

Title: *Vibrio cholerae* genomic diversity within and between patients

Inès Levade¹, Yves Terrat¹, Jean-Baptiste Leducq¹, Ana A. Weil^{2,3}, Leslie M. Mayo-Smith², Fahima Chowdhury⁴, Ashraful I. Khan⁴, Jacques Boncy⁵, Josiane Buteau⁵, Louise C. Ivers^{3,6,7}, Edward T. Ryan^{2,3,8}, Richelle C. Charles^{2,3}, Stephen B. Calderwood^{2,3,9}, Firdausi Qadri⁴, Jason B. Harris^{2,10}, Regina C. LaRocque^{2,3}, B. Jesse Shapiro^{1*}

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

²Division of Infectious Diseases, Massachusetts General Hospital, Boston, Massachusetts, USA

³Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

⁴Center for Vaccine Sciences, International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh

⁵National Public Health Laboratory, Ministry of Public Health and Population, Port-au-Prince, Haiti.

⁶Division of Global Health Equity, Brigham and Women's Hospital, Boston, Massachusetts USA

⁷Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA

⁸Department of Immunology & Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA

⁹Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, USA

¹⁰Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA

*For correspondence: jesse.shapiro@umontreal.ca

Abstract

Cholera is a severe, waterborne diarrheal disease caused by toxin-producing strains of the bacterium *Vibrio cholerae*. Comparative genomics has revealed "waves" of cholera transmission and evolution, in which clones are successively replaced over decades and centuries. However, the extent of *V. cholerae* genetic diversity within an epidemic or even within an individual patient is poorly understood. Here, we characterized *V. cholerae* genomic diversity at a micro-epidemiological level within and between individual patients from Bangladesh and Haiti. To capture within-patient diversity, we isolated multiple (8 to 20) *V. cholerae* colonies from each of eight patients, sequenced their genomes and identified point mutations and gene gain/loss events. We found limited but detectable diversity at the level of point mutations within hosts (zero to three single nucleotide variants within each patient), and comparatively higher gene content variation within hosts (at least one gain/loss event per patient, and up to 103 events in one patient). Much of the gene content variation appeared to be due to gain and loss of phage and plasmids within the *V. cholerae* population, with occasional exchanges between *V. cholerae* and other members of the gut microbiota. We also show that certain intra-host variants have phenotypic consequences. For example, the acquisition of a *Bacteroides* plasmid and nonsynonymous mutations in a sensor histidine kinase gene both reduced biofilm formation, an important trait for environmental survival. Together, our results show that *V. cholerae* is measurably evolving within patients, with possible implications for disease outcomes and transmission dynamics.

Author Summary

Vibrio cholerae is the etiological agent of cholera, a severe diarrheal disease endemic to Bangladesh and responsible for global outbreaks, including one ongoing in Haiti. Certain bacterial pathogens can evolve and diversify within the human host, often altering virulence and antibiotic resistance. However, most examples of within-host evolution have come from chronic infections, in which the pathogen has sufficient time to mutate and diversify, and little attention has been paid to more acute infections such as the one caused by *V. cholerae*. The goal of this study was to measure the extent of within-host evolution of *V. cholerae* within individual infected patients. By sequencing multiple bacterial isolates from each of eight patients from Bangladesh and Haiti, we found that cholera patients can harbor a diverse population of *V. cholerae*. As expected for an acute infection, this diversity is limited, ranging from zero to three point mutations (single nucleotide variants) per patient. However, gene gain/loss events are more prevalent than point mutations, occurring in every single patient, and sometimes involving the transfer of dozens of genes on plasmids. Even if rare, point mutations and gene gain/loss events may be maintained by natural selection, and can alter clinically- and environmentally-relevant phenotypes such as biofilm formation. Therefore, within-patient evolution has the potential to impact clinical and epidemiological outcomes. Together, our results demonstrate that within-patient evolution may be a general feature of both acute and chronic infections, and that gene gain/loss may be an important but under-appreciated feature of within-host evolution.

Introduction

Cholera is an acute diarrheal infection that remains a serious health threat in countries with limited access to clean water [1]. *Vibrio cholerae* is the causative agent of the disease and is a natural inhabitant of aquatic ecosystems [2], with more than 200 serogroups identified to date on the basis of their somatic O antigens [3,4]. Most *V. cholerae* serogroups are not pathogenic; only isolates in serogroup O1 (consisting of two biotypes known as “classical” and “El Tor” and the serotypes Ogawa and Inaba) and O139 have been identified as agents of cholera epidemics and pandemics [1].

Whole genome sequencing and population genomics have the potential to improve our understanding of the epidemiology, etiology and evolution of bacterial infectious diseases [5]. For example, comparisons of whole-genome sequences of *V. cholerae* strains from across the world, over the course of a century, clarified the history of the current pandemic [6] and showed that this pandemic is the result of a single clonal expansion of one *V. cholerae* O1 El Tor ancestor, accompanied by horizontal gene transfer (HGT) events involving toxin and antibiotic resistance genes [7]. More recently, comparative genomics have been applied to answer epidemiological questions, proving the Asian origin of the strain causing the ongoing Haitian cholera outbreak, which began in 2010 [8-11]. Using whole genome sequencing and single nucleotide polymorphism (SNP) analysis, Azarian et al. [12] compared 60 clinical and environmental isolates collected in Haiti from 2010 to 2012. They found that the 2011 and 2012 strains rapidly diverged from the 2010 ancestral strain that initiated the outbreak, suggesting evolution driven by positive selection in a new environment [12].

Viral pathogens can evolve and diversify within infected patients, with serious consequences for disease outcome [13], and certain bacterial pathogens have recently been shown to diversify within patients as well [14]. However, evolutionary and epidemiological studies are usually conducted with just one bacterial isolate taken as representative of the infection, even though within-patient diversity is important to consider, for several reasons [15-18]. Within-host evolution may impact the longer-term evolution and transmission potential of pathogens, particularly if there are fitness tradeoffs between evolution within and between hosts. For example, a study of one cholera patient from Haiti showed that phage-resistant *V. cholerae* mutants rose to high frequency within the patient due to positive selection imposed by phage predation[19]. This study showed how strong selection can shape *V. cholerae* diversity within patients, but the prevalence and extent of *V. cholerae* genetic diversity within patients remains unclear, and whether intra-host evolution is generally driven by selection.

As for many other bacterial pathogens, the prevailing orthodoxy is that *V. cholerae* infections are clonal, and essentially devoid of within-host genetic diversity. Although within-host *V. cholerae* populations have not been studied extensively, additional evidence suggests that within-host diversity does indeed exist, at the level of phase variation in the O antigen, or in variable number tandem repeat (VNTR) loci [4,20]. This diversity could arise by within-host evolution, or be due to infection by different strains that diverged before entering the host. *V. cholerae* is genetically diverse in aquatic ecosystems [21] and co-infections from diverse environmental strains are possible [22]. *V. cholerae* infections are acute, lasting only a few days before the patient either recovers or dies [1]. Therefore, there is limited time for within-host evolution (including mutation, recombination, and selection) to occur. It is therefore expected that

V. cholerae will experience less within-host evolution compared to more chronic bacterial infections with documented within-host evolution [23-28]. On the other hand, *V. cholerae* grows to large population sizes within the host (from 10^7 to 10^9 vibrios per gram of stool), dominating the gut microbiome [29,30]. If the effective population size within a host is large, many mutations are expected and natural selection will be efficient. However, *V. cholerae* likely experiences population bottlenecks upon infection and within the gut [31], which would reduce genetic diversity and reduce the efficiency of selection. In addition to point mutations, *V. cholerae* can undergo high rates of HGT [7,32,33], providing an additional potential source of within-host diversity. During an infection, *V. cholerae* could acquire genes from plasmids, phages, pathogenicity islands or genes from the gut microbiota, which appears to be a hotspot of HGT [34]. However, the extent of within-patient mutation, HGT, and natural selection are still poorly known for *V. cholerae*.

In this study, we quantified within-patient genetic diversity of *V. cholerae* infections. We characterized genomic diversity of *V. cholerae* within and between eight cholera patients, sequencing between eight and 20 isolate genomes per patient. We identified both intra-host single nucleotide variants (iSNVs) and gene gain/loss events within patients. As expected for an acute infection, few within-patient point mutations were detected, ranging from zero to three iSNVs per host. In contrast, we found a substantial amount of gene content variation: between five and 103 gene gains or losses within each patient. We infer that most diversity is due to within-host mutation rather than co-infection, and that HGT of mobile elements is more common than point mutations. In most patients, within-host evolution can be explained by neutral mutation, recombination and bottlenecks; however, one patient showed evidence for diversification

141 driven by positive selection, resulting in phenotypic variation among intra-host *V.*
142 *cholerae* isolates in their ability to form biofilms. Despite the relatively small numbers of
143 mutations and HGT events within hosts, these events may have important evolutionary
144 and phenotypic consequences for *V. cholerae* populations.

145

Results

Genomic characterization of within-patient isolates

We isolated and sequenced a total of 122 clinical strains recovered in 2013 from eight patients with culture-confirmed cholera. Five patients were from Dhaka, Bangladesh. Samples were collected during the same week at the end of 2013, and identified as patients B1, B2, B3, B4 and B5. The three other patients (H1, H2, H3) came from Artibonite, Haiti and presented with cholera, all during a three-day period in April 2013. For each patient, eight to 20 colonies were isolated from a single stool sample to assess within-patient genetic variation (Fig 1). Colonies were named with a C, followed by a number; for example, B1C1 corresponds to colony 1 from Bangladesh patient 1. All strains were identified as toxigenic *V. cholerae* O1 biotype El Tor, serotype Ogawa which was the prevailing serotype at each site during the entire study period. Each isolate was separately sequenced using Illumina technology to a minimum depth of 28X coverage of the MJ-1236 reference genome (mean coverage = 136X). We *de novo* assembled each genome, and also mapped reads to two closed, annotated *V. cholerae* reference genomes. After filtering for errors due to culture and sequencing (see Materials and Methods), we identified a total of 485 high quality single-nucleotide polymorphisms (hqSNPs) among our 122 isolates, distributed across the genome (S1 Fig), with a majority (471 hqSNPs) located in the branch between the Bangladeshi and the Haitian strains.

Within-patient single nucleotide variation

The set of hqSNPs included intra-host single nucleotide variants (iSNVs) in patients B1, B4, B5 and H1 (Fig 1). Isolates within patients B2, B3, H2 and H3 were isogenic, with no iSNVs detected using our quality filters. Patient B1 contained one intergenic iSNV, and patients B4 and B5 each contained a synonymous iSNV, each in a different gene (Table 1). Patient H1 contained three iSNVs, all of which were non-synonymous, and two of which occurred in the same gene, a sensor histidine kinase (Table 1). No small insertion/deletions (indels) were found within any patients.

To ensure that the relatively small number of iSNVs were not due to mutation during strain isolation and culture [35] or sequencing errors, we sub-cultured and sequenced 12 colonies from one isolate (B1C1) as a control (Fig 1). Applying the same filters as for SNP discovery in our patients, we did not detect any iSNVs among control strains, nor did we detect any SNP differences between replicate libraries prepared and sequenced using different platforms (Materials and Methods). This suggests that the few iSNVs identified within patients are unlikely to be culture or sequencing artifacts.

Based on the number (0-3) and frequencies of iSNVs per host, we used measures of genetic diversity (θ_w and π) to estimate within-host effective population size (N_e). We estimated that N_e within each patient ranged from 0 to 110 (S1 Table). Such a small N_e in a bacterial population is consistent with a recent population bottleneck, possibly during host colonization, or a recent selective sweep having purged most of the diversity within the population.

