to be universally present in PGs, including the major virulence gene clusters CTXΦ and
71 VPI-1, none of these genes are unique to PGs as they are also present in some EGs
72 (Supplementary Table 3). This observation is consistent with previous work showing that these
73 virulence genes are rapidly gained and lost in both EGs and PGs[10,33]. Therefore, there appear to
74 be no gene families whose presence can easily explain the origin of pandemic cholera, nor are
75 there strong boundaries to gene transfer between PGs and EGs.
76 Given the lack of PG-specific genes, we hypothesized that the environmental ancestor of
77 the PGs had a particular genomic background containing alleles of core genes – which we term
78 virulence adaptive polymorphisms (VAPs) – that served as “preadaptations” and enhanced its
79 potential to give rise to pandemic disease. We started our search for VAPs by identifying SNPs
80 in the aligned core of the 22 primary dataset genomes with one allele uniquely present in all
81 PGs and a different allele uniquely present in all EGs. We called this a “fixed” SNP pattern. We

82 identified 819 such fixed SNPs, which fall on the long branch separating PGs and EGs (Fig. 1a)
83 and are distributed across the genome (Fig. 1b; Supplementary Fig. 3). Some fixed SNPs could
84 have contributed to the evolution of the pandemic phenotype in PG (*i.e.* candidate VAPs, under
85 selection for increased virulence or host colonization), while others could be selectively neutral
86 hitchhikers on the PG genomic background. Using the McDonald-Kreitman test[35], we found
87 evidence for genome-wide positive selection during the divergence of PGs from EGs, due to an
88 excess of nonsynonymous changes at fixed sites, in both primary and replication datasets
89 (Supplementary Table 4). However, no individual gene showed evidence for selection after
90 correcting for multiple tests (Methods), rendering it difficult to identify candidate VAPs using
91 fixed SNPs. Nonetheless, fixed SNPs constitute only a modest fraction of possible SNP patterns
92 (Supplementary Table 2), and VAPs could also exist at other SNP sites that might shed light in
93 the evolutionary past of the ancestral PG.
94 We then searched for variations following what we termed a "mixed" SNP pattern, where
95 PGs encode one fixed allele and EGs encode a "mix" of two alleles: one that is unique to EGs
96 (as in our previous search) but also one that is fixed in PGs (Fig. 1c inset). We reasoned that
97 the existence of PG-like alleles segregating in contemporary environmental populations of *V.*
98 *cholerae* could be informative about: 1) pathogen emergence, because having an allele fixed in
99 PG suggests that it could also have been present in the environmental ancestor of PG; and 2)
100 the potential for pathogen reemergence, as PG-like alleles, and thus potential VAPs, are still
101 circulating in the environmental gene pool. We identified 39,171 mixed SNPs in the primary
102 dataset. Most genes contain few mixed SNPs (median of 3) but some genes contain dense
103 clusters, resulting in a mean of 10.3 mixed SNPs per gene. The replication dataset contained
104 greater genetic diversity, but showed the same pattern of a few genes containing many mixed
105 SNPs (Supplementary Table 2). Clusters of genes containing many mixed SNPs are visible
106 when plotted across the genome (Fig. 1c). Some of these clusters are known polymorphic
107 regions of the genome, such as the loci encoding the O-antigen and the flagellum. Many of

108 these clusters could be mutation hotspots, containing SNPs not directly relevant to virulence
109 adaptation. However, clusters of mixed SNPs do not visibly overlap with clusters of overall
110 polymorphism (Fig. 1c), indicating that accumulation of mixed SNPs cannot be explained only
111 by mutation hotspots.

112 To formally exclude mutation hotspots and focus on clusters of mixed SNPs shaped
113 mainly by natural selection instead of mutation and drift, we considered only genes with an
114 excess of nonsynonymous (NS) mixed SNPs relative to synonymous (S) mixed SNPs, which
115 control for the baseline mutation rate. The mixed SNP pattern by definition groups PG-like EGs
116 away from the other environmental strains and clusters them with the PGs. We reasoned that
117 genes with an elevated NS:S ratio at mixed SNP sites were more likely to show phenotypic
118 variations and have evolved under positive selection, possibly underlying preadaptations of the
119 PG ancestor. Using a threshold of ≥12 NS mixed SNPs per gene, mixed NS:S>1.78, and mixed
120 *dN/dS*>0.60 (respectively two standard deviations above the genome-wide medians) we
121 identified five genes as candidate VAPs in the primary dataset, three of which survived multiple
122 hypothesis correction, and two of which (*ompU* and hypothetical gene VCD_001600) were also
123 found in the replication dataset (Table 1). In contrast to the star-like genome-wide phylogeny
124 (Fig. 1a), each of these five gene trees support one or more EGs grouping with PGs (Fig. 1d
125 and Supplementary Fig. 4). Three of the gene trees, including the lipid A core O-antigen ligase,
126 support EGs LMA3894-4 and 12129 branching with PGs (Supplementary Table 1 and
127 Supplementary Fig. 4), consistent with these strains being O1-like[10,36]. The O-antigen ligase
128 gene was not identified as a VAP in the replication dataset, which is expected given the
129 absence of O1-like EGs in that dataset[33]. Three additional VAPs – all hypothetical proteins –
130 were identified in the replication dataset (Supplementary Table 5).

131 Among the candidate VAPs, the gene with the most significant excess of mixed
132 nonsynonymous SNPs in both datasets is *ompU* (Table 1). OmpU is an outer membrane porin
133 that has been shown to play numerous roles in the intestinal colonization of *V. cholerae*, making

134 it a compelling candidate for phenotypic characterization[37-41]. We observed that environmental
135 strains RC385, GBE0658 and GBE0428 have PG-like *ompU* alleles: they cluster near PGs in
136 the *ompU* gene tree, share an 11 amino acid N-terminus insertion with PGs (Supplementary
137 Fig. 5), and differ by only 15 to 45 amino acids (Supplementary Table 6). In contrast, the
138 environmental strain GBE1114 differs from PG by 80 amino acid changes in OmpU and
139 branches distantly from PG (Fig. 1d). We hypothesized that the PG-like alleles present in
140 environmental strains might confer properties conducive to virulence. To test this, we
141 constructed three different mutant strains each encoding one of three environmental alleles of
142 *ompU* (all from sympatric GBE strains) into the background of N16961, a clinical strain from the
143 current pandemic (Fig. 2). OmpU was detected on a protein gel stained with Coomassie blue
144 and through immunoblot in all the constructed strains, indicating that all the strains effectively
145 produce the environmental versions of the protein (Fig. 2a and S6). The band size differences
146 reflect the diversity in protein sizes among the three environmental OmpU variants
147 (Supplementary Fig. 5 and Supplementary Table 5).
148 We performed three sets of experiments to compare the phenotypes conferred by EG
149 and PG-like alleles of *ompU*. First, we determined the survival of these strains in the presence
150 of 0.4% bile, as it has been previously shown that OmpU confers resistance to this antimicrobial
151 compound[39]. The mutant strain encoding the *ompU* allele from the environmental strain
152 GBE1114 ($OmpU^{GBE1114}$) had diminished bile tolerance and its survival in the presence of bile is
153 similar to that of a deletion mutant (Fig 2b). $OmpU^{GBE0428}$ showed a statistically significant
154 decrease in survival in the presence of bile, closer to wild-type PG than to a deletion mutant
155 (Fig. 2b). In contrast, $OmpU^{GBE0658}$ show similar survival in the presence of bile as a PG strain
156 (Fig. 2b). These experiments indicate that some environmental alleles of *ompU* confer
157 properties beneficial for virulence. Second, we tested the survival of the mutants in the presence
158 of polymyxin B, as OmpU also confers resistance against this antibiotic[37]. The ability to tolerate
159 the antimicrobial effects of polymyxin B appears to be independent of which *ompU* allele is

160 encoded by *V. cholerae*, as the three strains encoding environmental alleles of *ompU* had a
161 similar survival rate (Fig. 2c). This experiment shows that OmpU^{GBE1114} is not simply a loss of
162 function mutant, and is not equivalent to a knockout. Third, we determined the intestinal
163 colonization of the *ompU* mutants by performing competition assays using the infant mouse
164 model of human infection. We found that OmpU^{GBE1114} had a colonization defect similar to
165 Δ*ompU* whereas OmpU^{GBE0658} was able to colonize similarly to the strain with the wild-type PG
166 allele (Fig. 2d). These results indicate that certain naturally occurring environmental alleles of
167 *ompU* confer properties that provide an advantage to *V. cholerae* during or prior to host
168 colonization, as would be expected for VAPs.

169 The *ompU*^{GBE1114} allele appears to be maladaptive for intestinal colonization; however, its
170 presence in several environmental isolates of *V. cholerae* prompted us to investigate its
171 possible adaptive role in the environment (Fig. 2e). *V. cholerae* forms biofilms on the surface of
172 biotic and abiotic environmental surfaces, [28,42-45] yet biofilm formation inside the host is thought
173 to impair intestinal colonization[28,41-43,45,46]. Strains with deletions in *ompU* have been shown to
174 form a more robust biofilm on abiotic surfaces[47]. We found that OmpU^{GBE1114} has higher biofilm
175 formation than the strain encoding the wild-type PG allele, similar to the Δ*ompU* strain (Fig. 2e).
176 Both OmpU^{GBE0658} and OmpU^{GBE0428} formed biofilm similar to the strains with the wild-type PG
177 allele (Fig. 2e). It therefore appears that there is an evolutionary trade-off between encoding the
178 PG-like or EG-like alleles of VAPs, as they seem to confer mutually exclusive traits: either
179 biofilm formation or bile resistance and host colonization. This suggests that environmental
180 strains can be divided into subgroups which, due to their contrasting lifestyles, differ in their
181 potential to give rise to pandemic clones.

182 In summary, we have determined that virulence adaptive polymorphisms are present in
183 the environment, and shown how these VAPs can be identified, based on two independent sets
184 of genomes. The top candidate VAP, *ompU*, was identified in both of our genomic datasets.
185 However, the replication dataset yielded three additional candidates (Supplementary Table 5),

186 suggesting the potential for other VAPs to be identified with further sampling of genetically
187 diverse environmental genomes.

188 Our experiments show that the *ompU* allele from some environmental strains, such as
189 GBE0658, confers properties that allow for host colonization equally as efficient as alleles from
190 clinical strains (Fig. 2). This leads to a natural question: Why have environmental strains with
191 PG-like alleles not emerged as pandemic cholera strains? One immediate answer is because
192 they might not encode the key virulence factors, CT and TCP. However, CTXΦ and VPI-1 have
193 been found in environmental *V. cholerae* strains that do not cause disease in humans[10,11,21-28]. It
194 therefore appears that a variety of virulence adaptive genes and alleles are circulating in the
195 environment, but only the PG group encodes the optimal combination of VAPs that allowed for
196 pandemic potential (Fig. 3). We propose a conceptual model in which VAPs circulate in a
197 diverse, recombining environmental gene pool, being maintained in the population through
198 various biotic and abiotic selective pressures (Fig. 3a). A new ecological opportunity occurs,
199 such as human consumption of brackish water or transient colonization of other animal hosts,
200 which leads to the proliferation and gradual enrichment in the population of clones encoding a
201 mosaic of VAPs (Fig. 3b). Finally, a genome encoding a critical combination of VAPs acquires
202 key virulence factors allowing it to emerge as a virulent, potentially pandemic clone (Fig. 3c).

203 Our model posits that VAPs are circulating in the environment prior to the acquisition of
204 key virulence factors. This is based on experimental evidence that a current pandemic strain,
205 N16961, which encodes the key virulence factors CT and TCP, cannot efficiently colonize the
206 mammalian intestine without a PG-like *ompU* allele. If virulence adaptive alleles of *ompU* are
207 indeed required prior to the acquisition of virulence factors, we would expect the same
208 phenotypes of PG-like and EG-like *ompU* alleles in the genomic background of a more deeply
209 branching PG isolate, such as classical *V. cholerae*, the cause of the first six cholera
210 pandemics. Indeed, we found that PG-like *ompU* alleles in the classical O395 background
211 conferred efficient host colonization (Supplementary Fig. 7), which is consistent with an *ompU*

212 VAP having played a role in the emergence of pathogenic *V. cholerae* prior to the acquisition of
213 key virulence factors.

