TITLE: A combination of metagenomic and cultivation approaches reveals hypermutator phenotypes within *Vibrio cholerae* infected patients.

AUTHORS: Inès Levade[1,8], Ashraful I. Khan[2], Fahima Chowdhury[2], Stephen B. Calderwood[3,4,5], Edward T. Ryan[3,4], Jason B. Harris[3,6], Regina C. LaRocque[3,4,], Taufiqur R. Bhuiyan[2], Firdausi Qadri[2], Ana A. Weil[7], B. Jesse Shapiro[1,8,9*]

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

[2]Center for Vaccine Sciences, International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh

[3]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA

[4]Department of Medicine, Harvard Medical School, Boston, MA USA

[5]Department of Microbiology, Harvard Medical School, Boston, MA USA

[6]Department of Pediatrics, Harvard Medical School, Boston, MA, USA

[7]Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, USA

[8]Department of Microbiology and Immunology, McGill University, Montreal, QC, Canada

[9]McGill Genome Centre, Montreal, QC, Canada

CORRESPONDING AUTHOR: jesse.shapiro@mcgill.ca

ABSTRACT

Vibrio cholerae can cause a range of symptoms, ranging from severe diarrhea to asymptomatic infection. Previous studies using whole genome sequencing (WGS) of multiple bacterial isolates per patient showed that *V. cholerae* can evolve modest genetic diversity during symptomatic infection. To further explore the extent of *V. cholerae* within-host diversity, we applied culture-based WGS and metagenomics to a cohort of both symptomatic and asymptomatic cholera patients from Bangladesh. While metagenomics allowed us to detect more mutations in symptomatic patients, WGS of cultured isolates was necessary to detect *V. cholerae* diversity in asymptomatic carriers, likely due to their low *V. cholerae* load. Using both metagenomics and isolate WGS, we report three lines of evidence that *V. cholerae* hypermutators evolve within patients. First, we identified nonsynonymous in *V. cholerae* DNA repair genes in five out of 11 patient metagenomes sequenced with sufficient coverage of the *V. cholerae* genome, and in one of three patients with isolate genomes sequenced. Second, mutations in DNA repair genes tended to be accompanied by an excess of intrahost single nucleotide variants (iSNVs). Third, these iSNVs were enriched in transversion mutations, a known hallmark of hypermutator phenotypes. While hypermutators appeared to generate mostly selectively neutral mutations, non-mutators showed signs of convergent mutation across multiple patients, suggesting *V. cholerae* adaptation within hosts. Our results highlight the power and limitations of metagenomics combined with isolate sequencing to characterize within-patient diversity in acute *V. cholerae* infections, while providing evidence for hypermutator phenotypes within cholera patients.

42 **IMPORTANCE**

43 Pathogen evolution within patients can impact phenotypes such as drug resistance and virulence,
44 potentially affecting clinical outcomes. *V. cholerae* infection can result in life-threatening
45 diarrheal disease, or asymptomatic infection. Here we describe whole-genome sequencing of *V.*
46 *cholerae* isolates and culture-free metagenomic sequencing from stool of symptomatic cholera
47 patients and asymptomatic carriers. Despite the typically short duration of cholera, we found
48 evidence for adaptive mutations in the *V. cholerae* genome that occur independently and
49 repeatedly within multiple symptomatic patients. We also identified *V. cholerae* hypermutator
50 phenotypes within several patients, which appear to generate mainly neutral or deleterious
51 mutations. Our work sets the stage for future studies of the role of hypermutators and within-
52 patient evolution in explaining the variation from asymptomatic carriage to symptomatic cholera.
53

54 KEYWORDS: *Vibrio cholerae,* cholera, metagenomics, within-patient evolution, hypermutation,
55 asymptomatic carriage
56

57 **INTRODUCTION**

58 Infection with *Vibrio cholerae,* the etiological agent of cholera, causes a clinical spectrum
59 of symptoms that range from asymptomatic colonization of the intestine to severe watery diarrhea
60 that can lead to death. Although absent from most resource-rich countries, this severe diarrheal
61 disease still plagues many developing nations. According to the WHO, there are an estimated 1.3
62 to 4.0 million cases of cholera each year, with 21,000 to 143,000 deaths worldwide (1). Cholera
63 predominantly occurs in endemic areas, but can also cause explosive outbreaks as seen in Haiti in
64 2010 or in Yemen, where over 2.2 million cases are suspected since 2016 (2, 3). Although
65 cholera vaccines have reduced disease in some areas, the increasing number of people lacking
66 access to sanitation and safe drinking water, the emergence of a pandemic lineage of *V. cholerae*
67 with increased virulence (4), and environmental persistence of this waterborne pathogen
68 underscore the need to understand and interrupt transmission of this disease.

69 Cholera epidemiology and evolutionary dynamics have been studied by high-throughput
70 sequencing technologies and new modeling approaches, at a global and local scale (5, 6). Yet,
71 many questions remain regarding asymptomatic carriers of *V. cholerae*, including their role and
72 importance in the transmission chain during an epidemic (7, 8). Numerous observational studies
73 have identified host factors that could impact the severity of symptoms, including lack of pre-
74 existing immunity, blood group O status, age, polymorphisms in genes of the innate immune
75 system, or variation in the gut microbiome (9–13).

76 Recent studies have shown that despite the acute nature of *V. cholerae* infection, which
77 typically lasts only a few days, genetic diversity can appear and be detected in a *V. cholerae*
78 population infecting individual patients (14, 15). In a previous study, we sampled multiple *V.*
79 *cholerae* isolates from each of eight patients (five from Bangladesh and three from Haiti) and
80 sequenced 122 bacterial genomes in total. Using stringent controls to guard against sequencing

81 errors, we detected a few (0-3 per patient) within-patient intra-host single nucleotide variants
82 (iSNVs), and a greater number of gene content variants (on the order of ~100 gene gain/loss
83 events within patients) (15). This variation may affect adaptation to the host environment, either
84 by resistance to phage predation (14) or by impacting biofilm formation (15).

85 Several pathogens are known to evolve within human hosts (16), and hypermutation has
86 been observed in some cases due to loss-of-function mutations in the mismatch repair machinery
87 (17–19). While these hypermutators may quickly acquire adaptive mutations, they also bear a
88 burden of deleterious mutations (20). For the population to survive the burden of deleterious
89 mutations, hypermutators may revert to a non-mutator state, or may recombine their adaptive
90 alleles into the genomes of non-mutators in the population (17, 21). The hypermutator phenotype
91 has been observed in vibrios in the aquatic environment (22), and induced in *V. cholerae* in an
92 experimental setting (23), but not clearly documented within infected patients. There is some
93 evidence for hypermutation in *V. cholerae* clinical strains isolated between 1961 and 1965 (24);
94 however, the authors recognized that these hypermutators could also have emerged during long-
95 term culture (25). It therefore remains unclear if hypermutators readily evolve within individual
96 cholera patients.

97 When within-patient pathogen populations are studied with culture-based methods, their
98 diversity may be underestimated because the culture process can select isolates more suited to
99 growth in culture, and due to undersampling of rare variants (26). In this study, we used a
100 combination of culture-free metagenomics and WGS of a limited number of cultured isolates to
101 characterize the within-patient diversity of *V. cholerae* in individuals with different clinical
102 syndromes ranging from symptomatic to asymptomatic infection. Using both approaches, we
103 report evidence of *V. cholerae* hypermutators within both symptomatic and asymptomatic
104 infected patients. These hypermutators are characterized by a high mutation rate, and

105 accumulation of an excess of likely neutral or deleterious mutations in the genome. Finally, we

106 provide evidence of adaptive mutations occurring in non-mutator *V. cholerae* infections.

107

108 **RESULTS**

109 **Taxonomic analyses of metagenomics sequences from *Vibrio cholerae* infected index cases**
110 **and household contacts.**

111 To evaluate the level of within-patient diversity of *Vibrio cholerae* populations infecting
112 symptomatic and asymptomatic patients in a cohort in Dhaka, Bangladesh, we used both culture-
113 based whole genome sequencing and culture-free shotgun metagenomic approaches (Fig. 1).
114 Cholera patients and their household contacts were enrolled, because household contacts of
115 cholera patients are known to be at high risk of *V. cholerae* infection during the week after an
116 index case of cholera occurs in the household (11). Household contact infections are known to
117 exhibit a range of clinical outcomes from asymptomatic to severe symptomatic disease. We
118 enrolled patients from February 2013 to May 2014, and collected demographic information,
119 obtained rectal swabs, immunologic measures, and symptom histories during 30 days of
120 prospective follow-up, as previously described (11–13). We excluded household contacts who
121 reported antibiotic use during the week prior to enrollment.

122 We performed metagenomic sequencing of 22 samples from 21 index cases and 11
123 samples from 10 household contacts infected with *Vibrio cholerae,* of which two remained
124 asymptomatic during the 30-day follow-up period (Table S1). After removal of reads mapping to
125 the human genome, we used Kraken2 and MIDAS to taxonomically classify the remaining reads
126 and identify samples with enough *Vibrio cholerae* reads to reconstruct genomes. Among
127 symptomatic patients (index cases and household contacts), 15 samples from 14 patients
128 contained enough reads to reconstruct the *Vibrio cholerae* genome with a mean depth of coverage
129 > 5X. Neither of the two asymptomatic patients had enough *Vibrio cholerae* reads in their
130 metagenomic sequences to reconstruct genomes by mapping or *de novo* assembly (mean

131 coverage depth <0.05X). We also detected reads from two *Vibrio* phages (ICP1 and ICP3) in
132 some of these samples (Table S1).
133
134 **Recovery of high quality *Vibrio cholerae* MAGs from metagenomic samples**
135
136 To reconstruct *Vibrio cholerae* metagenomic assembled genomes (MAGs) from the 11
137 samples with coverage depth > 10X, we *de novo* assembled each sample individually except for
138 patient E, for whom we co-assembled two samples from two consecutive sampling days. High
139 quality MAGs identified as *Vibrio cholerae* were obtained from each assembly, with no
140 redundancy, and completeness ranging from 91 to 100% (Table S2). We dereplicated the set of
141 bins and removed all but the highest quality genome from each redundant set, identifying the bin
142 from patient J as the best quality MAG overall, which was used as a *V. cholerae* reference for
143 read mapping and SNV calling.
144
145 ***Vibrio cholerae* within patient nucleotide diversity estimated from metagenomic data**
146 All metagenomes with *Vibrio cholerae* mean coverage depth >5X were mapped against the
147 dereplicated genome set, and we assessed within-patient genetic diversity using inStrain (27).
148 This program reconstructs the "strain cloud" of a bacterial population, by mapping metagenomic
149 reads to metagenome-assembled genomes (MAGs) and calculates the allele frequency of each
150 single nucleotide variant (SNV) in the population. To remove potential false positive SNVs (due
151 to sequencing errors or mis-mapping of reads belonging to other species) we applied stringent
152 filtering thresholds (Methods) and identified both single nucleotide polymorphisms (SNPs) that
153 varied between patients (Table S3), and intra-host single nucleotide variants (iSNVs) that varied
154 within patients (Table S4). We found a total of 39 SNPs between patients, and a range of two to

155 207 iSNVs within each patient metagenome (Table 1, Fig. 2). Given the wide variation in
156 coverage across samples, we checked for any bias toward detecting iSNVs in high-coverage
157 samples. We observed no correlation between the number of detected iSNVs and depth of
158 coverage (Fig. S3; $\rho = -0.12$, $P = 0.65$, Pearson correlation), suggesting no coverage bias, and
159 that diversity levels are comparable across samples.

160 Several mechanisms could account for the origins of the observed iSNVs, including *de novo*
161 mutation within a patient, co-infection by divergent *V. cholerae* strains, homologous
162 recombination, or sequencing errors. Despite stringent filters for iSNV calls in inStrain, some
163 iSNVs could be false-positives due to sequencing or read mapping errors. In patient E, sampled
164 on two consecutive days, we detected eight iSNVs on the first day, of which four were again
165 detected on the second day, along with 13 additional iSNVs. It would be unlikely for random
166 sequencing errors to occur in the exact same four sites on two consecutive days by chance alone,
167 therefore these iSNVs are likely true positives. The additional iSNVs detected at only one time
168 point could be sequencing errors, or could reflect iSNV allele frequency changes over time. In the
169 analyses that follow, we acknowledge that a subset of iSNVs could be false-positives, but assume
170 that this source of error is randomly distributed across samples and can thus be accounted for in
171 statistical tests.