Gene gain and loss within and between cholera patients

To characterize variation in gene content among the 122 sequenced isolates, we analyzed orthologous coding sequences from *de novo* assemblies. Of 3907 gene families identified, 3489 were defined as core (i.e. present in all genomes) while 401 (~10% of the total gene pool) were considered flexible, present in only a subset of genomes. Due to variation in sequencing coverage and nucleotide composition, genome assemblies may be incomplete, missing a subset of genes that are actually present [36,37]. To address this issue, we mapped reads back to the full gene catalogue, and considered a gene to be present when it was covered at 1X (Materials and Methods; S2 and S3 Fig). This filtering procedure revealed that, of the 401 genes initially classified as flexible, 252 were actually part of the core, leaving 155 *bona fide* flexible genes. Before filtering, we observed gene content variation among control isolates, but these false positives were removed using the 1X coverage filter. After filtering, between five and 103 flexible genes showed variation within patients (Table 2; Fig 2). Because gene content variation could be biased by different methods of library construction and sequencing, we sequenced 12 strains in duplicate using two different methods. We observed variation in the detection of the flexible genome for six of the duplicate sets (S4 Fig), likely because different library preparation methods are known to have different GC-content biases [38]. Using genomes from one method only (NEBNext/HiSeq), we identified between five and 67 variable genes per patient; using the other method (Nextera/MiSeq) we identified between zero and 62 genes (S2 Table). When a variable gene is not detected in a given method, we consider this a false negative. We conclude that methodological differences alone cannot explain the flexible genome variation within patients, and consider both methods combined for the remainder of the study.

Clustering the 155 flexible genes by their presence/absence profile across patients revealed four distinct categories of genes (Fig 2). Category A consists of genes present in Bangladesh (part of the Bangladesh core genome), but showing a patchy distribution in Haiti. Category B genes are fixed (present in all isolates) in patients B1 and B5, and patchy in other patients. Category C genes are fixed in Haiti and nearly absent from Bangladesh. Category D genes tend to be rare, often singletons only observed in a single isolate within either patient B1 or H2.

Several flexible genes corresponded to known mobile genetic elements. Notably, category A contains 61 gene families (39% of the flexible genome), all located on one single contig corresponding to the SXT Integrative Conjugative Element (ICE). Category B encompassed 49 gene families matching Kappa phage proteins. These putative phage genes were clustered together on large contigs of chromosome 1, and were fixed in patients B1 and B5 but variable among other patients, some of which contained complete phage sequences (patients B2 and B3). Category C contained 15 genes, including some present in the ICE, which mapped to at least five different contigs (depending on which isolate's assembly was considered), suggesting multiple gain/loss events or frequent rearrangements.

Over half of the flexible genome (80 genes) was annotated as hypothetical proteins, compared to the core genome that contained less than 3% hypotheticals. The flexible genome also contained ten transposases (6.5% of the flexible genome, compared to 1.4% of the core) and eight genes involved plasmid and viral replication, all potential mechanisms of HGT[39]. A complete list and annotation of flexible genes is given in Supplementary Table 5.

Variation in the flexible gene pool could arise from gene deletion, duplication, or horizontal gene transfer (HGT). To determine the extent of HGT across species boundaries, we identified the taxonomic affiliation of each flexible gene according to its best BLAST hit in GenBank. While 117 of the 155 of flexible genes were assigned to *V. cholerae*, the other 38 genes were assigned to other members of the family *Vibrionaceae* or even distantly related species of *Bacteroides* or *Staphylococcus* (Fig 2), suggesting HGT from these donors to *V. cholerae* in the gut. For example, a group of 20 genes present in isolate H2C3 (but absent in other isolates from patient H2) matched a plasmid previously identified in *Bacteroides* (Fig 2). These 20 genes of putative *Bacteroides* plasmid origin are among 25 singletons, present in only one isolate of patient H2. Similarly, the 5 singletons in patient B1 (Table 2) are all of putative *Staphylococcus* origin. As no other genes were assigned to either *Staphylococcus* or *Bacteroides*, it appears these putative HGT events are not due to contamination from these taxa. This suggests that cross-species HGT events are relatively rare and recent events, which may never achieve high frequency in the *V. cholerae* population, either within or across hosts. Together, these results suggest that most within-patient variation in gene content is due to gene flow, deletion or duplication within the *V. cholerae* population, with rare but detectable HGT from other species of bacteria, phages and plasmids in the gut microbiota.

***V. cholerae* evolution on different time scales.**

In order to place within-host variation in the context of longer-term *V. cholerae* evolution, and to distinguish within-patient mutation from coinfection events, we built a phylogeny of the 122 isolates (all from 2013) as well as 21 additional isolates obtained

from acute cholera patients sampled in Bangladesh from 2011 to 2013 (Fig 3). These clinical isolates were processed with the same methods as the within-host isolates, from the culture to the hqSNP discovery (Materials and Methods). With a total of 35 unique genotypes (one to four per patient), we computed a Bayesian phylogeny, estimated divergence times, and evolutionary rates using BEAST 1.8.3 [40]. We used an alignment of all the concatenated hqSNPs except those found in the ICE, a highly variable 100 kb region (S1 Fig) that undergoes frequent HGT and thus has a separate evolutionary history from the rest of the genome [6].

To test the degree of temporal structure in the phylogeny, we performed a root-to-tip linear regression ($R^2=0.65$, $p<0.0001$; S5 Fig), which suggested that our dataset is sufficiently clock-like to robustly estimate an evolutionary rate. The Bayesian tree was calibrated with sampling dates of the isolates ranging from March 2011 to December 2013 (Fig 3). We compared different molecular clock and demographic models (Materials and Methods), and found that the strict molecular clock and the Bayesian skyline plot models provided the best fit (S3 Table), in accordance with previous studies of *V. cholerae* [12,41]. Assuming a constant evolutionary rate across lineages, we estimated the evolutionary rate at 7.94×10^{-7} hqSNP site⁻¹ year⁻¹ (95% highest posterior density [HPD], 4.89×10^{-7} to 1.14×10^{-6}) or approximately 3.3 hqSNP year⁻¹ in the core genome (95% HPD), consistent with previous estimates [6].

We then estimated the ages of the most recent common ancestors (MRCAs) of the phylogenetic sub-lineages and clusters (S4 Table). Notably, four of the six strains isolated in 2011 from Bangladesh were found to be closer to the Haitian strains than to all the other Bangladeshi strains (Fig 3). We called this sub-lineage BDG-B, and estimated the age of the MRCA of these four strains and the Haitian isolates in

September 2005 (95% HPD, June 2002 to July 2008), which pre-dates the introduction of pathogenic *V. cholerae* in Haiti, and is consistent with its Asian origin [10,11]. The time of the MRCA of the strains isolated from the three Haitian patients was estimated at December 2012 (95% HPD, August 2012 to March 2013).

Based on the phylogeny (Fig 3), we sought to distinguish between scenarios of within-patient mutation or coinfection as causes of within-patient diversity. It is clear that isolates from the same patient always grouped together, and were never polyphyletic. This observation is consistent with each patient being colonized by a single clone, which subsequently diversified by mutation within the patient. The diversity between the eight patients (7 SNPs) was greater than the diversity within patients (0-3 iSNVs), which would be unlikely if strains were sampled by patients at random from an environmental gene pool (Table 1). If within-patient diversity was due to coinfection of the same patient by multiple different strains, we would expect these strains to share a MRCA before the date of infection, and certainly before the date stool was sampled. However, the MRCA of isolates from a single patient always overlapped with the date of sampling, suggesting that within-patient diversity is more likely due to within-patient mutation than to coinfection.

Signatures of natural selection on within-patient variants.

We next asked how regimes of natural selection on *V. cholerae* varied over time, within and between patients. We used the nonsynonymous (NS) to synonymous (S) substitution ratio to measure the strength of selection at the protein level. Over three years of evolution, we could not reject a neutral evolutionary model and found no evidence for variation in the NS:S ratio over time, considering only SNPs fixed between

patients (Supplementary Text). Another possibility is that selection acts over shorter evolutionary scales, by shaping intra-host diversity during acute infection. Under this scenario, we would expect NS:S ratios to differ significantly within and between hosts. For example, higher NS:S within than between hosts could be due to positive or balancing selection on NS mutations within hosts, or due to more efficient purifying selection (against deleterious NS mutations) between hosts. To test for such deviations from neutral evolution, we applied the McDonald-Kreitman test [42] to the eight hosts surveyed for within-host genetic variation (five from Bangladesh and three from Haiti). Despite the overall low number of SNPs and iSNVs observed, we found a significant excess of NS mutations between Bangladeshi patients (Fisher's exact test, Odds Ratio > 12, $p < 0.05$; Table 3), suggesting positive selection for the fixation of NS mutations between patients, or purifying selection against NS mutations within patients. In contrast, all three iSNVs observed in Haiti were NS, suggesting positive, balancing, or relaxed purifying selection within patients, although not statistically significant (Fisher's exact test, Odds Ratio < 0.32, $p = 1$; Table 2).

The three NS iSNVs observed in Haiti all occurred within a single patient, possibly driven by selective pressures specific to this patient (Table 1). To test whether this pattern of iSNVs was likely to have occurred at random or due to patient-specific selection, we performed permutations of iSNVs among hosts and estimated expected iSNVs frequencies (F) and number of NS iSNVs per host and region. We found a significant excess of iSNVs in Haiti ($F_{HTI} = 0.15$; $p < 0.05$; 10,000 permutations) and in patient H1 ($F_{H1} = 0.44$; $p < 0.01$), but not in Bangladesh ($F_{BGD} = 0.03$; $p > 0.05$) nor in any other patients ($F = 0-0.05$; $p > 0.05$). All iSNVs identified in Haiti and in patient H1 were nonsynonymous, which was significantly higher than expected by chance ($p < 0.01$ and

p<0.001, respectively; Fig 4). These results show that patient H1 has a significant excess of NS iSNVs compared to other patients. This suggests positive or balancing selection on NS iSNVs within patient H1, or relaxed purifying selection in H1 compared to other patients.

The three NS iSNVs in patient H1 occurred in two genes. The first gene, containing one iSNV, encodes a member of the tetracycline resistance (Tet^R) family of transcriptional regulators (NCBI: ACQ60802.1), known to be involved in the transcriptional control of multidrug efflux pumps and other pathways like quorum-sensing circuits or pathogenicity[41,43]. The other two non-synonymous mutations in patient H1 were located in the same gene (NCBI: ACQ61177), a sensor histidine kinase (HK) called *vprB*, which is required for resistance to the antimicrobial peptide polymyxin B [44]. Each of these two iSNVs occurs at a different site in the gene, each in a different isolate. Based on the fact that the major allele at each of these iSNV sites was present in both reference genomes (MJ1236 and 2010EL-1786), we inferred that the minor alleles (both at frequency 1/9 in patient H1; Table 1) were derived, presumably due to within-patient mutation. A comparison of the *vprB* (ACQ61177) protein sequence with its 465 closest orthologs revealed that the NS iSNVs identified in this gene modify peptides that are otherwise highly conserved across *Vibrio* species (S7 Fig), suggesting that these mutations may affect protein function.

Within-patient variants affect biofilm formation.

To identify possible phenotypes conferred by these within-patient variations, we performed three sets of experiments on the strains from patient H1 harboring the three NS point mutations, and the strain for which the *Bacteroides* plasmid was detected in

patient H2. First, to test whether intra-host variants had any effect on *V. cholerae* growth in liquid medium, we measured the growth rates of strains with derived iSNV alleles compared to one isogenic strain (with inferred ancestral alleles and no variation in the flexible genome) from the same patient, and found no significant difference in the growth rates between strains. Second, we tested for resistance to the antibiotic polymyxin B, as loss of *vprB* function has been previously associated with increased susceptibility to polymyxin B [44], and again found no difference between strains (S8 Fig). Third, we tested for biofilm formation, as it has been previously suggested that biofilm formation inside the host can impair intestinal colonization [45,46], and it is known that HKs in certain two-component systems can affect biofilm formation (48,49). As a biofilm-negative control, we used RBM, a biofilm knock-out strain of *V. cholerae* [47], sterile LB, or saline.

