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

AUTHORS: 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 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]Department of Biological Sciences, University of Montreal, Montreal, Quebec, 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

24 **CORRESPONDING AUTHOR** : jesse.shapiro@umontreal.ca

25

26 **RUNNING TITLE:** Stool metagenomics predicts cholera

ABSTRACT (word count: 196)

Background: 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.

Methods: 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.

Results: 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. Use of metagenomic features improved the precision and accuracy of prediction relative to 16S 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.

Conclusion: 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

MAIN TEXT (word count: 3407)

INTRODUCTION

Cholera is an acute diarrheal disease caused by *Vibrio cholerae*. Cholera 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 (1,2). Transmission of *V. cholerae* between household members commonly occurs through shared sources of contaminated food or water or through fecal-oral spread (3,4). The clinical spectrum of disease ranges from asymptomatic infection to severe watery diarrhea that can lead to fatal dehydration (5). 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 (6).

The composition of the gut bacterial community can protect against enteropathogenic infections (7), 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 (8). Animal and *in vitro* experiments have demonstrated that alteration of commensal-derived metabolite levels influenced host susceptibility by affecting *V. cholerae* growth or colonization (9-13).

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 (8,14). One study to date has characterized the gut microbiome of individuals exposed to *V. cholerae* to predict susceptibility to infection (15). 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 of

household contacts of patients who have cholera, a group known to have high risk of infection (4). Midani *et al* 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 sequencing does not allow identification of precise strains or the underlying genetic factors of the observed relationship.

To increase our understanding of the relationship between the gut microbiome and susceptibility to *V. cholerae*, we used shotgun metagenomics to analyze an expanded prospective cohort of persons exposed to *V. cholerae* in Bangladesh. We sequenced all DNA obtained from study participant rectal swabs (rather than only the 16S rRNA gene), and this resulted in improved predictive power, including identification of specific genes associated with remaining uninfected after exposure to *V. cholerae*. We also examined a larger cohort of samples to predict disease severity among infected contacts, albeit with lower power and precision than our susceptibility prediction. We also highlight several microbiome metabolic functions associated with protection against cholera.

METHODS

Sample collection, clinical outcomes and metagenomic sequencing

As described in (15), 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 (4,6). 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 (4,6). 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 Supplemental Methods.

Taxonomic/functional profiling and predictive model construction

We used MetaPhlAn2 (version 2.9) (16) for taxonomic profiling and HUMAnN2 (17) 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 (18), 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 (18). 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 (18). Further details are described in the Supplemental Methods.

Data availability.

After removal of human reads (Supplementary Methods), the sequence data has been deposited in NCBI under BioProject PRJNA608678.

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*. (15), 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 S1**) 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). 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.

We used the shotgun metagenomic DNA sequence reads from these samples 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 (18). 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.

Improved prediction compared to known factors impacting susceptibility

To put the metagenomic predictions in context, we next compared their predictive power and accuracy to clinical and demographic factors (**Table S1**). Three of these factors (age, baseline vibriocidal antibodies and blood group) are known to impact susceptibility to *V. cholerae* infection (6,15) (**Table S5**). The likelihood of developing infection was not predicted well using a RF model trained on the 7 features clinical and demographic factors (AUC=0.60, not significantly different from shuffled labels, 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.

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 (**Figure 3A**, **Figure S3A and Figure S4A**). 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 (**Figure 3B**, **Figure S3B and Figure S4B**), including three species of *Bifidobacterium* (indicated with asterisks in **Figure 3**). The top predictive species in contacts who developed symptomatic infection were *Clostridium ventriculi* (formerly *Sarcina ventriculi*), *Streptococcus parasanguinis* and members of the *Veillonella* genera. *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. 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 correctly predicting disease severity.

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 S7 and S8)**, 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 more likely to be associated with individuals who remained uninfected, while amino acid biosynthesis pathways and catabolic pathways were associated with individuals who developed infection (**Figures S9 and S10, Table S9**). We identified three iron-related genes associated with individuals that remained 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 (19-21). 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**).

