Université de Montréal

Évolution intra-hôte de *Vibrio cholerae* et interactions avec le microbiome intestinal

par Inès Levade

Département de sciences biologiques
Faculté des arts et des sciences

Thèse présentée
en vue de l'obtention du grade Philosophiae Doctor (Ph.D.)
en sciences biologiques

Mai, 2020

Membres du comité examinateur

Superviseur de thèse	Dr. Jesse Shapiro
Président-rapporteur	Dr. Sophie Breton
Membre du jury	Dr. François-Joseph Lapointe
Examinateur externe	Dr. Jennifer Ronholm

Résumé

Le choléra est une infection diarrhéique aiguë qui représente encore aujourd'hui un grave problème de santé publique dans les pays où l'accès à l'eau potable et un système d'assainissement adéquat ne peut pas être garanti. *Vibrio cholerae*, le pathogène bactérien responsable de cette maladie, peut provoquer toute une série de symptômes chez les individus infectés, allant d'une diarrhée intense conduisant à une déshydratation sévère, au portage asymptomatique de la bactérie. Bien que notre compréhension du choléra à une échelle macro-épidémiologique a considérablement été améliorée par le développement des techniques de séquençage à haut débit et par les avancées dans le domaine de la génomique bactérienne, aucune étude n'a encore été menée pour caractériser son évolution à l'échelle des individus infectés. De plus, le rôle des porteurs asymptomatiques au sein d'une épidémie et la raison derrière l'absence de symptômes chez ces individus infectés sont encore méconnus. L'objectif principal de cette thèse est donc de (1) caractériser la diversité génomique de *V. cholerae* au niveau des individus et des cercles familiaux, mais aussi (2) d'évaluer le rôle potentiel du microbiome intestinal dans la susceptibilité de contracter cette maladie entérique aiguë et de présenter des symptômes sévères.

Dans un premier temps, nous caractérisons la diversité génomique de colonies isolées à partir de patients symptomatiques. Le séquençage de génomes entiers de souches provenant de patients du Bangladesh et d'Haïti révèle que cette diversité sous la forme de mutations ponctuelles reste limitée, mais détectable au sein des hôtes. Une grande partie de la variation du contenu génétique semble être surtout due au gain et à la perte de phages et de plasmides au sein de la population de *V. cholerae*, avec des échanges occasionnels entre le pathogène et d'autres membres commensaux du microbiote intestinal. Cela contredit l'hypothèse couramment acceptée que les infections par *V. cholerae* sont majoritairement clonales, et confirme que le transfert horizontal de gènes est un facteur important dans l'évolution de *V. cholerae*. De plus, nos résultats montrent que certains de ces variants peuvent avoir un effet phénotypique, impactant par exemple la formation de biofilms, et peuvent être sélectionnés au sein des individus infectés.

Par la suite, nous appliquons une association de méthodes de séquençage de génomes entiers et de méthodes métagénomiques afin d'améliorer la détection des variants intra-hôte, à la fois chez des patients symptomatiques, mais aussi chez des porteurs asymptomatiques. Notre étude montre que l'approche métagénomique offre une meilleure résolution dans la détection de la diversité dans la population microbienne, mais reste difficile à appliquer chez des patients asymptomatiques, en raison du faible nombre de cellules de *V. cholerae* chez ces patients. Dans l'ensemble, nous constatons que le niveau de diversité au sein de la population bactérienne intra-hôte est similaire entre les patients symptomatiques et asymptomatiques. Nous détectons aussi la présence de souches hypermutantes chez certains patients. De plus, alors que les mutations chez les patients porteurs de phénotypes d'hypermutations ne semblent pas sous l'effet de la sélection, des signes d'évolution parallèle sont détectés chez les patients présentant un plus faible nombre de mutations, suggérant des mécanismes d'adaptation au sein de l'hôte. Nos résultats soulignent la puissance de la métagénomique combinée au séquençage de génomes entiers pour caractériser la diversité intra-hôte dans le cas d'une infection aiguë du choléra, mais aussi dans le cas de portage asymptomatique, tout en identifiant pour la première fois le phénotype d'hypermutation chez des patients infectés.

Finalement, nous nous intéressons aux facteurs liés à la susceptibilité à la maladie et à la sévérité des symptômes. Basée sur une étude récente utilisant le séquençage 16S pour montrer le lien potentiel entre le microbiome intestinal et la susceptibilité à l'infection par *V. cholerae*, nos analyses utilisent les méthodes de séquençage métagénomique sur les mêmes échantillons de cette précédente étude afin de caractériser les profils taxonomiques et fonctionnels du microbiome intestinal de contacts familiaux exposés à *V. cholerae*. Les échantillons sont prélevés avant l'infection de ces contacts familiaux et l'apparition ou non de symptômes, et sont analysés pour identifier des prédicteurs à la maladie symptomatique. Grâce à un algorithme d'apprentissage machine, nous pouvons identifier des espèces, des familles de gènes et des voies métaboliques du microbiome au moment de l'exposition à *V. cholerae* pour détecter des biomarqueurs potentiels corrélés avec les risques d'infection et la gravité des symptômes. Nos résultats montrent que l'utilisation du séquençage métagénomique améliore la précision et l'exactitude des prévisions par rapport au séquençage 16S. Nos analyses permettent aussi de prédire la gravité de la maladie, bien qu'avec une plus grande incertitude que la prédiction de l'infection. Des taxons bactériens des genres *Prevotella*

et *Bifidobacterium* ont été identifiées comme des marqueurs potentiels de protection contre l'infection, tout comme gènes impliqués dans le métabolisme du fer. Nos résultats soulignent le pouvoir de la métagénomique pour prédire l'évolution des maladies et identifient des espèces et des gènes spécifiques pouvant être impliqués dans des tests expérimentaux afin d'étudier les mécanismes liés au microbiome intestinal expliquant la potentielle protection contre le choléra.

Mots clés : *Vibrio cholerae*, choléra, évolution intra-hôte, génomique, métagénomique, hypermutation, microbiome, apprentissage machine

Abstract

Cholera is an acute diarrhoeal disease that remains a global threat to public health in countries where access to safe water and adequate sanitation cannot be guaranteed. *Vibrio cholerae*, the bacterial pathogen responsible for this disease, can cause a range of symptoms in infected individuals, from intense diarrhea leading to severe dehydration, to asymptomatic carriage of the bacteria. Although our understanding of cholera on a macro-epidemiological scale has been considerably improved by the development of high-throughput sequencing techniques and by advances in bacterial genomics, no studies have yet been conducted to characterize its evolution at the scale of infected individuals. Furthermore, the role of asymptomatic carriers in an epidemic and the reason behind the absence of symptoms in these infected individuals remains unknown. The main objective of this thesis is therefore to characterize the genomic diversity of *V. cholerae* at the level of individuals and households, but also to evaluate the potential role of the gut microbiome in the susceptibility to contract this acute enteric disease and to present severe symptoms. First, we characterize the genomic diversity of colonies isolated from symptomatic patients. The whole genome sequencing of strains from patients in Bangladesh and Haiti reveals that this diversity is detectable in the form of point mutations within hosts, but remains limited. Much of the variation detected within patients appears to be due to the gain and loss of phages and plasmids within the *V. cholerae* population, with occasional exchanges between the pathogen and other commensal members of the gut microbiota. These results challenge the commonly accepted assumption that *V. cholerae* infections are predominantly clonal, and confirm that horizontal gene transfer is an important factor in the evolution of *V. cholerae*. In addition, our results show that some of these variants may also have a phenotypic effect, for example by impacting biofilm formation, and can be selected within infected individuals.

Next, we apply a combination of whole genome sequencing and metagenomic approaches to improve the detection of intra-host variants, both in symptomatic patients and in asymptomatic carriers. Our study shows that the metagenomic approach offers a better resolution in the detection of the diversity in the microbial population, but remains difficult to apply in asymptomatic patients, due to the low number of *V. cholerae* cells in these individuals. Overall, we find that the level of diversity within the intra-host bacterial population is similar between symptomatic and asymptomatic patients. We also detect the presence of

hypermutator strains in some patients. In addition, while mutations in patients with hypermutator phenotypes did not appear to be driven by selection, signs of parallel evolution are detected in patients with fewer mutations, suggesting adaptive mechanisms within the host. Our results underline the power of metagenomics combined with whole genome sequencing to characterize intra-host diversity in acute cholera infection, but also in asymptomatic carriers, while identifying for the first time an hypermutator phenotype in infected patients.

Finally, we are interested in factors related to susceptibility to the disease and related to the severity of symptoms. Based on a recent study using 16S rRNA amplicon sequencing to show the potential link between the intestinal microbiome and susceptibility to *V. cholerae* infection, our study uses metagenomic sequencing methods on the same samples from this previous study to characterize the taxonomic and functional profiles of the gut microbiome of household contacts exposed to *V. cholerae*. Samples are collected prior to infection of these household contacts, and used to identify predictors of symptomatic disease. Using a machine learning algorithm, we can identify species, gene families and metabolic pathways in the microbiome at the time of exposure to *V. cholerae* to detect potential biomarkers correlated with risk of infection and symptom severity. Our results show that the use of metagenomic sequencing improves the precision and accuracy of predictions compared to 16S rRNA amplicon sequencing. Our analyses also predict disease severity, although with greater uncertainty than the prediction of infection. Bacterial taxa from the genera *Prevotella* and *Bifidobacterium* have been identified as potential markers of protection against infection, as well as genes involved in iron metabolism. Our results highlight the power of metagenomics to predict disease progression and identify specific species and genes that could be involved in experimental tests to study the mechanisms related to the microbiome explaining potential protection against cholera.

Keywords: *Vibrio cholerae*, cholera, within-host evolution, genomics, metagenomics, hypermutation, microbiome, machine learning

Table des matières

Liste des tableaux

Chapitre 2

Chapitre 3

mutators are underlined. Only genes and patients containing more than one mutated gene are shown.

Table 3: Nucleotide changes identified in core genes of the *V. cholerae* strains isolated from index cases (56.00, 57.00 and 58.00) and their asymptomatic contacts. Genome position is according to the MJ-1236 reference genome. Mutations segregating within patients are denoted iSNVs; Mutations fixed between patients are denoted 'Patient.' Nucleotide positions are based on the reference Vibrio cholerae MJ-1236 (CP001485.1, CP001486.1). Patient allele frequency shows the allele frequency of the alternative (minor) allele. Ref=Reference allele; Alt=Alternative allele. NS=non-synonymous; S=synonymous. Chr1=chromosome 1; Chr2=chromosome

Chapitre 4

Table 1: Assessment of prediction performance for a random forest (RF) model applied to the Midani 2018 and expanded cohorts. Species abundances, strain-specific markers presence/absence, relative abundance of Pfam-grouped gene families, and MetaCyc pathways were used as features. For each dataset, we applied a binary (uninfected vs. infected contacts) and a multi-class (asymptomatic vs. symptomatic vs. uninfected contacts) classifier and reported performance metrics for each dataset. Metrics obtained by the same classifier applied to the same datasets with shuffled class labels (random assignment of labels to samples) are also reported (shuffled). The margins of errors (95% confidence intervals) are reported in parenthesis.

Liste des figures

Chapitre 2

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.

Figure 4: Significant excess of non-synonymous iSNVs in patient H1. (a) Distribution of 122 *V. cholerae* isolates 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 isolates among regions). (b) Distribution of 122 *V. cholerae* isolates containing different iSNVs, or no detectable iSNVs, by patient. Patient H1 has a significant excess of isolates with Nsyn iSNVs ($p \leq 0.001$; 10,000 random permutations of mutations across patients; Supplementary Methods).

Figure 5: Biofilm formation of isolates from patients H1 and H2. Optical density was measured for four to 12 replicates of each isolate, 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**) Isolates 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**) Isolates 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.

Chapitre 3

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

Figure 2. Within patient *Vibrio cholerae* diversity from metagenomic data. (A) Minor allele frequency and distribution of intergenic, synonymous and non-synonymous iSNVs across the two Vibrio cholerae chromosomes for the 14 patients (B) Number and proportion of intergenic, synonymous and non-synonymous iSNVs for each patient (C) Transversion/Transition mutation spectrum of the iSNVs in the samples with more than 6 iSNVs. Error bars represent standard error of the mean across the group of samples with hypermutators and the group with all the other samples.

Figure 3. Phylogenetic and pan-genomic analysis from 48 *Vibrio cholerae* isolates from index cases and their asymptomatic contacts. Phylogenetic tree was inferred using the Maximum Parsimony method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches). Plain circles represent isolates from index cases and empty circles represent isolates from their asymptomatic contacts. Heat map of the gene presence-absence is based on the 102 genes of the flexible genome (color: present, colourless:absent). Each row corresponds to an isolate from the phylogenetic tree and each column represents an orthologous gene family. Each color represents an individual.

Chapitre 4

Figure 1. Study cohort in Dhaka, Bangladesh. After presentation of a *V. cholerae* culture-positive index case to the hospital on day 1, household contacts were enrolled on day 2. The expanded cohort includes the Midani 2018 cohort (Midani et al. 2018), with an addition of 33 samples from infected individuals (13 asymptomatic and 2 symptomatic).

Figure 2. Metagenomic features predict *V. cholerae* infection better than clinical and demographic features. Random forest prediction of infection status was applied to 7 clinical and demographic features, and compared with all species and all gene families (top row), as well as 30 selected species features from metagenomes and 60 selected gene family features (bottom row), or a combination of clinical, demographic and metagenomic features. Plots show receiver operating characteristic (ROC) curves (average across cross-validations) for the Midani 2018 dataset. Shuffled labels represent the prediction run on a dataset with a random assignment of infection outcomes. AUC = area under the curve.

Figure 3. Most important discriminating species of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Species associated with contacts that became infected (red) or remained uninfected (yellow) during follow-up. (B) Species associated with contacts who remained uninfected (yellow), or became infected asymptomatic (green), or symptomatic (red) during follow-up. The top 25 most important features are shown here; see Table S6 for the full list. Yellow lines connect species associated with uninfected individuals in both (A) and (B); red lines connect species associated with infection in (A) and symptomatic disease in (B); grey lines connect species associated with infection in (A) but asymptomatic infection in (B). Three species of *Bifidobacterium* are marked with an asterisk.

Figure 4. Most important discriminating gene families of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Genes families associated with contacts that became infected (red) or remained uninfected (yellow) during follow-up. (B) Genes families associated with contacts who remained uninfected (yellow), or became infected asymptomatic (green), or symptomatic

(red) during follow-up. The top 25 most important features are shown here; see Table S8 for the full list. Yellow lines connect species associated with uninfected individuals in both (A) and (B). Asterisks indicate genes involved in redox or iron metabolism.

Figure 5. Top predictive cellular pathways of the gut microbiome at the time of exposure to *V. cholerae* in the Midani 2018 cohort, annotated by their taxonomic contributors. The four top-ranked pathways associated with uninfected contacts (left column), contacts who developed asymptomatic infection (middle), and contacts who developed symptomatic infection (right column) are shown. Total bar height reflects $\log_{10}$-scaled community relative abundance of each pathways. The contributions of each genus to encoding these pathways are shown as stacked colors within each bar, linearly scaled within the total. See Table S9 for the complete list of pathways.

Liste des abréviations

AUC: area under the curve

DNA: deoxyribonucleic acid

FDR: false discovery rate

Fur: ferric uptake regulator

GWAS: genome-wide association studies

HGT: horizontal gene transfer

HK: histidine kinase

hqSNP: high quality single-nucleotide polymorphism

ICE: integrative conjugative element

iSNV: intra-host single nucleotide variant

LDA: linear discriminant analysis

MAG: metagenomic assembled genome

ML: machine learning

MMR: methyl-directed mismatch repair

MRCA: most recent common ancestor

NS: non-synonymous

OMS : Organisation mondiale de la Santé

ONU : Organisation des Nations Unies

OTU: operational taxonomic unit

PICI: phage-inducible chromosomal island

RF: random forest

ROS: reactive oxygen species

rRNA: ribosomal ribonucleic acid

S: synonymous

SCFA: short-chain fatty acid

SNP: single-nucleotide polymorphism

T6SS: Type VI secretion system

ROC: receiver operating characteristic

VCO: vaccins anticholériques oraux

Remerciements

Cette thèse n'aurait jamais vu le jour sans l'aide précieuse de mes collègues, mes amis et ma famille. Je profite donc de ces quelques lignes pour les remercier.

Tout d'abord, je remercie chaleureusement Sophie Breton, Jennifer Ronholm et François-Joseph Lapointe d'avoir accepté de réviser cette thèse.

Cette aventure n'aurait pas été possible sans mon directeur de thèse Jesse Shapiro. Merci à toi Jesse, pour m'avoir fait confiance et m'avoir accueilli comme une de tes premiers étudiants, pour tout le soutien que tu m'as apporté au cours de ces 6 années, pour ton optimisme sans faille qui m'a permis de tenir pendant les moments les plus difficiles et pour ton amour pour la science plus que contagieux. A chaque moment où je commençais à douter de moi-même ou du projet, il suffisait d'une rencontre avec toi pour que je ressorte avec le moral et une confiance en moi retrouvée. Je pense que peu d'étudiants gradués peuvent se vanter d'avoir une telle relation avec leur superviseur et avec du recul je peux vraiment dire que je n'aurais jamais pu rêver meilleur directeur de recherche que toi. Merci pour cette incroyable opportunité.

A big thank you to all my collaborators, and especially to Ana Weil, who has been a precious support throughout the different projects. Thank you, Ana, for your precious advice, your enthusiasm every time I presented new results and your help at every step. Even if you are not officially my co-director, I consider you a bit like it. Thank you for everything you have given me professionally.

Merci à tous les membres de mon lab mais plus particulièrement Julie et Nico pour votre aide au laboratoire, Jean-Baptiste et Masih pour vos précieux conseils, Naila, Naima et Olga pour l'entraide et les discussions formatrices.

Cette thèse m'a offert la chance de travailler avec de nombreuses personnes, et ces collaborations, toutes fructueuses et enrichissantes, ont permis l'élaboration de ce travail. Mais les rencontres que j'ai pu faire au sein du département ne m'ont pas seulement aidé à mener à bien ce travail, elles ont aussi changé ma vie. Je n'y ai pas juste rencontré de nouveaux collègues et amis, je m'y suis trouvé une famille.

Tout d'abord je m'y suis trouvé un grand frère. Yves, merci d'avoir été là tout au long du parcours et de m'avoir servi de mentor, sans toi je n'y serais clairement pas arrivé. Merci pour tes mots d'encouragement, tes conseils, ton aide sur les projets, les rigolades et

les sessions de chialage, et surtout, merci pour ton amitié. Te côtoyer ces dernières années n'a pas seulement fait de moi une meilleure scientifique, ça a fait de moi une meilleure personne, du moins je l'espère.

Eva, merci d'avoir été là tout au long du parcours. Merci en particulier pour ces derniers mois, je ne sais pas comment j'aurais pu finir d'écrire cette thèse en ces temps de pandémie sans t'avoir à mes côtés. En quittant la France, je pensais laisser derrière ces fameuses soirées à refaire le monde autour d'une bouteille de vin, mais ça c'était avant de te rencontrer ! J'espère qu'au cours des années à venir on pourra continuer à débattre pendant des heures sur la forme alors qu'en fait, on est quand même un peu d'accord sur le fond...Merci de me laisser faire partie intégrante de ta vie et de m'avoir introduit à Brieuc et Réglisse, ces deux êtres incroyables.

Un énorme merci à Allison et Catherine, mes colocataires de rêve. Partager son quotidien avec une étudiante au doctorat n'est pas toujours chose aisée, alors je te remercie pour ton soutien et ta patience Allison. Merci de m'avoir épaulé lors des moments les plus difficiles, et de m'avoir poussé à célébrer les moments forts. Catherine, il est assez difficile d'exprimer à quel point je te suis reconnaissante de m'avoir intégré dans ta vie comme tu l'as fait. Un gros merci à toi et toute ta famille pour tous ces Noëls québécois, qui m'ont permis de me sentir un peu moins seule de ce côté de l'Atlantique.

Quand je suis arrivé au département, j'ai eu la chance de tomber dans le meilleur de tous les cubicules. Merci à vous tous, Stéphanie pour tes belles maladresses et les aventures en canot, Cynthia pour ta force tranquille qui me sert toujours d'inspiration, et Marie-Pier pour ton rire si communicateur. Merci à toi et Vincent pour tous les beaux moments passés avec vous, et pour les moments à venir ! Et merci à toi Dan, pour ta bonne humeur sans faille et ton écoute. Sans les pauses café avec toi et Mario, les derniers mois de ma thèse auraient été bien plus difficiles.

Un gros merci à un autre duo de choc, Manu et Max. Merci pour les moments de partage, de rire et de débats parfois un peu trop enflammés. Merci de m'avoir supporté pendant ces 6 mois de vie commune Max, et d'être constamment là pour moi depuis. Je suis vraiment reconnaissante de t'avoir dans ma vie. À Manu, merci pour ton soutien et ta patience sans limites (surtout vis-à-vis de mes talents de conduite à gauche ou au karaoké), on repart à Skye quand tu veux ! Merci de m'avoir offert ta précieuse amitié et m'avoir

introduite aux personnes importantes dans ta vie. Parmi ces belles rencontres, j'en profite pour remercier Seb et Stef, merci à vous deux de m'avoir accueilli à la Mazana, une des plus belles places au monde. Et merci à toi Louis, pour les fous-rires badmintoniens et pour tout le reste. Vous me faites tous aimer le Québec un peu plus chaque jour, et me donner envie d'y rester.

Pour ceux que j'ai laissés sur le vieux continent, votre soutien c'est révélé être tout aussi important que celui des gens que j'ai rencontrés ici.

À Clio et Axelle, je n'y serais pas arrivé sans vous.

Clio, parce que c'est grâce à tes encouragements que je me suis dit que j'étais capable de faire une thèse. Les mois passés avec toi à Copenhague ont été parmi les plus beaux moments de ma vie, merci de m'avoir pris sous ton aile à ce moment-là et d'être avec moi encore aujourd'hui.

Axelle, merci pour cette incroyable amitié qui résistera à tout, contre vents et marées (littéralement). Il me faudrait des pages entières pour donner toutes les raisons pour lesquelles je suis reconnaissante de t'avoir dans ma vie et comment tu m'as aidé à garder le cap au cours de ces 6 années. Mais pas besoin, car ces raisons, tu les connais. Merci, tu es la meilleure.

Cécile et Mathilde, je ne peux qu'être reconnaissante de vous avoir encore là avec moi malgré le temps et la distance. Merci de continuer à me montrer ce que vous voyez et aimez en moi, même si je ne suis pas toujours une élève facile.

Un gros merci à tous les autres, ceux d'ici et d'ailleurs qui ont fait de ces 6 années une belle aventure : Audrey-Alexandra, Mathilde, Gwyneth, Dave, Adeline, Sarah, Pierre-Yves, Anissa, Jihane, Charlotte, Stefano, Alex, Nounou, Thibault, Antoine, Emily, Ivan, Xavier, Émilie, Guillaume, Arnaud, Andréa, Julia, Josh, Catherine, Mathieu, Fanny, Thomas, Richard, François, Martin, Marina et Marie-Émilie.

Finalement un merci spécial à mes parents, sans qui je n'aurais pas pu entreprendre une telle aventure. Merci d'avoir accepté de me voir partir aussi loin, et de m'avoir offert votre soutien inconditionnel. Sans oublier mes sœurs, et particulièrement Sarah, merci à toi, vivre au Québec ne serait pas une aussi belle expérience si tu n'étais pas là pour la partager avec moi.

"Death's Dispensary" par George Pinwell

Avant-propos

Une thèse sur le choléra au temps de la COVID-19

Dans son livre « Pandemic: Tracking Contagions, From Cholera to Ebola and Beyond » paru en 2016, Sonia Shah a écrit "The disease-causing microbe, or pathogen, that will cause the world's next pandemic lurks among us today. We don't know its name or where it comes from. But for now call it 'cholera's child', because what we do know is that it will likely follow the path that cholera blazed".

Au moment de rédiger ces lignes, rien ne semble contredire cette prédiction. Nous connaissons aujourd'hui le nom de ce nouveau pathogène, le virus SARS-CoV-2. Même si le doute subsiste toujours sur l'origine exacte du virus, une chose reste certaine: notre

vulnérabilité croissante face aux pandémies est intrinsèquement liée à la destruction accélérée des habitats naturels et la manière dont ceux-ci sont remplacés pour répondre à notre mode de consommation.

Le choléra est le parfait exemple de ce lien entre destruction des habitats, mode de vie et pandémies. La bactérie à l'origine de cette maladie est un des rares pathogènes (avec ceux à l'origine de la peste, de la tuberculose, de la grippe, de la variole ou encore du SIDA) qui, à l'époque moderne, ont provoqué des pandémies. Aujourd'hui, le choléra est surtout connu comme une maladie qui touche les pays en développement, mais cela n'a pas toujours été le cas.

En 1817, la Compagnie britannique des Indes orientales envoya des milliers d'ouvriers au fin fond des Sundarbans, une région reculée du delta du Gange, pour abattre la jungle et développer la riziculture. Cette destruction de la mangrove fut suffisante pour exposer les populations aux bactéries aquatiques présentes dans ces eaux saumâtres. Cette maladie diarrhéique avait été décrite en Inde et ses pays limitrophes depuis l'Antiquité, mais n'avait jamais dépassé ces frontières jusqu'à l'arrivée des colons européens. Bien que cette expansion coloniale combinée à la destruction de cet écosystème rare et fragile, fut à l'origine de cette première pandémie, elle a surtout été facilitée par les changements rapides liés à la révolution industrielle. Le développement de nouveaux modes de transport, ainsi que les conditions de surpopulation et d'insalubrité de ces villes en pleine croissance, ont permis à la bactérie de pénétrer l'Europe et l'Amérique du Nord. La maladie continua ainsi de circuler pendant deux siècles, causant un total de sept pandémies, et des dizaines de millions de victimes.

L'apparition d'épidémies répétées de choléra a représenté un défi de taille pour les institutions politiques et sociales des sociétés qu'elles touchaient. Seule une coopération internationale des institutions, une gouvernance locale efficace et une cohésion sociale jamais observée jusque-là ont permis de contenir la maladie au cours du siècle suivant la première pandémie. La découverte de son origine et de son remède le plus efficace - l'eau potable - a exigé des médecins et des scientifiques qu'ils transcendent leurs idées reçues sur l'origine et la propagation des maladies, ce qui mena à la naissance de l'épidémiologie moderne.

Nous ne pouvons pas encore estimer le réel impact sur nos modes de vie qu'aura à long terme la COVID-19. L'histoire des pandémies, et particulièrement celle(s) du choléra, témoigne de leur pouvoir transformationnel sur nos sociétés.

Chapitre 1 : Introduction

Choléra : l'hôte, le pathogène et le microbiome

Un fléau toujours d'actualité

Les maladies diarrhéiques sont une des principales causes de morbidité et la deuxième cause de mortalité chez les enfants de moins de cinq ans dans le monde (Kosek, Bern, et Guerrant 2003; Bryce et al. 2005). Parmi ces maladies, le choléra représente une des infections les plus foudroyantes, causant des vomissements et une diarrhée aiguë type « eau-de-riz » pouvant entrainer une perte hydrique allant jusqu'à 15L par jour (Harris et al. 2012). Touchant les enfants comme les adultes, cette diarrhée aqueuse aiguë sévère peut entrainer une déshydratation rapide de l'individu infecté, qui peut se traduire par un choc hypotensif, une insuffisance rénale et la mort dans les heures qui suivent l'apparition des premiers symptômes.

Provoquée par l'ingestion de la bactérie à gram négatif *Vibrio cholerae*, cette maladie infectieuse représente encore aujourd'hui un grave problème de santé publique dans de nombreuses parties du monde, ainsi qu'un indicateur de l'absence d'équité et de l'insuffisance du développement social. D'après des estimations, entre 1,3 et 4 millions de cas de choléra et 21 000 à 143 000 décès surviendraient chaque année dans les pays où l'accès à l'eau potable et à un système d'assainissement adéquat ne peut être garanti (Ali et al. 2015). Toutefois, la charge globale exacte du choléra est sous-estimée en raison de facteurs contributifs tels que le faible nombre de rapports, la surveillance épidémiologique limitée, l'utilisation de définitions de cas inappropriées, l'insuffisance des capacités de diagnostic des laboratoires et la réticence à la notification par crainte de conséquences économiques négatives. Selon l'Organisation mondiale de la santé (OMS), le risque de contracter la maladie concernerait près de 1,3 milliard de personnes dans 70 pays où le choléra est considéré comme endémique. Dans un contexte non endémique, les catastrophes naturelles ou provoquées par l'homme, les conflits armés et les déplacements de populations dans des camps de réfugiés surpeuplés représentent les facteurs de risque liés aux épidémies de choléra. En résultent des flambées épidémiques explosives, qui présentent des taux de létalité plus élevés qu'en contexte endémique. Ce fut notamment le cas en Haïti en 2010, où suite à un violent séisme et la

destruction d'une grande partie des infrastructures du pays, une épidémie de choléra a été déclarée, causant plus de 800 000 cas et entrainant la mort de plus de 10 000 personnes (Chin et al. 2011; Katz et al. 2013). En proie à une épidémie de choléra pour d'autres raisons, le Yémen affiche aujourd'hui un bilan tout aussi lourd. Depuis 2015, ce pays est déchiré par une guerre causant une des plus grandes crises alimentaires du monde, ainsi que l'effondrement des structures médicales et sanitaires. Depuis octobre 2016, il est également frappé par une épidémie de choléra à l'origine de plus de 500 000 cas et de près de 2 000 décès en seulement 4 mois (Firdausi Qadri, Islam, et Clemens 2017; Camacho et al. 2018). Le bilan atteint aujourd'hui plus de 2,2 millions de cas, ce qui en fait la pire crise humanitaire au monde d'après l'ONU.

Vibrio cholerae : l'agent étiologique du choléra

Vibrio cholerae est une des espèces pathogènes les plus étudiées de la famille des Vibrionaceae. C'est un pathogène facultatif, ce qui signifie qu'il ne dépend pas de son hôte pour sa survie et sa persistance à long-terme. Il est en effet naturellement présent à l'état saprophyte dans divers environnements aquatiques, tels que les océans, les estuaires, les rivières ou les lacs (Colwell 1996; Lutz et al. 2013). Dans son environnement naturel, *V. cholerae* est souvent retrouvé en association avec des mollusques à coquille, copépodes et autres crustacés, car il est capable d'utiliser la chitine comme source de carbone et d'azote (Harris et al. 2012; Lutz et al. 2013; Sakib, Reddi, et Almagro-Moreno 2018).

La classification de cette espèce bactérienne a longtemps été réalisée sur la base de l'antigène somatique O du lipopolysaccharide (LPS) présent à la surface de la bactérie, permettant d'identifier plus de 200 sérogroupes à ce jour (Harris et al. 2012). Parmi eux, seuls les sérogroupes O1 et O139 ont été associés aux symptômes du choléra. Historiquement, le sérogroupe O1 a causé la grande majorité des maladies et est divisé en plusieurs biotypes phénotypiquement distincts : le biotype El Tor et le biotype classique sont les plus courants. Les deux biotypes peuvent être subdivisés en deux sérotypes principaux : Inaba et Ogawa. Au cours de ces 25 dernières années, le biotype El Tor a largement remplacé le biotype classique et est à l'origine de la pandémie actuelle : la septième pandémie de choléra. Le sérogroupe O139 est apparu en 1992 et a été identifié comme la cause d'une épidémie de choléra durant cette période, mais est resté limité à certaines régions d'Asie

depuis. Certaines souches de *V. cholerae* non-O1 et non-O139 ont été identifiées dans des cas de gastro-entérites et diarrhées bénignes, mais ne peuvent pas causer de flambées épidémiques (Dziejman et al. 2005). La virulence et le caractère épidémique des souches pathogènes proviennent notamment de l'acquisition des gènes codant pour la toxine cholérique CT et pour le pilus de type IV, le TCP (Toxin Coregulated Pilus).

Physiopathologie, traitement et transmission du choléra

Vibrio cholerae infecte les individus via l'ingestion d'eau ou d'aliments contaminés ou directement par contamination fécale-orale. La dose infectieuse, déterminée au cours d'expérimentations sur des volontaires, est relativement élevée, de l'ordre de 10^8 à 10^{11} bactéries, mais peut être réduite de 10^4 à 10^6 dans des condition alcalines (Nelson et al. 2009). Une grande partie des vibrions ingérés est détruite par l'acidité gastrique, et seul un faible nombre d'organismes colonise l'intestin grêle puis traverse la couche de mucus tapissant la muqueuse intestinale. Les bactéries adhèrent alors aux entérocytes, se multiplient et, lorsqu'une certaine concentration de cellules est atteinte, produisent le pilus TCP, ce qui permet la formation de micro-colonies et la production de la toxine cholérique. Cette toxine va alors se fixer aux entérocytes et entrainer la sécrétion importante d'eau et d'électrolytes en direction de la lumière intestinale, entrainant une perte hydrique pouvant aller jusqu'à 15 à 20L par jour, décrite comme un cas sévère de choléra ou « cholera gravis » (Almagro-Moreno, Pruss, et Taylor 2015; Weil et Ryan 2018).

Sans traitement, le taux de mortalité peut atteindre 50% des cas, particulièrement chez les enfants et femmes enceintes (Nelson et al. 2009). La base du traitement contre la maladie reste la réhydratation, orale pour les cas les moins graves, ou par intraveineuse pour les patients souffrant d'une déshydratation sévère (définie par une perte de plus de 10% de la masse corporelle). Les patients présentant des symptômes sévères reçoivent également un traitement antibiotique afin de raccourcir la durée de la diarrhée, diminuer les quantités de liquide de réhydratation nécessaires et écourter la durée de l'excrétion des bacilles de *V. cholerae* dans leurs selles (Weil, Ivers, et Harris 2011).

La diarrhée produite par *V. cholerae* peut devenir un vecteur de transmission important. Une fois que la population bactérienne présente dans l'intestin grêle atteint une forte densité, les organismes se détachent de la surface intestinale pour s'échapper de l'hôte

(Qadri et al. 2004). Dans les cas les plus sévères, jusqu'à 10^9 organismes viables sont excrétés par ml de selles (Nelson et al. 2009), et sans traitement antibiotique adéquat, cette sécrétion peut se poursuivre pendant plusieurs jours. De plus, les bactéries quittant l'hôte présentent un phénotype hyperinfectieux pouvant durer jusqu'à 24 heures après l'excrétion dans les selles (Alam et al. 2005; Nelson et al. 2009). Cette infectiosité accrue contribue à la propagation de la maladie lors de flambées épidémiques.

Une infection par *V. cholerae* ne résulte pas toujours en une diarrhée aiguë sévère, la plupart des cas étant bénins voir même asymptomatiques, en particulier en contexte endémique (Weil et al. 2009). Les personnes infectées et ne présentant pas de symptômes peuvent cependant excréter des bactéries dans leurs selles jusqu'à deux semaines après leur infection. Plusieurs études effectuées sur des cohortes de contacts de patients infectés au Bangladesh et des estimations faites à partir de l'épidémie haïtienne ont montré que plus de la moitié des individus infectés par *V. cholerae* seraient asymptomatiques et que leur contribution à la chaine de transmission serait mal évaluée (King et al. 2008; Weil et al. 2009; Lewnard et al. 2016; Phelps, Simonsen, et Jensen 2019).

Susceptibilité à l'infection par Vibrio cholerae

Parmi les personnes exposées, divers facteurs liés à l'hôte influencent la susceptibilité à l'infection par *V. cholerae*. Le régime alimentaire et l'état nutritionnel de l'hôte semblent par exemple influencer cette susceptibilité. L'hypoacidité gastrique a été associée au risque de contracter la maladie, dans des cas où l'ingestion de seulement 10^4 organismes est suffisante pour déclencher une infection (Sack et al. 1998). De même, il a été montré que les personnes présentant une déficience en rétinol présentent plus de risques d'être infectés et de développer des symptômes sévères (Harris et al. 2008a).

D'autres études ont découvert que des facteurs génétiques peuvent influencer les risques d'infection et la sévérité des symptômes. Par exemple, chez différentes populations, le groupe sanguin O a été associé avec la forme sévère de la maladie et des taux de déshydratation plus élevés que pour les autres groupes sanguins (Harris et al. 2005). Récemment, des expériences *in vitro* ont démontré un effet plus puissant de la toxine cholérique dans l'induction de l'AMPc au niveau des entérocytes d'individus de groupe sanguin O comparativement à des individus de groupe sanguin A (Kuhlmann et al. 2016). Sur la base

de ces observations, et du fait que la prévalence du groupe sanguin O dans la population du delta du Gange est une des plus faibles au monde, le choléra a été proposé comme un facteur de sélection naturelle dans l'évolution humaine des populations de cette région où la maladie est endémique (Glass, Holmgren, et al. 1985). Plusieurs autres polymorphismes génétiques ont été associés à la susceptibilité au choléra et sont sous impact de la sélection naturelle chez la population du Bangladesh, notamment au niveau de gènes impliqués dans la réponse immunitaire innée (LaRocque et al. 2009; Karlsson et al. 2013).

En plus des facteurs d'immunité innée, l'immunité acquise semble aussi fortement influencer la susceptibilité à l'infection par la bactérie. L'infection par *V. cholerae* entraîne une immunité durable contre la réinfection, avec une protection pouvant aller de trois à dix ans (Levine et al. 1981; Glass et al. 1982). Cette durée semble cependant être plus courte chez les enfants, le jeune âge étant un autre facteur démontré de susceptibilité accrue (Weil et Ryan 2018). Des différences semblent cependant apparaître au niveau de cette immunité en fonction des différents sérotypes. Une étude plus récente effectuée au Bangladesh sur une période de 10 ans a démontré que bien que le sérotype Inaba de *V. cholerae* O1 confère une protection contre les réinfections récurrentes par les deux sérotypes, l'infection par le sérotype Ogawa ne confère qu'une protection contre le sérotype homologue (Ali et al. 2011).

La mesure la plus courante de la réponse immunitaire à *V. cholerae* est le titrage des anticorps vibriocides. En effet, l'infection par la bactérie induit chez l'homme la production d'anticorps vibriocides et des anticorps antitoxiques, et plusieurs études ont montré une bonne corrélation entre les titres en anticorps vibriocides et la protection contre le choléra (Glass, Svennerholm, et al. 1985; Clemens et al. 1991; Harris et al. 2008a). Cependant on sait maintenant que ces anticorps sériques n'ont pas d'effet direct sur l'immunité anticholérique. L'immunité protectrice contre le choléra résulte plutôt de la stimulation des réponses immunitaires de la muqueuse intestinale, principalement antibactériennes (anti-O1 LPS pour le choléra du sérogroupe O1 et anti-O139 LPS pour le choléra du sérogroupe O139), mais aussi antitoxines (Leung et al. 2012; Weil, Becker, et Harris 2019).

Néanmoins, les facteurs de risques directement liés à l'hôte n'expliquent que partiellement la variation dans les manifestations cliniques observées à la suite d'une exposition au pathogène (Harris et al. 2005; 2008a). Ainsi, il a été récemment montré que les microorganismes composant le microbiote intestinal ont un rôle potentiel dans la réponse

aux pathogènes responsables des infections entériques telles que le choléra (Weil, Becker, et Harris 2019).

Interactions entre Vibrio cholerae et le microbiome intestinal

Notre compréhension de l'impact de la population microbienne intestinale sur les interactions hôte-pathogène dans l'infection par *V. cholerae* est naissante. Dans le cas du choléra, il est depuis longtemps reconnu lors d'expérimentations sur des modèles animaux qu'il est nécessaire d'utiliser des antibiotiques afin de perturber la communauté bactérienne qui compose le microbiote, permettant par la suite la colonisation du système digestif par le pathogène (Freter 1955).

Une infection par *V. cholerae* et la déclaration de symptômes sévères a un effet considérable sur la composition du microbiome intestinal (David et al. 2015). En premier lieu, la sécrétion massive d'eau dans la lumière intestinale, due à la toxine cholérique, emporte le mucus protecteur où réside une grande partie du microbiote intestinal. Par la suite, l'appauvrissement de la communauté microbienne peut être exacerbée par les traitements contre le choléra, notamment l'ingestion de grandes quantités de solution de réhydratation orale et d'antibiotiques qui tuent certaines espèces de bactéries intestinales (Monira et al. 2013; David et al. 2015). Au premier stade de l'infection, *V. cholerae* est souvent retrouvé en grande majorité dans la population microbienne contenue dans les selles excrétées par le patient malade (Hsiao et al. 2014; David et al. 2015). Le microbiote intestinal met par la suite plusieurs semaines avant de retrouver sa composition basale.

Malgré l'impact de la colonisation *V. cholerae* sur sa composition lors de l'apparition de symptômes sévères, dans certains cas le microbiome intestinal pourrait aussi avoir un effet sur la susceptibilité à l'infection. Les résultats d'une étude prospective basée sur des contacts familiaux de patients atteints de choléra ont montré une forte association entre certains groupes de bactéries composant le microbiote au moment de l'exposition à *V. cholerae* et leur infection par le pathogène au cours des jours suivants (Midani et al. 2018). En utilisant un modèle d'apprentissage automatique, les taxons bactériens ont été classés en fonction de leur association avec une augmentation ou une diminution de la sensibilité à l'infection, et les 100 premiers taxons bactériens de ce classement ont permis de prédire l'infection avec une

précision plus élevée que des facteurs de risque cliniques, immunologiques et épidémiologiques connus pour le choléra.

Des études effectuées *in vitro* ou sur des modèles animaux ont identifié de potentiels mécanismes permettant aux bactéries commensales de l'intestin de résister à l'invasion de pathogènes (My Young Yoon, Lee, et Yoon 2014). Ces mécanismes incluent la compétition pour l'accès aux nutriments (Kamada et al. 2012), l'inhibition de l'adhésion et de la colonisation (Mack et al. 1999; Moroni et al. 2006; Medellin-Peña et Griffiths 2009; Karczewski et al. 2010; Collins et al. 2014), la stimulation du système immunitaire (Moal et al. 2002; Corr, Gahan, et Hill 2007) ou encore la production de molécules antimicrobiennes (Rea et al. 2010; Crost et al. 2011; Fukuda et al. 2011).

Lorsque *V. cholerae* colonise l'intestin grêle, il rencontre une couche de protéines hautement glycosylées (mucines) tapissant l'épithélium intestinal. Cette couche de mucus est également colonisée par de nombreuses bactéries commensales, et *V. cholerae* détecte la présence de ces bactéries et de leurs métabolites antibactériens par divers mécanismes. La détection de la mucine active notamment le système de sécrétion de type VI de *V. cholerae* (T6SS). Ce système de sécrétion fonctionne comme une seringue moléculaire, délivrant des toxines et effecteurs dans les cellules cibles, et se révèle être un facteur essentiel dans la colonisation de l'hôte et la compétition contre les bactéries du microbiome intestinal (MacIntyre et al. 2010; Fu, Waldor, et Mekalanos 2013; Bachmann et al. 2015). Certains de ces microbes ont cependant trouvé une parade contre cette attaque de la bactérie pathogène. En effet, des bactéries de l'intestin grêle sont capable de déshydroxyler les acides biliaires primaires – molécules normalement reconnues par *V. cholerae* comme indicateur de l'environnement de l'intestin grêle – en métabolites secondaires qui peuvent supprimer l'activité du T6SS. En métabolisant les acides biliaires et en masquant la "détection" de l'environnement de l'intestin grêle par le pathogène, ces commensaux peuvent ainsi favoriser leur propre survie (Bachmann et al. 2015). De plus, l'action de ces bactéries commensales peut avoir un impact sur l'hôte infecté. En effet, la détection de la bile et de ces métabolites par *V. cholerae* peut faire varier la motilité de la bactérie, ainsi que la production de la toxine cholérique et du pilus TCP, en fonction de la forme spécifique d'acide biliaire détectée (Gupta et Chowdhury 1997; Yang et al. 2013).

Un autre moyen utilisé par les bactéries commensales afin de parer la colonisation de *V. cholerae* semble être via la production d'auto-inducteurs. En effet lorsque *V. cholerae* est présent à une densité élevée, un quorum est détecté et les auto-inducteurs AI-2 et CAI-1 sont produits. Leur détection par des récepteurs histidine kinase permet une action coordonnée au niveau de la population, entraînant une modulation des facteurs de virulence, notamment une réduction de l'expression et de la production de la protéine TcpA, ce qui signale que *V. cholerae* doit se dissocier de la surface épithéliale (Higgins et al. 2007). Jusqu'à présent, la production de l'auto-inducteur CAI-1 n'a été observée que chez *V. cholerae*. Cependant dans une étude de 2014, des chercheurs ont constaté que la production d'AI-2 par les bactéries intestinales commensales pouvait bloquer l'expression de la virulence du pathogène (Hsiao et al. 2014). Après avoir identifié des espèces bactériennes associées avec le rétablissement du choléra, les chercheurs ont reconstitué une communauté composée de 14 bactéries associées chez des souris gnotobiotiques (souris dont tous les membres du microbiome sont connus et contrôlés), puis ont infecté ces souris avec *V. cholerae*. Ils ont alors constaté qu'une de ces espèces, *Blautia obeum* (anciennement *Ruminococcus obeum*), restreint la colonisation de *V. cholerae* via la production d'autoinduceurs AI-2. Ceci démontre un potentiel mécanisme de résistance de la colonisation de l'intestin par la perturbation de la détection du quorum de *V. cholerae* via un système signalisation inter-espèces (Hsiao et al. 2014).

D'autres membres du microbiome intestinal, notamment des espèces de *Lactobacilli* et de *Streptococcus*, ont été identifiés comme des antagonistes à la croissance de *V. cholerae*, via la production de métabolites inconnus (Silva et al. 2001). Dans une étude récente, des chercheurs ont démontré que des métabolites présents dans le surnageant de culture de sept isolats de *Lactobacillus* prélevé chez des enfants en bonne santé ont la capacité d'inhiber la formation du biofilm de *V. cholerae* de manière dépendante du pH, bien que la structure et la fonction des composés antimicrobiens dans ces études restent inconnues (Kaur et al. 2018). Les biofilms sont un facteur de virulence important pour la survie de *V. cholerae*, car à un certain stade de l'infection, ils peuvent faciliter l'adhésion à l'épithélium intestinal, et protègent l'agent pathogène de l'action des antibiotiques et de la prédation par d'autres espèces de microbes intestinaux (Almagro-Moreno, Pruss, et Taylor 2015; Toska, Ho, et Mekalanos 2018). Des études chez l'homme demeurent néanmoins nécessaires pour

déterminer si les auto-inducteurs ou autres métabolites provenant de microbes intestinaux ou d'espèces modifiées pourraient avoir un impact sur les manifestations cliniques de la maladie (Duan et March 2010).

Apport de la génomique dans l'étude de l'épidémiologie et l'évolution de Vibrio cholerae

La septième pandémie de choléra persiste depuis plus de 50 ans et a été ponctuée d'épisodes d'émergence et de réémergence de la maladie dans différentes régions du monde. Notre compréhension de l'évolution de *V. cholerae* et de sa transmission à cette échelle macro-épidémiologique a considérablement été améliorée par le développement des techniques de séquençage à haut débit et par les avancées dans le domaine de la génomique bactérienne. Au cours de la dernière décennie, de nombreuses études ont eu recours au séquençage de génomes entiers et à des analyses phylogénomiques et de génomique comparative, ce qui a permis de révéler comment le pathogène *V. cholerae* a pu évoluer au cours de plusieurs décennies, et à une échelle géographique continentale. Dans une de ces premières grandes études, des chercheurs ont séquencé les génomes de 23 souches pandémiques de *V. cholerae* isolées à partir de diverses sources étendues sur 98 ans et les ont comparées afin de révéler les divers mécanismes évolutifs qui ont pu conduire aux souches actuelles (Chun et al. 2009). Ils ont ainsi découvert que les différentes souches causant les épidémies actuelles dans différentes régions du monde sont le résultat de l'évolution rapide de plusieurs descendants d'un ancêtre de *V. cholerae* O1 El Tor, qui se traduirait par plusieurs évènements de transferts horizontaux de gènes (HGT) via transduction, conjugaison et transformation, permettant l'acquisition de nouveaux îlots de pathogénicité caractérisant les souches de la septième pandémie. D'autres études ont aussi confirmé la plasticité du génome de *V. cholerae*, par l'échange fréquent de nombreux gènes, fournissant un mécanisme rapide d'adaptation à des environnements locaux (Boucher et al. 2011). Cela démontre aussi que l'utilisation de génomes entiers dans ces études se montre bien plus informative dans la définition et classification des différentes souches pandémiques que l'utilisation des méthodes de classification classiques à partir des sérogroupes (Chun et al. 2009)

Au niveau de l'épidémiologie moléculaire, une première étude publiée en 2011 a comparé plusieurs souches provenant de différentes localisations et périodes, démontrant que la pandémie actuelle de choléra est en fait le résultat de trois grandes vagues de transmission (Mutreja et al. 2011). Toutes auraient pris leur origine dans le golfe du Bengale à partir des années 50 et ont résulté en une succession d'épidémies, toutes ayant eu lieu dans des zones géographiquement restreintes. Chacune de ces vagues de transmission a été associée à un événement majeur de transfert horizontal de gènes conduisant à une augmentation de la virulence ou une adaptation à l'environnement. Ces évènements de transfert horizontal de gènes ont ainsi permis de distinguer les différentes souches pandémiques qui se différencient par les gènes de toxine cholérique (CTX), les éléments intégratifs conjugatifs (ICE) et par d'autres éléments mobiles. En outre, les auteurs ont montré comment l'acquisition de la famille SXT/ICE d'éléments de résistance aux antibiotiques a façonné la propagation de la pandémie actuelle (Mutreja et al. 2011).

Deux études plus récentes, conduites par l'équipe de Nicolas Thompson, ont comparé les séquences de plus de 1000 souches cliniques et environnementales provenant d'Amérique du Sud et d'Afrique, afin de reconstruire la propagation spatiotemporelle du choléra à travers les deux continents depuis le début de la septième pandémie (Weill et al. 2017; Domman et al. 2017). En plus de reconfirmer l'origine asiatique des différentes vagues de transmission, ces deux études ont permis d'identifier au moins douze évènements d'introduction de la maladie en Afrique depuis 1970 et trois évènements d'introduction en Amérique centrale et du Sud depuis 1991. De plus, ces données se sont montrées conformes aux études épidémiologiques, qui ont démontré que les facteurs liés à l'homme jouent un rôle beaucoup plus important dans la dynamique du choléra en Afrique et en Amérique que les facteurs climatiques et environnementaux. Ces résultats contredisent l'hypothèse stipulant que les réservoirs environnementaux aquatiques seraient la principale source de choléra épidémique au niveau de ces deux continents, comme cela a été précédemment suggéré (Colwell 1996).

D'autres approches basées sur l'étude de génomes entiers ont aussi été utilisées pour mieux comprendre l'épidémiologie du choléra au cours d'épidémies récentes. L'épidémie ayant eu lieu à Haïti en 2010 en est le parfait exemple : une approche génomique en temps réel a permis de déterminer l'origine de l'épidémie. L'analyse génomique de plusieurs isolats haïtiens a montré que la souche de *V. cholerae* responsable de l'épidémie en Haïti différait des souches

impliquées dans les épisodes de choléra du reste de l'Amérique latine, mais ressemblait fortement à des souches d'Asie du Sud (Chin et al. 2011; Sealfon et al. 2012). Ces résultats sont concordants avec une autre étude ayant comparé des souches haïtiennes avec des souches népalaises et concluant sur une possible introduction de la maladie par un contingent de Casques bleus de l'Organisation des Nations Unies déployé depuis le Népal suite au tremblement de terre, dans une région qui n'avait pas vu de cas de choléra depuis un siècle (Hendriksen et al. 2011). Ces soldats népalais provenaient d'une zone touchée par une épidémie de choléra, mais ne présentaient pas les symptômes de la maladie. Cette découverte a mis en avant l'importance du rôle des porteurs asymptomatiques dans la chaine de transmission du choléra et de quelle manière ce rôle reste encore sous-estimé. Plus récemment, l'épidémie en cours au Yémen a été reliée à une sous-lignée El Tor originaire d'Asie du Sud ayant provoqué des épidémies en Afrique de l'Est avant d'apparaître dans ce pays. Les auteurs ont aussi montré que les isolats du Yémen présentaient un profil inhabituel de sensibilité à plusieurs antibiotiques couramment utilisés pour traiter le choléra, notamment la polymyxine B, dont la résistance est habituellement utilisée comme marqueur du biotype El Tor (Weill et al. 2019).

Les techniques de plus en plus développées de séquençage nouvelle génération et les méthodes de génomique des populations ont le potentiel d'améliorer notre compréhension de l'épidémiologie, de l'étiologie, et de l'évolution des maladies infectieuses bactériennes (Wilson 2012). Tout comme de nombreux autres pathogènes bactériens, *V. cholerae* possède la capacité d'évoluer rapidement par mutation et recombinaison, mais très peu de choses sont encore connues sur sa diversité génétique et son évolution à court terme, au sein de familles et au sein d'individus infectés. Une approche utilisant la génomique des populations pourrait permettre de mieux comprendre les mécanismes de la chaine de transmission de la bactérie, et de l'adaptation à son hôte à une échelle micro-épidémiologique, comme cela a été réalisé pour d'autres pathogènes (Didelot et al. 2016).

L'évolution intra-patient des bactéries pathogènes

Apport de la génomique dans l'étude de la diversification des populations bactériennes au sein de patients infectés

Depuis le développement des méthodes de séquençage nouvelle génération et leur application sur l'étude des pathogènes, une grande majorité des études s'est d'abord concentrée sur les dynamiques évolutives et épidémiologiques à des échelles géographiques et temporelles globales. De telles analyses se sont montrées efficaces dans la reconstruction de l'histoire des épidémies et dans l'identification de leurs sources, dans l'estimation de leur échelle de temps et de leurs routes de transmission. Elles ont aussi contribué à l'identification de gènes impliqués dans la virulence et la résistance aux antibiotiques chez les bactéries (Wilson 2012). Cependant ces études présentent une résolution limitée pour comprendre et mesurer les processus démographiques élémentaires et évolutifs qui se déroulent au sein des individus infectés. Ainsi, comprendre de quelles manières la variabilité génétique des pathogènes humains est modulée par les dynamiques épidémiques, les goulots d'étranglement suivant une transmission, ou encore l'environnement à l'intérieur de l'hôte, requiert des études menées à différentes échelles : au niveau de différentes populations touchées par des épidémies distinctes, de différents patients au sein d'une épidémie, et au niveau de la population microbienne d'un patient infecté. Les dynamiques microbiennes à l'échelle d'un seul hôte représentent un aspect fondamental de la maladie clinique, et leur étude se révèle être de première importance dans la compréhension de l'origine de la diversité globale observée à plus grande échelle.

On sait que de nombreux virus et bactéries évoluent rapidement au cours d'une infection et peuvent ensuite présenter une diversité génétique intra-hôte dans leur population. La plupart des études montrant l'existence de cette diversité se sont d'abord concentrées sur les virus à ARN humains tels que le VIH, la grippe, l'hépatite C et le virus de l'hépatite B (Pybus et Rambaut 2009) et ce pour des raisons techniques. En effet, la petite taille des génomes viraux (< 30 000 pb) et les taux élevés de mutation et, parfois, de recombinaison sont responsables de l'accumulation de différences génétiques détectables. Ces différences intra-hôtes sont potentiellement plus difficiles à détecter chez les pathogènes bactériens, en raison de génomes beaucoup plus grands et d'une variation génétique plus faible. Les

approches classiques de comparaison d'isolats bactériens, afin de comprendre l'émergence de variantes causées par la microévolution, reposaient en premier lieu sur la comparaison de quelques gènes seulement, via le typage de séquences multilocus (MLST) de gènes de ménage (Morelli et al. 2010) ou par empreinte génétique standard (RFLP et VNTR), sur des échelles de temps allant de plusieurs mois à des années (Pérez-Lago et al. 2011). Toutefois, les récents progrès dans le séquençage à haut débit et l'analyse de génomes entiers et de métagénomes ont considérablement amélioré la résolution des études de populations bactériennes et offrent une sensibilité accrue pour la détection de variantes génétiques rares (Didelot et al. 2016).

Premières études de la diversité génomique de populations bactériennes au sein d'un individu

La recherche menée jusqu'à aujourd'hui sur la diversité intra-patient de pathogènes bactériens humains a révélé le degré étonnant d'adaptabilité de ces organismes, ainsi que le potentiel impact de cette diversité sur la reconstruction des évènements de transmission de patient à patient. Principalement effectuées sur des bactéries impliquées dans des infections chroniques pouvant durer de quelques mois à plusieurs années, ces études ont permis de mettre en avant les différences dans les forces évolutives agissant au sein et entre patients pour différentes espèces pathogènes.

Majoritairement basées sur la comparaison de génomes entiers d'isolats prélevés simultanément ou longitudinalement sur les mêmes hôtes, ces recherches ont montré que le niveau de diversité intra-hôte varie grandement d'une espèce pathogène à une autre, allant d'environ 30 mutations ponctuelles par génome par an pour *Helicobacter pilori* (Kennemann et al. 2011; Didelot et al. 2013) à 10-8 mutations par génome par an pour *Klebsiella pneumoniae* (Mathers et al. 2015) et *Staphylococcus aureus* (Young et al. 2012), ou encore 1 mutation par génome par an pour *Escherichia coli* (Reeves et al. 2011). Cette diversité a aussi pu être détectée chez des bactéries présentant des taux d'évolution très faibles. En effet, une étude a montré que *Mycobacterium tuberculosis*, une espèce connue pour sa croissance, son faible taux de mutation et son absence de recombinaison, était capable de s'adapter rapidement à l'utilisation d'antibiotiques, développant une résistance à sept antibiotiques différents sur une période de 42 mois (Eldholm et al. 2014). Ces différences

entre espèces sont en partie dues à des différences de taille de génome mais aussi à des taux de mutation par site variables, en lien avec l'efficacité du système de réparation des erreurs de réplication de l'ADN (système MMR, mutation mismatch repair). Certains de ces gènes sont par exemple absents chez *Helicobacter pilori* (Tomb et al. 1997), expliquant en partie son taux d'évolution élevé. Il a aussi été observé que chez certaines populations bactériennes, un taux de mutation ponctuelle beaucoup plus élevé, appelé hypermutation, peut également se produire chez certaines souches lorsque le système MMR est perturbé (Jayaraman 2011; Jolivet-Gougeon et al. 2011). De telles souches hypermutatrices ont été détectées chez des espèces infectant les poumons de patients atteints de la fibrose kystique, telles que *Burkholderia dolosa* (Lieberman et al. 2014) ou *Pseudomonas aeruginosa* (Mena et al. 2008; Oliver et Mena 2010).

