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: Salvador.Almagro-Moreno@Dartmouth.edu

[#]These authors contributed equally

Abbreviations: PG, Phylocore genome; EG, Environmental genome; SNP, Single nucleotide polymorphism; HGT; Horizontal gene transfer; VAP, Virulence adaptive polymorphism; NS, Nonsynonymous; S, Synonymous; FI, Fixation Index

Abstract

Some microbes can transition from an environmental lifestyle to a pathogenic one[1-3]. This ecological switch typically occurs through the acquisition of horizontally acquired virulence genes[4-7]. 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 one of them reduced the ability of clinical strains to colonize a mammalian host, whereas two other alleles conferred efficient 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.

31 **Main text**

32 Numerous bacterial pathogens have emerged from environmental populations[1-3]. These virulent
33 clones evolve through the acquisition of toxins and host colonization factors[4-7]. Given that the
34 genes encoding these factors can often spread widely by horizontal gene transfer (HGT), it is
35 surprising that only a limited number of pathogenic clones have emerged from any given
36 bacterial species. As a model of how non-pathogenic gene pools give rise to pandemic clones,
37 we used *Vibrio cholerae*, a genetically diverse group of aquatic bacteria that include a confined
38 phylogenetic group, the phylocore genome (PG), that can cause the severe diarrheal disease
39 cholera in humans[8-11]. Virulence in *V. cholerae* PG is mainly determined by two virulence
40 factors: the cholera toxin (CT) and the toxin-coregulated pilus (TCP), which are encoded within
41 horizontally acquired genetic elements, the CTXΦ phage and the Vibrio Pathogenicity Island-1
42 (VPI-1) respectively[12-16]. Both gene clusters are always found in the PG group, however, they
43 are also encoded in some environmental populations of *V. cholerae*[8,9,17-24]. The PG group
44 evolved from a common ancestor and underwent clonal expansion, thus, harboring relatively
45 little genetic diversity other than sporadic swapping of CTXΦ and O-antigen alleles[8,9]. Seven
46 pandemics of cholera have been recorded to date, all caused by the PG group. The first six
47 pandemics were caused by strains of the classical biotype[10,11]; the current and seventh
48 pandemic is caused by strains of the El Tor biotype, and has spread across the globe in several
49 waves of transmission[25].

50 To investigate the evolutionary origins of pandemic clones and the potential for their
51 reemergence, we analyzed 43 *V. cholerae* genomes sequenced from clinical and environmental
52 samples (Methods; Supplementary Table 1). These were divided into a primary dataset of 22
53 genomes and a replication dataset of 22 genomes (containing additional single nucleotide
54 polymorphisms; SNPs; Supplementary Table 2), with one reference genome in common, MJ-
55 1236. In the primary dataset, we chose 7 PG genomes, including both classical and El Tor, to
56 represent the genetic diversity of the pandemic group. We compared these with 15 non-clinical

57 environmental genomes (EGs): 10 EGs from worldwide samples to include global diversity, and
58 five sympatric isolates from the Great Bay Estuary (GBE) in New Hampshire, USA, a region with
59 no recent history of cholera outbreaks[26].

60 Consistent with the results of previous studies[8,25-27], PG genomes form a distinct
61 monophyletic group compared to EGs, based on the aligned core genome (Fig. 1a). Other than
62 the PG group, there is little phylogenetic structure and the tree is star-like, with a similar pattern
63 being observed in both the primary and replication datasets (Supplementary Fig. 1).
64 Reticulations in the phylogenetic network indicate homoplasic mutation and/or recombination,
65 consistent with a large, genetically diverse and recombining *V. cholerae* population
66 (Supplementary Fig. 1). In terms of gene content, PGs appear to sample genes at random from
67 the environmental gene pool (Supplementary Fig. 2). In the primary dataset, there are only 12
68 genes, located in 5 clusters in the genome, present in all 7 PGs and absent in all 15 EGs; these
69 include the major virulence gene clusters CTXΦ and VPI-1 (Supplementary Table 3). However,
70 most of the 12 gene families were represented in both PGs and EGs in the replication dataset.
71 Overall, these results indicate that the PG-specific gene pool is small, and there is extensive
72 gene sharing between EGs and PGs.

73 Given a frequently recombining *V. cholerae* population[27,28], the small number of PG-
74 specific genes, and the mobile nature of its virulence factors, it remains puzzling why the ability
75 to cause pandemic cholera is limited to the PG group. Two possible scenarios that could explain
76 this confined distribution are: 1) The acquisition of virulence genes could have triggered a clonal
77 expansion of the first genomic background that acquired them (PG). Other clones that
78 subsequently acquired the virulence genes failed to emerge as global pandemics because that
79 niche had already been filled by PG; 2) The environmental ancestor of PG had a genomic
80 background that rendered them susceptible to becoming virulent by encoding a set of allelic
81 variations – which we term virulence adaptive polymorphisms (VAPs) – endowing them with
82 properties that enhanced their pathogenic potential.

83 To distinguish between the two scenarios, we first identified SNPs in the aligned core of
84 the 22 primary dataset genomes, with one allele fixed in all PGs and a different allele fixed in all
85 EGs. There are 819 such "fixed" SNPs, of which 714 (87%) fall within genes distributed across
86 the genome (Fig. 1b; Supplementary Fig. 3). Under the first scenario, these SNPs are
87 selectively neutral markers with no impact on virulence. Under the second scenario, at least
88 some of these SNPs would be under selection for increased virulence or host colonization.
89 Using the McDonald-Kreitman test[29], we found evidence for genome-wide positive selection
90 during the divergence of PG from EGs due to an excess of nonsynonymous changes in both
91 primary and replication datasets, supporting the second scenario and suggesting the existence
92 of VAPs in natural *V. cholerae* populations (Supplementary Table 4). No individual gene showed
93 evidence for selection after correcting for multiple tests (Methods), rendering it difficult to identify
94 VAPs at fixed SNP sites. However, fixed SNPs constitute only a modest fraction of possible
95 SNP patterns (Supplementary Table 2), and VAPs could also exist at other SNP sites. In
96 particular, SNPs with one allele fixed in PGs but with two alleles segregating in EGs could be
97 informative both about pathogen emergence, and about the potential for reemergence (*e.g.* the
98 extent of virulence adaptive alleles circulating in the environmental gene pool).
99 To identify specific genomic variants putatively encoded by the environmental ancestor
100 of PG but still segregating in contemporary natural populations of *V. cholerae*, we searched for
101 PG-like alleles present in EGs. Specifically, we defined a "mixed" SNP pattern at sites that were
102 fixed in PGs but polymorphic in EGs. We only considered dimorphic SNP sites, meaning that at
103 least one EG had the same allele fixed in the PGs (Fig. 1c). We identified 39,171 mixed SNPs in
104 the primary dataset, of which 34,348 (88%) fall within genes. Most genes contain few mixed
105 SNPs (median of 3) but some genes contain dense clusters, resulting in a mean of 10.3 mixed
106 SNPs per gene. The replication dataset contained greater genetic diversity, but showed the
107 same pattern (Supplementary Table 2). Clusters of genes containing many mixed SNPs are
108 visible when plotted across the genome (Fig. 1c). Some of these clusters are known

109 polymorphic regions of the genome, such as the loci encoding the O-antigen and the flagellum.
110 Many of these clusters could be mutation hotspots, containing mutations not directly relevant to
111 virulence adaptation. However, clusters of mixed SNPs do not visibly overlap with clusters of
112 overall polymorphism (Fig. 1c), indicating that accumulation of mixed SNPs cannot be explained
113 only by mutation hotspots.