172 Although it is difficult to distinguish *de novo* mutation from coinfection using metagenomic
173 data alone, the distribution of iSNV allele frequencies may provide clues. Specifically, under a
174 standard neutral coalescent model, a single evolving population or strain is expected to generate a
175 geometric distribution of iSNV frequencies, dominated by a peak of low-frequency mutations. In
176 contrast, simple mixtures of a few strains (*i.e.* coinfections) will produce distributions with one or
177 more peaks at intermediate allele frequencies (28). Examining the iSNV frequency distributions

178 in our data revealed four patients (A, B, H, and L) dominated by a single peak of low-frequency
179 alleles, and two patients (F and I) with a possible peak at intermediate frequency, consistent with
180 a mixture of strains at frequencies of roughly 0.15 and 0.85 (Fig. S1). Other patients had
181 distributions too noisy to interpret, often due to small number of iSNVs. The iSNV frequency
182 distributions of patients F and I are consistent with coinfection, but could equally be explained by
183 within-host balancing or negative frequency-dependent selection. Coinfections might be expected
184 to have lower ratios of nonsynonymous (NS) to synonymous (S) substitutions compared to within
185 host *de novo* mutations, since NS mutations are more likely to be deleterious and purged over
186 evolutionary time. However, the NS:S ratios in patients F and I were in the same range as those
187 observed in patients A, B, H, and L, providing no clear support for the coinfection hypothesis
188 (Table 1). We therefore cannot exclude coinfection as a source of iSNVs in a minority of
189 patients, although the evidence remains ambiguous.
190 Finally, iSNVs could arise via homologous recombination rather than point mutation. iSNVs
191 were distributed across the genome (Fig. 2A), rather than clustered in hotspots as would be
192 expected if iSNVs arose from recombination events (29). No iSNVs were observed at the exact
193 same nucleotide position in different patients, suggesting that iSNVs rarely spread by
194 homologous recombination, and are never precisely recurrent in our dataset.
195
196 **Evidence for *V. cholerae* hypermutators within patients**
197 In five of the six patients with a high number of iSNVs (>25), we identified nonsynonymous
198 (NS) mutations in genes involved in DNA mismatch repair pathways, including the DNA
199 polymerase II in patient D, or proteins of the methyl-directed mismatch repair (MMR) system in
200 patient F, I and K (Table 1). We henceforth define these five patients as containing potential
201 hypermutator lineages of *V. cholerae.* Although the precise functional consequences of these NS

202 mutations are unknown, they are potential loss-of-function mutations that could plausibly result
203 in hypermutator phenotypes (17). In the patient harboring the highest number of variants (Patient
204 F, 207 iSNVs), we detected two NS mutations in two different genes coding for proteins involved
205 in DNA repair: the DNA mismatch repair endonuclease MutL (17), and the nuclease SbcCD
206 subunit C (24)(30)(31). Even with such a high number of iSNVs, it is surprising to observe NS
207 mutations in two DNA repair genes in patient F (133 NS mutations in ~4000 genes; $p = 0.033$ for
208 a mutation in one gene and $p = 0.0011$ for two genes). In patient I, where we also detected a high
209 number of iSNVs, a NS mutation in the gene coding for the MutT/nudix protein, involved in the
210 repair of oxidative DNA damage (32), could also cause a strong hypermutation phenotype.
211 Patients D, H and K presented fewer iSNVs but also contained NS mutations in genes involved in
212 DNA damage repair (33–35). However, some of these genes have been shown to play less critical
213 roles in bacterial DNA repair than MutSLH (17, 36), which could lead to a weaker hypermutator
214 phenotype.
215 The patient with the second highest number of iSNVs, patient A, contained a high number of
216 intergenic variants (87 out of 96 iSNVs, Fig. 2B), but no apparent NS mutations in genes
217 involved in DNA repair; we therefore did not consider patient A as a hypermutator. This large
218 number of intergenic iSNVs could be caused by read mapping errors to a distantly related *V.*
219 *cholerae* reference genome; however, the same iSNV calls were obtained when using the MAG
220 from patient A as a reference genome. False iSNVs could also occur due to mismapping of reads
221 from different species to the *V. cholerae* genome. Although we took measures to exclude such
222 mismapping by removing reads mapped to 79 representative MAGs in our patient microbiomes,
223 and by excluding sites with aberrant high or low depth of coverage (Methods), we cannot exclude
224 the possibility that patient A contained a cryptic member of the gut microbiome that resulted in
225 mismapping.

226 Previous studies have noted mutational biases in hypermutators, such as an increase of
227 transition over transversion mutations in a *Burkholderia dolosa* mutator with a defective MutL
228 (18), or an excess of G:C→T:A transversions in a *Bacillus anthracis* hypermutator (37), and in
229 members of the gut microbiome (38). When we compared the spectrum of mutations observed in
230 suspected hypermutators to non-mutator samples, we found a significance difference (Chi-square
231 test, $P<0.01$) due to an apparent excess of G:C→T:A transversions in hypermutators (Fig. 2C;
232 Fig. S2). While not all NS mutations in DNA repair genes necessarily cause defects, we observed
233 an increase in transversion mutations concordant with the MMR gene mutated (Fig. S2). For
234 instance, it has been shown in other bacterial pathogens that mutations in *mutT* and *mutL* lead to a
235 strong mutator phenotype, increasing the rate of A:T→C:G and G:C →A:T transversions
236 respectively (34), which we observed in patients (F and I) containing these mutations (Table 1,
237 Fig. S2). Mutations in *mutM* were also previously associated with G:C →T:A mutations, as
238 observed in patient K (Fig. S2). More experiments are clearly needed to confirm the phenotypes
239 of these DNA repair mutants, but our results are largely consistent with known hypermutation
240 profiles.

241 Current theory suggests that hypermutators may be adaptive under novel or stressful
242 environmental conditions because they more rapidly explore the mutational space and are the first
243 to acquire adaptive mutations. However, hypermutation comes at the cost of the accumulation of
244 deleterious mutations. To test the hypothesis that hypermutation leads to fitness costs due to these
245 deleterious mutations, we used iRep (39) to estimate *V. cholerae* replication rates in each sample,
246 and test whether replication rate was negatively associated with the number of iSNVs. iRep infers
247 replication rates from MAGs and metagenomic reads (39). For instance, an iRep value of 2 would
248 indicate that most of the population is replicating one copy of its genome. In our data (Table 1),
249 iRep values varied from 1.23 (patient E at day 2) to 5.43 (patient D), and we did not find any

250 association between the replication rate of *Vibrio cholerae* and the number of iSNVs detected
251 within each subject (Fig. S3, Pearson correlation, $\rho = 0.15$, $P > 0.05$). This suggests that
252 deleterious mutations in hypermutators could be counterbalanced by adaptive mutations that
253 maintain growth. Alternatively, higher iRep values could be associated with larger *V. cholerae*
254 population sizes, which could support greater genetic diversity and yield a positive correlation
255 between iRep values and the number of iSNVs. These hypotheses merit testing in larger patient
256 cohorts.
257
258 **Convergent evolution suggests adaptation of non-mutator *V. cholerae* within patients**
259 While none of the patients shared iSNVs at the exact same nucleotide position, some
260 contained mutations in the same gene that occurred independently in more than one patient
261 (Table 2). These are examples of convergent evolution at the gene level. To determine whether
262 genes that acquired multiple mutations in independent patients could be under convergent
263 selection within the host, we performed permutation tests for hypermutator and non-mutator
264 samples separately (Methods). Such a test would identify consistent signatures of either positive
265 or relaxed purifying selection common to multiple hosts. Among the hypermutator samples, we
266 identified five genes with NS mutations in two or more patients (Table 2), which was not an
267 unexpectedly high level of convergence given the large number of mutations in hypermutators
268 (permutation test, P=0.97). That the P-value approaches 1 suggests either that the hypermutators
269 are actually selected against mutating the same genes in different patients, or – more likely – that
270 the permutation test is conservative. For the samples with no evidence of hypermutator
271 phenotypes, we identified two genes with NS mutations in two patients. The first gene, *hlyA*,
272 encodes a hemolysin that causes cytolysis by forming heptameric pores in human cell membranes

273 (40), while the second gene encodes a putative ABC transporter ferric-binding protein (Table 2).
274 Observing convergent mutations in two different genes is unexpected (permutation test, P=0.039)
275 in a test that is likely to be conservative. We also note that the three iSNVs in *hlyA* have
276 relatively high minor allele frequencies (0.22-0.43) compared to other convergent NS mutations
277 (median minor allele frequency of 0.11; Table 2) and to NS mutations overall (median of 0.12;
278 Table S4). Together, these analyses suggest that *V. cholerae* hypermutators produce NS
279 mutations that are predominantly deleterious or neutral, while evidence for within-patient
280 positive selection on certain genes in non-mutators merits further investigation.

281 To further explore differential selection at the protein level within and between patients,
282 we applied the McDonald-Kreitman test (41) to the 9 patients with no evidence for
283 hypermutation, and to the five patients harboring potential hypermutators. Based on whole-
284 genome sequences of *V. cholerae* isolates, we previously found an excess of NS mutations fixed
285 between patients in Bangladesh, based on a small sample of five patients (15). Here, based on
286 metagenomes from a larger number of patients, we found the opposite pattern of a slight excess
287 of NS mutations segregating as iSNVs within patients, consistent with slightly deleterious
288 mutations occurring within patients and purged over evolutionary time. However, the difference
289 between NS:S ratios within and between patients was not statistically significant (Fisher's exact
290 test, P>0.05; Table S5); thus, the evidence for differential selective pressures within versus
291 between cholera patients remains inconclusive.

292 Many NS mutations occurred in genes involved in transmembrane transport,
293 pathogenesis, response to antibiotics, secretion systems, chemotaxis, and metabolic processes
294 (Fig. S4). Both hypermutator samples (Fig. S4B) and non-mutators (Fig. S4C) have a high NS:S
295 ratio in genes of unknown function, while hypermutators have many NS mutations in
296 transmembrane proteins, which are absent in non-mutators. However, non-mutator samples have

297 more NS mutations in genes involved in pathogenesis and secretion systems. Most of the NS
298 mutations involved in pathogenesis were found in the gene *hlyA* (a target of convergent
299 evolution, mentioned above).
300
301 **Evidence for a hypermutator from isolate whole-genome sequencing**
302 In addition to metagenomic analyses, we performed whole genome sequencing of
303 multiple *Vibrio cholerae* clinical isolates from index cases and asymptomatic contacts (Fig. 1A)
304 from three households (56, 57, and 58, Table S1). As noted above, asymptomatic infected
305 contacts did not yield sufficient metagenomic reads to assemble the *V. cholerae* genome or call
306 iSNVs, but their stool cultures yielded colonies for whole-genome sequencing. The first
307 asymptomatic contact, 58.01, tested positive for *Vibrio cholerae* on day 4 after the presentation of
308 the index case to the hospital, and *Vibrio cholerae* was cultured from the stool on days 4, 6, 7 and
309 8. We sequenced five isolates respectively from day 4 and 6 samples, and four isolates from each
310 of the subsequent days. For households 56 and 57, five isolates were sequenced from each
311 sample, at day 1 for the index cases and day 2 for the asymptomatic carriers (Table S6).
312 The index case from household 58 (called 58.00 or patient N) was also included in the
313 metagenomic analysis described above, allowing a comparison between culture-dependent and -
314 independent assessments of within-patient diversity. We did not detect any iSNVs among the five
315 isolates sequenced from patient 58.00. In contrast, the metagenomic analysis of patient N
316 revealed seven iSNVs (Table 1), suggesting a potentially higher sensitivity for the detection of
317 rare variants, or possibly false-positive iSNVs inferred from metageomic reading mapping
318 compared to isolate sequencing. Distinguishing between these possibilities would require
319 sequencing a much larger number of colonies, which is beyond the scope of the current study.