Au sein de l'hôte humain, un autre facteur pouvant conduire à une plus grande diversification de la population bactérienne est le transfert horizontal de gènes (HGT). Par exemple, le processus de recombinaison homologue semble jouer un rôle important dans l'évolution de *Helicobacter pylori* au sein de son hôte, augmentant la diversité de la population lorsque plusieurs souches sont présentes (Kennemann et al. 2011; Cao et al. 2015). De même, il a été montré que la plasticité génomique liée à la perte ou au gain de gènes lors de la recombinaison non-homologue peut avoir rôle essentiel dans l'adaptation aux modifications des pressions de sélection, notamment dans la résistance aux antibiotiques. En effet, les gènes de résistance aux antibiotiques de nombreux agents pathogènes se retrouvent sur des éléments génétiques mobiles tels que les plasmides, les éléments transposables et les bactériophages (Rau et al. 2012; Stanczak-Mrozek et al. 2015).

Dynamiques évolutives au sein des individus infectés

Lorsqu'il est transmis à un nouvel hôte, un agent pathogène doit faire face à de nombreuses pressions de sélection, incluant les barrières physiques à la colonisation et à la transmission, le système immunitaire de l'hôte et la compétition avec le microbiote commensal, ou encore les traitements médicamenteux. La diversité intra-hôte est ainsi façonnée par la sélection purifiante et la sélection diversifiante, mais aussi par la dérive génétique. En effet dans le cas d'une population intra-hôte, la dérive est amplifiée par plusieurs facteurs, tels que le goulot d'étranglement lors de la transmission (Toft et

Andersson 2010), l'isolement de populations bactériennes distinctes dans différentes parties du corps au sein de l'individu (Jorth et al. 2015) ou par les fluctuations de taille de la population (Golubchik et al. 2013).

L'effet de la sélection naturelle sur la diversité intra-hôte peut prendre de nombreuses formes. À des échelles évolutives plus longues, on a observé que la sélection purifiante domine généralement le paysage évolutif, mais qu'elle peut être plus faible au sein des populations intra-hôtes, où l'isolement de la population ancestrale entraîne une plus grande dérive génétique et où dans certains cas, il y a un manque de temps disponible pour purger les mutations qui ne sont que légèrement délétères (Rocha et al. 2006). L'action de sélection positive diversifiante a été détectée dans de nombreux cas d'évolution intra-patient (Lieberman et al. 2011; E. P. Price et al. 2013; Marvig et al. 2015), témoignant de l'adaptabilité rapide de certains pathogènes opportunistes. Le faible nombre de mutations habituellement détecté au sein d'une population intra-hôte peut rendre la détection de la sélection difficile. Par exemple, dans une étude sur l'évolution intra-hôte de *Burkholderia dolosa*, les auteurs ont montré qu'un résultat du test de dN/dS (qui compare le ratio de substitutions non-synonymes (dN) à celui des substitutions synonymes (dS)) ne devait pas forcément être interprété comme signe de l'absence de sélection dans cette population (Lieberman et al. 2011). En effet, la sélection positive sur certains gènes et la sélection purifiante sur d'autres gènes s'annulent et fournissent dans ce cas un signal d'évolution neutre. À la place, les auteurs ont mis en avant des signes de sélection positive par l'observation d'une évolution convergente entre les populations bactériennes de différents patients, présentant des mutations non-synonymes sur les mêmes gènes. Ainsi la capacité de détecter une évolution parallèle et convergente s'est révélée être un outil particulièrement puissant pour identifier l'évolution adaptative dans ces populations intra-patient (Wong et Kassen 2011; Lieberman et al. 2011; E. P. Price et al. 2013; Marvig et al. 2015), particulièrement pour des pathogènes opportunistes. Mises ensemble, ces observations indiquent que l'effet dominant de la sélection naturelle chez l'hôte serait de conserver la fonctionnalité pour la majorité des gènes, l'adaptation n'agissant que sur certains gènes importants dans l'adaptation au nouvel hôte.

Impact de la diversité intra-patient sur la reconstruction des évènements de transmission

Actuellement, l'une des utilisations les plus courantes du séquençage de génomes bactériens entiers est dans la reconstruction des réseaux de transmission à travers le monde ou au niveau des épidémies (Croucher et Didelot 2015; Didelot et al. 2017; Gardy et Loman 2018). Dans ce contexte, la présence d'une diversité génétique intra-patient représente à la fois des défis et des opportunités pour la compréhension des dynamiques de transmission des pathogènes. D'une part, la présence d'une diversité au sein du patient signifie que la pratique consistant à utiliser la distance génomique entre des isolats uniques prélevés chez les individus pour déduire les réseaux de transmission peut être sérieusement biaisée (Didelot, Gardy, et Colijn 2014). En effet, à cause d'une potentielle diversification à l'intérieur de l'hôte, deux isolats d'un même patient peuvent être séparés par plus de mutations que les isolats de différents patients, et la topologie du réseau de transmission sera radicalement modifiée en fonction du choix d'isolats effectué.

Ainsi, lorsqu'elle est prise en compte, la diversité intra-patient peut aider à identifier les événements de transmission entre individus (Worby, Lipsitch, et Hanage 2014; Martin et al. 2018). Des séquences génomiques suffisamment diverses peuvent être utilisées pour inférer l'appartenance à une même épidémie (Snitkin et al. 2012; 2013; S. R. Harris et al. 2013) ou encore, exclure des événements de transmission potentiels (E. P. Price et al. 2013). Ainsi, l'intégration des données génomiques et épidémiologiques intra-hôte et inter-hôte peut permettre de mettre en évidence différentes voies de transmission. Dans certains cas, la diversité intra-hôte est le résultat de la co-infection de diverses souches d'origines différentes. Un arbre phylogénétique qui comprend de nombreux isolats de chaque patient dans une épidémie et qui est enraciné avec un isolat extérieur à l'épidémie peut aider à identifier les événements de transmission; les individus infectés par d'autres patients auront leurs isolats imbriqués dans la diversité observée chez cet autre patient. Parfois, ces multiples événements de transmission peuvent être détectés et, en utilisant des informations sur le contact entre les cas et leur exposition à des sources possibles d'infection, peuvent expliquer l'infection mixte observée (Snitkin et al. 2012; Eyre et al. 2013).

Toutefois, cette approche présente encore de nombreuses limitations : les niveaux de diversité faible dues au manque de temps pour accumuler les mutations avant la transmission

peuvent rendre impossible l'inférence de la direction de la transmission; les pressions de sélection au sein des hôtes infectés peuvent entrainer une évolution convergente pouvant conduire à des inférences phylogénétiques erronées; ou un échantillonnage insuffisant peut sous-estimer la diversité réelle et empêcher une inférence correcte des évènements de transmission (Didelot, Gardy, et Colijn 2014; Didelot et al. 2016). Cependant l'amélioration des méthodes de séquençage et la diminution des coûts permettent aujourd'hui d'obtenir une meilleure résolution pour la détection de cette diversité et de parer à ces limites (R. S. Lee et al. 2020).

La diversité génétique intra-hôte de Vibrio cholerae

Contrairement aux populations bactériennes impliquées dans des infections à long terme, la diversité intra-hôte de populations impliquées dans des infections aiguës, telles que le choléra, a été peu étudiée à ce jour (Erin P. Price et al. 2010). Cependant certaines études récentes ont montré que cette diversité existait chez *Vibrio cholerae*. Par exemple, il a été montré que les populations intra-hôtes de *Vibrio cholerae* présentaient une variabilité au niveau des loci VNTR, comme le montre une autre étude où 8 des 9 patients échantillonnés exhibaient au moins deux génotypes VNTR différents dans leurs échantillons de selles (Kendall et al. 2010).

Une autre étude a aussi montré qu'une variation de phase au niveau du génome de *Vibrio cholerae* pouvait rapidement générer de la diversité au niveau de l'antigène O, un élément majeur de la couche de lipopolysaccharides de la membrane externe, et qui sert de cible aux bactériophages et au système immunitaire (Seed et al. 2012). Des expérimentations *in vitro* ont montré que des variations au niveau de deux gènes impliqués dans la biosynthèse de l'antigène O avaient un impact sur la modulation de la résistance aux infections phagiques, et que ces variants étaient présents au sein d'hôtes humains (Seed et al. 2012). Une étude plus récente menée par la même équipe de recherche a découvert d'autres protéines de surface subissant des mutations intra-hôtes permettant à *Vibrio cholerae* d'échapper à la prédation des phages. Dans cette étude, les auteurs ont étudié les échantillons de selles de deux patients souffrants du choléra (un venant du Bangladesh, l'autre d'Haïti), et présentant tous deux une forte charge virale d'un vibriophage nommé ICP2. Dans chaque échantillon, les chercheurs ont découvert que certaines des bactéries étaient résistantes au phage, tandis que

les autres y étaient sensibles. Le séquençage du génome des clones bactériens résistants et des clones sensibles aux phages a permis de découvrir que la seule différence entre ces variants se trouvait dans un gène unique. Cependant, un gène bactérien différent a été muté chez chaque patient. Plusieurs mutants différents de chaque gène ont été trouvés, suggérant fortement que ces mutations se sont produites et ont ensuite été sélectionnées au sein du patient lors de l'infection.

Au sein du patient Haïtien, presque toutes les bactéries isolées étaient résistantes au phage (> 99%); et parmi les bactéries testées résistantes aux phages, toutes présentaient l'une des six mutations différentes dans un seul gène appelé *ompU*. La protéine *OmpU* forme un pore dans la membrane externe de la bactérie pour permettre aux nutriments d'être importés dans la cellule. Les bactéries ont besoin de cette protéine pour leur survie à la fois dans les hôtes humains et dans les eaux environnementales. Les mutants *OmpU* étant résistants à la prédation des phages, les résultats de Seed et al. ont permis d'identifier cette protéine comme le récepteur possible du phage ICP2.

Dans l'échantillon de selles du patient bangladais, 22% des isolats bactériens se sont montrés résistants à l'infection du phage ICP2; et les auteurs ont identifié quatre mutations au niveau du gène *toxR*, codant pour une protéine qui régule l'expression de nombreux gènes, y compris l'*OmpU*. Étant donné que ces mutants *ToxR* ne produisent pas le récepteur du phage - la protéine *OmpU* - cela confère une résistance à l'attaque des phages. Cependant, la protéine *ToxR* est aussi une protéine importante dans la régulation également les gènes de virulence de *Vibrio cholerae*, et des expériences sur des modèles murins ont montré que ces mutants *ToxR* sont incapables d'être transmis et de déclencher l'infection chez un autre individu (Seed et al. 2014).

Structure générale et objectifs de la thèse

Le **Chapitre 1** de cette thèse représente un survol des connaissances concernant les dynamiques entre *Vibrio cholerae,* son hôte humain, et les membres du microbiote intestinal, ainsi que de l'état des connaissances concernant l'évolution intra-patient des bactéries pathogènes. Comme présenté dans ce chapitre, les analyses génomiques ont clairement avancé notre compréhension de la propagation et de l'évolution de la bactérie *Vibrio cholerae* à l'échelle des continents touchés par la pandémie. Un nombre croissant d'études supporte l'apport majeur de la génomique des populations dans la compréhension des dynamiques intra-hôtes des bactéries et dans l'identification des mécanismes évolutifs impliqués dans la progression des maladies. Dans le cadre de ce doctorat, je me suis en premier lieu penchée sur la caractérisation de la diversité génétique de *Vibrio cholerae* au sein de patients infectés, en utilisant plusieurs méthodes basées sur le séquençage nouvelle génération. Par la suite je me suis intéressée au microbiome intestinal de personnes susceptibles d'être infectées par la bactérie, et à la possibilité de prédire l'infection en analysant la composition de ce microbiome.

Bien que déjà observée dans une première étude (Seed et al. 2014), la diversité génomique intra-patient de *Vibrio cholerae* n'avait pas été caractérisée à ce jour. Dans le **Chapitre 2**, je me suis penchée sur la caractérisation de cette diversité au sein de patients venant du Bangladesh et d'Haïti, deux lieux présentant des dynamiques épidémiologiques différentes (Levade et al. 2017). Les objectifs de ce chapitre étaient d'abord de (1) déterminer l'étendue de la diversité observée dans des populations bactériennes infectant différents patients, à partir du séquençage de génomes entiers d'isolats. (2) Nous avons par la suite identifié la nature de cette diversité (mutations ponctuelles? présence/absence de gènes?) et les régions du génomes affectées. (3) Enfin, nous avons testé les potentiels effets phénotypiques des variations détectées, et tenté de distinguer les pressions de sélection ou autres forces évolutives pouvant affecter la population bactérienne au sein de l'hôte.

Dans cette étude nous avons montré que l'évolution intra-hôte, bien que limitée, se produit effectivement lors des infections par le choléra, avec des conséquences phénotypiques pour la population de *V. cholerae*. Nous avons prélevé 8 à 20 isolats de *V. cholerae* sur chacun

des huit patients (cinq du Bangladesh et trois d'Haïti) et séquencé 122 génomes bactériens au total. En utilisant des contrôles rigoureux pour éviter les erreurs de séquençage, nous avons identifié des variants nucléotidiques intra-hôte (iSNV) de haute qualité et une variation du contenu génétique (événements de gain/perte de gènes chez les patients). Pour évaluer la variation phénotypique chez les patients, nous avons comparé les taux de croissance des isolats de *V. cholerae* en milieu liquide et effectué des tests pour la formation de biofilms, un trait important sur le plan clinique et environnemental.

Nous avons ainsi pu détecter un niveau de variation faible, mais mesurable chez les patients lors d'infections par le choléra (0 à 3 iSNV par patient), probablement dû à des mutations intra-hôte plutôt qu'à une co-infection par différentes souches. La plupart des iSNV se sont montrés sélectivement neutres, mais deux mutations non synonymes dans un gène codant pour une histidine kinase semblaient être favorisées par la sélection naturelle chez un seul patient originaire d'Haïti. Nous avons aussi observé que le gain/perte de gènes était la principale source de diversité intra-hôte, mettant en avant quelques cas de transferts horizontaux de gènes entre *V. cholerae* et d'autres espèces du microbiote intestinal. De plus, nous avons démontré que ces mutations intra-patient et événements de gain/perte de gènes peuvent avoir des conséquences phénotypiques, par exemple sur la capacité de la bactérie à produire des biofilms.

Dans le **Chapitre 3**, nous avons confirmé la diversité génétique de population intra-patient de *Vibrio cholerae,* cette fois-ci en utilisant la méthode métagénomique, consistant à séquencer tout le matériel génétique présent dans un échantillon sans passer par une étape de culture d'isolats bactériens. Cette méthode a été appliquée à la fois sur des échantillons provenant de patients symptomatiques et de patients asymptomatiques mais ne s'est révélée efficace que dans le premier cas, les patients asymptomatiques ne présentant pas suffisamment de cellules de *Vibrio cholerae* dans leurs selles. Les objectifs de ce chapitre étaient d'abord de (1) identifier les biais potentiels de la méthode basée sur la culture puis le séquençage d'isolats et déterminer si une meilleure résolution peut être obtenue avec la méthode métagénomique pour la détection de variants. (2) Nous avons tenté d'identifier les potentielles pressions de sélection pouvant agir sur les variants identifiés par la méthode

métagénomique. (3) Enfin, nous avons comparé la diversité observée au sein de population de *Vibrio cholerae* infectant des patients présentant des symptômes différents.

En ce qui concerne les patients symptomatiques, la méthode métagénomique a montré une meilleure résolution pour la détection de variants intra-hôtes sous la forme de mutations ponctuelles. Sur les 15 échantillons séquencés, 6 ont montré un nombre de mutations particulièrement élevé, bien plus que la variation décrite dans le Chapitre 2. Nous avons détecté des mutations dans des gènes impliqués dans la réparation des mésappariements de l'ADN, indiquant la présence de souches de *Vibrio cholerae* hypermutantes dans ces échantillons. Ces souches hypermutantes, décrites pour la première fois dans un contexte clinique au sein de patients infectés, ont cependant été trouvées à faible fréquence dans la population et ne semblaient pas être fortement sélectionnées par les pressions sélectives au sein de l'hôte. Au contraire nous avons détecté des signes d'évolution convergente au sein des patients ne présentant pas de phénotype d'hypermutation, deux patients présentant des mutations dans le même gène codant pour une toxine impliquée dans la virulence du pathogène. Ces résultats, combinés aux résultats du Chapitre 2 et l'étude de Seed et al, indiquent la nature idiosyncratique de l'action de la sélection au sein des patients infectés par *Vibrio cholerae.*

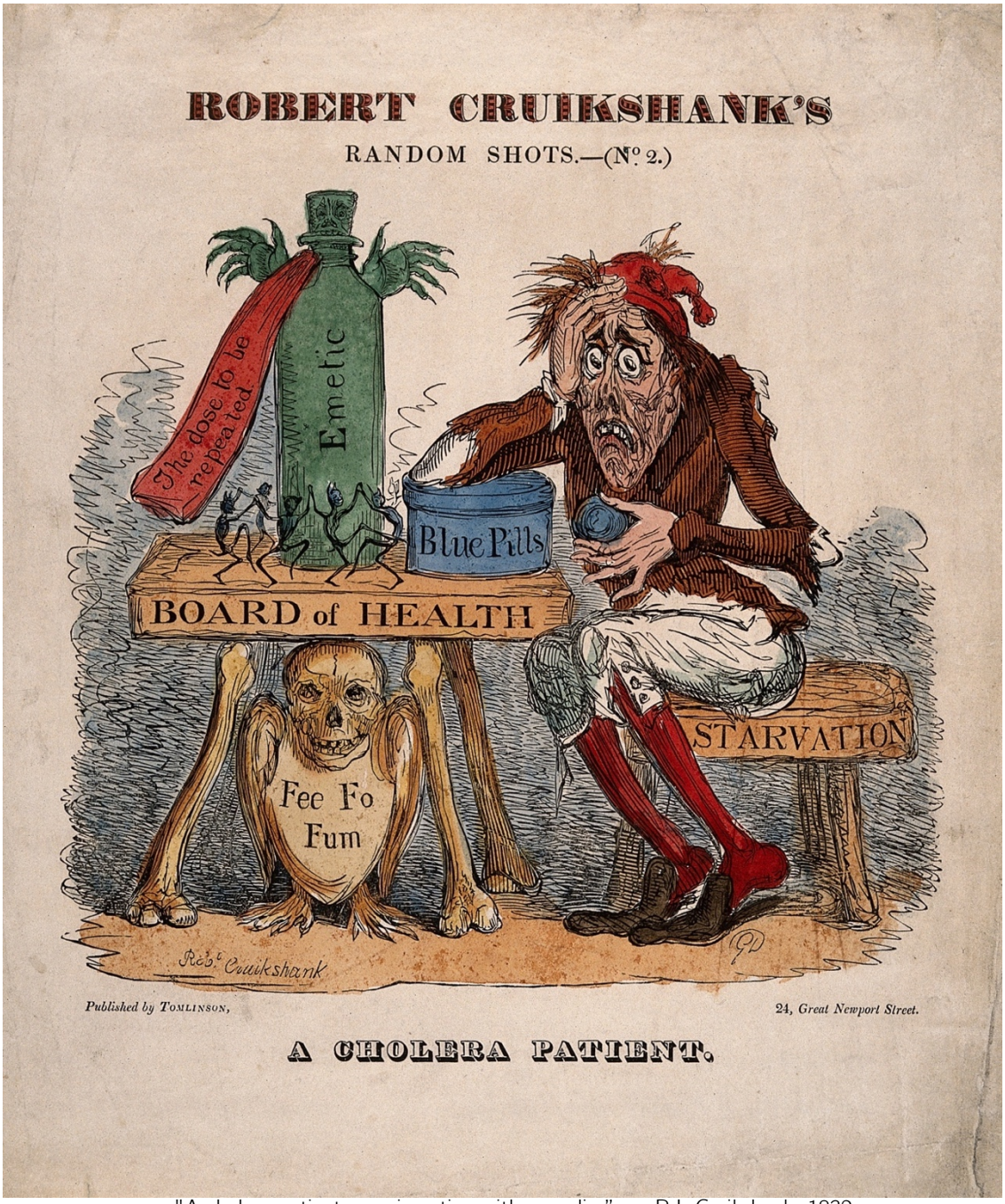
En utilisant la méthode de culture bactérienne suivit du séquençage de génome entier, nous avons pu montrer que les patients infectés par le pathogène, mais ne présentant pas de symptômes, pouvaient eux aussi être porteurs de plusieurs souches de *Vibrio cholerae*, incluant des souches hypermutantes. De plus, bien que n'excluant pas la possibilité que certains variants intra-patients soient plus virulents que d'autres et puissent expliquer une partie de la différence dans la sévérité des symptômes, nous avons montré que la diversité au sein de ces patients asymptomatiques, à la fois au niveau des mutations ponctuelles, mais aussi au niveau de la présence/absence de gènes, ne semblait pas être plus ou moins élevée que celle observée chez les patients symptomatiques.

Dans le **Chapitre 4**, nous nous sommes intéressés au microbiome intestinal et à sa composition taxonomique et fonctionnelle comme potentiel facteur de risque pour l'infection par *Vibrio cholerae* et pour la sévérité des symptômes. Ce travail s'appuie en grande partie sur une étude précédente réalisée par nos collaborateurs (Midani al. 2018), qui a montré que

la composition taxonomique du microbiome intestinal "de base" pouvait servir à la prédiction des risques d'infection après une exposition pour les contacts familiaux de patients. Cette étude a utilisé le séquençage d'amplicons de l'ARNr 16S pour étudier la composition taxonomique du microbiome intestinal à une résolution relativement faible, ce qui rendait difficile l'identification de bactéries potentiellement protectrices pour un suivi expérimental. Au cours de notre étude, avons appliqué le séquençage métagénomique et une méthode d'apprentissage machine à la même cohorte prospective échantillonnée au Bangladesh, afin de caractériser de manière plus précise les espèces bactériennes et voies métaboliques potentiellement associées à la protection contre le choléra.

Nous avons ainsi pu montrer que les données métagénomiques permettent une prévision plus précise de la susceptibilité à l'infection par *Vibrio cholerae* par rapport aux facteurs de risque cliniques connus (tel que l'âge, les titres d'anticorps ou le groupe sanguin) ou le séquençage de l'amplicon 16S. Cette étude a notamment permis la détection de souches spécifiques de *Prevotella* et de *Bifidobacterium* qui pourraient être impliquées dans la protection contre le choléra, mais aussi l'identification de familles de gènes et de voies cellulaires spécifiques - notamment le métabolisme et le transport du fer - comme mécanismes potentiels de protection.

Nous avons aussi pu démontrer qu'en plus du statut infecté/non infecté, la gravité de la maladie (maladie diarrhéique vs infection asymptomatique) pouvait également être prédite à partir du microbiome de base, mais avec une plus grande incertitude.



"A cholera patient experimenting with remedies" par R.I. Cruikshank, 1832

Chapitre 2 : Diversité génomique intra-hôte et inter-hôte de *Vibrio cholerae* au sein de patients infectés

Vibrio cholerae genomic diversity within and between patients

Inès Levade[1], Yves Terrat[1], Jean-Baptiste Leducq[1], Ana A. Weil[2,3], Leslie M. Mayo-Smith[2], Fahima Chowdhury[4], Ashraful I. Khan[4], Jacques Boncy[5], Josiane Buteau[5], Louise C. Ivers[3,6,7], Edward T. Ryan[2,3,8], Richelle C. Charles[2,3], Stephen B. Calderwood[2,3,9], Firdausi Qadri[4], Jason B. Harris[2,10], Regina C. LaRocque[2,3], B. Jesse Shapiro[1]

[1]Département de Sciences biologiques, Université de Montréal, Montréal, Québec, Canada

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

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

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

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

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

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

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

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

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

Published in Microbial Genomics 2017 Dec; 3(12): e000142.

DOI: 10.1099/mgen.0.000142

Résumé

Le choléra est une grave maladie diarrhéique d'origine hydrique causée par des souches la bactérie *Vibrio cholerae* capable de produire la toxine cholérique. Des études de génomique comparative ont révélé des "vagues" de transmission et d'évolution du choléra, au cours desquelles différentes souches ont été successivement remplacées au fil des décennies et des siècles. Cependant, l'étendue de la diversité génétique de *V. cholerae* au sein d'une épidémie ou même chez un patient individuel est mal comprise. Ici, nous avons caractérisé la diversité génomique de *V. cholerae* à un niveau micro-épidémiologique au sein et entre des patients individuels du Bangladesh et d'Haïti. Pour saisir la diversité au sein des patients, nous avons isolé plusieurs (8 à 20) colonies de *V. cholerae* chez chacun des huit patients, séquencé leur génome et identifié les mutations ponctuelles et les événements de gain/perte de gènes. Nous avons trouvé une diversité limitée mais détectable au niveau des mutations ponctuelles au sein des hôtes (de zéro à trois variants nucléotidiques uniques chez chaque patient), et une variation comparativement plus importante du contenu génétique au sein des hôtes (au moins un événement de gain/perte par patient, et jusqu'à 103 événements chez un patient). Une grande partie de la variation du contenu génétique semble être due au gain et à la perte de phages et de plasmides au sein de la population de *V. cholerae*, avec des échanges occasionnels entre *V. cholerae* et d'autres membres du microbiote intestinal. Nous montrons également que certaines variantes intra-hôtes ont des conséquences phénotypiques. Par exemple, l'acquisition d'un plasmide de *Bacteroides* et des mutations non synonymes dans un gène codant pour une protéine de détection de type histidine kinase ont montré une réduction dans la formation de biofilms, un trait important pour la survie de l'environnement. Ensemble, nos résultats montrent que *V. cholerae* évolue de manière mesurable chez les patients, avec des implications possibles sur l'évolution de la maladie et la dynamique de sa transmission.

Mots-clés : *Vibrio cholerae*, génomique comparative, évolution intra-hôte, transfert horizontal de gènes, biofilm

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 non-synonymous 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.

Keywords: *Vibrio cholerae*, comparative genomics, within-host evolution, horizontal gene transfer, biofilm

Impact statement

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*. By sequencing multiple bacterial isolates from each of eight patients from Bangladesh and Haiti, we found that cholera patients can harbour 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 feature of within-host evolution.

Introduction

Cholera is an acute diarrhoeal infection that remains a serious health threat in countries with limited access to clean water (Harris et al. 2012). *Vibrio cholerae* is the causative agent of the disease and is a natural inhabitant of aquatic ecosystems (Reidl and Klose 2002), with more than 200 serogroups identified to date on the basis of their somatic O antigens (Chatterjee et Chaudhuri 2003; Seed et al. 2012). 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 (Harris et al. 2012).

Whole genome sequencing and population genomics have the potential to improve our understanding of the epidemiology, aetiology and evolution of bacterial infectious diseases (Wilson 2012). For example, comparisons of whole-genome sequences of strains of *V. cholerae* from across the world, over the course of a century, clarified the history of the current pandemic (Mutreja et al. 2011) 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 (Chun et al. 2009). More recently, comparative genomics has been applied to answer epidemiological questions, proving the Asian origin of the strain causing the ongoing Haitian cholera outbreak, which began in 2010 (Chin et al. 2011; Hendriksen et al. 2011; Katz et al. 2013; Orata, Keim, et Boucher 2014). Using whole genome sequencing and single nucleotide polymorphism (SNP) analysis, Azarian et al. (2014) 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 (Azarian et al. 2014).

Viral pathogens can evolve and diversify within infected patients, with serious consequences for disease outcome (Pybus et Rambaut 2009), and certain bacterial pathogens have recently been shown to diversify within patients as well (Didelot et al. 2016). However, evolutionary and epidemiological studies have conventionally been conducted with just one bacterial isolate taken as representative of the infection, even though within-patient diversity is important to consider, for several reasons (Patra et al. 2012; Worby, Lipsitch, et Hanage 2014; Didelot, Gardy, et Colijn 2014; Croucher et Didelot 2015). Within-host evolution may

impact the longer-term evolution and transmission potential of pathogens, particularly if there are fitness trade-offs 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 (Seed et al. 2014). The 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 also whether intra-host evolution is generally driven by selection.

As for many other bacterial pathogens, the prevailing orthodoxy is that *V. cholerae* infections are essentially clonal, and essentially devoid of within-host genetic diversity. Although within-host populations of *V. cholerae* have not been studied extensively, 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 (Seed et al. 2012; Kendall et al. 2010). 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 (Faruque et al. 2004) and co-infections from diverse environmental strains are possible (Rashed et al. 2014). *V. cholerae* infections are acute, lasting only a few days before the patient either recovers or dies (Harris et al. 2012). Therefore, there is limited time for within-host evolution (including mutation, recombination and selection) to occur. On the other hand, measurable within-host evolution has been demonstrated over just 6 days in a *Burkholderia pseudomallei* infection (Limmathurotsakul et al. 2014). It is therefore expected that even if *V. cholerae* will experience less within-host evolution compared to more chronic bacterial infections with documented within-host evolution (Limmathurotsakul et al. 2014; Morelli et al. 2010; McAdam et al. 2011; Golubchik et al. 2013; Lieberman et al. 2014), diversity among isolates from the same patient could be detectable. Indeed, *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 (Nelson et al. 2009; David et al. 2015). 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 (Abel et al. 2015), which would reduce genetic diversity and reduce the efficiency of selection. In addition to point mutations, *V. cholerae* can undergo high rates of HGT (Chun et al. 2009; Keymer et Boehm 2011; Boucher et al. 2011), 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 hot-spot of HGT (LaRocque et al. 2005). However, the extent of within-patient mutation, HGT and natural selection are still poorly known for *V. cholerae*.

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

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. In Haiti, samples were collected from patients presenting to St. Marc's Hospital, Arbonite, with acute watery diarrhea in April 2013. See the Supplementary Methods for more details on sample collection.

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

Stools from both Haiti and Bangladesh were stored at Massachusetts General Hospital (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, 20 well-separated colonies were inoculated into 5 ml Luria-Bertani broth and grown at 37 °C overnight. For each colony, 1 ml 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 artefacts and sequencing errors. Bacterial stocks made from a single colony were grown in 1.5 ml LB media with agitation at 37 °C for 12 h. Colonies were named with a 'C', followed by a number; for example, B1C1 corresponds to colony 1 from Bangladesh patient 1. All isolates 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. Genomic DNA was extracted for each isolate using the Qiagen DNeasy Blood and Tissue kit, using 1.5 ml bacteria grown 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.

Whole genome sequencing

Each isolate was separately sequenced using Illumina technology to a minimum depth of 28× coverage of the MJ-1236 reference genome (mean coverage = 136×). From 122 isolates retrieved from the eight patients, 66 genomic libraries were constructed using the Nextera DNA library kit (Illumina) according to the manufacturer's protocol, 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 (New England Biolabs) and sequenced on the Illumina HiSeq 2500 (paired-end 125 bp) at the Genome Quebec sequencing platform (McGill University). Twelve isolates were sequenced in replicate using both methods. For details about isolates, sequencing and assembly see Table S1.

Each genome was *de novo* assembled, and reads were also mapped to two closed, annotated *V. cholerae* reference genomes. After filtering for errors due to culture and sequencing (see Supplementary Methods), we identified a total of 485 high-quality single-nucleotide variants (hqSNVs) and 22 indels among our 122 isolates, distributed across the genome (Figure S1), with a majority of hqSNVs (471) located in the branch between the Bangladeshi and the Haitian isolates. Among them, we characterized the intra-host single nucleotide variants (iSNVs) that varied among the isolates taken from a single patient (Table 1).

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 (Tajima 1989). We assumed a mutation rate of $\mu = 1/300$ per genome per generation (Drake et al. 1998). We report the estimated N_e for each patient in Table S2, 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 described by Tajima (Tajima 1989).

Characterization of the flexible genome

From assemblies, we annotated the genomes using the RAST pipeline (Aziz et al. 2008) with default parameters. Predicted proteins were used as input for the OrthoFinder

software (Emms et Kelly 2015) 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 genomes (<1 copy per genome). Due to variation in sequencing coverage and nucleotide composition, genome assemblies may be incomplete, missing a subset of genes that are actually present (Alkan, Sajjadian, et Eichler 2011; Denton et al. 2014). 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 (Supplementary Methods; Figs S2 and S3). Potential donors of horizontally transferred genes were identified using blast against NCBI GenBank, followed by phylogenetic analyses (Supplementary Methods).

Phylogenetic evolutionary inference and root-to-tip regression

For all phylogenetic analyses, we excluded the Integrative Conjugative Element (ICE), a highly variable 100-kb region (Figure S1) that undergoes frequent HGT and thus has a separate evolutionary history from the rest of the core genome (Mutreja et al. 2011). 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 (Taviani et al. 2009). A final alignment of 201 concatenated hqSNVs 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. Seaview v.4.5.4 (Gouy, Guindon, et Gascuel 2010) was used to generate a maximum-likelihood phylogeny, employing a general time reversible (GTR) substitution model with four rate classes and subtree pruning and regrafting (SPR) branch-swapping. All sites being variable in the alignment, we did not consider the proportion of invariable sites. To test the degree of temporal structure of our data we performed a root-to-tip linear regression using the TempEst software [44], which suggested that our dataset is sufficiently clock-like to robustly estimate an evolutionary rate ($R2 = 0.65$, $p < 0.0001$; Figure S4). Substitution rates and divergence times were then estimated using beast v.1.8.3 [45], with XML-input files manually modified to include both variable and invariable sites, for a total genome the size of the *V. cholerae* MJ-1236 reference. The Bayesian tree was computed and calibrated with sampling dates of the isolates ranging from March 2011 to December 2013. We compared different molecular clock and demographic

models (Supplementary Methods), and found that the strict molecular clock and the Bayesian skyline plot models provided the best fit (Table S3), in accordance with previous studies of *V. cholerae* (Azarian et al. 2014; Duchêne et al. 2016).

Tests for violations of neutral evolution

We conducted permutations of the distribution of non-synonymous, synonymous and intergenic SNVs across sites in the genome and branches of the phylogeny to identify any possible deviations from neutral evolution (Supplementary Methods). To investigate the role of natural selection within versus between patients, we performed the McDonald-Kreitman test (McDonald et Kreitman 1991) and also conducted permutations to identify any patients with an excess of non-synonymous iSNVs compared to random distribution of iSNVs across patients (Supplementary Methods).

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

To identify possible phenotypes conferred by the within-patient variations, we performed *in vitro* experiments on the isolates from patient H1 harbouring the three NS point mutations, and the isolate 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 the variable isolates compared to one isogenic isolate (with inferred ancestral alleles and no variation in the flexible genome) from the same patient. Isolates were grown in 4 ml LB broth with agitation at 30 °C, and optical densities were measured at 600 nm using a spectrophotometer every hour for 12 h. Second, to test for biofilm production, we grew the same isolates in 200 µl LB in a 96-well plate, without agitation at 30 °C for 48 h. Controls included empty wells, wells with LB only, and wells with a *V. cholerae* isolate with an in-frame deletion in the gene vpsA (*Vibrio* polysaccharide A), which results in reduction of biofilm production [48]. 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 12.

Polymyxin B MIC assay on patient H1 isolates

Isolates 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 E-test gradient strips (AB Biodisk) were applied to inoculated plates and incubated at 37°C, and the MICs were evaluated after 16 h.

Results

Within-patient single nucleotide variation

Among the 485 hqSNVs detected, we found intra-host single nucleotide variants (iSNVs) in patients B1, B4, B5 and H1, from whom we sequenced respectively 19, 17, 20 and 9 isolates (Figure 1). Isolates from patients B2, B3, H2 and H3 (20, 20, 8 and 9 isolates, respectively) were all isogenic within patients, with no iSNVs detected using our quality filters. Patient B1 contained one intergenic iSNV (with an allele frequency of 1/19), and patients B4 and B5 each contained a synonymous iSNV, each in a different gene (with respective allele frequencies of 1/17 and 1/20) (Table 1). Patient H1 contained three iSNV sites, all of which were non-synonymous, and two of which occurred in the same gene, a sensor histidine kinase (Table 1), and one of which was at frequency 2/9, for a total of 4/9 isolates containing iSNVs (Figure 1). Twenty-two small insertion/deletions (indels) were found to vary between patients (and specifically between patients sampled in different years or different countries), but we did not detect any indels that varied within patients.

To ensure that the relatively small number of iSNVs were not due to mutation during isolate isolation and culture (Draper et al. 2017) or sequencing errors, we sub-cultured and sequenced 12 colonies from one isolate (B1C1) as a control (Figure 1). Applying the same filters as for SNV discovery in our patients, we did not detect any iSNVs among control isolates, nor did we detect any SNV differences between replicate libraries prepared and sequenced using different platforms (Methods). This suggests that the few iSNVs identified within patients are unlikely to be culture or sequencing artefacts.

Based on the number (0-3) and frequencies of iSNVs per host, we used measures of genetic diversity (θW and π) to estimate within-host effective population size (N_e). We estimated that N_e within each patient ranged from 0 to 110 (Table S2). 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.

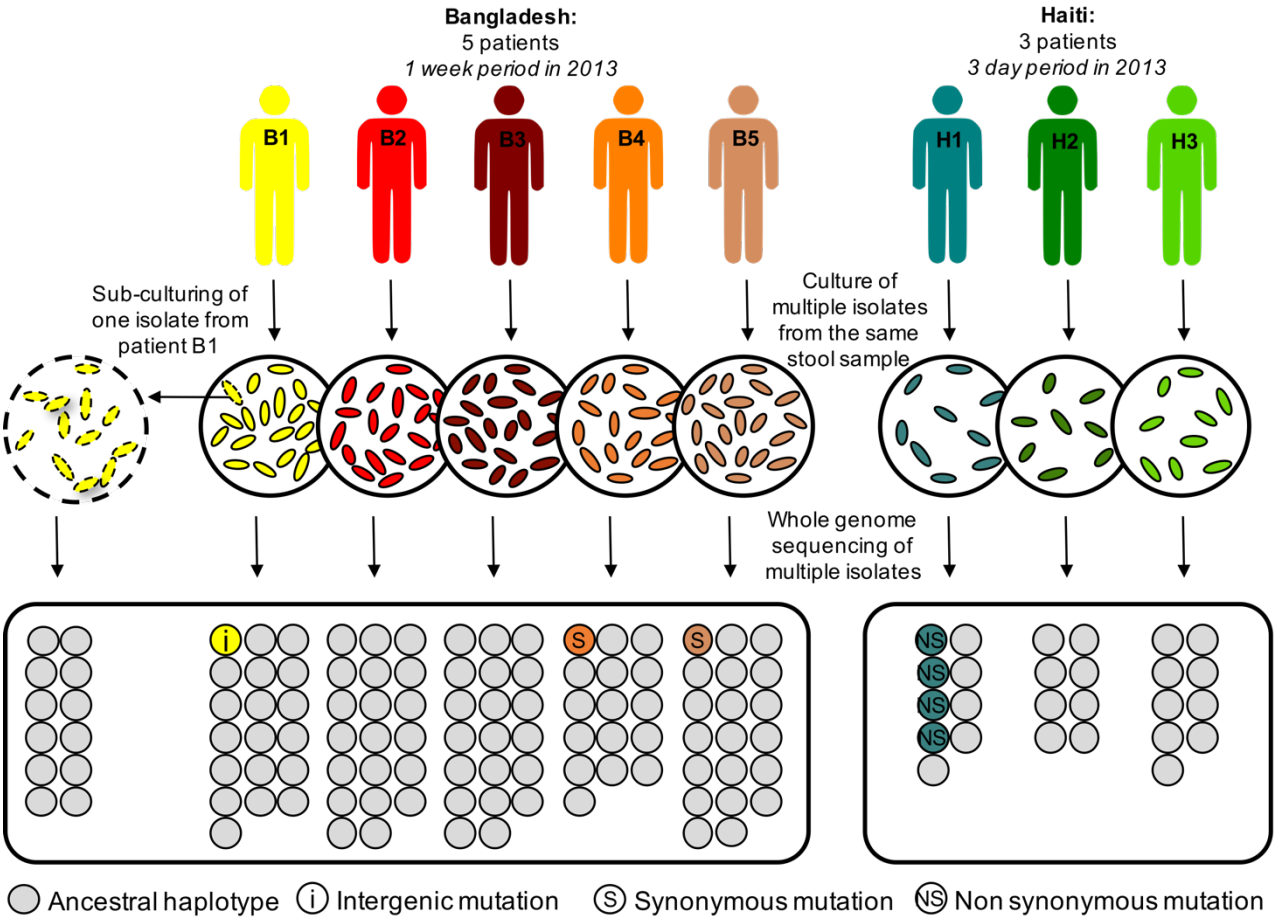


Figure 1. Culture and sequencing of *Vibrio cholerae* isolates from eight acutely infected patients. To study within-patient evolution, selective media was used to culture stool samples from five patients from Bangladesh (B1 to B5) and three patients from Haiti (H1 to H3). Between eight and 20 colonies were isolated from each patient and sequenced separately. For patient B1, we performed a sub-culture of one isolate (dotted outline) and sequenced 12 of these new isolates as a control for cultured-induced and sequencing artefacts. We independently called variants, compared them between isolates within each patient to identify the intra single nucleotide variants (iSNVs, coloured circles) and determined whether they were intergenic (i), synonymous (S), or non-synonymous (NS) mutations.

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

Region	Type	Isolates	Ref. nucleotide	Alt. nucleotide	Nucleotide position	NS/S	Protein	Ref. amino acid	Alt. amino acid	Gene annotation	Patient allele frequency
Bangladesh	iSNV	B1C19	C	T	CHR1, 746965	intergenic	-	-	-	-	1/19
Bangladesh	iSNV	B4C12	C	T	CHR1, 2549743	S	ACQ61457.1	D	D	RNA polymerase sigma-70 factor	1/17
Bangladesh	iSNV	B5C11	T	G	CHR1, 2764922	S	ACQ61649.1	V	V	toxin co-regulated pilus biosynthesis protein F	1/20
Haiti	iSNV	H1C5	G	T	CHR1, 2240431	NS	ACQ61177.1	R	L	sensor histidine kinase	1/9
Haiti	iSNV	H1C4 H1C8	G	A	CHR1, 1785021	NS	ACQ60802.1	R	H	TetR family transcriptional regulator	2/9
Haiti	iSNV	H1C6	G	T	CHR1, 2241580	NS	ACQ61177.1	R	L	sensor histidine kinase	1/9
Bangladesh	Patient	B2 + B3	T	C	CHR1, 1295440	intergenic	-	-	-	-	20/20 20/20
Bangladesh	Patient	B4	C	T	CHR1, 610212	NS	ACQ59736.1	G	R	TatD family hydrolase	17/17
Bangladesh	Patient	B4	C	T	CHR2, 1007076	NS	ACQ62879.1	A	T	hypothetical protein	17/17
Bangladesh	Patient	B2 + B3	G	A	CHR1, 1036008	intergenic	-	-	-	-	20/20 20/20
Bangladesh	Patient	B4	C	T	CHR1, 1400553	intergenic	-	-	-	-	17/17
Bangladesh	Patient	B4	G	A	CHR2, 350409	NS	ACQ62264.1	P	L	PTS system mannitol-specific EIICBA component	17/17
Bangladesh	Patient	B2 + B3	G	A	CHR1, 2301641	NS	ACQ61224.1	P	S	LacI family transcription regulator	20/20 20/20
Bangladesh	Patient	B1 + B5	C	T	CHR1, 359133	N	ACQ59516.1	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.' Nucleotide positions are based on the reference *Vibrio cholerae* MJ-1236 (CP001485.1, CP001486.1). Patient allele frequency shows the allele frequency of the alternative (minor) allele. Ref=Reference allele; Alt=Alternative allele. NS=nonsynonymous; S=synonymous. CHR1=chromosome 1; CHR2=chromosome 2.

Gene gain and loss within and between cholera patients

To characterize variation in gene content among the 122 sequenced isolates, we analysed orthologous coding sequences from *de novo* assemblies. We defined the core genome as the genes present in all isolates, and the flexible genome as the genes absent in some of the isolates. We defined a flexible genome of 155 genes that varied in their presence or absence across genomes (Methods). Some of these flexible genes varied in presence/absence within patients, ranging from five to 103 genes depending on the patient considered (Table 2, Figure 2).

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

Patients	#genes fixed within patients	#genes variable within patients	#singletons
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 isolate, 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.

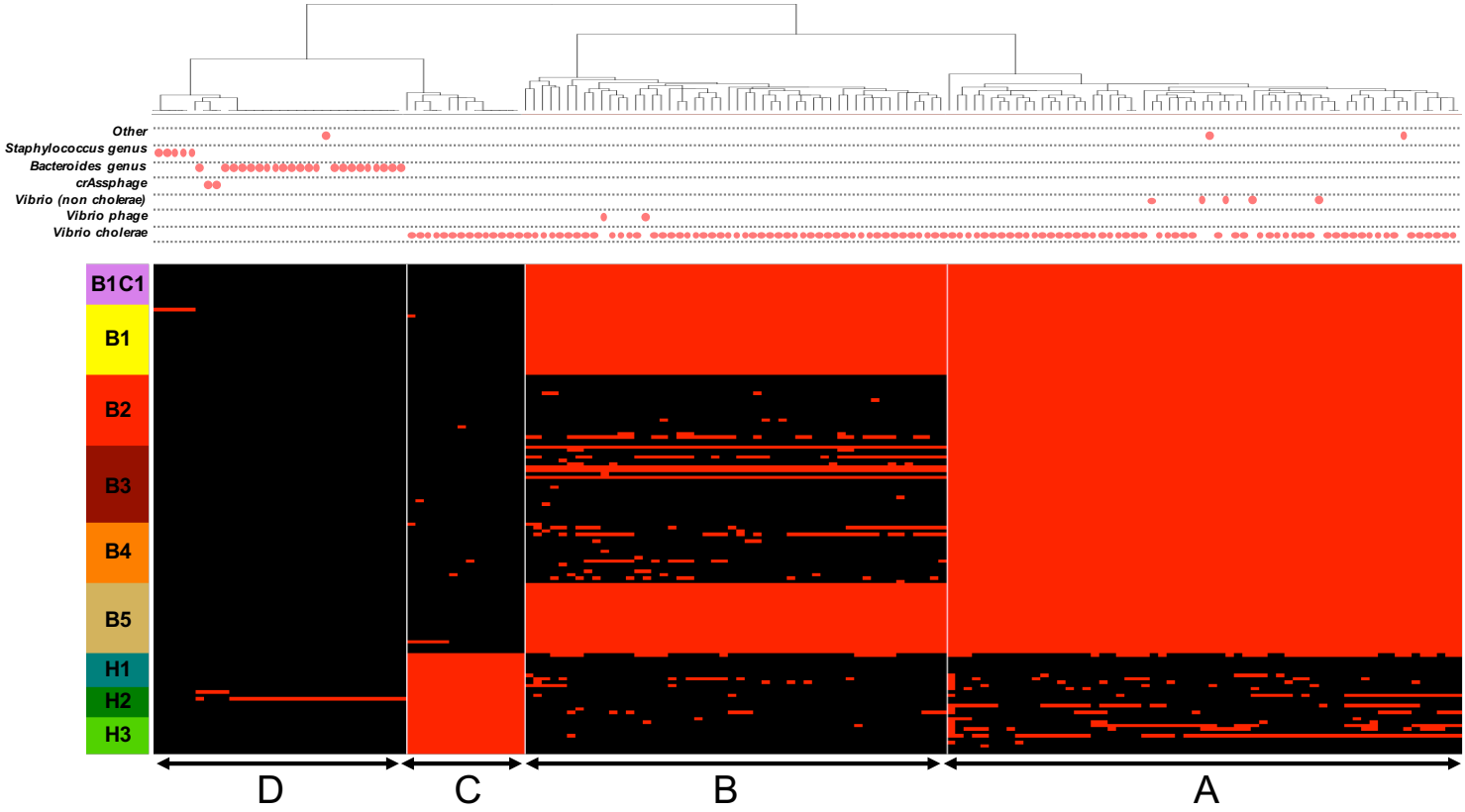
As two different methods of library construction and sequencing were used in this study, we sequenced twelve isolates using both methods. Among them, we observed variation in the detection of the flexible genome for six of the duplicate sets (Figure S5), likely because different library preparation methods may have different G+C content biases (Ross et al.

2013). 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 (Table S4). In six out of the eight patients studied, the NEBNext/HiSeq prep identified more within-patient variable genes than the Nextera/MiSeq prep. However, for the other two patients, the Nextera/MiSeq prep detected more variable genes; we therefore hesitate to draw general conclusions on which method performed best. 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 paper.

Clustering the 155 flexible genes by their presence/absence profile across patients revealed four distinct categories of genes (Figure 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 patient 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 (possibly gained/lost in a single event) corresponding to a 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 which contained less than 3% hypothetical proteins. The flexible genome also contained 10 transposases (6.5% of the flexible genome, compared to 1.4% of the core) and eight genes involved in plasmid and viral replication, all potential mechanisms of HGT (Darmon et Leach 2014). A complete list and annotation of flexible



genes is given in Table S5.

Figure 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 dots above the heatmap.

Variation in the flexible gene pool could arise from gene deletion, duplication, or HGT. To determine the extent of HGT across species boundaries, we identified the taxonomic affiliation of each flexible gene according its placement on a phylogeny of homologs from the GenBank database (Methods). While the majority (117 out of 155) of flexible genes were

assigned to *V. cholerae,* several were assigned to non-cholera vibrios or even distantly related species of *Bacteroides* or *Staphylococcus* (Figure 2). These genes had no BLAST hits to *Vibrio*, but numerous hits to *Bacteroides* or *Staphylococcus,* 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* (Figure 2). These 20 genes of putative *Bacteroides* sp. plasmid origin are among 25 singletons, present in only one isolate of patient H2. Similarly, the five singletons in patient B1 (Table 2) are all of putative *Staphylococcus* origin. Each of these 25 genes are located on a different contig where a single gene is predicted, except for two *Bacteroides* sp. genes, identified on the same contig. We are therefore unable to conclude whether the genes are integrated into a *V. cholerae* chromosome or as part of a plasmid. Aside from the genes suspected to have been transferred, no other reads from genomes of *V. cholerae* isolates mapped to *Staphylococcus* or *Bacteroides*, suggesting that putative HGT events were not due to contamination. 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 bacterial species, phages and plasmids in the gut microbiota. Owing to their low frequencies within patients, these cross-species HGTs are likely rare and recent events, which may never achieve high frequency in the *V. cholerae* population, either within or across hosts.

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 co-infection 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 (Figure 3). Assuming a constant evolutionary rate across lineages, we estimated the evolutionary rate at 7.94×10^{-7} hqSNV site^{-1} year^{-1} [95% highest posterior density (HPD), 4.89×10^{-7} to 1.14×10^{-6}] or approximately 3.3 hqSNV year^{-1} in the core genome (95% HPD), consistent with previous estimates (Mutreja et al. 2011).

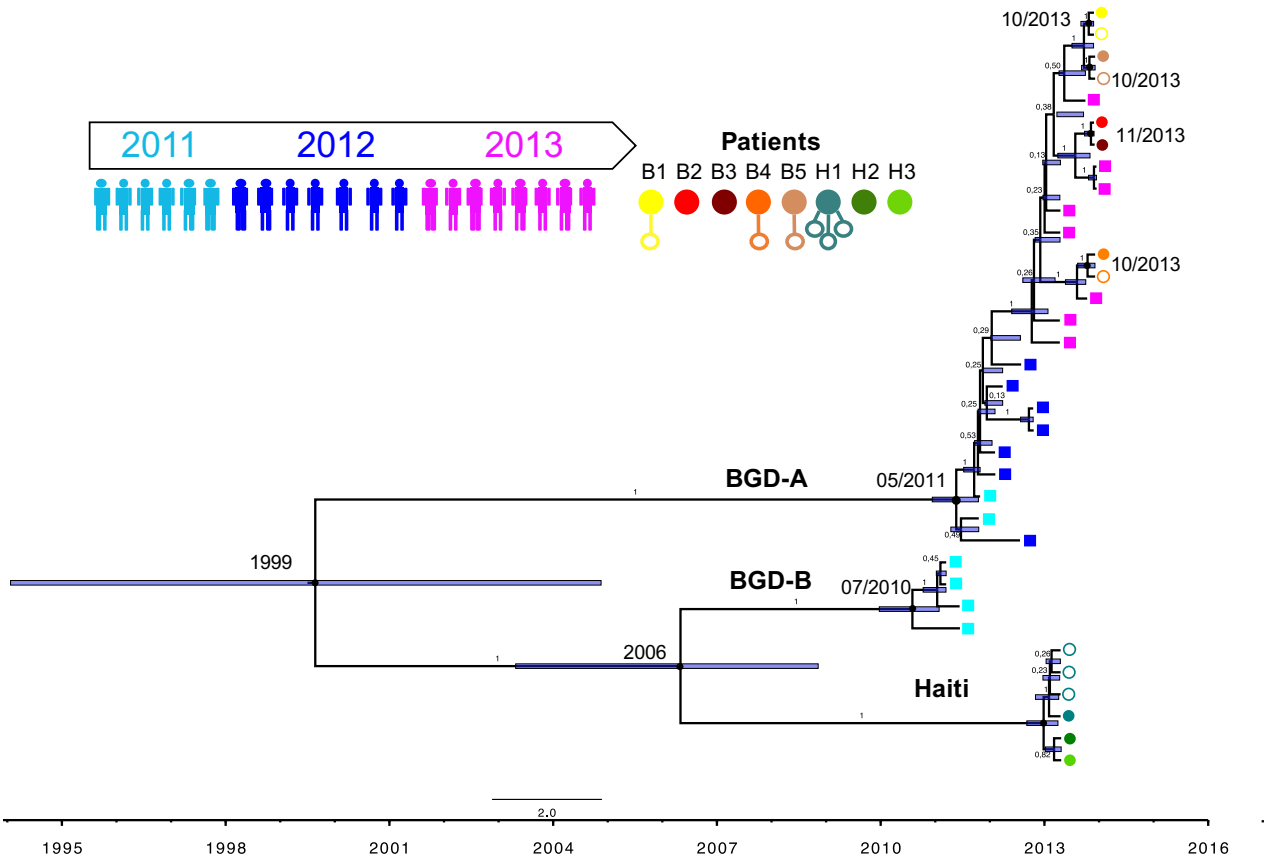


Figure 3. Bayesian phylogenic tree of 35 *V. cholerae* genotypes sampled over three years in Bangladesh and Haiti. The maximum clade credibility tree represents the genealogy of sequences in the study, reconstructed from concatenated hqSNVs, using BEAST. Colored squares (shades of blue and purple) represent 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.

We then estimated the ages of the most recent common ancestors (MRCAs) of the phylogenetic sub-lineages and clusters (Table S6). Notably, four of the six isolates from 2011 in Bangladesh were found to be closer to the Haitian isolates than to all the other Bangladeshi isolates (Figure 3). We called this sub-lineage BDG-B, and estimated the age of the MRCA of these four isolates 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 (Katz et al. 2013; Orata, Keim, et Boucher 2014). The time

of the MRCA of the isolates collected from the three Haitian patients was estimated at December 2012 (95% HPD, August 2012 to March 2013).

Based on the phylogeny, we sought to distinguish between scenarios of within-patient mutation or co-infection as causes of within-patient diversity. It is clear that isolates from the same patient always grouped together, and were never polyphyletic (Figure 3). 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 SNVs) was greater than the diversity within patients (0-3 iSNVs), which would be unlikely if cells of *V. cholerae* were sampled by patients at random from an environmental pool (Table 1). If within-patient diversity was due to co-infection 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 of stool sampling. 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 co-infection.

Signatures of natural selection on within-patient variants.

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 SNVs fixed between patients (Supplementary Note; Figure S7). 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 (McDonald et Kreitman 1991) to the eight hosts surveyed for within-host genetic variation (five from Bangladesh and three from Haiti). Despite the overall low number of SNVs 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 3).

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

	Bangladesh		Haiti	
Population	NS	S	NS	S
Polymorphic (within patient)	0	2	3	0
Fixed (between patients)	5	0	0	0
	Fisher exact test, p = 0.048		*Fisher exact test, p = 1*	

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

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 non-synonymous, which was significantly higher than expected by chance ($p < 0.01$ and $p < 0.001$, respectively; Figure 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 patient H1 compared to other patients.