For patient H1, we did not detect any significant difference in biofilm formation between strains with the ancestral iSNV allele (in isolate H1C1) and the derived allele in the transcriptional regulator gene (isolate H1C4). However, we found that H1C5 and H1C6, the two isolates with derived alleles in the HK gene *vprB*, produced significantly less biofilm than the other strains from the same patient (Fig 5A). Based on our hqSNP calls and flexible gene analysis, the genome of H1C5 was identical to H1C1, with the exception of the iSNV in the HK gene. Therefore, the difference in biofilm phenotypes is attributable to this iSNV. However, H1C6 differed from H1C1 by an iSNV in the HK gene, and also by the presence/absence of genes in the flexible genome (S4 Fig). However, this gene content variation did not measurably affect biofilm formation (Fig 5A).

379 In the case of patient H2, we showed that the presence of a plasmid of putative
 380 *Bacteroides* origin strongly affects biofilm production. Specifically, the plasmid-
 381 containing isolate (H2C3) produces approximately two-fold less biofilm than an isogenic
 382 control from the same patient (H2C5). H2C3 biofilm formation was indistinguishable from
 383 negative controls (Fig 5B). Together, these results show that both point mutations and
 384 plasmids segregating within patients can affect biofilm formation.

Discussion:

In this study, we surveyed the genetic diversity of *Vibrio cholerae* within infected patients. Using whole-genome sequencing, we analyzed 122 clinical isolates from eight cholera patients from Bangladesh and Haiti, and demonstrated that overall levels of within-patient variation are low for *V. cholerae* populations compared to more chronic bacterial pathogens, which routinely harbor more than 20 iSNVs per patient [24-28,48]. Even if rare, point mutations may be under selection within hosts, with phenotypic consequences. For example, we showed that intra-host mutations in a sensor histidine kinase gene reduced biofilm formation. In addition to point mutations, HGT plays a major role in *Vibrio cholerae* evolution and may represent the major source of genetic diversity, not only in the aquatic environment, but also in the human host – and with large effects on phenotypes like biofilm formation. Specifically, different mutations in a sensor histidine kinase and the acquisition of a plasmid both reduced the ability of *V. cholerae* to form biofilms, which could be advantageous during host colonization[44,45].

Gene content variation within patients and its functional consequences

While HGT is already well-characterized on longer epidemiological time scales, we show that it also occurs within individual patients. *V. cholerae* is known to undergo HGT via transformation [49], transduction [50] and conjugation [51,52]. HGT contributes substantially to drug resistance, pathogenicity and adaptation to different environments, via the acquisition of genomic islands, phages, transposons, ICEs and plasmids [7,53]. Our characterization of the flexible genome within patients used read mapping to confirm gene absences, reducing false positive inference of gene content variation (S2 and S3

Fig). We detected between five and 103 genes that varied in presence/absence within patients (Fig 2; Table 2). Each gene does not necessarily represent an independent gain/loss event; for example, the ICE contained 61 genes on a single contig, likely a single gain/loss event. Even under the conservative assumption that all gene content variation represents a single gain/loss event per patient, this still indicates at least one event per patient.

Some of the putative HGT events could have consequences for *V. cholerae* survival and virulence within the host. For instance, the group of 20 genes acquired by one *V. cholerae* isolate within patient H2, likely via a plasmid of *Bacteroides* origin, is associated with a two-fold reduction in biofilm formation (Fig 5). Among these 20 genes, we identified an antibiotic resistance gene, a haloacid dehalogenase protein that could impact pathogenicity [54], and a FtsY recognition signal protein that was shown to increase virulence in *Streptococcus* [55], among others (S4 Fig). Among the genes likely acquired from a *Staphylococcus* donor in isolate B1C2, three could potentially be involved in modulation of virulence (S4 Fig, S6 Table). First, a GNAT family acetyltransferase could promote virulence or increase antibiotic resistance [56,57]. Second, a putative phosphoenolpyruvate phosphotransferase has been demonstrated to play a role in biofilm formation in *Vibrio* [58-60]. Finally, the KdpC gene (a potassium-transporting ATPase) has been shown to modulate virulence in *Mycobacterium paratuberculosis* [61].

Not all gene gain/losses are due to HGT. Many can be explained by gene deletions, such as phage excision events. Deletions could explain much of the variation among genes in categories A and B (Fig 2), respectively corresponding to the ICE and Kappa phage. These elements are known to vary among *V. cholerae* strains, but here

we document likely excision events during human infection. Genes in category D tend to be singletons, present in just a single isolate, and with taxonomic affiliations well beyond *V. cholerae*, including *Bacteroides*, *Staphylococcus*, and crAssphage (Fig 2). Category D genes are most easily explained by cross-species HGT, as previously documented by Folster and colleagues, who identified a Haitian isolate having gained multidrug resistance through transfer of a plasmid from a species of *Enterobacteriaceae* [51]. Taken together, these results are consistent with the human gut being a hotspot of HGT [62], sometimes involving pathogens like *V. cholerae*.

Regimes of natural selection inferred from within-patient point mutations

The low levels of polymorphism (0-3 iSNVs per patient) we observed within cholera infections could be easily confounded with sequencing errors or mutations occurring during culture rather than within patients. Therefore, we developed filters for calling SNPs and gene gain/loss events that yielded zero variation among control colonies, suggesting low rates of false-positive variant calls and increasing confidence that the six total iSNVs (Fig 1; Table 1) did indeed vary within patients.

Point mutations detected within cholera patients could be the result of *de novo* mutations occurring within the patient, or a consequence of a co-infection from different strains that had diverged previous to the infection. Although we cannot formally exclude coinfections, our results are more consistent with *de novo* mutation within hosts. Isolates from the same patient were grouped together on the phylogeny (Fig 3), suggesting a clonal ancestor. This result is consistent with previous findings that cholera outbreaks are highly clonal [63].

Cycles of transmission from host to host, or from aquatic environment to host, can induce population bottlenecks, reducing the effective population size. Based on comparisons of distantly related genomes, N_e of *V. cholerae* has been estimated to be 4.78×10^8 [64]. This relatively large N_e reflects the high genetic diversity present in the aquatic environment, and over long evolutionary time scales. However, during transmission and intestinal colonization, the size of the *V. cholerae* population experiences drastic bottlenecks that could temporarily reduce N_e . Abel and colleagues showed that *V. cholerae* population sizes in rabbit models of infection ranged from 10^5 during the early phases of colonization to $\sim 10^2$ at the late phases of infection [31]. Our estimates of N_e based on iSNVs give values of 0 to $\sim 10^2$, consistent with population bottlenecks or selective sweeps purging diversity (S1 Table).

In addition to the small within-patient N_e , we found that the distribution of mutations, physically along the *V. cholerae* genome, and temporally along the phylogeny, could generally be explained by random neutral simulations (Supplementary text). However, we identified an excess of nonsynonymous mutations in one Haitian patient (H1), suggesting positive or diversifying selection on *V. cholerae* within this patient (Fig 4). Two of these mutations affected the same protein, a sensor protein-histidine kinase. This sensor protein histidine kinase (HK) is part of a two-component system known as VprAB, which has been shown to mediate glycine fixation in the lipid A domain of lipopolysaccharide molecules, which is necessary for resistance to the antimicrobial peptide polymyxin B [44]. Another two-component system, CarRS, is known to confer polymyxin B resistance, but also to negatively regulate biofilm formation [65,66]. Here, we found that derived alleles (presumed *de novo* mutations within the host) in the HK VprB did not appear to affect polymyxin B resistance (S8 Fig), but did

reduce the ability of *V. cholerae* to form biofilms (Fig 5). It has been previously suggested that biofilm formation may be beneficial for survival in the aquatic environment, but detrimental to survival or colonization of mammalian hosts [45,46]. Therefore, the derived iSNV alleles may have been selected for reduced biofilm formation.

Our study provides an opportunity to compare *V. cholerae* evolutionary dynamics between Bangladesh, where cholera has been endemic for hundreds or thousands of years [67], and Haiti, where it was introduced in 2010. Our results suggest that selective pressures on *V. cholerae* may differ between Haiti and Bangladesh, as previously proposed [12]. In Bangladesh, we observed an excess of nonsynonymous (NS) mutations between patients (Table 3), suggest positive selection on protein sequences between patients, or efficient purifying selection purging nonsynonymous mutations within patients. In contrast to Bangladesh, where zero NS iSNVs were observed, we observed three NS iSNVs in Haiti, all within the same patient. Such differences between Haiti and Bangladesh need confirmation in a larger sample, and if confirmed could have different explanations. For example, if Haitian patients are less likely than Bangladeshis to have had prior exposure and immunity to cholera, perhaps cholera infections could last longer, or support larger *V. cholerae* population sizes within patients in Haiti, allowing more efficient positive selection within patients. Further sequencing of intra-host *V. cholerae* genomes, ideally in combination with clinical data on infection durations and outcomes, will be needed to test this hypothesis.

Conclusion

In summary, we have shown that small but measurable genomic changes occur in the *V. cholerae* genome during human infections. Changes in flexible gene content appear to accumulate more quickly than point mutations, although point mutations may be targets of natural selection. Both gene content variation and point mutations can have consequences for the phenotypes of within-patient *V. cholerae* populations, including clinically- and environmentally relevant traits like biofilm formation. Future studies will be necessary to determine the role of intra-host diversity in the evolution of antibiotic resistance, host adaptation, and the severity of disease in infected patients.

Materials and methods

Enrollment

To study cholera within-host diversity, stool samples were collected from five patients (B1 to B5) from Dhaka, Bangladesh, and three patients (H1 to H3) from Artibonite, Haiti. Between eight and 20 *V. cholerae* colonies were isolated from each patient, as described below. Patients in Bangladesh were enrolled at the icddr,b (International Center for Diarrheal Disease Research, Bangladesh) Dhaka Hospital. The icddr,b cares for more than 120,000 patients annually including approximately 20,000 with cholera. Patients presenting during 2013 with acute watery diarrhea were eligible for inclusion in this study if stool cultures were positive for *V. cholerae* as the only pathogen, if they were between 2 and 60 years of age, resided in or around Dhaka, and were without major comorbid conditions. In Haiti, samples were collected from patients presenting to St. Marc's Hospital, Arbonite, Haiti, with acute watery diarrhea in April 2013.

In addition to these eight patients, we included 21 "Time Course" patients (TC01 to TC21) from a surveillance program conducted by the icddr,b, between 2011 and 2013. For the Time Course samples, only one isolate was sequenced per patient.

Sample Processing

In Bangladesh, diarrheal samples were examined by dark field microscopy and if positive for *V. cholerae* on presentation, then stool was cultured overnight. Samples with visible *V. cholerae* growth were serologically confirmed by slide agglutination with specific monoclonal antibodies for Ogawa or Inaba serotypes [68]. Confirmed cholera

stool was stored in glycerol at -80°C and shipped to Massachusetts General Hospital, Boston. In Haiti, fresh stool from suspected cholera patients was stored in glycerol at -80°C and shipped from St. Marc's Hospital to Massachusetts General Hospital, Boston. Stool from both Haiti and Bangladesh was stored at MGH at -80°C and then streaked directly onto thiosulfate-citrate-bile salts-sucrose agar (TCBS), a medium selective for *V. cholerae*. After overnight incubation, twenty well-separated colonies were inoculated into 5 mL Luria-Bertani broth and grown at 37°C overnight. For each colony, 1mL of broth culture was stored at -80°C with 30% glycerol until DNA extraction. For patient 1, one of the colonies was re-streaked on a new TCBS plate and 12 colonies were selected as a control for culture-induced artifacts and sequencing errors.

DNA extraction and whole genome sequencing.