114 To formally exclude mutation hotspots and focus on clusters of mixed SNPs likely to be
115 under natural selection, we considered only genes with an excess of nonsynonymous (NS)
116 mixed SNPs relative to synonymous (S) mixed SNPs, which control for the baseline mutation
117 rate. The mixed SNP pattern necessarily groups PG-like EGs away from the other
118 environmental strains and clusters them with the PGs. We reasoned that genes with an
119 elevated NS:S ratio at mixed SNP sites were more likely to show phenotypic variations and
120 have evolved under positive selection, possibly underlying host adaptations of the PG ancestor.
121 Using a threshold of ≥12 NS mixed SNPs per gene and mixed NS:S>1.78 (respectively two
122 standard deviations above the genome-wide medians) we identified five genes as candidate
123 VAPs in the primary dataset, three of which survived multiple hypothesis correction, and two of
124 which (*ompU* and hypothetical gene VCD_001600) were also found in the replication dataset
125 (Table 1). All five genes had NS:S greater than two (the neutral expectation, based on
126 approximately two nonsynonymous sites and one synonymous site per codon), suggesting
127 positive or diversifying selection rather than relaxed purifying selection. In contrast to the star-
128 like genome-wide phylogeny (Fig. 1a), each of these five gene trees support one or more EGs
129 grouping with PGs (Fig. 1d and Supplementary Fig. 4). Three of the gene trees, including the
130 lipid A core O-antigen ligase, support EGs LMA3894-4 and 12129 branching with PGs
131 (Supplementary Table 1 and Supplementary Fig. 4), consistent with these strains being O1-
132 like[8,30]. The O-antigen ligase gene was not identified as a VAP in the replication dataset, which
133 is to be expected given the absence of O1-like EGs in that dataset[27] (Table 1). Three additional

134 VAPs – all hypothetical proteins – were identified in the replication dataset (Supplementary
135 Table 5).

136 Among the candidate VAPs, the gene with the most significant excess of mixed NS
137 SNPs in both datasets is *ompU* (Table 1). OmpU is an outer membrane porin that has been
138 shown to play numerous roles in the intestinal colonization of *V. cholerae*, making it a
139 compelling candidate for phenotypic characterization[31-35]. We observed that environmental
140 strains RC385, GBE0658 and GBE0428 cluster near PGs in the *ompU* gene tree, resulting in
141 mixed SNP patterns, and share an 11 amino acid N-terminus insertion with PGs
142 (Supplementary Fig. 5), whereas GBE1114 is part of a separate clade with the rest of the EGs
143 (Fig. 1d). We hypothesized that the PG-like environmental alleles might confer properties
144 conducive to virulence. To test this, we constructed three different mutant strains each encoding
145 one of three environmental alleles of *ompU* into the background of N16961, a clinical strain from
146 the current pandemic (Fig. 2). OmpU was detected on a protein gel stained with Coomassie
147 blue in all the constructed strains, indicating that all the strains effectively produce the
148 environmental versions of the protein (Fig. 2a).

149 We performed three sets of experiments to compare the phenotypes conferred by EG
150 and PG-like alleles of *ompU*. First, we determined the survival of these strains in the presence
151 of 0.4% bile, as it has been previously shown that OmpU confers resistance to this antimicrobial
152 compound[33]. The mutant strain encoding the *ompU* allele from the environmental strain
153 GBE1114 ($OmpU^{GBE1114}$) had diminished bile tolerance and its survival in the presence of bile is
154 similar to that of a deletion mutant (Fig 2b). In contrast, the mutants encoding the PG-like
155 environmental alleles ($OmpU^{GBE0658}$ and $OmpU^{GBE0428}$) show similar survival in the presence of
156 bile as wild-type (Fig. 2b). These experiments indicate that some environmental alleles of *ompU*
157 confer properties beneficial for virulence. Second, we tested the survival of the mutants in the
158 presence of polymyxin B, as OmpU also confers resistance against this antibiotic[31]. The ability
159 to tolerate the antimicrobial effects of polymyxin B appears to be independent of which *ompU*

160 allele is encoded by *V. cholerae*, as the three strains encoding environmental alleles of *ompU*
161 had a similar survival rate (Fig. 2c). The three mutant strains had a consistent decrease in
162 survival when compared to wild-type; however, the difference was not statistically significant
163 (Fig. 2c). Third, we determined the intestinal colonization of the *ompU* mutants by performing
164 competition assays using the infant mouse model of human infection. We found that
165 $OmpU^{GBE1114}$ had a colonization defect similar to Δ*ompU* whereas $OmpU^{GBE0658}$ was able to
166 colonize similarly to wild-type (Fig. 2d). These results indicate that certain naturally occurring
167 environmental alleles of *ompU* confer properties that provide an advantage to *V. cholerae* prior
168 to host colonization.

169 The $ompU^{GBE1114}$ allele does not appear to be adaptive for intestinal colonization;
170 however, its presence in several environmental isolates of *V. cholerae* prompted us to
171 investigate its possible role in the environment (Fig. 2e). *V. cholerae* forms biofilms on the
172 surface of biotic and abiotic environmental surfaces[24,36-39]. Biofilm formation inside the host is
173 thought to be detrimental for successful intestinal colonization[24,35-37,39,40]. Strains with deletions
174 in *ompU* have been shown to form a more robust biofilm on abiotic surfaces[41]. We found that
175 $OmpU^{GBE1114}$ has higher biofilm formation than wild-type, similar to the Δ*ompU* strain (Fig. 2e).
176 Both $OmpU^{GBE0658}$ and $OmpU^{GBE0428}$ formed biofilm similar to the wild-type strain (Fig. 2e). It
177 therefore appears that there is an evolutionary trade-off between encoding the PG-like or EG-
178 like alleles of VAPs, as they seem to confer mutually exclusive traits. This suggests that
179 environmental strains can be divided into subgroups which, due to their contrasting lifestyles,
180 differ in their potential to become pathogenic.

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

185 (Supplementary Table 5), suggesting the potential for other VAPs to be identified with further
186 sampling of genetically diverse environmental genomes.

187 Our data show that the *ompU* allele from some environmental strains, such as
188 GBE0658, confers properties that allow for host colonization equally as efficient as alleles from
189 clinical strains (Fig. 2). This leads to a natural question: Why have environmental strains with
190 PG-like alleles not emerged as pandemic cholera strains? One immediate answer is because
191 they might not encode the key virulence factors CT and TCP. However, CTXΦ and VPI-1 have
192 been found in environmental *V. cholerae* strains that do not cause disease in humans[8,9,17-24]. It
193 therefore appears that a variety of virulence adaptive alleles are circulating in the environment,
194 but only the PG group encodes the optimal combination of VAPs that allowed for emergent
195 virulent properties and pandemic potential (Fig. 3). We propose a conceptual model in which
196 VAPs circulate in a diverse, recombining environmental gene pool, being maintained in the
197 population through various biotic and abiotic selective pressures (Fig. 3a). A new ecological
198 opportunity occurs, such as human consumption of brackish water or transient colonization of
199 other animal hosts, which leads to the proliferation and gradual enrichment in the population of
200 clones encoding a mosaic of VAPs (Fig. 3b). Finally, a genome encoding a critical combination
201 of VAPs acquires key virulence factors allowing it to emerge as a virulent, potentially pandemic
202 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 O395 background conferred

211 efficient host colonization (Supplementary Fig. 6), which is consistent with an *ompU* VAP having
212 played a role in the emergence of pathogenic *V. cholerae* prior to the acquisition of key
213 virulence factors.

214 Our model further posits 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[42]. When adaptation to the new niche
219 depends on few loci, it is more likely for recombination to assemble the right combination of
220 alleles in the same genome before. As more loci are involved, it becomes less likely to achieve
221 the optimal combination. Assuming the *V. cholerae* population undergoes approximately 100
222 recombination events per locus per generation[28], equivalent to a recombination rate of 10^{-4} in
223 the modelled population of size 10^6, an optimal combination of alleles at five loci could
224 conceivably evolve, but seven loci is very unlikely[42]. Therefore, if virulence depended on five loci
225 in the *V. cholerae* genome, the optimal combination of alleles would be expected to appear
226 repeatedly in nature. Given suitable ecological opportunities, it is then plausible that pandemic
227 *V. cholerae* could emerge multiple times, originating from outside the PG group. However, the
228 number of loci that are sufficient for the emergence of a virulent strain remains unknown and if it
229 was much greater than five, excluding horizontally acquired elements, pathogen emergence
230 would be naturally limited. Coincidentally, we identified between five and eight candidate VAPs
231 in our two datasets. However, these passed stringent filters and we suspect there might exist
232 additional VAPs in the genome, identifiable by further sampling and experimentation. We also
233 note that LMA3894-4, a PG-like environmental strain containing PG-like alleles at three out of
234 five candidate VAP loci (Supplementary Fig. 4), does not contain a predicted *ompU* ortholog,
235 suggesting a colonization-impaired phenotype. Similarly, strain 12129 also contains PG-like
236 alleles at these VAP loci, but has an EG-like allele of *ompU* (Fig 1D). Therefore, no genome

237 other than PGs contains a clinical-like combination of VAPs, due either to recombination or
238 selection limitation.

