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 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[1,12,13*], Ana A. Weil[14]

[1]Department of Biological Sciences, University of Montreal, Montreal, QC, 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]Department of Microbiology and Immunology, McGill University, Montreal, QC, Canada

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

24 [14]Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, USA

25

26 **CORRESPONDING AUTHOR** : jesse.shapiro@mcgill.ca

27

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

29

30 **SUMMARY:** Cholera infection and disease severity can be predicted using metagenomic

31 sequencing of the gut microbiome pre-infection in a prospective cohort, and suggests potentially

32 protective bacterial species and genes.

ABSTRACT

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

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 (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 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). Other 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 few bacterial species, or involve patients who already have symptomatic cholera (8,14). One study recently characterized the gut microbiome of healthy individuals exposed to *V. cholerae* (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 known to have high risk of infection: household contacts of confirmed cholera patients (4). Midani *et al* showed that microbiome composition at the time *V. cholerae* exposure to can predict infection with similar or better

81 accuracy as commonly measured host factors known to impact susceptibility. However, 16S
82 rRNA sequencing has limited taxonomic resolution and does not identify the genetic mechanisms
83 of protection.

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

91 ## METHODS

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

93 As described in (15), household contacts were enrolled within 6 hours of the presentation
94 of an index cholera case at the icddr,b (International Center for Diarrheal Disease Research,
95 Bangladesh) Dhaka Hospital. Index patients with severe acute diarrhea, a stool culture positive
96 for *V. cholerae*, age between 2 and 60 years old, and no major comorbid conditions were
97 recruited (4,6). A clinical assessment of symptoms in household contacts was conducted daily for
98 the 10-day period after presentation of the index case, and repeated on day 30. We collected
99 demographic information, rectal swabs, and blood samples for ABO typing and vibriocidal
100 antibody titers as described in the Supplementary Methods. During the observation period,
101 contacts were determined to be infected if any rectal swab culture was positive for *V. cholerae*
102 and/or if the contact developed diarrhea and a 4-fold increase in vibriocidal titer during the
103 follow-up period (4,6). Contacts with positive rectal swabs developing watery diarrhea were
104 categorized as symptomatic and those without diarrhea were considered asymptomatic (**Figure**
105 **1**). *V. cholerae* positive contacts (by culture or deep 16S amplicon sequencing (15)) at the time of
106 enrollment were excluded, in addition to contacts who reported antibiotic use or diarrhea during
107 the week prior to enrollment. DNA extraction was performed for the selected samples and used
108 for shotgun metagenomics sequencing. Details on cohorts, sequencing methods and sample
109 processing are described in Supplementary Methods.