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, the *Prevotella* genus was the major contributor to several iron metabolism related gene families (**Figure S12**). In other cases, the major contributors to protective gene families were unclassified (**Figure 5 and Figure S11**). 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.

DISCUSSION

Cholera continues to cause widespread disease in populations without access to safe water. 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.

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 data (15). 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(15). *Prevotella* species are hypothesized to be beneficial members of the microbiota in

292 healthy individuals in non-Westernized countries, and this species is a potential candidate for
293 follow up experimental studies in *V. cholerae* susceptibility (14,22,23).

294 Several species known to ferment mucin glycans into short chain fatty acids (SCFAs)
295 correlated with remaining uninfected, including *Eubacterium rectale*, *Ruminococcus gnavus* and
296 *Bacteroides vulgatus* (24,25). This finding is consistent with experiments of SCFAs applied to
297 animal models. *B. vulgatus* has been shown to inhibit *V. cholerae* colonization in mice, an effect
298 that was dependent upon SCFAs butyrate and propionate production (13). SCFAs are known to
299 impact immune cell development and attenuate inflammation by inhibiting histone deacetylases
300 and other mechanisms of altering gene expression (26-29).

301 All three *Bifidobacterium* species associated with contacts that developed infection were
302 also associated with asymptomatic rather than symptomatic disease (**Figure 3**), and prior work on
303 this genera supports several hypotheses for this relationship. First, *Bifidobacteria* are known to
304 produce the SCFA acetate that can protect against enteric infection in mice (33,34)(30). SCFAs
305 are also known to inhibit cholera toxin-related chloride secretion in the mouse gut, reducing
306 water and sodium loss, and have also been observed to increase cholera toxin-specific antibody
307 responses (31-33). *Bifidobacteria* are also major producers of lactate, a metabolite that has been
308 shown to impair *V. cholerae* biofilm formation, a function that can impact pathogen virulence
309 (12). Lastly, *B. bifidum* and *B. adolescentis* are known to reduce the activity of the *V. cholerae*
310 type VI secretion system through modification of bile acids (9).

311 The use of metagenomics also allowed us to identify bacterial functions that could impact
312 the ability of *V. cholerae* to compete and colonize the gut. For example, several gene families
313 involved in iron transport, iron regulation, and riboflavin conversion appeared among the top
314 twenty features associated with uninfected and asymptomatic individuals, suggesting that
315 competition for iron might be a protective mechanism of the gut microbiota against *V. cholerae*,

as is the case for other pathogens (7). Iron is often a limiting redox cofactor in the gut, and bacteria have evolved strategies to solubilize and internalize iron (34,35). 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 (34,36). 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 (37), 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. 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.

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, the Alfred P. Sloan Fellowship (LAD), 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 thank Meti Debela for technical assistance. 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). We declare that we have no competing financial interest.