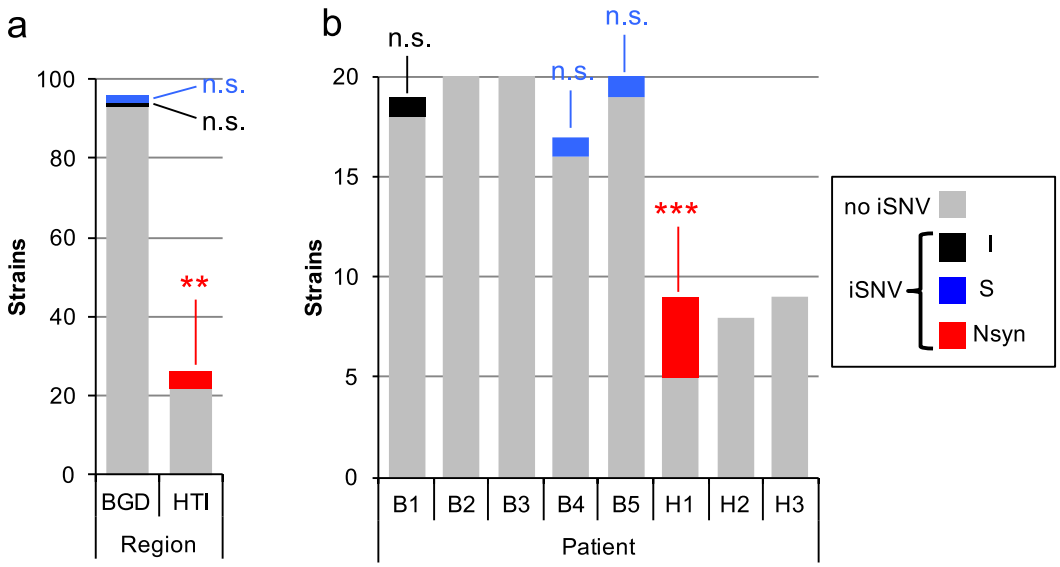


Figure 4. Significant excess of non-synonymous iSNVs in patient H1. (a) Distribution of 122 *V. cholerae* isolates containing different categories of iSNVs (I: intergenic; S: synonymous; Nsyn: non-synonymous) or no detectable iSNVs, according to geographic region (BGD: Bangladesh; HTI: Haiti). Patients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

The three NS iSNVs in patient H1 occurred in two genes. The first gene, containing one iSNV at position 1 785 021 on chromosome 1 of the MJ-1236 reference genome (Table 1) encodes a member of the tetracycline resistance (Tet^R) family of transcriptional regulators (NCBI accession number ACQ60802.1), known to be involved in the transcriptional control of multidrug efflux pumps and other pathways like quorum-sensing circuits or pathogenicity (Ramos et al. 2005). The other two NS mutations (positions 2240431 and 2241580 of chromosome 1) in patient H1 were located in the same gene (NCBI accession number ACQ61177), a sensor histidine kinase (HK) called *vprB*, which is required for resistance to the antimicrobial peptide polymyxin B (Herrera et al. 2014). 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 modify peptides that are otherwise highly conserved across species of the genus *Vibrio* (Figure S6), suggesting that these mutations may affect protein function.

Within-patient variants affect biofilm formation

We next asked whether any of the intra-host variants affected *V. cholerae* phenotypes. We focused on the non-synonymous SNVs in *vprB* which showed a signature of positive selection in patient H1, and on the plasmid of putative *Bacteroides* origin that varied in presence/absence within patient H2. First, we established that *V. cholerae* isolates (with or without plasmid, or with ancestral or derived *vprB* alleles) did not differ in growth rate in rich medium (Methods). As loss of *vprB* function has been previously associated with increased susceptibility to polymyxin B (Herrera et al. 2014), we also tested for resistance to the antibiotic polymyxin B, and again found no difference between isolates. Therefore, we chose to focus on biofilm formation, a trait which can impair intestinal colonization but might be beneficial in the aquatic environment (Almagro-Moreno, Pruss, et Taylor 2015; Shapiro et al. 2016), and is readily quantifiable. Furthermore, it is known that HKs in certain two-component systems can affect biofilm formation (Bilecen et Yildiz 2009; Bilecen et al. 2015) but the role of the *vprB* HK in particular is unknown.

We found that H1C5 and H1C6, the two isolates with derived alleles in the HK gene *vprB*, produced significantly less biofilm than the other isolates from the same patient (Figure 5a). Based on our hqSNV 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 (Figure S5). However, this gene content variation did not measurably affect biofilm formation (Figure 5a). In contrast to the differences in biofilm formation between *vprB* alleles, we did not detect any significant difference in biofilm formation between isolates from patient H1 with the ancestral iSNV allele (in isolate H1C1) or the derived allele in the transcriptional regulator gene (isolate H1C4). In summary, within-patient mutations in *vprB*, but not another mutation within the same patient, significantly reduced the ability of *V. cholerae* to form biofilms.

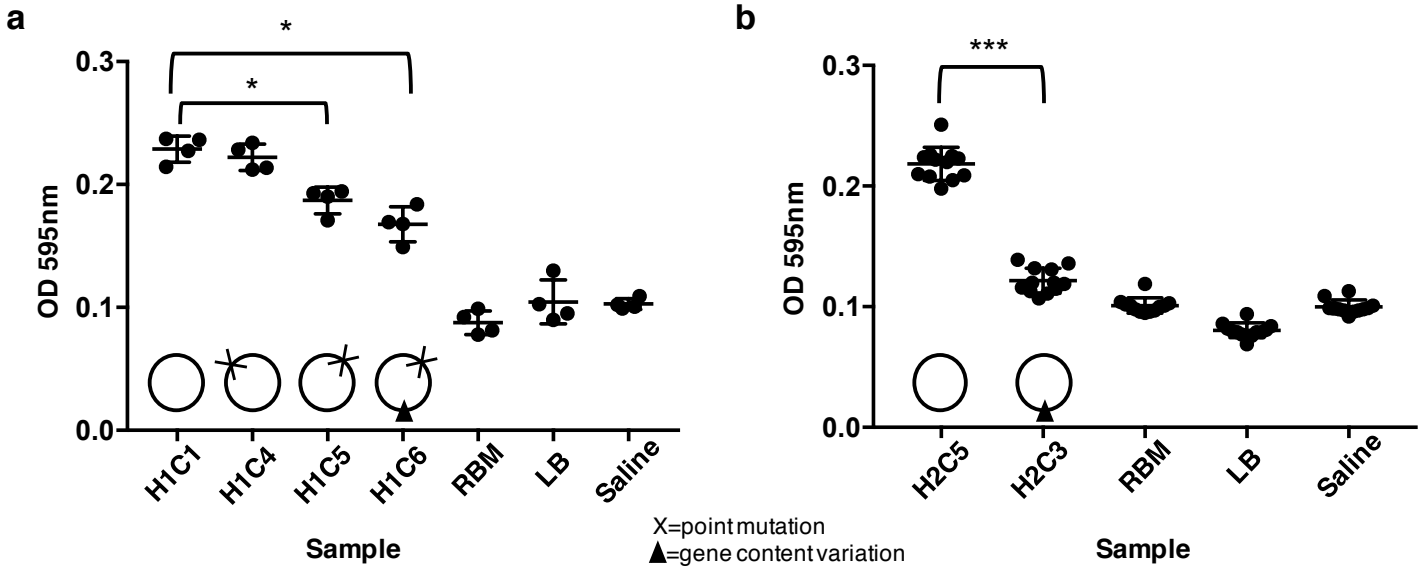


Figure 5. Biofilm formation of isolates from patients H1 and H2. Optical density was measured for four to 12 replicates of each isolate, after 48h of growth at 30°C. Statistical comparisons were made using a non-parametric Mann-Whitney test (*$p < 0.05$, ***$p < 0.0001$). Circles represent genomes with either variation in gene content (dark triangle) or iSNV variation (cross). (**a**) Isolates 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**) Isolates 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.

In the case of patient H2, we found that the presence of a plasmid of putative *Bacteroides* origin reduces biofilm formation even more strongly than point mutations in *vprB*. Specifically, the *Bacteroides* sp. plasmid-containing isolate (H2C3) produces approximately two-fold less biofilm than an isogenic control from the same patient H2C5 (Mann-Whitney test, $p < 0.0001$). The biofilm formation of H2C3 was indistinguishable from negative controls (Figure 5b). Together, these results show that both point mutations and 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 analysed 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 harbour more than 20 iSNVs per patient (Morelli et al. 2010; Price et al. 2013; Eldholm et al. 2014). 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 *V. 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 *Bacteroides* sp. plasmid both reduced the ability of *V. cholerae* to form biofilms, which could be advantageous during host colonization (Herrera et al. 2014; Almagro-Moreno, Pruss, et Taylor 2015).

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 (Borgeaud et al. 2015), transduction (Faruque et Mekalanos 2014) and conjugation (Folster et al. 2014; Hazen et al. 2010). HGT contributes substantially to drug resistance, pathogenicity and adaptation to different environments, via the acquisition of genomic islands, phages, transposons, ICEs and plasmids (Chun et al. 2009; Das et al. 2016). Our characterization of the flexible genome within patients used read mapping to confirm gene absences, reducing false-positive inference of gene content variation (Figs S2 and S3). We detected between five and 103 genes that varied in presence/absence within patients (Figure 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 twofold reduction in biofilm formation (Figure 5). Among these 20 genes, we identified an antibiotic resistance gene, a haloacid dehalogenase protein that could impact pathogenicity (Tribble et al. 2006), and a FtsY recognition signal protein that was shown to increase virulence in *Streptococcus* (Rosch et al. 2008). Among the genes likely acquired from a *Staphylococcus* donor in isolate B1C2, three could potentially be involved in modulation of virulence (Figure S5, Table S6). Firstly, a GNAT family acetyltransferase could promote virulence or increase antibiotic resistance (Rogers et al. 2007; Favrot, Blanchard, et Vergnolle 2016). Secondly, a putative phosphoenolpyruvate phosphotransferase has been demonstrated to play a role in biofilm formation in *Vibrio* (Houot et Watnick 2008; Houot, Chang, Pickering, et al. 2010; Houot, Chang, Absalon, et al. 2010). Finally, the KdpC gene (a potassium-transporting ATPase) has been shown to modulate virulence in *Mycobacterium paratuberculosis* (Shin et al. 2006).

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 (Figure 2), respectively corresponding to the ICE and Kappa phage. These elements are known to vary among *V. cholerae* genomes sampled over larger temporal and geographic scales (Grim et al. 2010), 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 (Figure 2). Category D genes are most easily explained by cross-species HGT, as previously documented by Folster and colleagues, who identified a Haitian *V. cholerae* isolate that gained multidrug resistance through transfer of a plasmid from a species of *Enterobacteriaceae* (Folster et al. 2014). Although to our knowledge, ours is the first description of HGT specifically between *V. cholerae* and *Bacteroides*, the genus *Bacteroides* is known to be involved in inter-species and inter-genus HGT in the human gut (Shoemaker et al. 2001; Huddleston 2014). Taken together, these results are consistent with the human gut being a hotspot of HGT (Smillie et al. 2011), sometimes involving pathogens like *V. cholerae.* Although we hesitate to speculate on the eventual fate of within-

patient gene gain/loss events, it appears that certain events (e.g. in the ICE or kappa phage) persist long enough to be observed as fixed differences between patients. The rare cross-species HGT events we observed were at low frequency (present in just one isolate per patient), suggesting they are neutral or slightly deleterious variants that will never attain high frequency – although their eventual fate is unknown.

Regimes of natural selection inferred from within-patient point mutations

The low levels of variation (0-3 iSNVs per patient) 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 SNVs and gene gain/loss events that yielded zero variation among control isolates, suggesting low rates of false-positive variant calls and increasing confidence that the six total iSNVs (Figure 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 co-infections (particularly of low-frequency strains not detectable by sequencing 8-20 isolates), our results are more consistent with *de novo* mutation within hosts. Isolates from the same patient were grouped together on the phylogeny (Figure 3), suggesting a recent clonal ancestor. This result is consistent with previous findings that cholera outbreaks are highly clonal (Rafique et al. 2016).

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 (Dillon et al. 2017). 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 (Abel et al. 2015). Our estimates of N_e based on iSNVs

give values of 0 to ~10^2, consistent with population bottlenecks or selective sweeps purging diversity (Table S1). The infectious dose of *V. cholerae* has been estimated to be 10^3-10^8 cells (Harris et al. 2012). Our low estimates of N_e indicate that (i) this infectious dose is genetically homogeneous, or that (ii) any pre-existing diversity is quickly purged by a bottleneck or selective sweep.

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 Note). However, we identified an excess of NS mutations in one Haitian patient (H1), suggesting positive or diversifying selection on *V. cholerae* within this patient (Figure 4). Patient H1, like all patients in this study, suffered from severe acute cholera, and we do not have access to further information about this patient which might explain the excess of NS mutations. 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 (Herrera et al. 2014). Another two-component system, CarRS, is known to confer polymyxin B resistance, but also to negatively regulate biofilm formation (Bilecen et Yildiz 2009; Bilecen et al. 2015). 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, but did reduce the ability of *V. cholerae* to form biofilms (Figure 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 (Almagro-Moreno, Pruss, et Taylor 2015; Shapiro et al. 2016). 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 (Boucher, Orata, et Alam 2015), 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 (Azarian et al. 2014). In Bangladesh, we observed an excess of NS mutations between patients (Table 3), suggesting positive selection on protein sequences between patients, or efficient purifying selection purging NS 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

We have shown that small but measurable changes occur in the *V. cholerae* genome during human infection. Changes in flexible gene content appear to accumulate more quickly than point mutations, although point mutations may also 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 – particularly in the ICE and mobile genetic elements – in the evolution of antibiotic resistance, host adaptation, and the severity of disease in infected patients.

Data bibliography

1. Levade *et al.* Sequence Reads Archive, SRP116359 (2017).
2. Levade *et al.* Github. 2017. https://github.com/ilevade/Vibrio_cholerae_within_patient_assemblies.

Funding information

This study was supported by CIHR (Canadian Institutes of Health Research) and the Canada Research Chairs program; the icddr,b: Centre for Health and Population Research; grants AI099243, AI103055, AI106878, AI058935, T32A1070611976 and K08AI123494 from the National Institutes of Health; and the Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program.

Acknowledgements

We thank Simone Perinet and Meri Debela for their technical assistance, Paula Watnick who contributed materials to the study, and Salvador Almagro-Moreno for constructive comments on the manuscript. We also thank the Zanmi Lasante staff at the enteric microbiology laboratory in St. Marc, Haiti. Finally, we are grateful to the people of Dhaka where our study was undertaken; to the field, laboratory and data management staff who provided 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).

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.

Supplementary data

Supplementary materials and methods

Enrollment and sample processing.

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. Diarrheal samples were then 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 (F. Qadri et al. 1994). 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.

Genome assembly

To exclude low-quality sequences, we filtered raw reads with Trimmomatic (Bolger, Lohse, et Usadel 2014). 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 isolate using IDBA_ud v1.1.1 (Peng et al. 2012).

Filtered reads were also mapped with Bowtie2 v2.2.5 (Langmead et Salzberg 2012) 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 (H. Li et al. 2009), 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 (Rissman et al. 2009) with default parameters.

Variant calling and annotation

We used SAMtools v1.3 and BCFtools v1.1 to call SNVs and indels from mapping, requiring a minimum mapping quality of 30 and a minimum base quality of 20. Resulting SNVs 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 (Garrison et Marth 2012). For each isolate, 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 isolate B1C1, to control for possible SNVs 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 SNVs as potential sequencing, mapping or assembly errors. Using these filters, we generated a list of high-quality SNVs (hqSNVs). From this list, we identified intra-host single nucleotide variants (iSNVs) as SNVs 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.

Characterization of the flexible and core genomes

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. 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 (Figure S2). A representative catalogue of the flexible genome protein sequences was built using the cd-hit program with a 90% similarity threshold (L. Fu et al. 2012). 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 (Figure S3) while observing no variability among the control isolates. 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.

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.

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 (Edgar 2004) and gene trees were constructed with FastTree (M. N. Price, Dehal, et Arkin 2009) using default parameters (Maximum likelihood with Jukes-Cantor model). We screened each gene tree 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* species); 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.

Molecular clock and demographic model comparison

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 six 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 (Baele et al. 2012). Divergence time, substitution rates and resulting tree were reported from the models with the highest marginal likelihood.

Tests for natural selection between patients

To distinguish between positive selection, purifying selection, or neutral evolution of protein-coding sequences, we considered variation in the proportion of non-synonymous hqSNVs (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 hqSNVs (excluding the ICE region, a mutation hotspot; Figure S1) 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 hqSNVs). 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 hqSNVs). 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 hqSNVs 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 three 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 (Tamura et al. 2011) (Figure S7b). We distinguished hqSNVs 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 hqSNVs 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 patients

To investigate the role of natural selection within *versus* between patients from Bangladesh and Haiti, we performed the McDonald-Kreitman test (McDonald et Kreitman 1991) 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 SNVs) 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 isolates collected from the eight patients to one of the four following haplotypes: H_0 as isolates having the most frequent haplotype found within each patient, and

assumed to be ancestral for that patient; H_{INT} as isolates having one intergenic iSNV; H_S as isolates having one synonymous iSNV and H_{NS} as isolates having one synonymous iSNV. (No isolates 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 isolates. When an iSNV was shared between two or more isolates within a patient, but not observed in other patients, we ensured that these isolates 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 isolates containing iSNVs / the total number of isolates 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 detected 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.

Supplementary results

Natural selection is not detectable in *V. cholerae* core genome over a three-year period.

The constant molecular clock of the *V. cholerae* core genome (Table S3) suggests that the substitution rate was constant over our three-year survey. This constant substitution rate may hide more complex evolutionary processes, like natural selection on protein-coding sequences occurring before and during host infection, which could result in an overall excess (positive selection) or a deficit (purifying selection) of non-synonymous (NS) mutations. We considered only SNVs fixed among hosts. Among the 136 hqSNVs identified in the core genome (iSNVs excluded), 88 were identified as NS mutations (percentage NS, p_{NS} = 64.7%), based on the annotation of the MJ1236 reference genome. We then asked whether these mutations represent a random subset of expected mutations in the MJ1236 reference genome under neutral evolution. To test this, we simulated 1,000 sets of 136 mutations distributed randomly along the MJ1236 core genome and, for each of these sets, calculated the percentage of NS mutations. We took the mean p_{NS} across these simulations as our estimate of the expected p_{NS} under neutral evolution (see Materials and Methods). We observed no significant difference between the observed (64.7%) and the expected p_{NS} (66.4 ± 4.1%, $p > 0.05$; Figure S7b). We observed significantly higher GC content at mutated positions (72.8%) and higher transition rate (75%) than expected by chance (47.2% and 33.3%, respectively; $p < 0.001$; 1,000 simulations). However, the difference between the observed and expected p_{NS} remained non-significant when incorporating the observed bias in GC content into the simulation (expected p_{NS} = 66.8 ± 3.9%, $p > 0.05$). Similarly, the observed p_{NS} was not significantly different than expected when the observed bias in transition rate (expected p_{NS} = 60.3 ± 4.2%, $p > 0.05$), or both GC and transition biases together were incorporated into the model (expected p_{NS} = 60.9 ± 4.2%, $p > 0.05$; Figure S7a). According to these results, we could not reject the hypothesis that protein-coding sequences in the *V. cholerae* core genome evolved under a neutral regime over a three-year period.

We next asked whether fixed core genome SNVs were distributed evenly across branches of the phylogeny. Specifically, we asked if 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 address these questions, we first defined three well-separated monophyletic clades based on the evolutionary tree of

the *V. cholerae* core genome (Figure 3), and corresponding to isolates from Haiti, isolates from Bangladesh sampled before October 2011 (BGD-B) and isolates from Bangladesh sampled after October 2011 (BGD-A). We distinguished hqSNVs that were fixed among clades (corresponding to long branches) from those that are variable within clades, the latter category corresponding to more recent mutations event (Figure 3). If differences in p_{NS} are observed among these clades, this could suggest that selection or substitution rates vary among clades and over time. To test this, we permuted SNVs among branches of the evolutionary tree and recalculated p_{NS} within and between clades. We observed no significant difference in p_{NS} among clades ($p > 0.05$; 10,000 permutations; Figure S7b), indicating that there is no evidence for differential selection on protein sequences among clades or over time.

Supplementary tables

Table S1. Sequencing data

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5761273/bin/mgen-3-142-s002.xlsx

Table S2. Estimation of effective population size (N_e) within each patient based on iSNVs numbers and frequencies.

Patients	#iSNVs	Patient allele frequency	Ne(S)	Ne(π)
B2 – B3 – H1 – H2	0	0	~0	~0
B1	1	0	44.12	16.5
B4	1	1/19	45	20
B5	1	1/20	43.3	15
H1	2	1/9 2/9	110.3	100

Table S3. Model comparison using path sampling to compute marginal likelihood estimations (MLE).

	Path-sampling log MLE	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Model 1	-5722274,221	-	10.81	-0.47	-1027.76	-286.81	-742.66
Model 2	-5722285,028	-10.81	-	-11.28	-1038.57	-297.62	-753.47
Model 3	-5722273,752	0.47	11.28	-	-1027.29	-286.34	-742.20
Model 4	-5721246,462	1027.76	1038.57	-1027.29	-	740.95	285.09
Model 5	-5721987,408	286.81	297.62	-286.34	-740.95	-	-455.85
Model 6	-5721531,555	742.66	753.47	-742.20	-285.09	455.85	-

Higher MLE values indicate better model fit. Bayes factors are reported, with positive values indicating better relative model fit of the row's model compared with the column's model. Model 1: Relaxed molecular clock, Bayesian skyline plot. Model 2: Relaxed molecular clock, Constant population size. Model 3: Relaxed molecular clock, Exponential population size. Model 4: Strict molecular clock, Bayesian skyline plot. Model 5: Strict molecular clock, Constant population size. Model 6: Strict molecular clock, Exponential population size.

Table S4. Flexible gene content variation within and between patients, according to sequencing methods.

Patients	#genes fixed within patients		#genes variable within patients	
	Nextera/MiSeq	NEBNext/HiSeq	Nextera/MiSeq	NEBNext/HiSeq
B1	111	111	0	6
B2	61	61	4	33
B3	61	61	50	51
B4	61	61	38	30
B5	111	111	0	5
H1	14	15	0	67
H2	14	14	62	31
H3	14	14	30	58

Table S5. Orthologous groups of the flexible genome

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5761273/bin/mgen-3-142-s003.xlsx

Table S6. Estimated divergence dates of *V. cholerae* sub-lineages and clusters.

Phylogenetic sub-lineages and clusters	Estimated divergence dates (day-month-year)	95% HPD interval	
BGD-A	17-05-2011	26-11-2010	05-10-2011
BGD-B	30-07-2010	09-12-2009	28-01-2011
Haiti - BDG-B	17-04-2006	24-04-2003	16-11-2008
Haiti	23-12-2012	03-09-2012	30-03-2013
Patient B1	22-10-2013	28-08-2013	26-11-2013
Patient B4	16-10-2013	11-08-2013	03-12-2013
Patient B5	27-10-2013	02-09-2013	04-01-2013
Patient H1	30-01-2013	05-11-2012	08-04-2013
Patient B2 - B3	07-11-2013	25-10-2013	28-11-2013

Supplementary Figures

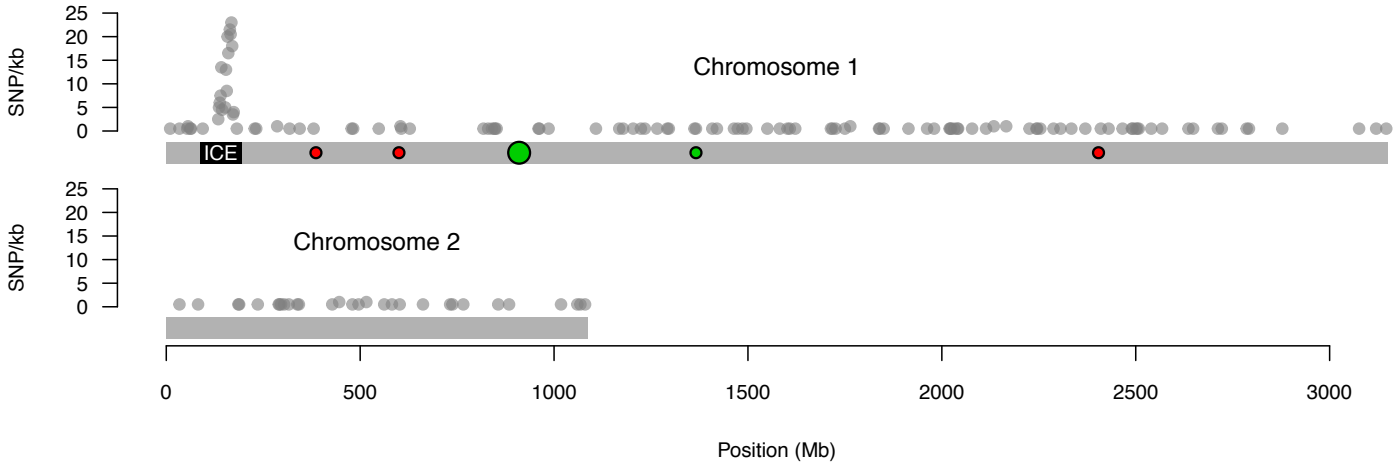


Figure S1. hqSNV distribution across the *V. cholerae* genome. The x-axis indicates the position in the MJ1236 reference genome (reverse complement). The height of grey dots on the y-axis indicate the number of hqSNVs per 1kb discrete window, which is higher in the Integrative Conjugative Element (ICE; black region). Coloured circles indicate iSNVs identified in Bangladeshi (red) and Haitian (green) isolates. Circle diameter is proportional to iSNV frequency per 2kb window (ranging from 1 to 2 iSNVs per 2kb).

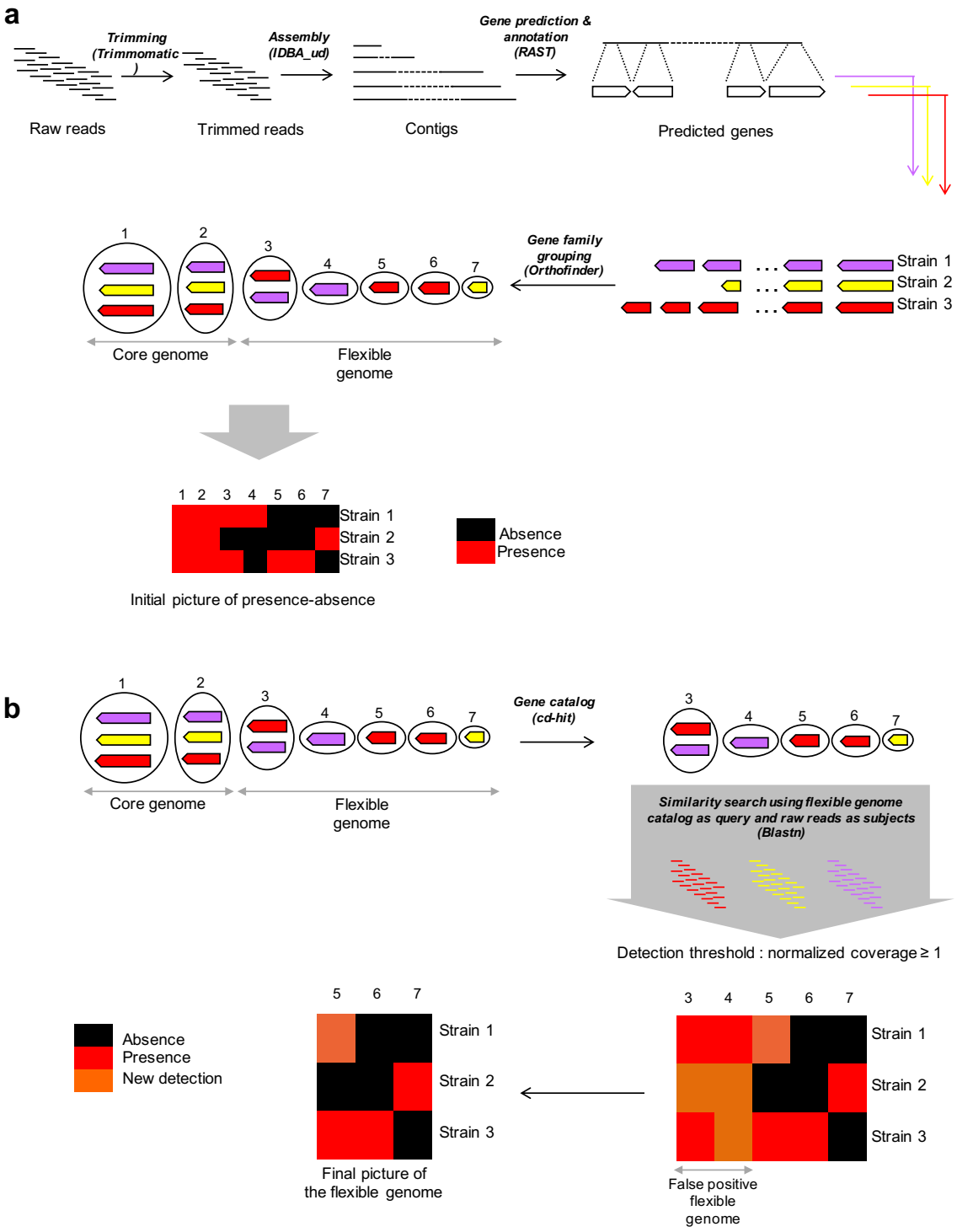


Figure S2. Flexible genome analysis pipeline. (a) From sequence reads to primary flexible genome analysis. (b) From a flexible gene catalogue to a final picture of the flexible genome.

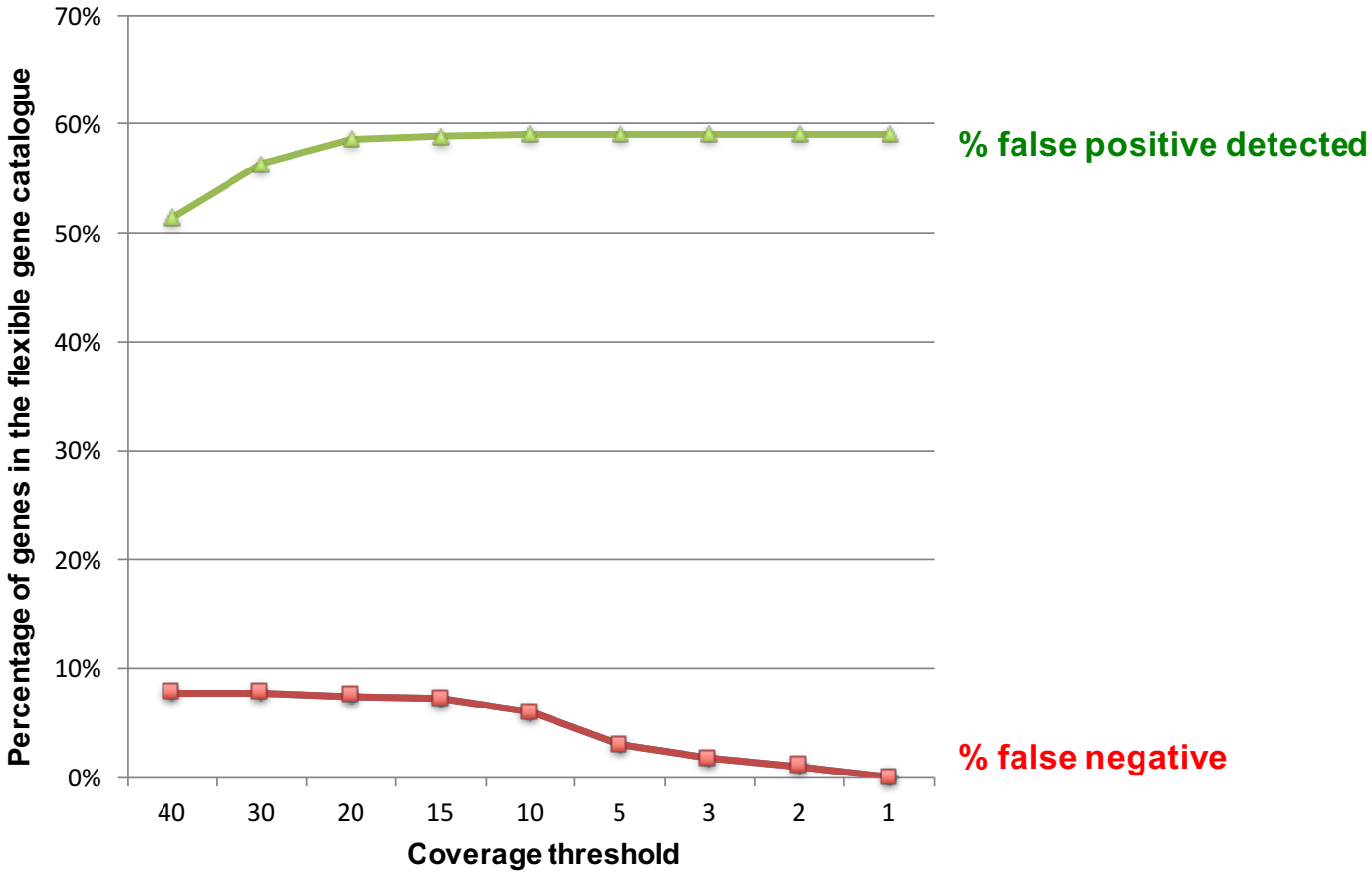


Figure S3. Coverage filters used to infer presence-absence of 401 putative flexible genes. False positive flexible genes (green) are detected in all isolates (at a given coverage threshold), and are therefore part of the core, not the flexible genome. False negatives are genes known to be present in at least one genome (because they are part of the gene catalogue; Figure S2) but are not identified in any genome at a given coverage threshold. Coverage is defined as the average coverage of a gene by sequence reads (Methods). Based on this analysis, we used a coverage filter of 1X to remove false positives, without suffering from false negatives.

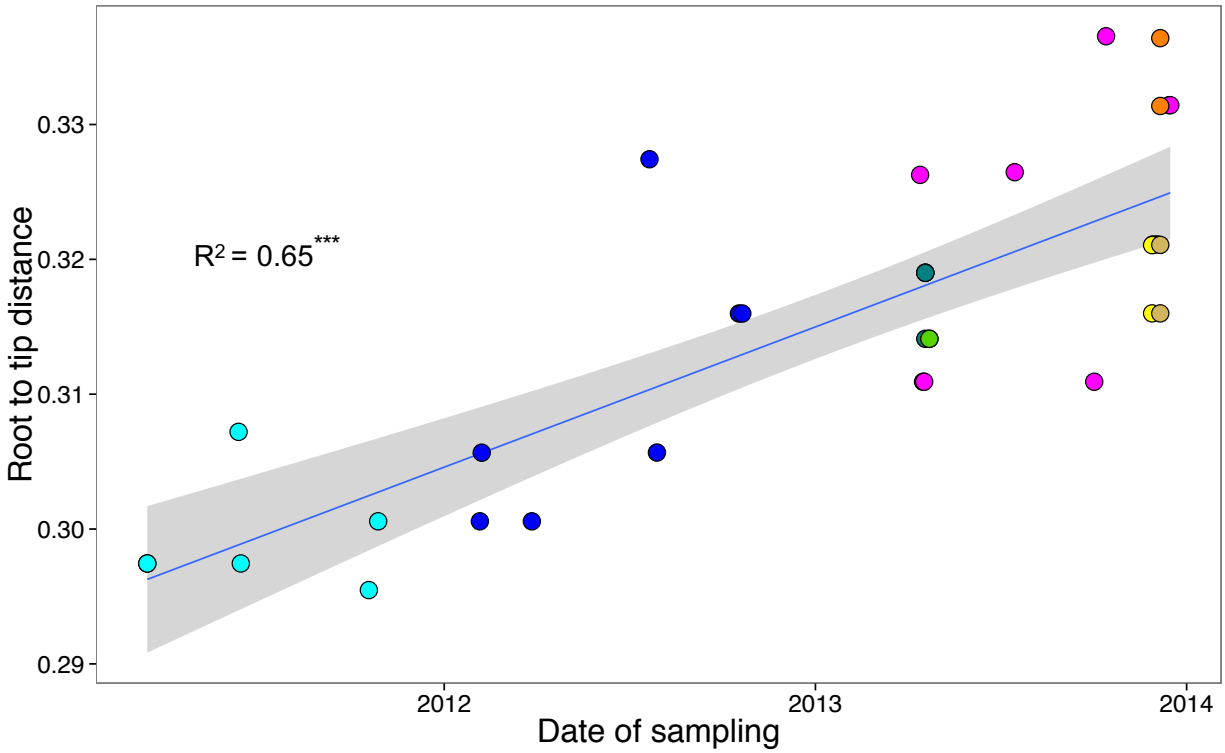
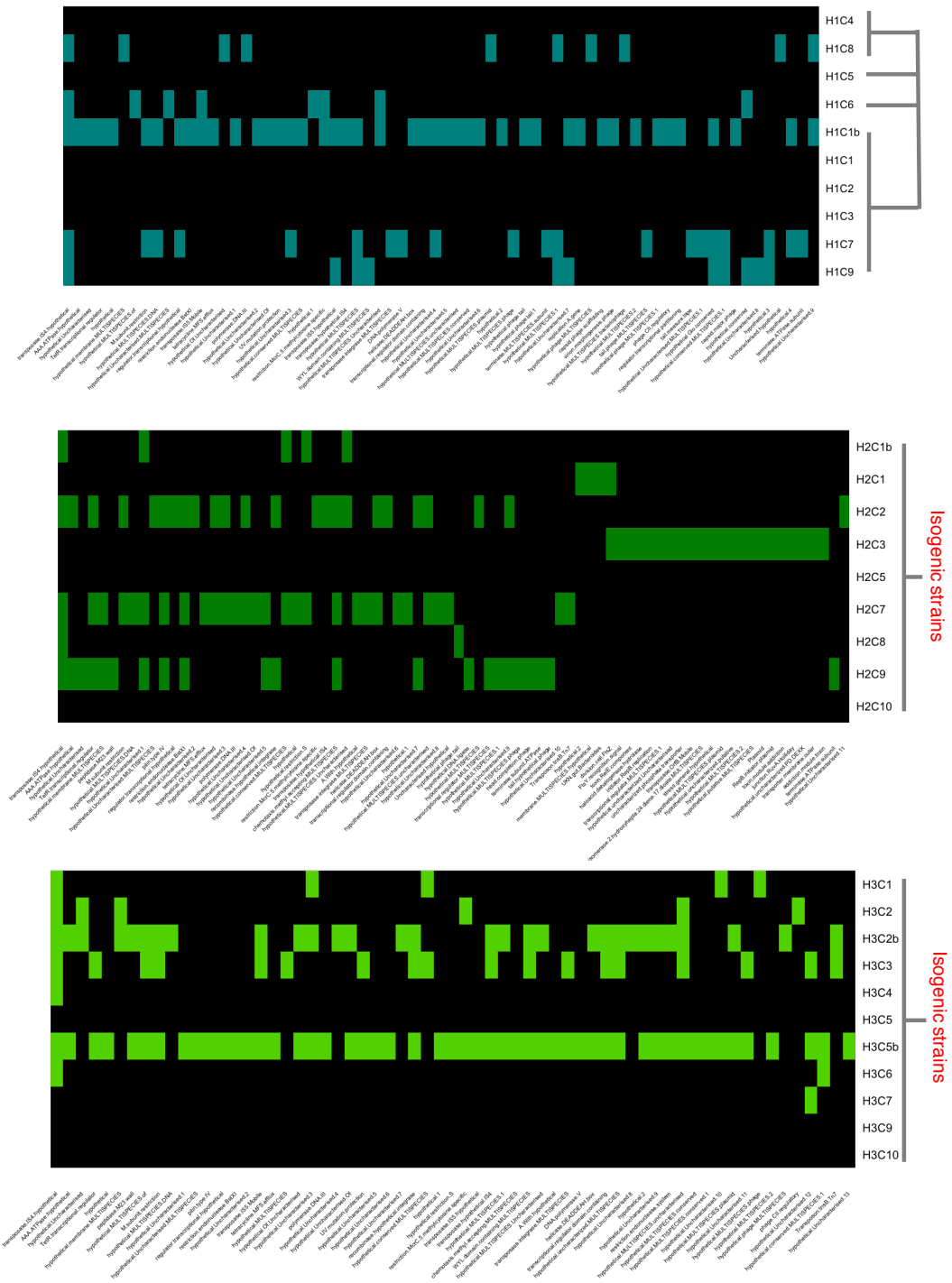
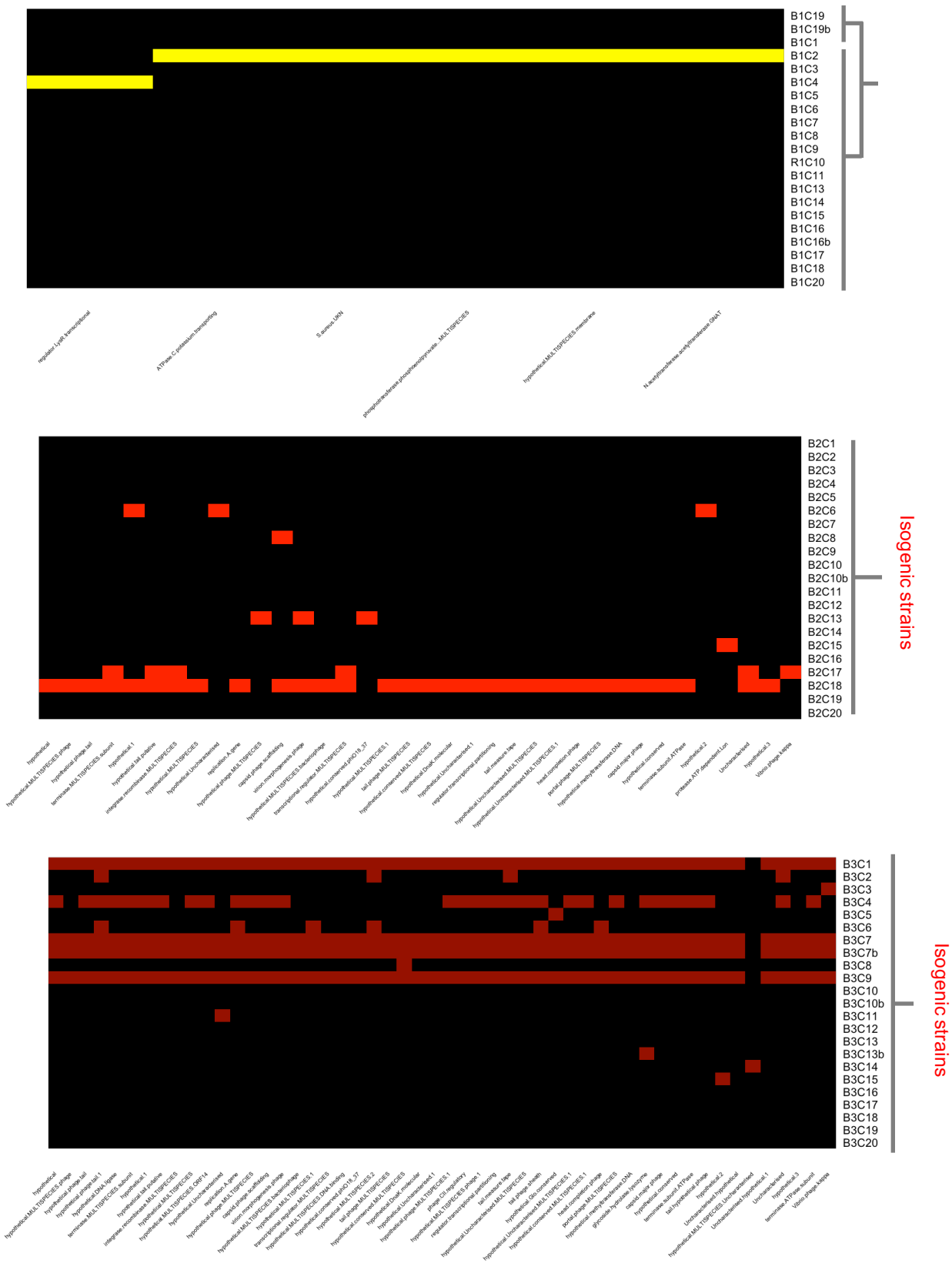


Figure S4. Positive significant correlation between root-to-tip distance and year of sampling. Regression of the root-to-tip genetic distance as a function of sampling time (year) for 35 *V. cholerae* isolates sampled from 2011 to 2013. Each point corresponds to a genotype, the blue dashed line represents the linear regression and in grey its 95% confidence interval. *** indicates linear regression p-value < 0.0001.


H1C4
H1C8
H1C5
H1C6
H1C1b
H1C1
H1C2
H1C3
H1C7
H1C9
H2C1b
H2C1
H2C2
H2C3
H2C5
H2C7
H2C8
H2C9
H2C10
Isogenic strains
H3C1
H3C2
H3C2b
H3C3
H3C4
H3C5
H3C5b
H3C6
H3C7
H3C9
H3C10
Isogenic strains


B1C19
B1C19b
B1C1
B1C2
B1C3
B1C4
B1C5
B1C6
B1C7
B1C8
B1C9
R1C10
B1C11
B1C13
B1C14
B1C15
B1C16
B1C16b
B1C17
B1C18
B1C20
regulator, LysR transcriptional
ATPase, C-potassium transporting
S.aureus, UKN
phosphoesterase,phosphoesterphrase , MULTISPECIES
hypothetical, MULTISPECIES, membrane
N-acetyltransferase, acetyltransferase, GNAT
B2C1
B2C2
B2C3
B2C4
B2C5
B2C6
B2C7
B2C8
B2C9
B2C10
B2C10b
B2C11
B2C12
B2C13
B2C14
B2C15
B2C16
B2C17
B2C18
B2C19
B2C20
Isogenic strains
hypothetical
hypothetical, MULTISPECIES, phage
hypothetical, phage, tail
terminase, MULTISPECIES, subunit
hypothetical, 1
hypothetical, tail, putative
integrase, recombinase, MULTISPECIES
hypothetical, MULTISPECIES
hypothetical, Uncharacterised
replication, A, gene
hypothetical, phage, MULTISPECIES
capsid, phage, scaffolding
virion, morphogenesis, phage
hypothetical, MULTISPECIES, bacteriophage
transcriptional, regulator, MULTISPECIES
hypothetical, conserved, phiO18_37
hypothetical, MULTISPECIES, 1
tail, phage, MULTISPECIES
hypothetical, conserved, MULTISPECIES
hypothetical, DnaK, molecular
hypothetical, Uncharacterised, 1
regulator, transcriptional, partitioning
tail, measure, tape
hypothetical, Uncharacterised, MULTISPECIES
hypothetical, Uncharacterised, MULTISPECIES, 1
head, completion, phage
portal, phage, MULTISPECIES
hypothetical, methyltransferase, N4
capsid, major, phage
hypothetical, conserved
terminase, subunit, ATPase
hypothetical, 2
protease, ATP-dependent, Lon
Uncharacterised
hypothetical, 3
Vibrio, phage, kappa
B3C1
B3C2
B3C3
B3C4
B3C5
B3C6
B3C7
B3C7b
B3C8
B3C9
B3C10
B3C10b
B3C11
B3C12
B3C13
B3C13b
B3C14
B3C15
B3C16
B3C17
B3C18
B3C19
B3C20
Isogenic strains
hypothetical
hypothetical, MULTISPECIES, phage
hypothetical, phage, tail
hypothetical, phage, tail, 1
hypothetical, DNA, ligase
terminase, MULTISPECIES, subunit
hypothetical, 1
hypothetical, tail, putative
integrase, recombinase, MULTISPECIES
hypothetical, MULTISPECIES
hypothetical, MULTISPECIES, GBF14
hypothetical, Uncharacterised
replication, A, gene
hypothetical, phage, MULTISPECIES
capsid, phage, scaffolding
virion, morphogenesis, phage
hypothetical, MULTISPECIES, bacteriophage
transcriptional, regulator, MULTISPECIES
hypothetical, MULTISPECIES, 1
hypothetical, conserved, DNA, binding
hypothetical, conserved, phiO18_37
tail, phage, MULTISPECIES, 2
hypothetical, conserved, MULTISPECIES
hypothetical, DnaK, molecular
hypothetical, Uncharacterised, 1
hypothetical, phage, MULTISPECIES, 1
phage, CII, regulatory
regulator, transcriptional, partitioning, phage, 1
tail, measure, tape
hypothetical, MULTISPECIES
tail, phage, sheath
hypothetical, Gp, conserved
hypothetical, Uncharacterised, MULTISPECIES, 1
hypothetical, conserved, MULTISPECIES, 1
head, completion, phage
portal, phage, MULTISPECIES
hypothetical, methyltransferase, DNA
glycoside, hydrolase, lysozyme
capsid, major, phage
hypothetical, conserved
terminase, subunit, ATPase
tail, hypothetical, phage
hypothetical, 2
Uncharacterised, hypothetical
hypothetical, MULTISPECIES, Uncharacterised
Uncharacterised, hypothetical, 1
Uncharacterised
hypothetical, 3
terminase, ATPase, subunit
Vibrio, phage, kappa

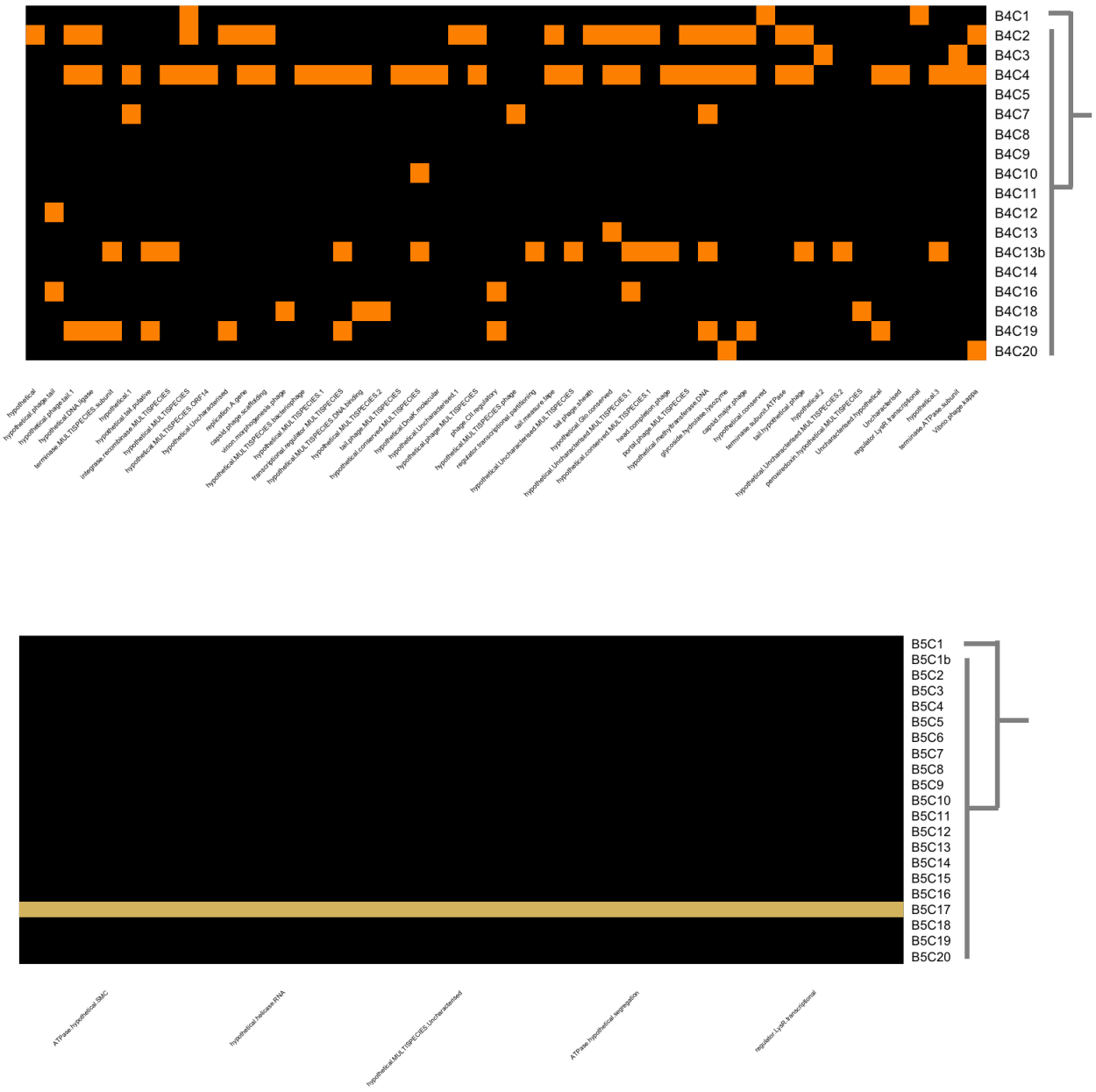


Figure S5. Detailed presence/absence profile of gene families for each patient. The flexible genes within each patient are shown, with each patient in a separate panel. Only genes that vary in presence/absence among isolates (y-axis) from a given patient are shown (x-axis) Color denotes gene presence; black denotes absence. Duplicate isolates sequenced with two different procedures (Nextera/Miseq or NEB/HiSeq) are shown with a "b" (indicating the NEB/HiSeq procedure). All other isolates were sequenced with only one procedure (Table S5). Gene annotations are provided in Table S4.

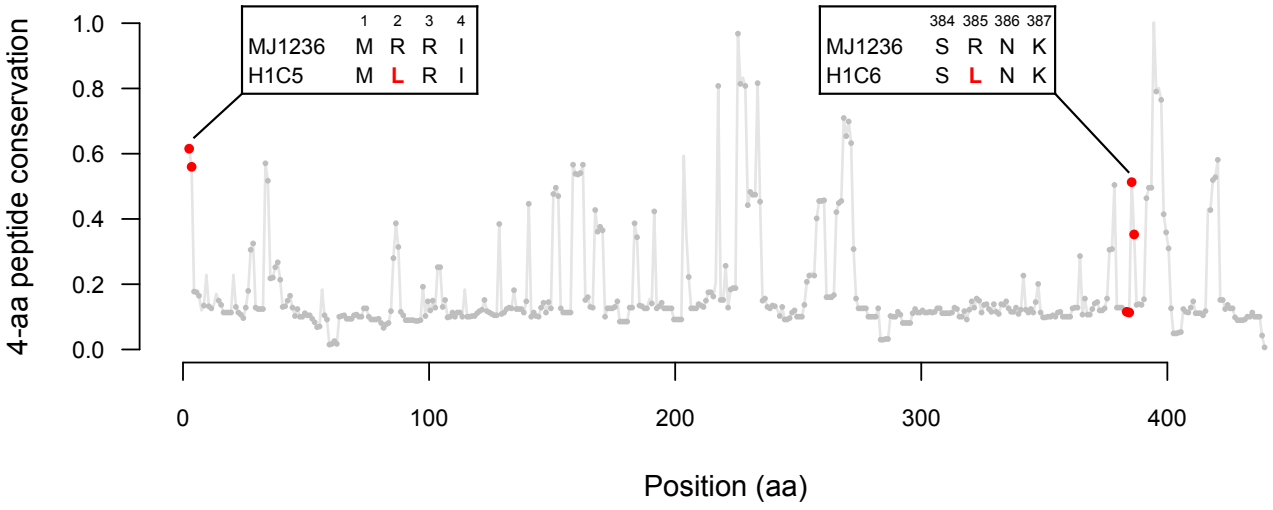


Figure S6. Non-synonymous iSNVs in a sensor histidine kinase in patient H1 affect highly conserved peptides. Distribution of peptide conservation scores (grey line; 4-amino-acid sliding windows) across the ACQ61177 protein sequence based on the comparison of 465 homologous proteins. The conservation score is defined as the proportion of homologs having the reference peptide (from *V. cholerae* MJ-1236). Grey dots indicate peptides that are not redundant across a single sequence. Red dots indicate peptides that are affected by NS iSNVs in patient H1. For each iSNV, details of the amino-acid reference (MJ1236) and mutated (H1C5 and H1C6) sequences are given for peptide with the highest conservation score (boxes).

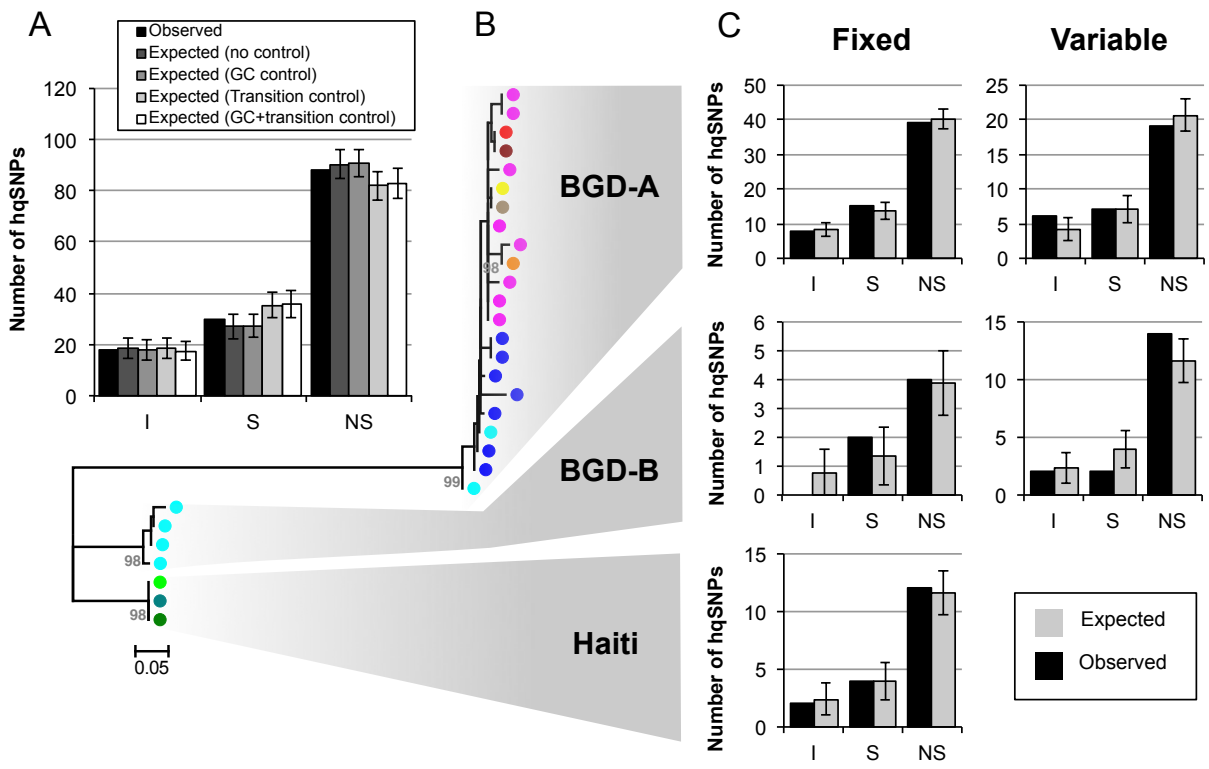


Figure S7. Neutral evolution of protein-coding sequences in the *V. cholerae* core genome over a three-year period. These analyses are based on 136 hqSNVs identified among 29 isolates (iSNVs excluded, and excluding SNVs in the ICE). (A) The numbers of intergenic (I), synonymous (S) and non-synonymous (NS) mutations observed among the 136 hqSNVs (black bars) are no different than numbers expected after random mutation of the MJ1236 reference genome, simulated under four different models (see legend on the top and Methods for details of simulations). For each model, bars represent the expected average values and error bars show standard deviations calculated over 1,000 simulations. (B) Three well-separated clades (BGD-A, BGD-B and Haiti) are supported by an evolutionary tree based on 136 hqSNVs. The scale represents the number of single nucleotide substitutions per site. Nodes supported by bootstrap values ≥98% are indicated in grey (Maximum composite likelihood, bootstrap test, 1,000 replicates). (C) The observed numbers (black bars) of I, S, and NS mutations fixed (between clades) and variable (within clades) are not significantly different than expected by chance (grey bars). Grey bars and errors bars represent average and standard deviation of I, S and NS mutations expected from 10,000 random permutations of hqSNVs across the tree.