239 Here we have described a framework for identifying loci that are present in a natural
240 population and confer properties beneficial for virulence prior to acquisition of essential
241 virulence genes and host colonization. This framework could be applied to other bacterial
242 pathogens that emerge as clonal offshoots from non-virulent relatives, including *Yersinia,*
243 *Salmonella,* and *Escherichia*, among others[1,6,7]. Pathogens that emerged through clonal
244 expansion limit our ability to dissect the genetic basis of their pathogenicity, because bacterial
245 genome-wide association studies lack power when the phenotype of interest has evolved only
246 once[43]. Our framework therefore provides a way forward to identify the genetic basis of
247 virulence, even in pathogens that evolved through clonal expansion, and begin to assess the
248 risk of pathogen emergence and reemergence from environmental gene pools.

249 **Methods**

250 **Genome sequencing.** DNA from clinical isolates (Bgd1, Bgd5, Bgd8, MQ1795) and
251 environmental isolates (GBE0428, GBE0658, GBE1068, GBE1114, GBE1173) was extracted
252 using the Gentra kit (QIAGEN) and purified using the MoBio PowerClean Pro DNA Clean-Up
253 Kit. Multiplexed genomic libraries were constructed using the Illumina-compatible Nextera DNA
254 Sample Prep kit following the manufacturer's instructions. Sequencing was performed with 250-
255 bp paired-end (v2 kit) reads on the illumina MiSeq.

256

257 **Genome assembly.** To exclude low-quality data, raw reads were filtered with Trimmomatic[44].
258 The 15 first bases of each reads were trimmed and reads containing at least one base with a
259 quality score of <30 were removed. *De novo* assembly was performed on resulting reads using
260 Ray v2.3.1[45].

261

262 **Genome alignment, annotation and SNP calling.** We used mugsy v.1 r.2.2[46] with default
263 parameters to align the primary dataset of 22 *V. cholerae* genomes (Supplementary Table 1).
264 From this alignment, we extracted dimorphic SNP sites and annotated genes according to MJ-
265 1236 as a reference genome. We replicated the alignment, annotation and SNP calling using 21
266 different *V. cholerae* genomes from Orata *et al*.[27] (plus the MJ-1236 reference, for a total of 22),
267 consisting of 15 environmental genomes (EGs: 1587, AM19226, MZ03, MZO2, V51, YB1A01,
268 YB1G06, YB2G01, YB3B05, YB4B03, YB4C07, YB4G05, YB4H02, YB5A06, YB7A09) and 7
269 phylocore genomes (PGs: 2010EL, 274080, BX330286, MAK757, MO10, V52, MJ-1236).

270
271 **Definition of orthologous groups.** Genomes were annotated using the RAST web server
272 (www.rast.nmpdr.org)[47]. Annotated genes were clustered into orthologous groups using
273 OrthoMCL (www.orthomcl.org)[48] with default parameters, yielding 2844 orthologous groups.

274

275 **Phylogenetic analysis.** We constructed a core genome phylogeny using the concatenated
276 alignment of 1031 single-copy orthologous protein-coding genes (present in exactly one copy in
277 each of the 22 primary dataset genomes). Each protein sequence was aligned with Muscle[49],
278 and the concatenated alignment was used to infer an approximate maximum likelihood
279 phylogeny with FastTree v. 2.1.8[50] using default parameters (Fig. 1a). Individual gene trees (Fig.
280 1d) were built in the same way. We constructed a neighbour-net of the 22 genomes using
281 SplitsTree v.4.10[51], based on dimorphic SNPs from the mugsy genome alignment, excluding
282 sites with gaps.
283
284 **Tests for selection.** We conducted a genome-wide version of the McDonald-Kreitman test[29] by
285 first counting the number of fixed nonsynonymous (fn), fixed synonymous (fs), polymorphic
286 nonsynonymous (pn), and polymorphic synonymous (ps) sites within each gene. We then
287 summed these values across all genes (FN, FS, PN, and PS) and calculated the genome-wide
288 Fixation Index, FI=(FN/FS)/(PN/PS). A fixation index greater than one suggests positive
289 selection between the ingroup and outgroup (in this case, between EGs and PGs). However,
290 care must be taken when computing a genome-wide FI because summing genes with different
291 amounts of substitutions and polymorphism can result in FI>1 in the absence of selection[52]. We
292 therefore performed 1000 permutations of the data, keeping the row totals (fn+fs and pn+ps)
293 and column totals (fn+pn and fs+ps) constant and recomputing FI. We used the mean FI from
294 the permutations as the expected value of FI under neutral evolution. To evaluate the
295 hypothesis that the observed FI was higher than expected, suggesting positive selection, we
296 computed a *P*-value as the fraction of permutations with FI greater or equal to the observed FI.
297 We repeated the test using polymorphism from either the PG group or the EG group
298 (Supplementary Table 3).
299 To identify individual genes under selection between PGs and EGs in the primary
300 dataset, we restricted our search to 87 genes with fn > 0.68 1 and fn:fs > 1.48 (respectively two

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

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

320

321 **Bacterial strains and plasmids.** *V. cholerae* O395 and *V. cholerae* N16961 were used as wild-
322 type strains of classical and El Tor biotypes respectively. Strains cultivated on solid medium
323 were grown on LB agar; strains in liquid media were grown in aerated LB broth at 37°C.
324 pKAS154 was used for allelic exchange[53]. When necessary, media was supplemented with
325 antibiotics to select for certain plasmids or strains of *V. cholerae* at the following concentrations:
326 gentamycin, 30μg/ml; kanamycin, 45μg/ml; polymyxin B, 50μg/ml; and streptomycin, 1 mg/ml.

327

328 **Strain construction.** In-frame deletions of genes of interest were constructed via homologous
329 recombination. PCR was used to amplify two approximately 500 bp fragments flanking the gene
330 and to introduce restriction sites. For exchange of environmental alleles the respective *ompU*
331 gene was also amplified and restriction sites introduced. The fragments were then cloned into a
332 restriction-digested plasmid using a three-segment ligation for in-frame deletion mutants and a
333 four-segment ligation for environmental allele exchange mutants. The resulting plasmid was
334 electroporated into *Escherichia coli* S17-1λpir. *E. coli* with the constructed plasmid was mated
335 with wild-type *V. cholerae* O395 or N16961, and allelic exchange was carried out by selection
336 on antibiotics, as described previously[53]. Potential mutants were screened using PCR: two
337 primers flanking the deletion construct were used to amplify chromosomal DNA isolated from
338 plated *V. cholerae*. The lengths of the PCR fragments were analyzed on 0.8% agarose gel for
339 gene deletions and putative deletions were subsequently confirmed by DNA sequencing.

340

341 **Protein electrophoresis and OmpU visualization.** Whole cell protein extracts were prepared
342 from cultures grown for overnight at 37°C in a rotary shaker. The extracts were subjected to
343 SDS-PAGE on 16% Tris Glycine gels (Invitrogen). OmpU bands were visualized after protein
344 gels were stained by Coomassie blue overnight.

345

346 **Survival assays.** *V. cholerae* strains were cultured overnight in LB broth at 37°C in a rotary
347 shaker. Overnight cultures were diluted 1:100 in LB and grown to an OD600 of 0.5. Cells were
348 pelleted and resuspended in either LB broth, LB containing 0.4% bile bovine (Sigma), or LB
349 containing 1000U/ml of polymyxin B (Sigma). Cultures were incubated for 1h at 37°C in a rotary
350 shaker. After incubation CFU/ml of each culture was calculated by plating dilutions in LB plates.

351 Survival was calculated by comparing the number of CFU/ml in LB plus treatment versus LB.
352 N≥6.
353
354 **Infant mouse competition assays.** Overnight cultures were diluted 1:100. Each test strain was
355 mixed in a 1:1 ratio with a Δ*lacZ* reference strain. Four- to 5-day-old CD-1 mice were inoculated
356 orogastrically with 50µl of the bacterial mixture. The intestines were harvested 24h post-
357 inoculation and homogenized in 4ml of LB broth containing 10% glycerol. The mixtures were
358 serially diluted and plated on LB agar plates supplemented with streptomycin and 5-bromo-4-
359 chloro-3-indolyl-D-galactopyramoside (X-Gal) (40µg/ml). The competition indices were
360 calculated as previously described by others, test (output CFUs/input CFUs)/reference (output
361 CFUs/ input CFUs).
362
363 **Biofilm assays.** 96-well plate assay. Cultures were incubated overnight at 30°C. 100µl of 1:100
364 dilutions of overnight cultures were placed per well in 96-well plates. Plates were left at 25°C for
365 24h. Liquid contents were discarded and plates were washed 2 times with LB. 200µl of 0.01%
366 crystal violet was added per well and incubated at room temperature for 5 minutes. Liquid
367 contents were discarded and plates were washed extensively with dH_2O. After the plates were
368 dry, biofilms were resuspended in 150µl of 50% acetic acid. Contents were transferred to a flat
369 bottom dish and quantitated in a microtiter plate reader at OD550. Values were plotted using
370 Prism software. N=15.
371