ETHICAL STATEMENT

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

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

FIGURES AND TABLES

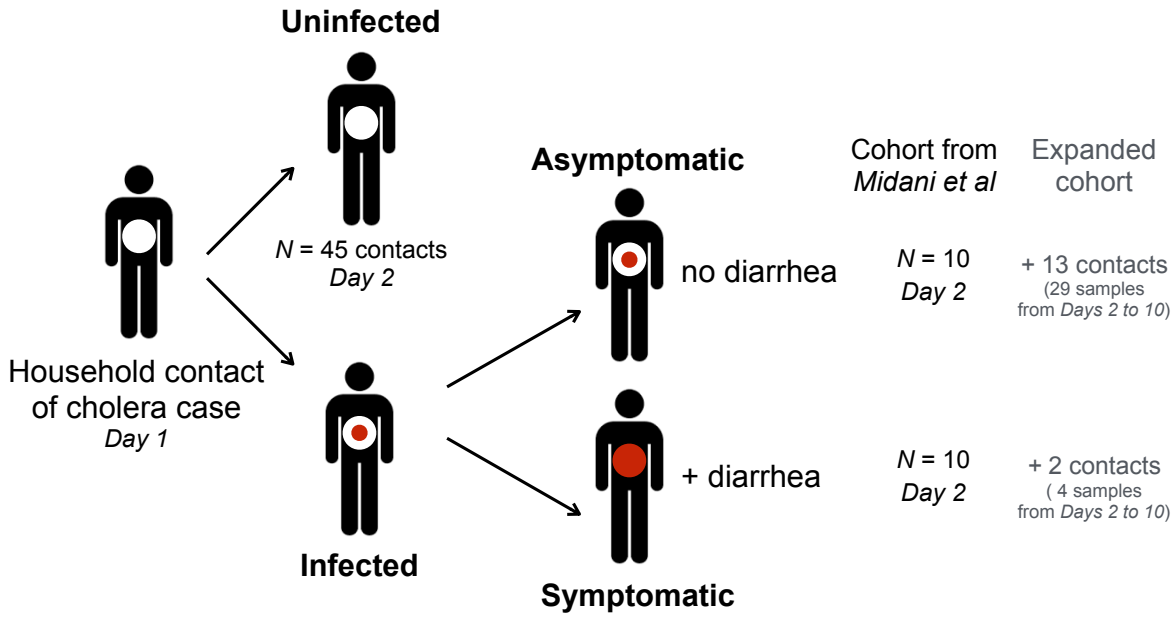
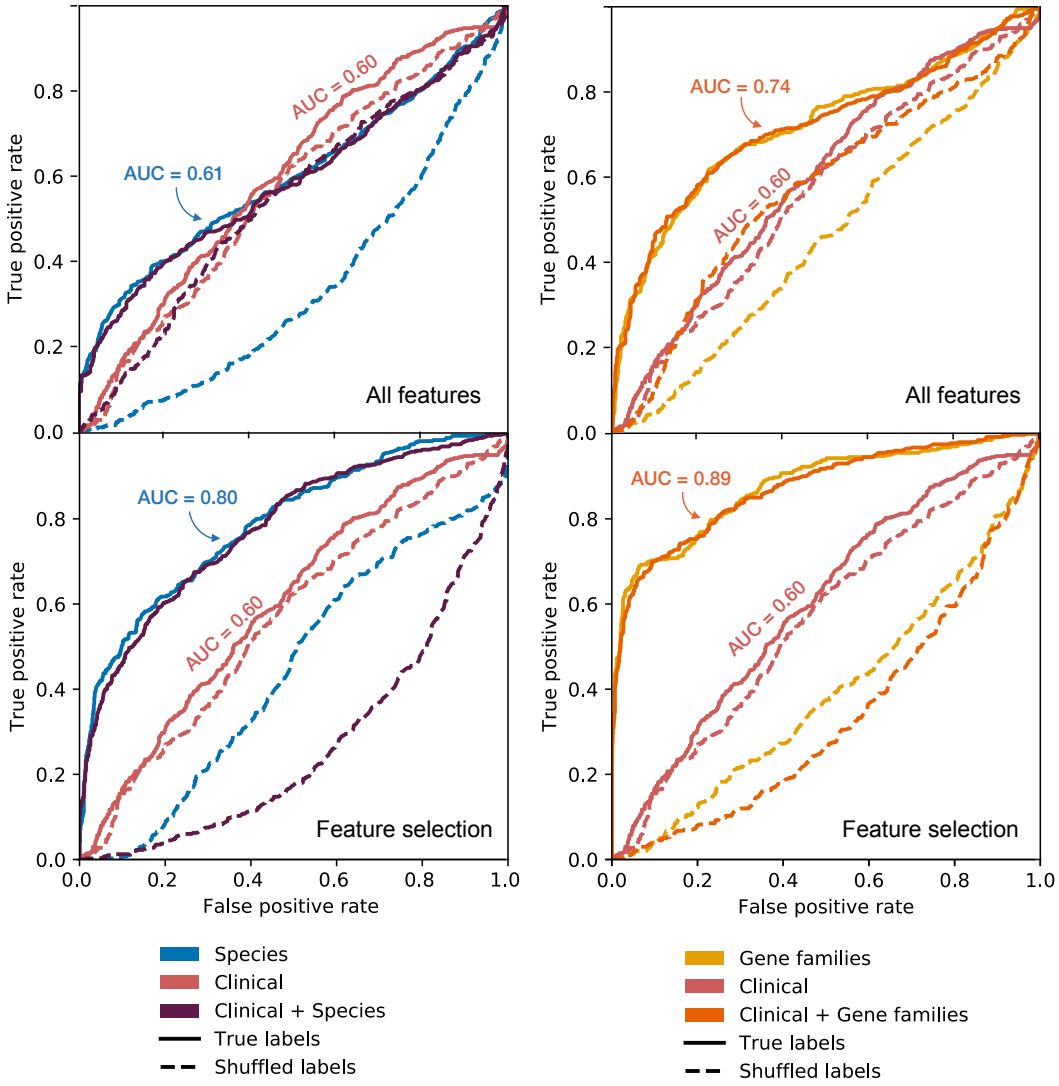