320 In contrast to metagenomes consisting of many unlinked reads, whole-genome sequencing
321 allows the reconstruction of a phylogeny describing the evolution of *V. cholerae* within and
322 between patients based on SNVs in the core genome (Fig. 3). As described previously (6),
323 isolates from members of the same household tended to cluster together. In index case 57.00, all
324 isolates were identical in terms of SNVs, with the exception of one isolate that was identical to
325 the five isolates sequenced from the asymptomatic contact from the same household, patient
326 57.01 (Table 3, Fig. 3). In the inferred phylogeny, isolates from the contact 57.01 are ancestral,
327 and the more common genotype in the index case 57.00 has one additional derived mutation. This
328 shared ancestral genotype between the two individuals was unexpected, and could suggest a
329 potential transmission event from the asymptomatic contact to the index case, followed by a
330 mutational event and the spreading of the new variant in the index case. The only mutation found
331 in four of the five isolates from the index case was a non-synonymous mutation in a gene coding
332 for a cyclic-di-GMP-modulating response regulator, which could have an impact on the
333 regulation of biofilm formation in the host (42). However, the direction of transmission (from
334 contact to index case) is only supported by one mutation, and therefore remains uncertain. The
335 genomes from household 57 are also similar or identical to genomes from household 56,
336 suggesting further caution in inferring transmission chains. Among the other index cases, we
337 found no iSNVs in patient 58.00 and two iSNVs in patient 56.00. One isolate from this patient
338 had a synonymous mutation in a hypothetical protein, and another isolate had a non-synonymous
339 mutation in a UDP-N-acetylglucosamine 4,6-dehydratase gene (Table 3). We detected iSNVs in
340 the other asymptomatic contacts, with one synonymous and one intergenic mutation in contact
341 58.02, and one non-synonymous mutation in one isolate from contact 56.01 (Table 3, Fig. 3).
342 Notably, we also found evidence for a hypermutator in contact 58.01. One isolate sampled
343 from this contact had the highest number of mutations seen in any branch in the phylogeny (five

344 NS mutations, all G:C→T:A transversions) which could be explained by a NS mutation in the
345 gene encoding MutS, another key component of the methyl-directed mismatch repair (MMR)
346 system (Table 3, Fig. 3). The mutation in this gene could explain the accumulation of a surprising
347 number of mutations in this isolate, which is likely a hypermutator with a characteristic
348 transversion bias. This contact contained no SNVs among the isolates sampled on days 4 and 6,
349 and we found this likely hypermutator isolate on day 7. However, the hypermutator was not
350 observed again at day eight, either due to the lower resolution in the detection of variants with the
351 WGS of cultured isolates, or the disappearance of this mutant from the population.
352
353 **Pan-genome analyses**
354 Whole-genome isolate sequencing also provides the opportunity to study variation in
355 gene content (the pangenome) within and between patients. We identified a total of 3478 core
356 genes common to all *V. cholerae* genomes, and 106 flexible genes present in some but not all
357 genomes (Fig. 3; Table S7). We also found an additional 251 genes present uniquely in isolate
358 56.00C4, assembled into one single contig identified as the genome of the lytic Vibrio phage
359 ICP1, which was assembled alongside the *Vibrio cholerae* genome. This phage contig contained
360 the ICP1 CRISPR/Cas system, which consists of two CRISPR loci (designated CR1 and CR2)
361 and six *cas* genes, as previously described (43, 44). These genes were excluded from subsequent
362 *V. cholerae* pan genome analyses.
363 Among the 106 flexible genes, some varied in presence/absence within a patient, ranging
364 from 36 to 88 genes gained or lost per patient (Table S7; Fig. 3). The majority of these flexible
365 genes (75%) were annotated as hypothetical, and several were transposases or prophage genes. A
366 large deletion of 24 genes was detected in patient 58.00, in an 18kb phage-inducible
367 chromosomal island (PICI) previously shown to prevent phage reproduction, and which is

368 targeted by the ICP1 CRISPR/Cas system (44). These PICI-like elements are induced during
369 phage infection, and interfere with phage reproduction via multiple mechanisms (45, 46). The
370 deletion of this PICI element in the *V. cholerae* genome may be a consequence of an ongoing
371 evolutionary arms race between *V. cholerae* and its phages.
372

373 **DISCUSSION**

374 Although within-patient *Vibrio cholerae* genetic diversity has been reported previously
375 (14, 15, 47, 48), our results confirmed that within-patient diversity is a common feature observed
376 in both symptomatic patients with cholera but also in a small sample of asymptomatically
377 infected individuals. In this study, we used a combination of metagenomic and WGS sequencing
378 technologies to characterize this within-patient diversity, revealing evidence for hypermutator
379 phenotypes in both symptomatic and asymptomatic infections.

380 In our previous study, we detected between zero and three iSNVs in cultured isolates
381 from patients with acute infection (15). In contrast, metagenomic analyses allowed us to detect
382 two iSNVs in the patient with the lowest level of diversity, but up to 207 iSNVs in another
383 individual (Table 1). In the only patient for which we were able to characterize *Vibrio cholerae*
384 intra-host diversity both from the metagenome and from cultured isolates, we did not identify any
385 iSNVs among five sequenced isolates, but detected 7 iSNVs from the metagenomic analyses.
386 This is could be due to false-positive iSNV calls inferred from metagenomes, but could also
387 represent a higher sensitivity to identify rare mutations (49). Our previous phylogenetic analysis
388 of *V. cholerae* isolate genomes also concluded that within-patient mutation was a more likely
389 source of variation than coinfection with multiple strains of *V. cholerae* (15). Although neither
390 isolate WGS or metagenomic analysis can fully exclude the possibility of coinfection (especially

391 involving very closely related strains), neither our earlier nor our present study provides strong
392 evidence for coinfection.

393 Despite its potential sensitivity to detect rare variants (26), metagenomics has limitations.
394 As already mentioned, some of the iSNVs inferred from metagenomes could be false-positives,
395 and this deserves further benchmarking. Within-sample diversity profiles cannot be established
396 for low-abundance microbes with insufficient sequence coverage (< 5X) and depth, and this level
397 of coverage is difficult to obtain in diverse microbial communities. In this study, only 48% of the
398 samples from patients with acute symptoms, known to harbor a high fraction of vibrios in their
399 stool (10^{10}-10^{12} vibrios per liter of stool), contained enough reads to reconstruct *Vibrio cholerae*
400 MAGs and to quantify within-patient diversity. Asymptomatic patients typically shed even less
401 *V. cholerae* in their stool (50), making it even more challenging to assemble their genomes using
402 metagenomics without depletion of host DNA or targeted sequence capture techniques (51, 52).
403

404 Hypermutation is defined as an excess of mutations due to deficiency in DNA mismatch
405 repair, and hypermutator strains have been described in diverse pathogenic infections and *in vivo*
406 experiments, including *Pseudomonas aeruginosa, Haemophilus influenzae* and *Streptococcus*
407 *pneumoniae* in cystic fibrosis patients, or *E. coli* in diverse habitats (17, 34, 53). In *Vibrio*
408 *cholerae,* a previous study of 260 clinical isolate genomes identified 17 isolates with an unusually
409 high number of SNPs uniformly distributed along the genome (24). Most of these genomes
410 contained mutations in one or more of four genes (*mutS, mutH, mutL* and *uvrD*) that play key
411 roles in DNA mismatch repair (24). These authors cautiously suggested that this apparent high
412 frequency of hypermutators could be associated with the rapid spread of the seventh cholera
413 pandemic, particularly because hypermutators may be a sign of population bottlenecks and recent
414 selective pressure. However, they also hypothesized that these high mutation rates could be

415 artefactual because the *V. cholerae* isolates had been maintained in stab cultures for many years.
416 It thus remains unclear if a hypermutator phenotype was derived within patients or during culture
417 (24, 25). Using our metagenomic approach, we provide evidence that hypermutators can indeed
418 emerge during infection, because DNA was extracted directly from patient samples without a
419 culture step in which mutations could have occurred during DNA replication. Using culture-
420 based WGS, using only a brief overnight culture, we report further evidence that hypermutators
421 occur in asymptomatic patients as well. An alternative explanation would be that hypermutators
422 evolved very recently prior to infection – a possibility which is difficult to exclude or test. Even if
423 the number of iSNVs is somewhat inflated by metagenomic sequencing, the concordance of
424 known DNA repair mutations with a transversion-skewed mutation profile is consistent with
425 current knowledge of hypermutation, and not easily explained by sequencing errors alone. Future
426 work will be required to determine any impacts of hypermutation on cholera disease severity or
427 transmission.

428 Hypermutator phenotypes are believed to be advantageous for the colonization of new
429 environments or hosts, allowing the hypermutator bacteria to generate adaptive mutations more
430 quickly, which leads to the more efficient exploitation of resources or increased resistance to
431 environmental stressful conditions, such as antibiotics (17, 20, 34, 53). However, this high
432 mutation rate can have a negative impact on fitness on the long term, with most of the mutations
433 being neutral or deleterious (20, 22, 54). A mouse model study showed that hypermutation can be
434 an adaptive strategy for *V. cholerae* to resist host-produced reactive oxygen induced stress, and
435 lead to a colonization advantage by increased catalase production and increased biofilm
436 formation (23). In our study of convergent evolution, we found no evidence for adaptive
437 mutations in the hypermutators. This could be because the signal from a small number of
438 adaptive mutations are obscured by overwhelming noise from a large number of neutral or

439 deleterious mutations. Further work is therefore needed to determine if *V. cholerae* mutators
440 produce adaptive mutations during human infection.
441
442 In contrast, we did find evidence for an excess of convergent mutations occurring
443 independently in the same genes in different patients, suggesting parallel adaptation in non-
444 mutator *V. cholerae* infections. Specifically, two patients contained mutations in the same
445 hemolysin gene, *hlyA*, which codes for a toxin that has both vacuolating and cytocidal activities
446 against a number of cell lines, including human intestinal cells (55), and is known to be an
447 important virulence factor in *Vibrio cholerae* El Tor O1 and a major target of immune responses
448 during acute infection (56, 57). Previous studies of within-patient *V. cholerae* evolution did not
449 identify mutations in *hlyA*, and instead identified different mutations possibly under selection for
450 biofilm formation (15) or phage resistance phenotypes (14). This lack of concordance could be
451 explained by relatively modest sample sizes of cholera patients in these studies, but could also
452 suggest that selective pressures may be idiosyncratic and person-specific across *Vibrio cholerae*
453 infections.
454 In conclusion, our results illustrate the potential and limitations of metagenomics as a
455 culture-independent approach for the characterization of within-host pathogen diversity. We also
456 provide evidence that hypermutators emerge within human *V. cholerae* infection, and their
457 evolutionary dynamics and relevance to disease progression merits further study.
458

459 ## MATERIALS AND METHODS

460 **Sample collection, clinical outcomes and metagenomic sequencing**

461 To study within-host diversity of *V. cholerae* during infection, we used stool and rectal swab
462 samples collected from cholera patients admitted to the icddr,b (International Center for Diarrheal

463 Disease Research, Bangladesh) Dhaka Hospital, and from their household contacts, as previously
464 described (12). Patients present to the icddr,b year-round with cholera, and cases peak during
465 biannual floods (58, 59). Index cases were defined as patients presenting to the hospital with
466 severe acute diarrhea and a stool culture positive for *V. cholerae*. Individuals who shared the
467 same cooking pot with an index patient for three or more days are considered household contacts
468 and were enrolled within 6 hours of the presentation of the index patient to the hospital. Rectal
469 swabs were collected each day during a ten-day follow up period after presentation of the index
470 case. Household contacts underwent daily clinical assessment of symptoms and collection of
471 blood for serological testing. Contacts were determined to be infected if any rectal swab culture
472 was positive for *V. cholerae* or if the contact developed diarrhea and a 4-fold increase in
473 vibriocidal titer during the follow-up period (10, 11). If they developed watery diarrhea during
474 the follow up period, contacts with positive rectal swabs were categorized as symptomatic and
475 those without diarrhea were considered asymptomatic. We excluded patients with age below two
476 and above 60 years old, or with major comorbid conditions (10, 11).
477
478 Fecal samples and rectal swabs from the day of infection and follow up timepoints were
479 collected and immediately placed on ice after collection and stored at -80°C until DNA
480 extraction. DNA extraction was performed with the PowerSoil DNA extraction kits (Qiagen)
481 after pre-heating to 65°C for 10 min and to 95°C for 10 min. Sequencing libraries were
482 constructed for 33 samples from 31 patients, for which we obtained enough DNA. We used the
483 NEBNext Ultra II DNA library prep kit and sequenced the libraries on the Illumina HiSeq 2500
484 (paired-end 125 bp) and the Illumina NovaSeq 6000 S4 (paired-end 150 bp) platforms at the
485 Genome Québec sequencing platform (McGill University).
486

487 **Metagenomic analyses**

488 ***Sequence preprocessing and assembly***

489 Sequencing fastq files were quality checked with FastQC

490 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We removed human and technical

491 contaminant DNA by aligning reads to the PhiX genome and the human genome (hg19) with

492 Bowtie2 (60) , and used the iu-filter-quality-minoche script of the illumina-utils program with

493 default parameters to filter the reads (61).

494

495 ***Taxonomic assignment***

496 Processed paired-end metagenomics sequences were classified using two taxonomic

497 profilers: Kraken2 v.2.0.8_beta (a k-mer matching algorithm) (62) and MIDAS v.1.3.0 (a read

498 mapping algorithm) (63). Kraken 2 examines the k-mers within a query sequence and uses the

499 information within those k-mers to query a database, then maps k-mers to the lowest common

500 ancestor (LCA) of all genomes known to contain a given k-mer. Kraken2 was run against a

501 reference database containing all RefSeq viral, bacterial and archaeal genomes (built in May

502 2019), with default parameters. MIDAS uses a panel of 15 single-copy marker genes present in

503 all of ~31,000 bacterial species included in its database to perform taxonomic classification, and

504 maps metagenomic reads to this database to estimate the read depth and relative abundance of

505 5,952 bacterial species. We identified metagenomic samples containing *V. cholerae* and

506 vibriophage reads, and computed the mean depth of coverage (number of reads per base-pair) of

507 the *V. cholerae* pangenome in the MIDAS database (Table 1).

508

509 ***Assembly and binning of Vibrio cholerae genomes***

510 To recover good quality metagenome-assembled genomes (MAGs) of *V. cholerae,* we
511 selected metagenomic samples with coverage >10X against the *V. cholerae* pangenome in the
512 MIDAS database, and used MEGAHIT v.1.2.9 (64) to perform *de novo* assembly. For 9 of the 11
513 selected samples, we independently assembled the genome of each sample, and co-assembled the
514 two remaining samples, which belong to the same patient (a symptomatic infected contact on
515 days 9 and 10). Contigs of <1.5 kb were discarded.