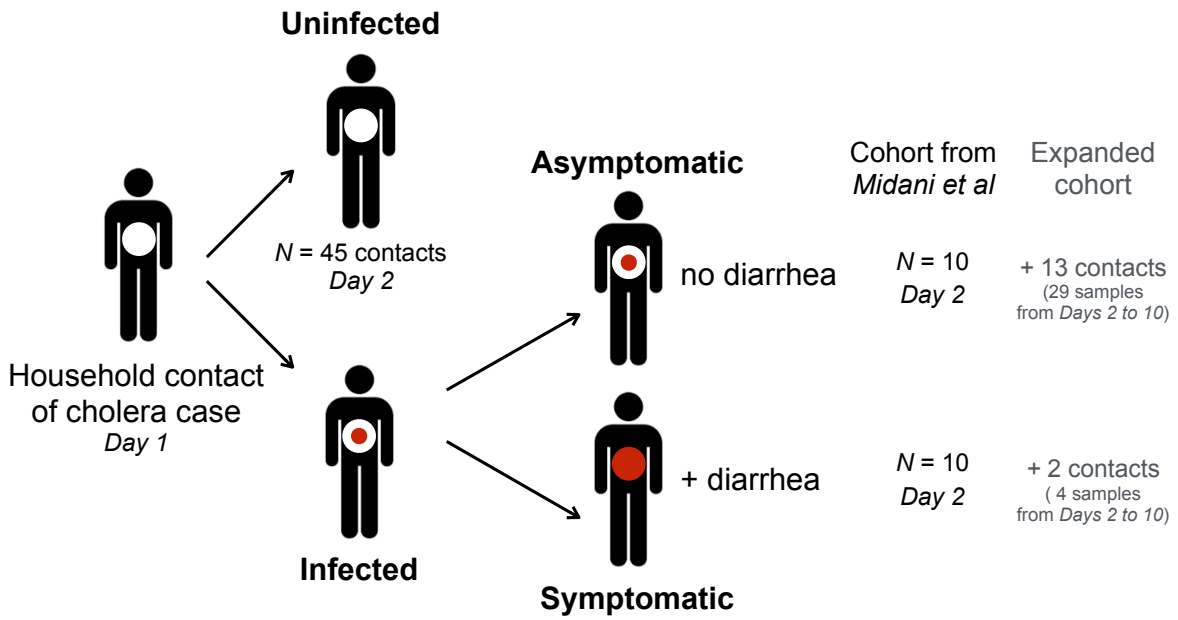


Figure 1. Study cohort in Dhaka, Bangladesh. After presentation of a *V. cholerae* culture-positive index case to the hospital on day 1, household contacts were enrolled on day 2. The expanded cohort includes the Midani 2018 cohort (15), with an addition of 33 samples from infected individuals (13 asymptomatic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

Taxonomic/functional profiling and predictive model construction

We used MetaPhlAn2 (version 2.9) (16) for taxonomic profiling and HUMAnN2 (17) to profile cellular pathways (from MetaCyc) and gene families (identified using the PFAM database). For identification of biomarkers of susceptibility and disease severity, we used MetAML (18) to apply 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 features types were compared 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 (18). We used a stratified 3-fold cross validation approach, splitting our dataset into validation and training sets (1/3 and 2/3 of samples, respectively) with the same infected:uninfected ratio. We used an embedded feature selection strategy to identify the most

useful features and improve model 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 Supplementary 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. 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 30-day follow-up period, 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, 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 place and time (2012-2014, Dhaka, Bangladesh). We used pre-infection samples in order to identify predictive features of disease outcomes in 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 samples. 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**).

		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.

164 **Predicting susceptibility to *V. cholerae* infection with Random Forests**

165 We first used an RF model to predict *V. cholerae* susceptibility (developing infection or
166 remaining uninfected) from baseline microbiome features (**Figure 1**). In the Midani 2018 cohort,
167 functional pathways and gene families predicted infection significantly better than random (Two-
168 sample *t*-tests comparing area under the curve [AUC] across 20 replicate 3 fold cross-validations;
169 $p < 0.05$) compared to data with shuffled (randomized) labels, and also predicted infection better
170 than species or strain features (**Tables 1 and S2**). Pathways and gene families had significantly
171 higher mean AUCs (0.71 and 0.74, respectively) compared to species or strains (0.61 and 0.62) (p
172 < 0.05; **Table 1**; **Figure S1, Table S3**).