References

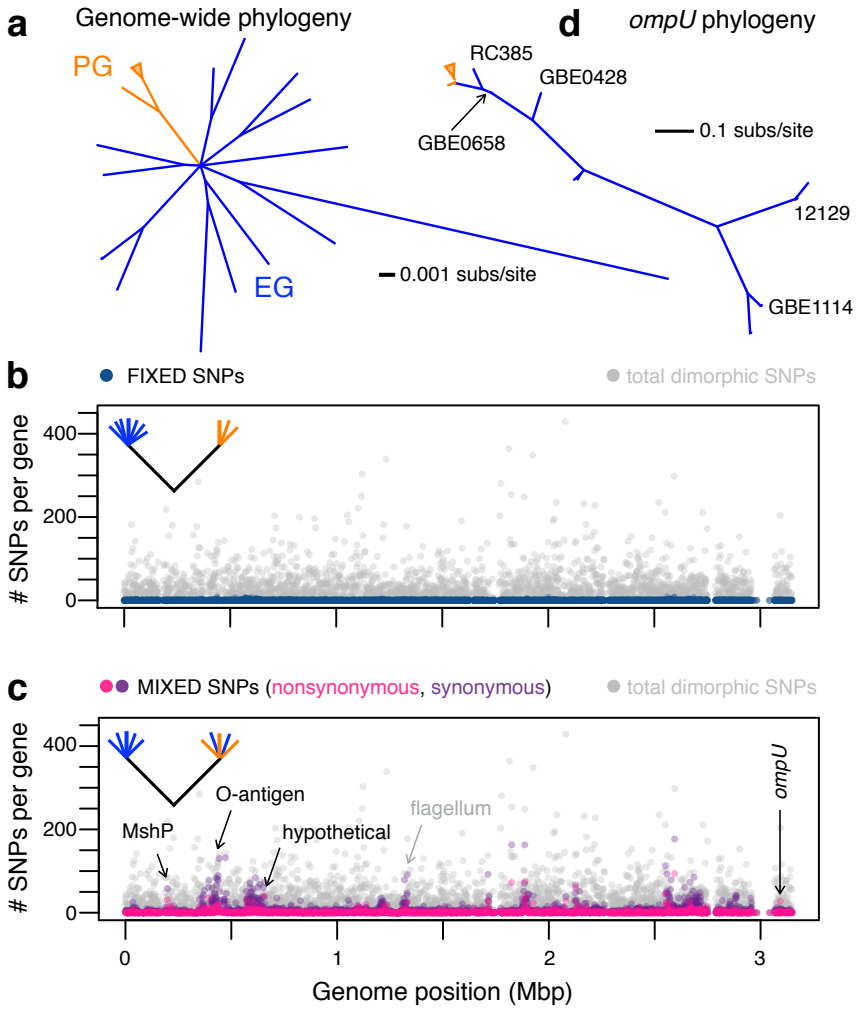
1. McNally, A., Thomson, N. R., Reuter, S. & Wren, B. W. 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.* **14,** 177–190 (2016).
2. Luo, C. *et al.* Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci. USA* **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. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
5. Shapiro, B. J. How clonal are bacteria over time? *Curr. Opin. Microbiol.* **31,** 116–123 (2016).
6. Groisman, E. A. & Ochman, H. Pathogenicity Islands: Bacterial Evolution in Quantum Leaps. *Cell* **87,** 791–794 (1996).
7. Ochman, H. Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. *Science* **292,** 1096–1099 (2001).
8. Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc. Natl. Acad. Sci. USA* **106,** 15442–15447 (2009).
9. Faruque, S. M. & Mekalanos, J. J. Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends Microbiol.* **11,** 505–510 (2003).
10. Kaper, J. B., Morris, J. G. & Levine, M. M. Cholera. *Clin. Microbiol. Rev.* **8,** 48–86 (1995).
11. Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T. & Calderwood, S. B. Cholera. *Lancet* **379,** 2466–2476 (2012).
12. 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).
13. 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).
14. Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272,** 1910–1914 (1996).
15. 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).
16. 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).
17. 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).
18. 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 Cell Biol.* **27,** 347–355 (2008).
19. 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).
20. Chakraborty, S. *et al.* Virulence genes in environmental strains of Vibrio cholerae. *Appl. Environ. Microbiol.* **66,** 4022–4028 (2000).
21. 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).
22. 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).
23. 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).
24. Almagro-Moreno, S. & Taylor, R. K. Cholera: Environmental Reservoirs and Impact on Disease Transmission. *Microbiol. Spect.* **1,** 149–165 (2013).
25. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477,** 462–465 (2011).
26. 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).
27. 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. Evo.l* **7,** 2941–2954 (2015).
28. Keymer, D. P. & Boehm, A. B. Recombination shapes the structure of an environmental *Vibrio cholerae* population. *Appl. Environ. Microbiol.* **77,** 537–544 (2011).

29. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).
30. 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).
31. Mathur, J. & Waldor, M. K. The *Vibrio cholerae* ToxR-regulated porin OmpU confers resistance to antimicrobial peptides. *Infect. Immun.* **72,** 3577–3583 (2004).
32. 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).
33. 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).
34. 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).
35. Almagro-Moreno, S., Pruss, K. & Taylor, R. K. Intestinal colonization dynamics of *Vibrio cholerae*. *PLoS Pathog.* **11,** e1004787 (2015).
36. 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).
37. 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).
38. Watnick, P. I. & Kolter, R. Steps in the development of a *Vibrio cholerae* El Tor biofilm. *Mol. Microbiol.* **34,** 586–595 (1999).
39. 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).
40. Hsiao, A., Liu, Z., Joelsson, A. & Zhu, J. *Vibrio cholerae* virulence regulator-coordinated evasion of host immunity. *Proc. Natl. Acad. Sci. USA* **103,** 14542–14547 (2006).
41. 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).
42. Friedman, J., Alm, E. J. & Shapiro, B. J. Sympatric Speciation: When Is It Possible in Bacteria? *PLoS ONE* **8,** e53539 (2013).
43. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* **25,** 17–24 (2015).
44. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
45. Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13,** R122 (2012).
46. Samuel V Angiuoli, S. L. S. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27,** 334–342 (2011).
47. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 1 (2008).
48. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
49. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
50. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5,** e9490 (2010).
51. Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21,** 255–265 (2004).
52. Shapiro, J. A. *et al.* Adaptive genic evolution in the Drosophila genomes. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 2271–2276 (2007).
53. Skorupski, K. & Taylor, R. K. Positive selection vectors for allelic exchange. *Gene* **169,** 47–52 (1996).
54. Son, M.S., Megli, C.J., Kovacikova, G., Qadri, F. & Taylor, R.K. Characterization of *Vibrio cholerae* O1 El Tor Biotype Variant Clinical Isolates from Bangladesh and Haiti, including a Molecular Genetic Analysis of Virulence Genes. *J Clinical Microbiology.* **49,** 3739-3749 (2011).

Acknowledgements

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

495 # Figures



496

497 **Figure 1. Comparative genomics reveals candidate virulence adaptive polymorphisms.**
498 **a,** Phylogeny of 22 *V. cholerae* genomes based on 1031 single-copy orthologs in the primary dataset. All
499 branches have local support values >0.99 except for very short, deep internal branches (resulting in the
500 star-like polytomy at the centre of the tree). Not all 22 genomes are visible because some have nearly
501 identical sequences (*e.g.* 6 of the 7 PG genomes are nearly identical, shown as an orange triangle;
502 GBE1173 and GBE1114 are nearly identical, as can be seen in Supplementary Figure 1). **b,** Distribution
503 of fixed SNPs across chromosome 1. (See Supplementary Fig. 3 for chromosome 2). Genome position is
504 according to the MJ-1236 reference genome. SNP-free regions (*e.g.* near 3 Mbp, the locus of the
505 integrative conjugative element) are part of the flexible genome, present in the reference but not the other
506 21 genomes. The schematic tree in the top left illustrates the fixed SNP pattern, in which one allele is
507 present in PGs and a different allele in EGs. **c**, Distribution of mixed SNPs across the genome. The
508 cartoon tree in the top left illustrates the mixed SNP pattern, in which one allele is fixed in PGs, and
509 another allele is polymorphic among EGs, with some EGs containing the PG-like allele. Black arrows
510 show candidate VAPs (Table 1). Grey arrow shows the flagellum as an example variable region not
511 containing candidate VAPs. **d**, *ompU* phylogeny. All visible branches have local support values >0.9
512 except for the branch separating RC385 and GBE0658, the branch grouping MJ-1236 and O395
513 together, and the branch grouping HE09 and VL426 together.