516 We extracted MAGs by binning of our metagenomic assemblies. Because no single
517 binning approach is superior in every case, with performance of the algorithms varying across
518 samples, we used different binning tools to recover MAGs. The quality of a metagenomic bin is
519 evaluated by its completeness (the level of coverage of a population genome), and the
520 contamination level (the amount of sequence that does not belong to this population from another
521 genome). These metrics can be estimated by counting the frequency of single-copy marker genes
522 within each bin (65). We inferred bins using CONCOCT v.1.1.0 (66), MaxBin 2 v.2.2.7 (67) and
523 MetaBAT 2 v.2.12.1 (68), with default parameters. We then used DAS_Tool v.1.1.1 on the
524 results of these three methods, to select a single set of non-redundant, high-quality bins per
525 sample (69). DAS_Tool is a bin consolidation tool, which predicts single-copy genes in all the
526 provided bin sets, aggregates bins from the different binning predictions, and extracts a more
527 complete consensus bin from each aggregate such that the resulting bin has the most single-copy
528 genes while having a reasonably low number of duplicate genes (69). We then used Anvi'o v.6.1
529 (70) to manually refine the bins with contamination higher than 10% and Centrifuge v.1.0.4_beta
530 (71) to determine the taxonomy of all bins in each sample, in order to identify *V. cholerae*
531 MAGs.

532 Bins with completeness > 60% and contamination <10% were first selected, and those
533 assigned to *V. cholerae* were further filtered (completeness > 90% and contamination <1% for

534 the *V. cholerae* bins). We dereplicated the entire set of bins with dRep v.2.2.3 using a minimum
535 completeness of 60%, the ANImf algorithm, 99% secondary clustering threshold, maximum
536 contamination of 10%, and 25% minimum coverage overlap, and obtained 79 MAGs displaying
537 the best quality and representing individual metagenomic species (MGS).
538
539 ***Detection of Vibrio cholerae genetic diversity within and between metagenomic samples***
540 We created a bowtie2 index of the 79 representative genomes from the dereplicated set,
541 including a single high-quality *Vibrio cholerae* MAG, and mapped reads from each sample to this
542 set. By including many diverse microbial genomes in the bowtie2 index, we aimed to avoid the
543 mismapping of reads from other species to the *V. cholerae* genome, and to reduce potential false
544 positive intra-host single nucleotide variant (iSNV) calls. As recommended, we used *Vibrio*
545 *cholerae* MAGs from the samples we under study rather than a genetically distant reference, as
546 read mapping to the most closely-related genome available is expected to reduce the rate of false
547 positive iSNV calls (72). We mapped the metagenomics reads of each sample with a *V. cholerae*
548 coverage value >5X (obtained with MIDAS) against the set of 79 MAGs, using Bowtie2 (60)
549 with the --very-sensitive parameters. We also used Prodigal (73) on the concatenated MAGs, in
550 order to predict open reading frames using default metagenomic settings.
551 We then used inStrain on the 15 selected samples
552 (https://instrain.readthedocs.io/en/latest/index.html). This program aims to identify and compare
553 the genetic heterogeneity of microbial populations within and between metagenomic samples
554 (27). "InStrain profile" was run on the mapping results, with the minimum percent identity of
555 read pairs to consensus set to 99%, minimum depth of coverage to call a variant of 5X, and
556 minimum allele frequency to confirm a SNV equal to 0.05. All non-paired reads were filtered
557 out, as well as reads with an identity value below 0.99. Coverage and breadth of coverage

558 (percentage of reference base pairs covered by at least one read) were computed for each
559 genome. InStrain identified both bialllelic and multiallelic SNV frequencies at positions where
560 phred30 quality filtered reads differ from the reference genome and at positions where multiple
561 bases were simultaneously detected at levels above the expected sequencing error rate. SNVs
562 were classified as non-synonymous, synonymous, or intergenic based on gene annotations, and
563 gene functions were recovered from the Uniprot database (74) and BLAST (75). Then, similar
564 filters to those described in (28) were applied to the detected SNVs. We excluded from the
565 analysis positions with very low or high coverage value D compared to the median coverage $\overline{D}$,
566 and positions within 100 bp of contig extremities. As sites with very low coverage could result
567 from a bias in sequencing or library preparation, and sites with higher coverage could arise from
568 mapping error or be the result of repetitive region or multi-copy genes not well assembled, we
569 masked sites in all the samples if D was $< 0.3\overline{D}$ and if D was $> 3\overline{D}$ in at least two samples.
570

571 ***Mutation spectrum of hypermutator and non-mutator samples***

572 For each sample, iSNVs were categorized into six mutation types based on the chemical
573 nature of the nucleotide changes (transitions or transversions). We combined all the samples with
574 hypermutators and compared them to the mutation spectrum of the non-mutators. The mutation
575 spectrum was significantly different between the hypermutator samples and the non-
576 hypermutator samples (Chi-square test, $P<0.01$). We then computed the mutation mean and
577 standard error of each of the six mutation types and compared the two groups (Fig. 2C).
578

579 ***Bacterial replication rate estimation***

580 Replication rates were estimated with the metric iRep (index of replication), which is
581 based on the measurement of the rate of the decrease in average sequence coverage from the

582 origin to the terminus of replication. iRep values (39) were calculated by mapping the sequencing
583 reads of each sample to the *V. cholerae* MAG assembled from that sample.
584
585 ***Tests for natural selection***
586 First, we identified signals of convergent evolution in the form of nonsynonymous iSNVs
587 occurring independently in the same gene in multiple patients. To assess the significance of
588 convergent mutations, we compared their observed frequencies to expected frequencies in a
589 simple permutations model. We ran separate permutations for non-mutators (two genes with
590 convergent mutations in at least two out of eight non-mutator samples, including only one time
591 point from the patient sampled twice, and excluding the outlier patient A with a large number of
592 intergenic iSNVs) and possible hypermutators (five genes with convergent mutations in at least
593 two out of five possible hypermutator samples). In each permutation, we randomized the
594 locations of the nonsynonymous mutations, preserving the observed number of nonsynonymous
595 mutations in each sample, and the observed distribution of gene lengths. For simplicity, we
596 assumed that 2/3 of nucleotide sites in coding regions were nonsynonymous. We repeated the
597 permutations 1000 times and estimated a *P*-value as the fraction of permutations yielding greater
598 than or equal to the observed number of genes mutated in two or more samples.
599 Second, we compared natural selection at the protein level within versus between patients,
600 using the McDonald-Kreitman test (41). We again considered hypermutators separately. Briefly,
601 the four counts (Pn, Ps, Dn, Ds) of between-patient divergence (D) vs. within-patient
602 polymorphism (P), and non-synonymous (n) vs synonymous (s) mutations were computed and
603 tested for neutrality using a Fisher's exact test (FDR corrected *P*-values < 0.05).
604
605 **Whole genome sequencing analyses**

606 ***Culture of Vibrio cholerae isolates***

607 We selected three of the households with asymptomatic infected contacts (households 56,
608 57, and 58) for within-patient diversity analysis using multiple *V. cholerae* colonies per
609 individual. Each index case was sampled on the day of presentation to the icddr,b, and
610 asymptomatic contacts positive for *V. cholerae* were sampled on the following day, except for
611 one contact (household 58, contact 02). This individual was only positive on day 4 following
612 presentation of the index case, and we collected samples and cultured isolates from day 4 to day
613 8. Stool samples collected from three index cases and their respective infected contacts were
614 streaked onto thiosulfate-citrate-bile salts-sucrose agar (TCBS), a medium selective for *V.*
615 *cholerae*. After overnight incubation, individual colonies were inoculated into 5 ml Luria-Bertani
616 broth and grown at 37 °C overnight. For each colony, 1 ml of broth culture was stored at −80 °C
617 with 30% glycerol until DNA extraction. We used the Qiagen DNeasy Blood and Tissue kit,
618 using 1.5 ml bacteria grown in LB media, to extract the genomic DNA. In order to obtain pure
619 gDNA templates, we performed a RNase treatment followed by a purification with the MoBio
620 PowerClean Pro DNA Clean-Up Kit.

621

622 ***Whole genome sequencing and preprocessing***

623 We prepared 48 sequencing libraries using the NEBNext Ultra II DNA library prep kit
624 (New England Biolabs) and sequenced them on the Illumina HiSeq 2500 (paired-end 125 bp)
625 platform at the Genome Québec sequencing platform (McGill University). Sequencing fastq files
626 were quality checked with FastQC, and Kraken2 was used to test for potential contamination with
627 other bacterial species (62).

628

629 ***Variant calling and phylogeny***

630 We mapped the reads for each sample to the MJ-1236 reference genome and called single
631 nucleotide polymorphisms (SNPs, fixed within patients) and single nucleotide variants (SNVs,
632 variable within patients) using Snippy v.4.6.0 (76), with default parameters. A concatenated
633 alignment of these core variants was generated, and an unrooted phylogenic tree was inferred
634 using maximum parsimony (MP) in MEGA X (77). The percentage of replicate trees in which the
635 associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the
636 branches. The MP tree was obtained using the Subtree-Pruning-Regrafting (SPR) algorithm with
637 search level 1 in which the initial trees were obtained by the random addition of sequences (10
638 replicates).
639
640 ***De novo assembly and pan genome analyses***
641 We *de novo* assembled genomes from each isolate using SPAdes v.3.14 on the short reads, with
642 default parameters (78) and used Prokka v1.14.6 (79) to annotate them. We constructed a pan-
643 genome from the resulting annotated assemblies combining Roary v.3.13.0 (80) and GenAPI
644 (81), identifying genes present in all isolates (core genome) and genes only present in some
645 isolates (flexible genome). The flexible genome and the phylogenetic tree were visualized with
646 Phandango v.1.1.0 (82).
647
648 **DATA AVAILABILITY**
649 All metagenomic sequence data are available in NCBI GenBank under BioProject
650 PRJNA668607, and isolate genome sequences under BioProject PRJNA668606.
651
652

FUNDING INFORMATION

This study was supported by a Canadian Institutes of Health Research Operating Grant to BJS, the icddr,b Centre for Health and Population Research, grants AI103055 (J.B.H and F.Q), AI106878 (E.T.R and F.Q.), AI058935 (E.T.R, S.B.C and F.Q.), T32A1070611976 and K08AI123494 (A.A.W.), and Emerging Global Fellowship Award TW010362 (T.R.B.), from the National Institutes of Health, and the Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program (R.C.C.).

ACKNOWLEDGEMENTS

We are grateful to the people of Dhaka where our study was undertaken; to the field, laboratory and data management staff who provided a tremendous effort to make the study successful; and to the people who provided valuable support in our study. The icddr,b gratefully acknowledges the Government of the People's Republic of Bangladesh; Global Affairs Canada (GAC); Swedish International Development Cooperation Agency (Sida) and the Department for International Development, (UKAid). We declare that we have no competing financial interest.

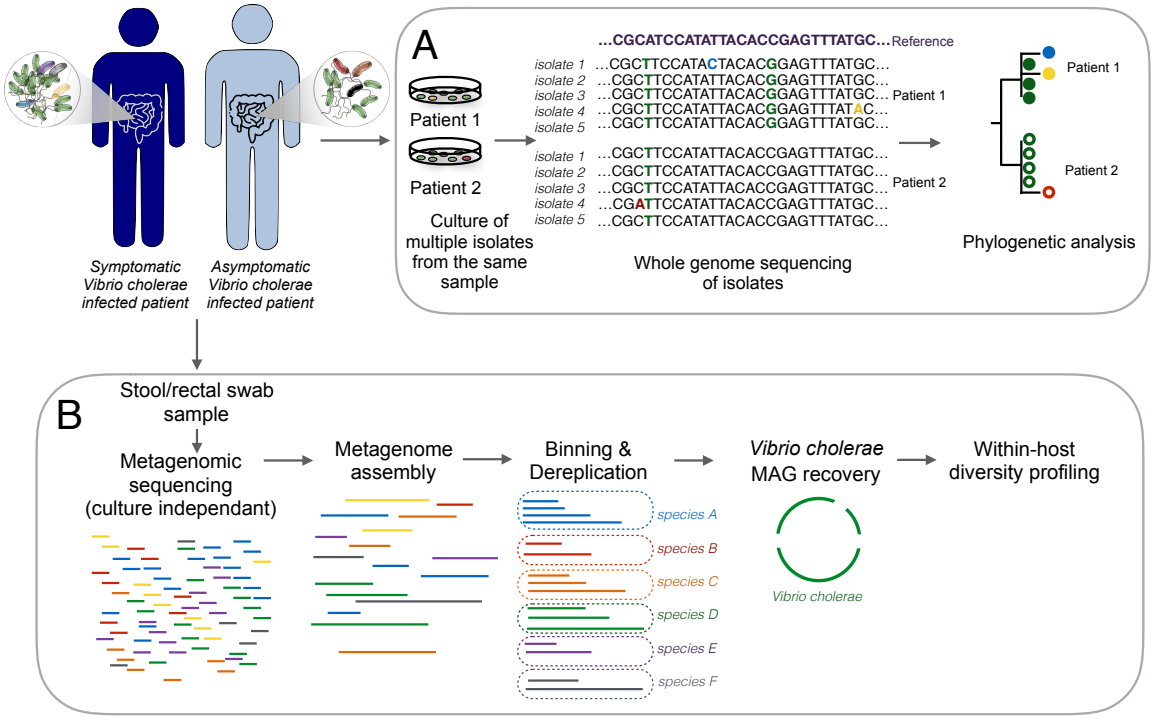
ETHICAL STATEMENT

The Ethical and Research Review Committees of the icddr,b and the Institutional Review Board of MGH reviewed the study. All adult subjects provided informed consent and parents/guardians of children provided informed consent. Informed consent was written.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

678 ## FIGURES AND TABLES



679
680 **Figure 1. Summary of the culture-dependent and the culture-free metagenomics workflows**
681 **for the characterization of the *Vibrio cholerae* within-patient diversity.** Stool or rectal swab
682 samples were collected from symptomatic and asymptomatic *Vibrio cholerae* infected individuals
683 and processed using two different approaches: (A) Culture, DNA extraction and whole genome
684 sequencing of multiple isolates per patient; (B) Genome-resolved metagenomics involves DNA
685 extraction directly from a microbiome sample followed by sequencing, assembly, genome
686 binning and dereplication to generate metagenome-assembled genomes (MAGs), and within-host
687 diversity profiling by mapping reads back to the MAGs.