'Monster Soup, commonly called Thames Water', by William Heath

Chapitre 3 : Association de méthodes métagénomiques et de culture bactérienne dans la détection de souches hypermutantes de *Vibrio cholerae* au sein de patients infectés

A combination of metagenomic and cultivation approaches reveals hypermutator phenotypes within Vibrio cholerae infected patients

Inès Levade[1], Ashraful I. Khan[2], Stephen B. Calderwood[3,4,5], Jason B. Harris[3,6], Regina C. LaRocque[3], Firdausi Qadri[2], Ana A. Weil[7], B. Jesse Shapiro[1]

[1]Département de Sciences biologiques, Université de Montréal, Montréal, Québec, 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, WA, USA

In prep.

Résumé

Vibrio cholerae peut provoquer toute une série de symptômes chez les patients infectés, allant de la diarrhée grave au portage asymptomatique. Des études antérieures utilisant le séquençage du génome entier (WGS) de plusieurs isolats bactériens par individu ont montré que *V. cholerae* présente une variation intra-patient sous forme de mutations ponctuelles et de gain/perte de gènes chez des patients infectés symptomatiques. Une telle évolution au sein du patient pourrait contribuer à expliquer la variabilité des symptômes observée chez les patients infectés, mais l'infection par *V. cholerae* chez des patients asymptomatiques a été peu étudiée jusqu'à présent. Nous avons donc combiné le WGS d'isolat cultivés et la métagénomique pour étudier l'évolution de *V. cholerae* chez une cohorte de patients atteints de choléra symptomatiques et asymptomatiques. Alors que la métagénomique nous a permis de détecter davantage de mutations chez les patients symptomatiques par rapport à l'approche dépendant de la culture, il était nécessaire de procéder à un WGS des isolats pour détecter la diversité de V. cholerae chez les porteurs asymptomatiques, car le séquençage métagénomique de V. cholerae est peu lu chez ces patients. Nos analyses ont révélé que les patients symptomatiques et asymptomatiques présentent un niveau similaire de diversité intra-patient, ainsi que la présence de souches hypermutatrices chez certains patients. Les bactéries hypermutatrices présentent généralement un avantage sélectif à court terme dans des conditions environnementales stressantes, mais finissent par accumuler un excès de mutations délétères. Nous montrons que cela est également susceptible d'être le cas lors d'infections individuelles de choléra. Contrairement à l'accumulation probable de mutations délétères chez les hypermutateurs, le *V. cholerae* non mutateur a montré des signes d'évolution convergente, suggérant que certains gènes sont sous l'action de la sélection positive chez plusieurs patients. Nos résultats présentent les avantages de l'utilisation de la métagénomique combinée au WGS pour caractériser la diversité intra-patient dans les cas d'infection aiguë du choléra et le portage asymptomatique, et présente l'hypermutation comme un mécanisme évolutif important mais sous-estimé lors de l'infection par *V. cholerae*.

Mots-clés : *Vibrio cholerae,* choléra, métagénomique, évolution intra-patient, hypermutation, patient asymptomatique

Abstract

Vibrio cholerae can cause a range of symptoms in infected patients, from severe diarrheal disease to asymptomatic carriage. Previous studies using whole genome sequencing (WGS) of multiple bacterial isolates per patient have shown that *V. cholerae* can experience a few mutations and several gene gain/loss events within symptomatic infected patients. Such within-patient evolution might help explain variable disease outcomes, but *V. cholerae* from asymptomatic patients has been understudied to date. Here, we combined culture-based population genomics and metagenomics to investigate *V. cholerae* within-patient evolution in a cohort of symptomatic and asymptomatic cholera patients. While metagenomics allowed us to detect more mutations in symptomatic patients compared to the culture-dependent approach, WGS of isolates was necessary to detect diversity in *V. cholerae* from asymptomatic carriers, owing to few *V. cholerae* metagenomic sequencing reads from these patients. Our analyses revealed that symptomatic and asymptomatic patients present a similar level of within-patient diversity, as well as the presence of hypermutator strains in some patients. Hypermutators are thought to enjoy a short-term selective advantage under stressful environmental conditions, but eventually to accumulate an excess of deleterious mutations. We show this is also likely to be the case during individual cholera infections. In contrast to the likely accumulation of deleterious mutations in hypermutators, the non-mutator *V. cholerae* showed signs of convergent evolution, suggesting certain genes under positive selection in multiple patients. Our results highlight the power of metagenomics combined with WGS to characterize within-patient diversity in acute cholera infection and asymptomatic carriage, and propose hypermutation as an important but underappreciated evolutionary mechanism during *V. cholerae* infection.

Keywords: *Vibrio cholerae*, cholera, metagenomics, within-patient evolution, hypermutation, asymptomatic patient

Introduction

Infection with *Vibrio cholerae,* the etiological agent of cholera, causes a clinical spectrum of symptoms that range from asymptomatic colonization of the intestine to severe watery diarrhea that can lead to death. Although absent from most resource-rich countries, this severe diarrheal disease is still plaguing many developing nations. According to the WHO, there is an estimation of 1.3 to 4.0 million cases of cholera each year, with 21,000 to 143,000 deaths worldwide (Ali et al. 2015). Cholera predominantly occurs in endemic areas, but can also cause explosive outbreaks as seen in Haiti in 2010 or in Yemen, where over 2.2 million suspected cases have occurred since 2016 (Weil, Ivers, et Harris 2011; Camacho et al. 2018). Even if a more widespread use of cholera vaccine and an improvement in its efficacity have improved the prevention of the disease in numerous locations, the increasing number of persons lacking access to sanitation and safe drinking water, the emergence of pandemic strains of *V. cholerae* with increased virulence (Satchell et al. 2016), and environmental persistence of this waterborne pathogen underscore the need to understand and interrupt the transmission of this disease.

Despite over a century of research on cholera, we lack a full understanding of the variation in the disease severity, which range from asymptomatic carriage to cholera gravis, which can lead to a fluid loss of up to 1 litre per hour (Nelson et al. 2009). Understanding of the epidemiology and evolutionary dynamics of this pathogen have been significantly improved by the use of next generation sequencing technologies and new modeling approaches, at a global and local scale (Weil et Ryan 2018; Domman et al. 2018). However many questions remain regarding asymptomatic carriers of *V. cholerae*, including their role and importance in the transmission chain during an epidemic (King et al. 2008; Phelps, Simonsen, et Jensen 2019), or even the potential differences in host response or bacterial mechanisms between symptomatic and asymptomatic patients leading to attenuation of symptoms. Numerous observational studies already identified host factors that could impact the severity of symptoms, including lack of pre-existing immunity, blood group O status, age, polymorphisms in genes of the innate immune system or gut microbiome composition (Harris et al. 2005; 2008b; Weil et al. 2009; Midani et al. 2018; Levade et al. 2020). However, *V. cholerae* population genetic variation within infected patients could also be important in

explaining the clinical manifestation of infection, and it is not clear if bacterial population in asymptomatic patients are similar than in patients presenting symptoms.

Recent studies have shown that despite the acute nature of the infection, which typically lasts only a few days, genetic diversity can appear and be detected in a *V. cholerae* population infecting individual patients (Seed et al. 2014; Levade et al. 2017). In a previous study, we sampled multiple *V. cholerae* isolates from each of eight patients (five from Bangladesh and three from Haiti) and sequenced 122 bacterial genomes in total. Using stringent controls to guard against sequencing errors, we detected few (0-3 per patient) within-patient intra-host single nucleotide variants (iSNVs), and a greater number of gene content variants (gene gain/loss events within patients) (Levade et al. 2017). This variation has been shown to have a potential effect for the adaptation of the host environment, either by resistance to phage predation (Seed et al. 2014) or with an impact on biofilm formation (Levade et al. 2017), but it is not clear yet if variation in the *V. cholerae* population could have an impact on the disease severity.

Mutations within the host have been shown to be adaptive for several pathogens (Didelot et al. 2016), and hypermutation phenotypes have been observed in some cases (Jolivet-Gougeon et al. 2011; Lieberman et al. 2013; Marvig et al. 2013). Hypermutation is a phenotype whereby a strain loses the function of its mismatch repair machinery and thus become a hypermutator. These hypermutators may be quick to acquire adaptive mutations, but also suffer a burden of deleterious mutations on the long term (Giraud et al. 2001). For the adaptive genotypes to be maintained in the population, each hypermutator can transmit its genes only by returning to a non-mutator state or by recombining its genes with non-mutator members of the population (Denamur et al. 2000; Jolivet-Gougeon et al. 2011). Such hypermutator phenotypes have been observed in vibrios in the aquatic environment (Chu et al. 2017), and induced in *V. cholerae* in an experimental setting (Wang et al. 2018), but never detected within infected patients. There were some evidence for hypermutation in *V. cholerae* clinical strains isolated between 1961 and 1965 (Didelot et al. 2015), however these strains had been maintained in stab cultures for many years, and hypermutators could have evolved during long-term culture (Eisenstark 2010). Therefore, it is still not clear if *V. cholerae* hypermutators actually arise within infected patients.

Moreover, when within-patient populations are studied with culture-based methods, the diversity may be underestimated, as the culture process can select isolates more suited to growth in culture, and due to undersampling of rare variants initially present in the bacterial population. To overcome these culture limitations, we performed whole genome shotgun metagenomic sequencing directly from stool samples collected from symptomatic and asymptomatic infected patients, and characterized the genetic heterogeneity of the *V. cholerae* population. Although resolution of these mixed populations into individual haplotypes in metagenomes is still an open computational challenge (Segata 2018), several methods have been recently developed to resolve strain-level variation from shotgun metagenomic data, but are often restrained to the detection of one consensus strain per sample (Nayfach et al. 2016; Scholz et al. 2016; Truong et al. 2017) or restricted to a few marker gene regions, which decrease the detection resolution (Scholz et al. 2016; Zolfo et al. 2017). Here we used the program InStrain (Olm et al. 2020), which can reconstruct the strain cloud of a bacterial population, by mapping metagenomic reads on metagenomic assembled genomes (MAGs) and identify the allele frequency of each variant in the population.

In this study, we used a combination of metagenomic analyses and whole genome sequencing of cultured isolates to characterize the within-patient diversity of *V. cholerae* patients with different clinical symptoms ranging from symptomatic to asymptomatic. Through metagenomic analysis, we found that previous culture-based analyses likely underestimated the true variation within infected hosts. However, asymptomatic carriers yielded too few metagenomic reads to assess within-patient variation, which could only be accessed using cultured *V. cholerae* isolates. Using this culture-based approach to compare symptomatic to asymptomatic contacts from three households, we found similar levels of within-patient diversity regardless of disease severity. The preliminary results should be replicated in larger sample sizes. Using both approaches, we also provide evidence for the presence of hypermutator *V. cholerae* strains within both symptomatic and asymptomatic infected patients. These hypermutators are characterized by a high mutation rate, and accumulation of an excess of likely neutral or deleterious mutations in the genome.

Materials and methods

Sample collection, clinical outcomes and metagenomic sequencing

To study within-host diversity of *Vibrio cholerae* during infection, we used stool and swab samples collected from index patients enrolled at the icddr,b (International Center for Diarrheal Disease Research, Bangladesh) Dhaka Hospital, and from their household contacts, as previously described (Midani et al. 2018). Index cases were defined as patients with severe acute diarrhea and a stool culture positive for *V. cholerae*. Individuals who shared the same cooking pot with an index patient for three or more days are considered as household contacts and were enrolled within 6 hours of the presentation of the index patient to the hospital. Rectal swabs were collected each day during a ten-days follow up period after presentation to the index case, coupled with a daily clinical assessment of symptoms, and collection of serological data from blood draw. Contacts were determined to be infected if any rectal swab culture was positive for *V. cholerae* and/or if the contact developed diarrhea and a 4-fold increase in vibriocidal titer during the follow-up period. If they developed watery diarrhea during this period, contacts with positive rectal swabs were categorized as symptomatic and those without diarrhea were considered asymptomatic. We excluded patients with age below 2 and above 60 years old, or with major comorbid conditions (Harris et al. 2008b ; Weil et al. 2009).

Fecal samples and rectal swabs from the day of infection and the following day were collected and immediately placed on ice after collection and stored at -80°C until DNA extraction. DNA extraction was performed with the PowerSoil DNA extraction kits (Qiagen) after pre-heating to 65°C for 10 min and to 95°C for 10 min. Sequencing libraries were constructed for 33 samples from 31 patients, for which we obtained enough DNA. We used the NEBNext Ultra II DNA library prep kit and sequenced the libraries on the Illumina HiSeq 2500 (paired-end 125 bp) and the Illumina NovaSeq 6000 S4 (paired-end 150 bp) at the Genome Québec sequencing platform (McGill University).

Metagenomic analyses

Sequence preprocessing and assembly

Sequencing fastq files were quality checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We removed human and technical contaminant DNA by aligning reads to the PhiX genome and the human genome (hg19) with Bowtie2 (Langmead et Salzberg 2012) , and used the iu-filter-quality-minoche script of the illumina-utils program with default parameters to filter the reads (Eren et al. 2013).

Taxonomic assignment

Processed paired-end metagenomics sequences were classified using two taxonomic profilers: Kraken2 v.2.0.8_beta (a k-mer matching algorithm) (Wood, Lu, et Langmead 2019) and MIDAS v.1.3.0 (a read mapping algorithm) (Nayfach et al. 2016). Kraken 2 examines the k-mers within a query sequence and uses the information within those k-mers to query a database, then maps k-mers to the lowest common ancestor (LCA) of all genomes known to contain a given k-mer. Kraken2 was run against a reference database containing all RefSeq viral, bacterial and archaeal genomes (built in May 2019), with default parameters. MIDAS uses a panel of 15 single-copy marker genes present in all of ~31,000 bacterial species included in its database to perform taxonomic classification and maps metagenomic reads to this database to estimate the read depth and relative abundance of 5,952 bacterial species. We identified metagenomic samples containing *V. cholerae* and Vibriophage reads, and computed the mean coverage (number of reads per base-pair) of the *V. cholerae* pangenome in the MIDAS database (Table 1).

Assembly and binning of *Vibrio cholerae* genomes

To recover good quality metagenomics assembled genomes (MAGs) of *V. cholerae,* we selected metagenomic samples with coverage >10X against the *V. cholerae* pangenome in the MIDAS database, and used MEGAHIT v.1.2.9 (D. Li et al. 2016) to perform *de novo* assembly. For 9 of the 11 selected samples, we independently assembled the genome of each

sample, and co-assembled the two remaining samples, which belong to the same patient (a symptomatic infected contact on days 9 and 10). Contigs of <1.5 kb were discarded.

We extracted MAGs by binning of our metagenomic assemblies. Because no single binning approach is superior in every case, with performance of the algorithms varying between samples and biotopes, we used different binning tools to recover MAGs of good quality. Quality of a metagenomic bins is evaluated by its completeness (the level of coverage of a population genome), and the contamination level (the amount of sequence that does not belong to this population from another genome). These metrics can be estimated by counting the frequency of single-copy marker genes within each bin (Parks et al. 2015). After mapping each sample against its assembly with Bowtie 2, we predicted bins using CONCOCT v.1.1.0 (Alneberg et al. 2014), MaxBin 2 v.2.2.7 (Wu, Simmons, et Singer 2016) and MetaBAT 2 v.2.12.1 (Kang et al. 2019), with default parameters. We then used DAS_Tool v.1.1.1 on the results of these three methods, to select a single set of non-redundant, high-quality bins per sample (Sieber et al. 2018). DAS_Tool is a bin consolidation tool, predicts single-copy genes in all the provided bin sets, aggregates bins from the different binning predictions, and extracts a more complete consensus bin from each aggregate such that the resulting bin has the most single-copy genes while having a reasonably low number of duplicate genes (Sieber et al. 2018). We then used Anvi'o v.6.1 (Eren et al. 2015) to manually refine the bins with contamination higher than 10% and Centrifuge v.1.0.4_beta (Kim et al. 2016) to determine the taxonomy of all bins in each sample, in order to identified *V. cholerae* MAGs.

Bins with completeness >60% and contamination <10% were selected in the final set of bins (completeness >90% and contamination <1% for the *V. cholerae* bins). We dereplicated the entire set of bins with dRep v.2.2.3 using a minimum completeness of 60%, the ANImf algorithm 99% secondary clustering threshold, maximum contamination of 10%, and 25% minimum coverage overlap, and obtained 79 MAGs displaying the best quality and representing individual metagenomic species (MGS).

Detection of *Vibrio cholerae* genetic diversity within and between metagenomic samples

We created a bowtie2 index of the 79 representative genomes from the dereplicated set, including a single high-quality *V. cholerae* MAG, and mapped reads from each sample to this set. By including many diverse microbial genomes in the bowtie2 index, we aimed to avoid the mismapping of reads from other species to the *V. cholerae* genome, and to reduce potential false positive intra-host single nucleotide variant (iSNV) calls. We mapped the metagenomics reads of each sample with *V. cholerae* a coverage value >5X (obtained with MIDAS) against the set of 79 MAGs, using Bowtie2 (Langmead et Salzberg 2012) with the --very-sensitive parameters. We also used Prodigal (Hyatt et al. 2010) on the concatenated MAGs, in order to predict open reading frames using default metagenomic settings.

We then used inStrain on the 15 selected samples (https://instrain.readthedocs.io/en/latest/index.html). This program aims to identify and compare the genetic heterogeneity of microbial populations within and between metagenomic samples (Olm et al. 2020). "InStrain profile" was run on the mapping results, with the minimum percent identity of read pair to consensus set to 99%, minimum coverage to call a variant of 5X, and minimum allele frequency to confirm a SNV equal to 0.05. All non-paired reads were filtered out, as well as reads with an identity value below 0.99. Coverage and breadth of coverage (percentage of reference base pairs covered by at least one read) were computed for each genome. InStrain identified both bialllelic and multiallelic SNVs frequencies at positions where phred30 quality filtered reads differ from the reference genome and at positions where multiple bases were simultaneously detected at levels above the expected sequencing error rate. SNVs were classified as non-synonymous, synonymous, or intergenic based on gene annotations, and gene functions were recovered using the Uniprot database (The UniProt Consortium 2019) and Blast (Madden 2003). Then, similar filters than the ones described in (Garud et al. 2019) were applied on the detected SNVs. We excluded from the analyses all positions with very low or high coverage value D compared to the median coverage $\overline{D}$, and positions within 100 bp at contigs extremities. As sites with very low coverage could result from a bias in sequencing or library preparation, and sites with higher coverage could arise from mapping error or be the result of repetitive region or multi-copy

genes not well assembled, we masked sites in all the samples if $D < 0.3\overline{D}$ and if $D > 3\overline{D}$ in at least two samples.

Mutation spectrum of hypermutator and non-mutator samples

For each sample, iSNVs were categorized into six mutation types, based on the chemical nature of the nucleotide changes (transitions or transversions). We combined all the samples with hypermutators, and compared with the mutation spectrum of the non-mutators. The mutation spectrum was significantly different between the hypermutator samples and the non-hypermutator samples (Chi-squared test, $p < 0.01$). We then computed the mutation mean and standard error of each of the six mutation types and compared the two groups (Figure 2C).

Bacterial replication rate

Replication rates were estimated with the metric iRep (index of replication), which is based on the measurement of the rate of the decrease in average sequence coverage from the origin to the terminus of replication. iRep values (Brown et al. 2016) were calculated by mapping the sequencing reads of each sample to the *V. cholerae* MAG assembled from that sample.

Tests for natural selection

First, we identified signals of convergent evolution in the form of nonsynonymous iSNVs occurring independently in the same gene in multiple patients. To assess the significance of convergent mutation, we compared their observed frequencies to expected frequencies in a simple permutation. We ran separate permutations for non-mutators (two genes with convergent mutations in at least two out of eight non-mutator samples, including only one time point from the patient sampled twice (the latest time point), and excluding the outlier patient A with a large number of intergenic iSNVs) and possible hypermutators (five genes with convergent mutations in at least two out of five possible hypermutator samples). In each permutation, we randomized the locations of the nonsynonymous mutations, preserving the observed number of nonsynonymous mutations in each sample, and the observed distribution of gene lengths. For simplicity, we assumed that 2/3 of nucleotide

sites in coding regions were nonsynonymous. We repeated the permutations 1,000 times and estimated a *p*-value as the fraction of permutations yielding greater than or equal to the observed number of genes mutated in two or more samples.

Second, we compared natural selection at the protein level within versus between patients, using the McDonald-Kreitman test (McDonald et Kreitman 1991). We again considered hypermutators separately. Briefly, the four counts (Pn, Ps, Dn, Ds) of between-patient divergence (D) vs. within-patient polymorphism (P), and non-synonymous (n) vs synonymous (s) mutations were computed and tested for neutrality using a Fisher's exact test (FDR corrected p-values < 0.05).

Whole genome sequencing analyses

Culture of *Vibrio cholerae* isolates

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

Whole genome sequencing and preprocessing

We prepared 48 sequencing libraries using the NEBNext Ultra II DNA library prep kit (New England Biolabs) and sequenced them on the Illumina HiSeq 2500 (paired-end 125 bp) at the Genome Québec sequencing platform (McGill University). Sequencing fastq files were

quality checked with FastQC, and Kraken2 was used to test for potential contamination with other bacterial species (Wood, Lu, et Langmead 2019).

Variant calling and phylogeny

We mapped the reads for each sample to the MJ-1236 reference genome and called single nucleotide polymorphisms (SNPs, fixed within patients) and single nucleotide variants (SNVs, variable within patients) using Snippy v.4.6.0 (Seemann 2015), with default parameters. A concatenated alignment of these core variants was generated (Figure 3), and an unrooted phylogenetic tree was inferred using maximum parsimony in MEGA X (Stecher, Tamura, et Kumar 2020). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) are shown next to the branches. The MP tree was obtained using the Subtree-Pruning-Regrafting (SPR) algorithm with search level 1 in which the initial trees were obtained by the random addition of sequences (10 replicates). The tree is drawn to scale, with branch lengths calculated using the average pathway method and are in the units of the number of changes over the whole sequence

De novo assembly and pan genome analyses

We *de novo* assembled genomes from each isolate using SPAdes v.3.11.1 on the short reads, with default parameters (Bankevich et al. 2012) and used Prokka v1.5 (Seemann 2014) to annotated them. We constructed a pan-genome from the resulting annotated assemblies using Roary v.3.13.0 (Page et al. 2015), identifying genes present in all isolate (core genome) and genes only present in some isolates (flexible genome). The flexible genome and the phylogenetic tree were visualized with Phandango v.1.1.0 (Hadfield et al. 2018).

Results

Taxonomic analyses of metagenomics sequences from Vibrio cholerae infected index cases and household contacts

To evaluate the level of within patient diversity of *V. cholerae* populations infecting symptomatic and asymptomatic patients, we used both culture-based whole genome sequencing and culture-free shotgun metagenomic approaches (Figure 1). We performed metagenomic sequencing of 22 samples from 21 index cases and 11 samples from 10 household contacts infected with *V. cholerae,* of which two stayed asymptomatic during the follow-up period. After removal of reads mapping to the human genome, we used Kraken2 and MIDAS to taxonomically classify the remaining reads, and identify samples with enough *V. cholerae* reads to reconstruct genomes (Table S1). Among symptomatic patients (index cases and household contacts), 15 samples from 14 patients contained enough reads to reconstruct the *V. cholerae* genome with a mean coverage >5X. Neither of the two asymptomatic patients had enough *V. cholerae* reads in their metagenomic sequences to reconstruct genomes by mapping or *de novo* assembly (mean coverage <0.05X). We also detected reads from two known *Vibrio* phages in some of these samples, previously identified as ICP1 and ICP3 (Table S1).

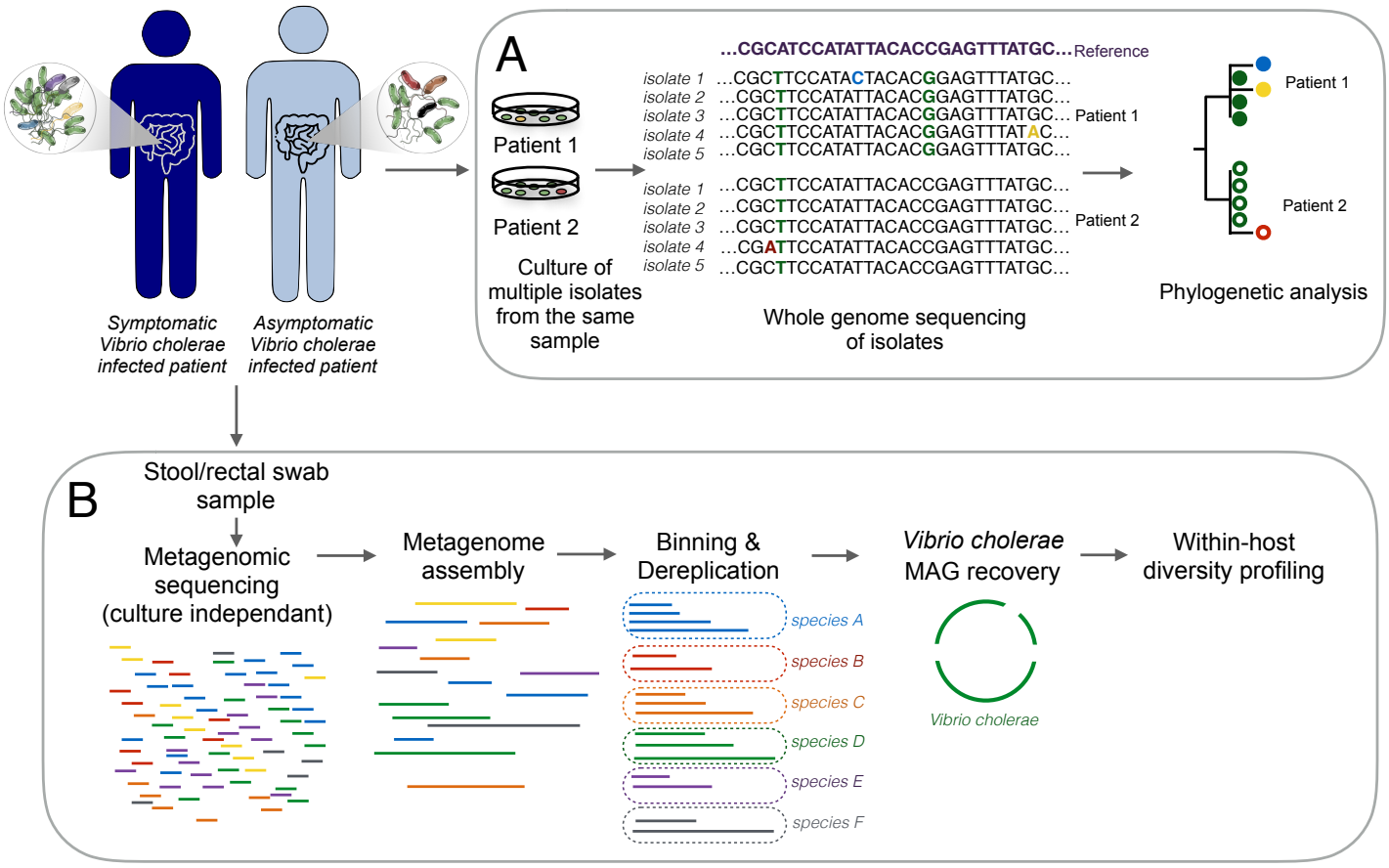


Figure 1. Summary of the culture-dependent and the metagenomics workflows for the characterization of the *Vibrio cholerae* within-patient diversity. Stool or rectal swab samples were collected from symptomatic and asymptomatic *V. cholerae* infected patients and processed using two different approaches: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

Recovery of high quality Vibrio cholerae MAGs from metagenomic samples

To reconstruct *V. cholerae* metagenomic assembled genomes (MAGs) from the 11 samples with coverage >10X, we *de novo* assembled each sample individually except for patient E, for whom we co-assembled two samples from two consecutive days. High quality MAGs identified as *V. cholerae* were obtained from each assembly, with no redundancy, and completeness ranging from 91 to 100% (Table S2). We dereplicated the set of bins and

removed all but the highest quality genome from each redundant set, identifying the bin from patient J as the best quality MAG overall.

Vibrio cholerae within patient nucleotide diversity estimated from metagenomic data

All metagenomes with *V. cholerae* mean coverage >5X were mapped against the dereplicated genome set, and we assessed within-patient genetic diversity using inStrain (Olm et al. 2020). We identified both single nucleotide polymorphisms (SNPs) that varied between patients (Table S3), and intra patient single nucleotide variants (iSNVs) that varied within patients (Table S4). We found a total of 39 SNPs between patients, and a range of two to 207 iSNVs within each sample (Table 1, Figure 2). Given the wide variation in coverage across samples, we checked for any bias toward detecting iSNVs in high-coverage samples. We observed no correlation between the number of detected iSNVs and coverage values (ρ = -0.12, $p > 0.05$, Pearson correlation), suggesting that diversity levels are comparable across samples.

Several mechanisms could account for the origins of the observed iSNVs, including *de novo* mutation within a patient, co-infection by divergent *V. cholerae* strains, or homologous recombination. Most iSNVs had low-frequency minor alleles (Figure S1), consistent with recent mutations occurring within individual patients, rather than co-infection by distantly related strains. No iSNVs were observed at the exact same nucleotide position in different patients, suggesting that iSNVs are rarely transmitted and never precisely recurrent in our dataset. In patient E, sampled on two consecutive days, we detected eight iSNVs on the first day, of which four were also detected on the second day, along with 13 additional iSNVs. This suggests that iSNV allele frequencies could fluctuate over time in the same, but more data are needed to test this hypothesis.

Table 1. Within patient *Vibrio cholerae* diversity profiles from 15 metagenomes.

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	-

Mutations segregating within patients are denoted iSNVs. The number of iSNVs and mean coverage values were computed with InStrain (Olm et al. 2020) and replication rate with iRep (Brown et al. 2016).

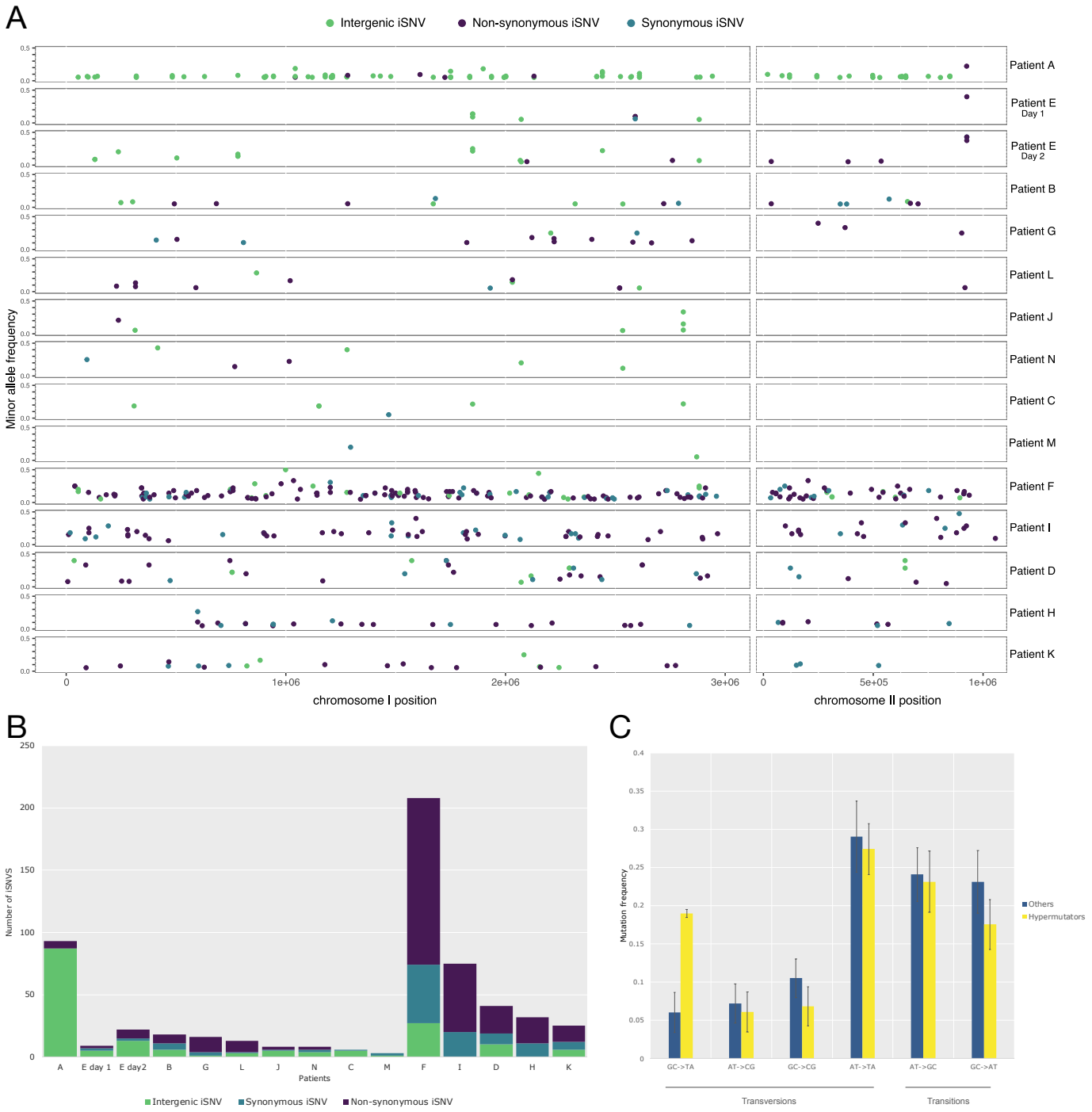


Figure 2. Within patient *Vibrio cholerae* diversity from metagenomic data. (A) Minor allele frequency and distribution of intergenic, synonymous and non-synonymous iSNVs across the two *V. cholerae* chromosomes for the 14 patients (B) Number and proportion of intergenic, synonymous and non-synonymous iSNVs for each patient (C) Transversion/Transition mutation spectrum of the iSNVs in the samples with more than 6 iSNVs. Error bars represent standard errors

Moreover, all the iSNVs were distributed across the genome (Figure 2A), rather than clustered as would be expected from recombination events (Croucher et al. 2011). Therefore, although we cannot strictly exclude co-infections or recombination events as sources of diversity, most of the observed iSNVs are consistent with *de novo* mutation within patients.

Evidence for Vibrio cholerae hypermutators within patients

Among five of the six patients with a high number of iSNVs (>25), we identified non-synonymous (NS) mutations in genes involved in DNA mismatch repair pathways, including the DNA polymerase II in patient D, or proteins of the methyl-directed mismatch repair (MMR) system in patient F, I and K (Table 1). These NS variants could explain why some samples seem to have a higher level of within host diversity, as it could indicate the presence of *V. cholerae* strains with a hypermutator phenotype (Jolivet-Gougeon et al. 2011). In the patient harboring the highest number of variants (Patient F, 207 iSNVs), we detected two NS mutations in two different genes involved in DNA repair: in the DNA mismatch repair endonuclease MutL (Jolivet-Gougeon et al. 2011), and in the nuclease SbcCD subunit C (Darmon et al. 2007; Lovett 2011; Didelot et al. 2015). The patient with the second highest number of iSNVs, patient A, contained a high number of intergenic variants (87 out of 96 iSNVs, Figure 2B), but no apparent NS mutations in genes involved in DNA repair. We obtained the same level of excessive variation in the intergenic regions while using the MAG obtained from the same patient (data not shown), excluding the possibility that this high number of mutations is due to mapping errors. In patient I, where we also detected a high number of iSNVs, a NS mutation in the gene coding for the MutT/nudix protein, involved in the repair of oxidative DNA damage (Lu et al. 2001), could also cause a strong hypermutation phenotype. Patient D, H and K presented less iSNVs but also showed NS mutations in genes involved in DNA damage repair (Foster et al. 1995; Lee, Sung, et Verdine 2019). However, some of these genes have been shown to play less critical roles in bacterial DNA repair than MutSLH (Kunkel et Erie 2005; Jolivet-Gougeon et al. 2011), which could lead to a weaker hypermutator phenotype for these *V. cholerae* strains. Studies described an excess of one type of mutations over the others in the case of hypermutator strains, like an increase of transition over transversion mutations for an hypermutator *Burkholderia dolosa* lineage with a NS mutation in MutL (Lieberman et al. 2013) or an excess of G:C→T:A transversions in

a *Bacillus anthracis* strong hypermutator (Zeibell et al. 2007) and in members of the gut microbiome (Zhao et al. 2019). When we looked at the spectrum of mutations in the samples containing weak and strong hypermutators (Figure 2C), we found that the mutation spectrum was significantly different between the group of hypermutator samples and the samples with no hypermutator and more than six iSNVs (Chi-square test, $p < 0.01$), and that difference was due to an apparent excess of G:C→T:A transversions in the hypermutators group (Figure 2C).

Current theory suggests that hypermutators may be adaptive under novel or stressful environmental conditions because they more rapidly explore the mutational space and are the first to acquire adaptive mutations. However, hypermutation comes at the cost of the accumulation of deleterious mutations. To test the hypothesis that hypermutation leads to fitness costs due to these deleterious mutations, we used iRep (Brown et al. 2016) to estimate *V. cholerae* replication rates in each sample, and test whether replication rate was negatively associated with the number of iSNVs. iRep infers replication rates from MAGs and metagenomic reads (Brown et al. 2016). For instance, an iRep value of 2 would indicate that most of the population is replicating one copy of its chromosome. In our data, iRep values varied from 1.23 (patient E) at day 2 to 5.43 (patient D), and we did not find a clear association between the replication rate of *V. cholerae* and the number of iSNVs detected within each subject (Figure S2B, $\rho = 0.15$, $p > 0.05$, Pearson correlation). This suggests that deleterious mutations in hypermutators could be counterbalanced by adaptive mutations that maintain growth. Higher iRep values could also be associated with larger *V. cholerae* population sizes, which could support greater genetic diversity and yield a positive correlation between iRep values and the number of iSNVs. These hypotheses merit testing in larger patient cohorts.

Tests for natural selection during Vibrio cholerae within patient evolution

While none of the patients shared iSNVs at the exact same nucleotide position, some contained convergent mutations in the same gene (Table 2). To determine whether genes that acquired multiple mutations could be under positive selection within the host, we performed permutation tests for hypermutator and non-mutator samples separately (Methods). Among the hypermutator samples, we identified five genes with NS mutations in

two or more patients (Table 2), which was not an unexpectedly high level of convergence given the large number of mutations in hypermutators (one-sided permutation test, 1000 permutations, $p = 0.97$). That the p-value approaches 1 suggests either that the hypermutators are actually selected to avoid mutating the same genes in different patients, or – more likely – that the permutation test is conservative. For the samples with no evidence for hypermutator phenotypes, we identified two genes with NS mutations in two patients. The first gene, *hlyA*, encodes a hemolysin that causes cytolysis by forming heptameric pores in human cell membranes, (Olson et Gouaux 2005), while the second gene encodes a putative ABC transporter ferric-binding protein (Table 2). Observing two convergent mutations in two different genes is somewhat unexpected (one sided permutation test, 1000 permutations, $p = 0.039$) in a test that is likely to be conservative. We also note that the three NS iSNVs in *hlyA* have minor alleles at relatively high frequencies (0.40, 0.43 and 0.22) compared to other convergent NS mutations (ranked 1 and 4 out of 17; Table 2) and compared to NS mutations overall (ranked 4 and 40 out of 290, median allele frequency of 0.12; Table S4). Together, these analyses suggest that *V. cholerae* hypermutators produce NS mutations that are likely deleterious or neutral, while evidence for within-patient positive selection on certain genes in non-mutators merits further investigation.

To test for differential selection at the protein level within and between patients, we applied the McDonald-Kreitman test (McDonald et Kreitman 1991) on the 9 patients with no hypermutator strains and on the five patients with potential hypermutators. Based on whole-genome sequences of *V. cholerae* isolates, we previously found an excess of NS mutations fixed between patients in Bangladesh, based on a small sample of five patients (Levade et al. 2017). Here, based on metagenomes from a larger number of patients, we found the opposite pattern of a slight excess of NS mutations segregating as iSNVs within patients, consistent with slightly deleterious mutations occurring within patients and purged over evolutionary time. However, the difference between NS:S ratios within and between patients was not statistically significant (Fisher's exact test, $p > 0.05$; Table S5); thus the evidence for differential selective pressures within versus between cholera patients remains inconclusive.

Table 2. Set of genes with mutations identified in more than one patient.

Protein	Patient A	Patient B	Patient D	Patient E	Patient F	Patient H	Patient I	Patient K
Hemolysin (VC cytolysin)	NS (0.22)	-	-	2 NS (~0.4)	-	-	-	-
2-aminoethylphosphonate ABC transporter ferric-binding protein	-	NS (0.05)	-	NS (0.05)	-	-	-	-
Peptidase B	-	-	NS (0.33)	-	-	-	NS (0.09)	-
Nuclease SbcCD subunit C	-	-	S (0.28)	-	NS (0.09)	-	-	-
C4-dicarboxylate transport sensor protein	-	-	-	-	NS (0.08)	-	NS (0.11)	-
zinc/cadmium/mercury/lead-transporting ATPase	-	-	-	-	NS (0.08)	-	-	NS (0.06)
hypothetical protein	-	-	-	-	NS (0.14)	-	-	NS (0.14)
hypothetical protein	-	-	-	-	NS (0.33)	NS (0.11)	-	-
Formamidopyrimidine-DNA glycosylase mutM	-	-	-	-	S (0.18)	-	-	NS (0.08)
Phosphoribosylformylglycinamidine synthase	-	-	-	-	-	-	NS (0.06)	S (0.08)

The presence of a synonymous or non-synonymous iSNV in each gene and each patient is indicated with S or NS, respectively, and the minor allele frequency is shown in parentheses. None of the mutation were found at the same nucleotide or codon position. Patients containing possible or likely hypermutators are underlined. Only genes and patients containing more than one mutated gene are shown

When we looked at the functional categories of genes with mutations in patients with hypermutators and patients with no hypermutation phenotype (Figure S3), we noticed that a lot of NS mutations occurred in genes whose function could have a positive impact on the survival of these strains (transmembrane transport proteins, pathogenesis, response to

antibiotics, secretion system, chemotaxis, other metabolic processes...), even if there is no strong sign of positive selection acting on these variants. Both groups, hypermutator samples (Figure S3B) and non-mutators samples (Figure S3C) have a high NS:S ratio in genes of unknown function and hypermutators samples have lots of NS mutations in transmembrane proteins, which is not the case in non-mutators. However non-mutator samples have more NS mutations in genes involved in pathogenesis and secretion systems. Most of the NS mutations involved in pathogenesis were found in the gene hlyA, which codes for a toxin that has been shown to have both vacuolating and cytocidal activities against a number of cell lines, including human intestinal cells (Tsou et Zhu 2010), and to be an important virulence factor in *V. cholerae* El Tor O1 (Olivier et al. 2007).

Whole genome sequencing of Vibrio cholerae isolates confirm hypermutator phenotypes and similar diversity levels in symptomatic and asymptomatic patients

In addition to metagenomic analyses, we performed the whole genome sequencing of multiple *V. cholerae* clinical isolates from index cases and asymptomatic contacts (Figure 1A) from three households (56, 57, and 58, Table S1). As noted above, asymptomatic carriers did not yield sufficient metagenomic reads to assemble the *V. cholerae* genome or call iSNVs, but they did yield colonies for whole-genome sequencing. The first asymptomatic contact, 58.01, tested positive to *V. cholerae* on day 4 after the presentation of the index case to the hospital, and was sampled on days 4, 6, 7 and 8, as *V. cholerae* was still present in the stools and culturable. We sequenced five isolates respectively from day 4 and 6 samples, and four isolates from each of the subsequent days. For households 56 and 57, five isolates were sequenced from each sample, at day 1 for the index cases and day 2 for the asymptomatic carriers (Table S6).

The index case from household 58 (called 58.00 or patient N) was also included in the metagenomic analysis described above, allowing a comparison between culture-dependent and culture-independent assessments of within-patient diversity. We did not detect any iSNVs in patient 58.00, as the five isolates sequenced were isogenic. In contrast, the metagenomic

analysis of patient N revealed seven iSNVs (Table 1), suggesting a higher sensitivity for the detection of rare variants.

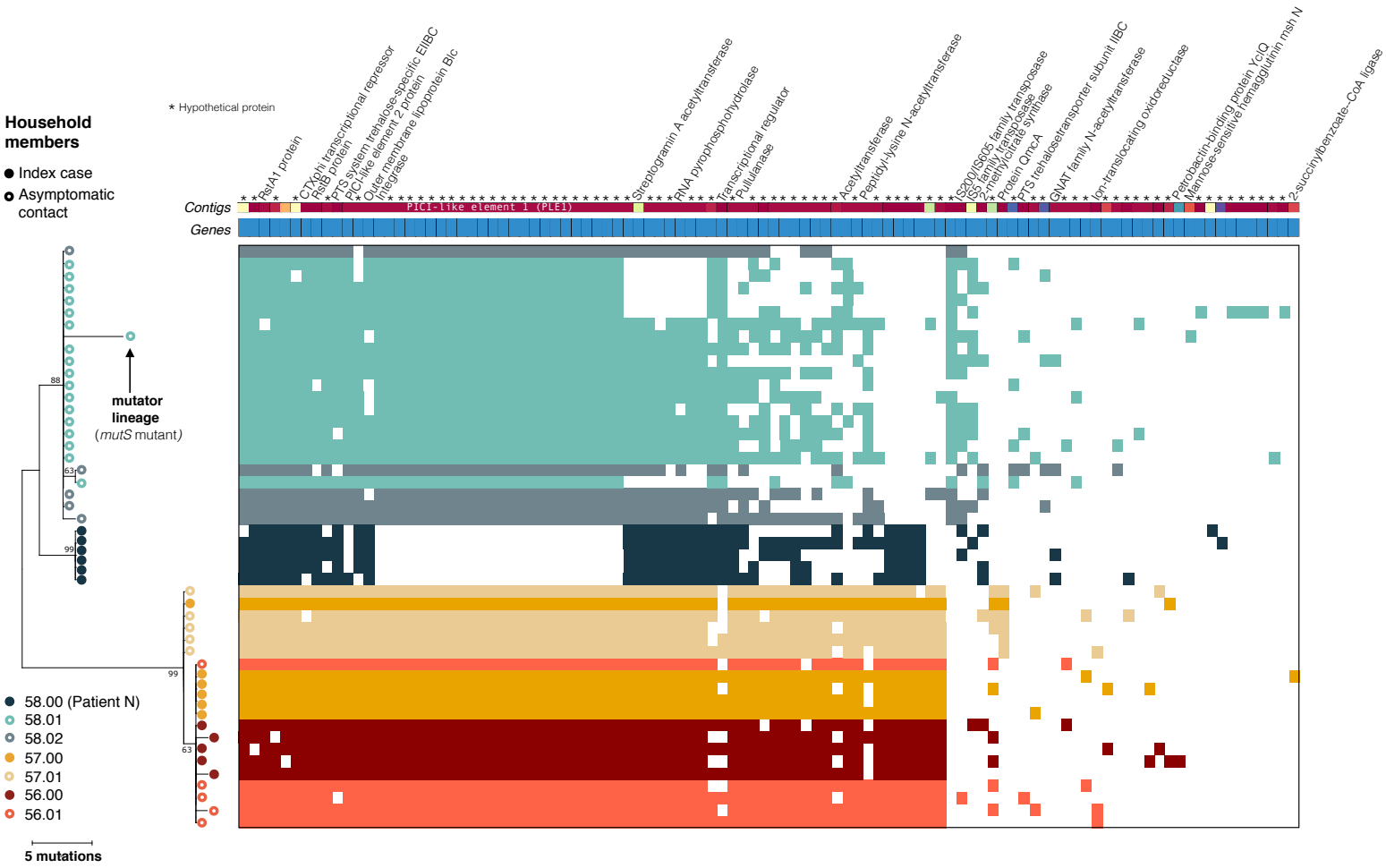


Figure 3. Phylogenetic and pan-genomic analysis from 48 *Vibrio cholerae* isolates from index cases and their asymptomatic contacts. Phylogenetic tree was inferred using the Maximum Parsimony method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) are shown next to the branches. Plain circles represent isolates from index cases and empty circles represent isolates from their asymptomatic contacts. Heat map of the gene presence-absence is based on the 102 genes of the flexible genome (color: present, colorless: absent). Each row corresponds to an isolate from the phylogenetic tree and each column represents an orthologous gene family. Each color represents an individual.

In contrast to metagenomes consisting of many unlinked reads, whole-genome sequencing allows the reconstruction of a phylogeny describing the evolution of *V. cholerae* within and between patients (Figure 3). As described previously (Domman et al. 2018), members of the same households tend to cluster together. In index case 57.00, four of the

isolates were isogenic, and one isolate was identical to the five isolates sequenced from the asymptomatic contact from the same household, patient 57.01 (Table 3, Figure 3). This shared genotype between the two individuals was unexpected, and could suggest a potential transmission event from the asymptomatic contact to the index case, followed by a mutational event and the spreading of the new variant in the *V. cholerae* population of the index case. The only mutation found in four of the five isolate from the index case was a non-synonymous mutation in a gene coding for a cyclic-di-GMP-modulating response regulator, which could have an impact on the regulation of biofilm formation in the host (Tischler et Camilli 2004). However, this scenario about a transmission event is only supported by one mutation, therefore remains uncertain. Among the other index cases, we found no iSNVs in patient 58.00 and two iSNVs in patient 56.00. One isolate from this patient had a synonymous mutation in a hypothetical protein, and another isolate had a non-synonymous mutation in a UDP-N-acetylglucosamine 4,6-dehydratase gene (Table 3). We detected iSNVs in the other asymptomatic contacts, with one synonymous and one intergenic mutation in contact 58.02, and one non-synonymous mutation in one isolate from contact 56.01 (Table 3, Figure 3).

Interestingly, we also found evidence for a hypermutator in contact 58.01. One isolate sampled from this contact had the highest number of mutations seen in on any branch in the phylogeny (five NS mutations) which could be explained by a NS mutation in the gene coding for the MutS protein, another key component of the methyl-directed mismatch repair (MMR) system (Table 3, Figure 3). Mutation in this gene could explain the accumulation of a surprising number of mutations in this isolate, which could also be referred as a hypermutator (Didelot et al. 2015; Chu et al. 2017). This contact presented no variants in the isolates sampled at day 4 and 6, but we found this hypermutator isolate on day 7. However, this genotype was not found at day eight, either due to the lower resolution in the detection of variants with the WGS of cultured isolates, or the disappearance of this strain from the population.

Type	Isolates	Mutation type	Nucleotide position in MJ-1236	Ref. amino acid	Alt. amino acid	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	-

Table 3. Nucleotide changes identified in core genes of the *V. cholerae* strains isolated from index cases (56.00, 57.00 and 58.00) and their asymptomatic contacts. Genome position is according to the MJ-1236 reference genome. Mutations segregating within patients are denoted iSNVs; Mutations fixed between patients are denoted 'Patient.' Nucleotide positions are based on the reference *V. cholerae* MJ-1236 (CP001485.1, CP001486.1). Patient allele frequency shows the allele frequency of the alternative (minor) allele. Ref=Reference allele; Alt=Alternative allele. NS=non-synonymous; S=synonymous. Chr1=chromosome 1; Chr2=chromosome

Pan genome analyses

Whole-genome isolate sequencing also provides the opportunity to study the variation in gene content (the pangenome) within and between patients. We identified a total of 3,523 core genes common to all *V. cholerae* genomes, and 102 flexible genes present in some but not all genomes (Figure 3; Table S7). We also found an additional 214 genes present uniquely in isolate 56.00C4, in one single contig identified as the lytic *Vibrio* phage ICP1, which was assembled with the *V. cholerae* genome. This phage contig contained the ICP1 CRISPR/Cas system, which consists of two CRISPR loci (designated CR1 and CR2) and six *cas* genes, as previously described (Seed et al. 2011; 2013). These genes were excluded from subsequent pan genome analyses.

Among the 102 flexible genes, some varied in presence/absence within a patient, ranging from twelve to 53 genes gained or lost per patient (Table S7; Figure 3). The majority of these flexible genes (78%) were annotated as hypothetical, and several were transposases or prophage genes. A large deletion of 24 genes was detected in patient 58.00, in a 18kb

phage-inducible chromosomal island (PICI) previously shown to prevent phage reproduction and is targeted by the ICP1 CRISPR/Cas system (Seed et al. 2013).

Discussion

Although the existence of multiple *V. cholerae* strains with closed but distinct genotypes has been previously reported in infected patients (Kendall et al. 2010; Seed et al. 2012; 2014; Levade et al. 2017), our results confirmed that within-patient genetic diversity of *V. cholerae* population is a common feature observed in symptomatic patients, but also in asymptomatic infected contacts. In this study, we used a combination of metagenomic and WGS sequencing technologies to characterize this within patient diversity, and which allowed us to describe for the first time multiple patients from the same cohort presenting various hypermutator phenotypes in the *V. cholerae* populations infecting them.

Notably, we showed that metagenomics can display a higher sensitivity in the detection of rare variants in the *V. cholerae* population. Indeed, in our previous observations we detected a level of intra-host diversity ranging from zero to three iSNVs in cultured isolates from patients with acute infection (Levade et al. 2017). In contrast, our metagenomic analyses allowed us to detect two iSNVs in the patient with the lowest level of diversity, but up to 207 iSNVs in another individual (Table 1). In the only patient for which we were able to characterize *V. cholerae* intra-host diversity both from the metagenome and from cultured isolates, we did not identify any iSNVs in the isolates, but detected 7 iSNVs from the metagenomic analyses, even with a coverage <10X. These results highlight one of the potential limitations of the culture-based approach for the study of within-host diversity of microbial population, which is the difficulty to recover rare members of the community. A cultured isolate collection is more likely dominated by the most abundant organisms of the population, or some strains with specific phenotype could be non-culturable in some conditions (Brenzinger et al. 2019).

However, despite better resolution in the detection of rare variants, the use of metagenomics has its limitations. Within-sample diversity profile cannot be established for low-abundance microbes that lack sufficient sequence coverage (<5X) and depth, and this level of coverage is difficult to obtain for most of the bacterial species in human samples. Moreover, although some bacterial populations can be near-clonal, others can be a more

complex strain mixtures, which increases the difficulty to generate high-quality genomes. And even when high quality MAGs are recovered in a case of low-complex population, identification of real variants is being challenged by low coverage, mismapped reads from other species, or SNVs frequency not substantially different from the expected sequencing error rate. In this study, only 48% of the samples collected on patients with acute symptoms, which usually harbour a high fraction of vibrios in their stool (10^{10}-10^{12} vibrios per litre), showed enough reads to reconstruct *V. cholerae* MAGs and conduct diversity analyses. As asymptomatic patients typically shed less vibrios in their stool compared to symptomatic patients (Nelson et al. 2009), analyses of *V. cholerae* using metagenomics on these clinical samples would require additional steps on the processing, like depletion of the host DNA or targeted sequence capture techniques, but these approaches also present limitations (Vezzulli et al. 2017; Bachmann et al. 2018). In our case, only the use of isolate culture allowed us to identify intra-host variation in asymptomatic patients, showing a level of variation similar to symptomatic patients using the same method. Culture and sequencing of clinical isolates also present the advantages to link the variants from the same strain and to place all the within-host strains in the right phylogenetic context, which is essential in the reconstruction of pathogen transmission events (Didelot et al. 2016; 2017; Gardy et Loman 2018). Moreover, even if pangenome analysis can be conducted from metagenomic samples by looking at which genes are present or absent within different strains of a species (Scholz et al. 2016; Nayfach et al. 2016), only a consensus profile can be generated and comparison of the gene sets is only possible between samples. The use of culture isolate sequencing allowed us to confirm that, as previously shown (Levade et al. 2017), changes in the flexible gene content accumulate more quickly than point mutations, even in asymptomatic patients. This approach also gave us insight in potential differences between patients, regarding *V. cholerae* protection strategies against phage predation, which would have been complicated to demonstrate with the only use of metagenomic sequencing. Indeed, the absence of a phage-inducible chromosomal island (PICI) was detected in one isolate in patient 58.00. PICI-like elements are induced during phage infection, and interfere with phage reproduction via multiple mechanisms (Ram et al. 2012; O'Hara et al. 2017). The deletion of this PICI element in the *V. cholerae* genome is a likely consequence of an ongoing evolutionary arms race between *V. cholerae* and its phages. Finally, the use of a culturable approach allowed us to confirm that

asymptomatic patients can be carriers of culturable bacteria for multiple days and to show that hypermutator strains are not only present in infected patients with strong symptoms.