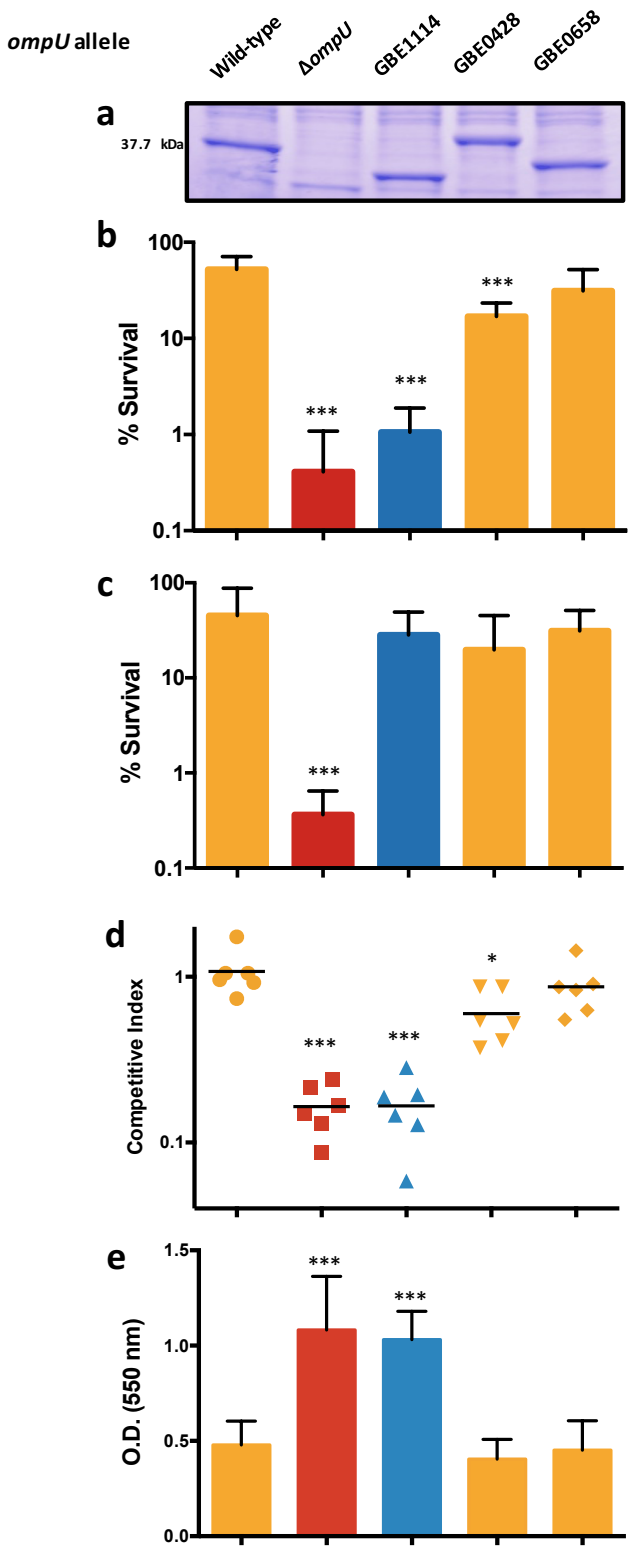
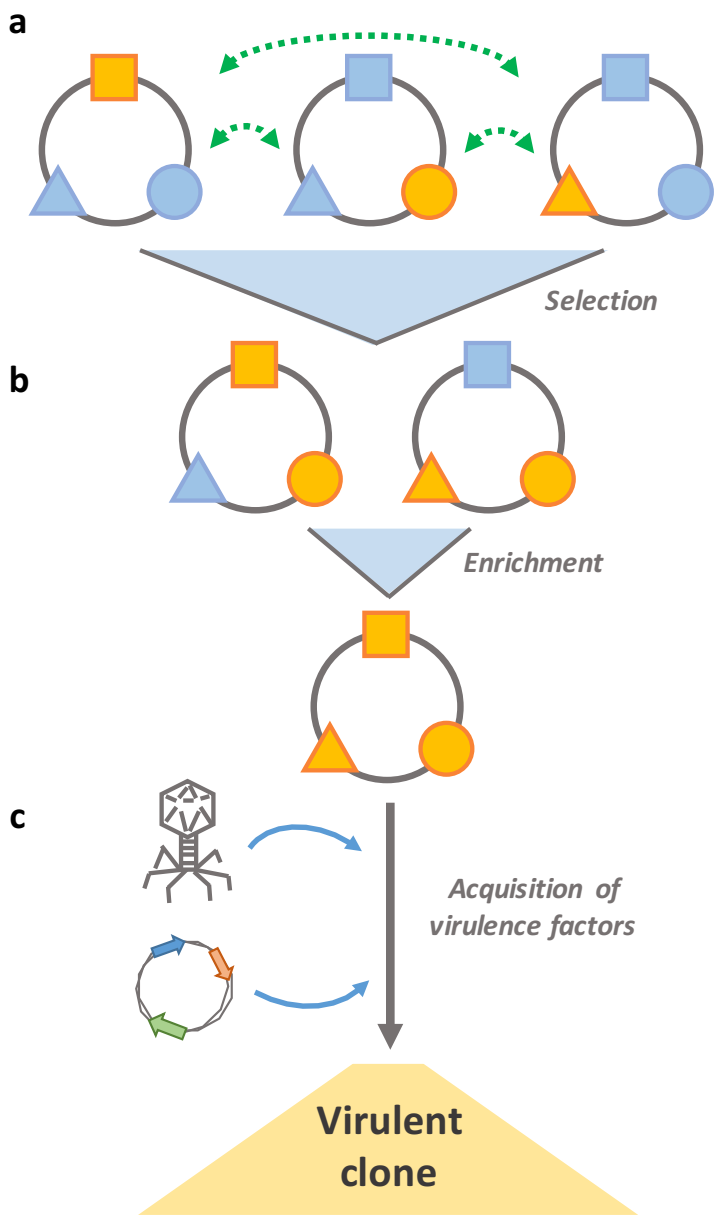


Figure 2. Phenotypic characterization of *ompU* alleles. a, OmpU production in clinical strains of *V. cholerae* encoding environmental alleles of *ompU*. Total protein lysates were run on a 16% Tris-glycine gel. OmpU bands were visualized after protein gels were stained with Coomassie blue. **b**, Survival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

514

515 **Figure 3**



516
517
518 **Figure 3. Model of pandemic clone emergence from an environmental gene pool.** We propose a
519 model that involves three events required for the emergence of pathogenic clones from environmental
520 populations. **a**, selection of VAPs. Virulence adaptive alleles circulate in naturally occurring populations
521 (orange symbols) and can be exchanged and mobilized through recombination (green dashed arrows).
522 Ecological events (temperature, nutrient availability, pH, etc.) lead to the selection of VAPs and an
523 increase in their distribution in environmental populations. **b**, enrichment of clones. A new ecological
524 opportunity occurs (human consumption of untreated waters, transient colonization of new environmental
525 hosts, etc.) which leads to the proliferation and enrichment in the population of clones encoding a mosaic
526 of VAPs. **c**, acquisition of virulence factors**.** A strain encoding a minimum set of VAPs required for host
527 colonization acquires the virulence factors that are necessary to produce a successful infection and
528 subsequently undergoes intra-host evolution and expansion.