Figure 1. Study cohort in Dhaka, Bangladesh. After presentation of a *V. cholerae* culture-positive index case to the hospital on day 1, household contacts were enrolled on day 2. The expanded cohort includes the Midani 2018 cohort (15), with an addition of 33 samples from infected individuals (13 asymptomatic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



366 **Figure 2. Metagenomic features predict *V. cholerae* infection better than clinical and**
367 **demographic features.** Random forest prediction of infection status was applied to 7 clinical and
368 demographic features, and compared with all species and all gene families (top row), as well as
369 30 selected species features from metagenomes and 60 selected gene family features (bottom
370 row), or a combination of clinical, demographic and metagenomic features. Plots show receiver
371 operating characteristic (ROC) curves (average across cross-validations) for the Midani 2018
372 dataset. Shuffled labels represent the prediction run on a dataset with a random assignment of
373 infection outcomes. AUC = area under the curve.

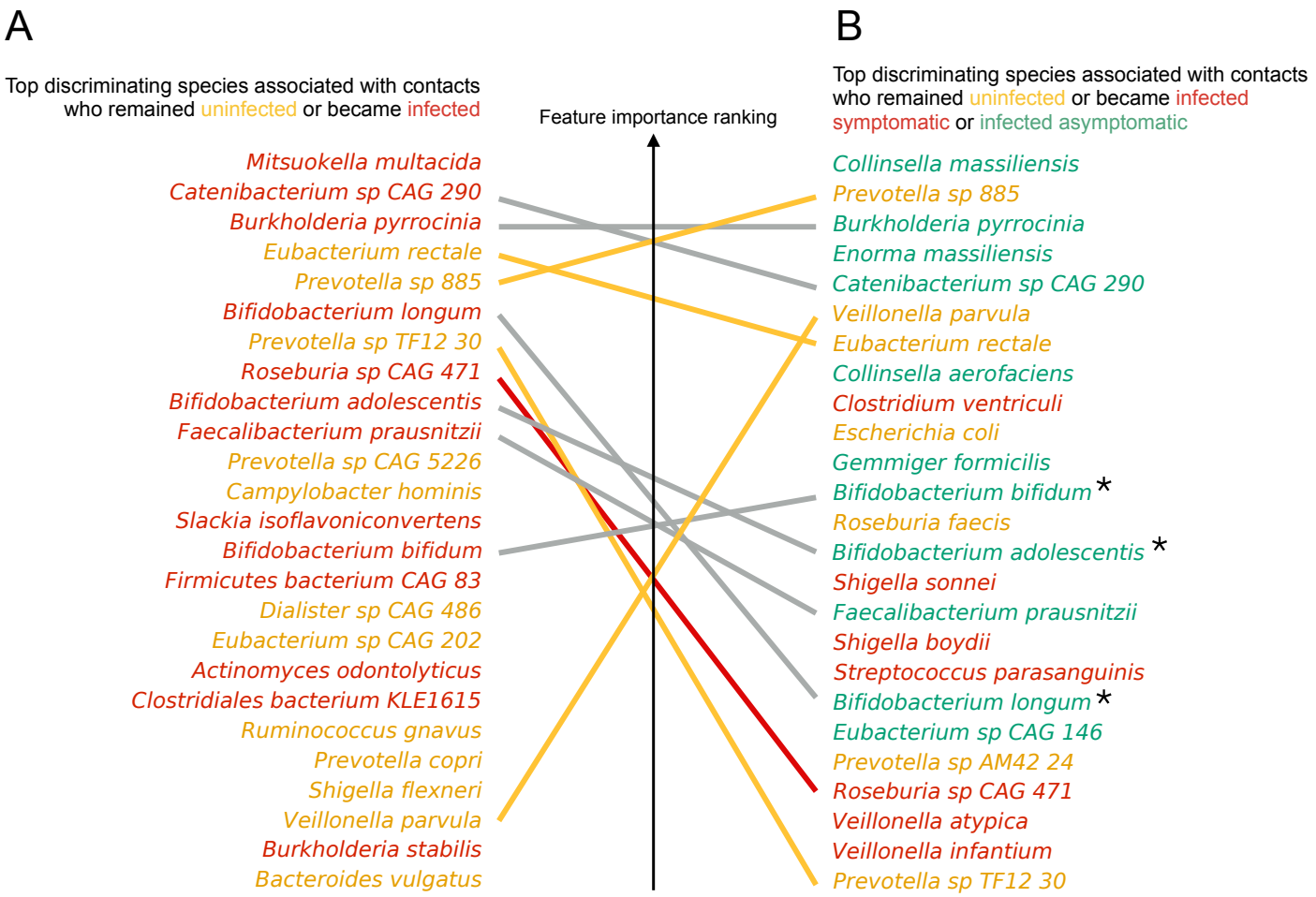


Figure 3. Most important discriminating species of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Species associated with contacts that became infected (red) or remained uninfected (yellow) during follow-up. (B) Species 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.