Hypermutation has been defined as an excess of mutations due to deficiency in DNA mismatch repair and hypermutator strains have been described during diverse pathogenic infections and *in vivo* experiments, including *P. aeruginosa, H. influenzae* and *Streptococcus pneumoniae* in cystic fibrosis patients, or *E. coli* in diverse habitats (Labat et al. 2005; Oliver et Mena 2010; Jolivet-Gougeon et al. 2011). In the case of *V. cholerae,* a study on clinical isolates described that among 260 genomes from lineage L2 of the seventh pandemic they analysed, 17 strains isolated between 1961 and 1965 presented a higher number of SNPs uniformly distributed along the genomes (Didelot et al. 2015). They showed that 14 of these 17 isolates possessed a total of 18 genetic variations in one or more of four genes (*mutS, mutH, mutL* and *uvrD*) that play a key role in the mismatch repair system, as well as 12 changes in ten other genes that can affect mismatch repair, in ten of the isolates (Didelot et al. 2015). They cautiously speculated that this apparent high frequency of hypermutators could be associated with the rapid spread of the seventh pandemic, particularly because hypermutators may be a sign of population bottlenecks and recent selective pressure. However, they also hypothesized that these old strains of *V. cholerae* could show an higher number of mutation because they had been maintained in stab cultures for many years, so it was unclear if this phenotype was derived in patients or during culture (Didelot et al. 2015; Eisenstark 2010). Using our metagenomic approach, we showed that this hypermutator phenotype was not a culture effect.

Moreover, it was also not previously clear if hypermutators emerged in patients or in the aquatic environment, or if they have an adaptive function in infected patients. More generally, hypermutator phenotype is believed to be advantageous for the colonization of new environments or hosts, allowing the hypermutator bacteria to generate adaptive mutations more quickly, which leads to the more efficient exploitation of the resources or a better resistance to environmental stressful conditions, like antibiotics (Giraud et al. 2001; Labat et al. 2005; Oliver et Mena 2010; Jolivet-Gougeon et al. 2011). However, this high mutation rate can have a negative impact on the fitness in the long term, with most of the mutations being neutral or deleterious, and functions dispensable in the current environment possibly compromised, impacting the survival in other environments (Funchain et al. 2000; Giraud et

al. 2001; Chu et al. 2017). In our case, the high majority of the mutations were at low frequency within patients with hypermutator phenotypes and we could not find a clear difference between the synonymous and non-synonymous of mutations, suggesting that they were accumulated as neutral mutations. A previous study conducted on a murine model, showed that hypermutation can be used as an adaptive strategy by the *V. cholerae* in order to resist host produced reactive oxygen species induced stress, and lead to a colonization advantage by increased catalase production and increased biofilm formation (Wang et al. 2018). However, even if a lot of NS mutations were found in genes that could be adaptive in the host environment, we did not detect any signs of selection on these mutations.

In patients with no hypermutator strains, we seemed to observe the same dynamics, where synonymous and non-synonymous mutations accumulate at similar rates within and between hosts, as expected for neutral genetic changes. We hypothesized that, in most cases, the acute nature of the infection provides limited time for selection to have detectable effects on within-patient diversity. However, we found evidence for convergent evolution in two patients harboring mutations in the same hemolysin gene involved in pathogenesis (Table 2). Previous studies have shown that selective pressures can act on the within host population, by selecting mutation that could have an impact on biofilm formation within the host (Levade et al. 2017) or in the context of phage predation (Seed et al. 2014). It seems that selective pressures may be idiosyncratic and person-specific in the case of *V. cholerae* infections, and more studies using experimental models would be necessary to characterized adaptive variation within the human host.

In conclusion, our results illustrate the potential of metagenomics as a culture-independent approach for the characterization of within-host diversity of a well-known pathogen, and how this diversity seems to be underestimated with the only use of traditional culture techniques. However, strain-level profiling remains a big challenge in shotgun metagenomic analyses (Segata 2018; Quince et al. 2017), and we demonstrated here that a combination of metagenomic and cultivation approaches can be a good way to overcome the limitations of each of these methods. Further studies on the short-term evolution of *V. cholerae*, the role of hypermutator strains within the host and the role of asymptomatic patients in the epidemic dynamics would then benefit from this approach.

Funding information

This study was supported by CIHR (Canadian Institutes of Health Research) and the Canada Research Chairs program (BJS), The icddr,b: Centre for Health and Population Research, grants AI099243 (J.B.H and L.C.I), 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.) 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 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.

Supplementary data

Supplementary tables

Supplementary tables are available on GitHub:
https://github.com/ilevade/Metagenomics_vibrio_cholerae_within_host_diversity

Supplementary figures

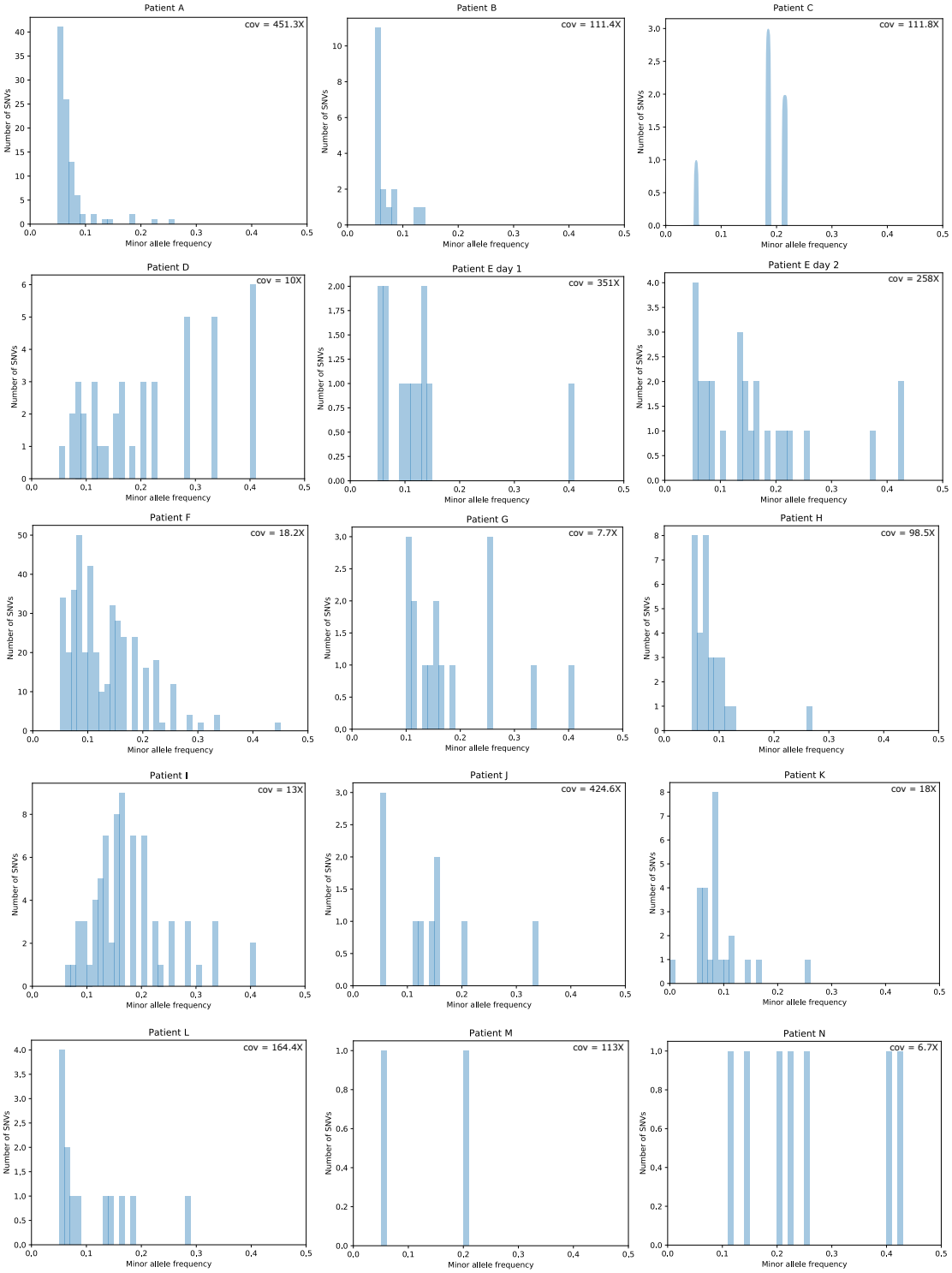


Figure S1. Minor allele frequency distribution for iSNVs in the 15 metagenomic samples. Allele frequencies and mean coverage values (cov) were computed with Instrain (Olm et al. 2020)

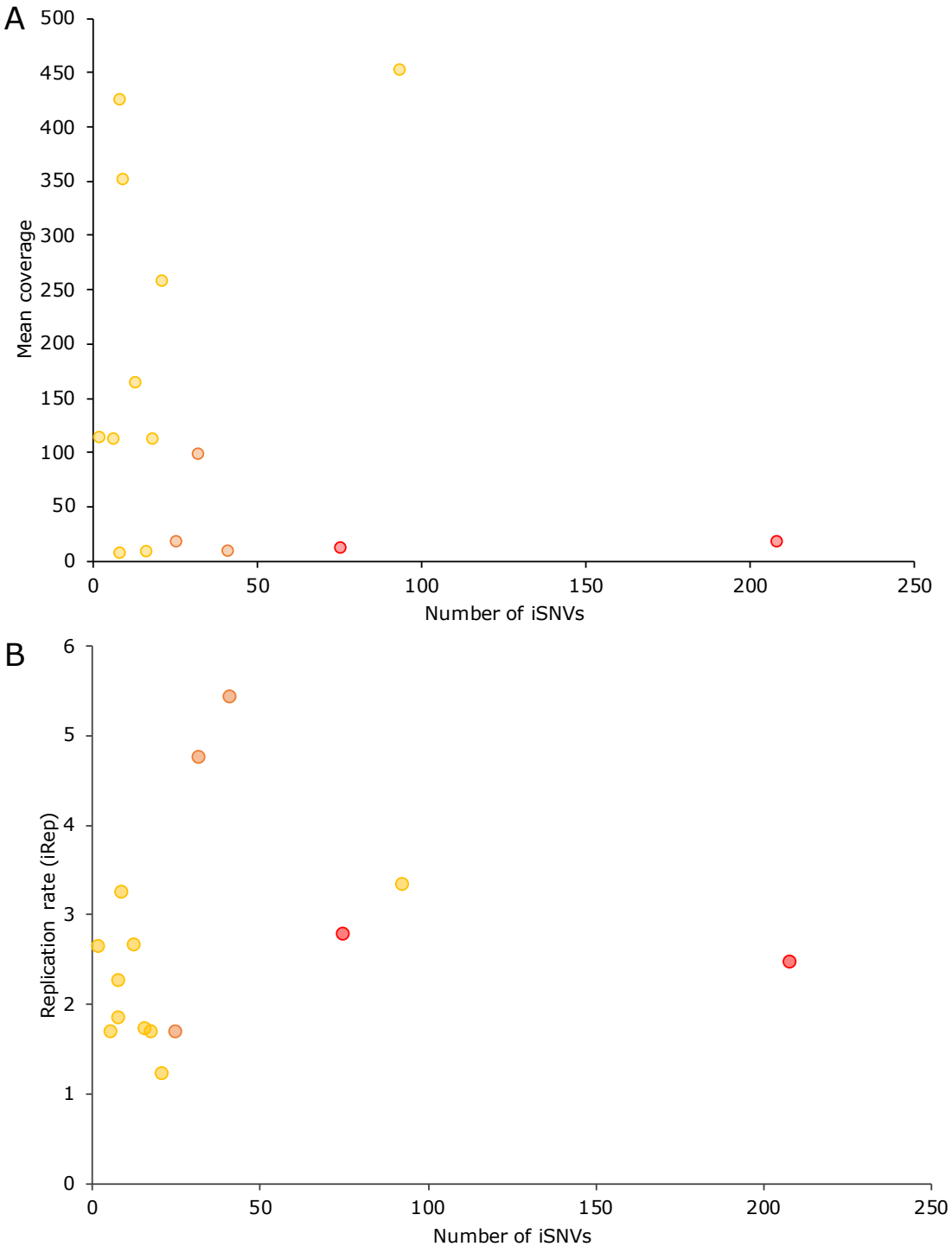


Figure S2. iSNVs numbers are not linked to depth of metagenomic read coverage (A) or replication rate (B). Patients with strong hypermutation phenotypes are represented in red, weak hypermutation phenotypes in orange, and other samples in yellow.

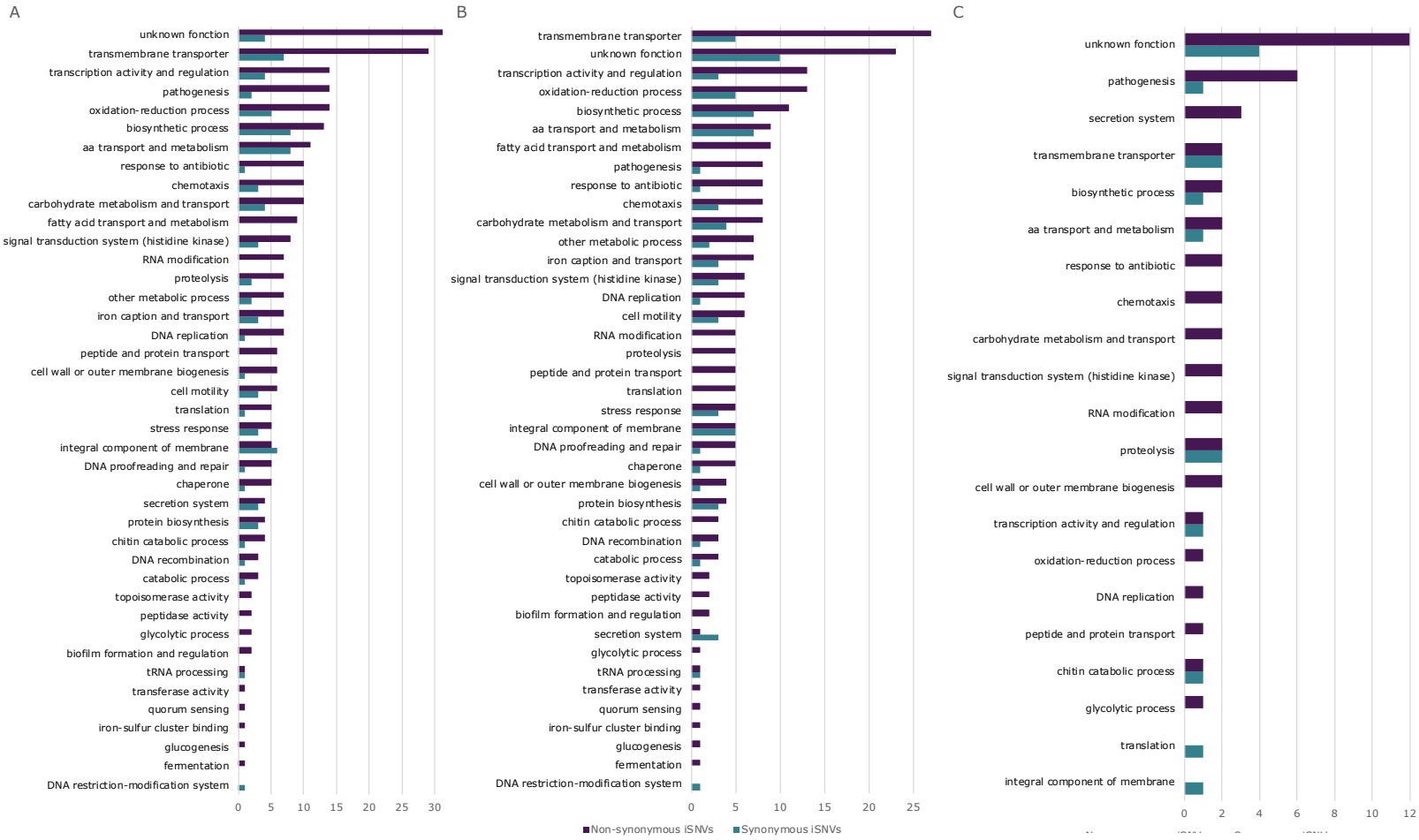
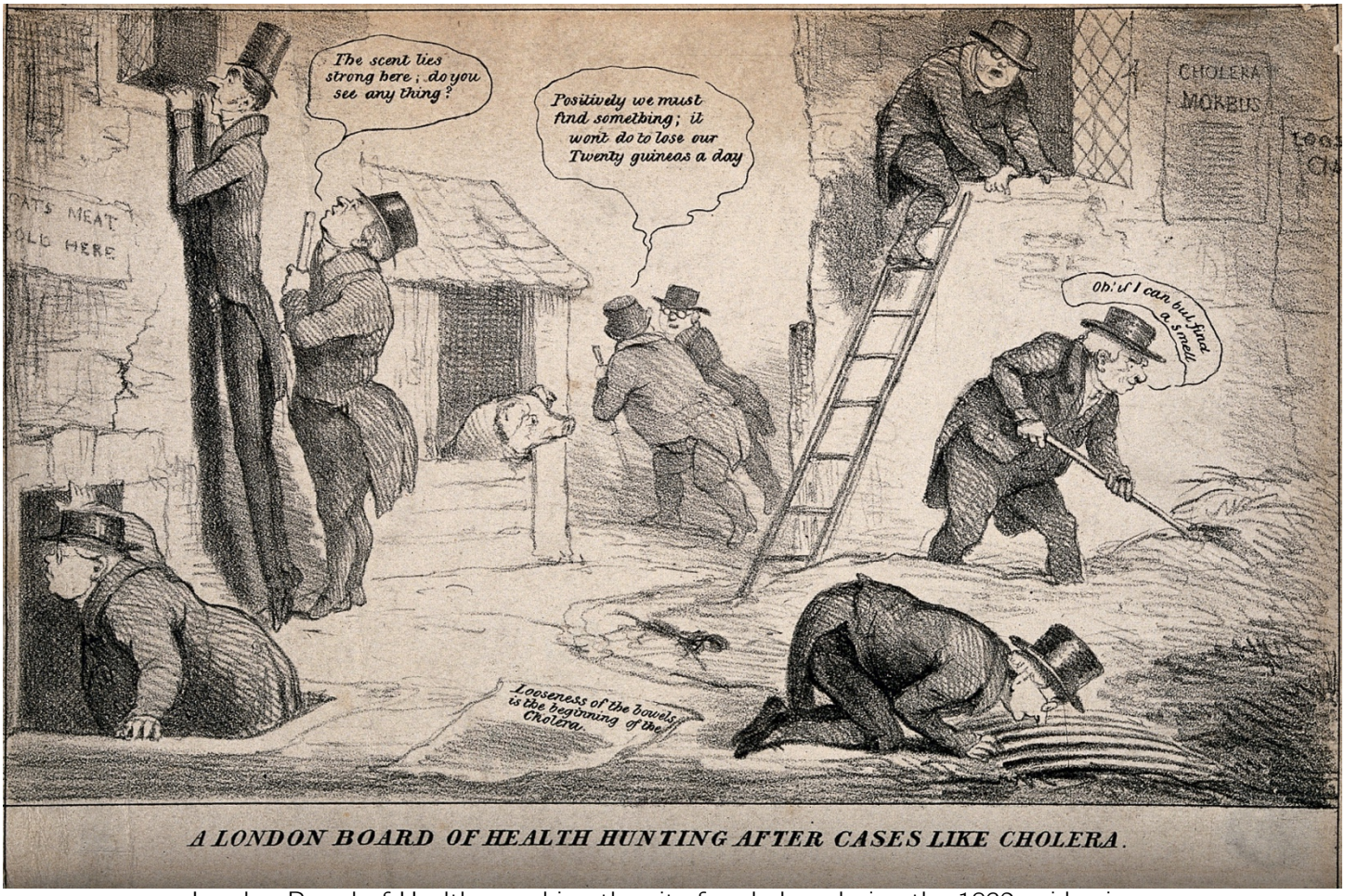


Figure S3. Functional categories of genes containing within-patient variants (iSNVs). The total number of synonymous and non-synonymous iSNVs in different functional categories was represented for each all patients (A), for the five patients with hypermutation phenotypes (B), and for the nine patients with no hypermutation phenotypes (C). Categories are ranked in descending order of the number of non-synonymous iSNVs.



London Board of Health searching the city for cholera during the 1832 epidemic. Lithograph, 1832. Credit: Wellcome Collection

Chapitre 4 : Étude d'une cohorte prospective et analyses métagénomiques pour la prédiction du risque d'infection et de la gravité des symptômes lors de l'exposition à *Vibrio cholerae*

Predicting Vibrio cholerae infection and disease severity using metagenomics in a prospective cohort study

Inès Levade[1], Morteza M. Saber[1], Firas Midani[2,3,4], Fahima Chowdhury[5], Ashraful I. Khan[5], Yasmin A. Begum[5], Edward T. Ryan[6,7,9], Lawrence A. David[2,3,4,8], Stephen B. Calderwood[6,7,10], Jason B. Harris[6,11], Regina C. LaRocque[6], Firdausi Qadri[5], B. Jesse Shapiro[1], Ana A. Weil[12]

[1]Département de Sciences biologiques, Université de Montréal, Montréal, Québec, Canada

[2]Program in Computational Biology and Bioinformatics, Duke University, Durham, NC, USA

[3]Center for Genomic and Computational Biology, Duke University, Durham, NC, USA

[4]Department of Molecular Genetics and Microbiology, Duke University, Durham, NC, USA

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

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

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

[8]Department of Biomedical Engineering, Duke University, Durham, NC, USA

[9]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA

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

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

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

Published in *The Journal of Infectious Diseases*, jiaa358,
https://doi.org/10.1093/infdis/jiaa358

Résumé

La susceptibilité à l'infection par *Vibrio cholerae* peut être influencée par le groupe sanguin, l'âge et l'immunité préexistante chez l'individu, mais ces facteurs n'expliquent que partiellement pourquoi certaines personnes d'un même foyer sont infectées et d'autres non. Une étude récente a utilisé le séquençage de l'amplicon de l'ARNr 16S pour quantifier la composition du microbiome intestinal et identifier les biomarqueurs prédictifs de l'infection, mais avec une résolution taxonomique limitée. Pour obtenir une meilleure résolution des facteurs microbiens intestinaux associés à la susceptibilité à l'infection par *V. cholerae* et identifier des prédicteurs de la maladie symptomatique, nous avons appliqué le séquençage métagénomique à une cohorte de contacts familiaux de patients atteints de choléra. Grâce à l'apprentissage machine, nous avons résolu les espèces, les souches, les familles de gènes et les voies cellulaires dans le microbiome au moment de l'exposition à *V. cholerae* afin d'identifier les marqueurs qui prédisent l'infection et les symptômes. Nous avons démontré que l'utilisation de caractéristiques métagénomiques peut améliorer la précision et l'exactitude des prédictions par rapport au séquençage 16S. Nous avons également prédit la gravité de la maladie, bien qu'avec une plus grande incertitude que notre prédiction de l'infection. Les espèces des genres *Prevotella* et *Bifidobacterium* prédisaient la protection contre l'infection, et les gènes impliqués dans le métabolisme du fer étaient également en corrélation avec la protection. Nos résultats soulignent l'apport de la métagénomique dans la prédiction de l'évolution des maladies et ont permis d'identifier des espèces et des gènes potentiels pouvant être tester expérimentalement afin d'étudier les mécanismes de protection contre le choléra liés au microbiome intestinal.

Mots-clés : *Vibrio cholerae*, choléra, microbiome, apprentissage machine, métagénomique

Abstract

Susceptibility to *Vibrio cholerae* infection is impacted by blood group, age, and pre-existing immunity, but these factors only partially explain who becomes infected. A recent study used 16S rRNA amplicon sequencing to quantify the composition of the gut microbiome and identify predictive biomarkers of infection with limited taxonomic resolution. To achieve increased resolution of gut microbial factors associated with *V. cholerae* susceptibility and identify predictors of symptomatic disease, we applied deep shotgun metagenomic sequencing to a cohort of household contacts of patients with cholera. Using machine learning, we resolved species, strains, gene families, and cellular pathways in the microbiome at the time of exposure to *V. cholerae* to identify markers that predict infection and symptoms. We demonstrated that the use of metagenomic features can improve the precision and accuracy of prediction relative to 16S rRNA amplicon sequencing. We also predicted disease severity, although with greater uncertainty than our infection prediction. Species within the genera *Prevotella* and *Bifidobacterium* predicted protection from infection, and genes involved in iron metabolism also correlated with protection. Our results highlight the power of metagenomics to predict disease outcomes and suggest specific species and genes for experimental testing to investigate mechanisms of microbiome-related protection from cholera.

Keywords: *Vibrio cholerae,* cholera, microbiome, machine learning, metagenomics

Introduction

Cholera is an acute diarrheal disease caused by *Vibrio cholerae*. It is a major public health threat worldwide that continues to cause major outbreaks, such as in Yemen, where over 1.7 million cases have been reported since 2016 (Ali et al. 2015; Camacho et al. 2018). Transmission of *V. cholerae* between household members commonly occurs through shared sources of contaminated food or water or through fecal-oral spread (Domman et al. 2018; Weil et al. 2009). The clinical spectrum of disease ranges from asymptomatic infection to severe watery diarrhea that can lead to fatal dehydration (Nelson et al. 2009). Host factors such as age, innate immune factors, blood group, or prior acquired immunity partially explain why some people are more susceptible to *V. cholerae* infection than others, but a substantial amount of the variation remains unexplained (Harris et al. 2008a).

The gut bacterial community can protect against enteropathogenic infections (Ubeda, Djukovic, et Isaac 2017; Baumgartner et al. 2020), and may explain some of the variation in *V. cholerae* susceptibility. Several studies have identified commensal bacteria and mechanisms that could be protective against *V. cholerae*. For instance, a species enriched in the gut microbiota of patients recovering from cholera, *Blautia obeum,* was found to interfere with *V. cholerae* pathogenicity through quorum-sensing inhibition in a mouse model (Hsiao et al. 2014). Animal and *in vitro* experiments have demonstrated that alteration of commensal-derived metabolite levels influenced host susceptibility by affecting *V. cholerae* growth or colonization (Bachmann et al. 2015; Mi Young Yoon et al. 2016; Kaur et al. 2018; Mao et al. 2018; You et al. 2019).

Studies of *V. cholerae* and the gut microbiota often focus on a limited number of bacterial species or involve patients who already have symptomatic cholera (Hsiao et al. 2014; David et al. 2015). One study to date has characterized the gut microbiome of individuals exposed to *V. cholerae* to predict susceptibility to infection (Midani et al. 2018). In this study, Midani *et al.* developed a machine learning model to predict susceptibility based on 16S rRNA gene amplicon sequencing of the gut microbiota in a group known to have high risk of infection: household contacts of confirmed cholera patients (Weil et al. 2009). Midani et al. (2018) showed that the microbiome composition at the time of exposure to *V. cholerae* can predict infection with similar or better accuracy as the commonly measured host factors

known to impact susceptibility. However, 16S rRNA amplicon sequencing has limited taxonomic resolution and does not identify the genetic mechanisms of protection.

Here we used shotgun metagenomics to analyze an expanded prospective cohort of persons exposed to V. cholerae in Bangladesh. Our metagenomic analysis yielded improved outcome predictions compared to 16S rRNA sequencing, and identified bacterial genes associated with remaining uninfected after exposure to V. cholerae. We are also able to predict disease severity among infected contacts, albeit with lower power and precision than susceptibility. Finally, we highlight several microbiome-encoded metabolic functions associated with protection against cholera.

Materials and methods

Sample collection, clinical outcomes and metagenomic sequencing

As described in Midani et al. (2018), household contacts were enrolled within 6 hours of the presentation of an index cholera case at the icddr,b (International Center for Diarrheal Disease Research, Bangladesh) Dhaka Hospital. Index patients with severe acute diarrhea, a stool culture positive for *V. cholerae*, age between 2 and 60 years old, and no major comorbid conditions were recruited (Harris et al. 2008a; Weil et al. 2009). A clinical assessment of symptoms in household contacts was conducted daily for the 10-day period after presentation of the index case, and repeated on day 30. We collected demographic information, rectal swabs, and blood samples for ABO typing and vibriocidal antibody titers as described in the Supplementary Methods. During the observation period, contacts were determined to be infected if any rectal swab culture was positive for *V. cholerae* and/or if the contact developed diarrhea and a 4-fold increase in vibriocidal titer during the follow-up period (Harris et al. 2008a; Weil et al. 2009). Contacts with positive rectal swabs developing watery diarrhea were categorized as symptomatic and those without diarrhea were considered asymptomatic (Figure 1). *V. cholerae* positive contacts (by culture or 16S testing) at the time of enrollment were excluded, in addition to contacts who reported antibiotic use or diarrhea during the week prior to enrollment. DNA extraction was performed for the selected samples and used for shotgun metagenomics sequencing. Detailed information on cohorts, sequencing methods and sample processing are described in Supplementary Methods.

Taxonomic/functional profiling and predictive model construction

We used MetaPhlAn2 (version 2.9) (Truong et al. 2015) for taxonomic profiling and HUMAnN2 (Franzosa et al. 2018) was used to profile cellular pathways (from MetaCyc) and gene families (identified using the PFAM database of protein families). See the Supplementary Methods for further details on the bioinformatic analyses. For identification of metagenomic biomarkers of susceptibility and disease severity, we used MetAML (Pasolli et al. 2016), a computational tool for metagenomics-based prediction tasks and for quantitative assessment of the strength of potential microbiome-phenotype associations. We applied a random forests (RF) classifier on species, pathways and gene-family relative abundances, as well as strain-specific markers presence/absence. Models constructed using each of these different types of features were compared to each other, to a random dataset with shuffled labels, and to a model constructed with clinical/demographic data, using two-sample, two-sided *t*-tests over 20 replicate cross-validation, as previously described (Pasolli et al. 2016). We used a stratified 3-fold cross validation approach, splitting our dataset into a validation set (1/3 of samples) and a training set (2/3 of samples) with the same infected:uninfected ratio. The model was first applied on all the features, then we used an embedded feature selection strategy to identify the most useful features in the model and improve its accuracy. Feature relative importance was computed using the mean decrease in impurity strategy, which calculates importance of each feature as the sum of the number of nodes (across all trees) that use the feature, proportional to the number of samples each of these nodes splits (Pasolli et al. 2016). Further details are described in the Supplementary Methods.

Results

Metagenomic sequencing of the gut microbiome in household contacts exposed to V. cholerae

We performed metagenomic sequencing of the gut microbiome in 65 contacts of cholera cases from a cohort described by Midani et al. (2018), from which sufficient DNA remained for shotgun metagenomic sequencing. Of these 65 contacts, referred to as the Midani 2018 cohort, 20 developed infection during the follow-up period, and 45 remained uninfected (Figure 1). Among the 20 contacts who became infected, 10 had no symptoms during the follow-up period (30 days), and were classified as asymptomatic, and 10 developed symptoms (Supplementary Methods). To increase our sample size, we surveyed an expanded cohort (Table S1a) by adding 33 samples to the Midani 2018 cohort, including 10 additional pre-infection samples from timepoints of contacts in the Midani 2018 cohort, and 23 samples from 16 newly enrolled contacts from the same population and the same time period (2012-2014, Dhaka, Bangladesh).

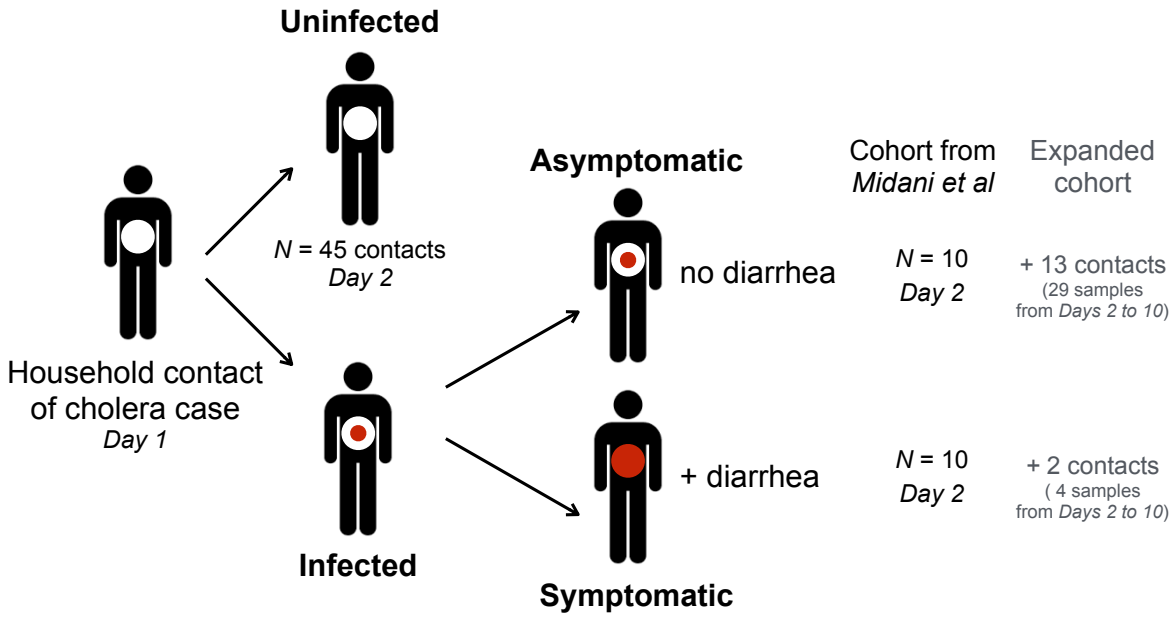


Figure 1. Study cohort in Dhaka, Bangladesh. After presentation of a *V. cholerae* culture-positive index case to the hospital on day 1, household contacts were enrolled on day 2. The expanded cohort includes the Midani 2018 cohort (15), with an addition of 33 samples from infected individuals (13 asymptomatic and 2 symptomatic).

We used pre-infection samples in order to identify factors predictive of disease outcomes and identify biomarkers in the microbiome of the Midani 2018 cohort, upon which we base the majority of our analyses. We also performed exploratory analyses on the expanded cohort to determine the potential for predictive models to be generalized to larger sample sizes.

Shotgun metagenomic DNA sequence reads from these samples were used to characterize four features of the microbiome: 1) relative abundances of microbial species, 2) the presence/absence of sub-species-level strains, 3) metabolic pathway relative abundances, and 4) gene family relative abundances (Table 1).

Predicting susceptibility to V. cholerae infection with Random Forest

We first used an RF model to predict *V. cholerae* susceptibility (developing infection or remaining uninfected) from baseline microbiome features (Figure 1). In the Midani 2018 cohort, functional pathways and gene families predicted infection significantly better than random (two-sample *t*-tests comparing area under the curve [AUC] across 20 replicate 3 fold cross-validations; $p < 0.05$) compared to data with shuffled (randomized) labels, and also predicted infection better than species or strain features (Table 1, Table S2). Pathways and gene families had significantly higher mean AUCs (0.71 and 0.74, respectively) compared to species or strains (0.61 and 0.62) ($p < 0.05$; Table 1; Figure S1, Table S3).

To determine the minimum number of metagenomic features required for accurate prediction, we repeated the analysis using smaller subsets of features. Using only 30 species, 60 gene families or pathways, or 200 strains achieved similar cross-validation AUC values (Figure S2). We then trained an RF model on this reduced number of selected features, yielding improved predictions for all feature types (Figure S1; Table S4). This suggests that only a limited number of strains, species, genes and pathways in the gut microbiome at the time of exposure are sufficient to predict *V. cholerae* susceptibility. For example, prediction using strain level markers after feature selection yielded an AUC of 0.95 (Table S4). However, such high AUC values should be treated with caution because the models can be overfit when a supervised feature selection step is applied on the same data used to train the model (Pasolli et al. 2016).

		RF – Cohort from Midani et al				RF – Expanded			
		Species abundance	Strain markers	Gene families	Pathways	Species abundance	Strain markers	Gene families	Pathways
	#features	705	54953	6810	443	807	62965	7514	461
Infected Vs Uninfected	Accuracy	0.73 (±0.02)	0.71 (±0.02)	0.76 (±0.02)	0.72 (±0.02)	0.76 (±0.03)	0.69 (±0.03)	0.80 (±0.02)	0.80 (±0.03)
	Precision	0.71 (±0.06)	0.68 (±0.06)	0.77 (±0.04)	0.70 (±0.05)	0.76 (±0.03)	0.70 (±0.03)	0.81 (±0.02)	0.81 (±0.03)
	F1	0.66 (±0.02)	0.64 (±0.03)	0.71 (±0.03)	0.66 (±0.03)	0. 75 (±0.03)	0.68 (±0.03)	0.80 (±0.02)	0.80 (±0.03)
	AUC	0.61 (±0.05)	0.62 (±0.04)	0.74 (±0.04)	0.71 (±0.04)	0.83 (±0.02)	0.76 (±0.03)	0.87 (±0.02)	0.88 (±0.02)
Shuffled	F1	0.55 (±0.04)	0.56 (±0.04)	0.56 (±0.04)	0.56 (±0.05)	0.40 (±0.03)	0.45 (±0.03)	0.48 (±0.03)	0.44 (±0.03)
	AUC	0.40 (±0.04)	0.57 (±0.04)	0.50 (±0.05)	0.50 (±0.04)	0.39 (±0.03)	0.52 (±0.03)	0.51 (±0.03)	0.46 (±0.03)
Asymptomatic vs Symptomatic vs Uninfected	Accuracy	0.70 (±0.02)	0.70 (±0.02)	0.69 (±0.01)	0.69 (±0.01)	0.68 (±0.01)	0.60 (±0.03)	0.69 (±0.02)	0.67 (±0.03)
	Precision	0.53 (±0.03)	0.53 (±0.03)	0.60 (±0.02)	0.59 (±0.02)	0.60 (±0.02)	0.53 (±0.03)	0.61 (±0.02)	0.59 (±0.02)
	F1	0.60 (±0.02)	0.59 (±0.02)	0.57 (±0.02)	0.57 (±0.02)	0.62 (±0.02)	0.55 (±0.03)	0.64 (±0.02)	0.62 (±0.02)
	AUC	NA	NA	NA	NA	NA	NA	NA	NA
Shuffled	F1	0.48 (±0.04)	0.49 (±0.04)	0.46 (±0.03)	0.55 (±0.03)	0.41 (±0.03)	0.35 (±0.03)	0.44 (±0.04)	0.37 (±0.03)
	AUC	NA	NA	NA	NA	NA	NA	NA	NA

Table 1. Assessment of prediction performance for a random forest (RF) model applied to the Midani 2018 and expanded cohorts. Species abundances, strain-specific markers presence/absence, relative abundance of Pfam-grouped gene families, and MetaCyc pathways were used as features. For each dataset, we applied a binary (uninfected vs. infected contacts) and a multi-class (asymptomatic vs. symptomatic vs. uninfected contacts) classifier and reported performance metrics for each dataset. Metrics obtained by the same classifier applied to the same datasets with shuffled class labels (random assignment of labels to samples) are also reported (shuffled). The margins of errors (95% confidence intervals) are reported in parenthesis.

Because we did not have a fully independent validation cohort (*e.g.* from another continent) to test our model, we decided to use the features selected from the Midani cohort to make predictions on the expanded dataset. Using the same features selected from the Midani 2018 training dataset, we made predictions on the expanded cohort and achieved AUCs between 0.89 and 0.93 for prediction of infection using the four types of features (Table S4). Again, because the expanded cohort partly overlaps with the Midani cohort, and includes some repeated samples from the same individuals on different study days, these results could also be prone to overfitting, but they demonstrate the potential for generalized predictions.

Finally, we repeated the RF analysis using all features in the expanded dataset and found that this increased predictive performance relative to the original Midani cohort (Figure S1). Once again, genes and pathways outperformed species and strains according to all metrics, with AUC reaching ~0.88 using cellular pathways (Table 1). This improvement in the expanded cohort also highlights the importance of using larger and more balanced datasets as input to predictive models.

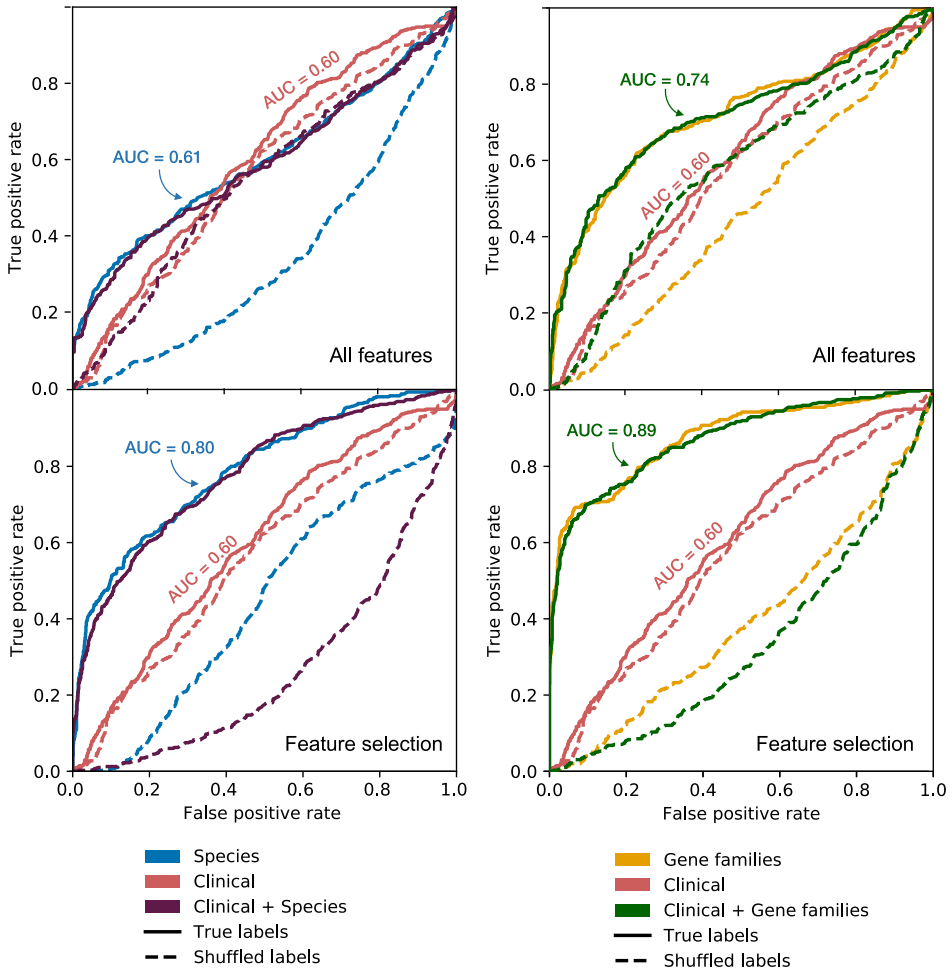


Figure 2. Metagenomic features predict *V. cholerae* infection better than clinical and demographic features. Random forest (RF) prediction of infection status was applied to seven clinical and demographic features, and compared with all species and all gene families (top row), as well as 30 selected species features from metagenomes and 60 selected gene family features (bottom row), or a combination of clinical, demographic and metagenomic features. Plots show receiver operating characteristic (ROC) curves (average across cross-validations) for the Midani 2018 dataset. Shuffled labels represent the prediction run on a dataset with a random assignment of infection outcomes. AUC = area under the curve.

Improved prediction compared to known factors impacting susceptibility

To put the metagenomic predictions in context, we compared their predictive power and accuracy to clinical and demographic factors (Table S1a). Three of these factors (age, baseline vibriocidal antibodies and blood group) are known to impact susceptibility to *V. cholerae* infection (6,15) and we used them to train RF models (Table S5). As expected, contacts who became infected tended to be younger and have lower baseline antibody titers than those who remained uninfected (Table S1b), but these small differences were not sufficient to train a significantly predictive model. An RF model trained on the seven clinical and demographic factors did not perform better than a random model with shuffled labels (AUC=0.60, p=0.66; Figure 2). Predictions were not improved using all species-level metagenomic features present at the time of exposure to *V. cholerae* (AUC=0.61), but significantly improved using a selected number of species (AUC=0.80, $p < 1 \times 10^{-7}$). The use of all gene families or a selected number of genes showed an increased predictive performance (AUC=0.74 and AUC=0.89 respectively; Figure 2) compared to species-level or clinical and demographic contact data ($p < 1 \times 10^{-7}$ for all comparisons). We again note the caveat that models with selected features may be overfit and represent an upper bound for predictive power. Even without feature selection, we found that gene families clearly provide superior predictions, and adding clinical data did not improve the predictions based on microbiome features alone (Figure 2). Together, these results demonstrate that gene families present in the gut microbiome at the time of exposure contain more information about *V. cholerae* susceptibility compared to species-level or clinical and demographic contact data.

Disease severity is more difficult to predict than likelihood of infection

To predict symptomatic disease among infected individuals (Figure 1), we divided samples into uninfected, symptomatic and asymptomatic groups and again applied the RF approach. We used the F1 score as a performance metric since it is well suited for uneven class distributions in our uninfected/symptomatic/asymptomatic comparison. Applied to the Midani 2018 cohort, this model predicted outcomes significantly better than random (shuffled labels) using species, strains or pathway data, but not gene families (Table 1; see Table S3 for *p*-values). However, the F1 scores for the symptomatic/asymptomatic predictions were systematically lower (mean scores ranging from of 0.57 to 0.60) than for the

infected/uninfected prediction (means ranging from 0.64 to 0.71). Using the expanded cohort, the scores were improved only slightly (Table 1). These results suggest that disease severity is predictable in principle, but with greater uncertainty than the simpler infection outcome.

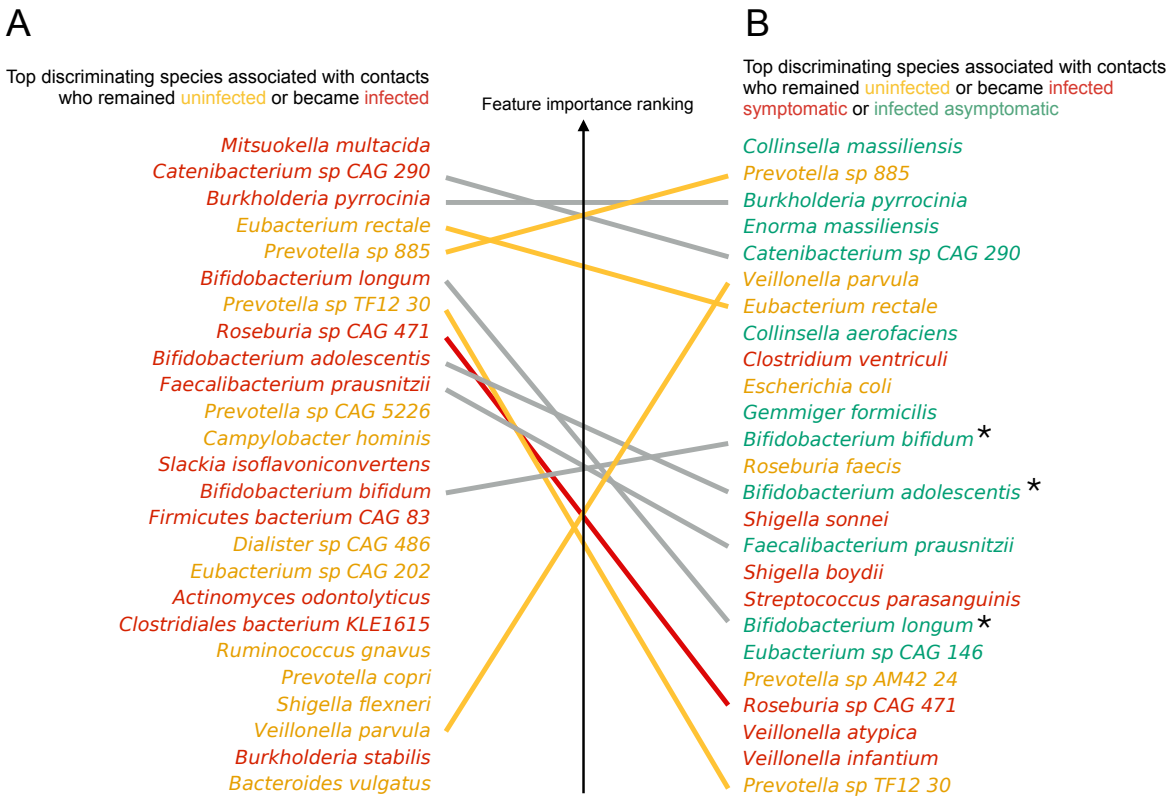


Figure 3. Most important discriminating species of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Species associated with contacts that became infected (red) or remained uninfected (yellow) during follow-up. (B) Species associated with contacts who remained uninfected (yellow), or became infected asymptomatic (green), or symptomatic (red) during follow-up. The top 25 most important features are shown here; see Table S6 for the full list. Yellow lines connect species associated with uninfected individuals in both (A) and (B); red lines connect species associated with infection in (A) and symptomatic disease in (B); grey lines connect species associated with infection in (A) but asymptomatic infection in (B). Three species of *Bifidobacterium* are marked with an asterisk.

Taxonomic biomarkers of disease susceptibility and severity

Predictive features in the gut microbiome identified to a species/strain or gene level allow the possibility of experimental follow-up to investigate mechanisms of the associations we observed. We characterized the most predictive species, pathways, and gene families

(Tables S6-S9). The most common discriminating species in individuals that remained uninfected during the follow-up period were *Eubacterium rectale, Campylobacter hominis, Ruminococcus gnavus, Bacteroides vulgatus, Veillonella parvula* and members of the *Prevotella* and *Eubacterium* genera (Figures 3A, S3A and S4A). These species are ranked by their importance score, which is effectively their relative weighting in the RF model. Several species associated with contacts that developed *V. cholerae* infection belonged to the genera *Bifidobacterium, Actinomyces* or *Collinsella*, and many of the species were also associated with asymptomatic infection (Figures 3B, S3B and S4B), including three species of *Bifidobacterium.* The top predictive species in contacts who developed symptomatic infection were *Clostridium ventriculi* (formerly *Sarcina ventriculi*), *Streptococcus parasanguinis* and members of *Veillonella*. *Shigella* species were also associated with the gut microbiome of persons who developed symptomatic *V. cholerae* infection, although persons enrolled in this study were *Shigella* stool-culture negative. *Shigella* identified by DNA presence in stool may be the result of recent or resolving infection, or may be present at subclinical levels due to ingestion of contaminated water. The features identified by the multivariate RF model were confirmed using univariate statistics for the uninfected/infected prediction (Figure S5), but the overlap was poorer for the uninfected/symptomatic/asymptomatic prediction (Figure S6). This is consistent with the difficulty of predicting disease severity.

In general, the most important species were selected by the model because of differences in relative abundance at baseline among uninfected/symptomatic/asymptomatic outcomes (Figure S7, S8). In rare cases, species presence/absence information was predictive. For example, *Ruminococcus gnavus,* is absent (near or below limit of detection) in most of the individuals who become infected, but present in many (but not all) of those who remain uninfected (Figure S7). Thus, there is no single, strong predictor of infection outcomes, but rather a probabilistic combination of many species, each of relatively modest predictive value.

Identification of functional biomarkers of disease susceptibility and severity

We also identified gene families in the gut microbiome of persons who remained uninfected during follow-up (Figures S9 and S10), with some of the top gene families involved

in DNA repair, transmembrane transporter activity, iron metabolism (indicated with asterisks in Figure 4), and genes of unknown function (Table S8). Long-chain fatty acid biosynthesis pathways (*e.g.* cis-vaccenate, gondoate and stearate) were associated with individuals who remained uninfected, while amino acid biosynthesis and catabolic pathways were associated with individuals who developed infection (Figures S11 and S12, Table S9). We identified three iron-related genes associated with remaining uninfected: (1) the ferric uptake regulator Fur, a major regulator of iron homeostasis, (2) thioredoxin, a redox protein involved in adaptation to oxidative and iron-deficiency stress, and (3) the TonB/ExbD/TolQR system, a ferric chelate transporter (Noinaj et al. 2010; Fillat 2014; Bi-ying Wang et al. 2019). In individuals who became infected and were asymptomatic, two genes involved in the conversion of riboflavin into catalytically active cofactors, the riboflavin kinase and the FAD synthetase, were found as the first and the third most discriminant features (Figure 4, Table S8).

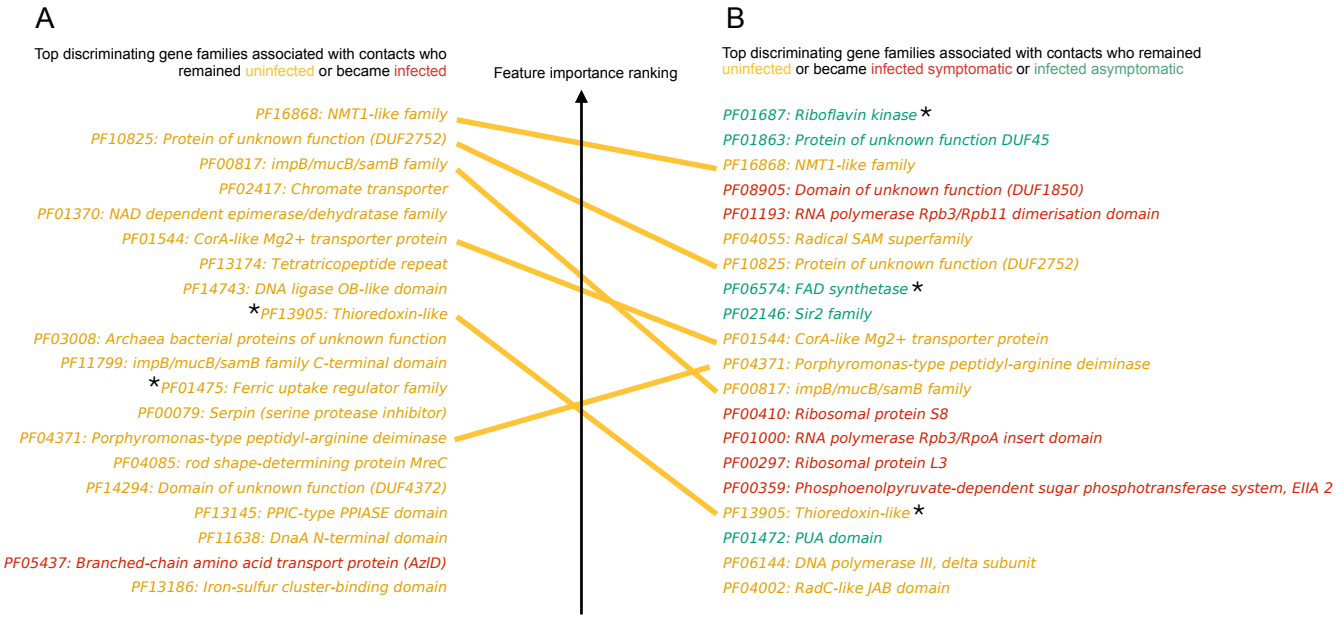


Figure 4. Most important discriminating gene families of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome.

(A) Genes families associated with contacts that became infected (red) or remained uninfected (yellow) during follow-up. (B) Genes families associated with contacts who remained uninfected (yellow), or became infected asymptomatic (green), or symptomatic (red) during follow-up. The top 25 most important features are shown here; see Table S8 for the full list. Yellow lines connect species associated with uninfected individuals in both (A) and (B). Asterisks indicate genes involved in redox or iron metabolism

We next asked which taxa in the microbiome likely encoded these genes. In some cases, specific taxonomic groups corresponded to discrete gene functions. For example, several iron metabolism-related gene families tend to be encoded by *Prevotella* genomes (Figure S14). In other cases, the major contributors to protective gene families were unclassified (Figures 5 and S13). These results partly explain why gene families or pathway features tend to outperform species-level features in predicting infection status – because predictive gene families are distributed across many species, including several with poor taxonomic annotation or families lacking representation in taxonomic databases.