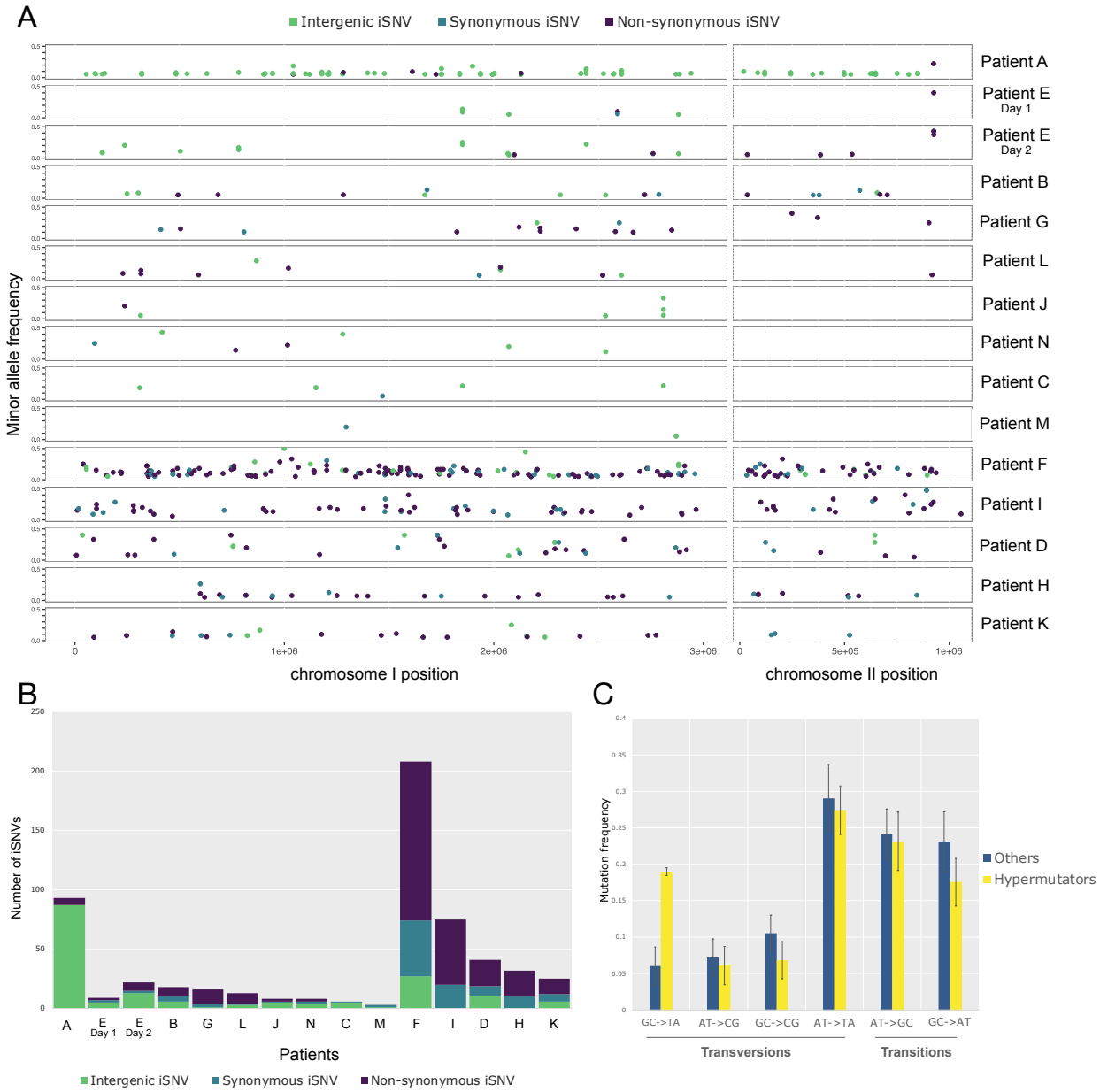
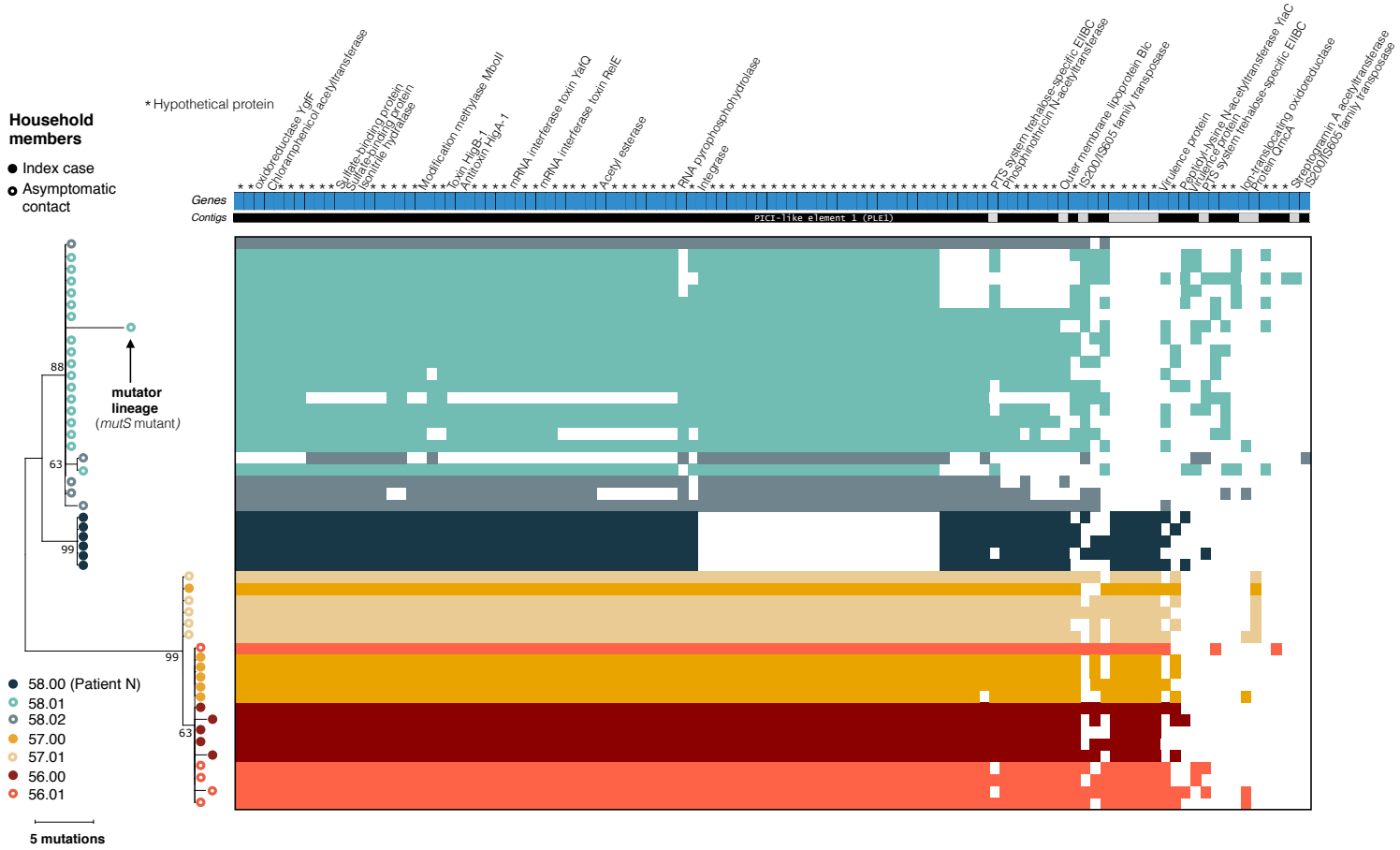


Figure 2. Within-patient *Vibrio cholerae* diversity quantified from metagenomic data. (A) Minor allele frequency and distribution of intergenic, synonymous and non-synonymous iSNVs across the two *Vibrio cholerae* chromosomes for 14 patients with over 5X coverage of the *V. cholerae* genome. (B) Numbers of intergenic, synonymous and non-synonymous iSNVs for each patient. (C) Spectrum of within-patient mutation in hypermutators compared to non-mutators.

694 Error bars represent standard error of the mean. Only samples with 6 or more iSNVs were
695 included to reduce noise from low counts.
696



697 **Figure 3. Phylogeny and pan-genome of 48 *Vibrio cholerae* isolates from index cases and**
698 **their asymptomatic contacts.** The phylogeny was inferred using Maximum Parsimony. The
699 percentage of replicate trees in which the associated taxa clustered together in the bootstrap test
700 (1000 replicates) are shown next to the branches. Filled circles represent isolates from index
701 cases and empty circles represent isolates from their asymptomatic contacts. The heatmap of gene
702 presence-absence is based on 106 genes in the flexible genome. Colored blocks in the heatmap
703 indicate gene presence; white indicates gene absence. Each row corresponds to an isolate from
704 the phylogenetic tree and each column represents an orthologous gene family. Each unique color
705 represents a different individual. Different contigs are represented as alternating blocks of black
706 and light grey.

Patient	Total number of iSNVs	Number of non-synonymous iSNVs	Number of synonymous iSNVs	Number of intergenic iSNVs	Mean coverage	iRep value	DNA repair and proofreading genes with NS mutation
Patient A	93	6	0	87	451.3X	3.34	-
Patient B	18	7	5	6	111.4X	1.7	-
Patient C	6	0	1	5	111.8X	1.7	-
Patient D	41	22	9	10	10X	5.43	DNA polymerase II
Patient E day 1	8	2	1	5	351X	3.25	-
Patient E day 2	21	7	1	13	258X	1.23	-
Patient F	207	133	47	27	18.2X	2.48	DNA mismatch repair endonuclease MutL Nuclease SbcCD subunit C
Patient G	16	12	3	1	7.7X	1.73	-
Patient H	32	21	11	0	98.5X	4.75	Excinuclease ABC subunit UvrB
Patient I	75	55	20	0	13X	2.79	MutT/nudix family protein
Patient J	6	1	0	5	424.6X	1.84	-
Patient K	25	13	6	6	18X	1.69	Formamidopyrimidine-DNA glycosylase mutM
Patient L	13	9	1	3	164.4X	2.67	-
Patient M	2	0	1	1	113X	2.65	-
Patient N	7	2	1	3	6.7X	2.27	-

707 **Table 1. Within-patient *Vibrio cholerae* diversity profiles from 15 metagenomes.** Mutations
708 segregating within patients are denoted iSNVs. The number of iSNVs and mean coverage values
709 were computed with InStrain (27) and replication rate with iRep (39).

Protein (Uniprot ID)	Patient A	Patient B	Patient D	Patient E	Patient F	Patient H	Patient I	Patient K
Hemolysin (VC cytolysin) (P09545)	NS (0.22)	-	-	3 NS (0.22-0.43)	-	-	-	-
2-aminoethylphosphonate ABC transporter ferric-binding protein (Q9KLY8)	-	NS (0.05)	-	NS (0.05)	-	-	-	-
Peptidase B (Q9KTX5)	-	-	NS (0.33)	-	-	-	NS (0.09)	-
Nuclease SbcCD subunit C (Q9KM67)	-	-	S (0.28)	-	NS (0.09)	-	-	-
C4-dicarboxylate transport sensor protein (Q9KN25)	-	-	-	-	NS (0.08)	-	NS (0.11)	-
zinc/cadmium/mercury/lead-transporting ATPase (Q9KT72)	-	-	-	-	NS (0.08)	-	-	NS (0.06)
hypothetical protein (A0A0H3AI44)	-	-	-	-	NS (0.14)	-	-	NS (0.14)
hypothetical protein (Q9KLL1)	-	-	-	-	NS (0.33)	NS (0.11)	-	-
Formamidopyrimidine-DNA glycosylase mutM (C3LQI3)	-	-	-	-	S (0.18)	-	-	NS (0.08)
Phosphoribosylformylglycinamidine synthase (Q9KTN2)	-	-	-	-	-	-	NS (0.06)	S (0.08)

710

711 **Table 2. Set of genes with convergent mutations identified in more than one patient.** The
712 presence of a synonymous or non-synonymous iSNV in each gene and each patient is indicated
713 with S or NS, respectively, and the minor allele frequency is shown in parentheses. None of the
714 mutations were found at the same nucleotide or codon position. Patients containing likely
715 hypermutators are underlined. Only genes and patients containing more than one mutated gene
716 are shown.