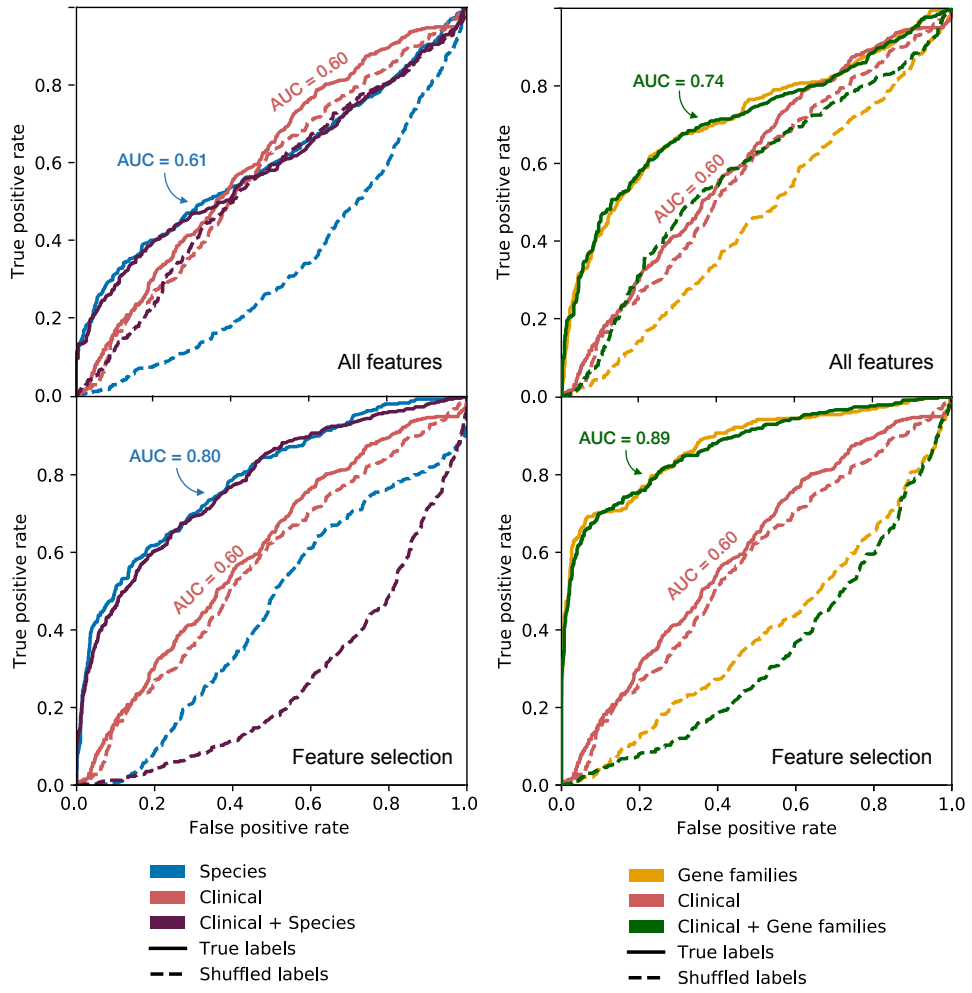
173 To determine the minimum number of metagenomic features required for prediction, we
174 repeated the analysis using smaller subsets of features. Using only 30 species, 60 gene families or
175 pathways, or 200 strains achieved similar cross-validation AUC values (**Figure S2**). We then
176 trained an RF model on this reduced number of selected features, yielding improved predictions
177 for all feature types (**Figure S1; Table S4**). This suggests that only a limited number of strains,
178 species, genes and pathways in the gut microbiome at the time of exposure are sufficient to
179 predict *V. cholerae* susceptibility. For example, prediction using strain-level markers after feature
180 selection yielded an AUC of 0.95 (**Table S4**). However, such high AUC values should be treated
181 with caution because the models can be overfit when a supervised feature selection step is applied
182 on the same data used to train the model (18). Because we did not have a fully independent
183 validation cohort (*e.g.* from another continent) to test our model, we decided to use the features
184 selected from the Midani cohort to make predictions on the Expanded dataset. Using the same
185 features selected from the Midani 2018 training dataset, we made predictions on the Expanded
186 cohort and achieved AUCs between 0.89 and 0.93 for prediction of infection using the four types
187 of features (**Table S4**). Again, because the expanded cohort partly overlaps with the Midani

188 cohort, and includes some repeated samples from the same individuals over time, these results
189 could also be prone to overfitting, but they demonstrate the potential for generalized predictions.
190 Finally, we repeated the RF analysis using all features in the expanded dataset, whichh
191 increased predictive performance relative to the original Midani cohort (**Figure S1**). Once again,
192 genes and pathways outperformed species and strains according to all metrics, with AUC
193 reaching ~0.88 using cellular pathways (**Table 1**). This improvement in the expanded cohort also
194 highlights the importance of using larger, more balanced datasets as input to predictive models.
195

196 **Improved prediction compared to known factors impacting susceptibility**

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

212 gene families clearly provide superior predictions, and adding clinical data did not improve the
213 predictions based on microbiome features alone (**Figure 2**). Together, these results demonstrate
214 that gene families present in the gut microbiome at the time of exposure contain more
215 information about *V. cholerae* susceptibility compared to species-level or clinical and
216 demographic contact data.



217

218 **Figure 2. Metagenomic features predict *V. cholerae* infection better than clinical and**
219 **demographic features.** Random forest prediction of infection status was applied to 7 clinical and
220 demographic features, and compared with all species and all gene families (top row), as well as
221 30 selected species features from metagenomes and 60 selected gene family features (bottom
222 row), or a combination of clinical, demographic and metagenomic features. Plots show receiver
223 operating characteristic (ROC) curves (average across cross-validations) for the Midani 2018
224 dataset. Shuffled labels represent the prediction run on a dataset with a random assignment of
225 infection outcomes. AUC = area under the curve.