Bacterial stocks made from a single colony were grown in 1.5mL LB media with agitation at 37°C for 12 hours. Genomic DNA was extracted for each isolate using the Qiagen DNeasy Blood® and Tissue kit, using 1.5mL bacteria grown up in LB media. In order to obtain pure gDNA templates, an RNase treatment was followed by a purification with the MoBio PowerClean® Pro DNA Clean-Up Kit.

From 122 isolates retrieved from the eight patients, 66 genomic libraries were constructed using the Nextera DNA library kit, according to the manufacturer's protocol (Illumina) and were sequenced with the 250-bp paired-end v2 kit on the Illumina MiSeq. The remaining 56 libraries were prepared using the NEBNext Ultra II DNA library prep kit and sequenced on the Illumina HiSeq 2500 (paired-end 125 bp) at the Genome Québec sequencing platform (McGill University). Twelve isolates were sequenced in replicate using both methods. For details about isolates, sequencing and assembly see S5 Table.

Genome assembly

To exclude low-quality sequences, we filtered raw reads with Trimmomatic [69]. The 15 first bases of each read were trimmed and reads containing at least one base with a quality score of <30 were removed. *De novo* assembly was then performed for each strain using IDBA_ud v1.1.1[70].

Filtered reads were also mapped with Bowtie2 v2.2.5[71] to a total of 11 references: two annotated reference genomes, one from Haiti (2010EL-1786, accession no. NC_016445.1 and NC_016446.1), one from Bangladesh (MJ1236, accession number NC_012667 and NC_012668), and nine assembled genomes (one from each patient and one from the sub-cultured control colonies: B1C1-06, B1C7, B2C12, B3C12, B4C5, B5C10, H1C5, H2C3, H3C2). PCR duplicates were removed from mapping using the MarkDuplicates function of PICARD TOOLS v1.130 (<https://broadinstitute.github.io/picard/>) and the SAMtools view utility [72], and realignment around indels was performed using the Genome Analysis Toolkit v3.1.1, with default settings. To facilitate the identification of homologous regions among the eleven reference genomes, MJ1236 and the nine *de novo* assembled genomes were aligned against the 2010EL-1736 genome using the “move contig” option in Mauve v. 2.4.0 [73], with default parameters.

Variant calling and annotation

We used SAMtools v1.3 and BCFtools v1.1 to call SNPs and indels from mapping, requiring a minimum mapping quality of 30 and a minimum base quality of 20. Resulting SNPs and indels were then filtered by quality score (<20), depth of coverage

(<10) and FQ scores (<0, lower values indicate agreement between reads) with VCFlib [74]. For each strain, we also filtered variable positions that were not retrieved in all the mappings against all 11 references, by performing reciprocal BLAST of the 50 nucleotides upstream and downstream of each variable position and comparing them. Only matches with >95% identity were kept and multiple matches were excluded as possible duplications and repeated elements. After applying these filters, we compared the genomes of the 12 replicate clones from the strain B1C1, to control for possible SNPs due to mutations during culture, or sequencing errors. We also removed positions that were called as variable when reads from one isolate were mapped to the assembly from the same isolate. We considered these SNPs as potential sequencing, mapping or assembly errors. Using these filters, we generated a list of high-quality SNPs (hqSNPs). From this list, we identified intra-host single nucleotide variants (iSNVs) as SNPs that were polymorphic among isolates from the same patient. No iSNVs were identified among the control colonies (the 12 replicate clones subcultured from B1C1).

Annotations were available for the MJ1236 reference genome and were retrieved from GenBank files (Chromosome 1: CP001485.1; chromosome 2: CP001486.1). Variants from the core genome were classified in three categories: intergenic (INT) when falling outside of a coding region, synonymous (S) when affecting the nucleotide sequence of at least one gene but not its amino-acid sequence; or non-synonymous (NS) when affecting the amino acid sequence of at least one gene.

Estimation of effective population sizes (N_e)

To estimate effective population size within each patient, we used the formula $N_e = \theta/2 \mu$, where N_e is the effective population size, θ is a measure of genetic diversity

and μ is the mutation rate [75]. We assumed a mutation rate of $\mu=1/300$ per genome per generation [76]. We report the estimated N_e for each patient in S1 Table, using both Waterson's estimator (θ_W or S) and Tajima's estimator (θ_T or π) as measures of genetic diversity. We calculate these estimators as describe in Tajima 1989 [75].

Characterization of the flexible genome

From assemblies, we annotated the genomes using the RAST pipeline [77] with default parameters. Predicted proteins were used as input for the OrthoFinder software [78] to predict orthologous gene families. These orthologous gene families were classified into three different categories: multiple gene families (>1 copy per genome, on average), single copy genes (exactly one copy per genome) and flexible genome (<1 copy per genome).

Presence or absence of genes

As absence of a given gene in a genome could be an artifact of the assembly process, we confirmed the absence of each gene family using the raw reads (Fig S2). A representative catalogue of the flexible genome protein sequences was built using the cd-hit program with a 90% similarity threshold [79]. We used sequences of the catalogue as queries for a blastn search on raw reads. We considered a gene family to be present in a given genome if the average coverage of the gene was greater or equal to 1X. This coverage threshold allowed us to detect every single gene in the gene catalogue (S3 Fig) while observing no variability among the control strains. To calculate coverage, we

summed the length of all reads matching a given query over a minimal length of 100 nucleotides and a minimal identity of 97%, and divided by the gene length.

Inference of flexible gene origins

In order to estimate the origin of the flexible genome gene pool, we performed an extended phylogenetic analysis of all 155 flexible gene families. The flexible gene catalogue served as query for a blastp search against the NCBI database. For each gene, we selected the first 200 hits matching with an E-value below 1E-05. The hit sequences were aligned using muscle with default parameters [80] and phylogenetic analysis performed with the FastTree algorithm using default parameters [81]. Finally, we screened nexus formatted trees from FastTree to identify the closest relative sequences of each gene in our dataset. This allowed us to classify the flexible genes into three mutually exclusive categories: first, genes whose closest relative originated from the *Vibrio cholerae* gene pool; second, genes whose closest relative is from *Vibrio* but not *cholerae* (i.e non-cholera *Vibrio* strains); and finally, the third category includes genes whose closest relative was outside the genus *Vibrio*. Trees were also displayed automatically using the FigTree java program for a manual inspection (<http://tree.bio.ed.ac.uk/software/figtree/>). To guard against false-positive inference of horizontal gene transfers from non-*V. cholerae*, a negative control was performed. We repeated the blastp/FastTree procedure using 155 genes extracted randomly from core genes in our study. As expected, these were all assigned *Vibrio cholerae* taxonomic affiliations.

Phylogenetic evolutionary inference and root-to-tip regression.

For all phylogenetic analyses, we did not consider the Integrative Conjugative Element (ICE), as the evolution of this region, a mutation hotspot, mostly reflects recombination events (S1 Fig). The ICE was defined as the region of MJ1236 chromosome 1 located between positions 87776 and 193789 (after reverse-complementation of MJ1236), according to a previous study [82]. A final alignment of 201 concatenated hqSNPs was generated from the core genome of the 35 genotypes (TC01-TC21 plus one to four unique genotypes per patient) and used for phylogenetic analysis. We used Seaview v.4.5.4 [83] to generate a maximum-likelihood phylogeny, employing a general time reversible (GTR) substitution model with four rate classes and SPR branch-swapping. All sites being variable in the alignment, we did not consider the proportion of invariable sites. To measure the extent of clock-like structure in our data set, we performed a linear regression of root-to-tip distances against dates of isolation, using the TempEst software [84]. Substitution rates and divergence times were then estimated using BEAST package v.1.8.3 [40], with XML-input files manually modified to specify the number of constant sites. For this analysis, we tested and compared both strict and uncorrelated lognormal molecular clock models and three coalescent models (exponential growth coalescent, constant-size coalescent and Bayesian Skyline demographic models), resulting in 6 possible model combinations. For all of them we used a GTR + G nucleotide substitution model and the sampling times for calibration. All of these combinations were run using 10,000,000 MCMC chains, with 10% burn-in and sampling every 5,000 generations. We used Tracer v.1.6 to ensure proper mixing, with all parameters having an effective sample size > 200. To select the best molecular clock and coalescent models, we estimate the marginal likelihoods for each combination via path-sampling, and we compared them with Bayes factors [85]. Divergence time,

substitution rates and resulting tree were reported from the models with the highest marginal likelihood.

Tests for natural selection over a three-year period

To distinguish between positive selection, purifying selection, or neutral evolution of protein-coding sequences, we considered variation in the proportion of non-synonymous hqSNPs (p_{NS}) in the *V. cholerae* core genome. Specifically, we evaluated how p_{NS} varied over time (a three-year period from 2011 to 2013) and among branches of the phylogenetic tree. We considered $N=136$ hqSNPs (excluding the ICE region, a mutation hotspot; S1 Fig) that varied among 21 isolates sampled over three years in Bangladesh (patients TC01-TC21) and the 122 isolates sampled from five patients from Bangladesh (patients B1-B5) and three from Haiti (Patient H1-H3). For these analyses, we excluded iSNVs by considering only the most frequent haplotype found within each patient, assumed to be ancestral.

We first tested whether the overall observed p_{NS} is to be expected under a simple neutral model of evolution. We performed 1,000 simulations of N mutations randomly distributed across the core genome of the MJ1236 reference. For each simulation, we re-estimated the relative proportions of intergenic (p_I), synonymous (p_S) and non-synonymous (p_{NS}) mutations, using annotations available for MJ1236 (GenBank: CP001485-6). We controlled for potential effects of genome-wide nucleotide composition by comparing simulations with or without imposed GC content of mutated positions (i.e. matching the GC content observed among the N real hqSNPs). We also controlled for any bias in the transition:transversion ratio by comparing simulations with or without imposed transition rate (i.e. matching the ratio observed among the N real hqSNPs). We

considered that the core genome evolved under positive selection when the observed p_{NS} was higher than in at least 97.5% of simulations. We considered that the core genome evolved under purifying selection when the observed p_{NS} was lower than in at least 97.5% of simulations. We failed to reject neutral evolution when the observed p_{NS} fell within the 95% range of simulations.

Second, we tested whether fixed core genome hqSNPs were distributed evenly across branches of the phylogeny. Specifically, we asked whether substitution rates differ between Bangladesh and Haiti, or between long internal branches and the shorter, more recent tips of the tree where selection may have had insufficient time to act. To do so, we defined 3 well-separated monophyletic clades based on the evolutionary tree of the *V. cholerae* core genome. The tree was built in MEGA5 using a maximum composite likelihood model [86] (S6B Fig). We distinguished hqSNPs that were fixed among clades (corresponding to long branches) from those that are variable within clades (the tips of the tree). We hypothesized that if differences in p_{NS} are observed between vs. within clades, this could suggest that selection or substitution rates vary among clades and over time. To test this, we performed 10,000 random permutations of hqSNPs among branches of the evolutionary tree, and for each permutation, we re-estimated p_I , p_S and p_{NS} within clade. For each monophyletic clade, we considered that the substitution rate was higher than expected by chance when the observed p_{NS} was higher than in at least 97.5% of permutations. We considered that the substitution rate was lower than expected by chance when the observed p_{NS} was lower than in at least 97.5% of permutations. We failed to reject neutral evolution when the observed p_{NS} fell within the 95% range of permutations.

Tests for natural selection within and between patients

To investigate the role of natural selection within versus between patients from Bangladesh and Haiti, we performed the McDonald-Kreitman test [42] to test the neutral hypothesis that nonsynonymous (NS) to synonymous (S) substitution ratios remained constant over evolutionary time (within vs. between hosts). Specifically, we computed the Fixation Index (equivalent to an odds ratio statistic) as the NS:S ratio between patients (fixed SNPs) divided by the NS:S ratio within patients (iSNVs). Significant deviations of the Fixation Index from neutral expectation were evaluated using Fisher's exact test.