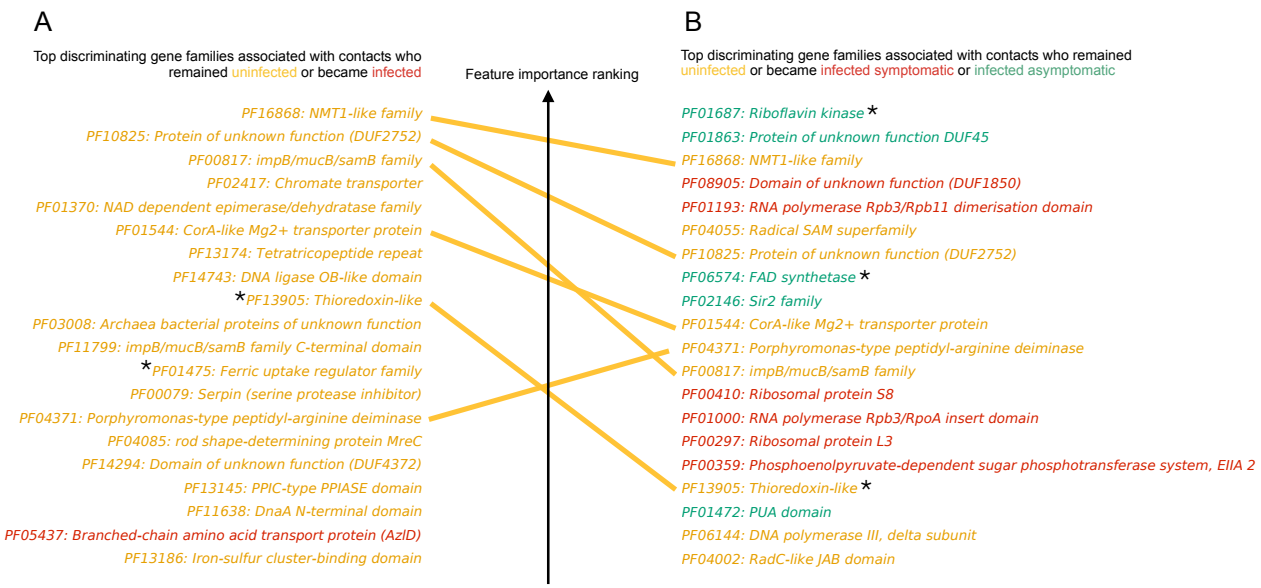
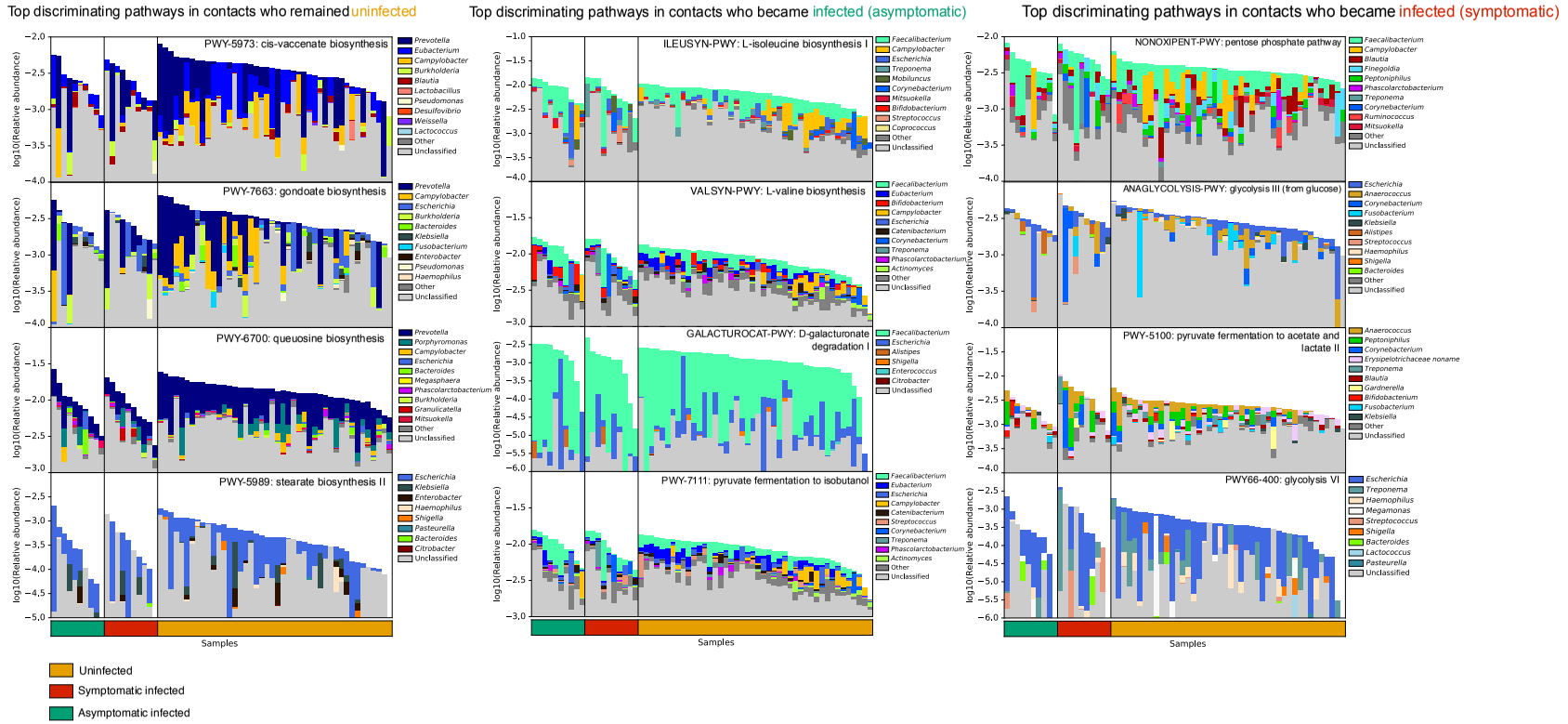