214 Our model further postulates that virulence adaptive alleles become enriched in the
215 environmental population. Such enrichment would be made possible if these alleles provided a
216 selective advantage in a newly available ecological niche, such as a human population
217 consuming brackish water[3]. In previous work, we modeled an evolving, recombining microbial
218 population that encounters a new ecological opportunity[48]. When adaptation to the new niche
219 depends on few loci under positive selection in the niche, it is more likely for recombination to
220 assemble the right combination of alleles in the same genome. These loci (akin to VAPs) could
221 contribute additively or synergistically to fitness. For example, the PG-like *ompU* alleles confer a
222 tenfold increase in fitness during host colonization (Fig. 2d). Other VAPs might contribute further
223 to this enhancement in host colonization. As more loci are involved in adaptation, it becomes
224 less likely to achieve the optimal combination. Assuming the *V. cholerae* population undergoes
225 approximately 100 recombination events per locus per generation (6.5 recombination events for
226 every point mutation)[34], equivalent to a recombination rate of 10^{-4} in the modelled population of
227 size 10^{6}, an optimal combination of alleles at five loci could conceivably evolve, but seven loci is
228 very unlikely[48]. Therefore, if virulence depended on five or fewer positively selected loci in the *V.*
229 *cholerae* genome, the optimal combination of alleles would be expected to appear repeatedly in
230 nature. Given suitable ecological opportunities, it is then plausible that pandemic *V. cholerae*
231 could emerge multiple times, originating from outside the PG group. However, the number of
232 loci that are sufficient for the emergence of a virulent strain remains unknown and if it was much
233 greater than five, pathogen emergence would be naturally limited. We identified eight candidate
234 VAPs in our two datasets, and if all eight are phenotypically confirmed to be VAPs, this number
235 of VAPs (<5) might naturally limit pandemic clone emergence. Furthermore, these eight
236 candidate VAPs passed stringent filters and we suspect there might exist additional VAPs in the
237 genome, identifiable by further sampling and experimentation. We also note that LMA3894-4, a

238 PG-like environmental strain containing PG-like alleles at three out of five candidate VAP loci
239 (Supplementary Fig. 4), does not contain a predicted *ompU* ortholog, suggesting a colonization-
240 impaired phenotype. Similarly, strain 12129 also contains PG-like alleles at these VAP loci, but
241 has an EG-like allele of *ompU* (Fig 1d). Therefore, no genome other than PGs contains a
242 clinical-like combination of VAPs, due either to recombination or selection limitation.

243 Here we have described a framework for identifying loci that are present in a natural
244 population and confer properties beneficial for virulence prior to acquisition of essential
245 virulence genes and host colonization. This framework could be applied to other bacterial
246 pathogens that emerge as clonal offshoots from non-virulent relatives, including *Yersinia,*
247 *Salmonella, Escherichia*, or other pathogenic *Vibrio* species[1,4,7-9]. Pathogens that emerged
248 through clonal expansion limit our ability to dissect the genetic basis of their pathogenicity,
249 because bacterial genome-wide association studies lack power when the phenotype of interest
250 has evolved only once[49]. Our framework therefore provides a way forward to identify the genetic
251 basis of virulence, even in pathogens that evolved through clonal expansion, and begin to
252 assess the risk of pathogen emergence and reemergence from environmental gene pools.

253 **Methods**

254 **Genome sequencing.** DNA from clinical isolates (Bgd1, Bgd5, Bgd8, MQ1795) and
255 environmental isolates (GBE0428, GBE0658, GBE1068, GBE1114, GBE1173) was extracted
256 using the Gentra kit (QIAGEN) and purified using the MoBio PowerClean Pro DNA Clean-Up
257 Kit. Multiplexed genomic libraries were constructed using the Illumina-compatible Nextera DNA
258 Sample Prep kit following the manufacturer's instructions. Sequencing was performed with 250-
259 bp paired-end (v2 kit) reads on the illumina MiSeq. Sequence reads will be made available in
260 Genbank SRA shortly.

261

262 **Genome assembly.** To exclude low-quality data, raw reads were filtered with Trimmomatic[50].
263 The 15 first bases of each reads were trimmed and reads containing at least one base with a
264 quality score of <30 were removed. *De novo* assembly was performed on the resulting reads
265 using Ray v2.3.1[51].

266

267 **Genome alignment, annotation and SNP calling.** We used mugsy v.1 r.2.2[52] with default
268 parameters to align the primary dataset of 22 *V. cholerae* genomes (Supplementary Table 1).
269 From this alignment, we extracted dimorphic SNP sites and annotated genes according to MJ-
270 1236 as a reference genome. We defined the core genome as locally colinear blocks (LCBs)
271 with all 22 genomes present in the alignment. We replicated the alignment, annotation and SNP
272 calling using 21 different *V. cholerae* genomes, mainly from Orata *et al*.[33], plus the MJ-126
273 reference (Supplementary Table 1).

274
275 **Definition of orthologous groups.** Genomes were annotated using the RAST web server
276 (www.rast.nmpdr.org)[53]. Annotated genes were clustered into orthologous groups using
277 OrthoMCL (www.orthomcl.org)[54] with default parameters, yielding 2844 orthologous groups.

278

279 **Phylogenetic analysis.** We constructed a core genome phylogeny using the concatenated
280 alignment of 1031 single-copy orthologous protein-coding genes (present in exactly one copy in
281 each of the 22 primary dataset genomes). Each protein sequence was aligned with Muscle[55],
282 and the concatenated alignment was used to infer an approximate maximum likelihood
283 phylogeny with FastTree v. 2.1.8[56] using default parameters (Fig. 1a). Individual gene trees (Fig.
284 1d) were built in the same way. We constructed a neighbour-net of the 22 genomes using
285 SplitsTree v.4.10[57], based on dimorphic SNPs from the mugsy genome alignment, excluding
286 sites with gaps.
287
288 **Tests for selection.** We conducted a genome-wide version of the McDonald-Kreitman test[35] by
289 first counting the number of fixed nonsynonymous (fn), fixed synonymous (fs), polymorphic
290 nonsynonymous (pn), and polymorphic synonymous (ps) sites within each gene. We then
291 summed these values across all genes (FN, FS, PN, and PS) and calculated the genome-wide
292 Fixation Index, FI=(FN/FS)/(PN/PS). A fixation index greater than one suggests positive
293 selection between the ingroup and outgroup (in this case, between EGs and PGs). However,
294 care must be taken when computing a genome-wide FI because summing genes with different
295 amounts of substitutions and polymorphism can result in FI>1 in the absence of selection[58]. We
296 therefore performed 1000 permutations of the data, keeping the row totals (fn+fs and pn+ps)
297 and column totals (fn+pn and fs+ps) constant and recomputing FI. We used the mean FI from
298 the permutations as the expected value of FI under neutral evolution. To evaluate the
299 hypothesis that the observed FI was higher than expected, suggesting positive selection, we
300 computed a *P*-value as the fraction of permutations with FI greater or equal to the observed FI.
301 We repeated the test using polymorphism from either the PG group or the EG group
302 (Supplementary Table 3).
303 To identify individual genes under selection between PGs and EGs in the primary
304 dataset, we restricted our search to 87 genes with fn > 0.68 1 and fn:fs > 1.48 (respectively two

305 standard deviations about the genome-wide medians). We then calculated the gene-specific FI
306 and assessed its significance with a Fisher exact test. We found no genes with FI significantly
307 greater than one, after Bonferroni correction for 87 tests. Similarly, in the replication dataset, we
308 restricted our search to 26 genes with fn > 1.5 and fn:fs > 1.70, none of which had significantly
309 high FI after correction for multiple tests.

310 To identify genes with an excess of nonsynonymous mixed SNPs (likely due to selection
311 for amino acid changes), we restricted our search to five genes with ≥ 12 NS mixed SNPs per
312 gene and mixed NS:S > 1.78 (respectively two standard deviations above the genome-wide
313 medians). We used a one-sided binomial test to assess whether the observed NS:S ratio for
314 each gene was significantly greater than the genome-wide median NS:S ratio of 0.5 (after
315 adding a pseudocount of one to both NS and S). Three out of the five genes had a significantly
316 high mixed NS:S ratio ($P < 0.05$) after Bonferroni correction for five tests (Table 1). We repeated
317 this procedure in the replication dataset, identifying genes with ≥ 18 NS mixed SNPs per gene
318 and mixed NS:S > 1.65 (respectively two standard deviations above the genome-wide
319 medians). We used a one-sided binomial test to assess whether the observed NS:S ratio for
320 each gene was significantly greater than the genome-wide median NS:S ratio of 0.33 (after
321 adding a pseudocount of one to both NS and S). The results of these tests are shown for genes
322 also identified in the primary dataset (Table 1) and three additional genes identified in the
323 replication dataset (Supplementary Table 5).

324

325 **Bacterial strains and plasmids.** *V. cholerae* O395 and *V. cholerae* N16961 were used as wild-
326 type strains of classical and El Tor biotypes respectively. Strains cultivated on solid medium
327 were grown on LB agar; strains in liquid media were grown in aerated LB broth at 37°C.
328 pKAS154 was used for allelic exchange[59]. When necessary, media was supplemented with
329 antibiotics to select for certain plasmids or strains of *V. cholerae* at the following concentrations:
330 gentamycin, 30μg/ml; kanamycin, 45μg/ml; polymyxin B, 50μg/ml; and streptomycin, 1 mg/ml.

331

332 **Strain construction.** In-frame deletions and exchange of *ompU* alleles in both O395 and
333 N16961 biotypes were constructed via homologous recombination[59]. For *ompU* deletions, PCR
334 was used to amplify two 500 bp fragments flanking the *ompU* gene and to introduce restriction
335 sites for cloning into the suicide vector pKAS154. For exchange of environmental *ompU* alleles,
336 the respective allele was also amplified using the extracted DNA from each environmental
337 strain. Restriction sites were introduced in the primers. The fragments were then cloned into a
338 restriction-digested suicide plasmid, pKAS154, using a four-segment ligation for each
339 environmental allele exchange mutants (plasmid, *ompU* flanking fragments and environmental
340 *ompU* allele). The resulting plasmids were electroporated into *Escherichia coli* S17-1λpir. *E. coli*
341 with the constructed inserts. Different sizes of *ompU* genes were confirmed, sequenced and
342 compared to the genome sequences of environmental strains. For both deletion and exchange
343 mutants, plasmids with the insert of interest were mated with wild-type *V. cholerae* O395 or
344 N16961, and allelic exchange was carried out by selection on antibiotics as described
345 previously[59]. For a more detailed description of allelic exchange please refer to Skorupski and
346 Taylor 1998[59]. Potential mutants were screened using PCR: two primers flanking the deletion
347 construct were used to amplify chromosomal DNA isolated from plated *V. cholerae*. The lengths
348 of the PCR fragments were analyzed on 0.8% agarose gel for gene deletions and putative
349 deletions and allele exchange were subsequently confirmed by DNA sequencing.

350

351 **OmpU visualization and immunobloting.** Whole cell protein extracts were prepared from
352 cultures grown for overnight at 37°C in a rotary shaker. The extracts were subjected to SDS-
353 PAGE on 16% Tris Glycine gels (Invitrogen). OmpU bands were visualized after protein gels
354 were stained by Coomassie blue overnight. Prior to staining the gels were transferred to
355 nitrocellulose membranes using iBlot (Invitrogen). The membranes were blocked O/N in Tris-

356 Buffered Saline, 3% BSA. Primary OmpU antibodies were diluted 1:10,000 in TBST (Tris-
357 Buffered Saline, 0.5% Tween-20). Membranes were incubated with primary antibodies for 2
358 hours at room temperature. After incubation, the membranes were washed with TBST four
359 times. Goat anti-rabbit secondary antibodies (BioRad) were diluted 1:10,000 in TBST and
360 incubated for 30 minutes at room temperature. The membranes were washed 4 times with TBS
361 (Tris-Buffered Saline). Reactive protein bands were detected via ECL (Amersham).
362
363 **Survival assays.** *V. cholerae* strains were cultured overnight in LB broth at 37°C in a rotary
364 shaker. Overnight cultures were diluted 1:100 in LB and grown to an OD600 of 0.5. Cells were
365 pelleted and resuspended in either LB broth, LB containing 0.4% bile bovine (Sigma), or LB
366 containing 1000U/ml of polymyxin B (Sigma). Cultures were incubated for 1h at 37°C in a rotary
367 shaker. After incubation CFU/ml of each culture was calculated by plating dilutions in LB plates.
368 Survival was calculated by comparing the number of CFU/ml in LB plus treatment versus LB.
369 N≥6. No samples were excluded.
370
371 **Infant mouse competition assays.** Overnight cultures were diluted 1:100. Each test strain was
372 mixed in a 1:1 ratio with a Δ*lacZ* reference strain. Four- to five-day-old CD-1 mice (*Mus*
373 *musculus*) from several mixed litters were randomly inoculated orogastrically in blinded
374 experiments with 50μl of the bacterial mixture. Sex of the animals was not inspected prior to
375 inoculations. The intestines were harvested 24h post-inoculation and homogenized in 4ml of LB
376 broth containing 10% glycerol. The mixtures were serially diluted and plated on LB agar plates
377 supplemented with streptomycin and 5-bromo-4-chloro-3-indolyl-D-galactopyramoside (X-Gal)
378 (40μg/ml). The competition indices were calculated as previously described by others, test
379 (output CFUs/input CFUs)/reference (output CFUs/ input CFUs) and the sample size selected
380 was appropriate for statistical analysis. No samples or animals were excluded.