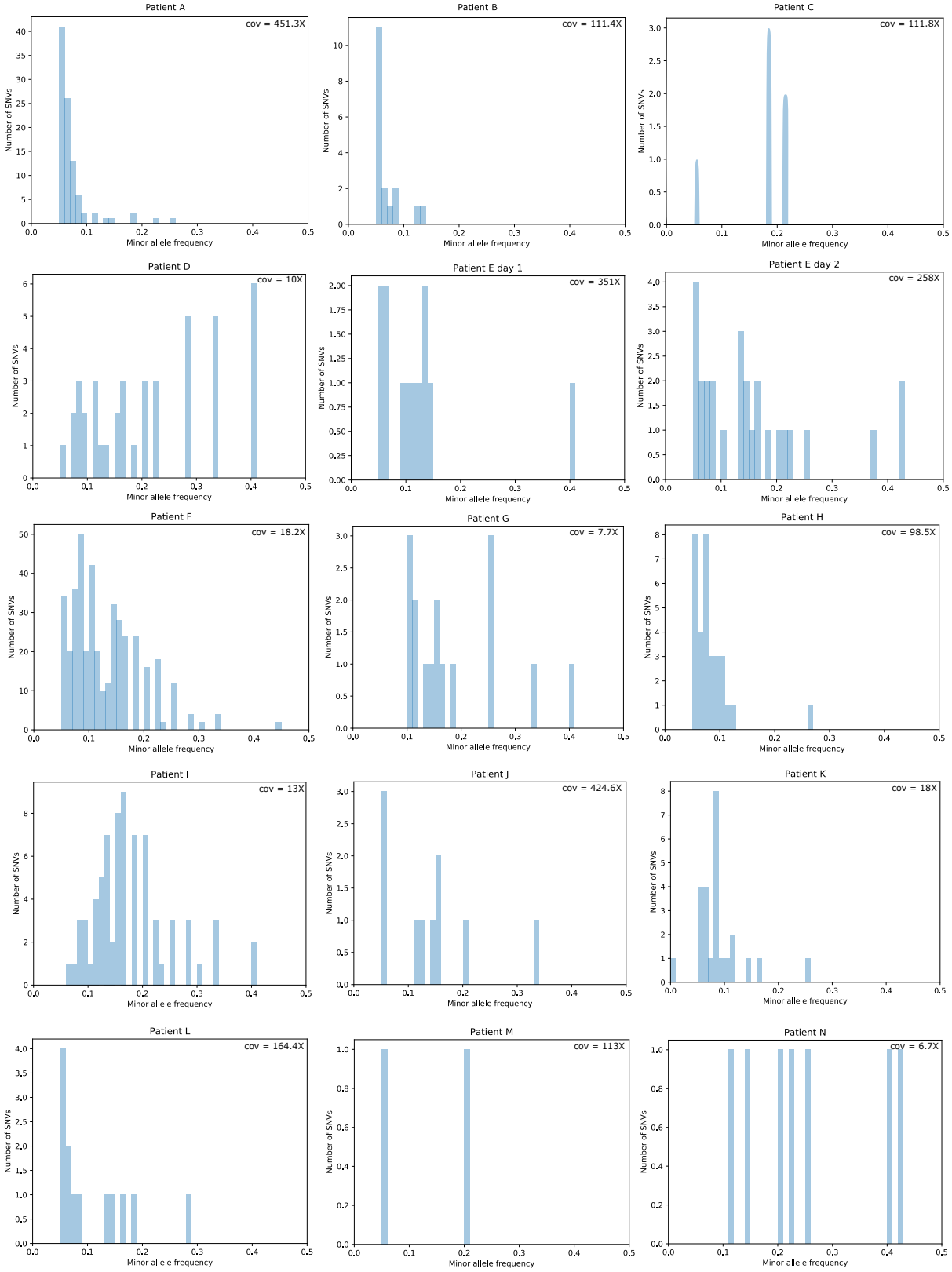
Type	Isolates	Mutation type	Nucleotide position in MJ-1236	Ref. nucleotide	Alt. nucleotide	Gene annotation	Metagenomic samples with same variant
iSNV	58.01d7C1	NS	Chr1,53054	G	A	DNA mismatch repair protein MutS	-
SNP	Households 56 and 57	S	Chr1, 198988	G	A	MSHA biogenesis protein MshQ	-
iSNV	58.01d7C1	NS	Chr1, 209665	G	A	MSHA biogenesis protein MshN	-
iSNV	56.00C4	NS	Chr1,374172	C	T	UDP-N-acetylglucosamine 4,6-dehydratase	-
SNP	Household 58	NS	Chr1,410638	G	A	Phosphopantetheine adenylyltransferase	Patients M, N
SNP	Households 56 and 57	NS	Chr1,754154	C	T	1,4-dihydroxy-2-naphthoate polyprenyltransferase	-
SNP	Household 58	S	Chr1,841538	C	T	SSU ribosomal protein S4p	Patients L, M, N
SNP	Household 58	S	Chr1,1315021	T	G	Exported zinc metalloprotease YfgC precursor	Patients L, M, N
iSNV	58.02C1	S	Chr1,1576083	C	A	Periplasmic thiol:disulfide oxidoreductase DsbB	-
SNP	Patient 58.00	NS	Chr1,1689779	A	C	Sigma-54 dependent transcriptional regulator	-
SNP	Contacts 58.01 and 58.02	NS	Chr1,2301641	G	A	Putative membrane protein	-
iSNV	58.01d7C1	NS	Chr1,1744854	C	T	Hypothetical protein	-
SNP	Contacts 58.01 and 58.02	NS	Chr1,2262202	A	G	Serine transporter	-
SNP	Households 56 and 57	NS	Chr1,2301641	C	T	LacI family DNA-binding transcriptional regulator	Patients D, J, K
iSNV	57.00C5	NS	Chr1,2509468	C	T	cyclic-di-GMP-modulating response regulator	-
iSNV	56.01C1	NS	Chr1,2588496	C	T	Amidophosphoribosyltransferase	-
iSNV	58.01d7C1	NS	Chr1,2693815	C	T	PTS system, trehalose-specific IIB component	-
SNP	Household 58	NS	Chr1,2806858	A	T	Citrate lyase alpha chain	Patients L, M, N
iSNV	56.00C1	S	Chr1,3037471	A	G	Hypothetical protein	-
SNP	Patient 58.00	NS	Chr1,3059131	C	T	DNA polymerase V (UmuC)	-
SNP	Households 56 and 57	NS	Chr1,3095039	G	A	Outer membrane protein OmpU	Patients D, F, G, I ,J, K
SNP	Contacts 58.01 and 58.02	S	Chr1,3105102	C	T	Glutamate-1-semialdehyde aminotransferase	-
iSNV	58.01d7C1	NS	Chr1,528409	C	T	Vibriolysin, extracellular zinc protease	-

717 **Table 3. Nucleotide changes identified in core genes of the *V. cholerae* isolates from index**

718 **cases (56.00, 57.00 and 58.00) and their asymptomatic contacts.** Genome position is according

719 to the MJ-1236 reference genome (CP001485.1, CP001486.1). Mutations segregating within
720 patients are denoted iSNVs; mutations fixed between patients are denoted SNP. Patient allele
721 frequency shows the allele frequency of the alternative (minor) allele. Ref=Reference allele;
722 Alt=Alternative allele. NS=non-synonymous; S=synonymous. Chr1=chromosome 1;
723 Chr2=chromosome
724

725 **SUPPLEMENTARY FIGURES**



726 **Figure S1. Minor allele frequency distributions for iSNVs in 15 metagenomic samples.**

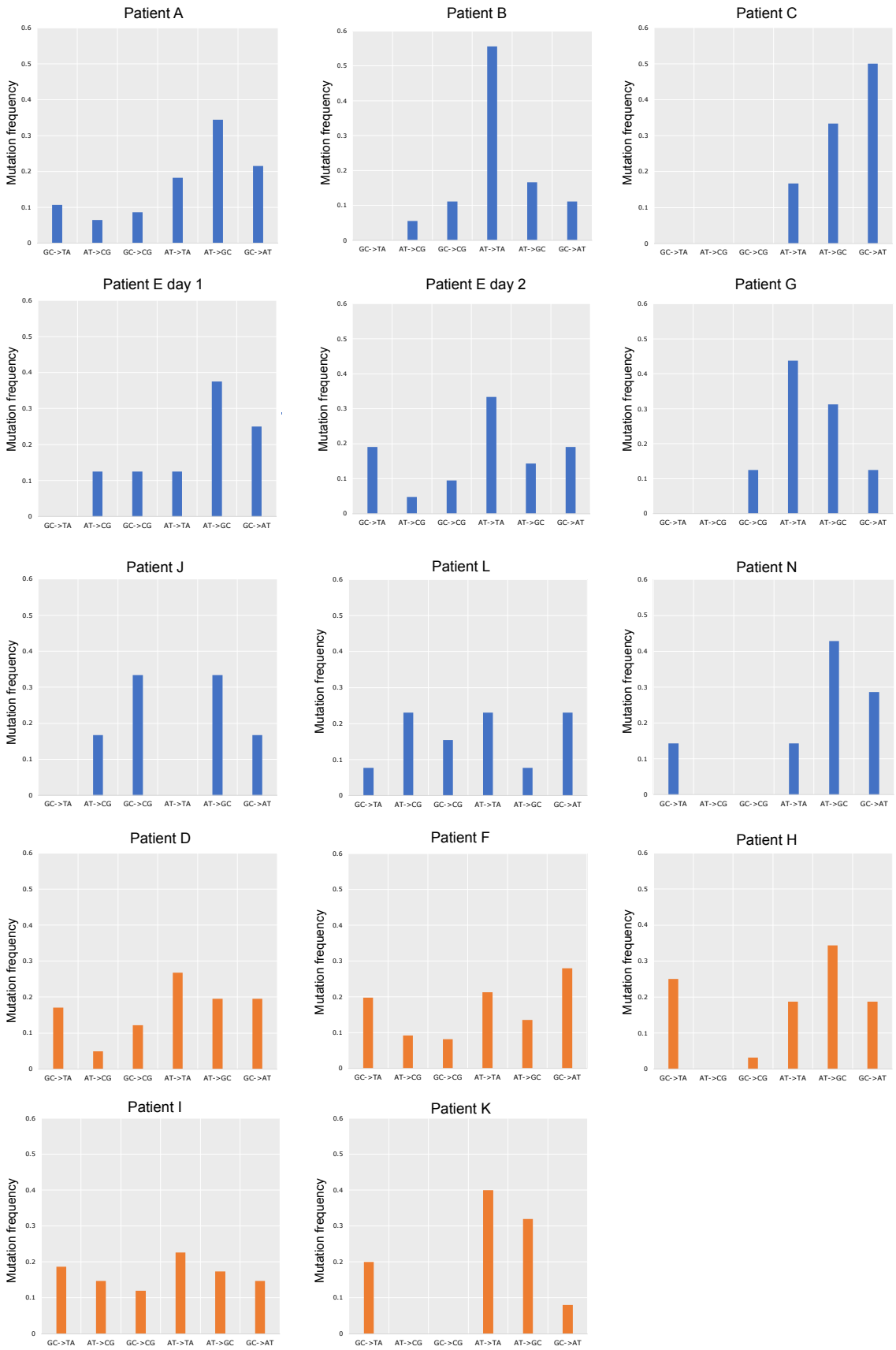
727 Allele frequencies and mean coverage values (cov) were computed with Instrain (27)

728

729

730

731


Patient A
Patient B
Patient C
Patient E day 1
Patient E day 2
Patient G
Patient J
Patient L
Patient N
Patient D
Patient F
Patient H
Patient I
Patient K
Mutation frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
GC->TA
AT->CG
GC->CG
AT->TA
AT->GC
GC->AT