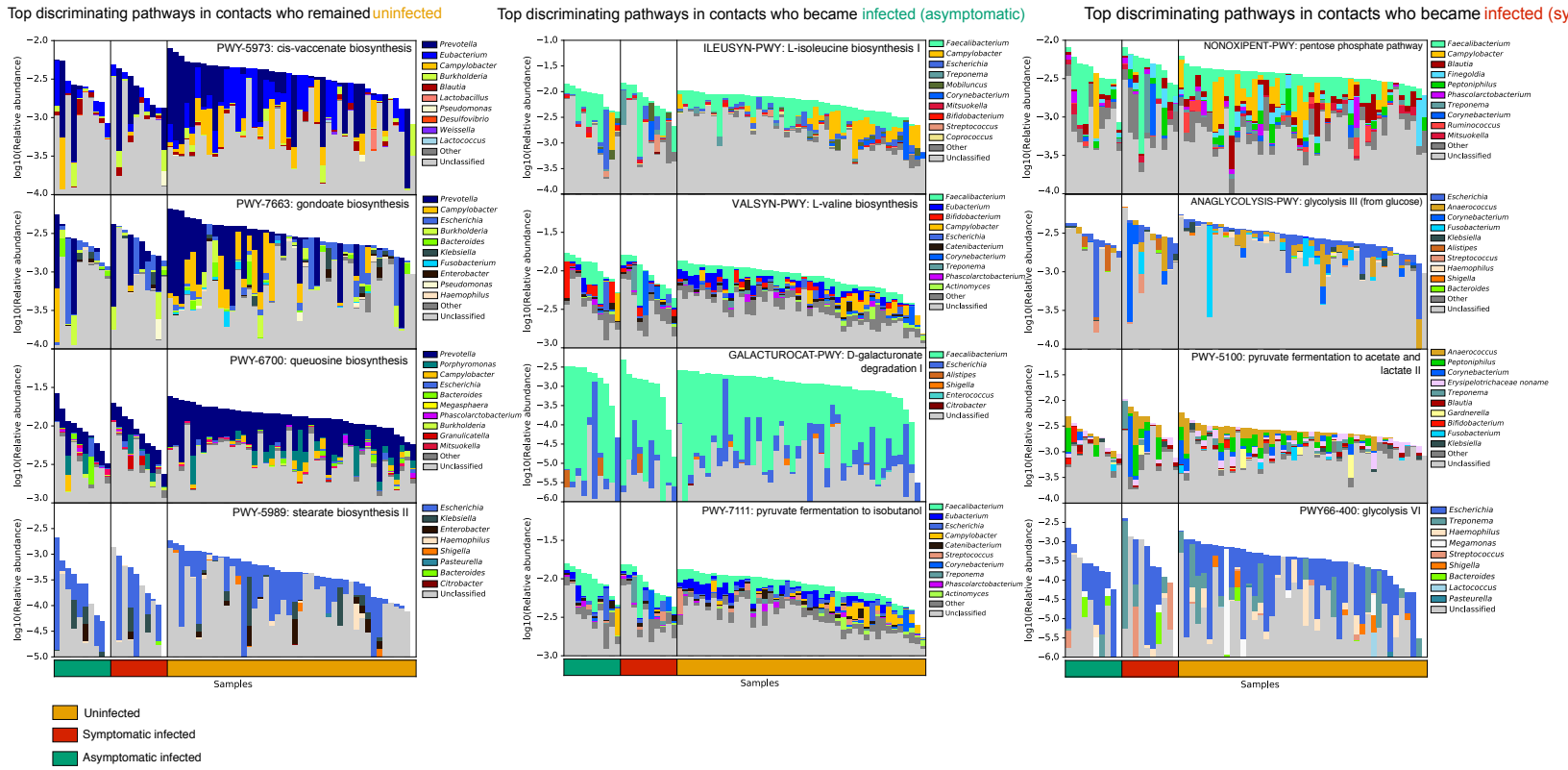


Figure 5. Top predictive cellular pathways of the gut microbiome at the time of exposure to *V. cholerae* in the Midani 2018 cohort, annotated by their taxonomic contributors. The four top-ranked pathways associated with uninfected contacts (left column), contacts who developed asymptomatic infection (middle), and contacts who developed symptomatic infection (right column) are shown. Total bar height reflects $\log_{10}$-scaled community relative abundance of each pathways. The contributions of each genus to encoding these pathways are shown as stacked colors within each bar, linearly scaled within the total. See Table S9 for the complete list of pathways

Discussion

The gut microbiome is a potentially modifiable host risk factor for cholera, and identification of specific genes and strains correlated with susceptibility is needed for experimental testing to understand the mechanisms of observed correlations. Compared to a previous study using a single marker gene, shotgun metagenomics provides this degree of resolution, potentially to the species and strain level, and to the level of individual genes and cellular functions. We found that gene families in the gut microbiome at the time of exposure to *V. cholerae* were more predictive of susceptibility compared to taxonomic or clinical and demographic information. Selecting a subset of the most informative features further improved predictions, but using these selected features may lead to overfitting. This suggests an upper limit to predictive power that requires validation in larger, independent cohorts.

Using a machine learning method to identify the most important factors contributing to our model, we selected 30 bacterial species from 65 samples to estimate which contacts became infected, and predicted outcomes with similar success rates as previously reported with 16S amplicon sequencing data (Midani et al. 2018). Prediction of infection was substantially improved by using gene families or metabolic pathways, highlighting the benefits of using richer metagenomic data. Selecting a subset of the most informative features further improved predictions, but using these selected features may lead to overfitting. This suggests an upper limit to predictive power that requires validation in larger, independent cohorts.

Most of the top predictive biomarkers (using both species and gene families) were associated with remaining uninfected after exposure to *V. cholerae*, and many of these biomarkers were consistently identified (Figures 3 and 4). An example is the genus *Prevotella,* including several strains within *Prevotella* sp. 885, identified only at the genus level in a previous study (Midani et al. 2018). *Prevotella* species are hypothesized to be beneficial members of the microbiota in healthy individuals in non-Westernized countries, and this species is a potential candidate for follow up experimental studies in *V. cholerae* susceptibility (David et al. 2015; Kovatcheva-Datchary et al. 2015; Tett et al. 2019).

Several species known to ferment mucin glycans into short chain fatty acids (SCFAs) correlated with remaining uninfected, including *Eubacterium rectale*, *Ruminococcus gnavus* and *Bacteroides vulgatu*s (Crost et al. 2013; Tailford et al. 2015). This finding is consistent with experiments of SCFAs applied to animal models. *B. vulgatus* has been shown to inhibit

V. cholerae colonization in mice, an effect that was dependent upon SCFAs butyrate and propionate production (You et al. 2019). SCFAs are known to impact immune cell development and attenuate inflammation by inhibiting histone deacetylases and other mechanisms of altering gene expression (Sun et O'Riordan 2013; Koh et al. 2016; Louis et Flint 2017; Fachi et al. 2019).

All three *Bifidobacterium* species associated with contacts that developed infection were also associated with asymptomatic rather than symptomatic disease (Figure 3), and prior work on this genus supports several hypotheses for this relationship. First, *Bifidobacteria* are known to produce the SCFA acetate that can protect against enteric infection in mice (Rabbani et al. 1999; Fukuda et al. 2011; Sepúlveda Cisternas, Salazar, et García-Angulo 2018). SCFAs are also known to inhibit cholera toxin-related chloride secretion in the mouse gut, reducing water and sodium loss, and have also been observed to increase cholera toxin-specific antibody responses (Rabbani et al. 1999; Canani et al. 2011; W. Yang et al. 2019). *Bifidobacteria* are also major producers of lactate, a metabolite that has been shown to impair *V. cholerae* biofilm formation, a function that can impact pathogen virulence (Kaur et al. 2018). Lastly, *B. bifidum* and *B. adolescentis* are known to reduce the activity of the *V. cholerae* type VI secretion system through modification of bile acids (Bachmann et al. 2015).

Metagenomics also allowed us to identify bacterial functions that could impact the ability of *V. cholerae* to compete and colonize the gut. For example, several gene families involved in iron transport, iron regulation, and riboflavin conversion appeared among the top twenty features associated with uninfected and asymptomatic individuals, suggesting that competition for iron might be a protective mechanism of the gut microbiota against *V. cholerae*, as is the case for other pathogens (Ubeda, Djukovic, et Isaac 2017). Iron is often a limiting redox cofactor in the gut, and bacteria have evolved strategies to solubilize and internalize iron (Sepúlveda Cisternas, Salazar, et García-Angulo 2018, Rivera-Chávez et Mekalanos 2019). Riboflavin (another major redox cofactor in bacteria) and iron levels are reciprocally regulated in *V. cholerae*, and more generally, riboflavin may allow *V. cholerae* to overcome iron limitation in the gut (Sepúlveda-Cisternas et al. 2018; Rivera-Chávez et Mekalanos 2019). A gut microbiota more competitive for iron could be an important factor in resistance to *V. cholerae* colonization or virulence. Further work is thus needed to

understand mechanisms of how the enrichment of these microbiome genes may protect people after exposure to *V. cholerae*.

Our results are currently not generalizable beyond the study cohort in Dhaka, Bangladesh, since a similar cohort in another geographic location is not available. As with any association-based study (Sepúlveda-Cisternas et al. 2018; Rivera-Chávez and Mekalanos 2019), it is unknown if any of the metagenomic features that correlate with protection from *V. cholerae* infection are causal, and many may be markers of clinical or environmental factors that themselves impact susceptibility.

Despite our deep sequencing and collection of standard cholera risk factors, our study was unable to measure all potentially relevant environmental or clinical risk factors. In line with recent studies in Dhaka, we assume that *V. cholerae* transmission occurs mainly within households (Domman et al. 2018) and did not consider how the mode of transmission (*e.g.* waterborne or not) might affect outcomes. It has also been noted that microbiome-disease associations may be poorly portable across human populations (Schmidt, Raes, et Bork 2018). For instance, we identified species of *Prevotella* as protective features in Bangladesh, but *Prevotella* is much less abundant and less diverse in Western countries (Tett et al. 2019). It thus remains to be seen if protective gene features (*e.g.* iron metabolism) are encoded in other species of the microbiome outside endemic areas like Bangladesh, or if people outside these areas are simply at greater risk for cholera. Further experimental characterization of metagenomic features correlated with protection from infection or symptoms are needed to understand if factors we identified impact *V. cholerae* pathogenesis or host responses to infection. Ultimately, the strains and functionalities identified have the potential to inform microbiota-based therapeutics to ameliorate or prevent disease. Our results show the power of metagenomic data from the gut microbiome to predict health outcomes such as susceptibility to infection and disease severity.

Supplementary data

Supplementary materials and methods

Clinical outcomes and metadata collection

All study participants lived in Dhaka city, Bangladesh. Multiple household contacts could be enrolled from one household, and only one index case was enrolled per household. Household contacts were defined as individuals who shared the same cooking pot with an index patient for three or more days. Rectal swab samples were obtained from contacts beginning the day after enrollment of the index case. During the follow up period for contacts, two sets of rectal swabs were collected. The first set were collected during daily home visits, when report of symptoms was also collected, and was used for *V. cholerae* culture. This data was used for the classification of clinical outcomes. The second set of rectal swabs collected over the first ten days, and day 30, also at the home visit, were used for DNA extraction to conduct the microbiome analyses. Infection status was determined using the first set of rectal swab samples, serologic data from the blood draw, and report of symptoms collected during the observation period. Blood samples collected during the same days from contacts were drawn at the International Center for Diarrheal Disease Research, Bangladesh, in Dhaka, Bangladesh. Symptomatic *V. cholerae* infection was defined by a contact reporting diarrhea within 3 days of a rectal swab positive for *V. cholerae* during the follow-up home visits, or by a four-fold increase in vibriocidal titer with diarrhea during the follow-up period. Two symptomatic contacts were rectal swab negative for *V. cholerae* but showed a four-fold increase in vibriocidal titer and reported diarrhea, and then were determined to be infected based on our inclusion criteria. *V. cholerae* infection (rectal swab positive) was defined as asymptomatic if no diarrhea was reported. All contacts in our study that developed infection had had the same serotype of *V. cholerae* as the infected household case (either Inaba or Ogawa).

Midani and Expanded cohort description

We used metagenomic reads from 65 of the 76 contacts from a previously published study by our collaborators, and these contacts are referred to as the Midani 2018 cohort (Midani et al. 2018). Eleven of the contacts from this study had insufficient DNA to perform metagenomic sequencing. Each sample from this cohort was sampled on day 2, the day of

the enrollment of household contacts, one day after the presentation of the household index case at the hospital. To increase power in our study, we added household contacts samples from additional households, and also added samples from additional time points of Midani 2018 cohort contacts (see Table S1). This larger group of samples is referred to as the Expanded cohort, and is comprised of the 65 samples from the Midani 2018 cohort plus 33 additional samples from 16 infected contacts (Figure 1). The 16 infected new contacts in the expanded cohort have multiple samples from different days that are prior to when infection occurred, and therefore have multiple "pre-infection" samples (Table S1). Enrollment and sample collection were identical for the two groups of samples collected. In total, 129 household contacts were initially enrolled. Eighteen contacts were excluded due to a positive rectal swab at the enrollment evaluation, and three were excluded due to recent antibiotic use. Nine contacts reported ambiguous clinical outcomes based on clinical and history evaluation (i.e. report of diarrhea with no serologic or culture evidence of *V. cholerae* infection). Six additional enrollees were excluded due to report of diarrhea during the week prior to enrollment. Among the remaining contacts, ten resulted with DNA evidence of *V. cholerae* infection despite a rectal swab culture negative for *V. cholerae* at the time of enrollment. Lastly, 13 contacts were excluded due to failure to amplify DNA from rectal swabs or sequencing failure.

DNA extraction and sequencing

Fecal samples and rectal swabs were collected during home visit and immediately placed on ice and stored at -80°C until DNA extraction. For each sample, DNA was extracted as previously described (Midani et al. 2018). Briefly, samples were processed using PowerSoil DNA extraction kits (Qiagen) after pre-heating to 65°C for 10 min and to 95°C for 10 min. A total of 98 samples were selected for library construction with NEBNext Ultra II DNA library prep kit and sequenced on the Illumina HiSeq 2500 (paired-end 125 bp) and the Illumina NovaSeq 6000 S4 (paired-end 150 bp) at the Genome Québec sequencing platform (McGill University).

Sequence preprocessing

Sequencing fastq files were quality checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For removal of human and technical contaminant DNA, metagenomic shotgun sequences were aligned to the PhiX genome and the human genome (hg19) with Bowtie2 (Langmead et Salzberg 2012). Host-decontaminated fastq files were then quality filtered using Trimmomatic, a computational tool for removing low quality data (Bolger, Lohse, et Usadel 2014), discarding all reads shorter than 75 nucleotides and with quality Phred score less than 20.

Taxonomic and functional profiling

We performed taxonomic profiling of archaea, bacteria and microbial eukaryotes using MetaPhlAn2 (version 2.9) (Truong et al. 2015), with default parameters. Briefly, MetaPhlAn2 maps shotgun metagenomic sequencing reads against a precomputed database of clade-specific markers to produce a robust estimate of the taxonomic clades present in the microbiome sample and estimate their relative abundance. In addition to species-level relative abundance, we used MetaPhlAn2 to characterize the presence/absence of strain-specific markers for each sample. Species abundances are real numbers in the range [0,1] while markers assume binary values.

Species-resolved functional profiling was completed using HMP Unified Metabolic Analysis Network 2 (HUMAnN2, version 2.8.1) (Franzosa et al. 2018), which maps reads to a sample-specific reference database from the pangenomes of the subset of species detected in the sample by MetaPhlAn2, quantifying gene presence and abundance on a per-species basis. A translated search is then performed against a UniRef-based protein sequence catalog for reads that fail to map to one of the detected species. The results are abundance profiles of gene families (UniRef90s) stratified by species contributing to those genes, which can be regrouped in higher level grouping categories (Pfam domains in this data set). Finally, gene families were further analyzed to reconstruct metabolic pathways abundance according to the MetaCyc pathway catalog (Franzosa et al. 2018).

Statistical analyses

Univariate analyses was performed to identify discriminatory biomarkers (species, protein families and pathways) using linear discriminant analysis (LDA) effect sizes (LEfSe)

(Segata et al. 2011). LEfSe first uses the non-parametric factorial Kruskal-Wallis test to detect features with significant differential abundance for each class of interest, then uses LDA to estimate the effect size of each differential abundant feature (Segata et al. 2011). In our case, the significance threshold was set at $p = 0.05$ and an LDA effect size of 2 was used for discriminative features.

Random-forest based machine learning approach

MetAML (Pasolli et al. 2016) was applied to four types of quantitative profiles: taxonomic species-level relative abundances, strain-specific markers presence/absence patterns, MetaCyc pathways and gene-families grouped by Pfam domains relative abundances. In each case, we used Random Forest (RF) as classifier and set the following parameters as previously described by (Thomas et al. 2019): the number of estimator trees was equal to 1000 and the quality of split at each tree node was evaluated using Shannon entropy. The minimum number of samples per leaf and the number of features per tree were respectively set as 1 and 30%, except for the strain-specific markers presence/absence profile, where the number of features equal the square root of the total number of features, due to the higher number of features. One of the advantages of RF models is that they can intrinsically integrate binary and multi-class classification problems, and implicitly provide a list of the most informative features to the predictive model. Another advantage of RFs is the built-in feature selection step during the model generation phase, which allows a selection of a reduced subset of the most important features for discriminating between classes. Adding a feature selection step to a model is a useful way to remove irrelevant features, especially in datasets with high dimensionality. Feature selection can also reduce overall training time and the risk of overfitting (Pasolli et al. 2016).

The feature relative importance values were used to perform an embedded feature selection strategy, implemented as described in Pasolli et al. (2016), in addition to the RF model, for two (Uninfected vs Infected contacts) and three classes (Asymptomatic infected vs Symptomatic infected vs Uninfected contacts). Feature selection benefits include removal of irrelevant features, especially in datasets with high dimensionality, to decrease the overall training time and reduce the risk of model overfitting (Zhou et Gallins 2019). Each cohort (the Midani_2018 and the expanded cohort) was analyzed independently for the RF model,

using a stratified cross-validation approach. In the feature selection model, we chose the minimum number of features that maximized the accuracy in order to generate a final model on this limited number of features (Figure S2). The same set of selected features determined by the Midani_2018 dataset was used for the two cohorts in this approach. In all cases, the sensitivity and accuracy of the models generated were tested by performing stratified 3-fold cross validations, repeated and averaged on 20 independent runs. Finally, the results obtained for the original classification problem were compared with those obtained by a random classifier (denoted in the paper as Shuffled). For this purpose, we shuffled randomly the labels of all the samples, and used these same settings. Difference of performance between classifiers and statistical significance were calculated as described in Pasolli et al. (2016).

The following performance metrics were reported: 1) the overall accuracy (i.e., the proportion of outcomes correctly predicted), 2) the precision (i.e., the number of correct positive samples divided by the number of samples predicted as positive), 3) the recall (i.e. the true positive rate, or power), and 4) the F1 score, which is the harmonic mean of precision and recall. For binary classification problems, class posterior probabilities were used to plot the Receiver Operating Characteristic (ROC) curve, which represents the true positive rate (i.e., the recall) against the false positive rate (i.e., the number of wrong positive samples divided by the total number of non-positive samples). From the ROC curve, the program computes the area under the curve (AUC), which can be interpreted as the probability that a randomly selected positive sample will have a higher classification result than a randomly selected negative one. The AUC ranges in [0, 1], where 0.5 corresponds to random prediction.

Supplementary tables

Supplementary tables are available on Github:
https://github.com/ilevade/Metagenomics_vibrio_cholerae_susceptibility_ML

Supplementary figures

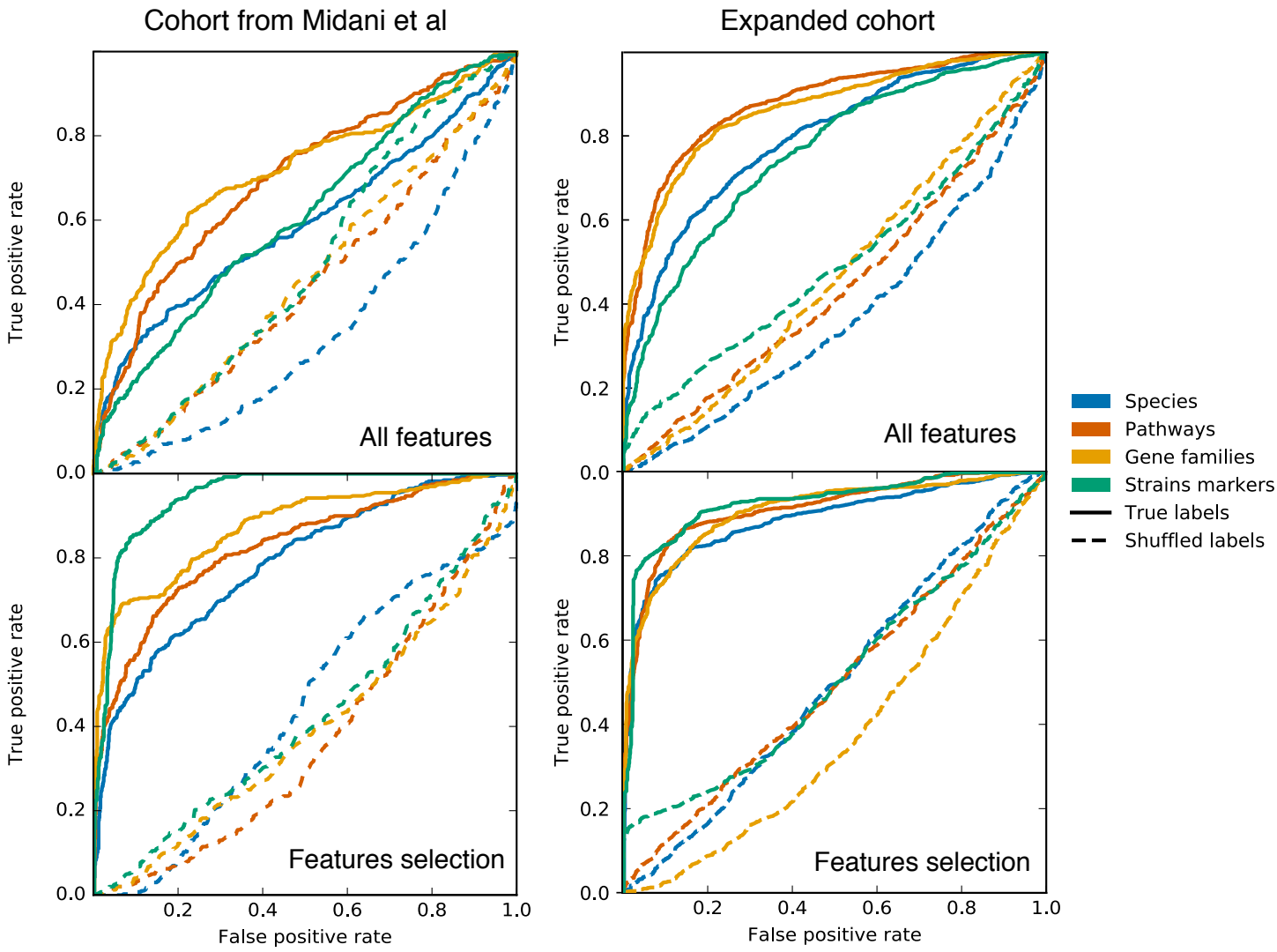


Figure S1. Classification result from species abundance, strain-level markers presence/absence, pathways and gene family abundances. Four different classifications of microbiome information were used to predict disease susceptibility (Uninfected vs Infected) using a random forest algorithm and stratified 3-fold cross validation. Plots show average ROC curves by using the four type of features for two different datasets (the Midani 2018 data and the expanded dataset). Random Forest was applied on all features (top row) and on a set of selected features (bottom row).

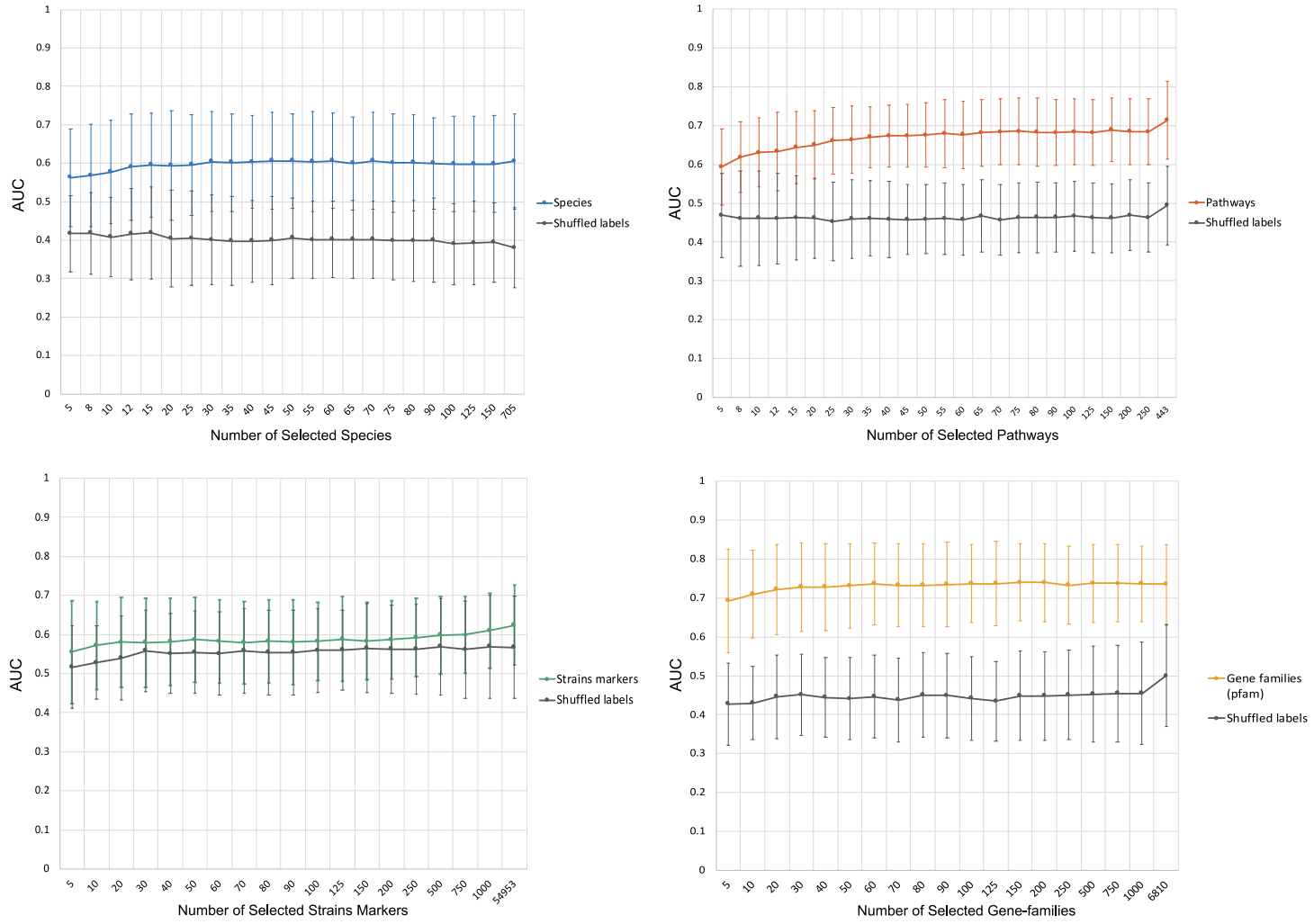


Figure S2. Identification of a minimal number of features for each type of biomarker. Prediction performances (AUC values) at increasing number of microbial species, pathways, strains markers and genes families, obtained by retraining the random forest model on the top-ranking features identified with a first random forest model training with a 3-fold cross-validation approach.

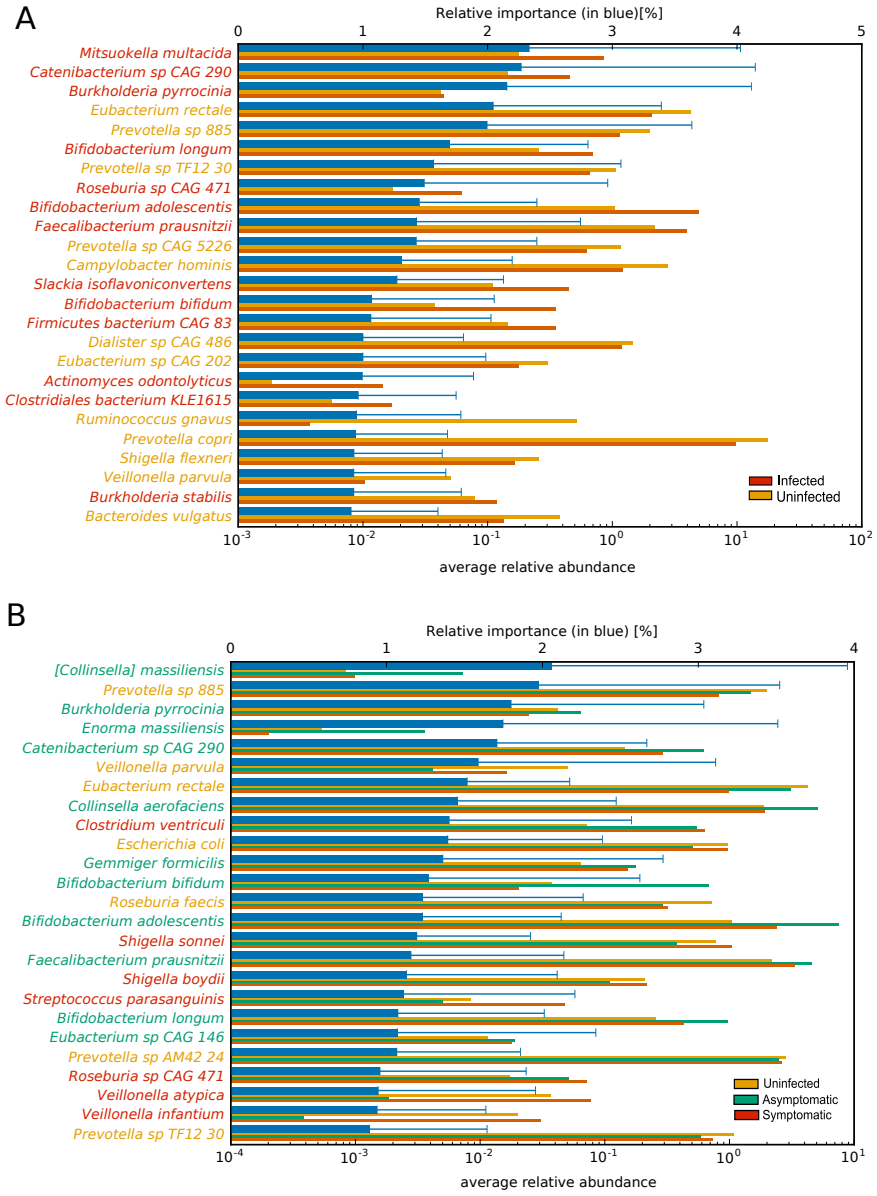


Figure S3. Most important discriminating species of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome.

(A) Species associated with contacts that became infected (or remained uninfected) during follow-up. (B) Species associated with contacts that became uninfected/asymptomatic/symptomatic during follow-up. For each species reported on the vertical axis, the top bar (blue) corresponds to the species relative importance (with standard deviation) and the other bars refer to the average relative abundance. The top 25 most important features are shown here; See Table S6 for the full list. Feature relative importance was computed using the Mean Decrease in Impurity strategy, as described in the Methods.

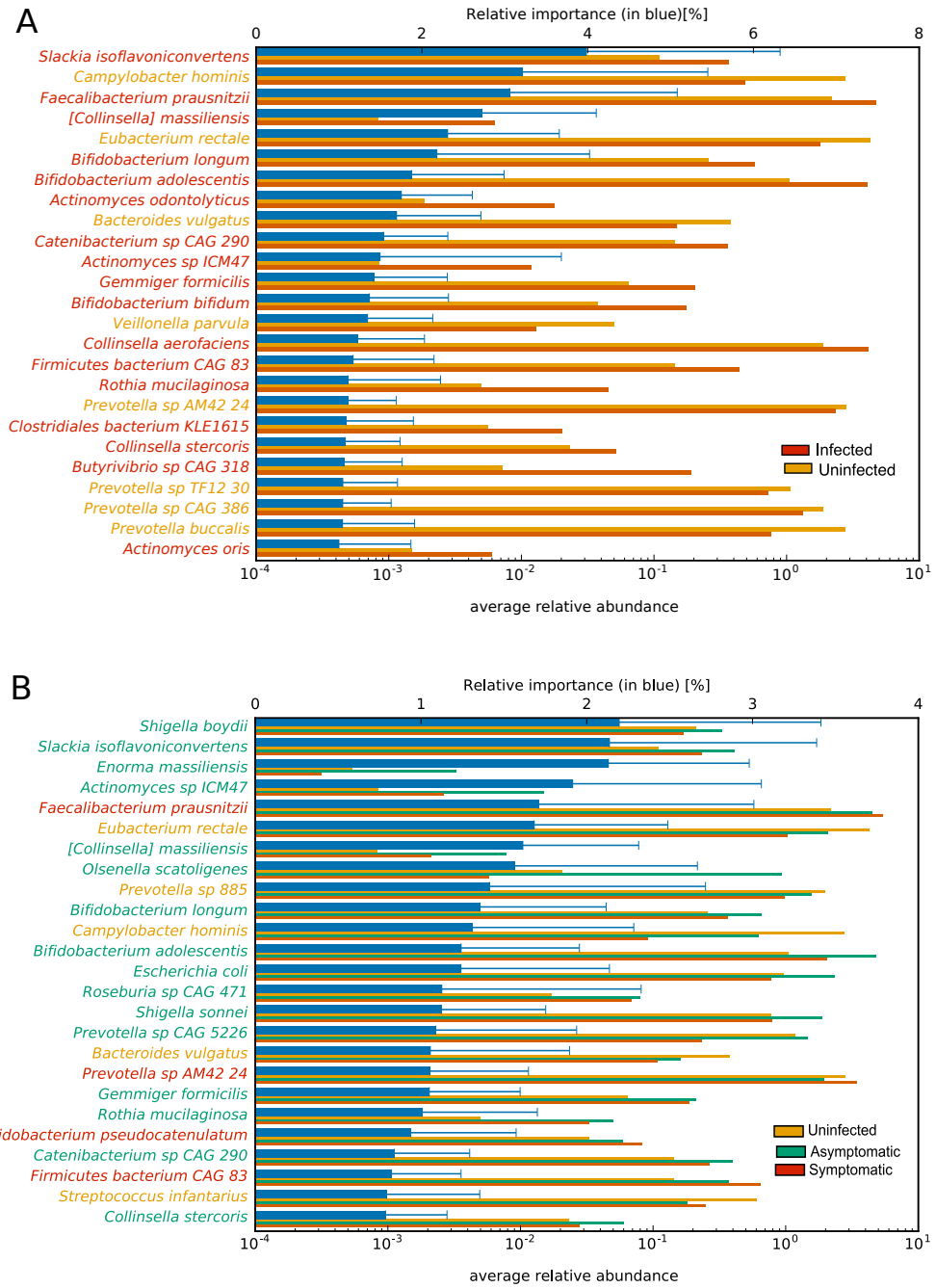


Figure S4. Most important discriminating species identified with the random forest algorithm on the expanded dataset for (A) contacts that became infected or remained uninfected and (B) contacts that remained uninfected, or became infected and were asymptomatic vs symptomatic. For each species reported on the vertical axis, the top bar (in blue) corresponds to the feature relative importance (with standard deviation) and the other bars refer to the average relative abundance. Feature relative importance (blue) is computed using the mean decrease impurity strategy.

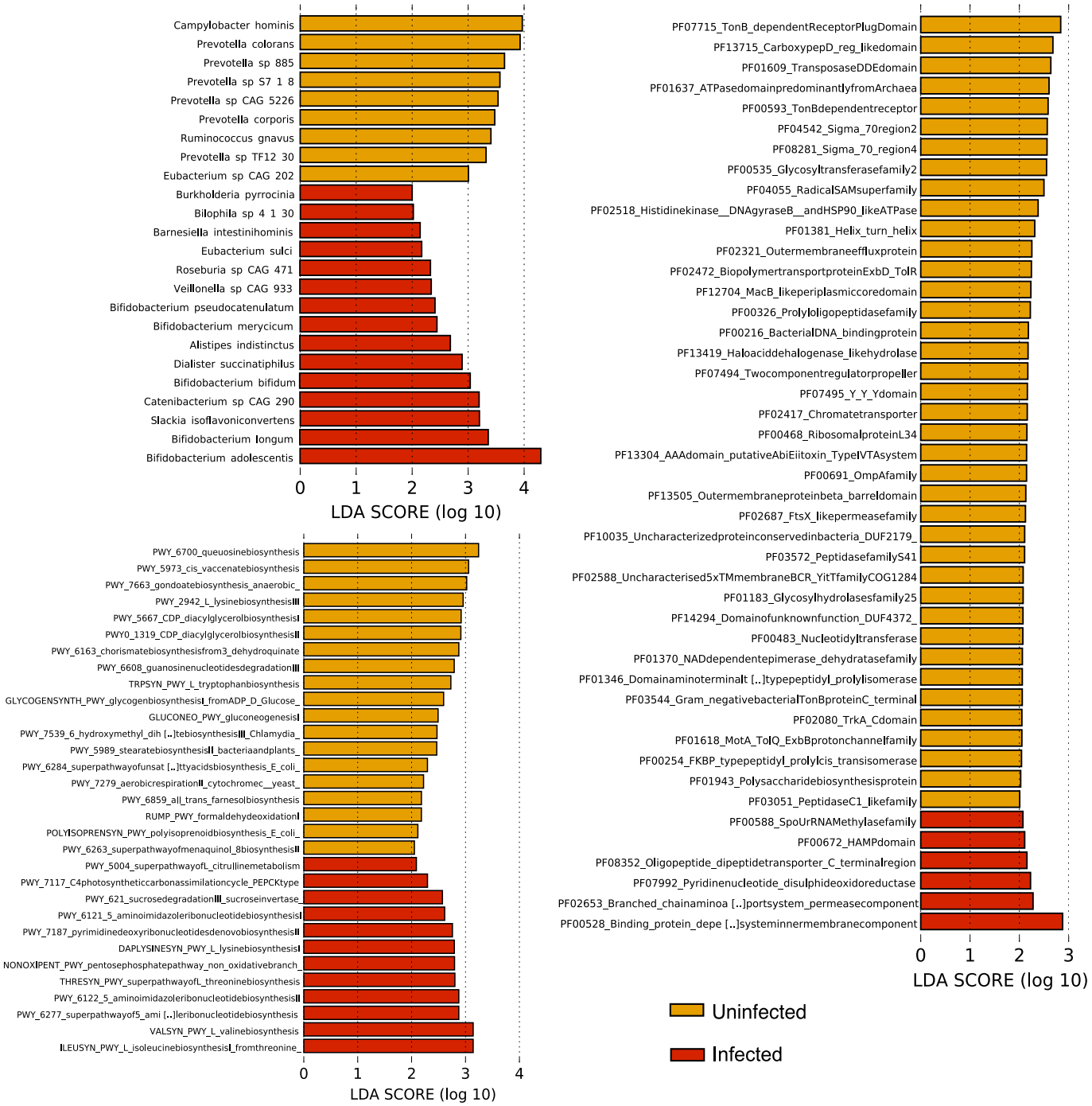


Figure S5. Linear discriminant analysis (LDA) scores computed for differentially abundant species, pathways and genes families in the fecal microbiomes of samples from the Midani 2018 dataset for two categories. The cohort is divided in two categories: controls who remained uninfected (yellow) and controls who became infected (red). Length of the bar indicates effect size associated with a feature. $p = 0.05$ for the Kruskal-Wallis H statistic; features with LDA score > 2 are shown.

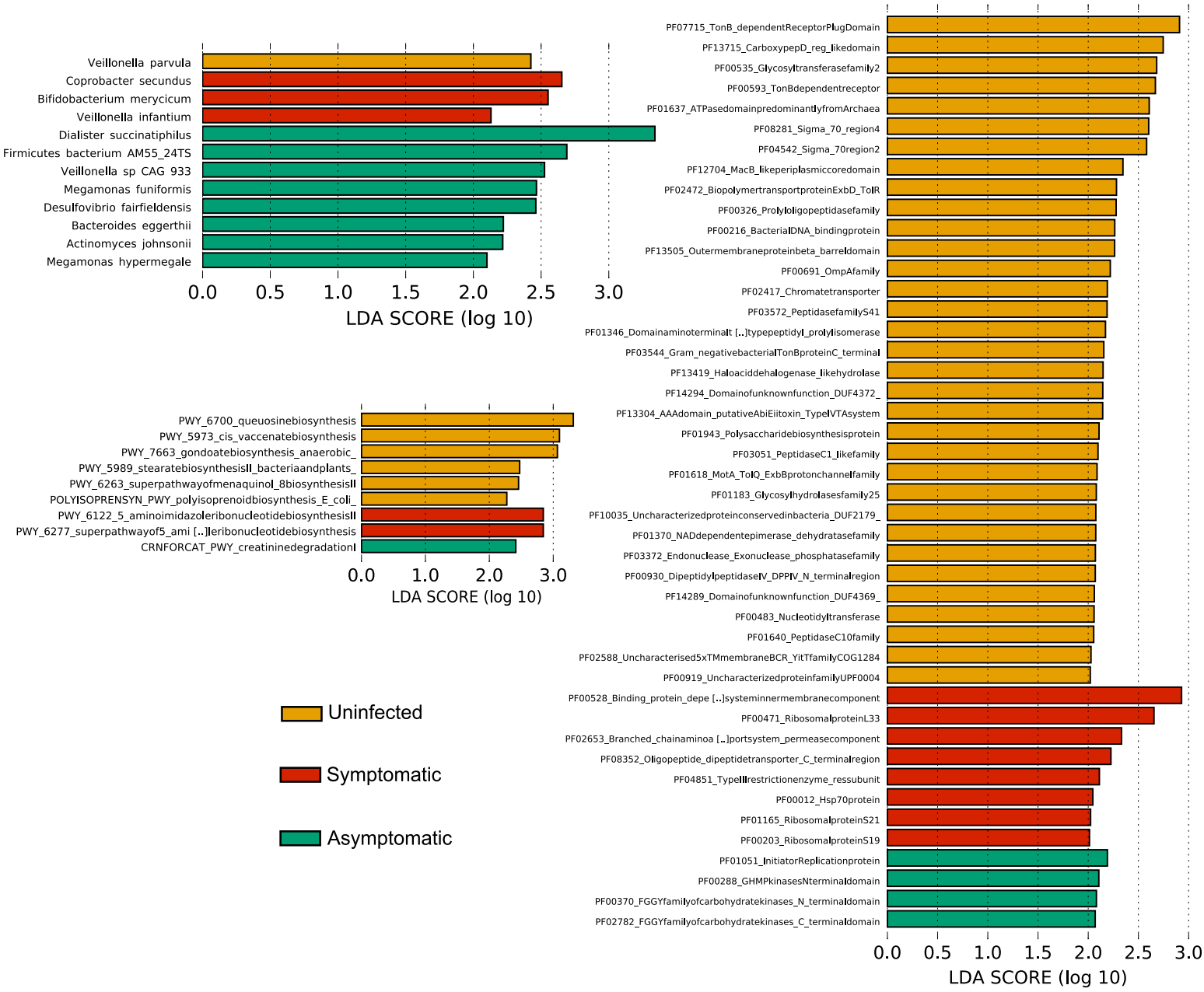


Figure S6. Linear discriminant analysis (LDA) scores computed for differentially abundant species, pathways and genes families in the fecal microbiomes of samples from the Midani 2018 dataset for three categories. The cohort is divided in three categories: controls who remained uninfected (yellow), controls who became infected and symptomatic (red), and controls who became infected but stayed asymptomatic (green). Length indicates effect size associated with a feature. $p = 0.05$ for the Kruskal-Wallis H statistic; features with LDA score > 2 are shown.

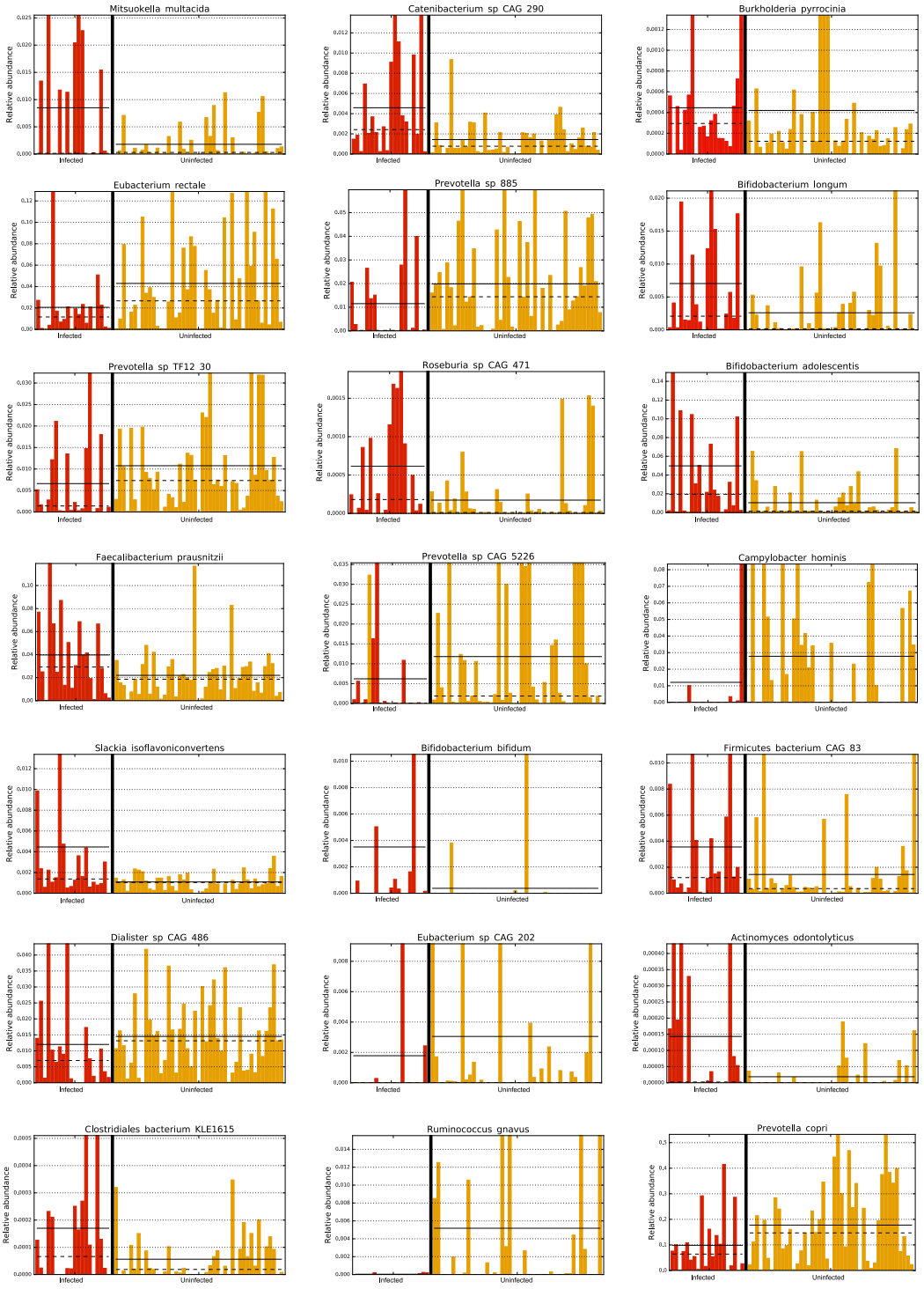


Figure S7. Relative abundance of the top 21 most important discriminating species of the gut microbiome of contacts at the time of exposure to V. cholerae identified in the Midani 2018 dataset for two classes (Uninfected vs Infected). The straight line indicates the group

means, and the dotted line indicates the group medians.

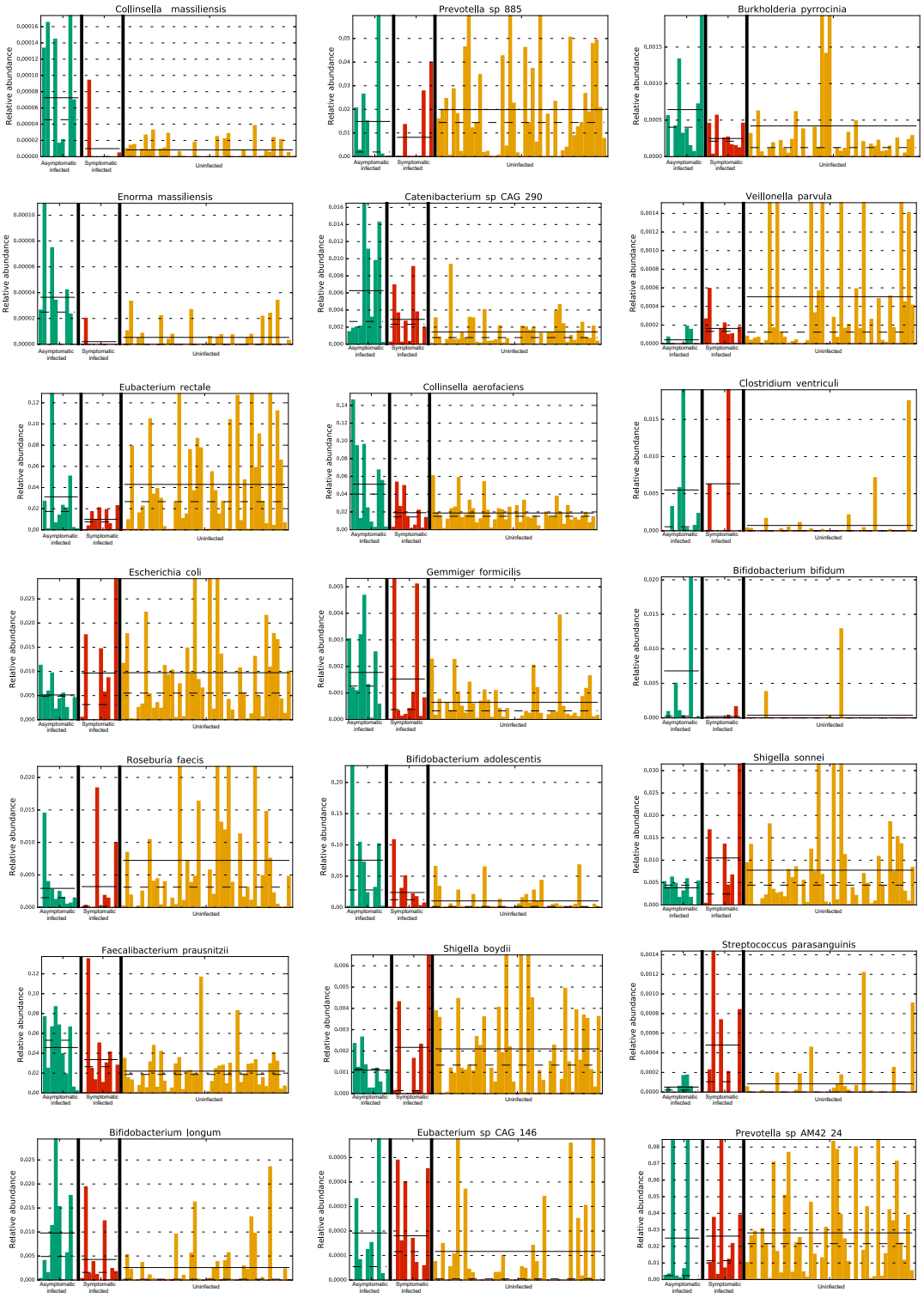


Figure S8. Relative abundance of the top 21 most important discriminating species in the gut microbiome of contacts at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset for three classes (Uninfected vs Asymptomatic Infected and Symptomatic

Infected). The straight line indicates the group means, and the dotted line indicates the group medians

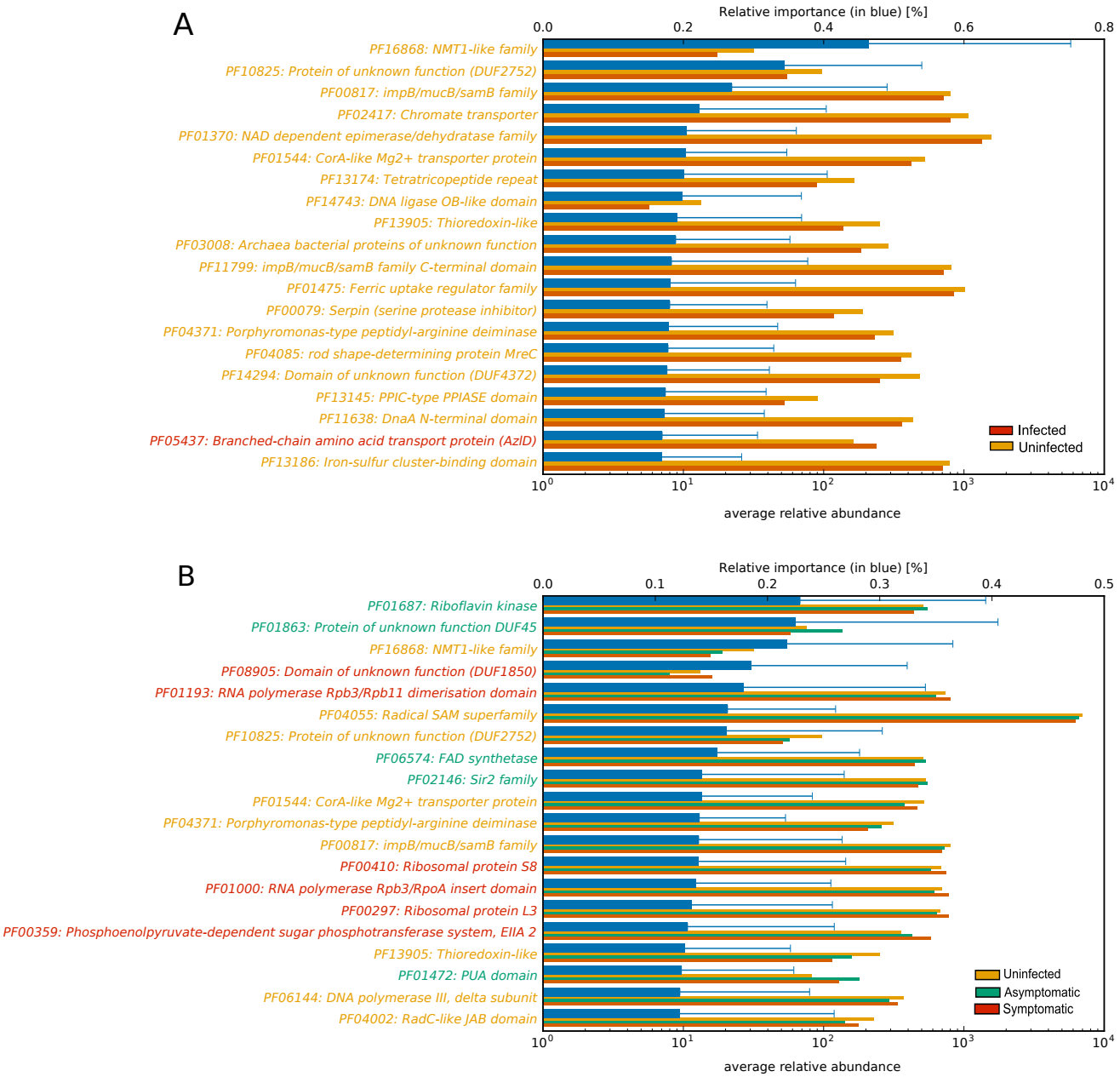


Figure S9. Most important discriminating gene families of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Genes associated with contacts that became uninfected/infected during follow-up. (B) Genes associated with contacts that became uninfected/asymptomatic/symptomatic during follow-up. For each gene-family reported on the vertical axis, the top bar (in blue) corresponds to the feature relative importance (with standard deviation) and the other bars refer to the average relative abundance (copies per million). The top 25 most important features are shown here; See Table S8 for the full list.

Feature relative importance was computed using the mean decrease in impurity strategy, as described in the Methods.

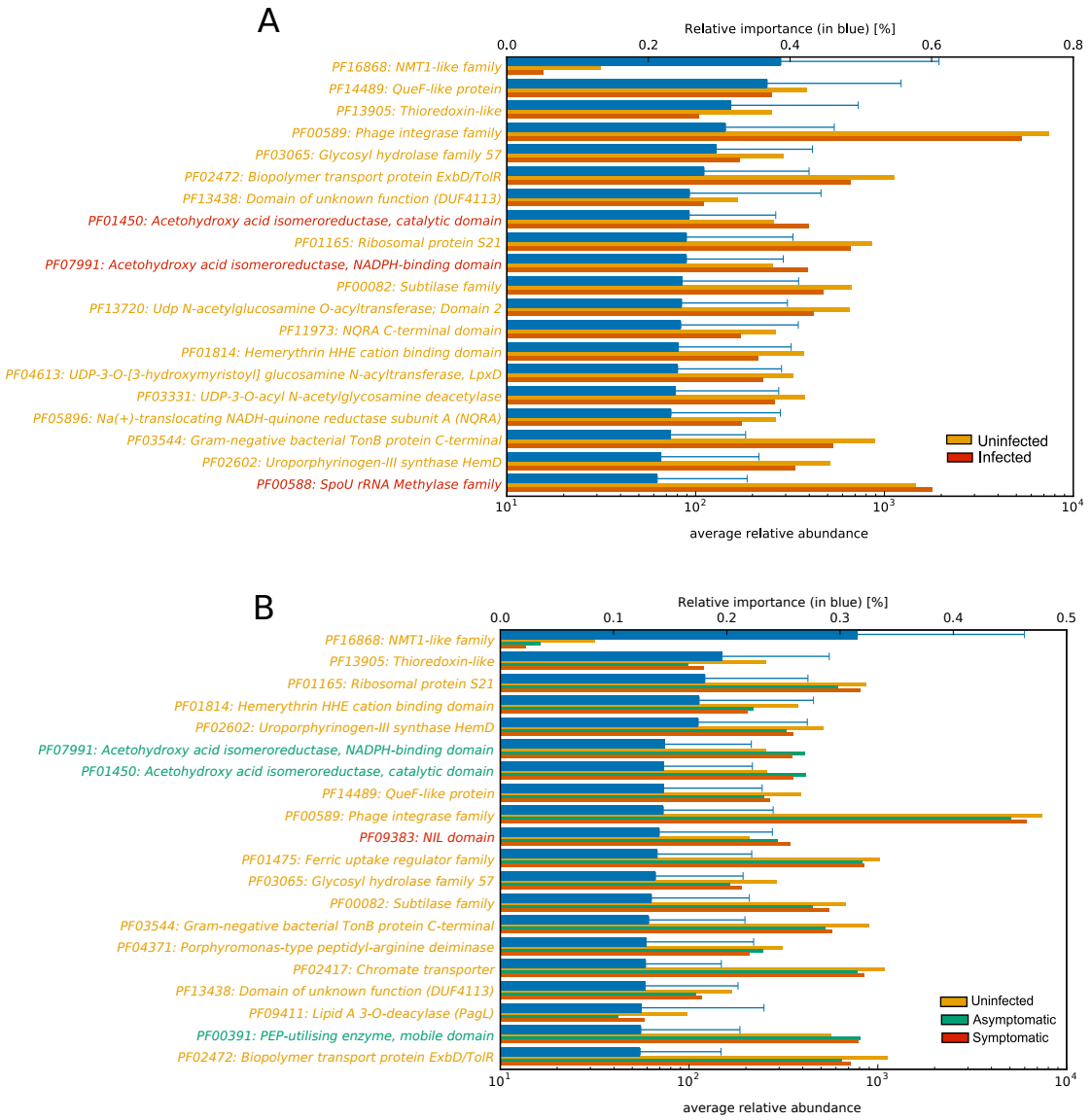


Figure S10. Most important discriminating gene-families (grouped by Pfam domain) identified with the random forest algorithm on the Expanded dataset for (A) contacts that became infected or remained uninfected and (B) contacts that remained uninfected, or became infected and were asymptomatic vs symptomatic. For each gene-families reported on the vertical axis, the top bar (in blue) corresponds to the feature relative importance (with standard deviation) and the other bars refer to the average relative abundance (copies per million). Feature relative importance (blue) is computed using the mean decrease impurity strategy.