226 **Disease severity is more difficult to predict than likelihood of infection**

227 To predict symptomatic disease among infected individuals (**Figure 1**), we divided
228 samples into uninfected, symptomatic and asymptomatic groups and again applied the RF
229 approach. We used the F1 score as a performance metric since it is well suited for uneven class
230 distributions in our uninfected/symptomatic/asymptomatic comparison. Applied to the Midani
231 2018 cohort, this model predicted outcomes significantly better than random (shuffled labels)
232 using species, strains or pathway data, but not gene families (**Table 1;** see **Table S3** for *p*-
233 values). However, the F1 scores for the symptomatic/asymptomatic predictions were
234 systematically lower (mean scores ranging from of 0.57 to 0.60) than for the infected/uninfected
235 prediction (means ranging from 0.64 to 0.71). Using the expanded cohort, the scores were
236 improved only slightly (**Table 1**). These results suggest that disease severity is predictable in
237 principle, but with greater uncertainty than the infection outcome.

238

239 **Taxonomic biomarkers of disease susceptibility and severity**

240 Predictive features in the gut microbiome identified to a species/strain or gene level allow
241 the possibility of experimental follow-up to investigate mechanisms of the associations we
242 observed. We characterized the most predictive species, pathways, and gene families (**Tables S6-**
243 **S9**). The most common discriminating species in individuals that remained uninfected during the
244 follow-up period were *Eubacterium rectale, Campylobacter hominis, Ruminococcus gnavus,*
245 *Bacteroides vulgatus, Veillonella parvula* and members of the *Prevotella* and *Eubacterium*
246 genera (**Figures 3A**, **S3A and S4A**). These species are ranked by their importance score, which
247 is effectively their relative weighting in the RF model. Several species associated with contacts
248 that developed *V. cholerae* infection belonged to the genera *Bifidobacterium, Actinomyces* or
249 *Collinsella*, and many of the species were also associated with asymptomatic infection (**Figures**

250 **3B**, **S3B and S4B**), including three species of *Bifidobacterium.* The top predictive species in
251 contacts who developed symptomatic infection were *Clostridium ventriculi* (formerly *Sarcina*
252 *ventriculi*), *Streptococcus parasanguinis* and members of *Veillonella*. *Shigella* species were also
253 associated with the gut microbiome of persons who developed symptomatic *V. cholerae*
254 infection, although persons enrolled in this study were *Shigella* stool-culture negative. *Shigella*
255 identified by DNA presence in stool may be the result of recent or resolving infection, or may be
256 present at subclinical levels due to ingestion of contaminated water. The features identified by the
257 multivariate RF model were confirmed using univariate statistics for the uninfected/infected
258 prediction (**Figure S5)**, but the overlap was poorer for the uninfected/symptomatic/asymptomatic
259 prediction (**Figure S6)**. This is consistent with the difficulty of predicting disease severity.
260 In general, the most important species were selected by the model because of differences
261 in relative abundance at baseline among uninfected/symptomatic/asymptomatic outcomes
262 (**Figure S7, S8**). In rare cases, species presence/absence information was predictive. For
263 example, *Ruminococcus gnavus,* is absent (near or below limit of detection) in most of the
264 individuals who become infected, but present in many (but not all) of those who remain
265 uninfected (**Figure S7**). Thus, there is no single, strong predictor of infection outcomes, but
266 rather a probabilistic combination of many species, each of relatively modest predictive value.

