

Metabolomic networks connect host-microbiome processes to human *Clostridioides difficile* infections

John I. Robinson, ... , Peter J. Mucha, Jeffrey P. Henderson

J Clin Invest. 2019. <https://doi.org/10.1172/JCI126905>.

Research Article

Gastroenterology

Infectious disease

Clostridioides difficile infection (CDI) accounts for a substantial proportion of deaths attributable to antibiotic-resistant bacteria in the United States. Although *C. difficile* can be an asymptomatic colonizer, its pathogenic potential is most commonly manifested in patients with antibiotic-modified intestinal microbiomes. In a cohort of 186 hospitalized patients, we showed that host and microbe-associated shifts in fecal metabolomes had the potential to distinguish patients with CDI from those with non-*C. difficile* diarrhea and *C. difficile* colonization. Patients with CDI exhibited a chemical signature of Stickland amino acid fermentation that was distinct from those of uncolonized controls. This signature suggested that *C. difficile* preferentially catabolizes branched chain amino acids during CDI. Unexpectedly, we also identified a series of noncanonical, unsaturated bile acids that were depleted in patients with CDI. These bile acids may derive from an extended host-microbiome dehydroxylation network in uninfected patients. Bile acid composition and leucine fermentation defined a prototype metabolomic model with potential to distinguish clinical CDI from asymptomatic *C. difficile* colonization.

Find the latest version:

<https://jci.me/126905/pdf>



Metabolomic networks connect host-microbiome processes to human *Clostridioides difficile* infections

John I. Robinson,¹ William H. Weir,² Jan R. Crowley,¹ Tiffany Hink,¹ Kimberly A. Reske,¹ Jennie H. Kwon,¹ Carey-Ann D. Burnham,³ Erik R. Dubberke,¹ Peter J. Mucha,² and Jeffrey P. Henderson¹

¹Center for Women's Infectious Disease Research, Division of Infectious Diseases, Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri, USA. ²Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, and Curriculum in Bioinformatics & Computational Biology, University of North Carolina, Chapel Hill, North Carolina, USA. ³Department of Pathology and Immunology, Washington University in St. Louis, St. Louis, Missouri, USA.

***Clostridioides difficile* infection (CDI) accounts for a substantial proportion of deaths attributable to antibiotic-resistant bacteria in the United States. Although *C. difficile* can be an asymptomatic colonizer, its pathogenic potential is most commonly manifested in patients with antibiotic-modified intestinal microbiomes. In a cohort of 186 hospitalized patients, we showed that host and microbe-associated shifts in fecal metabolomes had the potential to distinguish patients with CDI from those with non-*C. difficile* diarrhea and *C. difficile* colonization. Patients with CDI exhibited a chemical signature of Stickland amino acid fermentation that was distinct from those of uncolonized controls. This signature suggested that *C. difficile* preferentially catabolizes branched chain amino acids during CDI. Unexpectedly, we also identified a series of noncanonical, unsaturated bile acids that were depleted in patients with CDI. These bile acids may derive from an extended host-microbiome dehydroxylation network in uninfected patients. Bile acid composition and leucine fermentation defined a prototype metabolomic model with potential to distinguish clinical CDI from asymptomatic *C. difficile* colonization.**

Introduction

Each year in the United States, over 450,000 cases of *Clostridioides difficile* infection (CDI) are associated with over 29,000 associated deaths, with attributable costs of over \$2 billion (1). CDI is the most common healthcare-associated infection in US hospitals, and most cases start outside of the hospital setting (2). Although antimicrobial exposures are clearly a critical CDI risk factor, the mechanisms contributing to this association are incompletely understood.