381

382 **Biofilm assays.** 96-well plate assay. Cultures were incubated overnight at 30°C. 100µl of 1:100
383 dilutions of overnight cultures were placed per well in 96-well plates. Plates were left at 25°C for
384 24h. Liquid contents were discarded and plates were washed 2 times with LB. 200µl of 0.01%
385 crystal violet was added per well and incubated at room temperature for 5 minutes. Liquid
386 contents were discarded and plates were washed extensively with dH_2O. After the plates were
387 dry, biofilms were resuspended in 150µl of 50% acetic acid. Contents were transferred to a flat
388 bottom dish and quantitated in a microtiter plate reader at OD550. Values were plotted using
389 Prism software. N=15.
390

391 **Supplementary Text**

392 **Choice of genomes and power considerations**

393 In the primary dataset, we sought a sample that spanned the genetic diversity within both PG
394 and EG groups. The PG group contains two clades, PG-1 and PG-2 (as defined by Chun et al.
395 2009), and later confirmed by Mutreja et al. [15], who defined lineages L2,3,4,5,6 and 8 within PG-
396 1, and L1 and 7 are PG-2). To span the genetic diversity of PG, the primary dataset contained 6
397 genomes from PG-1 (including representatives of all 3 waves of the 7th pandemic defined by
398 Mutreja et al.: N16961 from Wave 1, MJ1236 from Wave 2, and Bgd1, Bgd5, Bgd8 and MQ1795
399 from Wave 3) and one genome from PG-2 (O395, from Mutreja's L1). The relationships among
400 all PG genomes used in this study is illustrated in Supplementary Figure 8.

401 We found that adding PGs does not significantly affect our detection of fixed or mixed
402 SNPs (Supplementary Figure 9, panels a and b). This is because PGs are very closely related,
403 and adding new genomes does not add much genetic diversity. As a result, the effect of
404 including a PG-2 strain (O395, the deepest branch within the PG clade) has a much bigger
405 impact on the detection of fixed and mixed SNPs than the number of genomes included
406 (compare cyan and dark blue points in Supplementary Figure 9). Whether O395 is included or
407 not, the number of fixed SNPs declines very slightly as new genomes are added. This is
408 expected because SNPs that in reality are polymorphic (not fixed) within PG are more likely to
409 be falsely identified as fixed when fewer genomes are sampled. Similarly, the number of mixed
410 SNPs identified declines as more PGs are added. This is because mixed SNPs must by
411 definition be fixed in PGs, but polymorphic in EGs. Together, these results show that, unless
412 new deep-branching lineages (like PG-2) are identified within PG, sampling additional genomes
413 should not greatly affect the dataset of fixed or mixed SNPs. Conversely, EG sampling has a
414 large effect: adding more EGs reduces the number of fixed SNPs identified and increases the
415 number of mixed SNPs, because polymorphic sites are more likely to be sampled as more
416 genetic diversity within EGs is added (Supplementary Figure 9, panels c and d). However, both

417 the number of fixed and mixed SNPs reach a plateau at around 12 sampled EGs, suggesting
418 that additional sampling is unlikely to yield more mixed SNPs (which would increase power to
419 detect putative VAPs) or to greatly reduce the number of false-positive fixed SNPs. Therefore, it
420 seems that our power to detect fixed or mixed SNPs is unlikely to change significantly with
421 further sampling.

422 Despite this encouraging result, we nevertheless considered a replication dataset of 15
423 different EGs and 6 different PGs. Like the primary dataset, the replication dataset also
424 contained 6 genomes from PG-1 (including 7th pandemic strains MJ1236, MO10, and 2010EL,
425 as well as pre-7th pandemic strains BX330286, 274080, and MAK757, from Mutreja et al.'s L8,
426 L3, and L5, respectively) and one genome from PG-2 (V52, from Mutreja's L7). Our motivation
427 for considering the replication dataset was, first, that Orata et al.[33] recently identified five clonal
428 complexes on long branches which would be expected to add additional genetic diversity to the
429 EG group, and second, that the power to detect SNPs depends not just on their existence (*i.e.*
430 the actual genetic diversity), but also the quality of the alignment (*i.e.* if true SNPs are present in
431 the alignment). Therefore, the quality of genome assembly and alignment should also affect our
432 ability to detect SNPs. While the primary dataset generally contained assemblies of good quality
433 (median of 55.5 contigs and N50 of 251 Kbp; Supplementary Table 7), a few EG assemblies
434 were poorer (*e.g.* 314 contigs in strain 62339; 269 contigs in CT536993). In contrast, the
435 replication dataset contained no genomes with >150 contigs, yielding a substantially less
436 fragmented alignment (with fewer and longer locally colinear blocks compared to the primary
437 dataset) with a larger core genome (Supplementary Table 8). Due to a combination of additional
438 genetic diversity and a better alignment, the number of mixed SNPs more than doubled in the
439 replication dataset compared to the primary dataset (Supplementary Table 2). This suggests
440 that our power estimates (Supplementary Figure 9) are not entirely realistic, and sampling
441 additional EGs with high-quality assemblies will likely yield more mixed SNPs. However, power
442 to identify mixed SNPs is not the same as power to identify VAPs, which was the goal of our

443 study. Despite identifying more mixed SNPs, the replication dataset only revealed three new
444 candidate VAPs (Supplementary Table 5) and validated some of those found in the primary
445 dataset (Table 1). Therefore, while additional sampling of EGs is likely to yield more mixed
446 SNPs, we expect the number of new VAPs identified to be modest.

References

1. McNally, A., Thomson, N. R., Reuter, S. & Wren, B. W. 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.* **14,** 177–190 (2016).
2. Luo, C. *et al.* Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 7200–7205 (2011).
3. Boucher, Y., Orata, F. D. & Alam, M. The out-of-the-delta hypothesis: dense human populations in low-lying river deltas served as agents for the evolution of a deadly pathogen. *Front. Microbiol.* **6,** L19401 (2015).
4. Reen, F. J., Almagro-Moreno, S., Ussery, D. & Boyd, E. F. The genomic code: inferring Vibrionaceae niche specialization. *Nat. Rev. Microbiol.* **4,** 697–704 (2006).
5. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
6. Shapiro, B. J. How clonal are bacteria over time? *Curr. Opin. Microbiol.* **31,** 116–123 (2016).
7. Groisman, E. A. & Ochman, H. Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87,** 791–794 (1996).
8. Ochman, H. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292,** 1096–1099 (2001).
9. Cui, Y. *et al.* Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen *Vibrio parahaemolyticus*. *Mol. Biol. Evol.* **32,** 1396–1410 (2015).
10. Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 15442–15447 (2009).
11. Faruque, S. M. & Mekalanos, J. J. Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends Microbiol.* **11,** 505–510 (2003).
12. Kaper, J. B., Morris, J. G. & Levine, M. M. Cholera. *Clin. Microbiol. Rev.* **8,** 48–86 (1995).
13. Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T. & Calderwood, S. B. Cholera. *Lancet* **379,** 2466–2476 (2012).
14. Boucher, Y. Sustained local diversity of *Vibrio cholerae* O1 biotypes in a previously cholera-free country. *mBio* **7,** e00570–16 (2016).
15. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477,** 462–465 (2011).
16. De, S. N., Bhattacharya, K. & Sarkar, J. K. A study of the pathogenicity of strains of Bacterium coli from acute and chronic enteritis. *J. Pathol. Bacteriol.* **71,** 201–209 (1956).
17. Taylor, R. K., Miller, V. L., Furlong, D. B. & Mekalanos, J. J. Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proc. Natl. Acad. Sci. U.S.A.* **84,** 2833–2837 (1987).
18. Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272,** 1910–1914 (1996).
19. Karaolis, D. K. *et al.* A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 3134–3139 (1998).
20. Almagro-Moreno, S., Murphy, R. A. & Boyd, E. F. How genomics has shaped our understanding of the evolution and emergence of pathogenic *Vibrio cholerae*. In *Genomes of Foodborne and Waterborne Pathogens* 85–99 (ASM Press, 2011).
21. Faruque, S. M. *et al.* Analysis of clinical and environmental strains of nontoxigenic *Vibrio cholerae* for susceptibility to CTXPhi: molecular basis for origination of new strains with epidemic potential. *Infect. Immun.* **66,** 5819–5825 (1998).
22. Rahman, M. H. *et al.* Distribution of genes for virulence and ecological fitness among diverse *Vibrio cholerae* population in a cholera endemic area: tracking the evolution of pathogenic strains. *DNA and Cell Biology* **27,** 347–355 (2008).
23. Faruque, S. M. *et al.* Genetic diversity and virulence potential of environmental *Vibrio cholerae* population in a cholera-endemic area. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 2123–2128 (2004).
24. Chakraborty, S. *et al.* Virulence genes in environmental strains of *Vibrio cholerae*. *Appl. Environ. Microbiol.* **66,** 4022–4028 (2000).
25. Rivera, I. N., Chun, J., Huq, A., Sack, R. B. & Colwell, R. R. Genotypes associated with virulence in environmental isolates of *Vibrio cholerae. Appl. Environ. Microbiol.* **67,** 2421–2429 (2001).
26. Mukhopadhyay, A. K., Chakraborty, S., Takeda, Y., Nair, G. B. & Berg, D. E. Characterization of VPI pathogenicity island and CTXphi prophage in environmental strains of *Vibrio cholerae*. *J. Bacteriol.* **183,** 4737–4746 (2001).
27. Gennari, M., Ghidini, V. & Lleo, M. M. Virulence genes and pathogenicity islands in environmental Vibrio strains non-pathogenic to humans. *FEMS Microbiol. Ecol.* **82,** 563–573 (2012).
28. Almagro-Moreno, S. & Taylor, R. K. Cholera: Environmental reservoirs and impact on disease transmission. *Microbiol. Spect.* **1,** 149–165 (2013).