We then tested whether iSNVs are equally distributed among patients, and if any patient contained an excess (possibly due to positive or balancing selection) or a deficit (possibly due to efficient purifying selection) of NS iSNVs. To do so, we performed permutations of iSNVs among the eight patients and estimated expected iSNV frequencies (F) and p_{NS} per patient (B1-B5; H1-H3) and region (Bangladesh and Haiti). We first assigned each of the 122 strains collected from the eight patients to one of the four following haplotypes: H_0 as strains having the most frequent haplotype found within each patient, and assumed to be ancestral for that patient; H_{INT} as strains having one intergenic iSNV; H_S as strains having one synonymous iSNV and H_{NS} as strains having one synonymous iSNV. (No strains were observed with more than one iSNV, so these haplotypes are sufficient to model the observed intra-patient diversity). We then performed 10,000 random permutations of the four haplotypes among the 122 strains. When an iSNV was shared between two or more strains within a patient, but not observed in other patients, we ensured that these strains were always assigned to the same patient during permutations. For each simulation and for each patient or region,

we reported the total iSNV relative frequency (F = number of strains containing iSNVs / the total number of strains sequenced for that patient or region) and p_{NS} , defined as above. For each region and each patient, we considered that F and p_{NS} were higher or lower than expected by chance when the observed values were respectively higher or lower than in at least 97.5% of permutations. We concluded that F and p_{NS} were consistent with our neutral model when they fell within the 95% range of permutations.

Sensor histidine kinase protein conservation analysis

Of the three NS iSNVs in patient H1, two occur in the same gene, a predicted sensory histidine kinase (NCBI accession number ACQ61177, from the reference genome MJ-1236). We sought to determine whether these two NS iSNVs occurred in conserved or variable peptides. To do so, we retrieved the 500 best matches (top BLAST hits) for the ACQ61177 protein sequence in NCBI GenBank using BLASTp. From these 500 homologous sequences, we removed identical (duplicate) sequences, and those that were truncated at the N or the C terminus, resulting in 465 unique homologous sequences. We then determined whether 4-amino-acid (4aa) peptides surrounding the mutated residues were conserved among these sequences. We defined a simple conservation score as the proportion of homologs having the reference peptide (from *V. cholerae* MJ1236). This score could be influenced by a biased sample of sequences in GenBank, and thus represents a rough estimate of conservation. In order to minimize the effect of peptide convergence, we did not consider 4aa motifs that were found at least twice in at least one sequence, which was not the case for any of the peptide motifs affected by the two observed iSNVs.

Liquid culture and biofilm growth assays of isolates from patients H1 and H2

We performed *in vitro* experiments to identify phenotypic consequences of iSNVs in patient H1, or flexible gene content variation in patient H2. Isolates were grown in 4 ml of LB broth with agitation at 30°C, and optical densities were measured at 600 nm using a spectrophotometer every hour for 12 hours, revealing no significant differences in growth between strains. To test for biofilm production, we grew the same strains in 200 µl of LB in a 96-well plate, without agitation at 30°C for 48 hours. Controls included empty wells, wells with LB only, and wells with a *V. cholerae* strain with an in-frame deletion in the *vpsA* (Vibrio polysaccharide A) gene, which results in reduction of biofilm production[47]. A 0.1% solution of Crystal Violet was used to stain for biofilm adherent to the well. Biofilms were dissolved in ethanol at the end of the assay, and the optical density was measured at 595 nm using spectrophotometry. Experiments were performed in replicates of four to twelve.

Polymyxin B MIC assay on patient H1 isolates

Strains H1C1, H1C5 and H1C6 were grown overnight at 30°C on LB agar and then cultures were diluted 1:100 in fresh LB medium. Cells were grown to mid-exponential growth and diluted 1:10, and an aliquot was plated on LB agar. Polymyxin B Etest gradient strips (AB Biodisk) were applied onto inoculated plates and incubated at 37°C, and the MICs were evaluated after 16 h.

795

796 **Acknowledgments**

797 We thank Simone Perinet and Meri Debela for their technical assistance, Paula Watnick
798 who contributed materials to the study, and Salvador Almagro-Moreno for constructive
799 comments on the manuscript. We also thank the Zanmi Lasante staff at the enteric
800 microbiology laboratory in St. Marc, Haiti. Finally, we are grateful to the people of Dhaka
801 where our study was undertaken; to the field, laboratory and data management staff who
802 provided tremendous effort to make the study successful; and to the people who
803 provided valuable support in our study. ICDDR,B gratefully acknowledges the
804 Government of the People's Republic of Bangladesh; Global Affairs Canada (GAC);
805 Swedish International Development Cooperation Agency (Sida) and the Department for
806 International Development, (UKAid). We declare that we have no competing financial
807 interest.

808 **Funding information**

811 This study was supported by CIHR (Canadian Institutes of Health Research) and the
812 Canada Research Chairs program, the ICDDR,B: Centre for Health and Population
813 Research, grants AI099243 (J.B.H and L.C.I), AI103055 (J.B.H and F.Q), AI106878
814 (E.T.R and F.Q.), AI058935 (E.T.R, S.B.C and F.Q.), T32A1070611976 and
815 K08AI123494 (A.A.W.) from the National Institutes of Health, and the Robert Wood
816 Johnson Foundation Harold Amos Medical Faculty Development Program (R.C.C.).

Figures

Fig 1. Culture and sequencing of *Vibrio cholerae* isolates from eight acutely infected patients. To study within-patient evolution, we cultured stool samples from five patients from Bangladesh (B1 to B5) and three patients from Haiti (H1 to H3), on selective media. We isolated between eight and 20 colonies from each patient and sequenced them separately. For patient B1, we performed a sub-culture of one isolate (dotted outline) and sequenced twelve of these new isolates, as a control for cultured-induced and sequencing artifacts. We independently called variants, compared them between isolates within each patient to identify the intra single nucleotide variants (iSNVs, colored circles) and determined whether they were intergenic (i), synonymous (S), or non-synonymous (NS) mutations.

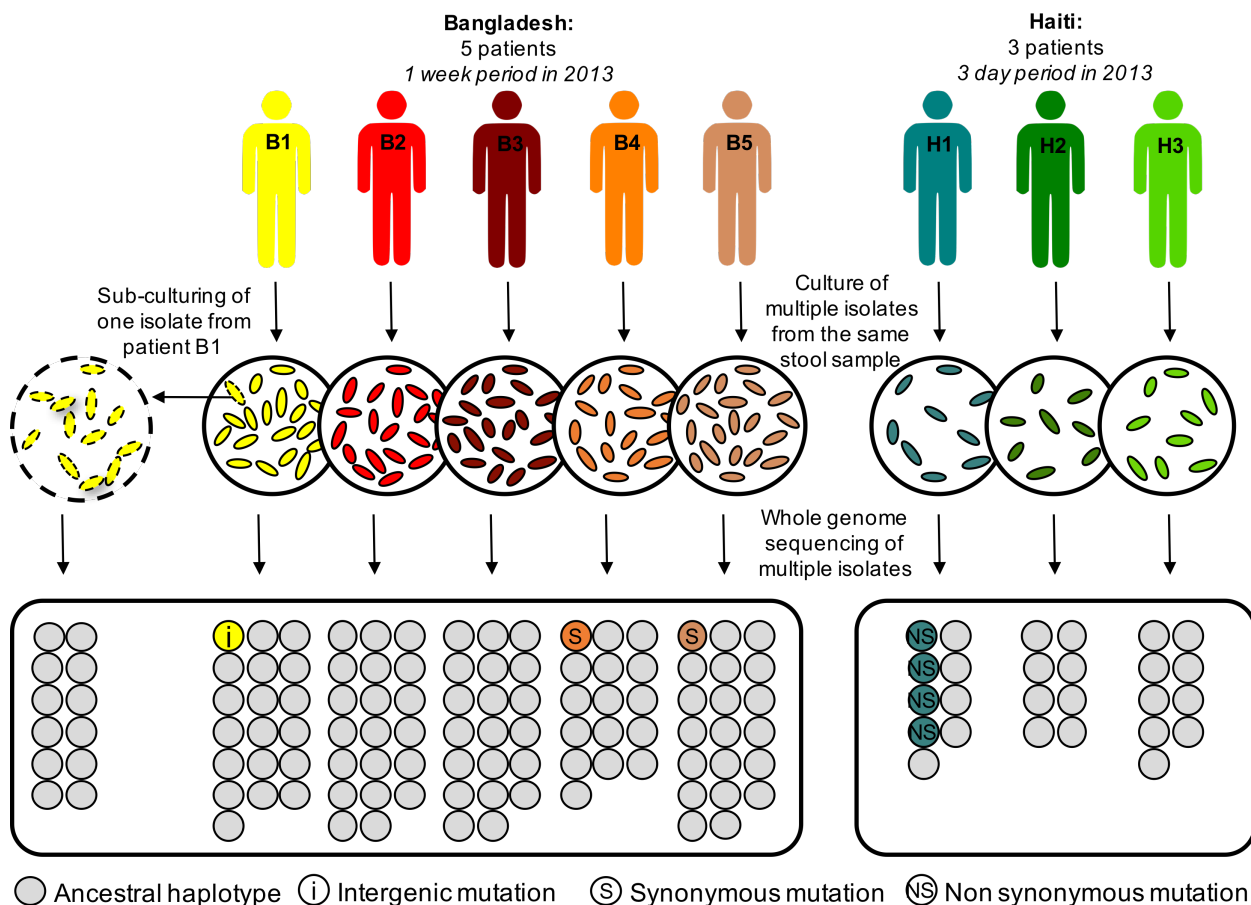


Fig 2. Presence/absence profile and taxonomic affiliation of gene families in the flexible genome. Red in the heatmap indicates gene presence; black indicates absence. Each column shows the presence/absence profile for a unique gene family. The heatmap is ordered by patient along the vertical axis. B1C1 is the control, subcultured from B1, and contains no flexible genome variation. The horizontal axis is ordered by hierarchical clustering, yielding four clusters: A, B, C and D. The taxonomic affiliation of each gene family (best BLAST hit) is indicated with red dots above the heatmap.