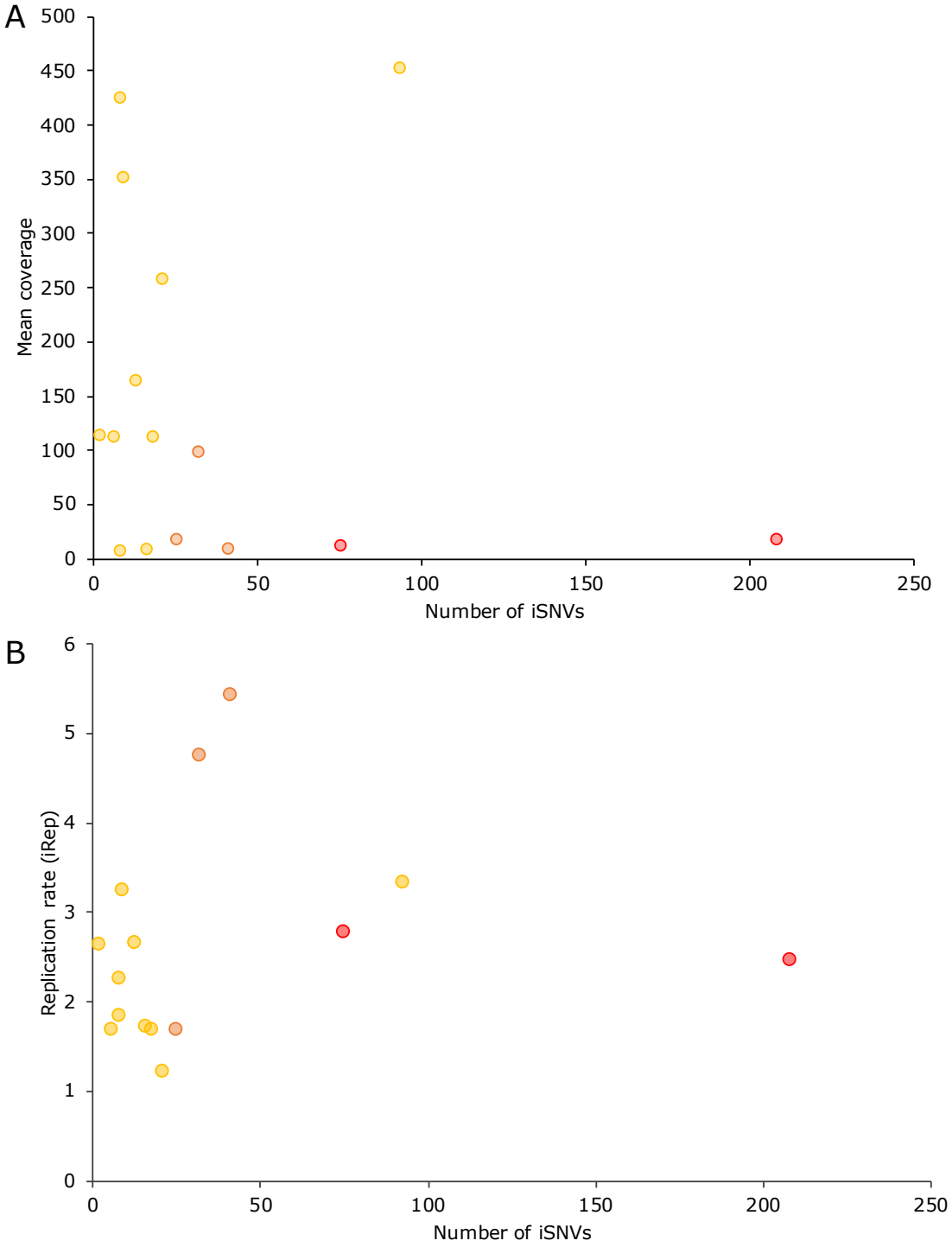
732

733 **Figure S2.** Spectrum of within-patient mutation (iSNVs) in patients with potential hypermutators

734 (orange; bottom two rows) compared to others (blue; top three rows). Only samples with 6 or

735 more iSNVs were included to reduce noise from low counts.

736



737
738
739 **Figure S3. iSNVs numbers are not linked to average depth of read coverage (A) or**
740 **replication rate (B).** Patients with strong potential hypermutation phenotypes are represented in
741 red, weak hypermutation phenotypes in orange, and other samples in yellow.

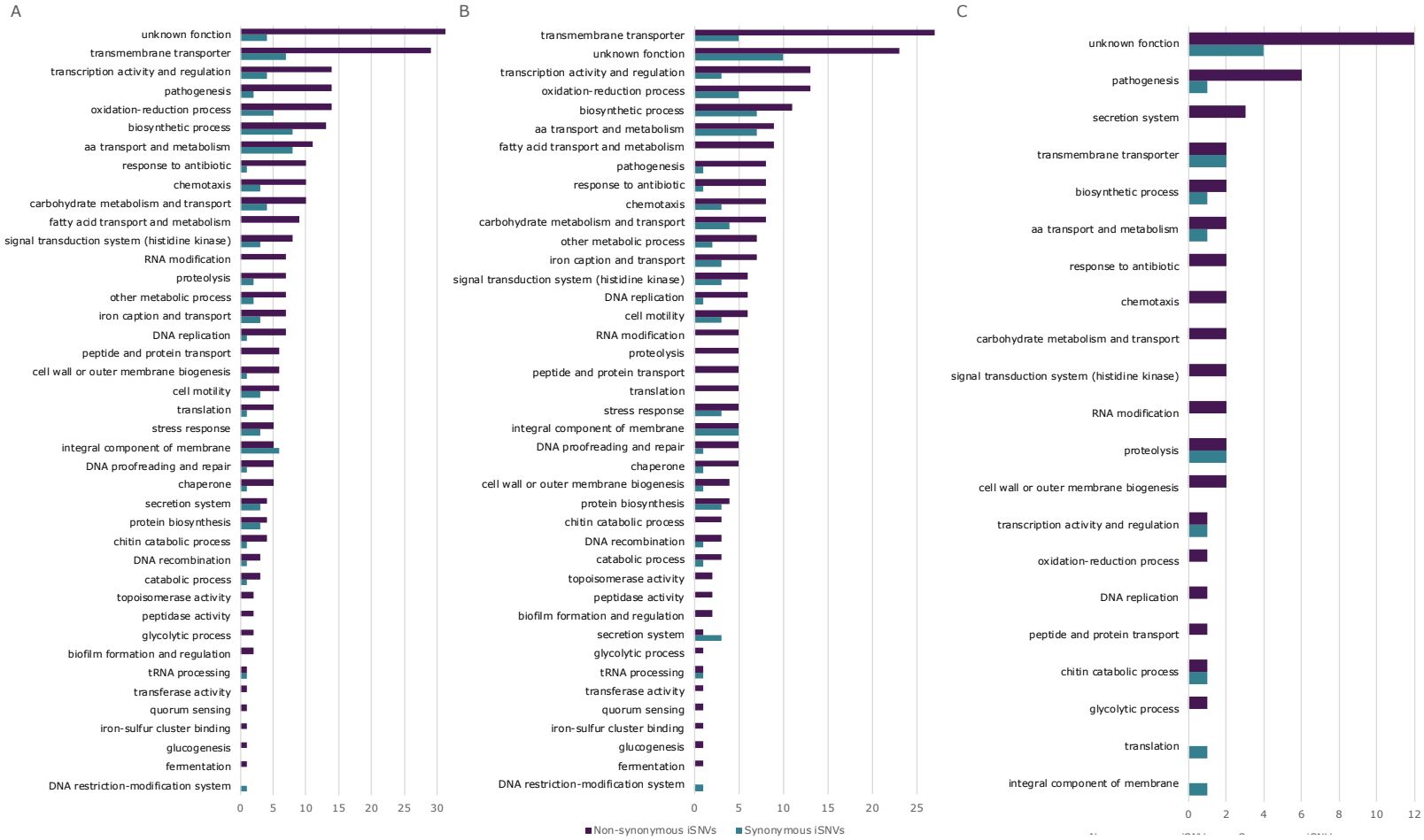


Figure S4. Functional categories of genes containing within-patient variants (iSNVs). The total number of synonymous and non-synonymous iSNVs in different functional categories was represented for all patients (A), for the 5 patients with hypermutation phenotypes (B), and for the 9 patients without hypermutation phenotypes (C). Categories are ranked in descending order of the number of non-synonymous iSNVs.

762
763 ## REFERENCES
764

765 1. Ali M, Nelson AR, Lopez AL, Sack DA. 2015. Updated Global Burden of Cholera in
766 Endemic Countries. PLoS Negl Trop Dis 9:e0003832.

767 2. Weil AA, Ivers LC, Harris JB. 2011. Cholera: Lessons from Haiti and Beyond. Current
768 Infectious Disease Reports 14:1–8.

769 3. Camacho A, Bouhenia M, Alyusfi R, Alkohlani A, Naji MAM, de Radiguès X, Abubakar
770 AM, Almoalmi A, Seguin C, Sagrado MJ, Poncin M, McRae M, Musoke M, Rakesh A,
771 Porten K, Haskew C, Atkins KE, Eggo RM, Azman AS, Broekhuijsen M, Saatcioglu MA,
772 Pezzoli L, Quilici M-L, Al-Mesbahy AR, Zagaria N, Luquero FJ. 2018. Cholera epidemic in
773 Yemen, 2016–18: an analysis of surveillance data. The Lancet Global Health 6:e680–e690.

774 4. Satchell KJF, Jones CJ, Wong J, Queen J, Agarwal S, Yildiz FH. 2016. Phenotypic Analysis
775 Reveals that the 2010 Haiti Cholera Epidemic Is Linked to a Hypervirulent Strain. Infection
776 and Immunity 84:2473–2481.

777 5. Weil AA, Ryan ET. 2018. Cholera: recent updates. Current Opinion in Infectious Diseases
778 31:455–461.

779 6. Domman D, Chowdhury F, Khan AI, Dorman MJ, Mutreja A, Uddin MI, Paul A, Begum
780 YA, Charles RC, Calderwood SB, Bhuiyan TR, Harris JB, LaRocque RC, Ryan ET, Qadri
781 F, Thomson NR. 2018. Defining endemic cholera at three levels of spatiotemporal
782 resolution within Bangladesh. 7. Nature Genetics 50:951–955.

783 7. King AA, Ionides EL, Pascual M, Bouma MJ. 2008. Inapparent infections and cholera
784 dynamics. 7206. Nature 454:877–880.

785 8. Phelps MD, Simonsen L, Jensen PKM. 2019. Individual and household exposures
786 associated with cholera transmission in case–control studies: a systematic review. Tropical
787 Medicine & International Health 24:1151–1168.

788 9. Harris JB, Khan AI, LaRocque RC, Dorer DJ, Chowdhury F, Faruque ASG, Sack DA, Ryan
789 ET, Qadri F, Calderwood SB. 2005. Blood Group, Immunity, and Risk of Infection with
790 *Vibrio cholerae* in an Area of Endemicity. Infection and Immunity 73:7422–7427.

791 10. Harris JB, LaRocque RC, Chowdhury F, Khan AI, Logvinenko T, Faruque ASG, Ryan ET,
792 Qadri F, Calderwood SB. 2008. Susceptibility to *Vibrio cholerae* Infection in a Cohort of
793 Household Contacts of Patients with Cholera in Bangladesh. PLoS Negl Trop Dis 2.

794 11. Weil AA, Khan AI, Chowdhury F, LaRocque RC, Faruque ASG, Ryan ET, Calderwood SB,
795 Qadri F, Harris JB. 2009. Clinical Outcomes in Household Contacts of Patients with
796 Cholera in Bangladesh. Clin Infect Dis 49:1473–1479.

797 12. Midani FS, Weil AA, Chowdhury F, Begum YA, Khan AI, Debela MD, Durand HK, Reese
798 AT, Nimmagadda SN, Silverman JD, Ellis CN, Ryan ET, Calderwood SB, Harris JB, Qadri
799 F, David LA, LaRocque RC. 2018. Human Gut Microbiota Predicts Susceptibility to *Vibrio*
800 *cholerae* Infection. J Infect Dis 218:645–653.

801 13. Levade I, Saber MM, Midani FS, Chowdhury F, Khan AI, Begum YA, Ryan ET, David LA,
802 Calderwood SB, Harris JB, LaRocque RC, Qadri F, Shapiro BJ, Weil AA. 2021. Predicting

803 *Vibrio cholerae* Infection and Disease Severity Using Metagenomics in a Prospective
804 Cohort Study. The Journal of Infectious Diseases 223:342–351.

805 14. Seed KD, Yen M, Shapiro BJ, Hilaire IJ, Charles RC, Teng JE, Ivers LC, Boncy J, Harris
806 JB, Camilli A. 2014. Evolutionary consequences of intra-patient phage predation on
807 microbial populations. eLife 3:e03497.

808 15. Levade I, Terrat Y, Leducq J-B, Weil AA, Mayo-Smith LM, Chowdhury F, Khan AI,
809 Boncy J, Buteau J, Ivers LC, Ryan ET, Charles RC, Calderwood SB, Qadri F, Harris JB,
810 LaRocque RC, Shapiro BJ. 2017. *Vibrio cholerae* genomic diversity within and between
811 patients. Microb Genom 3.

812 16. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of
813 bacterial pathogens. Nature Publishing Group 14:150–162.