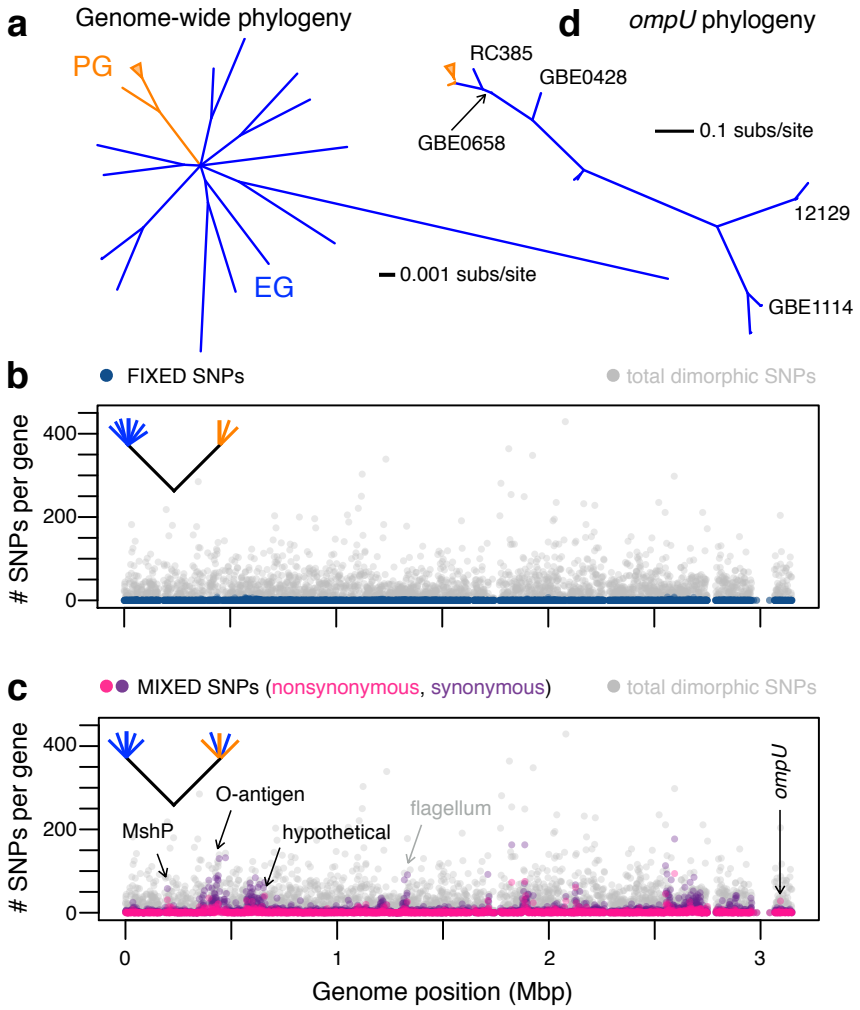
29. Dalsgaard, A. *et al.* Characterization of *Vibrio cholerae* non-O1 serogroups obtained from an outbreak of diarrhea in Lima, Peru. *J. Clin. Microbiol.* **33,** 2715–2722 (1995).
30. Bag, P. K. *et al.* Putative virulence traits and pathogenicity of *Vibrio cholerae* Non-O1, Non-O139 isolates from surface waters in Kolkata, India. *Appl. Environ. Microbiol.* **74,** 5635–5644 (2008).
31. Dziejman, M. *et al.* Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 3465–3470 (2005).
32. Schuster, B. M. *et al.* Ecology and genetic structure of a northern temperate *Vibrio cholerae* population related to toxigenic isolates. *Appl. Environ. Microbiol.* **77,** 7568–7575 (2011).
33. Orata, F. D. *et al.* The dynamics of genetic interactions between *Vibrio metoecus* and *Vibrio cholerae*, two close relatives co-occurring in the environment. *Genome Biol. Evol.* **7,** 2941–2954 (2015).
34. Keymer, D. P. & Boehm, A. B. Recombination shapes the structure of an environmental *Vibrio cholerae* population. *Appl. Environ. Microbiol.* **77,** 537–544 (2011).
35. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).
36. Pérez Chaparro, P. J. *et al.* Whole genome sequencing of environmental *Vibrio cholerae* O1 from 10 nanograms of DNA using short reads. *J. Microbiol. Methods* **87,** 208–212 (2011).
37. Mathur, J. & Waldor, M. K. The *Vibrio cholerae* ToxR-regulated porin OmpU confers resistance to antimicrobial peptides. *Infect. Immun.* **72,** 3577–3583 (2004).
38. Provenzano, D., Lauriano, C. M. & Klose, K. E. Characterization of the role of the ToxR-modulated outer membrane porins OmpU and OmpT in *Vibrio cholerae* virulence. *J. Bacteriol.* **183,** 3652–3662 (2001).
39. Provenzano, D., Schuhmacher, D. A., Barker, J. L. & Klose, K. E. The virulence regulatory protein ToxR mediates enhanced bile resistance in *Vibrio cholerae* and other pathogenic Vibrio species. *Infect. Immun.* **68,** 1491–1497 (2000).
40. Merrell, D. S., Bailey, C., Kaper, J. B. & Camilli, A. The ToxR-mediated organic acid tolerance response of *Vibrio cholerae* requires OmpU. *J. Bacteriol.* **183,** 2746–2754 (2001).
41. Almagro-Moreno, S., Pruss, K. & Taylor, R. K. Intestinal colonization dynamics of *Vibrio cholerae*. *PLoS Pathog.* **11,** e1004787 (2015).
42. Yildiz, F. H. & Schoolnik, G. K. *Vibrio cholerae* O1 El Tor: identification of a gene cluster required for the rugose colony type, exopolysaccharide production, chlorine resistance, and biofilm formation. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 4028–4033 (1999).
43. Watnick, P. I., Fullner, K. J. & Kolter, R. A role for the mannose-sensitive hemagglutinin in biofilm formation by *Vibrio cholerae* El Tor. *J. Bacteriol.* **181,** 3606–3609 (1999).
44. Watnick, P. I. & Kolter, R. Steps in the development of a *Vibrio cholerae* El Tor biofilm. *Mol. Microbiol.* **34,** 586–595 (1999).
45. Marsh, J. W. & Taylor, R. K. Genetic and transcriptional analyses of the *Vibrio cholerae* mannose-sensitive hemagglutinin type 4 pilus gene locus. *J. Bacteriol.* **181,** 1110–1117 (1999).
46. Hsiao, A., Liu, Z., Joelsson, A. & Zhu, J. *Vibrio cholerae* virulence regulator-coordinated evasion of host immunity. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 14542–14547 (2006).
47. Valeru, S. P., Wai, S. N., Saeed, A., Sandström, G. & Abd, H. ToxR of *Vibrio cholerae* affects biofilm, rugosity and survival with *Acanthamoeba castellanii*. *BMC Res Notes* **5,** 33 (2012).
48. Friedman, J., Alm, E. J. & Shapiro, B. J. Sympatric speciation: when is it possible in bacteria? *PLoS ONE* **8,** e53539 (2013).
49. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* **25,** 17–24 (2015).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
51. Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13,** R122 (2012).
52. Angiuoli, S. V. & Salzberg S. L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27,** 334–342 (2011).
53. Aziz, R. K. *et al.* The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* **9,** 1 (2008).
54. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
56. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5,** e9490 (2010).
57. Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21,** 255–265 (2004).
58. Shapiro, J. A. *et al.* Adaptive genic evolution in the Drosophila genomes. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 2271–2276 (2007).

569 59. Skorupski, K. & Taylor, R. K. Positive selection vectors for allelic exchange. *Gene* **169,** 47–52 (1996).
570 60. Son, M. S., Megli, C. J., Kovacikova, G., Qadri, F. & Taylor, R. K. Characterization of *Vibrio cholerae* O1 El
571 Tor biotype variant clinical isolates from Bangladesh and Haiti, including a molecular genetic analysis of
572 virulence genes. *J. Clin. Microbiol.* **49,** 3739–3749 (2011).
573 61. Nair, G. B. *et al.* New variants of *Vibrio cholerae* O1 biotype El Tor with attributes of the classical biotype from
574 hospitalized patients with acute diarrhea in Bangladesh. *J. Clin. Microbiol.* **40,** 3296–3299 (2002).
575

576 **Acknowledgements**

577 The authors would like to thank the anonymous reviewers for their thoughtful comments and
578 suggestions. We would also like to thank Otto Cordero, Yves Terrat, Nicolas Tromas and
579 Britney Privett for constructive comments on the manuscript. We thank Lawrence Shelven for
580 his highly valuable technical assistance. BJS was supported by a Canada Research Chair and
581 the Canadian Institutes for Health Research. RKT was supported by a National Institutes of
582 Health grants AI039654 and AI025096. SAM was supported by startup funds from the Burnett
583 School of Biomedical Sciences at the University of Central Florida and Dartmouth College's E.
584 E. Just Postdoctoral Fellowship.

585 # Figures



586

587 **Figure 1. Comparative genomics reveals candidate virulence adaptive polymorphisms.**
588 **a,** Phylogeny of 22 *V. cholerae* genomes based on 1031 single-copy orthologs in the primary dataset. All
589 branches have local support values >0.99 (based on FastTree's approximate likelihood ratio test) except
590 for very short, deep internal branches (resulting in the star-like polytomy at the centre of the tree). Not all
591 22 genomes are visible because some have nearly identical sequences (*e.g.* 6 of the 7 PG genomes are
592 nearly identical, shown as an orange triangle; GBE1173 and GBE1114 are nearly identical, as can be
593 seen in Supplementary Figure 1). **b,** Distribution of fixed SNPs across chromosome 1. (See
594 Supplementary Fig. 3 for chromosome 2). Genome position is according to the MJ-1236 reference
595 genome. SNP-free regions (*e.g.* near 3 Mbp, the locus of the integrative conjugative element) are part of
596 the flexible genome, present in the reference but not the other 21 genomes. The schematic tree in the top
597 left illustrates the fixed SNP pattern, in which one allele is present in PGs and a different allele in EGs. **c**,
598 Distribution of mixed SNPs across the genome. The cartoon tree in the top left illustrates the mixed SNP
599 pattern, in which one allele is fixed in PGs, and another allele is polymorphic among EGs, with some EGs
600 containing the PG-like allele. Black arrows show candidate VAPs (Table 1). Grey arrow shows the
601 flagellum as an example variable region not containing candidate VAPs. **d**, *ompU* phylogeny. All visible
602 branches have local support values >0.9 except for the branch separating RC385 and GBE0658, the
603 branch grouping MJ-1236 and O395 together, and the branch grouping HE09 and VL426 together.

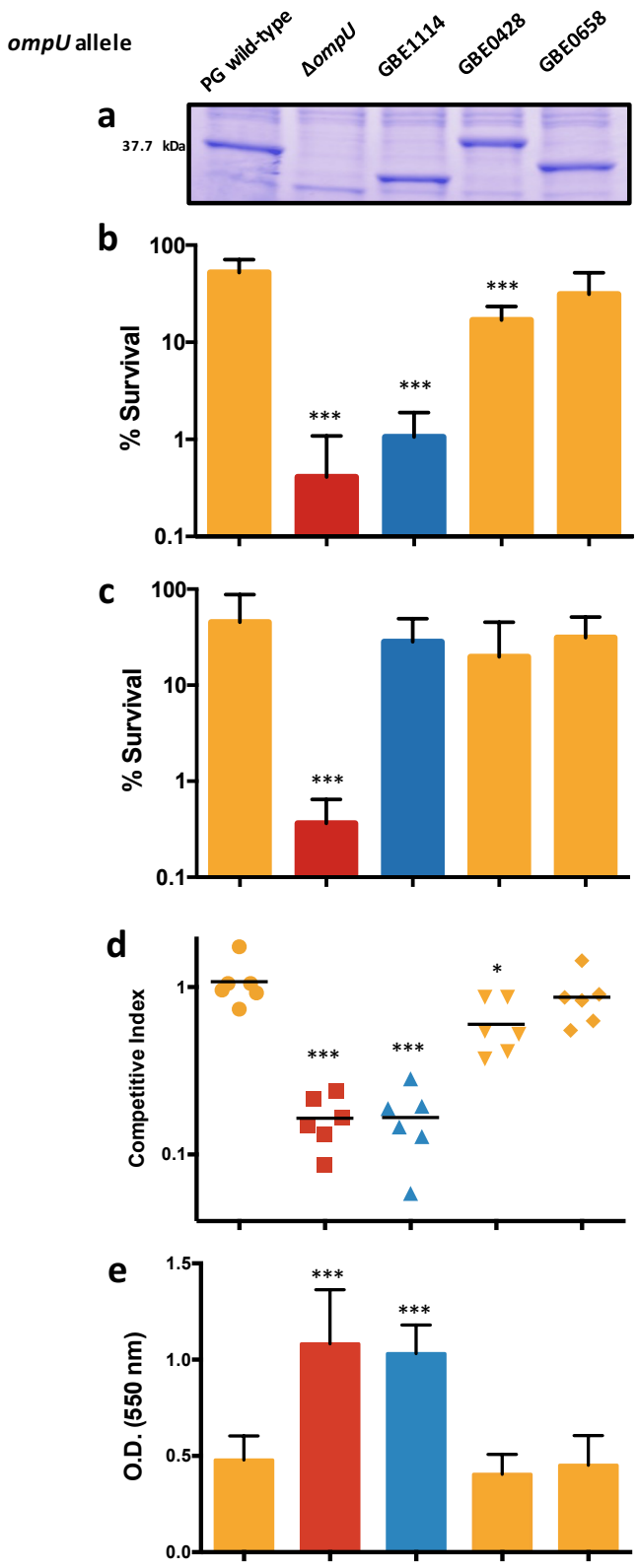
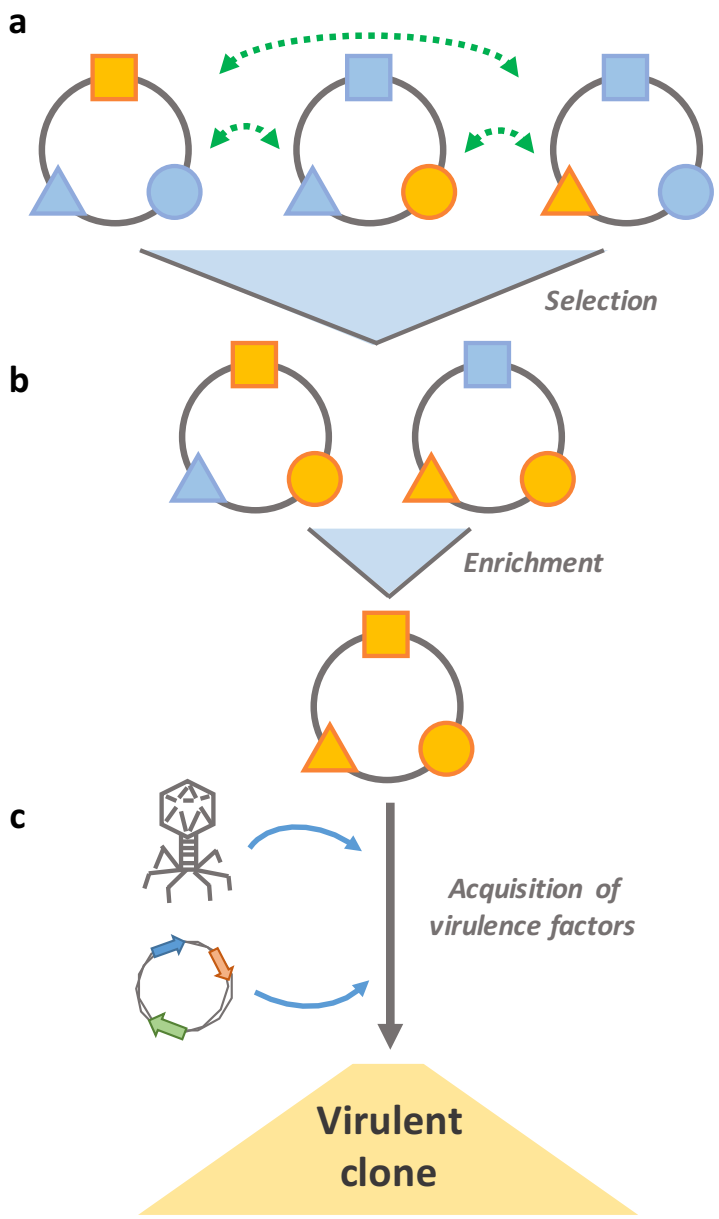


Figure 2. Phenotypic characterization of *ompU* alleles. a, OmpU production in clinical strains of *V. cholerae* encoding environmental alleles of *ompU*. Total protein lysates were run on a 16% Tris-glycine gel. OmpU bands were visualized after protein gels were stained with Coomassie blue. **b**, Survival of *ompU* mutants in the presence of bile (n=7) or **c**, polymyxin B (n=6). **d**, Colonization of the small intestine of *ompU* mutant strains (n=6). **e**, Biofilm formation of *ompU* mutant strains on an abiotic surface (n=15). Yellow bars and symbols, PG-like allele; red bars and squares, Δ*ompU*; blue bars and triangles, EG-like allele. Center values represent the mean and error bars the standard deviation. Variance between the groups was similar. Statistical comparisons were made using student's *t*-test. *$P < 0.05$, ***$P < 0.001$.