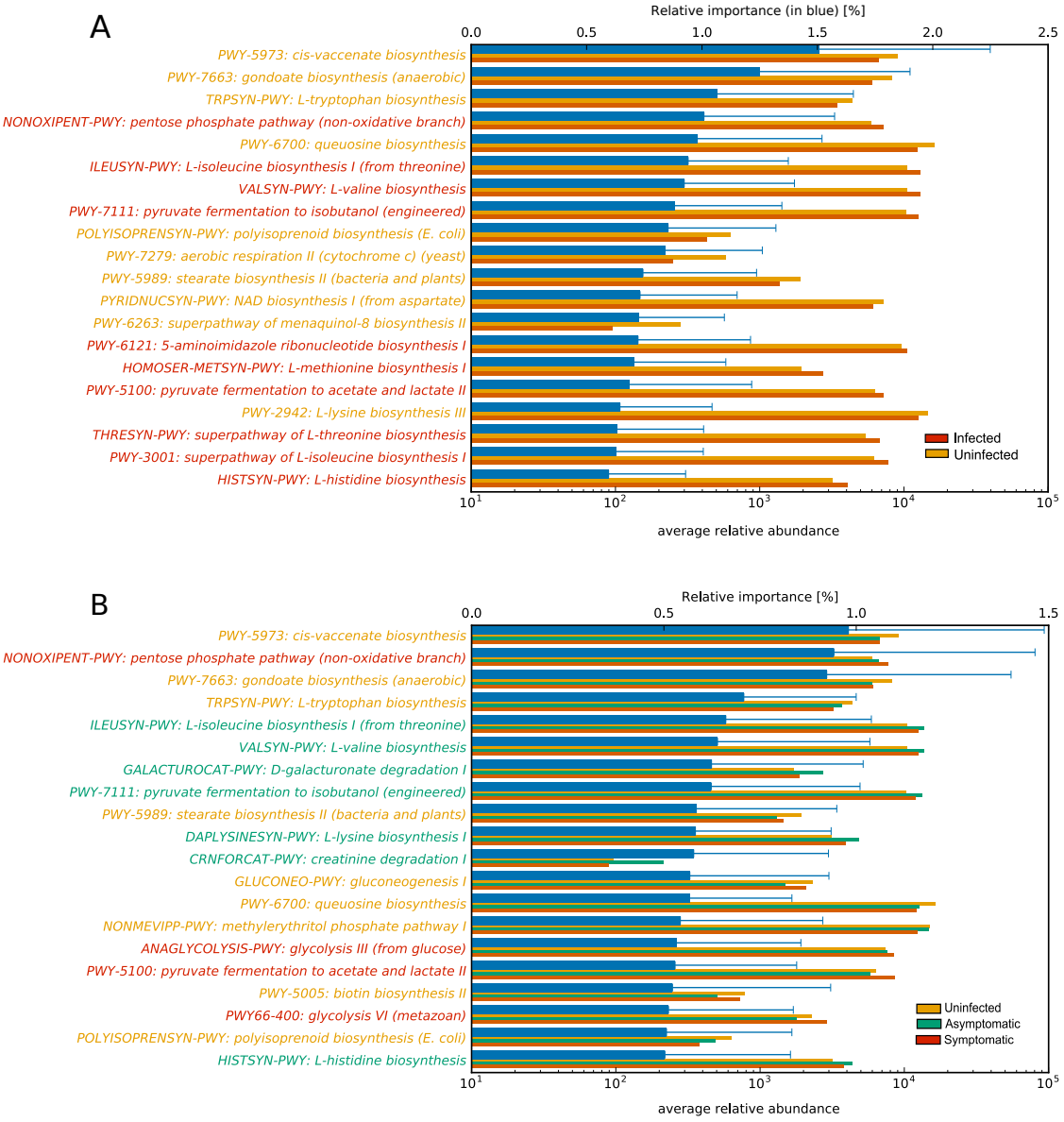


Figure S11. Most important discriminating pathways identified with the random forest algorithm on the Midani 2018 dataset for (A) contacts that became infected or remained uninfected and (B) contacts that remained uninfected or became infected and were asymptomatic vs symptomatic. For each pathway reported on the vertical axis, the top bar (blue) corresponds to the feature relative importance (with standard deviation) and other bars refer to the average relative abundance (copies per million). Feature relative importance (blue) is computed using the mean decrease impurity strategy.

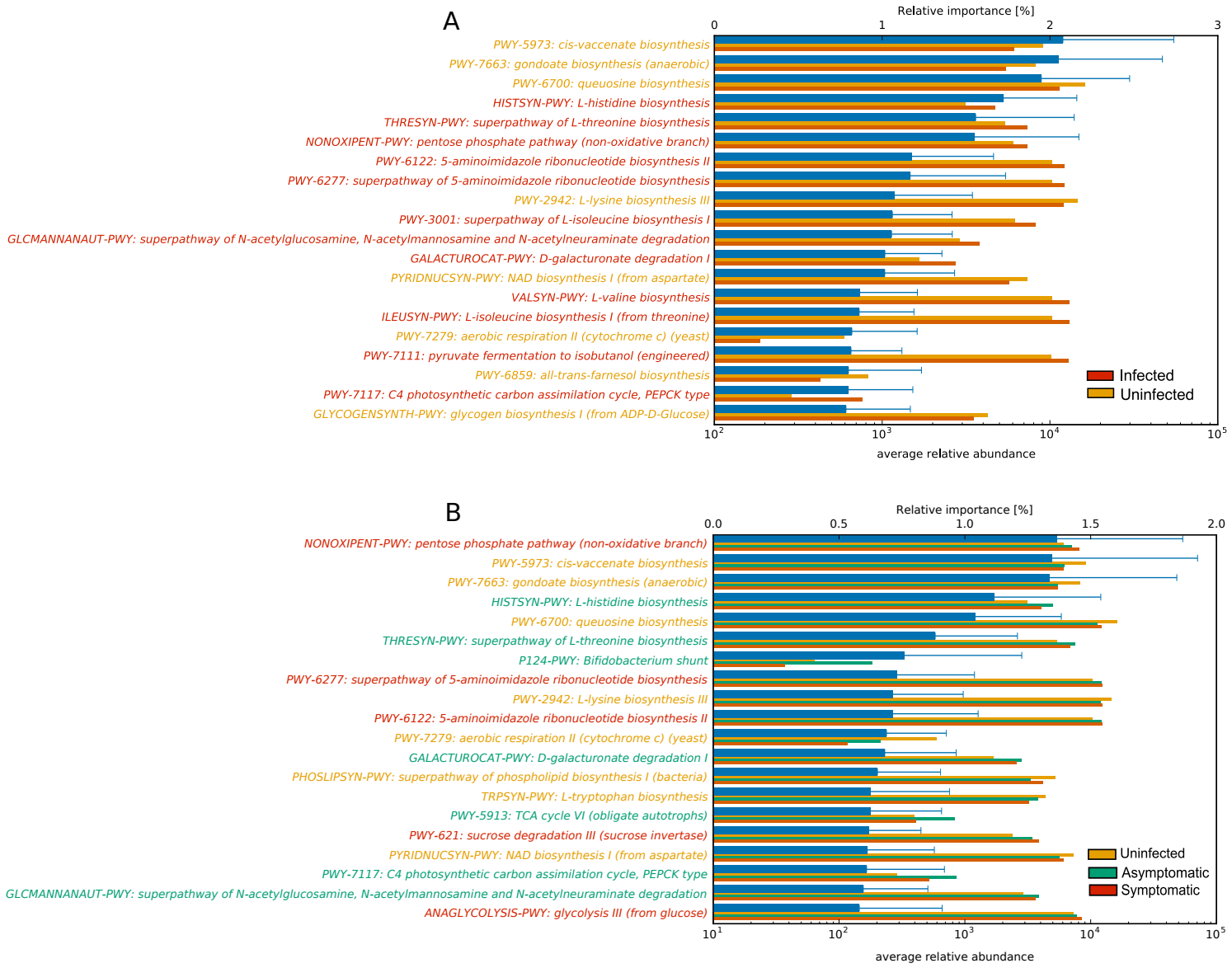


Figure S12. Most important discriminating pathways identified with the random forest algorithm on the expanded dataset for (A) contacts that became infected or remained uninfected and (B) contacts that remained uninfected, or became infected and were asymptomatic vs symptomatic. For each pathway reported on the vertical axis, the top bar (in blue) corresponds to the feature relative importance (with standard deviation) and the other bars refer to the average relative abundance (copies per million). Feature relative importance (blue) is computed using the mean decrease impurity strategy.

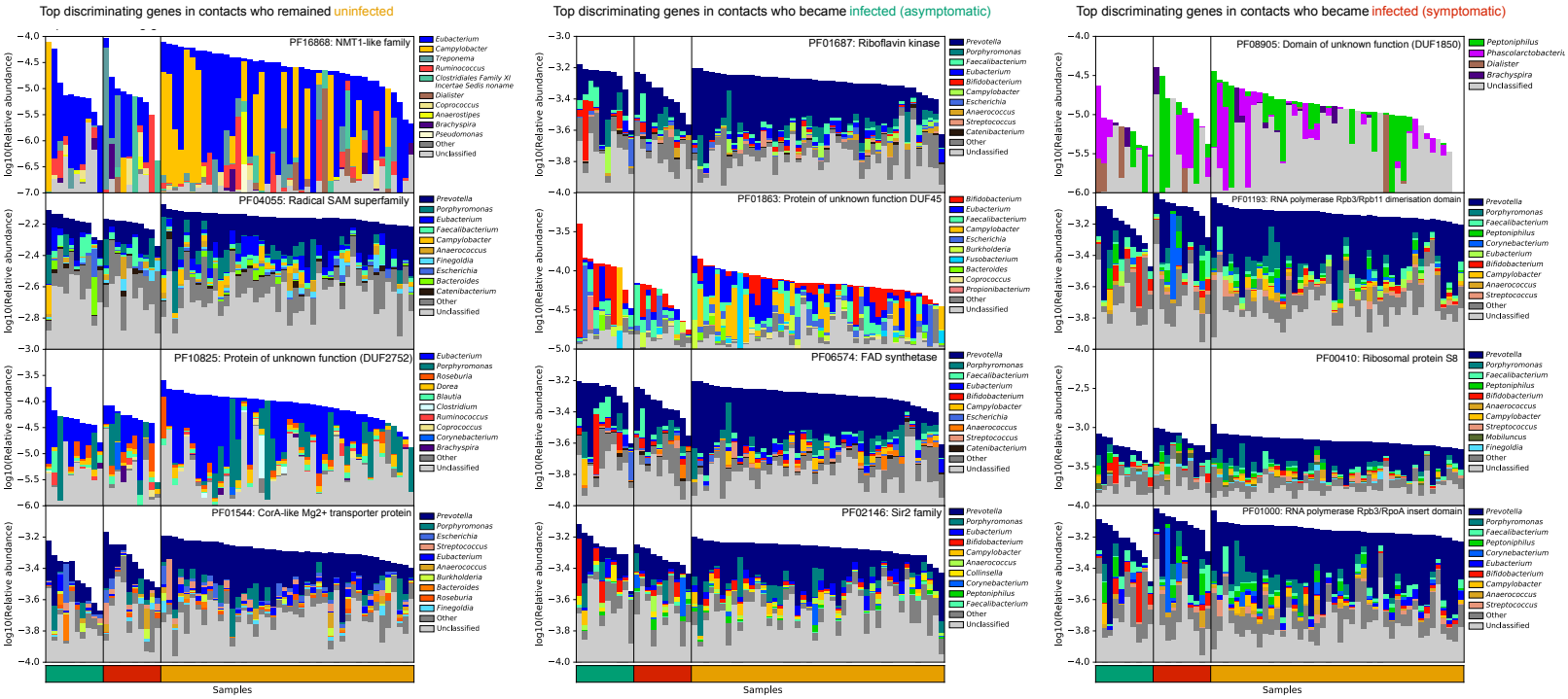


Figure S13. Top predictive gene families (Pfam domain grouping) for each class (uninfected, asymptomatic and symptomatic infected contacts), annotated by their taxonomic contributors. Total bar height reflects log10-scaled community abundance. Genera contributions are linearly scaled within total.

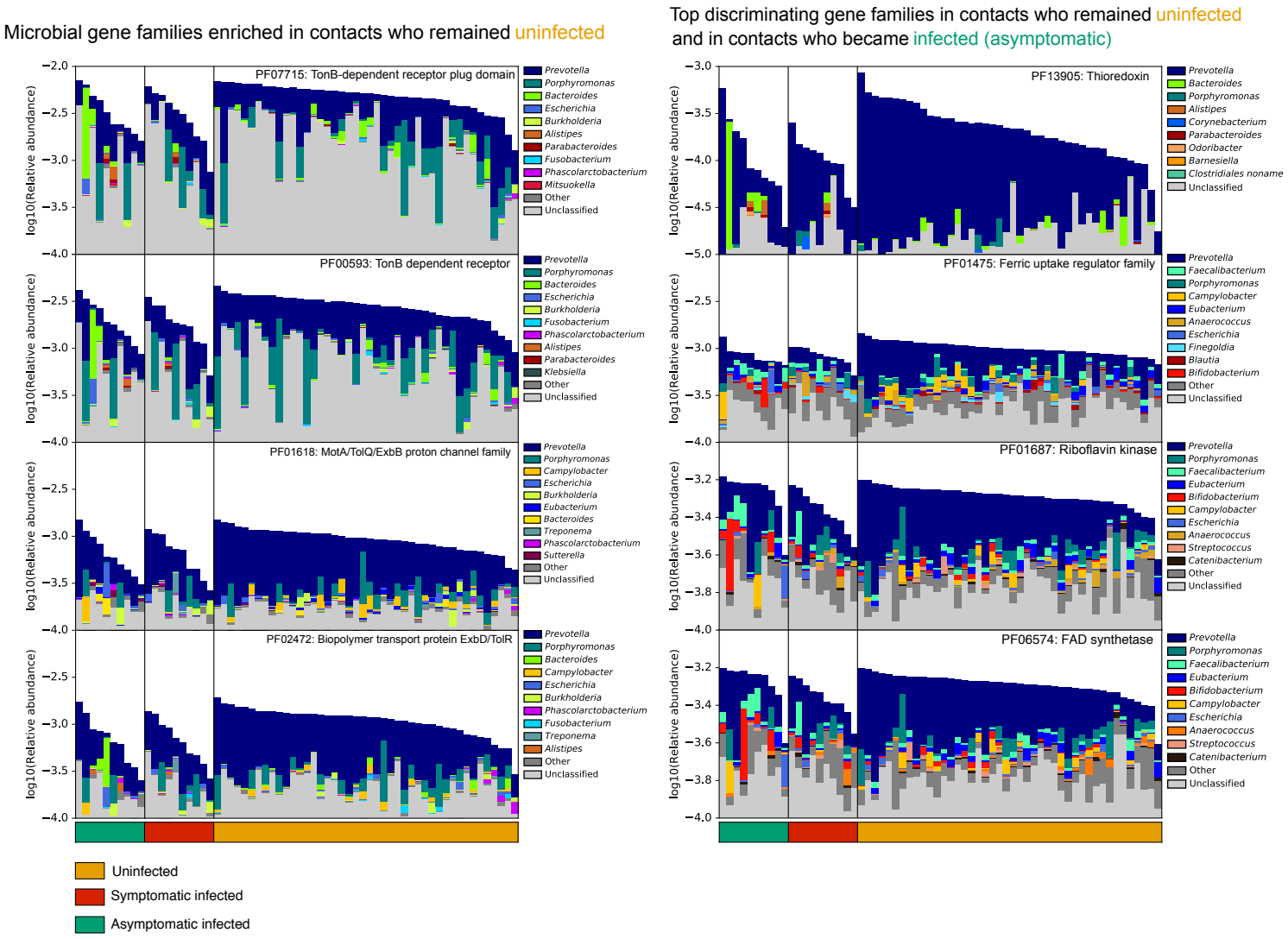


Figure S14. Enriched and top predictive gene families (Pfam domain grouping) involved in iron metabolism, annotated by their taxonomic contributors. Gene families involved with iron metabolism that were differentially abundant in contacts who remained uninfected were identified with LEfSe are represented on the left (shown in Figure S2). On the right are the top predictive gene families involved with iron metabolism identified with MetAML. Total bar height reflects log10-scaled community abundance. Genera contributions are linearly scaled within total.

Conclusion et perspectives futures

Comme résumé dans le chapitre 1, l'étude de la diversité génétique de populations bactériennes au sein de patients infectés et des dynamiques évolutives affectant cette diversité se révèle essentielle dans la compréhension de la virulence et de l'adaptabilité des pathogènes impliqués dans des infections chroniques. Ces connaissances sont en effet une première étape pour mettre en place des stratégies thérapeutiques ou prophylactiques prenant en compte les pressions sélectives identifiées et leur impact sur l'évolution de ces pathogènes. Ces applications représentent un défi majeur, mais indispensable, particulièrement dans un contexte où la résistance aux antibiotiques constitue aujourd'hui l'une des plus graves menaces pesant sur la santé mondiale, la sécurité alimentaire et le développement (Spellberg et al 2013).

Les recherches effectuées au cours de mon doctorat ont démontré que les approches visant à étudier la diversité intra-hôte des bactéries pathogènes et ses impacts peuvent aussi être appliquées dans le cadre d'infection aiguë telles que le choléra, ce qui n'était pas clair jusqu'à aujourd'hui. Les chapitres 2 et 3 de ma thèse auront permis d'en apprendre un peu plus sur l'évolution de ce pathogène au sein du patient qu'il infecte et vont permettre d'orienter de futures recherches menées dans un contexte expérimental, clinique et épidémiologique. L'étude des porteurs asymptomatiques en particulier, reste de première importance.

Bien que nous ayons montré dans les chapitres 3 et 4 que la variabilité génomique de *Vibrio cholerae* observée chez des hôtes asymptomatiques semble équivalente à celle retrouvée au niveau des patients symptomatiques et que le microbiome intestinal a un rôle potentiel dans la sévérité des symptômes, des études plus approfondies effectuées sur une cohorte comprenant un plus grand nombre de porteurs asymptomatiques sera nécessaire pour confirmer ces résultats. Le séquençage d'un plus grand nombre d'isolats issus de ces patients pourrait permettre d'identifier des variants impliqués dans la sévérité des symptômes, la réponse immunitaire ou la compétition avec le microbiome intestinal, via l'utilisation de méthodes d'étude d'association pangénomique (GWAS).

Jusqu'à aujourd'hui, l'utilisation du séquençage métagénomique dans un contexte clinique s'est majoritairement concentrée sur le diagnostic de pathologies dont l'agent ne

peut pas être identifié par des méthodes classiques (Gardy et Loman 2018), ou sur l'étude du microbiome humain et de son association avec certaines pathologies (Baohong Wang et al. 2017; Quince et al. 2017; Schmidt, Raes, et Bork 2018; Liang et al. 2018). Cette méthode présente encore de nombreuses limites quant à son application de manière routinière dans un contexte clinique ou en recherche fondamentale. Par exemple, il est encore coûteux de séquencer et d'analyser un grand nombre de métagénomes sans avoir accès à des installations de séquençage et de calcul adéquates. De plus, la faible proportion des cellules de l'agent pathogène dans les échantillons séquencés peut parfois rendre sa détection et caractérisation impossible (Gardy et Loman 2018; Garud et Pollard 2020). Dans nos travaux de recherche présentés dans le chapitre 2 nous avons montré que l'association du séquençage d'isolats cultivés et de la méthode métagénomique présente de nombreux avantages dans l'étude de l'évolution intra-hôte de pathogènes bactériens. Une combinaison de ces méthodes a récemment été appliquée pour l'étude de l'évolution intra-individuelle de membres du microbiome intestinal (Zhao et al. 2019; Lugli et al. 2019). Cependant, notre étude représente le premier exemple d'application de ces deux approches combinées pour l'analyse de la variation intra-hôte dans le cas d'une infection aiguë. Le développement et l'avancement continu des méthodes de séquençage, tels que le séquençage de cellules uniques, le séquençage de longs reads appliqué à des isolats ou des échantillons (Quince et al. 2017) ou encore la capture de génomes entiers (Vezzulli et al. 2017), permettront une étude plus approfondie de l'évolution de *Vibrio cholerae* au sein de son hôte humain particulièrement dans un contexte infectieux asymptomatique. Cependant, la mise en place d'un système de prélèvement d'échantillons visant à l'étude de ces porteurs asymptomatique présente de nombreux défis et reste très difficile à mettre en place. Notre cohorte, bien que de petite taille, reste un exemple unique de plan expérimental pour l'étude de ces individus infectés et demeure difficilement reproductible.

Ainsi, nos résultats présentés dans le chapitre 4 ne sont actuellement pas généralisables au-delà de notre cohorte d'étude à Dhaka, au Bangladesh, car une cohorte similaire dans un autre lieu géographique n'est pour le moment pas disponible. Comme pour toute étude par association (Schmidt, Raes, et Bork 2018), on ignore si certaines des caractéristiques métagénomiques qui sont en corrélation avec la protection contre l'infection par *V. cholerae* sont causales, car beaucoup peuvent être des marqueurs de facteurs cliniques

ou environnementaux qui ont eux-mêmes un impact sur la susceptibilité. Une caractérisation expérimentale plus poussée des caractéristiques métagénomiques corrélées avec la protection contre l'infection ou les symptômes est nécessaire pour comprendre si les facteurs que nous avons identifiés ont un impact sur la pathogenèse de *V. cholerae* ou sur les réponses de l'hôte à l'infection. Cependant, notre étude peut servir de base pour de futures études expérimentales visant à comprendre la manière dont les membres du microbiome intestinal peuvent influencer la susceptibilité à l'infection et la gravité de la maladie, ainsi que les interventions thérapeutiques et prophylactiques potentielles.

Le futur de la lutte contre le choléra

Une stratégie de lutte intitulée « Mettre fin au choléra: une feuille de route jusqu'à 2030 » a été lancée en 2017 par le Groupe spécial mondial de lutte contre le choléra. Cette stratégie qui relève d'une vingtaine de pays où le choléra est trouvée sous forme endémique vise à faire baisser de 90% le nombre des décès dus au choléra et à éliminer la maladie d'ici 2030. La stratégie offre un cadre pour des plans d'action nationaux mettant l'accent sur des axes de lutte comprenant: (1) une détection précoce et intervention rapide contre les flambées épidémiques, et (2) une approche multisectorielle intégrant une surveillance renforcée, la vaccination, la mise en place de mesure d'assainissement de l'eau et d'hygiène pour prévenir le choléra dans les points chauds des pays endémiques. Rien qu'en 2018, l'OMS a annoncé une diminution de 60% des cas reportés dans ces pays, notamment en Haïti, en Somalie et en République démocratique du Congo.

Toujours d'après l'OMS, plus d'un milliard de personnes n'ont toujours pas accès à une source d'eau sûre et près de 2,6 milliards de personnes n'ont pas accès à des systèmes d'assainissement des eaux satisfaisants. De plus, la disponibilité de l'eau douce sur la planète est de plus en plus menacée par l'utilisation des terres pour l'agriculture intensive, la déforestation, les catastrophes naturelles liées aux changements climatiques et la consommation accrue d'eau douce en raison de la croissance démographique et le développement de l'industrie. Pour ces différentes raisons, et à cause de la persistance du *Vibrio cholerae* dans les environnements aquatiques, les vaccins anticholériques oraux (VCO) sont un des moyens privilégiés à l'heure actuelle pour prévenir le choléra et atteindre les objectifs de l'OMS d'ici 2030. À ce jour, il existe deux vaccins anticholériques oraux (VCO)

utilisés dans les campagnes de vaccination de masse: SHANCHOL® et EUVICHOL®, deux vaccins bivalents constitués de cellules entières tuées de *V. cholerae* O1 (classique et El Tor) et O139. Il sont tous deux administrés en deux doses à toute personne de plus de 1 an, avec un intervalle d'au moins 14 jours entre les deux doses, présentent une plus grande efficacité que les vaccins à une dose, même si cette efficacité semble variable entre les différentes populations testées (Levine 2010; Franke et al. 2018).

Comme ces vaccins sont absorbés et agissent sur la réponse immunitaire au niveau de la surface des muqueuses intestinales, où résident une grande partie des membres du microbiome intestinal, ceux-ci pourraient avoir un impact significatif sur les réponses immunitaires à la vaccination. Cela a conduit à l'hypothèse que la composition du microbiome intestinal est un facteur hôte non pris en compte qui pourrait déterminer en partie les réponses immunitaires variables aux VCO et expliquer la variabilité de l'efficacité des vaccins (Weil, Becker, et Harris 2019). De plus, dans certains cas des antibiotiques sont administrés avant la vaccination (Cooper et al. 2000; Bhuiyan et al. 2014), ce qui pourrait altérer le microbiome intestinal et affecter la réponse vaccinale. Ainsi, des recherches supplémentaires sont nécessaires pour comprendre la dynamique et les avantages des traitements affectant le microbiome intestinal au moment de la vaccination.

De la même manière, des études devraient être effectuées pour évaluer l'impact de la prise de probiotiques sur l'infection par *V. cholerae*, ou sur l'efficacité de la réponse vaccinale. Deux études avec groupes contrôles randomisés menées sur des effectifs de petite taille visant à évaluer l'effet de la prise de probiotiques avant la prise d'un VCO n'ont pas montré d'effets probants sur l'immunogénicité vaccinale (Paineau et al. 2008; Matsuda et al. 2011). Cependant les études sur l'effet des probiotiques sur la santé humaine restent encore limitées par le choix des souches bactériennes, la pureté, la dose et le moment de l'administration, ou encore par le fait de savoir si la souche probiotique persiste et colonise l'individu ou ne fait que passer dans le système digestif (Zimmermann et Curtis 2018). Un suivi de notre étude présentée dans le chapitre 4 pourrait être utile pour la sélection et le test de souches probiotiques d'intérêt dans des essais cliniques.

Beaucoup reste encore à faire pour améliorer notre compréhension de ce pathogène et des interactions avec son hôte et le microbiome, et j'ose espérer que ma thèse, qui se

retrouve à l'interface entre la recherche fondamentale et la recherche clinique, aura permis d'ajouter une pierre à l'édifice dans la lutte contre le choléra.

Références

Abel, Sören, Pia Abel zur Wiesch, Hsiao-Han Chang, Brigid M Davis, Marc Lipsitch, et Matthew K Waldor. 2015. « Sequence Tag-Based Analysis of Microbial Population Dynamics. » *Nature Methods* 12 (3): 223–6– 3. https://doi.org/10.1038/nmeth.3253.

Alam, Ashfaqul, Regina C. LaRocque, Jason B. Harris, Cecily Vanderspurt, Edward T. Ryan, Firdausi Qadri, et Stephen B. Calderwood. 2005. « Hyperinfectivity of Human-Passaged *Vibrio cholerae* Can Be Modeled by Growth in the Infant Mouse ». *Infection and Immunity* 73 (10): 6674-79. https://doi.org/10.1128/IAI.73.10.6674-6679.2005.

Ali, Mohammad, Michael Emch, Jin Kyung Park, Mohammad Yunus, et John Clemens. 2011. « Natural Cholera Infection–Derived Immunity in an Endemic Setting ». *The Journal of Infectious Diseases* 204 (6): 912-18. https://doi.org/10.1093/infdis/jir416.

Ali, Mohammad, Allyson R. Nelson, Anna Lena Lopez, et David A. Sack. 2015. « Updated Global Burden of Cholera in Endemic Countries ». Édité par Justin V. Remais. *PLOS Neglected Tropical Diseases* 9 (6): e0003832. https://doi.org/10.1371/journal.pntd.0003832.

Alkan, Can, Saba Sajjadian, et Evan E Eichler. 2011. « Limitations of Next-Generation Genome Sequence Assembly. » *Nature Methods* 8 (1): 61–65. https://doi.org/10.1038/nmeth.1527.

Almagro-Moreno, Salvador, Kali Pruss, et Ronald K Taylor. 2015. « Intestinal Colonization Dynamics of *Vibrio Cholerae* ». *PLOS Pathogens* 11 (5): e1004787–11. https://doi.org/10.1371/journal.ppat.1004787.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, et Christopher Quince. 2014. « Binning Metagenomic Contigs by Coverage and

Composition ». *Nature Methods* 11 (11): 1144–1146. https://doi.org/10.1038/nmeth.3103.

Azarian, Taj, Afsar Ali, Judith A Johnson, David Mohr, Mattia Prosperi, Nazle M Veras, Mohammed Jubair, et al. 2014. « Phylodynamic Analysis of Clinical and Environmental *Vibrio Cholerae* Isolates from Haiti Reveals Diversification Driven by Positive Selection ». *MBio* 5 (6): e01824–14–11. https://doi.org/10.1128/mBio.01824-14.

Aziz, Ramy K, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, et al. 2008. « The RAST Server: Rapid Annotations Using Subsystems Technology. » *BMC Genomics* 9 (1): 75. https://doi.org/10.1186/1471-2164-9-75.

Bachmann, Nathan L., Rebecca J. Rockett, Verlaine Joy Timms, et Vitali Sintchenko. 2018. « Advances in Clinical Sample Preparation for Identification and Characterization of Bacterial Pathogens Using Metagenomics ». *Frontiers in Public Health* 6 (décembre). https://doi.org/10.3389/fpubh.2018.00363.

Bachmann, Verena, Benjamin Kostiuk, Daniel Unterweger, Laura Diaz-Satizabal, Stephen Ogg, et Stefan Pukatzki. 2015. « Bile Salts Modulate the Mucin-Activated Type VI Secretion System of Pandemic *Vibrio Cholerae* ». *PLOS Neglected Tropical Diseases* 9 (8): e0004031. https://doi.org/10.1371/journal.pntd.0004031.

Baele, Guy et al. 2012. « Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty ». *Molecular Biology and Evolution* 29(9): 2157-67.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, et al. 2012. « SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing ». *Journal of Computational Biology* 19 (5): 455–477. https://doi.org/10.1089/cmb.2012.0021.

Baumgartner, Michael, Florian Bayer, Katia R. Pfrunder-Cardozo, Angus Buckling, et Alex R. Hall. 2020. « Resident Microbial Communities Inhibit Growth and Antibiotic-Resistance Evolution of Escherichia Coli in Human Gut Microbiome Samples ». *PLOS Biology* 18 (4): e3000465. https://doi.org/10.1371/journal.pbio.3000465.

Bhuiyan, Taufiqur R., Feroza K. Choudhury, Farhana Khanam, Amit Saha, Md. Abu Sayeed, Umme Salma, Anna Lundgren, David A. Sack, Ann-Mari Svennerholm, et Firdausi Qadri. 2014. « Evaluation of Immune Responses to an Oral Typhoid

Vaccine, Ty21a, in Children from 2 to 5 Years of Age in Bangladesh ». *Vaccine* 32 (9): 1055-60. https://doi.org/10.1016/j.vaccine.2014.01.001.

Bilecen, Kivanc, Jiunn C N Fong, Andrew Cheng, Christopher J Jones, David Zamorano-Sánchez, et Fitnat H Yildiz. 2015. « Polymyxin B Resistance and Biofilm Formation in *Vibrio Cholerae* Are Controlled by the Response Regulator CarR ». *Infection and Immunity* 83 (3): 1199–1209. https://doi.org/10.1128/IAI.02700-14.

Bilecen, Kivanc, et Fitnat H Yildiz. 2009. « Identification of a Calcium-Controlled Negative Regulatory System Affecting *Vibrio Cholerae* Biofilm Formation ». *Environmental Microbiology* 11 (8): 2015–2029. https://doi.org/10.1111/j.1462-2920.2009.01923.x.

Bolger, Anthony M, Marc Lohse, et Bjoern Usadel. 2014. « Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. » *Bioinformatics* 30 (15): 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Borgeaud, Sandrine, Lisa C Metzger, Tiziana Scrignari, et Melanie Blokesch. 2015. « The Type VI Secretion System of *Vibrio Cholerae* Fosters Horizontal Gene Transfer ». *Science* 347 (6217): 63–67. https://doi.org/10.1126/science.1260064.

Boucher, Y., O. X. Cordero, A. Takemura, D. E. Hunt, K. Schliep, E. Bapteste, P. Lopez, C. L. Tarr, et M. F. Polz. 2011. « Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global *Vibrio cholerae* Populations ». *mBio* 2 (2): e00335-10-e00335-10. https://doi.org/10.1128/mBio.00335-10.

Boucher, Yan, Fabini D. Orata, et Munirul Alam. 2015. « The Out-of-the-Delta Hypothesis: Dense Human Populations in Low-Lying River Deltas Served as Agents for the Evolution of a Deadly Pathogen ». *Frontiers in Microbiology* 6. https://doi.org/10.3389/fmicb.2015.01120.

Brenzinger, Susanne, Lizah T. van der Aart, Gilles P. van Wezel, Jean-Marie Lacroix, Timo Glatter, et Ariane Briegel. 2019. « Structural and Proteomic Changes in Viable but Non-culturable *Vibrio cholerae* ». *Frontiers in Microbiology* 10 (avril). https://doi.org/10.3389/fmicb.2019.00793.

Brown, Christopher T., Matthew R. Olm, Brian C. Thomas, et Jillian F. Banfield. 2016. « Measurement of Bacterial Replication Rates in Microbial Communities ». *Nature Biotechnology* 34 (12): 1256-63. https://doi.org/10.1038/nbt.3704.

Bryce, Jennifer, Cynthia Boschi-Pinto, Kenji Shibuya, et Robert E Black. 2005. « WHO Estimates of the Causes of Death in Children ». *The Lancet* 365 (9465): 1147-52. https://doi.org/10.1016/S0140-6736(05)71877-8.

Camacho, Anton, Malika Bouhenia, Reema Alyusfi, Abdulhakeem Alkohlani, Munna Abdulla Mohammed Naji, Xavier de Radiguès, Abdinasir M Abubakar, et al. 2018. « Cholera Epidemic in Yemen, 2016–18: An Analysis of Surveillance Data ». *The Lancet Global Health* 6 (6): e680-90. https://doi.org/10.1016/S2214-109X(18)30230-4.

Canani, Roberto Berni, Margherita Di Costanzo, Ludovica Leone, Monica Pedata, Rosaria Meli, et Antonio Calignano. 2011. « Potential beneficial effects of butyrate in intestinal and extraintestinal diseases ». *World Journal of Gastroenterology : WJG* 17 (12): 1519-28. https://doi.org/10.3748/wjg.v17.i12.1519.

Cao, Qizhi, Xavier Didelot, Zhongbiao Wu, Zongwei Li, Lihua He, Yunsheng Li, Ming Ni, et al. 2015. « Progressive Genomic Convergence of Two *Helicobacter Pylori* Strains during Mixed Infection of a Patient with Chronic Gastritis ». *Gut* 64 (4): 554-61. https://doi.org/10.1136/gutjnl-2014-307345.

Chatterjee, S N, et Keya Chaudhuri. 2003. « Lipopolysaccharides of *Vibrio Cholerae* ». *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1639 (2): 65–79. https://doi.org/10.1016/j.bbadis.2003.08.004.

Chin, Chen-Shan, Jon Sorenson, Jason B. Harris, William P. Robins, Richelle C. Charles, Roger R. Jean-Charles, James Bullard, Dale R. Webster, Andrew Kasarskis, et Paul Peluso. 2011. « The origin of the Haitian cholera outbreak strain ». *New England Journal of Medicine* 364 (1): 33–42.

Chu, Nathaniel D., Sean A. Clarke, Sonia Timberlake, Martin F. Polz, Alan D. Grossman, et Eric J. Alm. 2017. « A Mobile Element in MutS Drives Hypermutation in a Marine Vibrio ». *MBio* 8 (1). https://doi.org/10.1128/mBio.02045-16.

Chun, Jongsik, Christopher J. Grim, Nur A. Hasan, Je Hee Lee, Seon Young Choi, Bradd J. Haley, Elisa Taviani, Yoon-Seong Jeon, Dong Wook Kim, et Jae-Hak Lee. 2009. « Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae* ». *Proceedings of the National Academy of Sciences* 106 (36): 15442–15447.

Clemens, John D., Frederik van Loon, David A. Sack, J. Chakraborty, M. R. Rao, Faruque Ahmed, Jeffrey R. Harris, et al. 1991. « Field Trial of Oral Cholera Vaccines in

Bangladesh: Serum Vibriocidal and Antitoxic Antibodies as Markers of the Risk of Cholera ». *The Journal of Infectious Diseases* 163 (6): 1235-42. https://doi.org/10.1093/infdis/163.6.1235.

Collins, James W., Kristie M. Keeney, Valerie F. Crepin, Vijay A. K. Rathinam, Katherine A. Fitzgerald, B. Brett Finlay, et Gad Frankel. 2014. « Citrobacter Rodentium : Infection, Inflammation and the Microbiota ». *Nature Reviews Microbiology* 12 (9): 612-23. https://doi.org/10.1038/nrmicro3315.

Colwell, Rita R. 1996. « Global Climate and Infectious Disease: The Cholera Paradigm». *Science* 274 (5295): 2025-31. https://doi.org/10.1126/science.274.5295.2025.

Cooper, Philip J., Martha E. Chico, Genevieve Losonsky, Carlos Sandoval, Ivan Espinel, Rajeshwari Sridhara, Marcelo Aguilar, et al. 2000. « Albendazole Treatment of Children with Ascariasis Enhances the Vibriocidal Antibody Response to the Live Attenuated Oral Cholera Vaccine CVD 103-HgR ». *The Journal of Infectious Diseases* 182 (4): 1199-1206. https://doi.org/10.1086/315837.

Corr, Sinead C., Cormac G. M. Gahan, et Colin Hill. 2007. « Impact of Selected Lactobacillus and Bifidobacterium Species on Listeria Monocytogenes Infection and the Mucosal Immune Response ». *FEMS Immunology & Medical Microbiology* 50 (3): 380-88. https://doi.org/10.1111/j.1574-695X.2007.00264.x.

Crost, E. H., E. H. Ajandouz, C. Villard, P. A. Geraert, A. Puigserver, et M. Fons. 2011. « Ruminococcin C, a New Anti-Clostridium Perfringens Bacteriocin Produced in the Gut by the Commensal Bacterium Ruminococcus Gnavus E1 ». *Biochimie* 93 (9): 1487-94. https://doi.org/10.1016/j.biochi.2011.05.001.

Crost, Emmanuelle H., Louise E. Tailford, Gwenaelle Le Gall, Michel Fons, Bernard Henrissat, et Nathalie Juge. 2013. « Utilisation of Mucin Glycans by the Human Gut Symbiont Ruminococcus Gnavus Is Strain-Dependent ». *PLOS ONE* 8 (10): e76341. https://doi.org/10.1371/journal.pone.0076341.

Croucher, Nicholas J, et Xavier Didelot. 2015. « The Application of Genomics to Tracing Bacterial Pathogen Transmission ». *Current Opinion in Microbiology*,: 62-67. https://doi.org/10.1016/j.mib.2014.11.004.

Croucher, Nicholas J., Simon R. Harris, Christophe Fraser, Michael A. Quail, John Burton, Mark van der Linden, Lesley McGee, et al. 2011. « Rapid Pneumococcal Evolution in Response to Clinical Interventions ». *Science* 331 (6016): 430-34. https://doi.org/10.1126/science.1198545.

Darmon, Elise, et David R. F. Leach. 2014. « Bacterial Genome Instability ». *Microbiology and Molecular Biology Reviews : MMBR* 78 (1): 1-39. https://doi.org/10.1128/MMBR.00035-13.

Darmon, Elise, Manuel A. Lopez-Vernaza, Anne C. Helness, Amanda Borking, Emily Wilson, Zubin Thacker, Laura Wardrope, et David R. F. Leach. 2007. « SbcCD Regulation and Localization in *Escherichia Coli* ». *Journal of Bacteriology* 189 (18): 6686-94. https://doi.org/10.1128/JB.00489-07.

Das, Bhabatosh, Gururaja P Pazhani, Anirban Sarkar, Asish K Mukhopadhyay, G Balakrish Nair, et Thandavarayan Ramamurthy. 2016. « Molecular Evolution and Functional Divergence of *Vibrio Cholerae* ». *Current Opinion in Infectious Diseases* 29 (5): 520–527. https://doi.org/10.1097/QCO.0000000000000306.

David, Lawrence A, Ana Weil, Edward T Ryan, Stephen B Calderwood, Jason B Harris, Fahima Chowdhury, Yasmin Begum, Firdausi Qadri, Regina C LaRocque, et Peter J Turnbaugh. 2015. « Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans ». *MBio* 6 (3): e00381–15. https://doi.org/10.1128/mBio.00381-15.

Denamur, Erick, Guillaume Lecointre, Pierre Darlu, Olivier Tenaillon, Cécile Acquaviva, Chalom Sayada, Ivana Sunjevaric, et al. 2000. « Evolutionary Implications of the Frequent Horizontal Transfer of Mismatch Repair Genes ». *Cell* 103 (5): 711-21. https://doi.org/10.1016/S0092-8674(00)00175-6.

Denton, James F, Jose Lugo-Martinez, Abraham E Tucker, Daniel R Schrider, Wesley C Warren, et Matthew W Hahn. 2014. « Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies ». *PLoS computational biology* 10 (12): e1003998. https://doi.org/10.1371/journal.pcbi.1003998.

Didelot, Xavier, Christophe Fraser, Jennifer Gardy, et Caroline Colijn. 2017. « Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks ». *Molecular Biology and Evolution* 34 (4): 997-1007. https://doi.org/10.1093/molbev/msw275.

Didelot, Xavier, Jennifer Gardy, et Caroline Colijn. 2014. « Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data ». *Molecular Biology and Evolution* 31 (7): 1869-79. https://doi.org/10.1093/molbev/msu121.

Didelot, Xavier, Sandra Nell, Ines Yang, Sabrina Woltemate, Schalk van der Merwe, et Sebastian Suerbaum. 2013. « Genomic Evolution and Transmission of *Helicobacter*

Pylori in Two South African Families ». *Proceedings of the National Academy of Sciences* 110 (34): 13880-85. https://doi.org/10.1073/pnas.1304681110.

Didelot, Xavier, Bo Pang, Zhemin Zhou, Angela McCann, Peixiang Ni, Dongfang Li, Mark Achtman, et Biao Kan. 2015. « The Role of China in the Global Spread of the Current Cholera Pandemic ». *PLOS Genetics* 11 (3): e1005072. https://doi.org/10.1371/journal.pgen.1005072.

Didelot, Xavier, A Sarah Walker, Tim E Peto, Derrick W Crook, et Daniel J Wilson. 2016. « Within-Host Evolution of Bacterial Pathogens. » *Nature Publishing Group* 14 (3): 150–162. https://doi.org/10.1038/nrmicro.2015.13.

Dillon, Marcus M, Way Sung, Robert Sebra, Michael Lynch, et Vaughn S Cooper. 2017. « Genome-Wide Biases in the Rate and Molecular Spectrum of Spontaneous Mutations in *Vibrio cholerae* and *Vibrio fischeri* ». *Molecular Biology and Evolution* 34 (1): 93–109. https://doi.org/10.1093/molbev/msw224.

Domman, Daryl, Fahima Chowdhury, Ashraful I. Khan, Matthew J. Dorman, Ankur Mutreja, Muhammad Ikhtear Uddin, Anik Paul, et al. 2018. « Defining Endemic Cholera at Three Levels of Spatiotemporal Resolution within Bangladesh ». *Nature Genetics* 50 (7): 951-55. https://doi.org/10.1038/s41588-018-0150-8.

Domman, Daryl, Marie-Laure Quilici, Matthew J. Dorman, Elisabeth Njamkepo, Ankur Mutreja, Alison E. Mather, Gabriella Delgado, et al. 2017. « Integrated View of *Vibrio Cholerae* in the Americas ». *Science* 358 (6364): 789-93. https://doi.org/10.1126/science.aao2136.

Drake, John W, Brian Charlesworth, Deborah Charlesworth, et James F Crow. 1998. « Rates of Spontaneous Mutation ». *Genetics* 148 (4): 1667–1686. https://doi.org/10.2307/2410123.

Draper, Jenny L, Lori M Hansen, David L Bernick, Samar Abedrabbo, Jason G Underwood, Nguyet Kong, Bihua C Huang, et al. 2017. « Fallacy of the Unique Genome: Sequence Diversity within Single *Helicobacter Pylori* Strains ». *MBio* 8 (1): e02321–16. https://doi.org/10.1128/mBio.02321-16.

Duan, Faping, et John C. March. 2010. « Engineered Bacterial Communication Prevents *Vibrio Cholerae* Virulence in an Infant Mouse Model ». *Proceedings of the National Academy of Sciences* 107 (25): 11260-64. https://doi.org/10.1073/pnas.1001294107.

Duchêne, Sebastian, Kathryn E Holt, François-Xavier Weill, Simon Le Hello, Jane Hawkey, David J Edwards, Mathieu Fourment, et Edward C Holmes. 2016. « Genome-scale rates of evolutionary change in bacteria ». *Microbial Genomics* 2 (11). https://doi.org/10.1099/mgen.0.000094.

Dziejman, Michelle, Davide Serruto, Vincent C. Tam, Derek Sturtevant, Pornphan Diraphat, Shah M. Faruque, M. Hasibur Rahman, et al. 2005. « Genomic characterization of non-O1, non-O139 Vibrio cholerae reveals genes for a type III secretion system ». *Proceedings of the National Academy of Sciences of the United States of America* 102 (9): 3465-70. https://doi.org/10.1073/pnas.0409918102.

Edgar, Robert C. 2004. « MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput ». *Nucleic Acids Research* 32(5): 1792-97.

Eisenstark, Abraham. 2010. « Genetic Diversity among Offspring from Archived *Salmonella enterica* ssp. enterica Serovar Typhimurium (Demerec Collection): In Search of Survival Strategies ». *Annual Review of Microbiology* 64 (1): 277-92. https://doi.org/10.1146/annurev.micro.091208.073614.

Eldholm, Vegard, Gunnstein Norheim, Bent von der Lippe, Wibeke Kinander, Ulf R Dahle, Dominique A Caugant, Turid Manns\a aker, Anne Torunn Mengshoel, Anne Ma Dyrhol-Riise, et Francois Balloux. 2014. « Evolution of Extensively Drug-Resistant *Mycobacterium Tuberculosis* from a Susceptible Ancestor in a Single Patient ». *Genome Biology* 15 (11): 490. https://doi.org/10.1186/s13059-014-0490-3.

Emms, David M., et Steven Kelly. 2015. « OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy ». *Genome Biology* 16 (1). https://doi.org/10.1186/s13059-015-0721-2.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, et Tom O. Delmont. 2015. « Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data ». *PeerJ* 3: e1319. https://doi.org/10.7717/peerj.1319.

Eren, A. Murat, Joseph H. Vineis, Hilary G. Morrison, et Mitchell L. Sogin. 2013. « A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology ». *PLOS ONE* 8 (6): e66643. https://doi.org/10.1371/journal.pone.0066643.

Eyre, David W., Madeleine L. Cule, David Griffiths, Derrick W. Crook, Tim E. A. Peto, A. Sarah Walker, et Daniel J. Wilson. 2013. « Detection of Mixed Infection from

Bacterial Whole Genome Sequence Data Allows Assessment of Its Role in *Clostridium difficile* Transmission ». Édité par Christophe Fraser. *PLoS Computational Biology* 9 (5): e1003059. https://doi.org/10.1371/journal.pcbi.1003059.

Fachi, José Luís, Jaqueline de Souza Felipe, Laís Passariello Pral, Bruna Karadi da Silva, Renan Oliveira Corrêa, Mirella Cristiny Pereira de Andrade, Denise Morais da Fonseca, et al. 2019. « Butyrate Protects Mice from *Clostridium Difficile*-Induced Colitis through an HIF-1-Dependent Mechanism ». *Cell Reports* 27 (3): 750-761.e7. https://doi.org/10.1016/j.celrep.2019.03.054.

Faruque, Shah M, Nityananda Chowdhury, M Kamruzzaman, Michelle Dziejman, M Hasibur Rahman, David A Sack, G Balakrish Nair, et John J Mekalanos. 2004. « Genetic Diversity and Virulence Potential of Environmental *Vibrio Cholerae* Population in a Cholera-Endemic Area ». *Proceedings of the National Academy of Sciences* 101 (7): 2123–2128. https://doi.org/10.1073/pnas.0308485100.

Faruque, Shah M, et John J Mekalanos. 2014. « Phage-Bacterial Interactions in the Evolution of Toxigenic *Vibrio Cholerae* ». *Virulence* 3 (7): 556–565. https://doi.org/10.4161/viru.22351.

Favrot, Lorenza, John S Blanchard, et Olivia Vergnolle. 2016. « Bacterial GCN5-Related N-Acetyltransferases: From Resistance to Regulation ». *Biochemistry*, février. https://doi.org/10.1021/acs.biochem.5b01269.

Fillat, María F. 2014. « The FUR (Ferric Uptake Regulator) Superfamily: Diversity and Versatility of Key Transcriptional Regulators ». *Archives of Biochemistry and Biophysics* 546 (mars): 41-52. https://doi.org/10.1016/j.abb.2014.01.029.

Folster, Jason P, Lee Katz, Andre McCullough, Michele B Parsons, Kristen Knipe, Scott A Sammons, Jacques Boncy, Cheryl Lea Tarr, et Jean M Whichard. 2014. « Multidrug-Resistant IncA/C Plasmid in *Vibrio cholerae* from Haiti ». *Emerging Infectious Diseases* 20 (11): 1951–1953. https://doi.org/10.3201/eid2011.140889.

Foster, P. L., G. Gudmundsson, J. M. Trimarchi, H. Cai, et M. F. Goodman. 1995. « Proofreading-Defective DNA Polymerase II Increases Adaptive Mutation in *Escherichia Coli* ». *Proceedings of the National Academy of Sciences* 92 (17): 7951-55. https://doi.org/10.1073/pnas.92.17.7951.

Franke, Molly F, Ralph Ternier, J Gregory Jerome, Wilfredo R Matias, Jason B Harris, et Louise C Ivers. 2018. « Long-Term Effectiveness of One and Two Doses of a

Killed, Bivalent, Whole-Cell Oral Cholera Vaccine in Haiti: An Extended Case-Control Study ». *The Lancet Global Health* 6 (9): e1028-35. https://doi.org/10.1016/S2214-109X(18)30284-5.

Franzosa, Eric A., Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, et al. 2018. « Species-Level Functional Profiling of Metagenomes and Metatranscriptomes ». *Nature Methods* 15 (11): 962–968. https://doi.org/10.1038/s41592-018-0176-y.

Freter, Rolf. 1955. « The Fatal Enteric Cholera Infection in the Guinea Pig, Achieved by Inhibition of Normal Enteric Flora ». *The Journal of Infectious Diseases* 97 (1): 57-65. https://doi.org/10.1093/infdis/97.1.57.

Fu, Limin et al. 2012. « CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data ». *Bioinformatics* 28(23): 3150-52.

Fu, Yang, Matthew K. Waldor, et John J. Mekalanos. 2013. « Tn-Seq Analysis of *Vibrio Cholerae* Intestinal Colonization Reveals a Role for T6SS-Mediated Antibacterial Activity in the Host ». *Cell Host & Microbe* 14 (6): 652-63. https://doi.org/10.1016/j.chom.2013.11.001.

Fukuda, Shinji, Hidehiro Toh, Koji Hase, Kenshiro Oshima, Yumiko Nakanishi, Kazutoshi Yoshimura, Toru Tobe, et al. 2011. « Bifidobacteria Can Protect from Enteropathogenic Infection through Production of Acetate ». *Nature* 469 (7331): 543-47. https://doi.org/10.1038/nature09646.

Funchain, Pauline, Annie Yeung, Jean Lee Stewart, Rose Lin, Malgorzata M. Slupska, et Jeffrey H. Miller. 2000. « The Consequences of Growth of a Mutator Strain of *Escherichia Coli* as Measured by Loss of Function Among Multiple Gene Targets and Loss of Fitness ». *Genetics* 154 (3): 959-70.

Gardy, Jennifer L., et Nicholas J. Loman. 2018. « Towards a Genomics-Informed, Real-Time, Global Pathogen Surveillance System ». *Nature Reviews Genetics* 19 (1): 9-20. https://doi.org/10.1038/nrg.2017.88.

Garrison, Erik, et Gabor Marth. 2012. « Haplotype-based variant detection from short-read sequencing ». *arXiv.org*. https://arxiv.org/abs/1207.3907.

Garud, Nandita R., Benjamin H. Good, Oskar Hallatschek, et Katherine S. Pollard. 2019. « Evolutionary Dynamics of Bacteria in the Gut Microbiome within and across

Hosts ». *PLOS Biology* 17 (1): e3000102. https://doi.org/10.1371/journal.pbio.3000102.

Garud, Nandita R., et Katherine S. Pollard. 2020. « Population Genetics in the Human Microbiome ». *Trends in Genetics* 36 (1): 53-67. https://doi.org/10.1016/j.tig.2019.10.010.

Giraud, Antoine, Ivan Matic, Olivier Tenaillon, Antonio Clara, Miroslav Radman, Michel Fons, et François Taddei. 2001. « Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut ». *Science* 291 (5513): 2606-8. https://doi.org/10.1126/science.1056421.

Glass, Roger I., Stan Becker, M. Imdadul Huq, Barbara J. Stoll, M. U. Khan, Michael H. Merson, John V. Lee, et Robert E. Black. 1982. « Endemic Cholera in Rural Bangladesh, 1966–1980 ». *American Journal of Epidemiology* 116 (6): 959-70. https://doi.org/10.1093/oxfordjournals.aje.a113498.

Glass, Roger I., Jan Holmgren, Charles E. Haley, M. R. Khan, Annmari Svennerholm, Barbara J. Stoll, K. M. Belayet Hossain, Robert E. Black, M. Yunus, et Dhiman Barua. 1985. « Predisposition for Cholera of Individuals with O Blood Group Possible Evolutionary Significance». *American Journal of Epidemiology* 121 (6): 791-96. https://doi.org/10.1093/oxfordjournals.aje.a114050.

Glass, Roger I., Ann-Mari Svennerholm, M. R. Khan, Shamsul Huda, M. Imdadul Huq, et Jan Holmgren. 1985. « Seroepidemiological Studies of El Tor Cholera in Bangladesh: Association of Serum Antibody Levels with Protection ». *The Journal of Infectious Diseases* 151 (2): 236-42. https://doi.org/10.1093/infdis/151.2.236.

Golubchik, Tanya, Elizabeth M. Batty, Ruth R. Miller, Helen Farr, Bernadette C. Young, Hanna Larner-Svensson, Rowena Fung, et al. 2013. « Within-Host Evolution of *Staphylococcus aureus* during Asymptomatic Carriage ». Édité par Ramy K. Aziz. *PLoS ONE* 8 (5): e61319. https://doi.org/10.1371/journal.pone.0061319.

Gouy, Manolo, Stéphane Guindon, et Olivier Gascuel. 2010. « SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building ». *Molecular Biology and Evolution* 27 (2): 221–224. https://doi.org/10.1093/molbev/msp259.

Grim, Christopher J, Nur A Hasan, Elisa Taviani, Bradd Haley, Jongsik Chun, Thomas S Brettin, David C Bruce, et al. 2010. « Genome Sequence of Hybrid *Vibrio Cholerae* O1 MJ-1236, B-33, and CIRS101 and Comparative Genomics with *V. Cholerae* ».

Journal of Bacteriology 192 (13): 3524–3533. https://doi.org/10.1128/JB.00040-10.

Gupta, S, et R Chowdhury. 1997. « Bile Affects Production of Virulence Factors and Motility of *Vibrio Cholerae*. » *Infection and Immunity* 65 (3): 1131-34. https://doi.org/10.1128/IAI.65.3.1131-1134.1997.

Hadfield, James, Nicholas J. Croucher, Richard J. Goater, Khalil Abudahab, David M. Aanensen, et Simon R. Harris. 2018. « Phandango: An Interactive Viewer for Bacterial Population Genomics ». *Bioinformatics* 34 (2): 292-93. https://doi.org/10.1093/bioinformatics/btx610.

Harris, Jason B., Ashraful I. Khan, Regina C. LaRocque, David J. Dorer, Fahima Chowdhury, Abu S. G. Faruque, David A. Sack, Edward T. Ryan, Firdausi Qadri, et Stephen B. Calderwood. 2005. « Blood Group, Immunity, and Risk of Infection with *Vibrio Cholerae* in an Area of Endemicity ». *Infection and Immunity* 73 (11): 7422-27. https://doi.org/10.1128/IAI.73.11.7422-7427.2005.

Harris, Jason B., Regina C. LaRocque, Fahima Chowdhury, Ashraful I. Khan, Tanya Logvinenko, Abu S. G. Faruque, Edward T. Ryan, Firdausi Qadri, et Stephen B. Calderwood. 2008a. « Susceptibility to *Vibrio Cholerae* Infection in a Cohort of Household Contacts of Patients with Cholera in Bangladesh ». *PLOS Neglected Tropical Diseases* 2 (4): e221. https://doi.org/10.1371/journal.pntd.0000221.