CDIs arise following ingestion of *C. difficile* spores, which germinate in the intestinal tract and give rise to metabolically active, Gram-positive rods that colonize the colon. These vegetative forms secrete toxins whose effects upon the colonic epithelium give rise to a spectrum of intestinal symptoms ranging from diarrhea to a life-threatening pseudomembranous colitis. *C. difficile* persist with assistance from extensive antibiotic resistance that enables proliferation in patients whose intestinal microbiomes have been altered by broad-spectrum antibiotic exposure. Individual differences in CDI susceptibility and

severity are also substantial, such that some patients harboring *C. difficile* do not benefit from *C. difficile*-directed antibiotic therapies (3, 4). The mechanistic bases for these individual differences are poorly understood. Adaptations to different chemical environments in the intestine may markedly affect the pathogenic potential of *C. difficile* (5).

C. difficile is regarded as an opportunistic colonizer that is susceptible to suppression by healthy intestinal microbiomes. A number of candidate metabolic functions may contribute to this suppressive activity (3). One such function is the ability of healthy microbiomes to convert primary bile acids (e.g., cholic and chenodeoxycholic acids), which have been shown to promote spore germination in vitro, to secondary bile acids (e.g., deoxycholic and lithocholic acids, respectively). Both gnotobiotic or antibiotic-exposed mice exhibit diminished secondary bile acid production, though it is unclear which bile acid changes have causative, in addition to correlative, relationships with CDI susceptibility in humans (6–9). Recent work in an ex vivo murine model shows that secondary bile acids inhibit *C. difficile* germination and growth, although these effects are partially strain-specific (8, 9). Multiple bile acid transformation pathways are plausible in humans, who possess distinctive bile acids and microbiome compositions that may contribute to CDI risk in species-specific ways.

Current approaches to CDI diagnosis require compatible laboratory findings in the context of attributable clinical symptoms (e.g., diarrhea, abdominal pain, megacolon) (3). Two laboratory diagnostic approaches currently predominate in US hospitals, one based on nucleic acid amplification-based identification of toxigenic *C. difficile* and the other based on enzyme immunoas-

► **Related Commentary:** <https://doi.org/10.1172/JCI130008>

Conflict of interest: ERD receives grants from Rebiotix, grants and personal fees from Pfizer and Merck, and personal fees from Valneva, Rebiotix, Achaeaogen, Biofire, Abbott, and Synthetic Biologics. CDB receives grants from bioMerieux, Cepheid, and Luminex, grants and personal fees from Accelerate Diagnostics, and personal fees from BioRad and the Journal of Clinical Microbiology.

Submitted: January 2, 2019; **Accepted:** June 11, 2019; **Published:** August 12, 2019.

Reference information: J Clin Invest. <https://doi.org/10.1172/JCI126905>.

Table 1. Demographics of patient cohorts, including a summary of *C. difficile* ribotypes

Characteristic	CDI (toxigenic culture ⁺ , toxin EIA ⁺) n = 62	Asymptomatically colonized (toxigenic culture ⁺ , toxin EIA ⁺) n = 62	Not colonized (toxigenic culture ⁻ , toxin EIA ⁻) n = 62
Ward at stool collection			
Ward	50 (81%)	50 (81%)	50 (81%)
Oncology	9 (15%)	9 (15%)	9 (15%)
ICU	3 (5%)	3 (5%)	3 (5%)
Age, categorical			
≤45	10 (16%)	17 (27%)	8 (13%)
>45–≤65	20 (32%)	22 (36%)	19 (31%)
>65–≤85	27 (44%)	21 (34%)	28 (45%)
>85	5 (8%)	2 (3%)	7 (11%)
Female	35 (57%)	35 (57%)	29 (47%)
Race			
Caucasian	49 (79%)	37 (60%)	44 (71%)
African American	9 (15%)	22 (36%)	14 (23%)
Other	0 (0%)	1 (2%)	1 (2%)
Unknown	4 (7%)	2 (3%)	3 (5%)
Toxin status			
Toxin A	62 (100%)	62 (100%)	NA
Toxin B	62 (100%)	61 (98%)	NA
Binary toxin	26 (42%)	9 (15%)	NA
<i>C. difficile</i> strain^a			
027	22 (36%)	8 (13%)	NA
106/174	11 (18%)	6 (10%)	NA
WU27	6 (10%)	1 (2%)	NA
002	4 (7%)	3 (5%)	NA
001	4 (7%)	6 (10%)	NA
WU8	3 (5%)	1 (2%)	NA
014/020	2 (3%)	9 (15%)	NA
017	2 (3%)	2 (3%)	NA
WU2	2 (3%)	4 (7%)	NA
WU30	1 (2%)	2 (3%)	NA
WU36	0 (0%)	4 (7%)	NA
075	0 (0%)	4 (7%)	NA
015/046	0 (0%)	2 (3%)	NA
070	0 (0%)	2 (3%)	NA
Other strain ^b	5 (8%)	8 (13%)	NA