604

605 **Figure 3**



606
607
608 **Figure 3. Model of pandemic clone emergence from an environmental gene pool.** We propose a
609 model that involves three events required for the emergence of pathogenic clones from environmental
610 populations. **a**, selection of VAPs. Virulence adaptive alleles circulate in naturally occurring populations
611 (orange symbols) and can be exchanged and mobilized through recombination (green dashed arrows).
612 Ecological events (temperature, nutrient availability, pH, etc.) lead to the selection of VAPs and an
613 increase in their distribution in environmental populations. **b**, enrichment of clones. A new ecological
614 opportunity occurs (human consumption of untreated waters, transient colonization of new environmental
615 hosts, etc.) which leads to the proliferation and enrichment in the population of clones encoding a mosaic
616 of VAPs. **c**, acquisition of virulence factors**.** A strain encoding a minimum set of VAPs required for host
617 colonization acquires the virulence factors that are necessary to produce a successful infection and
618 subsequently undergoes intra-host evolution and expansion.

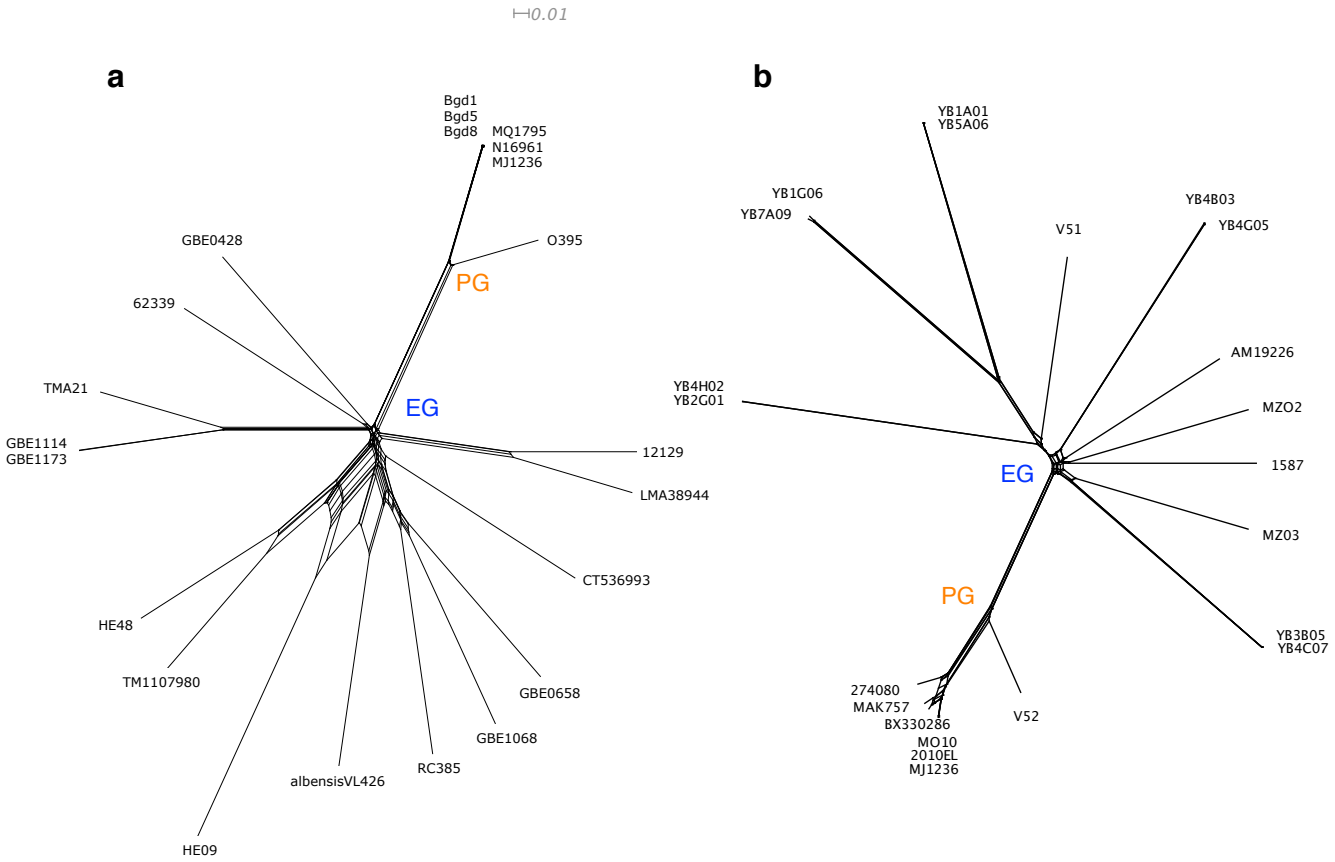
619 # Tables

620 **Table 1. Characteristics of five predicted VAPs with an excess of nonsynonymous mixed**
621 **SNPs.**

	Gene ID (VCD #)	Annotation	Gene length (bp)	Total # SNPs	Mixed *dN*	Mixed *dS*	Mixed *dN/dS*	Mixed NS	Mixed S	*P*
1	003778	outer membrane protein OmpU	1053	204 (110)	0.037 (0.018)	0.045 (0.033)	0.83 (0.53)	28 (13)	11 (8)	0.0047* (0.0068*)
2	001600	hypothetical	258	21 (26)	0.082 (0.12)	0.081 (0.065)	1.01 (1.82)	15 (22)	4 (3)	0.0096* (2.32e-8*)
3	001013	hypothetical	642	19 (13)	0.034 (0.006)	0.029 (0.012)	1.18 (0.55)	15 (2)	4 (1)	0.0096* (n.s.)
4	001209	MSHA biogenesis protein MshP	432	85 (63)	0.047 (0.025)	0.045 (0.054)	1.04 (0.46)	14 (7)	4 (5)	0.0154 (0.066)
5	001230	lipid A core O-antigen ligase	1794	75 (47)	0.0098 (0.001)	0.013 (0.004)	0.76 (0.18)	12 (0)	5 (1)	0.0717 (n.s.)

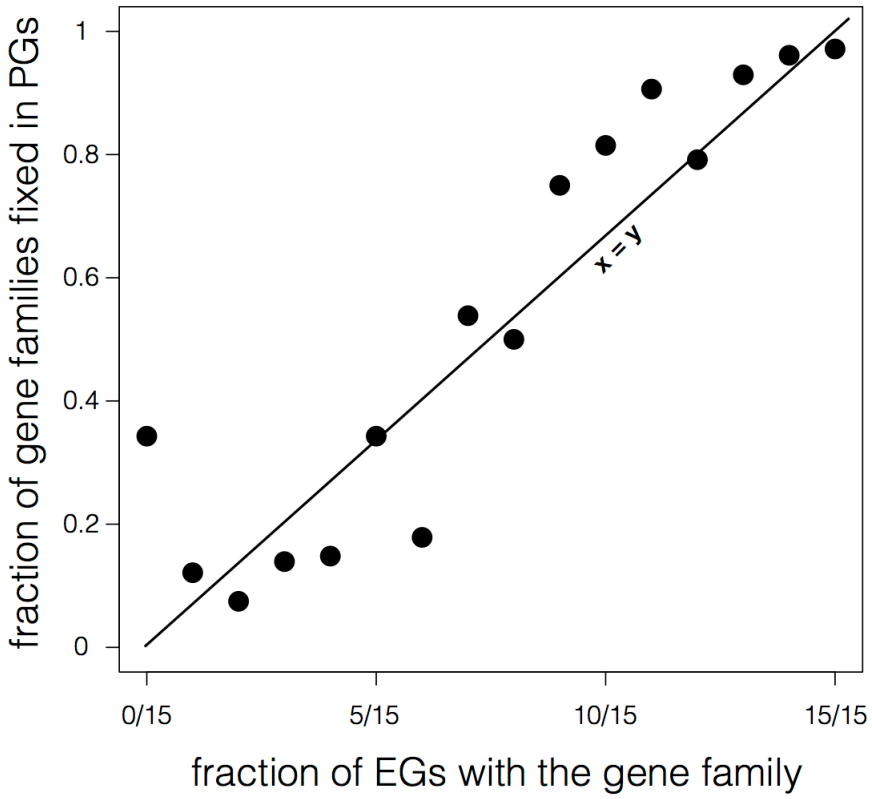
622 NS=nonsynonymous; S=synonymous. *dN* = number of nonsynonymous SNPs per
623 nonsynonymous site; *dS* = number of synonymous SNPs per synonymous site. Mixed SNPs are
624 polymorphic in EGs but fixed in PGs, meaning that at least one EG contains a PG-like allele.
625 The genes listed have mixed NS, mixed NS:S, and mixed *dN/dS* each over two standard
626 deviations above their respective genome-wide medians in the primary dataset. A one-sided
627 binomial test determined if the mixed NS:S ratio was greater than the expected genome-wide
628 median value of 0.5 per gene (uncorrected *P*-values shown; asterisks (*) indicate $P < 0.05$ after
629 Bonferroni correction for five tests). Numbers in parentheses are for the replication dataset, with
630 an expected genome-wide median mixed N:S of 0.33. *P*-values greater than 0.1 are denoted as
631 not significant (n.s.). Genes 1, 2, 4, and 5 are indicated with arrows on Figure 1C. Gene 3 is on
632 chromosome 2 (Supplementary Figure 3).

633 # Supplementary Figures



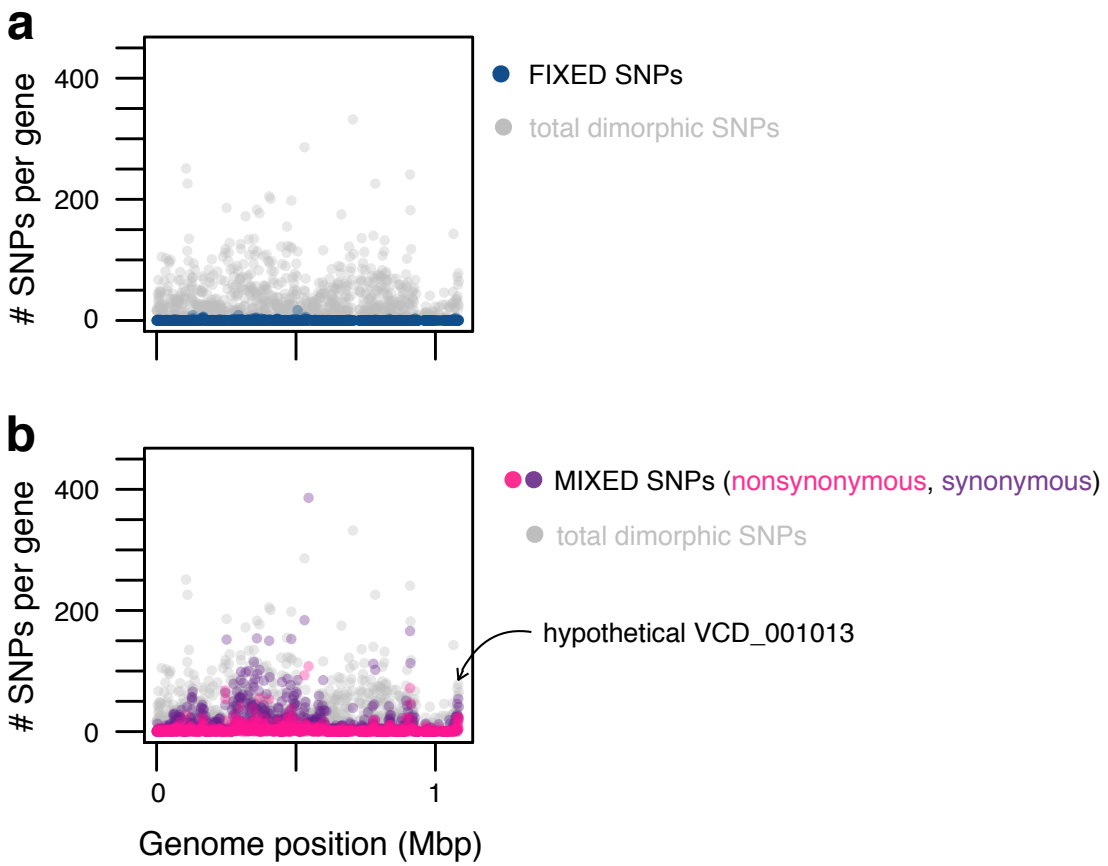
634

635 **Supplementary Figure 1. Splitstree showing relationship of environmental (EG) and pandemic**
636 **(PG) *V. cholerae* genomes. a,** The neighbour-net is based on dimorphic sites in the alignment of 22
637 genomes in the primary dataset, excluding sites with gaps. Only alignment blocks (locally colinear blocks
638 produced by mugsy) including all 22 genomes were included, yielding 126,099 dimorphic sites. **b,** The
639 neighbour-net based on 142,797 dimorphic sites in an alignment of 22 different genomes in the
640 replication dataset. The general star-like topology remains the same, with PGs clustering closely together
641 at the end of one long branch.
642