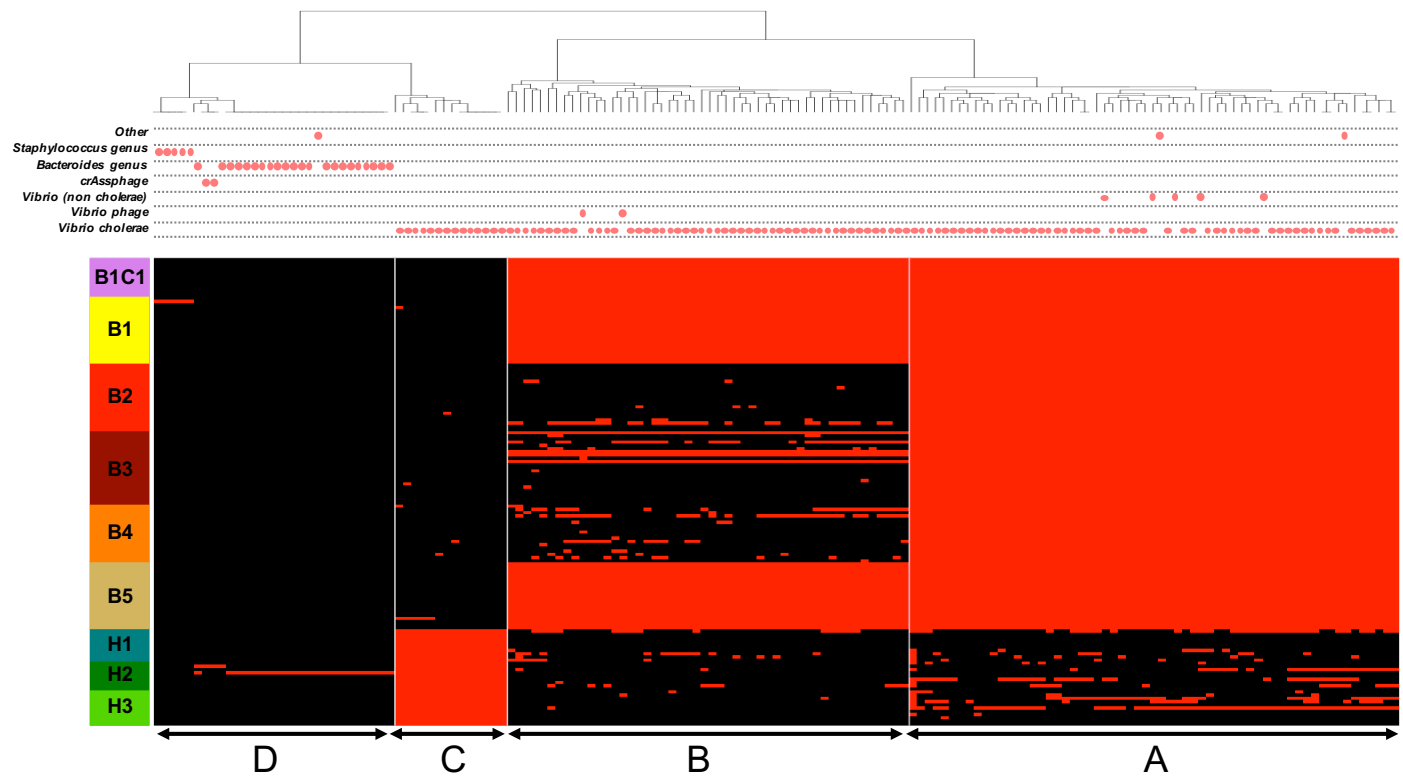


Fig 3. Bayesian phylogenetic tree of 35 *V. cholerae* genotypes sampled over three years in Bangladesh and Haiti. The maximum clade credibility tree represents the genealogy of sequences in the study, reconstructed from concatenated hqSNPs, using BEAST. Colored squares (shades of blue and purple) represent the time-course isolates collected from Bangladeshi patients from March 2011 to December 2013 (one isolate per patient). Patients for whom we measured intra-host variation (B1-B5 and H1-H3) are shown as circles. Filled circles indicate the putative ancestral genotype, and empty circles indicate putatively derived iSNVs. The median node age and divergence date in months and years are indicated at the nodes. The blue bars represent the 95% HPD intervals for divergence time estimates, and posterior probabilities are represented on the branches.

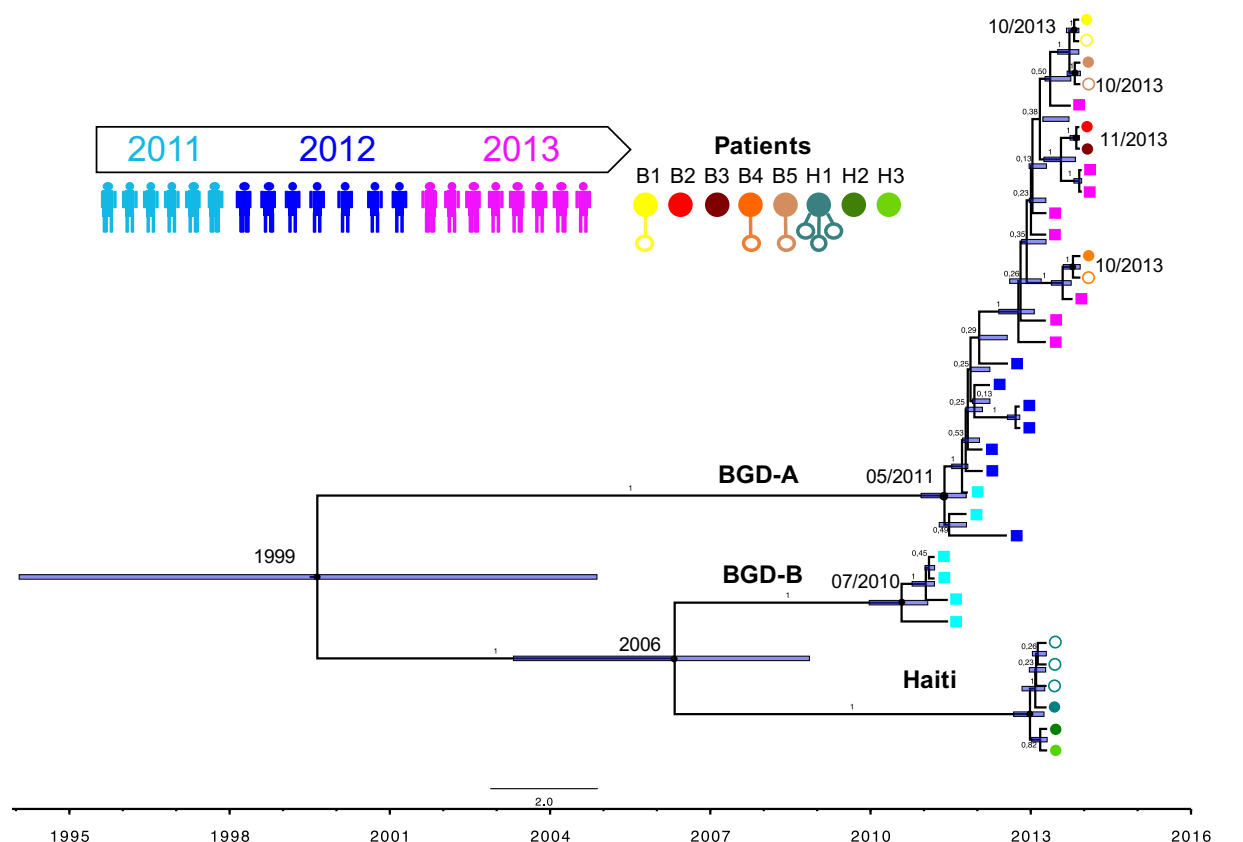


Fig 4. Significant excess of non-synonymous iSNVs in patient H1. (A) Distribution of 122 *V. cholerae* strains containing different categories of iSNVs (I: intergenic; S: synonymous; Nsyn: non-synonymous) or no detectable iSNVs, according to geographic region (BGD: Bangladesh; HTI: Haiti). Patients from Haiti have a significant excess of Nsyn iSNVs (red; $p \leq 0.01$; 10,000 random permutations of strains among regions). (B) Distribution of 122 *V. cholerae* strains containing different iSNVs, or no detectable iSNVs, by patient. Patient H1 has a significant excess of strains with Nsyn iSNVs ($p \leq 0.001$; 10,000 random permutations of strains among patients).

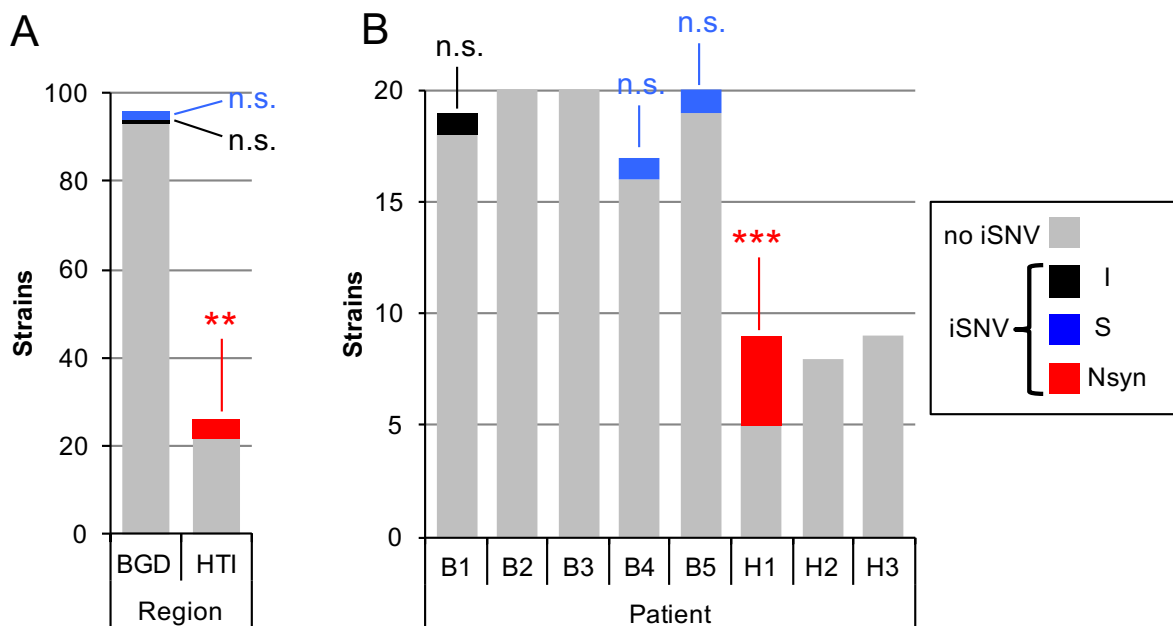
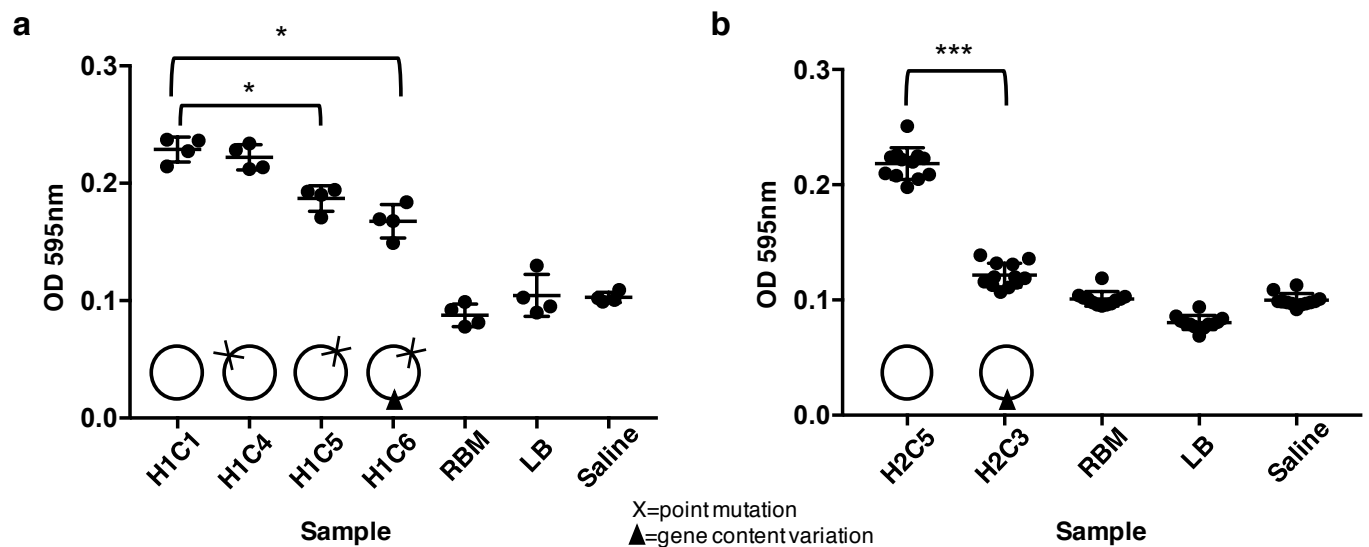


Fig 5. Biofilm formation of strains from patients H1 and H2. Optical density was measured for four to 12 replicates of each strain, after 48h of growth at 30°C. Statistical comparisons were made using a non-parametric Mann-Whitney test ($*P < 0.05$, $***P < 0.0001$). Circles represent genomes with either variation in gene content (dark triangle) or iSNV variation (cross). (a) Strains from patient H1. Isolate H1C1 represents the ancestral genotype, H1C4 has a nonsynonymous mutation in a transcriptional regulator gene, and H1C5 and H1C6 have different nonsynonymous mutations in the same gene, the histidine kinase gene. (b) Strains from patient H2. Isolate H2C5 represents the ancestral genotype, with no variation in the gene content, and H2C3 harbors a plasmid. RBM is a biofilm knockout strain, and LB and saline are negative controls.