814 17. Jolivet-Gougeon A, Kovacs B, Le Gall-David S, Le Bars H, Bousarghin L, Bonnaure-Mallet
815 M, Lobel B, Guillé F, Soussy C-J, Tenke P. 2011. Bacterial hypermutation: clinical
816 implications. Journal of Medical Microbiology, 60:563–573.

817 18. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. 2013.
818 Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a
819 record of selective pressures. Nature Genetics 46:82–87.

820 19. Marvig RL, Johansen HK, Molin S, Jelsbak L. 2013. Genome Analysis of a Transmissible
821 Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct
822 Evolutionary Paths of Hypermutators. PLOS Genetics 9:e1003741.

823 20. Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, Taddei F. 2001. Costs and
824 Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut. Science
825 291:2606–2608.

826 21. Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I,
827 Rothstein R, Elion J, Taddei F, Radman M, Matic I. 2000. Evolutionary Implications of the
828 Frequent Horizontal Transfer of Mismatch Repair Genes. Cell 103:711–721.

829 22. Chu ND, Clarke SA, Timberlake S, Polz MF, Grossman AD, Alm EJ. 2017. A Mobile
830 Element in mutS Drives Hypermutation in a Marine Vibrio. mBio 8.

831 23. Wang H, Xing X, Wang J, Pang B, Liu M, Larios-Valencia J, Liu T, Liu G, Xie S, Hao G,
832 Liu Z, Kan B, Zhu J. 2018. Hypermutation-induced in vivo oxidative stress resistance
833 enhances *Vibrio cholerae* host adaptation. PLOS Pathogens 14:e1007413.

834 24. Didelot X, Pang B, Zhou Z, McCann A, Ni P, Li D, Achtman M, Kan B. 2015. The Role of
835 China in the Global Spread of the Current Cholera Pandemic. PLOS Genetics 11:e1005072.

836 25. Eisenstark A. 2010. Genetic Diversity among Offspring from Archived *Salmonella enterica*
837 ssp. enterica Serovar Typhimurium (Demerec Collection): In Search of Survival Strategies.
838 Annu Rev Microbiol 64:277–292.

839 26. Anyansi C, Straub TJ, Manson AL, Earl AM, Abeel T. 2020. Computational Methods for
840 Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. Front
841 Microbiol 11.

842 27. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF.
843 2021. inStrain profiles population microdiversity from metagenomic data and sensitively
844 detects shared microbial strains. Nature Biotechnology 1–10.

845 28. Garud NR, Good BH, Hallatschek O, Pollard KS. 2019. Evolutionary dynamics of bacteria
846 in the gut microbiome within and across hosts. PLOS Biology 17:e3000102.

847 29. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, Linden M van der, McGee L,
848 Gottberg A von, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas
849 D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP,
850 Bentley SD. 2011. Rapid Pneumococcal Evolution in Response to Clinical Interventions.
851 Science 331:430–434.

852 30. Lovett ST. 2011. The DNA exonucleases of Escherichia coli. EcoSal Plus 4.

853 31. Darmon E, Lopez-Vernaza MA, Helness AC, Borking A, Wilson E, Thacker Z, Wardrope
854 L, Leach DRF. 2007. SbcCD Regulation and Localization in *Escherichia coli*. Journal of
855 Bacteriology 189:6686–6694.

856 32. Lu A-L, Li X, Gu Y, Wright PM, Chang D-Y. 2001. Repair of oxidative DNA damage. Cell
857 Biochem Biophys 35:141–170.

858 33. Foster PL, Gudmundsson G, Trimarchi JM, Cai H, Goodman MF. 1995. Proofreading-
859 defective DNA polymerase II increases adaptive mutation in Escherichia coli. PNAS
860 92:7951–7955.

861 34. Oliver A, Mena A. 2010. Bacterial hypermutation in cystic fibrosis, not only for antibiotic
862 resistance. Clinical Microbiology and Infection 16:798–808.

863 35. Lee S-J, Sung R-J, Verdine GL. 2019. Mechanism of DNA Lesion Homing and Recognition
864 by the Uvr Nucleotide Excision Repair System. Research. Research article.

865 36. Kunkel TA, Erie DA. 2005. Dna mismatch repair. Annu Rev Biochem 74:681–710.

866 37. Zeibell K, Aguila S, Shi VY, Chan A, Yang H, Miller JH. 2007. Mutagenesis and Repair in
867 *Bacillus anthracis*: the Effect of Mutators. Journal of Bacteriology 189:2331–2338.

868 38. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ,
869 Alm EJ. 2019. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host &
870 Microbe 25:656-667.e8.

871 39. Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication
872 rates in microbial communities. 12. Nature Biotechnology 34:1256–1263.

873 40. Olson R, Gouaux E. 2005. Crystal Structure of the *Vibrio cholerae* Cytolysin (VCC) Pro-
874 toxin and its Assembly into a Heptameric Transmembrane Pore. Journal of Molecular
875 Biology 350:997–1016.

876 41. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in
877 Drosophila. Nature 351:652–654.

878 42. Tischler AD, Camilli A. 2004. Cyclic diguanylate (c-di-GMP) regulates *Vibrio cholerae*
879 biofilm formation. Mol Microbiol 53:857–869.

880 43. Seed KD, Bodi KL, Kropinski AM, Ackermann H-W, Calderwood SB, Qadri F, Camilli A.
881 2011. Evidence of a Dominant Lineage of *Vibrio cholerae*-Specific Lytic Bacteriophages
882 Shed by Cholera Patients over a 10-Year Period in Dhaka, Bangladesh. mBio 2.

883 44. Seed KD, Lazinski DW, Calderwood SB, Camilli A. 2013. A bacteriophage encodes its own
884 CRISPR/Cas adaptive response to evade host innate immunity. 7438. Nature 494:489–491.

885 45. Ram G, Chen J, Kumar K, Ross HF, Ubeda C, Damle PK, Lane KD, Penadés JR, Christie
886 GE, Novick RP. 2012. Staphylococcal pathogenicity island interference with helper phage
887 reproduction is a paradigm of molecular parasitism. PNAS 109:16300–16305.

888 46. O'Hara BJ, Barth ZK, McKitterick AC, Seed KD. 2017. A highly specific phage defense
889 system is a conserved feature of the *Vibrio cholerae* mobilome. PLOS Genetics
890 13:e1006838.

891 47. Seed KD, Faruque SM, Mekalanos JJ, Calderwood SB, Qadri F, Camilli A. 2012. Phase
892 Variable O Antigen Biosynthetic Genes Control Expression of the Major Protective Antigen
893 and Bacteriophage Receptor in *Vibrio cholerae* O1. PLoS Pathogens 8:e1002917.

894 48. Kendall EA, Chowdhury F, Begum Y, Khan AI, Li S, Thierer JH, Bailey J, Kreisel K,
895 Tacket CO, LaRocque RC, Harris JB, Ryan ET, Qadri F, Calderwood SB, Stine OC. 2010.
896 Relatedness of *Vibrio cholerae* O1/O139 isolates from patients and their household
897 contacts, determined by multilocus variable-number tandem-repeat analysis. Journal of
898 Bacteriology 192:4367–4376.

899 49. Brenzinger S, van der Aart LT, van Wezel GP, Lacroix J-M, Glatter T, Briegel A. 2019.
900 Structural and Proteomic Changes in Viable but Non-culturable *Vibrio cholerae*. Front
901 Microbiol 10.

902 50. Nelson EJ, Harris JB, Glenn Morris J, Calderwood SB, Camilli A. 2009. Cholera
903 transmission: the host, pathogen and bacteriophage dynamic. Nature Reviews Microbiology
904 7:693–702.

905 51. Bachmann NL, Rockett RJ, Timms VJ, Sintchenko V. 2018. Advances in Clinical Sample
906 Preparation for Identification and Characterization of Bacterial Pathogens Using
907 Metagenomics. Front Public Health 6.

908 52. Vezzulli L, Grande C, Tassistro G, Brettar I, Höfle MG, Pereira RPA, Mushi D, Pallavicini
909 A, Vassallo P, Pruzzo C. 2017. Whole-Genome Enrichment Provides Deep Insights into
910 *Vibrio cholerae* Metagenome from an African River. Microb Ecol 73:734–738.

911 53. Labat F, Pradillon O, Garry L, Peuchmaur M, Fantin B, Denamur E. 2005. Mutator
912 phenotype confers advantage in *Escherichia coli* chronic urinary tract infection
913 pathogenesis. FEMS Immunol Med Microbiol 44:317–321.

914 54. Funchain P, Yeung A, Stewart JL, Lin R, Slupska MM, Miller JH. 2000. The Consequences
915 of Growth of a Mutator Strain of *Escherichia coli* as Measured by Loss of Function Among
916 Multiple Gene Targets and Loss of Fitness. Genetics 154:959–970.

917 55. Tsou AM, Zhu J. 2010. Quorum Sensing Negatively Regulates Hemolysin Transcriptionally
918 and Posttranslationally in *Vibrio cholerae*. Infection and Immunity 78:461–467.

919 56. Olivier V, Haines GK, Tan Y, Satchell KJF. 2007. Hemolysin and the Multifunctional
920 Autoprocessing RTX Toxin Are Virulence Factors during Intestinal Infection of Mice with
921 *Vibrio cholerae* El Tor O1 Strains. Infection and Immunity 75:5035–5042.

922 57. Weil AA, Arifuzzaman M, Bhuiyan TR, LaRocque RC, Harris AM, Kendall EA, Hossain
923 A, Tarique AA, Sheikh A, Chowdhury F, Khan AI, Murshed F, Parker KC, Banerjee KK,
924 Ryan ET, Harris JB, Qadri F, Calderwood SB. 2009. Memory T-Cell Responses to *Vibrio*
925 *cholerae* O1 Infection. Infection and Immunity 77:5090–5096.

926 58. Ferdous F, Ahmed S, Farzana FD, Das J, Malek MA, Das SK, Salam MA, Faruque ASG.
927 2015. Aetiologies of diarrhoea in adults from urban and rural treatment facilities in
928 Bangladesh. Epidemiology & Infection 143:1377–1387.

929 59. Azman AS, Lauer SA, Bhuiyan TR, Luquero FJ, Leung DT, Hegde ST, Harris JB, Paul KK,
930 Khaton F, Ferdous J, Lessler J, Salje H, Qadri F, Gurley ES. 2020. *Vibrio cholerae* O1
931 transmission in Bangladesh: insights from a nationally representative serosurvey. Lancet
932 Microbe 1:e336–e343.

933 60. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature
934 Methods 9:357–359.

935 61. Eren AM, Vineis JH, Morrison HG, Sogin ML. 2013. A Filtering Method to Generate High
936 Quality Short Reads Using Illumina Paired-End Technology. PLOS ONE 8:e66643.

937 62. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2.
938 Genome Biol 20:257.

939 63. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics
940 pipeline for strain profiling reveals novel patterns of bacterial transmission and
941 biogeography. Genome Res 26:1612–1625.

942 64. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016.
943 MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced
944 methodologies and community practices. Methods (San Diego, Calif) 102:3–11.

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

948 66. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ,
949 Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and
950 composition. Nature Methods 11:1144–1146.

951 67. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to
952 recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607.

953 68. Kang D, Li F, Kirton ES, Thomas A, Egan RS, An H, Wang Z. 2019. MetaBAT 2: an
954 adaptive binning algorithm for robust and efficient genome reconstruction from
955 metagenome assemblies. preprint, PeerJ Preprints.

956 69. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018.
957 Recovery of genomes from metagenomes via a dereplication, aggregation and scoring
958 strategy. 7. Nature Microbiology 3:836–843.

959 70. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015.
960 Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

961 71. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive
962 classification of metagenomic sequences. Genome Research 26:1721–1729.

963 72. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoesser N, Peto TEA,
964 Crook DW, Walker AS. 2020. Genomic diversity affects the accuracy of bacterial single-
965 nucleotide polymorphism–calling pipelines. Gigascience 9.

966 73. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
967 prokaryotic gene recognition and translation initiation site identification. BMC
968 Bioinformatics 11:119.

969 74. The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic
970 Acids Res 47:D506–D515.

971 75. Madden T. 2003. The BLAST Sequence Analysis ToolThe NCBI Handbook [Internet].
972 National Center for Biotechnology Information (US).

973 76. Seemann T. 2015. Snippy: fast bacterial variant calling from NGS reads.

974 77. Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA)
975 for macOS. Mol Biol Evol 37:1237–1239.

976 78. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
977 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,
978 Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its
979 Applications to Single-Cell Sequencing. Journal of Computational Biology 19:455–477.

980 79. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford,
981 England) 30:2068–2069.

982 80. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D,
983 Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis.
984 Bioinformatics 31:3691–3693.

985 81. Gabrielaite M, Marvig RL. 2020. GenAPI: a tool for gene absence-presence identification in
986 fragmented bacterial genome sequences. BMC Bioinformatics 21:320.

987 82. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. 2018.
988 Phandango: an interactive viewer for bacterial population genomics. Bioinformatics
989 34:292–293.

990
991