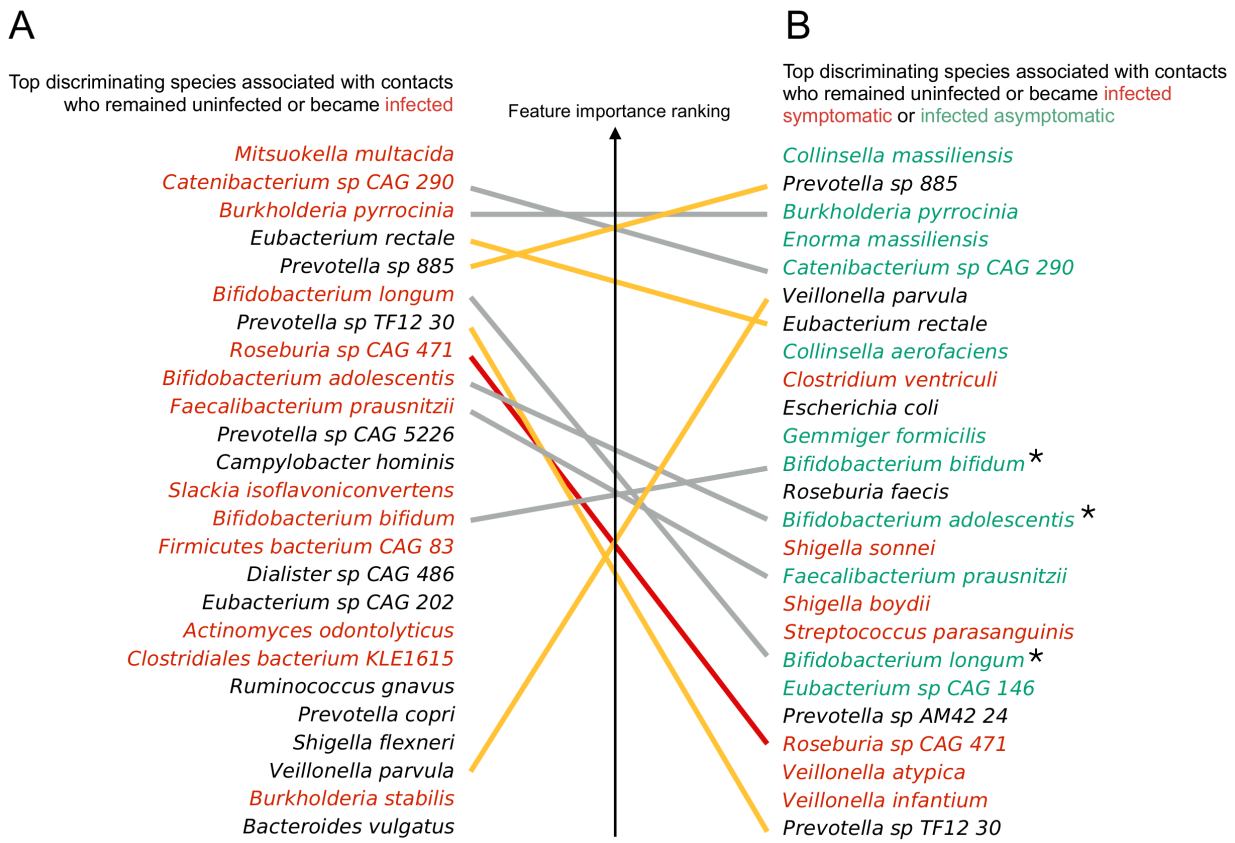


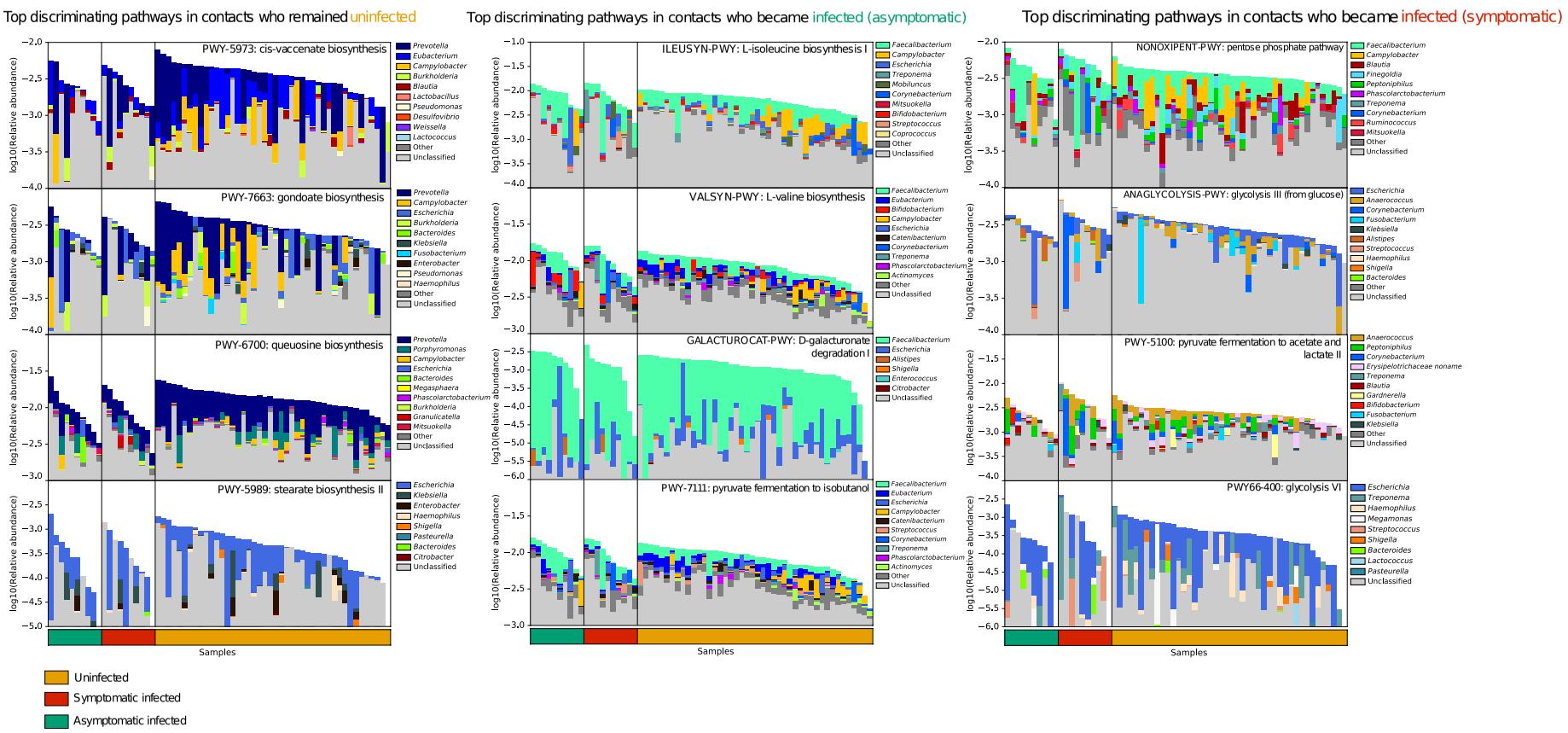
Figure 3. Most important discriminating species of the gut microbiome at the time of exposure to *V. cholerae* identified in the Midani 2018 dataset, classified by clinical outcome. (A) Species associated with contacts that became infected (red) or remained uninfected (black) during follow-up. (B) Species associated with contacts who remained uninfected (black), or became infected asymptomatic (green), or symptomatic (red) during follow-up. The top 25 most important features for discriminating between classes in the RF model 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.

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

284 pathways (*e.g.* cis-vaccenate, gondoate and stearate) were associated with individuals who
285 remained uninfected, while amino acid biosynthesis and catabolic pathways were associated with
286 individuals who developed infection (**Figures S11 and S12, Table S9**). We identified three iron-
287 related genes associated with remaining uninfected: (1) the ferric uptake regulator Fur, a major
288 regulator of iron homeostasis, (2) thioredoxin, a redox protein involved in adaptation to oxidative
289 and iron-deficiency stress, and (3) the TonB/ExbD/TolQR system, a ferric chelate transporter
290 (19-21). In individuals who became infected but asymptomatic, two genes involved in the
291 conversion of riboflavin into catalytically active cofactors, the riboflavin kinase and the FAD
292 synthetase, were found as the first and the third most discriminant features (**Figure 4, Table S8**).
293 We next asked which taxa in the microbiome likely encoded these genes. In some cases,
294 specific taxonomic groups corresponded to discrete gene functions. For example, several iron
295 metabolism-related gene families tend to be encoded by *Prevotella* genomes (**Figure S14**). In
296 other cases, the major contributors to protective gene families were unclassified (**Figures 5 and**
297 **S13**). These results partly explain why gene families or pathway features tend to outperform
298 species-level features in predicting infection status – because predictive gene families are
299 distributed across many species, including several with poor taxonomic annotation or families
300 lacking representation in taxonomic databases.
301

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 (black) during follow-up. (B) Genes families associated with contacts who remained uninfected (black), or became infected asymptomatic (green), or symptomatic (red) during follow-up. The top 25 most important features for discriminating between classes in the RF model 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.