Tables

Table 1. Nucleotide and amino acid changes identified in the *V. cholerae* core genome.

Region	Type	Strains	Reference nucleotide	Alternative nucleotide	NS/S	Reference amino acid	Alternative amino acid	Gene annotation	Patient allele frequency
Bangladesh	iSNV	B1C19	G	A	intergenic	-	-	-	1/19
Bangladesh	iSNV	B4C12	G	A	S	D	D	RNA polymerase sigma factor	1/17
Bangladesh	iSNV	B5C11	A	C	S	V	V	toxin co-regulated pilus biosynthesis protein F	1/20
Haiti	iSNV	H1C5	C	A	NS	R	L	sensor histidine kinase	1/9
Haiti	iSNV	H1C4 H1C8	C	T	NS	R	H	TetR family transcriptional regulator	2/9
Haiti	iSNV	H1C6	C	A	NS	R	L	sensor histidine kinase	1/9
Bangladesh	Patient	B2 + B3	A	G	intergenic	-	-	-	20/20 20/20
Bangladesh	Patient	B4	G	A	NS	G	R	TatD family hydrolase	17/17
Bangladesh	Patient	B4	G	T	NS	A	T	hypothetical protein	17/17
Bangladesh	Patient	B2 + B3	C	T	intergenic	-	-	-	20/20 20/20
Bangladesh	Patient	B4	G	A	intergenic	-	-	-	17/17
Bangladesh	Patient	B4	C	T	NS	P	L	PTS system mannitol-specific EIICBA component	17/17
Bangladesh	Patient	B2 + B3	C	T	NS	P	S	LacI family transcription regulator	20/20 20/20
Bangladesh	Patient	B1 + B5	A	G	N	A	V	bifunctional purine biosynthesis protein PurH	19/19 20/20

Mutations segregating within patients are denoted iSNVs; Mutations fixed between patients are

denoted 'Patient.' Patient allele frequency shows the allele frequency of the alternative (minor)

allele. NS=nonsynonymous; S=synonymous.

Table 2. Flexible gene content variation within and between patients.

Patients	#genes fixed within	#genes variable within	#singletons
	patients	patients	
B1	111	11	5
B2	61	35	0
B3	61	51	0
B4	61	49	0
B5	111	5	0
H1	14	68	0
H2	14	103	25
H3	14	63	0

Singletons are defined as genes only found in one strain, and are also counted as variable genes within patients. Genes fixed within patients are present in all isolates from a patient, but are absent in at least one other isolate in the study.

Table 3. McDonald-Kreitman test for differential selection within and between patients.

Population	Bangladesh		Haiti	
	Nsyn.	Syn.	Nsyn.	Syn.
Polymorphic (within patient)	0	2	3	0
Fixed (between patients)	5	0	0	0
<i>Fisher exact test, p = 0.048</i>		<i>Fisher exact test, p = 1</i>		

Counts of non-synonymous (Nsyn.) and synonymous (Syn) polymorphic sites (within patient iSNVs) and fixed sites (between patients) for Bangladeshi and Haitian patients.

References

1. Harris JB, LaRocque RC, Qadri F, Ryan ET, Calderwood SB. Cholera. *Lancet*. 2012;379: 2466–2476. doi:10.1016/S0140-6736(12)60436-X
2. Reidl J, Klose KE. *Vibrio cholerae* and cholera: out of the water and into the host. *FEMS Microbiol Rev*. 2002;26: 125–139. doi:10.1111/j.1574-6976.2002.tb00605.x
3. Chatterjee SN, Chaudhuri K. Lipopolysaccharides of *Vibrio cholerae*. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2003;1639: 65–79. doi:10.1016/j.bbadis.2003.08.004
4. Seed KD, Faruque SM, Mekalanos JJ, Calderwood SB, Qadri F, Camilli A. Phase variable O antigen biosynthetic genes control expression of the major protective antigen and bacteriophage receptor in *Vibrio cholerae* O1. *PLoS Pathog*. 2012;8: e1002917–13. doi:10.1371/journal.ppat.1002917
5. Wilson DJ. Insights from Genomics into Bacterial Pathogen Populations. Chitnis CE, editor. *PLoS Pathog*. 2012;8: e1002874–9. doi:10.1371/journal.ppat.1002874
6. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*. 2011;477: 462–465. doi:10.1038/nature10392
7. Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, et al. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci USA*. 2009;106: 15442–15447. doi:10.1073/pnas.0907787106
8. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The origin of the Haitian cholera outbreak strain. *N Engl J Med*. 2011;364: 33–42. doi:10.1056/NEJMoa1012928
9. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, et al. Population Genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. *mBio*. 2011;2: e00157–11–e00157–11. doi:10.1128/mBio.00157-11
10. Katz LS, Petkau A, Beaulaurier J, Tyler S, Antonova ES, Turnsek MA, et al. Evolutionary Dynamics of *Vibrio cholerae* O1 following a Single-Source Introduction to Haiti. *mBio*. 2013;4: e00398–13–e00398–13. doi:10.1128/mBio.00398-13
11. Orata FD, Keim PS, Boucher Y. The 2010 Cholera Outbreak in Haiti: How Science Solved a Controversy. *PLoS Pathog*. 2014;10: e1003967–5. doi:10.1371/journal.ppat.1003967
12. Azarian T, Ali A, Johnson JA, Mohr D, Prosperi M, Veras NM, et al. Phylodynamic Analysis of Clinical and Environmental *Vibrio cholerae* Isolates from Haiti Reveals Diversification Driven by Positive Selection. *mBio*. 2014;5: e01824–14–11. doi:10.1128/mBio.01824-14
13. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*. 2009;10: 540–550. doi:10.1038/nrg2583

- 933 14. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial
934 pathogens. 2016;14: 150–162. doi:10.1038/nrmicro.2015.13
- 935 15. Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders Accurate
936 Reconstruction of Transmission Networks from Genomic Distance Data. PLoS Comput
937 Biol. 2014;10: e1003549. doi:10.1371/journal.pcbi.1003549
- 938 16. Didelot X, Gardy J, Colijn C. Bayesian Inference of Infectious Disease Transmission from
939 Whole-Genome Sequence Data. Molecular Biology and Evolution. 2014;31: 1869–1879.
940 doi:10.1093/molbev/msu121
- 941 17. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen
942 transmission. Current Opinion in Microbiology. 2015;23: 62–67.
943 doi:10.1016/j.mib.2014.11.004
- 944 18. Patra R, Chattopadhyay S, De R, Ghosh P, Ganguly M, Chowdhury A, et al. Multiple
945 infection and microdiversity among *Helicobacter pylori* isolates in a single host in India.
946 PLoS ONE. 2012;7: e43370. doi:10.1371/journal.pone.0043370
- 947 19. Seed KD, Yen M, Shapiro BJ, Hilaire IJ, Charles RC, Teng JE, et al. Evolutionary
948 consequences of intra-patient phage predation on microbial populations. eLife. 2014;3:
949 W400–10. doi:10.7554/eLife.03497
- 950 20. Kendall EA, Chowdhury F, Begum Y, Khan AI, Li S, Thierer JH, et al. Relatedness of
951 *Vibrio cholerae* O1/O139 isolates from patients and their household contacts, determined
952 by multilocus variable-number tandem-repeat analysis. Journal of Bacteriology. ;
953 2010;192: 4367–4376. doi:10.1128/JB.00698-10
- 954 21. Faruque SM, Chowdhury N, Kamruzzaman M, Dziejman M, Rahman MH, Sack DA, et al.
955 Genetic diversity and virulence potential of environmental *Vibrio cholerae* population in a
956 cholera-endemic area. Proceedings of the National Academy of Sciences. 2004;101:
957 2123–2128. doi:10.1073/pnas.0308485100
- 958 22. Rashed SM, Azman AS, Alam M, Li S, Sack DA, Morris JG Jr., et al. Genetic Variation of
959 *Vibrio cholerae* during Outbreaks, Bangladesh, 2010–2011. Emerg Infect Dis. 2014;20:
960 54–60. doi:10.3201/eid2001.130796
- 961 23. Al-Hajj SAM, Akkerman O, Parwati I, al-Gamdi S, Rahim Z, van Soolingen D, et al.
962 Microevolution of *Mycobacterium tuberculosis* in a Tuberculosis Patient. Journal of
963 Clinical Microbiology. ; 2010;48: 3813–3816. doi:10.1128/JCM.00556-10
- 964 24. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, et al. Microevolution
965 of *Helicobacter pylori* during Prolonged Infection of Single Hosts and within Families.
966 PLoS Genet. 2010;6: e1001036. doi:10.1371/journal.pgen.1001036
- 967 25. McAdam PR, Holmes A, Templeton KE, Fitzgerald JR. Adaptive Evolution of
968 *Staphylococcus aureus* during Chronic Endobronchial Infection of a Cystic Fibrosis
969 Patient. Schrijver I, editor. PLoS ONE. 2011;6: e24301–5.
970 doi:10.1371/journal.pone.0024301
- 971 26. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Lerner-Svensson H, et al. Within-
972 Host Evolution of *Staphylococcus aureus* during Asymptomatic Carriage. Aziz RK, editor.

- 973 PLoS ONE. 2013;8: e61319–14. doi:10.1371/journal.pone.0061319
- 974 27. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, et al. Genetic
975 variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of
976 selective pressures. Nat Genet. 2014;46: 82–87. doi:10.1038/ng.2848
- 977 28. Price EP, Sarovich DS, Mayo M, Tuanyok A, Drees KP, Kaestli M, et al. Within-Host
978 Evolution of Burkholderia pseudomallei over a Twelve-Year Chronic Carriage Infection.
979 mBio. 2013;4: e00388–13–e00388–13. doi:10.1128/mBio.00388-13
- 980 29. Nelson EJ, Harris JB, Morris JG, Calderwood SB, Camilli A. Cholera transmission: the
981 host, pathogen and bacteriophage dynamic. Nat Rev Microbiol 2009;: 1–10.
982 doi:10.1038/nrmicro2204
- 983 30. David LA, Weil A, Ryan ET, Calderwood SB, Harris JB, Chowdhury F, et al. Gut Microbial
984 Succession Follows Acute Secretory Diarrhea in Humans. mBio. 2015;6: e00381–15.
985 doi:10.1128/mBio.00381-15
- 986 31. Abel S, Abel zur Wiesch P, Chang H-H, Davis BM, Lipsitch M, Waldor MK. Sequence tag-
987 based analysis of microbial population dynamics. Nature Methods. 2015;12: 223–6– 3 p
988 following 226. doi:10.1038/nmeth.3253
- 989 32. Keymer DP, Boehm AB. Recombination Shapes the Structure of an Environmental Vibrio
990 cholerae Population. Applied and Environmental Microbiology. 2011;77: 537–544.
991 doi:10.1128/AEM.02062-10
- 992 33. Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Baptiste E, et al. Local Mobile
993 Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global Vibrio
994 cholerae Populations. mBio. 2011;2: e00335–10–e00335–10. doi:10.1128/mBio.00335-10
- 995 34. LaRocque RC, Harris JB, Dziejman M, Li X, Khan AI, Faruque ASG, et al. Transcriptional
996 Profiling of Vibrio cholerae Recovered Directly from Patient Specimens during Early and
997 Late Stages of Human Infection. Infection and Immunity. 2005;73: 4488–4493.
998 doi:10.1128/IAI.73.8.4488-4493.2005
- 999 35. Draper JL, Hansen LM, Bernick DL, Abedrabbo S, Underwood JG, Kong N, et al. Fallacy
1000 of the Unique Genome: Sequence Diversity within Single Helicobacter pylori Strains.
1001 mBio. 2017;8: e02321–16. doi:10.1128/mBio.02321-16
- 1002 36. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence
1003 assembly. Nature Methods. 2011;8: 61–65. doi:10.1038/nmeth.1527
- 1004 37. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive
1005 Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Comput
1006 Biol. 2014;10: e1003998. doi:10.1371/journal.pcbi.1003998
- 1007 38. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing
1008 and measuring bias in sequence data. Genome Biol. 2013;14: R51. doi:10.1186/gb-2013-
1009 14-5-r51
- 1010 39. Darmon E, Leach DRF. Bacterial Genome Instability. Microbiology and Molecular Biology
1011 Reviews. ; 2014;78: 1–39. doi:10.1128/MMBR.00035-13