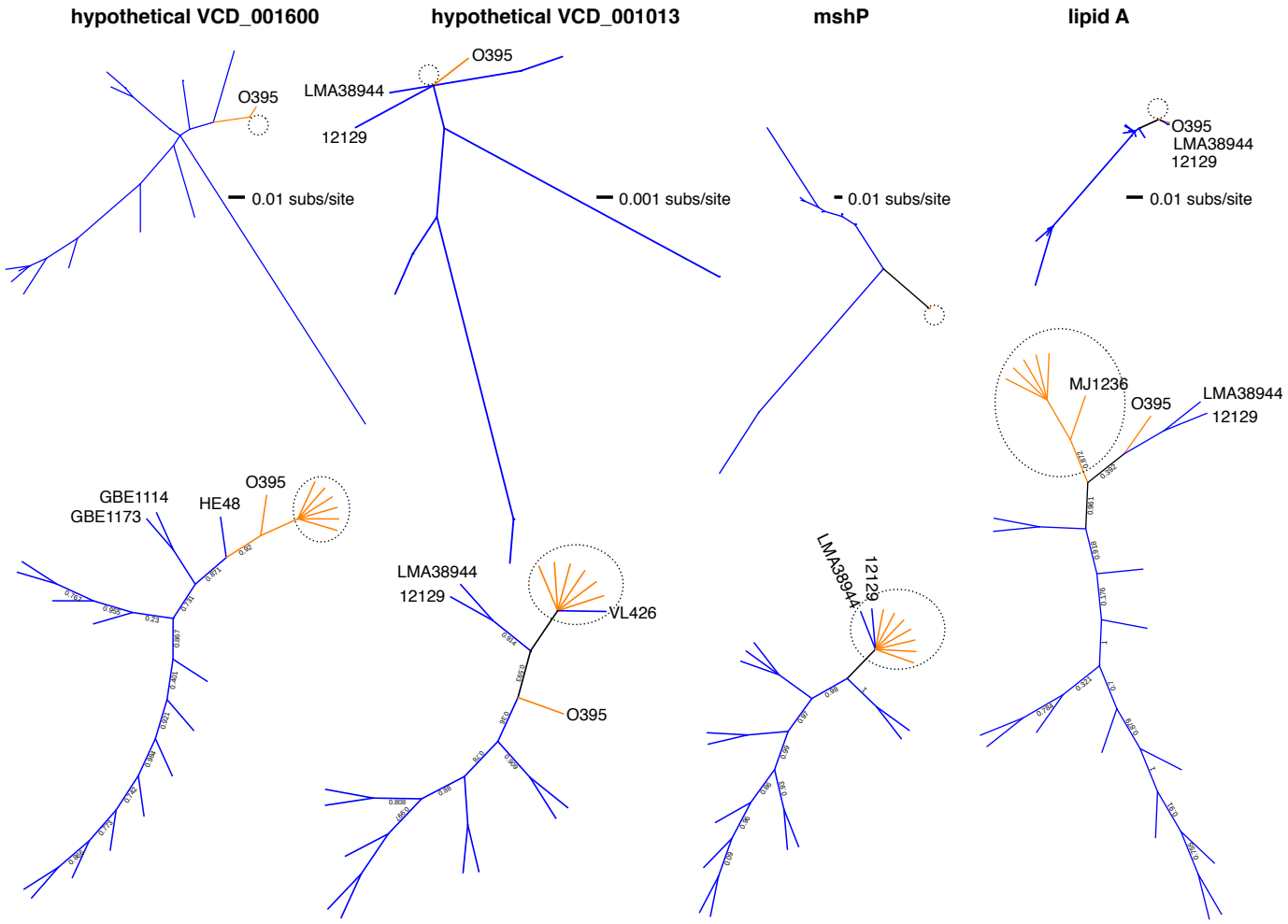
643

644 **Supplementary Figure 2. PG strains sample genes randomly from the environmental pool.** Gene
645 families from OrthoMCL were binned by their frequency in environmental genomes (EGs), ranging from
646 being present in zero to all 15 EGs in the primary dataset (x-axis). Within each of these bins, we
647 calculated the fraction of gene families fixed (*e.g.* present in all seven) phylocore genomes (PGs). For
648 example, the point in the top right includes 1817 gene families present in all 15 EGs, of which 1765 are
649 also present in all seven PGs, yielding a fraction of 0.97 fixed in PGs. The points fall closely along the x=y
650 line, with an outlier due to an excess of gene families (12 out of 35) fixed in PGs but absent in EGs
651 (present in 0/15 EGs). The observation that the fixation probability (y-axis) scales approximately linearly
652 with gene frequency in the environment (x-axis) suggests that PGs sample genes approximately
653 randomly from the environmental pool. The 12 gene families fixed in PGs but absent in EGs seem to
654 depart from the random expectation, suggesting a role for selection (Supplementary Table 2).
655

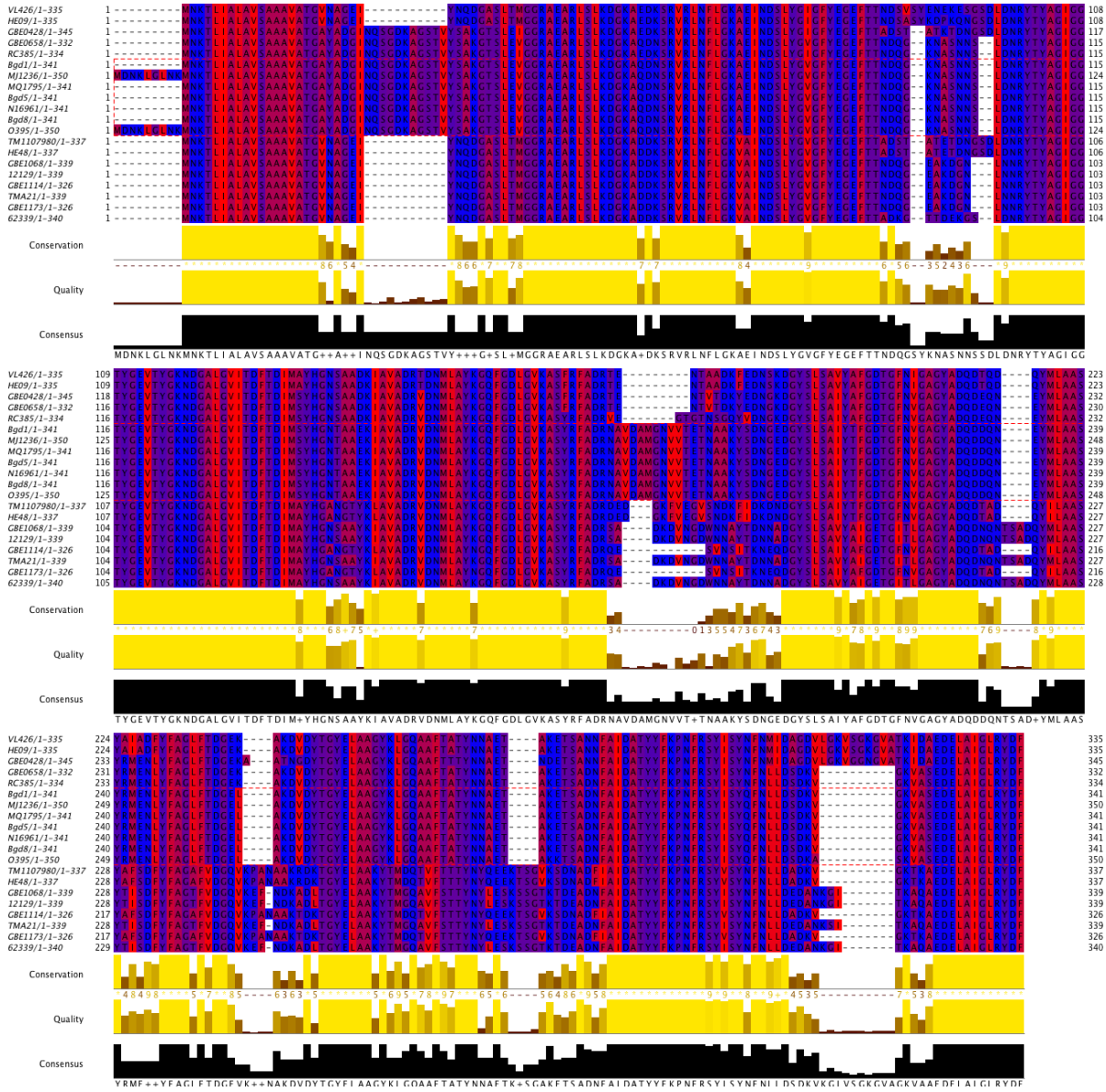


656

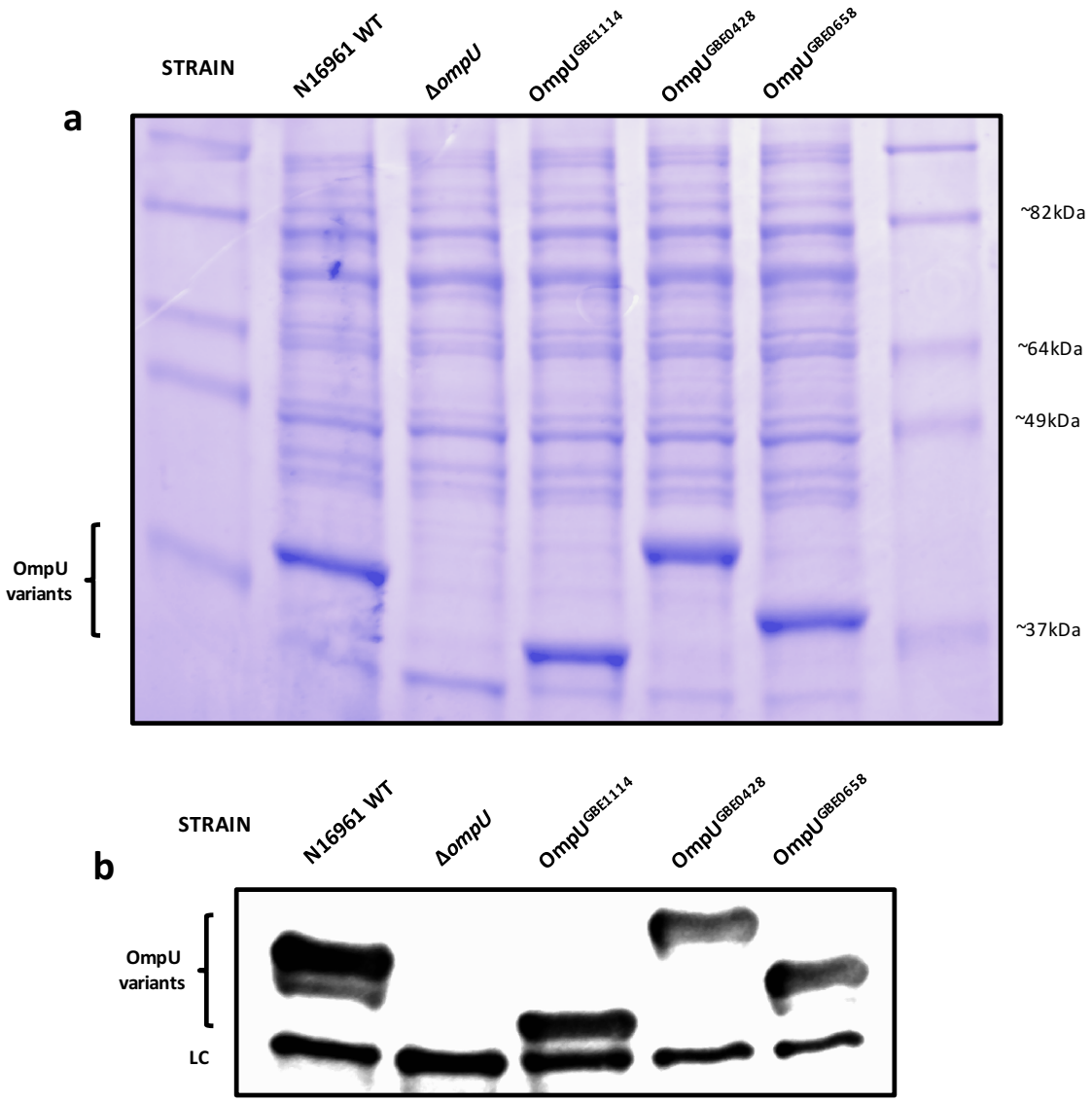
657 **Supplementary Figure 3. Distribution of fixed and mixed SNPs across chromosome 2. a,**
658 Distribution of fixed SNPs across chromosome 2 in the primary dataset. **b**, Distribution of mixed SNPs
659 across chromosome 2. Genome position is according to the MJ-1236 reference genome. See Fig. 1b and
660 1c in the main text for descriptions of fixed and mixed SNP sites. Arrow shows a candidate VAP (Table 1).
661