Figure 4. Most important discriminating gene families of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Genes families associated with contacts that became infected (red) or remained uninfected (yellow) during follow-up.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



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

Tables

		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.

REFERENCES

1. Ali M, Nelson AR, Lopez AL, Sack DA. Updated Global Burden of Cholera in Endemic Countries. PLoS Negl Trop Dis. **2015**; 9(6):e0003832.

2. Camacho A, Bouhenia M, Alyusfi R, et al. Cholera epidemic in Yemen, 2016-18: an analysis of surveillance data. Lancet Glob Health. **2018**; 6(6):e680–e690.

3. Domman D, Chowdhury F, Khan AI, et al. Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. Nat Genet. **2018**; :1–10.

4. Weil AA, Khan AI, Chowdhury F, et al. Clinical Outcomes in Household Contacts of Patients with Cholera in Bangladesh. Clin Infect Dis. **2009**; 49(10):1473–1479.

5. Nelson EJ, Harris JB, Morris JG, Calderwood SB, Camilli A. Cholera transmission: the host, pathogen and bacteriophage dynamic. Nature. **2009**; :1–10.

6. Harris JB, LaRocque RC, Chowdhury F, et al. Susceptibility to *Vibrio cholerae* Infection in a Cohort of Household Contacts of Patients with Cholera in Bangladesh. PLoS Negl Trop Dis. **2008**; 2(4):e221–8.

7. Ubeda C, Djukovic A, Isaac S. Roles of the intestinal microbiota in pathogen protection. Clinical & Translational Immunology. **2017**; 6(2).

8. Hsiao A, Ahmed AMS, Subramanian S, et al. Members of the human gut microbiota involved in recovery from *Vibrio cholerae* infection. Nature. **2014**; 515(7527):423–426.

9. Bachmann V, Kostiuk B, Unterweger D, Diaz-Satizabal L, Ogg S, Pukatzki S. Bile Salts Modulate the Mucin-Activated Type VI Secretion System of Pandemic *Vibrio cholerae*. PLoS Negl Trop Dis. **2015**; 9(8):e0004031.

10. Yoon MY, Min KB, Lee K-M, et al. A single gene of a commensal microbe affects host susceptibility to enteric infection. Nature Communications. **2016**; 7:1–11.

11. Mao N, Cubillos-Ruiz A, Cameron DE, Collins JJ. Probiotic strains detect and suppress cholera in mice. Science Translational Medicine. **2018**; 10(445).

12. Kaur S, Sharma P, Kalia N, Singh J, Kaur S. Anti-biofilm Properties of the Fecal Probiotic Lactobacilli Against Vibrio spp. Front Cell Infect Microbiol. **2018**; 8.

13. You JS, Yong JH, Kim GH, et al. Commensal-derived metabolites govern *Vibrio cholerae* pathogenesis in host intestine. Microbiome.. **2019**; 7(1):1–18.

14. David LA, Weil A, Ryan ET, et al. Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. mBio **2015**; 6(3):e00381–15.

15. Midani FS, Weil AA, Chowdhury F, et al. Human Gut Microbiota Predicts Susceptibility to *Vibrio cholerae* Infection. J Infect Dis. **2018**; 218(4):645–653.

16. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. **2015**; 12(10):902–903.

17. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods. **2018**; 15(11):962–968.

18. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. **2016**; 12(7):e1004977.

19. Fillat MF. The FUR (ferric uptake regulator) superfamily: diversity and versatility of key transcriptional regulators. Arch Biochem Biophys. **2014**; 546:41–52.

20. Wang B-Y, Huang H-Q, Li S, et al. Thioredoxin H (TrxH) contributes to adversity adaptation and pathogenicity of *Edwardsiella piscicida*. Vet Res. **2019**; 50(1):1–13.

21. Noinaj N, Guillier M, Barnard TJ, Buchanan SK. TonB-dependent transporters: regulation, structure, and function. Annu Rev Microbiol. **2010**; 64:43–60.

22. Tett A, Huang KD, Asnicar F, et al. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host and Microbe. **2019**; 26(5):666–679.e7.

23. Kovatcheva-Datchary P, Nilsson A, Akrami R, et al. Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of Prevotella. Cell Metab. **2015**; 22(6):971–982.

24. Crost EH, Tailford LE, Le Gall G, Fons M, Henrissat B, Juge N. Utilisation of mucin glycans by the human gut symbiont *Ruminococcus gnavus* is strain-dependent. PLoS ONE. **2013**; 8(10):e76341.

25. Tailford LE, Crost EH, Kavanaugh D, Juge N. Mucin glycan foraging in the human gut microbiome. Front Genet. **2015**; 6.

26. Sun Y, O'Riordan MXD. Regulation of bacterial pathogenesis by intestinal short-chain Fatty acids. Adv Appl Microbiol. **2013**; 85:93–118.

27. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. Cell. **2016**; 165(6):1332–1345.

28. Louis P, Flint HJ. Formation of propionate and butyrate by the human colonic microbiota. Environmental Microbiology. **2017**; 19(1):29–41.

29. Fachi JL, Souza Felipe J de, Pral LP, et al. Butyrate Protects Mice from *Clostridium difficile*-Induced Colitis through an HIF-1-Dependent Mechanism. Cell Reports. **2019**. p. 750–761.e7.

30. Fukuda S, Toh H, Hase K, et al. Bifidobacteria can protect from enteropathogenic infection through production of acetate. Nature. **2011**; 469(7331):543–547.

31. Canani RB, Costanzo MD, Leone L, Pedata M, Meli R, Calignano A. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol. **2011**; 17(12):1519–1528.

32. Yang W, Xiao Y, Huang X, et al. Microbiota Metabolite Short-Chain Fatty Acids Facilitate Mucosal Adjuvant Activity of Cholera Toxin through GPR43. The Journal of Immunology. **2019**; 203(1):282–292.

33. Rabbani GH, Albert MJ, Rahman H, Chowdhury AK. Short-Chain Fatty Acids Inhibit Fluid and Electrolyte Loss Induced by Cholera Toxin in Proximal Colon of Rabbit In Vivo. Dig Dis Sci. **1999** 44(8):1547–1553.

34. Sepúlveda Cisternas I, Salazar JC, García-Angulo VA. Overview on the Bacterial Iron-Riboflavin Metabolic Axis. Front Microbiol. **2018**; 9:1478.

35. Rivera-Chávez F, Mekalanos JJ. Cholera toxin promotes pathogen acquisition of host-derived nutrients. Nature. **2019**; 572(7768):244–248.

36. Sepúlveda Cisternas I, Aguirre LL, Flores AF, de Ovando IVS, García-Angulo VA. Transcriptomics reveals a cross-modulatory effect between riboflavin and iron and outlines responses to riboflavin biosynthesis and uptake in *Vibrio cholerae*. Scientific Reports. **2018**; 8(1):1–14.

37. Schmidt TSB, Raes J, Bork P. The Human Gut Microbiome: From Association to Modulation. Cell. **2018**; 172(6):1198–1215.