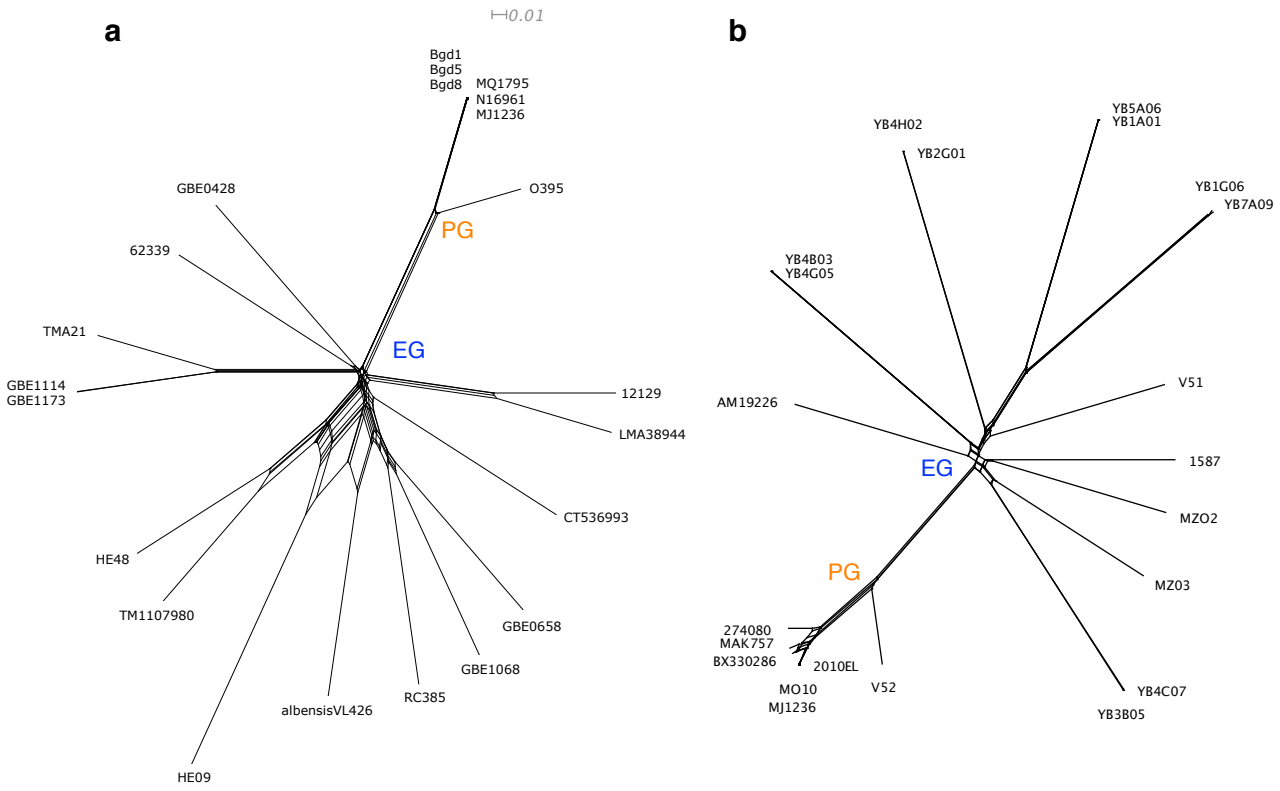
529 # Tables

530 **Table 1. Characteristics of five predicted VAPs with an excess of nonsynonymous mixed**
531 **SNPs.**

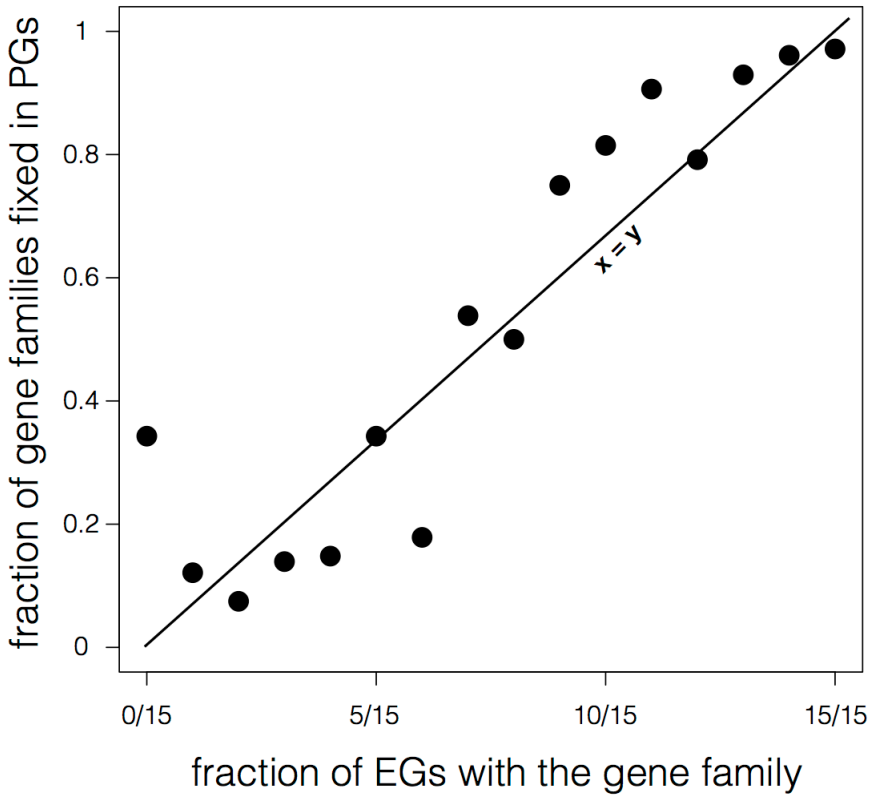
	Gene ID (VCD #)	Annotation	Gene length (bp)	Total # SNPs	Fixed NS	Fixed S	Mixed NS	Mixed S	*P*
1	003778	outer membrane protein OmpU	1053	204 (110)	0 (0)	0 (0)	28 (13)	11 (8)	0.0047* (0.0068*)
2	001600	hypothetical	258	21 (26)	0 (0)	1 (1)	15 (22)	4 (3)	0.0096* (2.32e-8*)
3	001013	hypothetical	642	19 (13)	0 (0)	0 (0)	15 (2)	4 (1)	0.0096* (n.s.)
4	001209	MSHA biogenesis protein MshP	432	85 (63)	0 (0)	0 (0)	14 (7)	4 (5)	0.0154 (0.066)
5	001230	lipid A core O-antigen ligase	1794	75 (47)	0 (0)	0 (1)	12 (0)	5 (1)	0.0717 (n.s.)

532 NS=nonsynonymous; S=synonymous. The genes listed have mixed NS and mixed NS:S both
533 over two standard deviations above the genome-wide median in the primary dataset. A binomial
534 test determined if the mixed NS:S ratio was greater than the expected genome-wide median
535 value of 0.5 per gene (uncorrected *P*-values shown; asterisks (*) indicate $P < 0.05$ after
536 Bonferroni correction for five tests). Numbers in parentheses are for the replication dataset, with
537 an expected genome-wide median mixed N:S of 0.33. *P*-values greater than 0.1 are denoted as
538 not significant (n.s.). Genes 1, 2, 4, and 5 are indicated with arrows on Figure 1C. Gene 3 is on
539 chromosome 2 (Supplementary Figure 3).