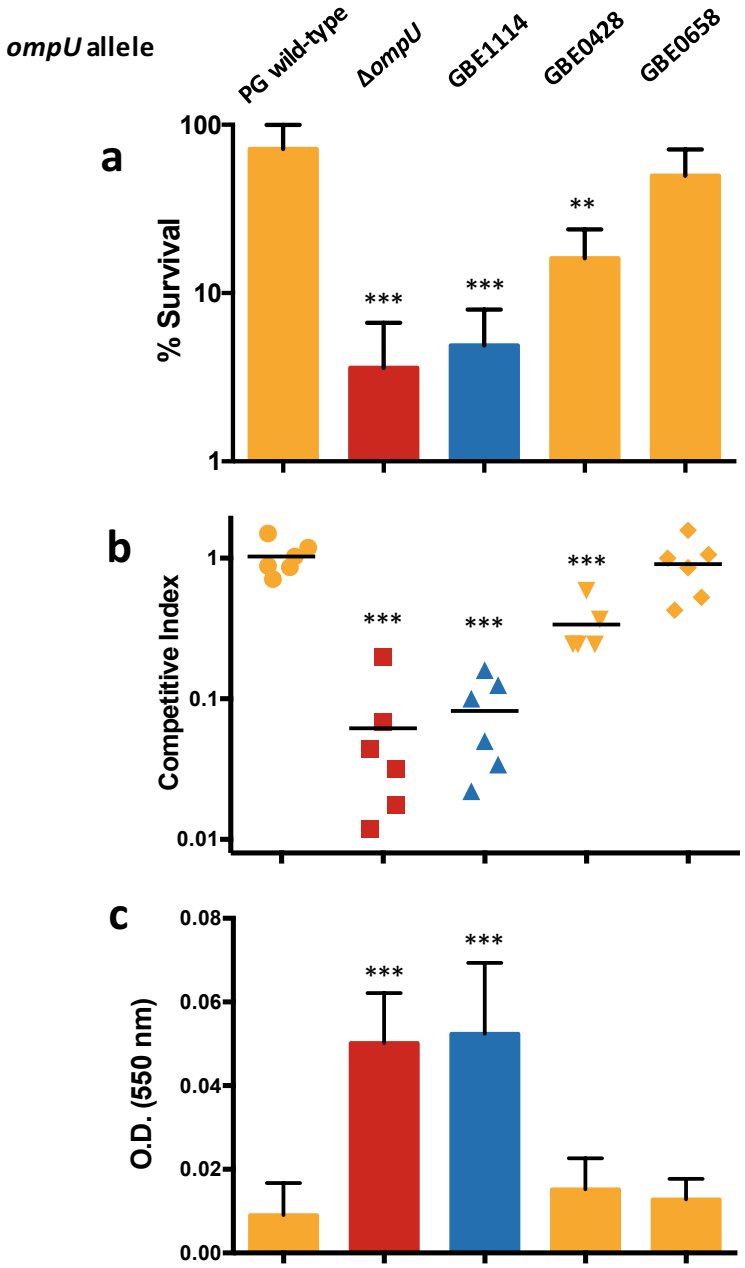
Supplementary Figure 4. Gene trees for additional candidate VAPs. Gene trees are shown for the four other candidate VAPs listed in Table 1, excluding *ompU* (Figure 1d). Top panels show branch lengths to scale; bottom panels show local support values and branches not to scale. Branches leading to EGs are blue; branches leading to PGs are in orange. The dotted circle indicates a clade containing most of the PGs (sometimes without O395 and sometimes with EGs, as indicated) which have identical sequences and are not visible in trees with branch lengths to scale (top). The lipid A core O-antigen ligase (VCD_001230) and *ompU* (Figure 1d) were grouped into orthologous gene families, which were aligned with Muscle and used to infer trees with FastTree. The other three genes were not grouped into gene families and were extracted from the mugsy alignment, realigned with Muscle and used to infer trees with FastTree (Methods).



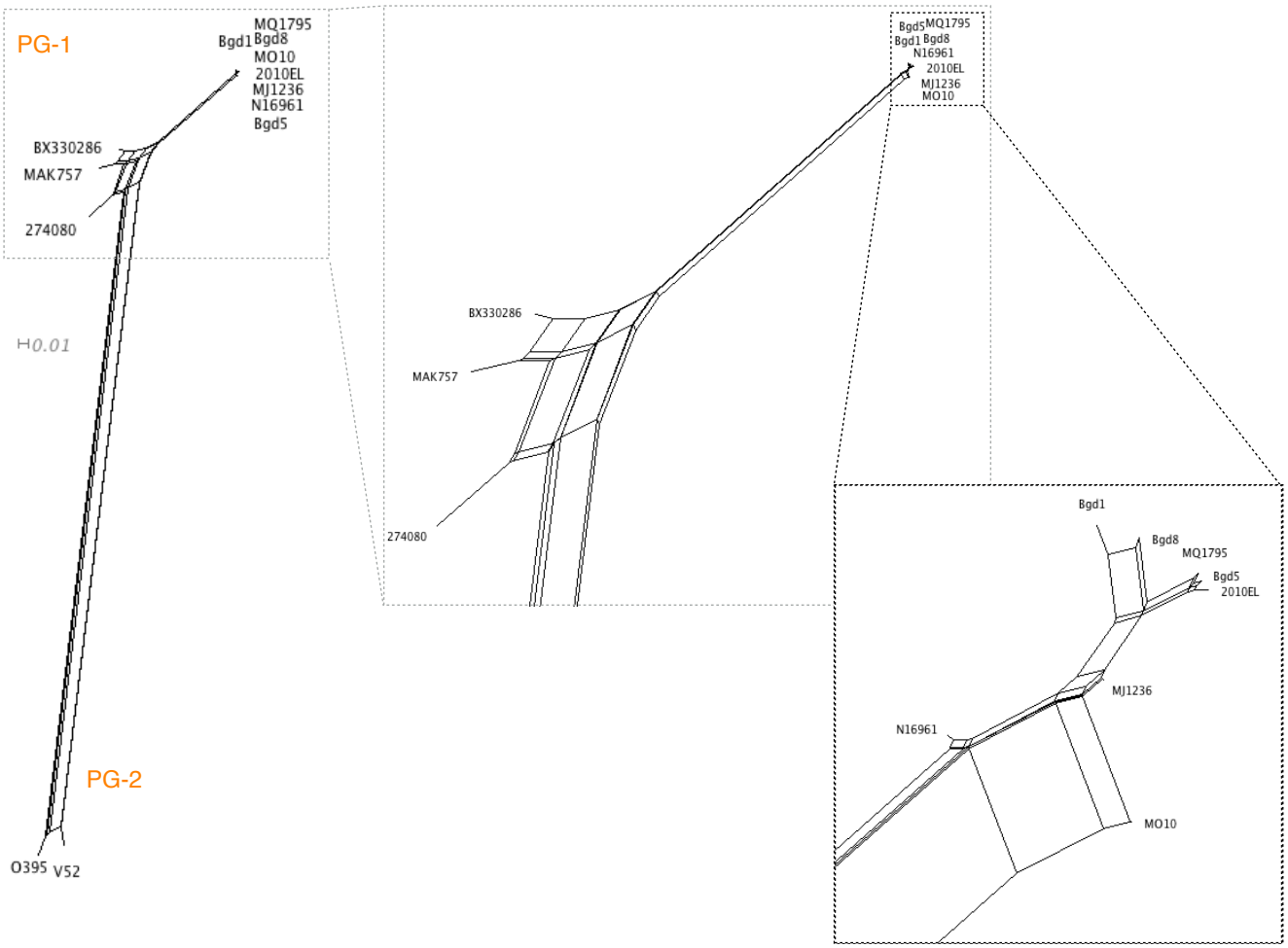
Supplementary Figure 5. Alignment of *ompU* sequences. The orthologous group containing OmpU was aligned with Muscle and visualized in Jalview. Amino acids are color-coded by hydrophobicity. The seven PG strains are shown in the dashed red box. Strain names are followed by the size of their ompU variant (in amino acids), following the dash character.



Supplementary Figure 6. OmpU production in clinical strains of *V. cholerae* encoding environmental alleles of *ompU*. Total protein lysates were run on a 16% Tris-glycine gel. **a**, gels were stained with Coomassie blue. WT, wild-type. **b**, gels were transferred to a nitrocellulose membrane. Protein samples were analyzed by immunoblotting with antibodies against OmpU. LC, loading control.

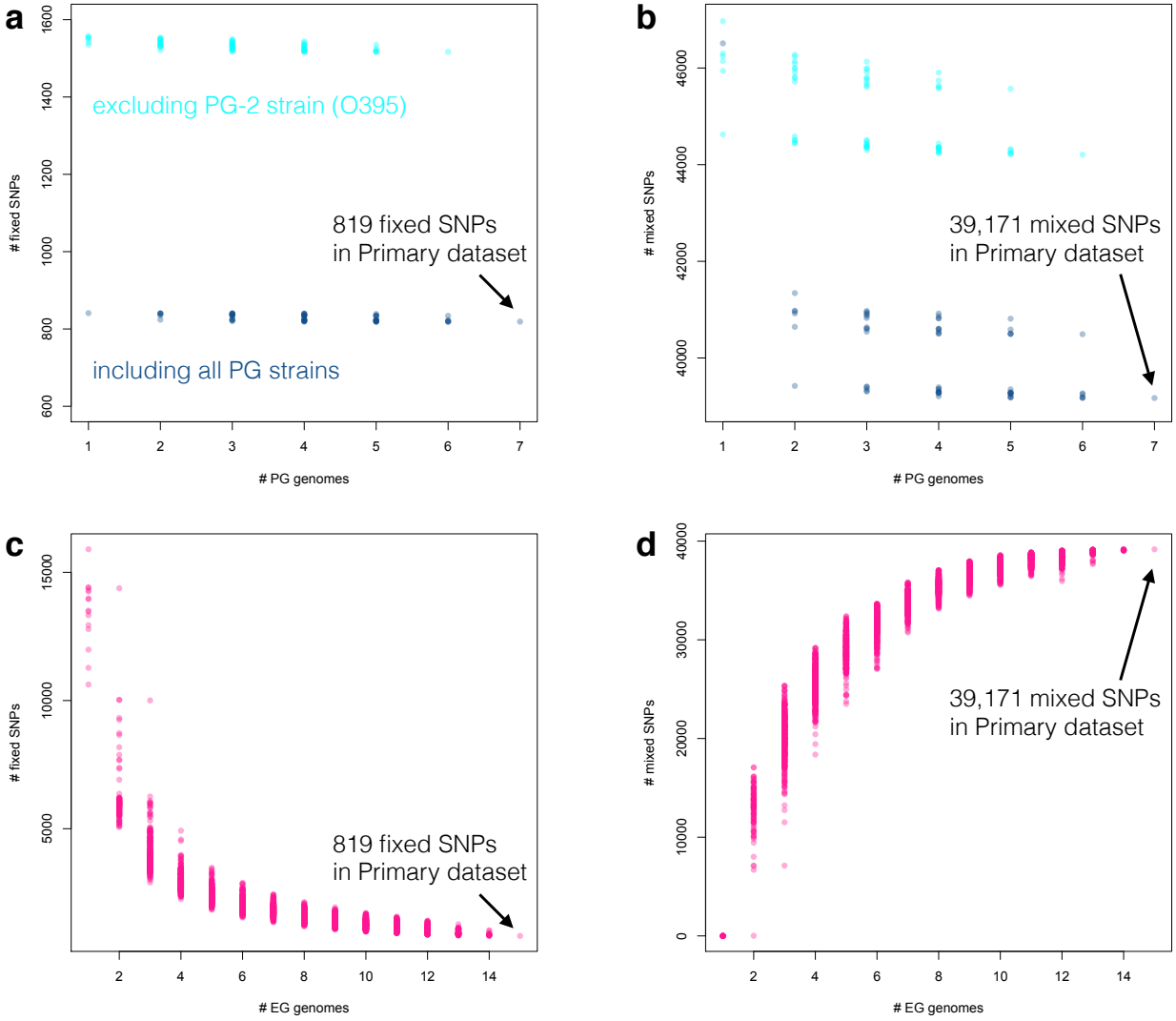


Supplementary Figure 7. Phenotypic characterization of *ompU* alleles in classical (O395) background. a, Survival of *ompU* mutants in the presence of bile (n=6). **b**, Colonization of the small intestine of *ompU* mutant strains (n≥5). **c**, Biofilm formation of *ompU* mutant strains on an abiotic surface (n=15). Yellow bars and symbols, PG-like allele; red bars and squares, Δ*ompU*; blue bars and triangles, EG-like allele. Center values represent the mean and error bars the standard deviation. Variance between the groups was similar. Statistical comparisons were made using student's *t*-test. $^{**}P < 0.01$, $^{***}P < 0.001$. Error bars represent standard deviation



Supplementary Figure 8. Splitstree showing relationship of all PG *V. cholerae* genomes used in this study. The neighbour-net is based on dimorphic sites in the alignment of all 13 PG genomes in the primary and replication datasets combined, excluding sites with gaps. Only alignment blocks (locally colinear blocks produced by mugsy) including all 13 genomes were included.