Harris, Jason B, Regina C LaRocque, Firdausi Qadri, Edward T Ryan, et Stephen B Calderwood. 2012. « Cholera. » *Lancet (London, England)* 379 (9835): 2466–2476. https://doi.org/10.1016/S0140-6736(12)60436-X.

Harris, Simon R, Edward JP Cartwright, M Estée Török, Matthew TG Holden, Nicholas M Brown, Amanda L Ogilvy-Stuart, Matthew J Ellington, et al. 2013. « Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study ». *The Lancet Infectious Diseases* 13 (2): 130-36. https://doi.org/10.1016/S1473-3099(12)70268-2.

Hazen, Tracy H., Li Pan, Ji-Dong Gu, et Patricia A. Sobecky. 2010. « The Contribution of Mobile Genetic Elements to the Evolution and Ecology of Vibrios ». *FEMS Microbiology Ecology* 74 (3): 485-99. https://doi.org/10.1111/j.1574-6941.2010.00937.x.

Hendriksen, Rene S., Lance B. Price, James M. Schupp, John D. Gillece, Rolf S. Kaas, David M. Engelthaler, Valeria Bortolaia, et al. 2011. « Population Genetics of

Vibrio Cholerae from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak ». *MBio* 2 (4). https://doi.org/10.1128/mBio.00157-11.

Herrera, Carmen M., Alexander A. Crofts, Jeremy C. Henderson, S. Cassandra Pingali, Bryan W. Davies, et M. Stephen Trent. 2014. « The *Vibrio Cholerae* VprA-VprB Two-Component System Controls Virulence through Endotoxin Modification ». *MBio* 5 (6). https://doi.org/10.1128/mBio.02283-14.

Higgins, Douglas A., Megan E. Pomianek, Christina M. Kraml, Ronald K. Taylor, Martin F. Semmelhack, et Bonnie L. Bassler. 2007. « The Major *Vibrio Cholerae* Autoinducer and Its Role in Virulence Factor Production ». *Nature* 450 (7171): 883-86. https://doi.org/10.1038/nature06284.

Houot, Laetitia, Sarah Chang, Cedric Absalon, et Paula I Watnick. 2010. « *Vibrio Cholerae* Phosphoenolpyruvate Phosphotransferase System Control of Carbohydrate Transport, Biofilm Formation, and Colonization of the Germfree Mouse Intestine ». *Infection and Immunity* 78 (4): 1482–1494. https://doi.org/10.1128/IAI.01356-09.

Houot, Laetitia, Sarah Chang, Bradley S Pickering, Cedric Absalon, et Paula I Watnick. 2010. « The Phosphoenolpyruvate Phosphotransferase System Regulates *Vibrio Cholerae* Biofilm Formation through Multiple Independent Pathways ». *Journal of Bacteriology* 192 (12): 3055–3067. https://doi.org/10.1128/JB.00213-10.

Houot, Laetitia, et Paula I Watnick. 2008. « A Novel Role for Enzyme I of the *Vibrio Cholerae* Phosphoenolpyruvate Phosphotransferase System in Regulation of Growth in a Biofilm ». *Journal of Bacteriology* 190 (1): 311–320. https://doi.org/10.1128/JB.01410-07.

Hsiao, Ansel, A. M. Shamsir Ahmed, Sathish Subramanian, Nicholas W. Griffin, Lisa L. Drewry, William A. Petri, Rashidul Haque, Tahmeed Ahmed, et Jeffrey I. Gordon. 2014. « Members of the Human Gut Microbiota Involved in Recovery from *Vibrio Cholerae* Infection ». *Nature* 515 (7527): 423-26. https://doi.org/10.1038/nature13738.

Huddleston, Jennifer R. 2014. « Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes ». *Infection and Drug Resistance* 7 (juin): 167-76. https://doi.org/10.2147/IDR.S48820.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, et Loren J. Hauser. 2010. « Prodigal: Prokaryotic Gene Recognition and Translation

Initiation Site Identification ». *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Jayaraman, R. 2011. « Hypermutation and Stress Adaptation in Bacteria ». *Journal of Genetics* 90 (2): 383-91. https://doi.org/10.1007/s12041-011-0086-6.

Jolivet-Gougeon, Anne, Bela Kovacs, Sandrine Le Gall-David, Hervé Le Bars, Latifa Bousarghin, Martine Bonnaure-Mallet, Bernard Lobel, François Guillé, Claude-James Soussy, et Peter Tenke. 2011. « Bacterial hypermutation: clinical implications ». *Journal of Medical Microbiology,* 60 (5): 563-73. https://doi.org/10.1099/jmm.0.024083-0.

Jorth, Peter, Benjamin J. Staudinger, Xia Wu, Katherine B. Hisert, Hillary Hayden, Jayanthi Garudathri, Christopher L. Harding, et al. 2015. « Regional Isolation Drives Bacterial Diversification within Cystic Fibrosis Lungs ». *Cell Host & Microbe* 18 (3): 307-19. https://doi.org/10.1016/j.chom.2015.07.006.

Kamada, Nobuhiko, Yun-Gi Kim, Ho Pan Sham, Bruce A. Vallance, José L. Puente, Eric C. Martens, et Gabriel Núñez. 2012. « Regulated Virulence Controls the Ability of a Pathogen to Compete with the Gut Microbiota ». *Science* 336 (6086): 1325-29. https://doi.org/10.1126/science.1222195.

Kang, Dongwan, Feng Li, Edward S Kirton, Ashleigh Thomas, Rob S Egan, Hong An, et Zhong Wang. 2019. « MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies ». Preprint. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27522v1.

Karczewski, Jurgen, Freddy J. Troost, Irene Konings, Jan Dekker, Michiel Kleerebezem, Robert-Jan M. Brummer, et Jerry M. Wells. 2010. « Regulation of human epithelial tight junction proteins by *Lactobacillus plantarum* in vivo and protective effects on the epithelial barrier ». *American Journal of Physiology-Gastrointestinal and Liver Physiology* 298 (6): G851-59. https://doi.org/10.1152/ajpgi.00327.2009.

Karlsson, Elinor K., Jason B. Harris, Shervin Tabrizi, Atiqur Rahman, Ilya Shlyakhter, Nick Patterson, Colm O'Dushlaine, Stephen F. Schaffner, Sameer Gupta, et Fahima Chowdhury. 2013. « Natural selection in a bangladeshi population from the cholera-endemic ganges river delta ». *Science translational medicine* 5 (192): 192ra86–192ra86.

Katz, Lee S., Aaron Petkau, John Beaulaurier, Shaun Tyler, Elena S. Antonova, Maryann A. Turnsek, Yan Guo, et al. 2013. « Evolutionary Dynamics of *Vibrio Cholerae* O1

Following a Single-Source Introduction to Haiti ». *MBio* 4 (4): e00398-13. https://doi.org/10.1128/mBio.00398-13.

Kaur, Sumanpreet, Preeti Sharma, Namarta Kalia, Jatinder Singh, et Sukhraj Kaur. 2018. « Anti-Biofilm Properties of the Fecal Probiotic Lactobacilli Against *Vibrio Spp.* » *Frontiers in Cellular and Infection Microbiology* 8. https://doi.org/10.3389/fcimb.2018.00120.

Kendall, Emily A, Fahima Chowdhury, Yasmin Begum, Ashraful I Khan, Shan Li, James H Thierer, Jason Bailey, et al. 2010. « Relatedness of *Vibrio Cholerae* O1/O139 Isolates from Patients and Their Household Contacts, Determined by Multilocus Variable-Number Tandem-Repeat Analysis. » *Journal of Bacteriology* 192 (17): 4367–4376. https://doi.org/10.1128/JB.00698-10.

Kennemann, Lynn, Xavier Didelot, Toni Aebischer, Stefanie Kuhn, Bernd Drescher, Marcus Droege, Richard Reinhardt, et al. 2011. « *Helicobacter Pylori* Genome Evolution during Human Infection ». *Proceedings of the National Academy of Sciences* 108 (12): 5033-38. https://doi.org/10.1073/pnas.1018444108.

Keymer, D P, et A B Boehm. 2011. « Recombination Shapes the Structure of an Environmental *Vibrio Cholerae* Population ». *Applied and Environmental Microbiology* 77 (2): 537–544. https://doi.org/10.1128/AEM.02062-10.

Kim, Daehwan, Li Song, Florian P. Breitwieser, et Steven L. Salzberg. 2016. « Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences ». *Genome Research* 26 (12): 1721–1729. https://doi.org/10.1101/gr.210641.116.

King, Aaron A., Edward L. Ionides, Mercedes Pascual, et Menno J. Bouma. 2008. « Inapparent Infections and Cholera Dynamics ». *Nature* 454 (7206): 877-80. https://doi.org/10.1038/nature07084.

Koh, Ara, Filipe De Vadder, Petia Kovatcheva-Datchary, et Fredrik Bäckhed. 2016. « From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites ». *Cell* 165 (6): 1332-45. https://doi.org/10.1016/j.cell.2016.05.041.

Kosek, Margaret, Caryn Bern, et Richard L. Guerrant. 2003. « The Global Burden of Diarrhoeal Disease, as Estimated from Studies Published between 1992 and 2000 ». *Bulletin of the World Health Organization* 81 (janvier): 197-204. https://doi.org/10.1590/S0042-96862003000300010.

Kovatcheva-Datchary, Petia, Anne Nilsson, Rozita Akrami, Ying Shiuan Lee, Filipe De Vadder, Tulika Arora, Anna Hallen, Eric Martens, Inger Björck, et Fredrik Bäckhed. 2015. « Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of Prevotella ». *Cell Metabolism* 22 (6): 971-82. https://doi.org/10.1016/j.cmet.2015.10.001.

Kuhlmann, F. Matthew, Srikanth Santhanam, Pardeep Kumar, Qingwei Luo, Matthew A. Ciorba, et James M. Fleckenstein. 2016. « Blood Group O–Dependent Cellular Responses to Cholera Toxin: Parallel Clinical and Epidemiological Links to Severe Cholera ». *The American Journal of Tropical Medicine and Hygiene* 95 (2): 440-43. https://doi.org/10.4269/ajtmh.16-0161.

Kunkel, Thomas A., et Dorothy A. Erie. 2005. « Dna mismatch repair ». *Annual Review of Biochemistry* 74 (1): 681-710. https://doi.org/10.1146/annurev.biochem.74.082803.133243.

Labat, Francoise, Olivier Pradillon, Louis Garry, Michel Peuchmaur, Bruno Fantin, et Erick Denamur. 2005. « Mutator Phenotype Confers Advantage in *Escherichia Coli* Chronic Urinary Tract Infection Pathogenesis ». *FEMS Immunology & Medical Microbiology* 44 (3): 317-21. https://doi.org/10.1016/j.femsim.2005.01.003.

Langmead, Ben, et Steven L Salzberg. 2012. « Fast Gapped-Read Alignment with Bowtie 2. » *Nature Methods* 9 (4): 357–359. https://doi.org/10.1038/nmeth.1923.

LaRocque, R. C., P. Sabeti, P. Duggal, F. Chowdhury, A. I. Khan, L. M. Lebrun, J. B. Harris, E. T. Ryan, F. Qadri, et S. B. Calderwood. 2009. « A Variant in Long Palate, Lung and Nasal Epithelium Clone 1 Is Associated with Cholera in a Bangladeshi Population ». *Genes & Immunity* 10 (3): 267-72. https://doi.org/10.1038/gene.2009.2.

LaRocque, Regina C., Jason B. Harris, Michelle Dziejman, Xiaoman Li, Ashraful I. Khan, Abu S. G. Faruque, Shah M. Faruque, et al. 2005. « Transcriptional Profiling of *Vibrio Cholerae* Recovered Directly from Patient Specimens during Early and Late Stages of Human Infection ». *Infection and Immunity* 73 (8): 4488-93. https://doi.org/10.1128/IAI.73.8.4488-4493.2005.

Lee, Robyn S, Jean-François Proulx, Fiona McIntosh, Marcel A Behr, et William P Hanage. 2020. « Previously undetected super-spreading of Mycobacterium tuberculosis revealed by deep sequencing ». Édité par Miles P Davenport, Eduardo Franco, et Conor J Meehan. *eLife* 9 (février): e53245. https://doi.org/10.7554/eLife.53245.

Lee, Seung-Joo, Rou-Jia Sung, et Gregory L. Verdine. 2019. « Mechanism of DNA Lesion Homing and Recognition by the Uvr Nucleotide Excision Repair System ». Research article. Research. 2019. https://spj.sciencemag.org/research/2019/5641746/.

Leung, Daniel T., Fahima Chowdhury, Stephen B. Calderwood, Firdausi Qadri, et Edward T. Ryan. 2012. « Immune responses to cholera in children ». *Expert Review of Anti-infective Therapy* 10 (4): 435-44. https://doi.org/10.1586/eri.12.23.

Levade, Inès, Morteza M. Saber, Firas Midani, Fahima Chowdhury, Ashraful I. Khan, Yasmin A. Begum, Edward T. Ryan, et al. 2020. « Predicting *Vibrio Cholerae* Infection and Disease Severity Using Metagenomics in a Prospective Cohort Study ». *BioRxiv*, février, 2020.02.25.960930. https://doi.org/10.1101/2020.02.25.960930.

Levade, Inès, Yves Terrat, Jean-Baptiste Leducq, Ana A. Weil, Leslie M. Mayo-Smith, Fahima Chowdhury, Ashraful I. Khan, et al. 2017. « *Vibrio cholerae* genomic diversity within and between patients ». *Microbial Genomics* 3 (12). https://doi.org/10.1099/mgen.0.000142.

Levine, M. M., R. E. Black, M. L. Clements, L. Cisneros, D. R. Nalin, et C. R. Young. 1981. « Duration of Infection-Derived Immunity to Cholera ». *The Journal of Infectious Diseases* 143 (6): 818-20. https://doi.org/10.1093/infdis/143.6.818.

Levine, Myron M. 2010. « Immunogenicity and efficacy of oral vaccines in developing countries: lessons from a live cholera vaccine ». *BMC Biology* 8 (1): 129. https://doi.org/10.1186/1741-7007-8-129.

Lewnard, Joseph A., Marina Antillón, Gregg Gonsalves, Alice M. Miller, Albert I. Ko, et Virginia E. Pitzer. 2016. « Strategies to Prevent Cholera Introduction during International Personnel Deployments: A Computational Modeling Analysis Based on the 2010 Haiti Outbreak ». *PLOS Medicine* 13 (1): e1001947. https://doi.org/10.1371/journal.pmed.1001947.

Li, H et al. 2009. « The Sequence Alignment/Map Format and SAMtools ». *Bioinformatics* 25(16): 2078–2079.

Li, Dinghua, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, et Tak-Wah Lam. 2016. « MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven by Advanced Methodologies and Community Practices ». *Methods* 102: 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.

Liang, Dachao, Ross Ka-Kit Leung, Wenda Guan, et William W. Au. 2018. « Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities ». *Gut Pathogens*. https://doi.org/10.1186/s13099-018-0230-4.

Lieberman, Tami D, Kelly B Flett, Idan Yelin, Thomas R Martin, Alexander J McAdam, Gregory P Priebe, et Roy Kishony. 2013. « Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures ». *Nature Genetics* 46 (1): 82-87. https://doi.org/10.1038/ng.2848.

Lieberman, Tami D., Jean-Baptiste Michel, Mythili Aingaran, Gail Potter-Bynoe, Damien Roux, Michael R. Davis Jr, David Skurnik, et al. 2011. « Parallel Bacterial Evolution within Multiple Patients Identifies Candidate Pathogenicity Genes ». *Nature Genetics* 43 (12): 1275-80. https://doi.org/10.1038/ng.997.

Limmathurotsakul, Direk, Matthew T G Holden, Paul Coupland, Erin P Price, Narisara Chantratita, Vanaporn Wuthiekanun, Premjit Amornchai, Julian Parkhill, et Sharon J Peacock. 2014. « Microevolution of *Burkholderia Pseudomallei* during an Acute Infection. » *Journal of Clinical Microbiology* 52 (9): 3418–3421. https://doi.org/10.1128/JCM.01219-14.

Louis, Petra, et Harry J. Flint. 2017. « Formation of Propionate and Butyrate by the Human Colonic Microbiota ». *Environmental Microbiology* 19 (1): 29-41. https://doi.org/10.1111/1462-2920.13589.

Lovett, Susan T. 2011. « The DNA exonucleases of *Escherichia coli* ». *EcoSal Plus* 4 (2). https://doi.org/10.1128/ecosalplus.4.4.7.

Lu, A.-Lien, Xianghong Li, Yesong Gu, Patrick M. Wright, et Dau-Yin Chang. 2001. « Repair of Oxidative DNA Damage ». *Cell Biochemistry and Biophysics* 35 (2): 141-70. https://doi.org/10.1385/CBB:35:2:141.

Lugli, Gabriele Andrea, Christian Milani, Sabrina Duranti, Giulia Alessandri, Francesca Turroni, Leonardo Mancabelli, Danilo Tatoni, Maria Cristina Ossiprandi, Douwe van Sinderen, et Marco Ventura. 2019. « Isolation of novel gut bifidobacteria using a combination of metagenomic and cultivation approaches ». *Genome Biology* 20 (1): 96. https://doi.org/10.1186/s13059-019-1711-6.

Lutz, Carla, Martina Erken, Parisa Noorian, Shuyang Sun, et Diane McDougald. 2013. « Environmental Reservoirs and Mechanisms of Persistence of *Vibrio Cholerae* ». *Frontiers in Microbiology* 4. https://doi.org/10.3389/fmicb.2013.00375.

MacIntyre, Dana L., Sarah T. Miyata, Maya Kitaoka, et Stefan Pukatzki. 2010. « The *Vibrio Cholerae* Type VI Secretion System Displays Antimicrobial Properties ». *Proceedings of the National Academy of Sciences* 107 (45): 19520-24. https://doi.org/10.1073/pnas.1012931107.

Mack, David R., Sonia Michail, Shu Wei, Laura McDougall, et Michael A. Hollingsworth. 1999. « Probiotics inhibit enteropathogenic *E. coli* adherence in vitro by inducing intestinal mucin gene expression ». *American Journal of Physiology-Gastrointestinal and Liver Physiology* 276 (4): G941-50. https://doi.org/10.1152/ajpgi.1999.276.4.G941.

Madden, Tom. 2003. *The BLAST Sequence Analysis Tool. The NCBI Handbook [Internet]*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK21097/.

Mao, Ning, Andres Cubillos-Ruiz, D. Ewen Cameron, et James J. Collins. 2018. « Probiotic Strains Detect and Suppress Cholera in Mice ». *Science Translational Medicine* 10 (445). https://doi.org/10.1126/scitranslmed.aao2586.

Martin, Michael A., Robyn S. Lee, Lauren A. Cowley, Jennifer L. Gardy, et William P. Hanage. 2018. « Within-host *Mycobacterium tuberculosis* diversity and its utility for inferences of transmission ». *Microbial Genomics* 4 (10). https://doi.org/10.1099/mgen.0.000217.

Marvig, Rasmus Lykke, Helle Krogh Johansen, Søren Molin, et Lars Jelsbak. 2013. « Genome Analysis of a Transmissible Lineage of *Pseudomonas Aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators ». *PLOS Genetics* 9 (9): e1003741. https://doi.org/10.1371/journal.pgen.1003741.

Marvig, Rasmus Lykke, Lea Mette Sommer, Søren Molin, et Helle Krogh Johansen. 2015. « Convergent Evolution and Adaptation of *Pseudomonas Aeruginos*a within Patients with Cystic Fibrosis ». *Nature Genetics* 47 (1): 57-64. https://doi.org/10.1038/ng.3148.

Mathers, Amy J., Nicole Stoesser, Anna E. Sheppard, Louise Pankhurst, Adam Giess, Anthony J. Yeh, Xavier Didelot, et al. 2015. « *Klebsiella Pneumoniae* Carbapenemase (KPC)-Producing *K. Pneumoniae* at a Single Institution: Insights

into Endemicity from Whole-Genome Sequencing ». *Antimicrobial Agents and Chemotherapy* 59 (3): 1656-63. https://doi.org/10.1128/AAC.04292-14.

Matsuda, F., M. I. Chowdhury, A. Saha, T. Asahara, K. Nomoto, A. A. Tarique, T. Ahmed, M. Nishibuchi, A. Cravioto, et F. Qadri. 2011. « Evaluation of a Probiotics, Bifidobacterium Breve BBG-01, for Enhancement of Immunogenicity of an Oral Inactivated Cholera Vaccine and Safety: A Randomized, Double-Blind, Placebo-Controlled Trial in Bangladeshi Children under 5 Years of Age ». *Vaccine* 29 (10): 1855-58. https://doi.org/10.1016/j.vaccine.2010.12.133.

McAdam, Paul R., Anne Holmes, Kate E. Templeton, et J. Ross Fitzgerald. 2011. « Adaptive Evolution of *Staphylococcus aureus* during Chronic Endobronchial Infection of a Cystic Fibrosis Patient ». Édité par Iris Schrijver. *PLoS ONE* 6 (9): e24301. https://doi.org/10.1371/journal.pone.0024301.

McDonald, John H., et Martin Kreitman. 1991. « Adaptive Protein Evolution at the Adh Locus in Drosophila ». *Nature* 351 (6328): 652-54. https://doi.org/10.1038/351652a0.

Medellin-Peña, Maira J., et Mansel W. Griffiths. 2009. « Effect of Molecules Secreted by Lactobacillus Acidophilus Strain La-5 on *Escherichia Coli* O157:H7 Colonization ». *Applied and Environmental Microbiology* 75 (4): 1165-72. https://doi.org/10.1128/AEM.01651-08.

Mena, A., E. E. Smith, J. L. Burns, D. P. Speert, S. M. Moskowitz, J. L. Perez, et A. Oliver. 2008. « Genetic Adaptation of *Pseudomonas Aeruginosa* to the Airways of Cystic Fibrosis Patients Is Catalyzed by Hypermutation ». *Journal of Bacteriology* 190 (24): 7910-17. https://doi.org/10.1128/JB.01147-08.

Midani, Firas S., Ana A. Weil, Fahima Chowdhury, Yasmin A. Begum, Ashraful I. Khan, Meti D. Debela, Heather K. Durand, et al. 2018. « Human Gut Microbiota Predicts Susceptibility to *Vibrio Cholerae* Infection ». *The Journal of Infectious Diseases* 218 (4): 645-53. https://doi.org/10.1093/infdis/jiy192.

Moal, V. Liévin-Le, R. Amsellem, A. L. Servin, et M.-H. Coconnier. 2002. « *Lactobacillus Acidophilus* (Strain LB) from the Resident Adult Human Gastrointestinal Microflora Exerts Activity against Brush Border Damage Promoted by a Diarrhoeagenic *Escherichia Coli* in Human Enterocyte-like Cells ». *Gut* 50 (6): 803-11. https://doi.org/10.1136/gut.50.6.803.

Monira, Shirajum, Shota Nakamura, Kazuyoshi Gotoh, Kaori Izutsu, Haruo Watanabe, Nur Haque Alam, Takaaki Nakaya, et al. 2013. « Metagenomic profile of gut microbiota in children during cholera and recovery ». *Gut Pathogens* 5 (1): 1. https://doi.org/10.1186/1757-4749-5-1.

Morelli, Giovanna, Xavier Didelot, Barica Kusecek, Sandra Schwarz, Christelle Bahlawane, Daniel Falush, Sebastian Suerbaum, et Mark Achtman. 2010. « Microevolution of *Helicobacter pylori* during Prolonged Infection of Single Hosts and within Families ». Édité par Harmit S. Malik. *PLoS Genetics* 6 (7): e1001036. https://doi.org/10.1371/journal.pgen.1001036.

Moroni, Olivier, Ehab Kheadr, Yvan Boutin, Christophe Lacroix, et Ismaïl Fliss. 2006. « Inactivation of Adhesion and Invasion of Food-Borne Listeria Monocytogenes by Bacteriocin-Producing Bifidobacterium Strains of Human Origin ». *Applied and Environmental Microbiology* 72 (11): 6894-6901. https://doi.org/10.1128/AEM.00928-06.

Mutreja, Ankur, Dong Wook Kim, Nicholas R. Thomson, Thomas R. Connor, Je Hee Lee, Samuel Kariuki, Nicholas J. Croucher, et al. 2011. « Evidence for Several Waves of Global Transmission in the Seventh Cholera Pandemic ». *Nature* 477 (7365): 462-65. https://doi.org/10.1038/nature10392.

Nayfach, Stephen, Beltran Rodriguez-Mueller, Nandita Garud, et Katherine S. Pollard. 2016. « An Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial Transmission and Biogeography ». *Genome Research* 26 (11): 1612-25. https://doi.org/10.1101/gr.201863.115.

Nelson, Eric J., Jason B. Harris, J. Glenn Morris, Stephen B. Calderwood, et Andrew Camilli. 2009. « Cholera transmission: the host, pathogen and bacteriophage dynamic ». *Nature Reviews Microbiology* 7 (10): 693-702. https://doi.org/10.1038/nrmicro2204.

Noinaj, Nicholas, Maude Guillier, Travis J. Barnard, et Susan K. Buchanan. 2010. « TonB-Dependent Transporters: Regulation, Structure, and Function ». *Annual Review of Microbiology* 64 (1): 43-60. https://doi.org/10.1146/annurev.micro.112408.134247.

O'Hara, Brendan J., Zachary K. Barth, Amelia C. McKitterick, et Kimberley D. Seed. 2017. « A Highly Specific Phage Defense System Is a Conserved Feature of the *Vibrio Cholerae* Mobilome ». *PLOS Genetics* 13 (6): e1006838. https://doi.org/10.1371/journal.pgen.1006838.

Oliver, A., et A. Mena. 2010. « Bacterial Hypermutation in Cystic Fibrosis, Not Only for Antibiotic Resistance ». *Clinical Microbiology and Infection* 16 (7): 798-808. https://doi.org/10.1111/j.1469-0691.2010.03250.x.

Olivier, Verena, G. Kenneth Haines, Yanping Tan, et Karla J. Fullner Satchell. 2007. « Hemolysin and the Multifunctional Autoprocessing RTX Toxin Are Virulence Factors during Intestinal Infection of Mice with *Vibrio Cholerae* El Tor O1 Strains ». *Infection and Immunity* 75 (10): 5035-42. https://doi.org/10.1128/IAI.00506-07.

Olm, Matthew R., Alexander Crits-Christoph, Keith Bouma-Gregson, Brian Firek, Michael J. Morowitz, et Jillian F. Banfield. 2020. « InStrain Enables Population Genomic Analysis from Metagenomic Data and Rigorous Detection of Identical Microbial Strains ». *BioRxiv*, janvier, 2020.01.22.915579. https://doi.org/10.1101/2020.01.22.915579.

Olson, Rich, et Eric Gouaux. 2005. « Crystal Structure of the *Vibrio Cholerae* Cytolysin (VCC) Pro-Toxin and Its Assembly into a Heptameric Transmembrane Pore ». *Journal of Molecular Biology* 350 (5): 997-1016. https://doi.org/10.1016/j.jmb.2005.05.045.

Orata, Fabini D, Paul S Keim, et Yan Boucher. 2014. « The 2010 Cholera Outbreak in Haiti: How Science Solved a Controversy ». *PLOS Pathogens* 10 (4): e1003967. https://doi.org/10.1371/journal.ppat.1003967.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, et Julian Parkhill. 2015. « Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis ». *Bioinformatics* 31 (22): 3691-93. https://doi.org/10.1093/bioinformatics/btv421.

Paineau, Damien, Didier Carcano, Greg Leyer, Sylviane Darquy, Marie-Alexandra Alyanakian, Guy Simoneau, Jean-François Bergmann, Dominique Brassart, Francis Bornet, et Arthur C. Ouwehand. 2008. « Effects of Seven Potential Probiotic Strains on Specific Immune Responses in Healthy Adults: A Double-Blind, Randomized, Controlled Trial ». *FEMS Immunology & Medical Microbiology* 53 (1): 107-13. https://doi.org/10.1111/j.1574-695X.2008.00413.x.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, et Gene W. Tyson. 2015. « CheckM: Assessing the Quality of Microbial Genomes

Recovered from Isolates, Single Cells, and Metagenomes ». *Genome Research* 25 (7): 1043-55. https://doi.org/10.1101/gr.186072.114.

Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, et Nicola Segata. 2016. « Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights ». *PLOS Computational Biology* 12 (7): e1004977. https://doi.org/10.1371/journal.pcbi.1004977.

Patra, Rajashree, Santanu Chattopadhyay, Ronita De, Prachetash Ghosh, Mou Ganguly, Abhijit Chowdhury, T Ramamurthy, G B Nair, et Asish K Mukhopadhyay. 2012. « Multiple Infection and Microdiversity among *Helicobacter Pylori* Isolates in a Single Host in India. » *PLOS ONE* 7 (8): e43370. https://doi.org/10.1371/journal.pone.0043370.

Peng, Yu, Henry C M Leung, S M Yiu, et Francis Y L Chin. 2012. « IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth. » *Bioinformatics* 28(11): 1420–1428.

Pérez-Lago, Laura, Marta Herranz, Miguel Martínez Lirola, Emilio Bouza, et Darío García de Viedma. 2011. « Characterization of Microevolution Events in *Mycobacterium Tuberculosis* Strains Involved in Recent Transmission Clusters ». *Journal of Clinical Microbiology* 49 (11): 3771-76. https://doi.org/10.1128/JCM.01285-11.

Phelps, Matthew D., Lone Simonsen, et Peter K. M. Jensen. 2019. « Individual and Household Exposures Associated with Cholera Transmission in Case–Control Studies: A Systematic Review ». *Tropical Medicine & International Health* 24 (10): 1151-68. https://doi.org/10.1111/tmi.13293.

Price, E. P., D. S. Sarovich, M. Mayo, A. Tuanyok, K. P. Drees, M. Kaestli, S. M. Beckstrom-Sternberg, et al. 2013. « Within-Host Evolution of *Burkholderia pseudomallei* over a Twelve-Year Chronic Carriage Infection ». *mBio* 4 (4): e00388-13-e00388-13. https://doi.org/10.1128/mBio.00388-13.

Price, Erin P., Heidie M. Hornstra, Direk Limmathurotsakul, Tamara L. Max, Derek S. Sarovich, Amy J. Vogler, Julia L. Dale, et al. 2010. « Within-Host Evolution of *Burkholderia pseudomallei* in Four Cases of Acute Melioidosis ». *PLoS Pathogens* 6 (1). https://doi.org/10.1371/journal.ppat.1000725.

Price, Morgan N, Paramvir S Dehal, et Adam P Arkin. 2009. « FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix. » *Molecular Biology and Evolution* 26(7): 1641–1650.

Pybus, Oliver G., et Andrew Rambaut. 2009. « Evolutionary analysis of the dynamics of viral infectious disease ». *Nature Reviews Genetics* 10 (8): 540-50. https://doi.org/10.1038/nrg2583.

Qadri, F. et al. 1994. « Production, Characterization, and Application of Monoclonal Antibodies to *Vibrio Cholerae* O139 Synonym Bengal. » *Clinical and Diagnostic Laboratory Immunology* 1(1): 51-54.

Qadri, F., T. R. Bhuiyan, K. K. Dutta, R. Raqib, M. S. Alam, N. H. Alam, A.-M. Svennerholm, et M. M. Mathan. 2004. « Acute Dehydrating Disease Caused by *Vibrio Cholerae* Serogroups O1 and O139 Induce Increases in Innate Cells and Inflammatory Mediators at the Mucosal Surface of the Gut ». *Gut* 53 (1): 62-69. https://doi.org/10.1136/gut.53.1.62.

Qadri, Firdausi, Taufiqul Islam, et John D. Clemens. 2017. « Cholera in Yemen — An Old Foe Rearing Its Ugly Head ». *New England Journal of Medicine* 377 (21): 2005-7. https://doi.org/10.1056/NEJMp1712099.

Quince, Christopher, Alan W Walker, Jared T Simpson, Nicholas J Loman, et Nicola Segata. 2017. « Shotgun Metagenomics, from Sampling to Analysis ». *Nature Biotechnology* 35 (9): 833-44. https://doi.org/10.1038/nbt.3935.

Rabbani, G. H., M. John Albert, Hamidur Rahman, et Asis Kumar Chowdhury. 1999. « Short-Chain Fatty Acids Inhibit Fluid and Electrolyte Loss Induced by Cholera Toxin in Proximal Colon of Rabbit In Vivo ». *Digestive Diseases and Sciences* 44 (8): 1547-53. https://doi.org/10.1023/A:1026650624193.

Rafique, Raisa, Mahamud-Ur Rashid, Shirajum Monira, Zillur Rahman, Md Toslim Mahmud, Munshi Mustafiz, K M Saif-Ur-Rahman, et al. 2016. « Transmission of Infectious *Vibrio cholerae* through Drinking Water among the Household Contacts of Cholera Patients (CHoBI7 Trial) ». *Frontiers in microbiology* 7 (1120): 4096. https://doi.org/10.3389/fmicb.2016.01635.

Ram, Geeta, John Chen, Krishan Kumar, Hope F. Ross, Carles Ubeda, Priyadarshan K. Damle, Kristin D. Lane, José R. Penadés, Gail E. Christie, et Richard P. Novick. 2012. « Staphylococcal Pathogenicity Island Interference with Helper Phage Reproduction Is a Paradigm of Molecular Parasitism ». *Proceedings of the National Academy of Sciences* 109 (40): 16300-305. https://doi.org/10.1073/pnas.1204615109.

Ramos, Juan L, Manuel Martínez-Bueno, Antonio J Molina-Henares, Wilson Terán, Kazuya Watanabe, Xiaodong Zhang, María Trinidad Gallegos, Richard Brennan, et Raquel Tobes. 2005. « The TetR Family of Transcriptional Repressors ». *Microbiology and Molecular Biology Reviews* 69 (2): 326–356. https://doi.org/10.1128/MMBR.69.2.326-356.2005.

Rashed, Shah M, Andrew S Azman, Munirul Alam, Shan Li, David A Sack, J Glenn Morris Jr., Ira Longini, et al. 2014. « Genetic Variation of *Vibrio cholerae* during Outbreaks, Bangladesh, 2010–2011 ». *Emerging Infectious Diseases* 20 (1): 54–60. https://doi.org/10.3201/eid2001.130796.

Rau, Martin H, Rasmus Lykke Marvig, Garth D Ehrlich, Søren Molin, et Lars Jelsbak. 2012. « Deletion and Acquisition of Genomic Content during Early Stage Adaptation of *Pseudomonas Aeruginosa* to a Human Host Environment ». *Environmental Microbiology* 14 (8): 2200–2211. https://doi.org/10.1111/j.1462-2920.2012.02795.x.

Rea, Mary C., Clarissa S. Sit, Evelyn Clayton, Paula M. O'Connor, Randy M. Whittal, Jing Zheng, John C. Vederas, R. Paul Ross, et Colin Hill. 2010. « Thuricin CD, a Posttranslationally Modified Bacteriocin with a Narrow Spectrum of Activity against *Clostridium Difficile* ». *Proceedings of the National Academy of Sciences* 107 (20): 9352-57. https://doi.org/10.1073/pnas.0913554107.

Reeves, Peter R., Bin Liu, Zhemin Zhou, Dan Li, Dan Guo, Yan Ren, Connie Clabots, Ruiting Lan, James R. Johnson, et Lei Wang. 2011. « Rates of Mutation and Host Transmission for an *Escherichia Coli* Clone over 3 Years ». *PLOS ONE* 6 (10): e26907. https://doi.org/10.1371/journal.pone.0026907.

Reidl, Joachim, et Karl E Klose. 2002. « *Vibrio Cholerae* and Cholera: Out of the Water and into the Host ». *FEMS Microbiology Reviews* 26 (2): 125–139. https://doi.org/10.1111/j.1574-6976.2002.tb00605.x.

Rissman, Anna I et al. 2009. « Reordering Contigs of Draft Genomes Using the Mauve Aligner. » *Bioinformatics* 25(16): 2071–2073.

Rivera-Chávez, Fabian, et John J. Mekalanos. 2019. « Cholera Toxin Promotes Pathogen Acquisition of Host-Derived Nutrients ». *Nature* 572 (7768): 244-48. https://doi.org/10.1038/s41586-019-1453-3.

Rocha, Eduardo P C, John Maynard Smith, Laurence D Hurst, Matthew T G Holden, Jessica E Cooper, Noel H Smith, et Edward J Feil. 2006. « Comparisons of DN/DS

Are Time Dependent for Closely Related Bacterial Genomes. » *Journal of Theoretical Biology* 239 (2): 226–235. https://doi.org/10.1016/j.jtbi.2005.08.037.

Rogers, P. David, Teresa T. Liu, Katherine S. Barker, George M. Hilliard, B. Keith English, Justin Thornton, Edwin Swiatlo, et Larry S. McDaniel. 2007. « Gene Expression Profiling of the Response of *Streptococcus Pneumoniae* to Penicillin ». *Journal of Antimicrobial Chemotherapy* 59 (4): 616-26. https://doi.org/10.1093/jac/dkl560.

Rosch, Jason W., Beth Mann, Justin Thornton, Jack Sublett, et Elaine Tuomanen. 2008. « Convergence of Regulatory Networks on the Pilus Locus of *Streptococcus Pneumoniae* ». *Infection and Immunity* 76 (7): 3187-96. https://doi.org/10.1128/IAI.00054-08.

Ross, Michael G, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, et David B Jaffe. 2013. « Characterizing and Measuring Bias in Sequence Data ». *Genome Biology* 14 (5): R51. https://doi.org/10.1186/gb-2013-14-5-r51.

Sack, David A., Carol O. Tacket, Mitchell B. Cohen, R. Bradley Sack, Genevieve A. Losonsky, Janet Shimko, James P. Nataro, et al. 1998. « Validation of a Volunteer Model of Cholera with Frozen Bacteria as the Challenge ». *Infection and Immunity* 66 (5): 1968-72.

Sakib, S. Nazmus, Geethika Reddi, et Salvador Almagro-Moreno. 2018. « Environmental Role of Pathogenic Traits in Vibrio Cholerae ». *Journal of Bacteriology* 200 (15). https://doi.org/10.1128/JB.00795-17.

Satchell, Karla J. F., Christopher J. Jones, Jennifer Wong, Jessica Queen, Shivani Agarwal, et Fitnat H. Yildiz. 2016. « Phenotypic Analysis Reveals That the 2010 Haiti Cholera Epidemic Is Linked to a Hypervirulent Strain ». *Infection and Immunity* 84 (9): 2473-81. https://doi.org/10.1128/IAI.00189-16.

Schmidt, Thomas S. B., Jeroen Raes, et Peer Bork. 2018. « The Human Gut Microbiome: From Association to Modulation ». *Cell* 172 (6): 1198-1215. https://doi.org/10.1016/j.cell.2018.02.044.

Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, et Nicola Segata. 2016. « Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics ». *Nature Methods* 13 (5): 435-38. https://doi.org/10.1038/nmeth.3802.

Sealfon, Rachel, Stephen Gire, Crystal Ellis, Stephen Calderwood, Firdausi Qadri, Lisa Hensley, Manolis Kellis, Edward Ryan, Regina LaRocque, et Jason Harris. 2012. « High depth, whole-genome sequencing of cholera isolates from Haiti and the Dominican Republic ». *BMC genomics* 13 (1): 468.

Seed, Kimberley D., Kip L. Bodi, Andrew M. Kropinski, Hans-Wolfgang Ackermann, Stephen B. Calderwood, Firdausi Qadri, et Andrew Camilli. 2011. « Evidence of a Dominant Lineage of *Vibrio Cholerae*-Specific Lytic Bacteriophages Shed by Cholera Patients over a 10-Year Period in Dhaka, Bangladesh ». *MBio* 2 (1). https://doi.org/10.1128/mBio.00334-10.

Seed, Kimberley D., Shah M. Faruque, John J. Mekalanos, Stephen B. Calderwood, Firdausi Qadri, et Andrew Camilli. 2012. « Phase Variable O Antigen Biosynthetic Genes Control Expression of the Major Protective Antigen and Bacteriophage Receptor in *Vibrio cholerae* O1 ». Édité par Karla J. F. Satchell. *PLoS Pathogens* 8 (9): e1002917. https://doi.org/10.1371/journal.ppat.1002917.

Seed, Kimberley D., David W. Lazinski, Stephen B. Calderwood, et Andrew Camilli. 2013. « A Bacteriophage Encodes Its Own CRISPR/Cas Adaptive Response to Evade Host Innate Immunity ». *Nature* 494 (7438): 489-91. https://doi.org/10.1038/nature11927.

Seed, Kimberley D, Minmin Yen, B Jesse Shapiro, Isabelle J Hilaire, Richelle C Charles, Jessica E Teng, Louise C Ivers, Jacques Boncy, Jason B Harris, et Andrew Camilli. 2014. « Evolutionary Consequences of Intra-Patient Phage Predation on Microbial Populations. » *ELife* 3 (août): e03497. https://doi.org/10.7554/eLife.03497.

Seemann, Torsten. 2014. « Prokka: Rapid Prokaryotic Genome Annotation ». *Bioinformatics (Oxford, England)* 30 (14): 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Seemann, Torsten. 2015. « Snippy: fast bacterial variant calling from NGS reads ». 2015. https://github.com/tseemann/snippy.

Segata, Nicola. 2018. « On the Road to Strain-Resolved Comparative Metagenomics ». *MSystems* 3 (2): e00190–17, /msystems/3/2/msys.00190–17.atom. https://doi.org/10.1128/mSystems.00190-17.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, et Curtis Huttenhower. 2011. « Metagenomic biomarker discovery and

explanation ». *Genome Biology* 12 (6): R60. https://doi.org/10.1186/gb-2011-12-6-r60.

Sepúlveda Cisternas, Ignacio, Juan C. Salazar, et Víctor A. García-Angulo. 2018. « Overview on the Bacterial Iron-Riboflavin Metabolic Axis ». *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.01478.

Sepúlveda-Cisternas, Ignacio, Luis Lozano Aguirre, Andrés Fuentes Flores, Ignacio Vásquez Solis de Ovando, et Víctor Antonio García-Angulo. 2018. « Transcriptomics Reveals a Cross-Modulatory Effect between Riboflavin and Iron and Outlines Responses to Riboflavin Biosynthesis and Uptake in Vibrio Cholerae ». *Scientific Reports* 8 (1): 1-14. https://doi.org/10.1038/s41598-018-21302-3.

Shapiro, B. Jesse, Inès Levade, Gabriela Kovacikova, Ronald K. Taylor, et Salvador Almagro-Moreno. 2016. « Origins of Pandemic *Vibrio Cholerae* from Environmental Gene Pools ». *Nature Microbiology* 2 (3): 1-6. https://doi.org/10.1038/nmicrobiol.2016.240.

Shin, Sung Jae, Chia-wei Wu, Howard Steinberg, et Adel M Talaat. 2006. « Identification of Novel Virulence Determinants in *Mycobacterium Paratuberculosis* by Screening a Library of Insertional Mutants ». *Infection and Immunity* 74 (7): 3825–3833. https://doi.org/10.1128/IAI.01742-05.

Shoemaker, N B, H Vlamakis, K Hayes, et A A Salyers. 2001. « Evidence for Extensive Resistance Gene Transfer among *Bacteroides* Spp. and among *Bacteroides* and Other Genera in the Human Colon. » *Applied and Environmental Microbiology* 67 (2): 561–568. https://doi.org/10.1128/AEM.67.2.561-568.2001.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, et Jillian F. Banfield. 2018. « Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy ». *Nature Microbiology* 3 (7): 836-43. https://doi.org/10.1038/s41564-018-0171-1.

Silva, S.H., E.C. Vieira, R.S. Dias, et J.R. Nicoli. 2001. « Antagonism against *Vibrio cholerae* by diffusible substances produced by bacterial components of the human faecal microbiota ». *Journal of Medical Microbiology,* 50 (2): 161-64. https://doi.org/10.1099/0022-1317-50-2-161.

Smillie, Chris S., Mark B. Smith, Jonathan Friedman, Otto X. Cordero, Lawrence A. David, et Eric J. Alm. 2011. « Ecology Drives a Global Network of Gene Exchange

Connecting the Human Microbiome ». *Nature* 480 (7376): 241-44. https://doi.org/10.1038/nature10571.

Snitkin, Evan S., Adrian M. Zelazny, Jyoti Gupta, NISC Comparative Sequencing Program, Tara N. Palmore, Patrick R. Murray, et Julia A. Segre. 2013. « Genomic Insights into the Fate of Colistin Resistance and *Acinetobacter Baumannii* during Patient Treatment ». *Genome Research* 23 (7): 1155-62. https://doi.org/10.1101/gr.154328.112.

Snitkin, Evan S., Adrian M. Zelazny, Pamela J. Thomas, Frida Stock, NISC Comparative Sequencing Program, David K. Henderson, Tara N. Palmore, et Julia A. Segre. 2012. « Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella Pneumoniae* with Whole-Genome Sequencing ». *Science Translational Medicine* 4 (148): 148ra116-148ra116. https://doi.org/10.1126/scitranslmed.3004129.

Stanczak-Mrozek, Kinga I., Anusha Manne, Gwenan M. Knight, Katherine Gould, Adam A. Witney, et Jodi A. Lindsay. 2015. « Within-Host Diversity of MRSA Antimicrobial Resistances ». *Journal of Antimicrobial Chemotherapy* 70 (8): 2191-98. https://doi.org/10.1093/jac/dkv119.

Stecher, Glen, Koichiro Tamura, et Sudhir Kumar. 2020. « Molecular Evolutionary Genetics Analysis (MEGA) for MacOS ». *Molecular Biology and Evolution* 37 (4): 1237-39. https://doi.org/10.1093/molbev/msz312.

Sun, Yvonne, et Mary X. D. O'Riordan. 2013. « Chapter Three - Regulation of Bacterial Pathogenesis by Intestinal Short-Chain Fatty Acids ». In *Advances in Applied Microbiology*, édité par Sima Sariaslani et Geoffrey M. Gadd, 85:93-118. Academic Press. https://doi.org/10.1016/B978-0-12-407672-3.00003-4.

Tailford, Louise E., Emmanuelle H. Crost, Devon Kavanaugh, et Nathalie Juge. 2015. « Mucin Glycan Foraging in the Human Gut Microbiome ». *Frontiers in Genetics* 6. https://doi.org/10.3389/fgene.2015.00081.

Tajima, F. 1989. « Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. » *Genetics* 123 (3): 585-95.

Tamura, Koichiro et al. 2011. « MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods ». *Molecular Biology and Evolution* 28(10): 2731-39.

Taviani, Elisa, Christopher J Grim, Jongsik Chun, Anwar Huq, et R R Colwell. 2009. « Genomic Analysis of a Novel Integrative Conjugative Element in *Vibrio Cholerae*. » *FEBS Letters* 583 (22): 3630–3636. https://doi.org/10.1016/j.febslet.2009.10.041.

Tett, Adrian, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, et al. 2019. « The Prevotella Copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations ». *Cell Host & Microbe* 26 (5): 666-679.e7. https://doi.org/10.1016/j.chom.2019.08.018.

The UniProt Consortium. 2019. « UniProt: A Worldwide Hub of Protein Knowledge ». *Nucleic Acids Research* 47 (D1): D506-15. https://doi.org/10.1093/nar/gky1049.

Thomas, Andrew Maltez, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, et al. 2019. « Metagenomic Analysis of Colorectal Cancer Datasets Identifies Cross-Cohort Microbial Diagnostic Signatures and a Link with Choline Degradation ». *Nature Medicine* 25 (4): 667-78. https://doi.org/10.1038/s41591-019-0405-7.

Tischler, Anna D., et Andrew Camilli. 2004. « Cyclic diguanylate (c-di-GMP) regulates *Vibrio cholerae* biofilm formation ». *Molecular microbiology* 53 (3): 857-69. https://doi.org/10.1111/j.1365-2958.2004.04155.x.

Toft, Christina, et Siv G. E. Andersson. 2010. « Evolutionary Microbial Genomics: Insights into Bacterial Host Adaptation ». *Nature Reviews Genetics* 11 (7): 465-75. https://doi.org/10.1038/nrg2798.

Tomb, Jean-F., Owen White, Anthony R. Kerlavage, Rebecca A. Clayton, Granger G. Sutton, Robert D. Fleischmann, Karen A. Ketchum, et al. 1997. « The Complete Genome Sequence of the Gastric Pathogen Helicobacter Pylori ». *Nature* 388 (6642): 539-47. https://doi.org/10.1038/41483.

Toska, Jonida, Brian T. Ho, et John J. Mekalanos. 2018. « Exopolysaccharide Protects *Vibrio Cholerae* from Exogenous Attacks by the Type 6 Secretion System ». *Proceedings of the National Academy of Sciences* 115 (31): 7997-8002. https://doi.org/10.1073/pnas.1808469115.

Tribble, Gena D, Song Mao, Chloe E James, et Richard J Lamont. 2006. « A Porphyromonas Gingivalis Haloacid Dehalogenase Family Phosphatase Interacts with Human Phosphoproteins and Is Important for Invasion. » *Proceedings of the*

National Academy of Sciences 103 (29): 11027–11032. https://doi.org/10.1073/pnas.0509813103.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, et Nicola Segata. 2015. « MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling ». *Nature Methods* 12 (10): 902–903. https://doi.org/10.1038/nmeth.3589.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, et Nicola Segata. 2017. « Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes ». *Genome Research* 27 (4): 626–638. https://doi.org/10.1101/gr.216242.116.

Tsou, Amy M., et Jun Zhu. 2010. « Quorum Sensing Negatively Regulates Hemolysin Transcriptionally and Posttranslationally in Vibrio Cholerae ». *Infection and Immunity* 78 (1): 461-67. https://doi.org/10.1128/IAI.00590-09.

Ubeda, Carles, Ana Djukovic, et Sandrine Isaac. 2017. « Roles of the intestinal microbiota in pathogen protection ». *Clinical & Translational Immunology* 6 (2): e128. https://doi.org/10.1038/cti.2017.2.

Vezzulli, L., C. Grande, G. Tassistro, I. Brettar, M. G. Höfle, R. P. A. Pereira, D. Mushi, A. Pallavicini, P. Vassallo, et C. Pruzzo. 2017. « Whole-Genome Enrichment Provides Deep Insights into Vibrio Cholerae Metagenome from an African River ». *Microbial Ecology* 73 (3): 734-38. https://doi.org/10.1007/s00248-016-0902-x.

Wang, Baohong, Mingfei Yao, Longxian Lv, Zongxin Ling, et Lanjuan Li. 2017. « The Human Microbiota in Health and Disease ». *Engineering* 3 (1): 71-82. https://doi.org/10.1016/J.ENG.2017.01.008.

Wang, Bi-ying, Hui-qin Huang, Shuang Li, Ping Tang, Hao-fu Dai, Jian-an Xian, Dong-mei Sun, et Yong-hua Hu. 2019. « Thioredoxin H (TrxH) contributes to adversity adaptation and pathogenicity of Edwardsiella piscicida ». *Veterinary Research* 50 (1): 26. https://doi.org/10.1186/s13567-019-0645-z.

Wang, Hui, Xiaolin Xing, Jipeng Wang, Bo Pang, Ming Liu, Jessie Larios-Valencia, Tao Liu, et al. 2018. « Hypermutation-Induced in Vivo Oxidative Stress Resistance Enhances *Vibrio Cholerae* Host Adaptation ». *PLOS Pathogens* 14 (10): e1007413. https://doi.org/10.1371/journal.ppat.1007413.

Weil, Ana A., Rachel L. Becker, et Jason B. Harris. 2019. « *Vibrio Cholerae* at the Intersection of Immunity and the Microbiome ». *MSphere* 4 (6). https://doi.org/10.1128/mSphere.00597-19.

Weil, Ana A., Louise C. Ivers, et Jason B. Harris. 2011. « Cholera: Lessons from Haiti and Beyond ». *Current Infectious Disease Reports* 14 (1): 1-8. https://doi.org/10.1007/s11908-011-0221-9.

Weil, Ana A., Ashraful I. Khan, Fahima Chowdhury, Regina C. LaRocque, A. S. G. Faruque, Edward T. Ryan, Stephen B. Calderwood, Firdausi Qadri, et Jason B. Harris. 2009. « Clinical Outcomes in Household Contacts of Patients with Cholera in Bangladesh ». *Clinical Infectious Diseases* 49 (10): 1473-79. https://doi.org/10.1086/644779.

Weil, Ana A., et Edward T. Ryan. 2018. « Cholera: Recent Updates ». *Current Opinion in Infectious Diseases* 31 (5): 455–461. https://doi.org/10.1097/QCO.0000000000000474.

Weill, François-Xavier, Daryl Domman, Elisabeth Njamkepo, Abdullrahman A. Almesbahi, Mona Naji, Samar Saeed Nasher, Ankur Rakesh, et al. 2019. « Genomic Insights into the 2016–2017 Cholera Epidemic in Yemen ». *Nature* 565 (7738): 230-33. https://doi.org/10.1038/s41586-018-0818-3.

Weill, François-Xavier, Daryl Domman, Elisabeth Njamkepo, Cheryl Tarr, Jean Rauzier, Nizar Fawal, Karen H. Keddy, et al. 2017. « Genomic History of the Seventh Pandemic of Cholera in Africa ». *Science* 358 (6364): 785-89. https://doi.org/10.1126/science.aad5901.

Wilson, Daniel J. 2012. « Insights from Genomics into Bacterial Pathogen Populations ». *PLoS Pathog* 8 (9): e1002874. https://doi.org/10.1371/journal.ppat.1002874.

Wong, Alex, et Rees Kassen. 2011. « Parallel evolution and local differentiation in quinolone resistance in Pseudomonas aeruginosa ». *Microbiology,* 157 (4): 937-44. https://doi.org/10.1099/mic.0.046870-0.

Wood, Derrick E., Jennifer Lu, et Ben Langmead. 2019. « Improved Metagenomic Analysis with Kraken 2 ». *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.

Worby, Colin J, Marc Lipsitch, et William P Hanage. 2014. « Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic

Distance Data ». *PLoS computational biology* 10 (3): e1003549. https://doi.org/10.1371/journal.pcbi.1003549.

Wu, Yu-Wei, Blake A. Simmons, et Steven W. Singer. 2016. « MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets ». *Bioinformatics* 32 (4): 605–607. https://doi.org/10.1093/bioinformatics/btv638.

Yang, Menghua, Zhi Liu, Chambers Hughes, Andrew M. Stern, Hui Wang, Zengtao Zhong, Biao Kan, William Fenical, et Jun Zhu. 2013. « Bile Salt–Induced Intermolecular Disulfide Bond Formation Activates *Vibrio Cholerae* Virulence ». *Proceedings of the National Academy of Sciences* 110 (6): 2348-53. https://doi.org/10.1073/pnas.1218039110.

Yang, Wenjing, Yi Xiao, Xiangsheng Huang, Feidi Chen, Mingming Sun, Anthony J. Bilotta, Leiqi Xu, et al. 2019. « Microbiota Metabolite Short-Chain Fatty Acids Facilitate Mucosal Adjuvant Activity of Cholera Toxin through GPR43 ». *The Journal of Immunology* 203 (1): 282-92. https://doi.org/10.4049/jimmunol.1801068.

Yoon, Mi Young, Kyung Bae Min, Kang-Mu Lee, Yujin Yoon, Yaeseul Kim, Young Taek Oh, Keehoon Lee, et al. 2016. « A Single Gene of a Commensal Microbe Affects Host Susceptibility to Enteric Infection ». *Nature Communications* 7 (1): 1-11. https://doi.org/10.1038/ncomms11606.

Yoon, My Young, Keehoon Lee, et Sang Sun Yoon. 2014. « Protective Role of Gut Commensal Microbes against Intestinal Infections ». *Journal of Microbiology* 52 (12): 983-89. https://doi.org/10.1007/s12275-014-4655-2.

You, Jin Sun, Ji Hyun Yong, Gwang Hee Kim, Sungmin Moon, Ki Taek Nam, Ji Hwan Ryu, Mi Young Yoon, et Sang Sun Yoon. 2019. « Commensal-derived metabolites govern *Vibrio cholerae* pathogenesis in host intestine ». *Microbiome* 7 (1): 132. https://doi.org/10.1186/s40168-019-0746-y.

Young, Bernadette C., Tanya Golubchik, Elizabeth M. Batty, Rowena Fung, Hanna Larner-Svensson, Antonina A. Votintseva, Ruth R. Miller, Heather Godwin, Kyle Knox, et Richard G. Everitt. 2012. « Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease ». *Proceedings of the National Academy of Sciences* 109 (12): 4550–4555.

Zeibell, Krystle, Sharon Aguila, Vivian Yan Shi, Andrea Chan, Hanjing Yang, et Jeffrey H. Miller. 2007. « Mutagenesis and Repair in *Bacillus Anthracis*: The Effect of

Mutators ». *Journal of Bacteriology* 189 (6): 2331-38. https://doi.org/10.1128/JB.01656-06.

Zhao, Shijie, Tami D. Lieberman, Mathilde Poyet, Kathryn M. Kauffman, Sean M. Gibbons, Mathieu Groussin, Ramnik J. Xavier, et Eric J. Alm. 2019. « Adaptive Evolution within Gut Microbiomes of Healthy People ». *Cell Host & Microbe* 25 (5): 656-667.e8. https://doi.org/10.1016/j.chom.2019.03.007.

Zhou, Yi-Hui, et Paul Gallins. 2019. « A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction ». *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.00579.

Zimmermann, Petra, et Nigel Curtis. 2018. « The Influence of Probiotics on Vaccine Responses – A Systematic Review ». *Vaccine* 36 (2): 207-13. https://doi.org/10.1016/j.vaccine.2017.08.069.

Zolfo, Moreno, Adrian Tett, Olivier Jousson, Claudio Donati, et Nicola Segata. 2017. « MetaMLST: Multi-Locus Strain-Level Bacterial Typing from Metagenomic Samples ». *Nucleic Acids Research* 45 (2): e7. https://doi.org/10.1093/nar/gkw837.