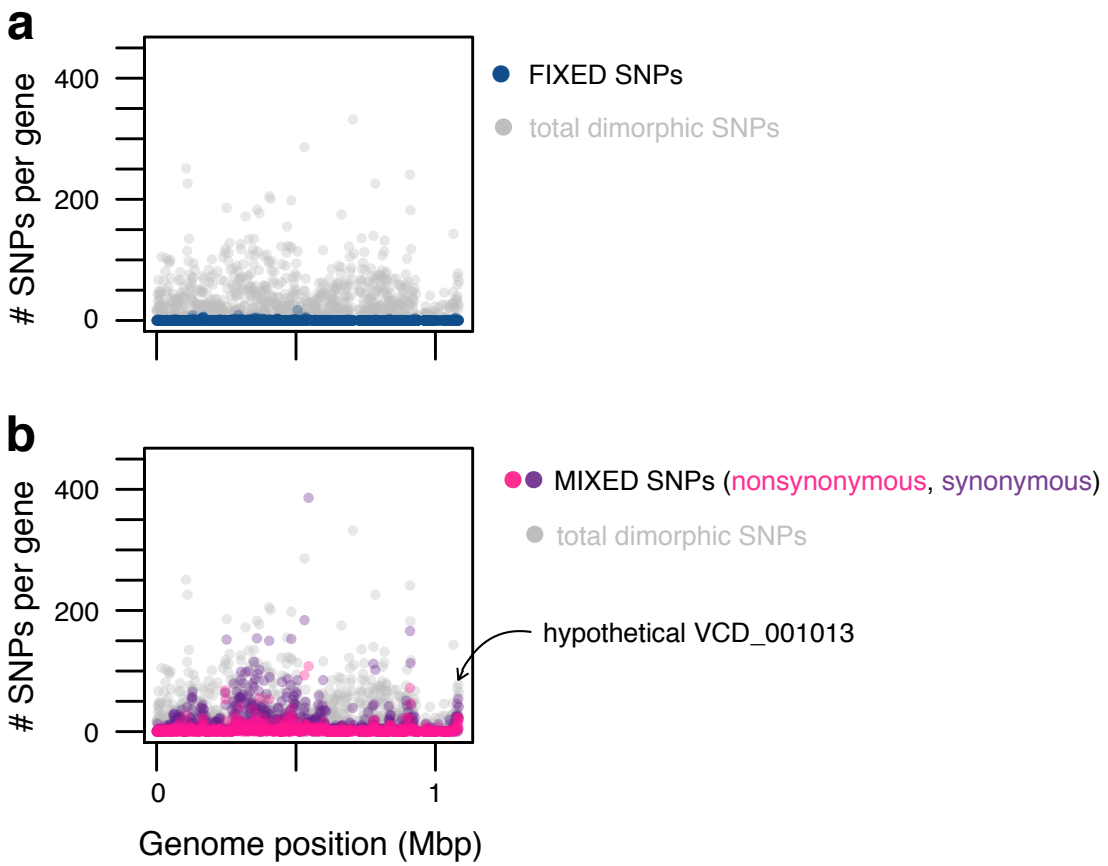
Supplementary Figures



Supplementary Figure 1. Splitstree showing relationship of environmental (EG) and pandemic (PG) *V. cholerae* genomes. a, The neighbour-net is based on dimorphic sites in the alignment of 22 genomes in the primary dataset, excluding sites with gaps. Only alignment blocks (locally colinear blocks produced by mugsy) including all 22 genomes were included, yielding 126,099 dimorphic sites. **b,** The neighbour-net based on 142,797 dimorphic sites in an alignment of 22 different genomes in the replication dataset (from Orata *et al.*[27]). The general star-like topology remains the same, with PGs clustering closely together at the end of one long branch.

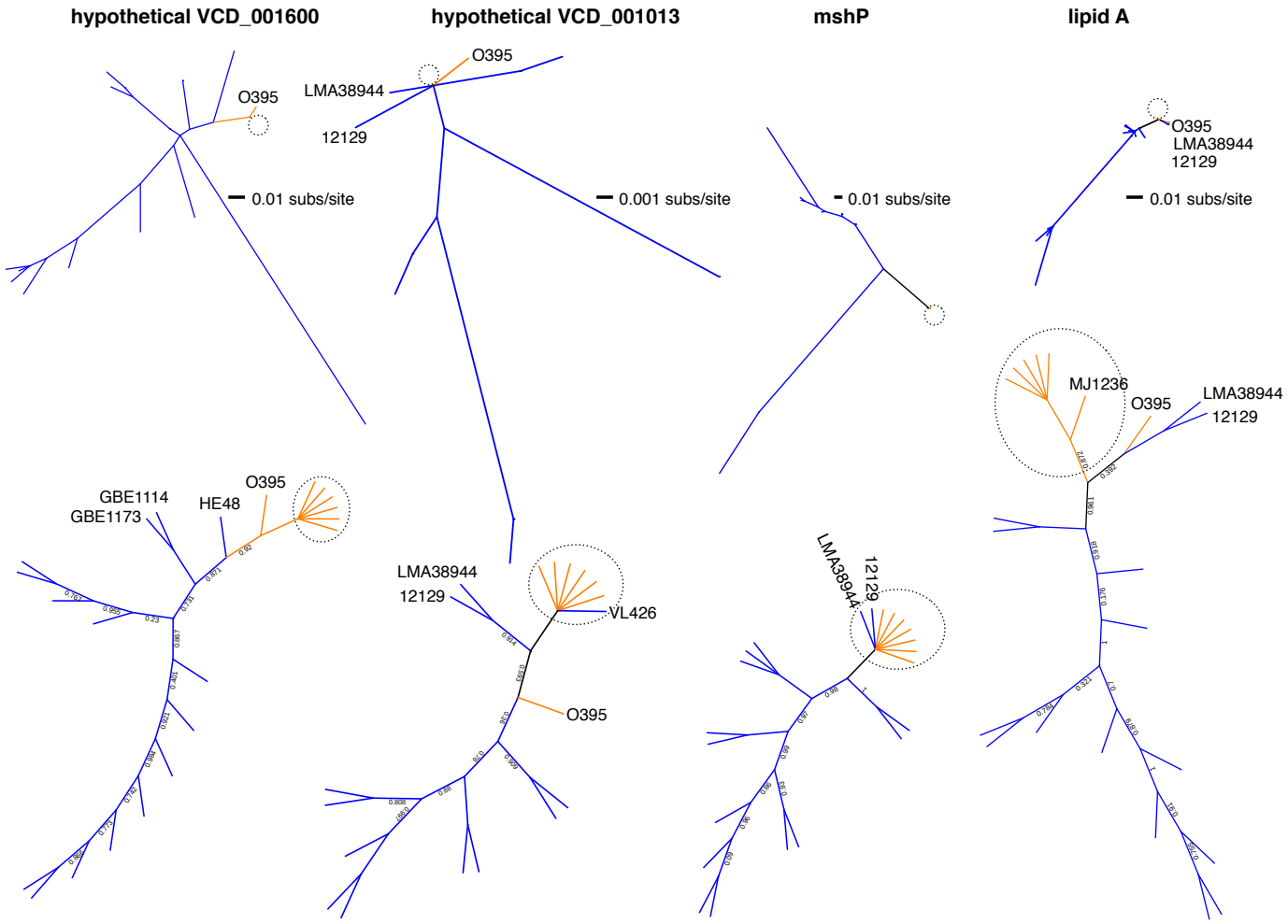


Supplementary Figure 2. PG strains sample genes randomly from the environmental pool. Gene families from OrthoMCL were binned by their frequency in environmental genomes (EGs), ranging from being present in zero to all 15 EGs in the primary dataset (x-axis). Within each of these bins, we calculated the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