- 1012 40. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and
1013 the BEAST 1.7. *Molecular Biology and Evolution*. 2012;29: 1969–1973.
1014 doi:10.1093/molbev/mss075
- 1015 41. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale
1016 rates of evolutionary change in bacteria. *Microbial Genomics*. 2016;2.
1017 doi:10.1099/mgen.0.000094
- 1018 42. Adaptive Protein Evolution at the Adh Locus in *Drosophila* - ProQuest. *Nature*. 1991.
- 1019 43. Ramos JL, Martínez-Bueno M, Molina-Henares AJ, Terán W, Watanabe K, Zhang X, et
1020 al. The TetR Family of Transcriptional Repressors. *Microbiology and Molecular Biology*
1021 *Reviews*. ; 2005;69: 326–356. doi:10.1128/MMBR.69.2.326-356.2005
- 1022 44. Herrera CM, Crofts AA, Henderson JC, Pingali SC, Davies BW, Trent MS. The *Vibrio*
1023 *cholerae* VprA-VprB Two-Component System Controls Virulence through Endotoxin
1024 Modification. *mBio*. 2014;5: e02283–14. doi:10.1128/mBio.02283-14
- 1025 45. Almagro-Moreno S, Pruss K, Taylor RK. Intestinal Colonization Dynamics of *Vibrio*
1026 *cholerae*. Bliska JB, editor. *PLoS Pathog*. 2015;11: e1004787–11.
1027 doi:10.1371/journal.ppat.1004787
- 1028 46. Shapiro BJ, Levade I, Kovacicova G, Taylor RK, Almagro-Moreno S. Origins of pandemic
1029 *Vibrio cholerae* from environmental gene pools. *Nature Microbiology*. 2016;; 1–6.
1030 doi:10.1038/nmicrobiol.2016.240
- 1031 47. Smith DR, Maestre-Reyna M, Lee G, Gerard H, Wang AHJ, Watnick PI. In situ proteolysis
1032 of the *Vibrio cholerae* matrix protein RbmA promotes biofilm recruitment. *Proceedings of*
1033 *the National Academy of Sciences*. 2015;112: 10491–10496.
1034 doi:10.1073/pnas.1512424112
- 1035 48. Eldholm V, Norheim G, Lippe von der B, Kinander W, Dahle UR, Caugant DA, et al.
1036 Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible
1037 ancestor in a single patient. *Genome Biol*. 2014;15: 490. doi:10.1186/s13059-014-0490-3
- 1038 49. Borgeaud S, Metzger LC, Scrignari T, Blokesch M. The type VI secretion system of *Vibrio*
1039 *cholerae* fosters horizontal gene transfer. *Science*. American Association for the
1040 Advancement of Science; 2015;347: 63–67. doi:10.1126/science.1260064
- 1041 50. Faruque SM, Mekalanos JJ. Phage-bacterial interactions in the evolution of toxigenic
1042 *Vibrio cholerae*. *Virulence*. 2014;3: 556–565. doi:10.4161/viru.22351
- 1043 51. Folster JP, Katz L, McCullough A, Parsons MB, Knipe K, Sammons SA, et al. Multidrug-
1044 Resistant IncA/C Plasmid in *Vibrio cholerae* from Haiti. *Emerg Infect Dis*. 2014;20: 1951–
1045 1953. doi:10.3201/eid2011.140889
- 1046 52. Hazen TH, Pan L, Gu JD, Sobecky PA. The contribution of mobile genetic elements to the
1047 evolution and ecology of *Vibrios*. *FEMS Microbiology Ecology*. 2010;74: 485–499.
1048 doi:10.1111/j.1574-6941.2010.00937.x
- 1049 53. Das B, Pazhani GP, Sarkar A, Mukhopadhyay AK, Nair GB, Ramamurthy T. Molecular
1050 evolution and functional divergence of *Vibrio cholerae*. *Current Opinion in Infectious*

1051 Diseases. 2016;29: 520–527. doi:10.1097/QCO.0000000000000306

1052 54. Tribble GD, Mao S, James CE, Lamont RJ. A *Porphyromonas gingivalis* haloacid
1053 dehalogenase family phosphatase interacts with human phosphoproteins and is important
1054 for invasion. *Proceedings of the National Academy of Sciences*. 2006;103: 11027–11032.
1055 doi:10.1073/pnas.0509813103

1056 55. Rosch JW, Mann B, Thornton J, Sublett J, Tuomanen E. Convergence of Regulatory
1057 Networks on the Pilus Locus of *Streptococcus pneumoniae*. *Infection and Immunity*. ;
1058 2008;76: 3187–3196. doi:10.1128/IAI.00054-08

1059 56. Rogers PD, Liu TT, Barker KS, Hilliard GM, English BK, Thornton J, et al. Gene
1060 expression profiling of the response of *Streptococcus pneumoniae* to penicillin. *J*
1061 *Antimicrob Chemother*. 2007;59: 616–626. doi:10.1093/jac/dkl560

1062 57. Favrot L, Blanchard JS, Vergnolle O. Bacterial GCN5-Related N-Acetyltransferases: From
1063 Resistance to Regulation. *Biochemistry*. 2016. doi:10.1021/acs.biochem.5b01269

1064 58. Houot L, Watnick PI. A Novel Role for Enzyme I of the *Vibrio cholerae*
1065 Phosphoenolpyruvate Phosphotransferase System in Regulation of Growth in a Biofilm.
1066 *Journal of Bacteriology*. 2008;190: 311–320. doi:10.1128/JB.01410-07

1067 59. Houot L, Chang S, Pickering BS, Absalon C, Watnick PI. The Phosphoenolpyruvate
1068 Phosphotransferase System Regulates *Vibrio cholerae* Biofilm Formation through Multiple
1069 Independent Pathways. *Journal of Bacteriology*. 2010;192: 3055–3067.
1070 doi:10.1128/JB.00213-10

1071 60. Houot L, Chang S, Absalon C, Watnick PI. *Vibrio cholerae* Phosphoenolpyruvate
1072 Phosphotransferase System Control of Carbohydrate Transport, Biofilm Formation, and
1073 Colonization of the Germfree Mouse Intestine. *Infection and Immunity*. 2010;78: 1482–
1074 1494. doi:10.1128/IAI.01356-09

1075 61. Shin SJ, Wu C-W, Steinberg H, Talaat AM. Identification of Novel Virulence Determinants
1076 in *Mycobacterium paratuberculosis* by Screening a Library of Insertional Mutants.
1077 *Infection and Immunity*. 2006;74: 3825–3833. doi:10.1128/IAI.01742-05

1078 62. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a
1079 global network of gene exchange connecting the human microbiome. *Nature*. 2011;480:
1080 241–244. doi:10.1038/nature10571

1081 63. Rafique R, Rashid M-U, Monira S, Rahman Z, Mahmud MT, Mustafiz M, et al.
1082 Transmission of Infectious *Vibrio cholerae* through Drinking Water among the Household
1083 Contacts of Cholera Patients (CHoBI7 Trial). *Front Microbiol*. 2016;7: 4096.
1084 doi:10.3389/fmicb.2016.01635

1085 64. Dillon MM, Sung W, Sebra R, Lynch M, Cooper VS. Genome-Wide Biases in the Rate
1086 and Molecular Spectrum of Spontaneous Mutations in *Vibrio cholerae* and *Vibrio fischeri*.
1087 *Molecular Biology and Evolution*. 2017;34: 93–109. doi:10.1093/molbev/msw224

1088 65. Bilecen K, Yildiz FH. Identification of a calcium-controlled negative regulatory system
1089 affecting *Vibrio cholerae* biofilm formation. *Environmental Microbiology*. 2009;11: 2015–

1090 2029. doi:10.1111/j.1462-2920.2009.01923.x

1091 66. Bilecen K, Fong JCN, Cheng A, Jones CJ, Zamorano-Sánchez D, Yildiz FH. Polymyxin B
1092 Resistance and Biofilm Formation in *Vibrio cholerae* Are Controlled by the Response
1093 Regulator CarR. *Infection and Immunity*. 2015;83: 1199–1209. doi:10.1128/IAI.02700-14

1094 67. Boucher Y, Orata FD, Alam M. The out-of-the-delta hypothesis: dense human populations
1095 in low-lying river deltas served as agents for the evolution of a deadly pathogen. *Front*
1096 *Microbiol*. 2015;6. doi:10.3389/fmicb.2015.01120

1097 68. Qadri F, Azim T, Chowdhury A, Hossain J, Sack RB, Albert MJ. Production,
1098 characterization, and application of monoclonal antibodies to *Vibrio cholerae* O139
1099 synonym Bengal. *Clin Diagn Lab Immunol*. ; 1994;1: 51–54.

1100 69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
1101 data. *Bioinformatics*. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

1102 70. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell
1103 and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 2012;28:
1104 1420–1428. doi:10.1093/bioinformatics/bts174

1105 71. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*.
1106 2012;9: 357–359. doi:10.1038/nmeth.1923

1107 72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1108 Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25: 2078–2079.
1109 doi:10.1093/bioinformatics/btp352

1110 73. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of
1111 draft genomes using the Mauve aligner. *Bioinformatics*. 2009;25: 2071–2073.
1112 doi:10.1093/bioinformatics/btp356

1113 74. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing.
1114 arXiv.org. 2012.

1115 75. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA
1116 polymorphism. *Genetics*. 1989.

1117 76. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of Spontaneous Mutation.
1118 *Genetics*. 1998;148: 1667–1686. doi:10.2307/2410123

1119 77. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server:
1120 rapid annotations using subsystems technology. *BMC Genomics*. 2008;9: 75.
1121 doi:10.1186/1471-2164-9-75

1122 78. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
1123 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*.
1124 *Genome Biology*; 2015;: 1–14. doi:10.1186/s13059-015-0721-2

1125 79. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
1126 sequencing data. *Bioinformatics*. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565

- 1127 80. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space
1128 complexity. BMC Bioinformatics. 2004;5: 113. doi:10.1186/1471-2105-5-113
- 1129 81. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with
1130 profiles instead of a distance matrix. Molecular Biology and Evolution. 2009;26: 1641–
1131 1650. doi:10.1093/molbev/msp077
- 1132 82. Taviani E, Grim CJ, Chun J, Huq A, Colwell RR. Genomic analysis of a novel integrative
1133 conjugative element in *Vibrio cholerae*. FEBS Lett. 2009;583: 3630–3636.
1134 doi:10.1016/j.febslet.2009.10.041
- 1135 83. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User
1136 Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and
1137 Evolution. 2010;27: 221–224. doi:10.1093/molbev/msp259
- 1138 84. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of
1139 heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2:
1140 vew007–7. doi:10.1093/ve/vew007
- 1141 85. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the
1142 accuracy of demographic and molecular clock model comparison while accommodating
1143 phylogenetic uncertainty. Molecular Biology and Evolution. 2012;29: 2157–2167.
1144 doi:10.1093/molbev/mss084
- 1145 86. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular
1146 Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and
1147 Maximum Parsimony Methods. Molecular Biology and Evolution. 2011;28: 2731–2739.
1148 doi:10.1093/molbev/msr121

1149