Top discriminating pathways in contacts who remained uninfected
Top discriminating pathways in contacts who became infected (asymptomatic)
Top discriminating pathways in contacts who became infected (symptomatic)
PWY-5973: cis-vaccenate biosynthesis
PWY-7663: gondoate biosynthesis
PWY-6700: queuosine biosynthesis
PWY-5989: stearate biosynthesis II
ILEUSYN-PWY: L-isoleucine biosynthesis I
VALSYN-PWY: L-valine biosynthesis
GALACTUROCAT-PWY: D-galacturonate degradation I
PWY-7111: pyruvate fermentation to isobutanol
NONOXIPENT-PWY: pentose phosphate pathway
ANAGLYCOLYSIS-PWY: glycolysis III (from glucose)
PWY-5100: pyruvate fermentation to acetate and lactate II
PWY66-400: glycolysis VI
log10(Relative abundance)
Samples
Uninfected
Symptomatic infected
Asymptomatic infected

314 **Figure 5. Top predictive cellular pathways of the gut microbiome at the time of exposure to**
315 ***V. cholerae*** in the Midani 2018 cohort, annotated by their taxonomic contributors. The four top-
316 ranked pathways associated with uninfected contacts (left column), contacts who developed
317 asymptomatic infection (middle), and contacts who developed symptomatic infection (right
318 column) are shown. Total bar height reflects log10-scaled community relative abundance of each
319 pathways. The contributions of each genus to encoding these pathways are shown as stacked
320 colors within each bar, linearly scaled within the total. See Table S9 for the complete list of
321 pathways

322 **DISCUSSION**

323

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

334 Most of the top predictive biomarkers were associated with remaining uninfected after
335 exposure to *V. cholerae*. An example is the genus *Prevotella,* including several strains within
336 *Prevotella* sp. 885, identified only at the genus level in a previous study (15). *Prevotella* species
337 are hypothesized to be beneficial members of the microbiota in healthy individuals in non-
338 Westernized countries, and this species is a potential candidate for follow-up experimental
339 studies in *V. cholerae* susceptibility (14,22,23).

340 Several species known to ferment mucin glycans into short chain fatty acids (SCFAs)
341 correlated with remaining uninfected, including *Eubacterium rectale*, *Ruminococcus gnavus* and
342 *Bacteroides vulgatu*s (24,25). This finding is consistent with experiments of SCFAs applied to
343 animal models. *B. vulgatus* has been shown to inhibit *V. cholerae* colonization in mice, an effect
344 that was dependent upon SCFA production (13). SCFAs are known to impact immune cell

345 development and attenuate inflammation by inhibiting histone deacetylases and other
346 mechanisms of altering gene expression (26-29).
347 All three *Bifidobacterium* species associated with contacts that developed infection were
348 also associated with asymptomatic rather than symptomatic disease, and prior work on this genus
349 supports several hypotheses for this relationship. First, *Bifidobacteria* are known to produce the
350 SCFA acetate that can protect against enteric infection in mice (30,33,34). SCFAs are also known
351 to inhibit cholera toxin-related chloride secretion in the mouse gut, reducing water and sodium
352 loss, and have been observed to increase cholera toxin-specific antibody responses (31-33).
353 *Bifidobacteria* are also major producers of lactate, a metabolite that has been shown to impair *V.*
354 *cholerae* biofilm formation, a function that can impact virulence (12). Lastly, *B. bifidum* and *B.*
355 *adolescentis* are known to reduce the activity of *V. cholerae* type VI secretion systems through
356 modification of bile acids (9).
357 Metagenomics also allowed us to identify bacterial functions that could impact the ability
358 of *V. cholerae* to compete and colonize the gut. For example, several gene families involved in
359 iron transport, iron regulation, and riboflavin conversion appeared among the top twenty features
360 associated with uninfected and asymptomatic individuals, suggesting that competition for iron
361 might be a protective mechanism of the gut microbiota against *V. cholerae*, as in other pathogens
362 (7). Iron is often a limiting redox cofactor in the gut, and bacteria have evolved strategies to
363 solubilize and internalize iron (34,35). Riboflavin (another major redox cofactor in bacteria) and
364 iron levels are reciprocally regulated in *V. cholerae*, and riboflavin may allow *V. cholerae* to
365 overcome iron limitation in the gut (34,36). A gut microbiota more competitive for iron could
366 therefore help resist *V. cholerae* colonization or reduce its virulence. Further work is thus needed
367 to understand mechanisms of how the enrichment of these genes may protect people after
368 exposure to *V. cholerae*.