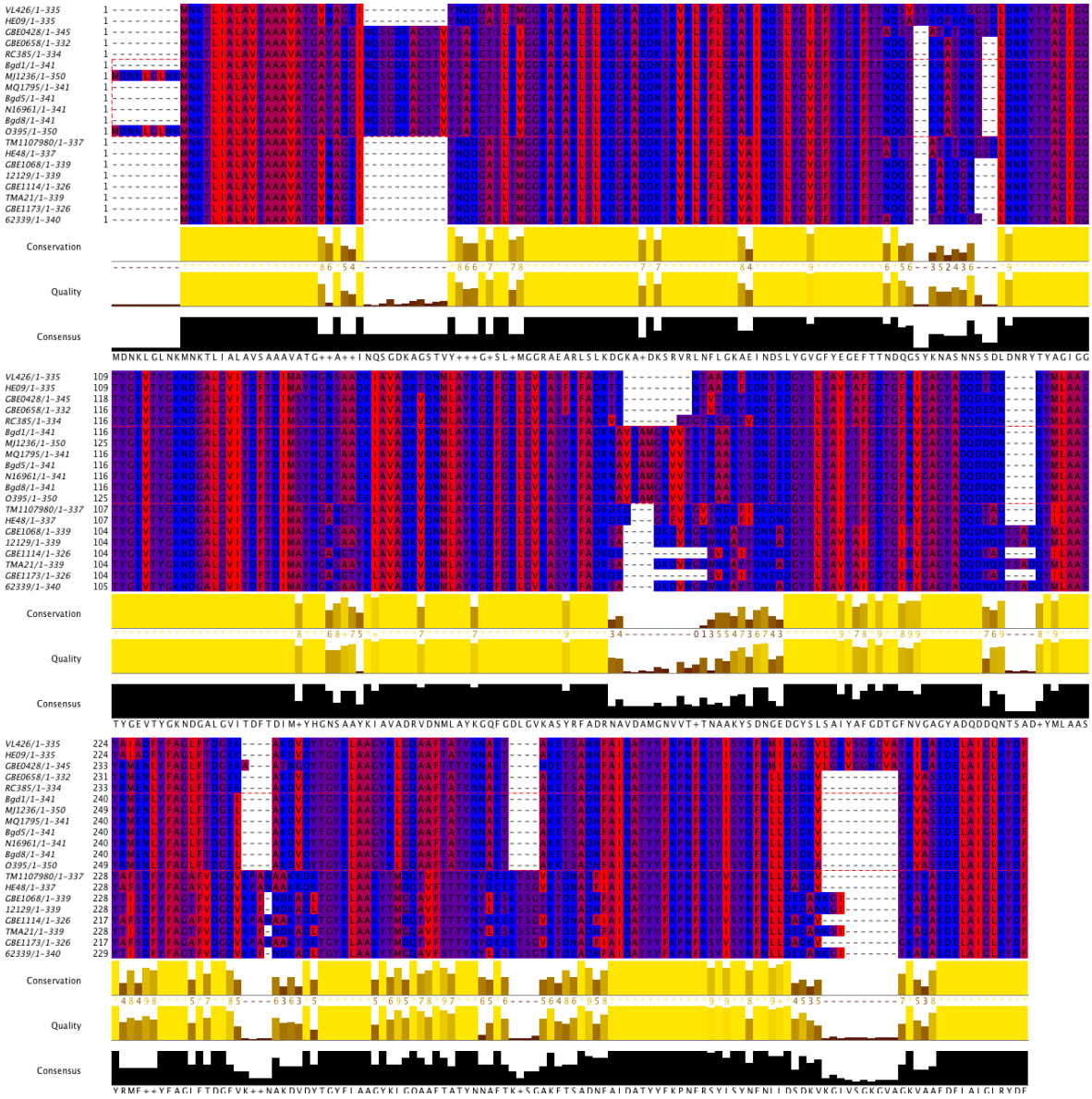


563

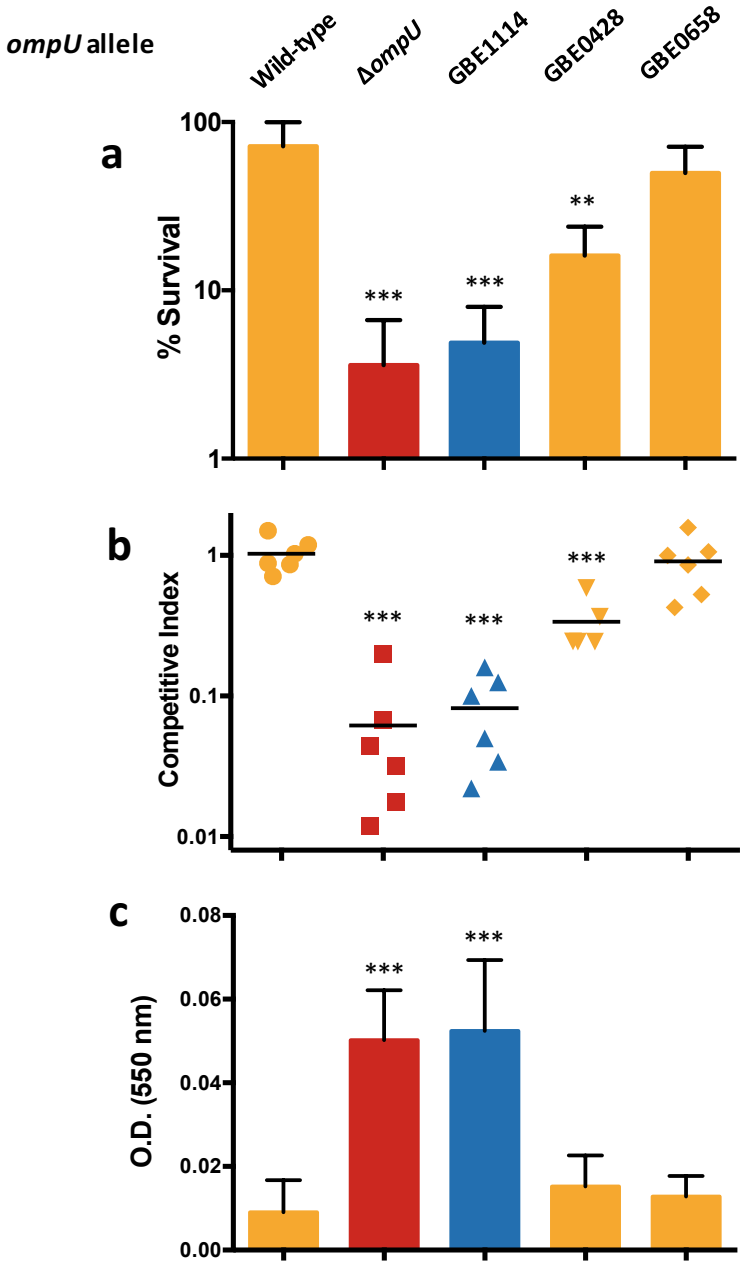
564 **Supplementary Figure 3. Distribution of fixed and mixed SNPs across chromosome 2. a,**
565 Distribution of fixed SNPs across chromosome 2 in the primary dataset. **b**, Distribution of mixed SNPs
566 across chromosome 2. Genome position is according to the MJ-1236 reference genome. See Fig. 1b and
567 1c in the main text for descriptions of fixed and mixed SNP sites. Arrow shows a candidate VAP (Table 1).
568



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.



586
587
588 **Supplementary Figure 6. Phenotypic characterization of *ompU* alleles in classical background. a**,
589 Survival of *ompU* mutants in the presence of bile. **b**, Colonization of the small intestine of *ompU* mutant
590 strains. **c**, Biofilm formation of *ompU* mutant strains on an abiotic surface. Yellow bars and symbols, PG-
591 like allele; red bars and squares, Δ*ompU*; blue bars and triangles, EG-like allele. Statistical comparisons
592 were made using student's *t*-test. $^{**}P < 0.01$, $^{***}P < 0.001$.

593 ## Supplementary tables

594 **Supplementary Table 1. *V. cholerae* genomes used in this study.** The 22 genomes in the
595 primary dataset are listed here; for a list of genomes in the replication dataset, see Methods and
596 Orata *et al*.[27].

Short name	PG/EG	Source	Accession number; source of strain
Bgd1	PG	O1 Ogawa El Tor, Bangladesh clinical	This study; Son et al.[54]
Bgd5	PG	O1 Inaba El Tor, Bangladesh clinical	This study; Son et al.[54]
Bgd8	PG	O1 Ogawa El Tor, Bangladesh clinical	This study; Son et al.[54]
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; Son et al.[54]
O395	PG	O1 Ogawa Classical, India clinical	CP000626/CP000627
GBE0428	EG	non-O1, USA oyster	This study; Schuster et al.[26]
GBE0658	EG	non-O1, USA water	This study; Schuster et al.[26]
GBE1068	EG	non-O1, USA oyster	This study; Schuster et al.[26]
GBE1114	EG	non-O1, USA water	This study; Schuster et al.[26]
GBE1173	EG	non-O1, USA sediment	This study; Schuster et al.[26]
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

597

598 **Supplementary Table 2. Summary statistics of SNPs in both datasets.** fixN = fixed
599 nonsynonymous SNPs; fixS = fixed synonymous SNPs; fmxN = mixed nonsynonymous SNPs;
600 mixS = mixed synonymous SNPs; sd = standard deviation.
601

	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

602

603 **Supplementary Table 3. Gene families present in all seven phylocore genomes (PGs) and**
604 **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

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

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

609 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

610
611 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

612
613 Total genomewide counts of classes of mutations: FN = fixed nonsynonymous, FS = fixed
614 synonymous, PN = polymorphic nonsynonymous, PS = polymorphic synonymous. Fixed
615 indicates that one allele is present in all PGs and a different allele in all EGs. Fixation Index (FI)
616 = (FN/FS)/(PN/PS)

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

	Gene ID (VCD #)	Annotation	Gene length (bp)	Total # SNPs	Fixed NS	Fixed S	Mixed NS	Mixed S	*P*
1	001506	hypothetical	331	21 (7)	0 (0)	0 (0)	20 (0)	1 (0)	4.11e-9 (n.s.)
2	003509	hypothetical; possible pseudogene	684	77 (35)	0 (0)	0 (0)	31 (4)	18 (0)	1.74e-5 (0.0625)
3	000213	hypothetical	3261	149 (118)	0 (0)	0 (0)	19 (54)	7 (64)	3.99e-5 (n.s.)

619 NS=nonsynonymous; S=synonymous. The genes listed have mixed NS and mixed NS:S both
620 over two standard deviations above the genome-wide median in the replication dataset. A
621 binomial test determined if the mixed NS:S ratio was greater than the expected genome-wide
622 median value of 0.33 per gene. Numbers in parentheses are for the primary dataset, with an
623 expected genome-wide median mixed N:S of 0.5. *P*-values greater than 0.1 are denoted as not
624 significant (n.s.). Note that the only predicted VAP gene with a significant *P*-value in both
625 primary (Table 1) and replication datasets with mixed N:S less than two (the neutral expectation,
626 based on approximately two nonsynonymous sites and one synonymous site per codon) is
627 VCD_003509, annotated in MJ-1236 as a predicted pseudogene. This is consistent with
628 approximately neutral evolution of pseudogenes.