703



704
705
706 **Supplementary Figure 9. The number of identified fixed and mixed SNPs reach a plateau as**
707 **additional genomes are sampled.** We repeated our pipeline to identify fixed (**a, c**) and mixed (**b, d**)
708 SNPs as increasing numbers of PG genomes (**a, b**) or EG genomes (**c, d**) are sampled from the primary
709 dataset. Each point represents a different combination of the 7 PGs or 15 EGs used in the study.
710
711

712

713 # Supplementary tables

714 **Supplementary Table 1. *V. cholerae* genomes used in this study.**

715 **Primary dataset:**

Short name	PG/EG	Source	Accession number; source of strain
Bgd1	PG	O1 Ogawa El Tor, Bangladesh clinical	This study; Son et al.[60]
Bgd5	PG	O1 Inaba El Tor, Bangladesh clinical	This study; Son et al.[60]
Bgd8	PG	O1 Ogawa El Tor, Bangladesh clinical	This study; Son et al.[60]
N16961	PG	O1 Inaba El Tor, Bangladesh clinical	AE003852-3
MJ1236	PG	O1 Inaba El Tor, Bangladesh clinical	CP001485-6
MQ1795	PG	O1 Inaba El Tor, Bangladesh clinical	This study; Nair et al.[61]
O395	PG	O1 Ogawa Classical, India clinical	CP000626/CP000627
GBE0428	EG	non-O1, USA oyster	This study; Schuster et al.[32]
GBE0658	EG	non-O1, USA water	This study; Schuster et al.[32]
GBE1068	EG	non-O1, USA oyster	This study; Schuster et al.[32]
GBE1114	EG	non-O1, USA water	This study; Schuster et al.[32]
GBE1173	EG	non-O1, USA sediment	This study; Schuster et al.[32]
12129	EG	O1 Inaba El Tor, Australia water	ACFQ00000000
LMA38944	EG	O1, Brazil water	CP002555-6
CT536993	EG	unknown serogroup, Brazil sewage	ADAL01000000
RC385	EG	O135, USA plankton	AAKH02000000
VL426	EG	non-O1/O139 Albensis, UK water	ACHV00000000

HE09	EG	unknown serogroup, Haiti water	AFOP01000000
TM1107980	EG	O1 Ogawa El Tor, Brazil sewage	ACHW0000000
HE48	EG	unknown serogroup, Haiti water	AFOR01000000
TMA21	EG	non-O1/O139, Brazil water	ACHY00000000
62339	EG	non-O1/O139, Bangladesh water	AAWG00000000

716

717 **Replication dataset:**

Short name	PG/EG	Source	Accession number; source of strain
2010EL	PG	Artibonite, Haiti clinical	CP003069-70
274080	PG	Gulf Coast, USA water	AAUT00000000
BX330286	PG	Australia water	ACIA00000000
MAK757	PG	Sulawesi, Indonesia clinical	AAUS00000000
MO10	PG	Madras, India clinical	AAKF00000000
V52	PG	Sudan clinical	AAKJ00000000
MJ1236	PG	O1 Inaba El Tor, Bangladesh clinical	CP001485-6
1587	EG	Lima, Peru clinical	AAUR00000000
AM19226	EG	Bangladesh clinical	AATY00000000
MZO3	EG	Bangladesh clinical	AAUU00000000
MZO2	EG	Bangladesh clinical	AAWF00000000
V51	EG	USA clinical	AAKI00000000
YB1A01	EG	Oyster pond, MA, USA water	LBCL00000000
YB1G06	EG	Oyster pond, MA, USA water	LBFV00000000
YB2G01	EG	Oyster pond, MA, USA water	LBFY00000000
YB3B05	EG	Oyster pond, MA, USA water	LBGB00000000
YB4B03	EG	Oyster pond, MA, USA water	LBGD00000000

YB4C07	EG	Oyster pond, MA, USA water	LBGE00000000
YB4G05	EG	Oyster pond, MA, USA water	LBGG00000000
YB4H02	EG	Oyster pond, MA, USA water	LBGI00000000
YB5A06	EG	Oyster pond, MA, USA water	LBGJ00000000
YB7A09	EG	Oyster pond, MA, USA water	LBGM00000000

718

719 **Supplementary Table 2. Summary statistics of SNPs in both datasets.** fixN = fixed
720 nonsynonymous SNPs; fixS = fixed synonymous SNPs; mixN = mixed nonsynonymous SNPs;
721 mixS = mixed synonymous SNPs; *dN* = number of nonsynonymous SNPs per nonsynonymous
722 site; *dS* = number of synonymous SNPs per synonymous site; sd = standard deviation.
723

	Dataset	
	Primary	**Replication**
total SNPs	136,160	146,309
total fixed	819	2,772
total mixed	39,171	86,370
total SNPs in genes	121,254	130,035
total fixN in genes	210	597
total fixS in genes	504	1,865
total mixN in genes	6,982	14,418
total mixS in genes	27,366	62,509
proportion fixed SNPs in genes	0.87	0.89
proportion mixed SNPs in genes	0.88	0.89
mean mixed SNPs per gene	10.3	22.6
median mixed SNPs per gene	3	12
median mixN per gene	0	2
sd mixN per gene	5.85	7.92
median mixN + 2sd	11.70	17.85
median mixN/mixS per gene	0.50	0.33
sd mixN/mixS per gene	0.64	0.66
median mixN/mixS + 2sd	1.78	1.65
median fixN per gene	0	0
sd fixN per gene	0.34	0.75
median fixN + 2sd	0.68	1.50
median fixN/fixS per gene	1	1
sd fixN/fixS per gene	0.24	0.35
median fixN/fixS + 2sd	1.48	1.70
median *dN/dS* per gene	0.17	0.12
sd *dN/dS* per gene	0.22	0.22
median *dN/dS* + 2sd	0.60	0.56

724

725 **Supplementary Table 3. Gene families present in all seven phylocore genomes (PGs) and**
726 **absent in all 15 environmental genomes (EGs) in the primary dataset.**

Locus	Gene ID	Annotation	Presence in replication dataset
1	VCA0790	Possible integrase (RefSeq)	Present in all 7 PGs + 2 EGs
	VCA0793	Phage regulatory protein, Rha-like (IPR019104)	Not assigned to gene family
	VCA0795	Site-specific recombinase PinR (COG1961)	Not assigned to gene family
2	VCA1042	Integral membrane protein CcmA (COG1664)	Present in all 7 PGs + 1 EG
	VCA1043	tagE protein (RefSeq)	Not assigned to gene family
3 (CTX)	VC1462	CTX phage RstB (IPR010008)	Present in 6 PGs + 2 EGs
4 (VPI-1)	VC0824	Peroxiredoxin Tpx (COG2077)	Present in all 7 PGs + 1 EG
	VC0831	Toxin-coregulated pilus biosynthesis protein TcpC	Present in all 7 PGs + 4 EGs
	VC0835	Toxin-coregulated pilus biosynthesis protein T	Present in all 7 PGs + 3 EGs
	VC0836	Toxin-coregulated pilus biosynthesis protein E	Present in all 7 PGs + 4 EGs
	VC0838	TCP virulence regulatory protein TcpN	Present in all 7 PGs + 4 EGs
5 (VSP-2)	VCA0483	Hypothetical protein (RefSeq)	Not assigned to gene family

727 Gene IDs are from the *V. cholerae* O1 biovar El Tor N16961 reference genome. Gene order and
728 annotations are nearly identical in *V. cholerae* MJ-1236. The 5 loci are defined as clusters of
729 adjacent or nearly adjacent genes (0-6 genes apart in either reference genome)

730 **Supplementary Table 4. Genome-wide McDonald-Kreitman test between PGs and EGs.**

731 Primary dataset:

Polymorphism from:	FN	FS	FN/FS	PN	PS	PN/PS	Obs. FI	Exp. FI	*P*(obs > exp)
PG	210	504	0.42	1022	4433	0.23	1.81	1.76	0.12
EG				6982	27366	0.26	1.63	1.06	<0.001

732
733 Replication dataset:

Polymorphism from:	FN	FS	FN/FS	PN	PS	PN/PS	Obs. FI	Exp. FI	*P*(obs > exp)
PG	597	1865	0.32	2694	14342	0.19	1.70	1.67	0.068
EG				14418	62509	0.23	1.39	0.98	<0.001

734
735 Total genomewide counts of classes of mutations: FN = fixed nonsynonymous, FS = fixed
736 synonymous, PN = polymorphic nonsynonymous, PS = polymorphic synonymous. Fixed
737 indicates that one allele is present in all PGs and a different allele in all EGs. Fixation Index (FI)
738 = (FN/FS)/(PN/PS). *P*(obs > exp) is the *P*-value computed from permutations as the fraction of
739 permutations in which the FI was greater than observed in the observed data.

740 **Supplementary Table 5. Characteristics of three additional predicted VAPs with an**
741 **excess of nonsynonymous mixed SNPs in the replication dataset.**

	Gene ID (VCD #)	Annotation	Gene length (bp)	Total # SNPs	Mixed *dN*	Mixed *dS*	Mixed *dN/dS*	Mixed NS	Mixed S	*P*
1	001506	hypothetical	331	21 (7)	0.122 (0)	0.034 (0)	3.58 (n.a.)	20 (0)	1 (0)	4.11e-9 (n.s.)
2	003509	hypothetical; possible pseudogene	684	77 (35)	0.063 (0.010)	0.110 (0.006)	0.57 (1.70)	31 (4)	18 (0)	1.74e-5 (0.0625)
3	000213	hypothetical	3261	149 (118)	0.008 (0.022)	0.010 (0.080)	0.83 (0.28)	19 (54)	7 (64)	3.99e-5 (n.s.)

742 NS=nonsynonymous; S=synonymous. *dN* = number of nonsynonymous SNPs per
743 nonsynonymous site; *dS* = number of synonymous SNPs per synonymous site. The genes
744 listed have mixed NS, mixed NS:S, and mixed *dN/dS* each over two standard deviations above
745 their respective genome-wide medians in the replication dataset. A one-sided binomial test
746 determined if the mixed NS:S ratio was greater than the expected genome-wide median value
747 of 0.33 per gene. Numbers in parentheses are for the primary dataset, with an expected
748 genome-wide median mixed N:S of 0.5. *P*-values greater than 0.1 are denoted as not significant
749 (n.s.).
750

751 **Supplementary Table 6. Amino acid differences between ompU alleles used in this study.**
752 Single amino acid substitutions are shown above the diagonal; the number of indels and their
753 total sizes are shown below. Differences are based on the ompU protein alignment
754 (Supplementary Figure 5). All PG1 genomes have identical ompU protein sequences.
755

	PG1	**GBE0658**	**GBE0428**	**GBE1114**
PG1	0	15aa	45aa	80aa
GBE0658	1 indel (9aa)	0	30aa	78aa
GBE0428	4 indels (22aa)	3 indels (13aa)	0	86aa
GBE1114	5 indels (31aa)	5 indels (22aa)	6 indels (33aa)	0

756
757

758 **Supplementary Table 7. Genome assembly statistics.** The mean number of contigs and N50
759 are shown, broken down by dataset, and within dataset by PGs (7 genomes) and EGs (15
760 genomes). Median values are shown in parentheses.
761

Dataset	# contigs	N50
Primary		
PGs	39.2 (53)	1,431,696 (246,647)
EGs	90.9 (58)	648,274 (255,577)
Total	74.5 (55.5)	897,545 (251,112)
Replication		
PGs	45.7 (43)	1,144,254 (227,382)
EGs	66.6 (57)	294,710 (285,424)
Total	60.0 (56.5)	565,019 (285,397)

762

763 **Supplementary Table 8. Genome-wide alignment statistics.** The number and total length of
764 locally colinear blocks are reported for mugsy alignments of the primary dataset, the replication
765 dataset, and an alignment of all 13 PG genomes from both datasets. The N50 was calculated
766 based on the mean size of each LCB, because LCBs can contain gaps. Core LCBs were
767 defined as those containing a sequence from each of the genomes in the dataset.
768

Dataset	# LCBs	Total LCBs length (bp) [pangenome]	Total LCBs N50 (bp)	# core LCBs	Core LCBs length (bp)	Core LCBs N50 (bp)
Primary (22 genomes)	11,015	6,952,342	5,900	627	3,302,728	8,566
Replication (22 genomes)	8,742	6,819,882	13,638	264	3,572,057	26,306
All PGs (13 genomes)	4,336	4,598,913	33,850	209	3,755,985	38,635

769