369 Our results are currently not generalizable beyond the study cohort in Dhaka, Bangladesh,
370 since a similar cohort in another geographic location is not available. As with any association-
371 based study (37), it is unknown if any of the metagenomic features that correlate with protection
372 from *V. cholerae* infection are causal, and many may be markers of clinical or environmental
373 factors that themselves impact susceptibility. Despite our deep sequencing and collection of
374 standard cholera risk factors, our study was unable to measure all potentially relevant
375 environmental or clinical risk factors. In line with recent studies in Dhaka, we assume that *V.*
376 *cholerae* transmission occurs mainly within households (3) and did not consider how the mode of
377 transmission (*e.g.* waterborne or not) might affect outcomes. It has also been noted that
378 microbiome-disease associations may be poorly portable across human populations (37). For
379 instance, we identified species of *Prevotella* as protective features in Bangladesh, but *Prevotella*
380 is much less abundant and less diverse in Western countries (22). It thus remains to be seen if
381 protective gene features (*e.g.* iron metabolism) are encoded in other species of the microbiome
382 outside endemic areas like Bangladesh, or if people outside these areas are simply at greater risk
383 for cholera. Further experimental characterization of metagenomic features correlated with
384 protection from infection or symptoms are needed to understand if factors we identified impact *V.*
385 *cholerae* pathogenesis or host responses to infection. Ultimately, the strains and functionalities
386 identified have the potential to inform microbiota-based therapeutics to ameliorate or prevent
387 disease. Our results show the power of metagenomic data from the gut microbiome to predict
388 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.

Supplementary tables S1-S9 are available at:
https://figshare.com/articles/Supplementary_Tables_-_Levade_et_al_2020/12440417

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.

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

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

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

456 18. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of
457 Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. **2016**;
458 12(7):e1004977.

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

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

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

465 22. Tett A, Huang KD, Asnicar F, et al. The Prevotella copri Complex Comprises Four
466 Distinct Clades Underrepresented in Westernized Populations. Cell Host and Microbe.
467 **2019**; 26(5):666–679.e7.

468 23. Kovatcheva-Datchary P, Nilsson A, Akrami R, et al. Dietary Fiber-Induced Improvement
469 in Glucose Metabolism Is Associated with Increased Abundance of Prevotella. Cell Metab.
470 **2015**; 22(6):971–982.

471 24. Crost EH, Tailford LE, Le Gall G, Fons M, Henrissat B, Juge N. Utilisation of mucin
472 glycans by the human gut symbiont *Ruminococcus gnavus* is strain-dependent. PLoS ONE.
473 **2013**; 8(10):e76341.

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

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

478 27. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From Dietary Fiber to Host
479 Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. Cell. **2016**;
480 165(6):1332–1345.

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

483 29. Fachi JL, Souza Felipe J de, Pral LP, et al. Butyrate Protects Mice from *Clostridium*
484 *difficile*-Induced Colitis through an HIF-1-Dependent Mechanism. Cell Reports. **2019**. p.
485 750–761.e7.

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

488 31. Canani RB, Costanzo MD, Leone L, Pedata M, Meli R, Calignano A. Potential beneficial
489 effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol. **2011**;
490 17(12):1519–1528.

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

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

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

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

501 36. Sepúlveda Cisternas I, Aguirre LL, Flores AF, de Ovando IVS, García-Angulo VA.
502 Transcriptomics reveals a cross-modulatory effect between riboflavin and iron and outlines
503 responses to riboflavin biosynthesis and uptake in *Vibrio cholerae*. Scientific Reports.
504 **2018**; 8(1):1–14.

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

507