^a*C. difficile* strain is given as ribotype (3 digits), or WU strain type if there was no match to a Cardiff-ECDC collection ribotype strain. ^bListed as other strain if only a single isolate was recovered.

say detection of *C. difficile* exotoxins. The relative merits of these approaches have been debated (10). Direct detection of *C. difficile* (by culture or nucleic acid amplification) may be particularly susceptible to false-positive results for CDI due to detection of inactive spores, while there is concern that toxin detection tests are insufficiently sensitive and may yield false-negative results in some patients with CDI. Prospects for new approaches to improve diagnostic accuracy are therefore of interest.

To better understand the relationship between the intestinal metabolome and CDI in humans, we conducted fecal metabolomic profiling of hospitalized patients with diarrheal symptoms at an academic medical center. The cohort consists of patients with

a toxigenic culture positive for *C. difficile*, either with or without a positive toxin enzyme immunoassay (EIA) result, alongside matched, uncolonized controls. To characterize fecal metabolomes, we used untargeted gas chromatography–mass spectrometry (GC-MS), which permits robust chemical identification of metabolites and dietary compounds. Using multivariate analyses, we resolved multiple CDI-associated metabolites with microbial, host, and dietary origins. A distinctive short chain fatty acid (SCFA) series implicates extensive anaerobic amino acid metabolism by *C. difficile* in some colonized subjects. A novel, noncanonical bile acid correlation network not previously described in CDI susceptibility was also resolved. These and other results are consistent with numerous host-pathogen interactions that shape the relationship between patients and *C. difficile*. This allowed us to assemble a metabolomic definition of CDI from biochemical indices based on *C. difficile*-associated amino acid fermentation and host bile acid metabolism. These results may direct new CDI therapeutic and diagnostic efforts toward clinically relevant targets.

Results

Clinical cohort. Diarrheal specimens meeting inclusion and exclusion criteria were cultured for *C. difficile*. All *C. difficile* isolates recovered in culture were characterized for the presence of toxins *tcdA*, *tcdB*, *cdtA*, and *cdtB* by multiplex PCR, and underwent PCR ribotyping as previously described (11–14). Of the 8931 available stool specimens (Supplemental Figure 1; supplemental material available online with this article; <https://doi.org/10.1172/JCI126905DS1>), 2829 were eligible for chart review, through which an additional 2206 were excluded, yielding 622 stool specimens meeting inclusion and exclusion criteria. From these specimens, we assembled a 186-person cohort split into 3 groups of 62 patients matched by age and hospital location. These groups were defined by laboratory results: toxigenic culture–positive and toxin enzyme immunoassay–positive (using the Wampole/TechLab Tox A/B II assay during routine clinical testing, Cx⁺/EIA⁺), toxigenic culture–positive and toxin enzyme immunoassay–negative (Cx⁺/EIA⁻), and toxigenic culture–negative and toxin enzyme immunoassay–negative (Cx⁻/EIA⁻) controls. Cohort demographics and clinical characteristics are shown in Table 1.

Fecal metabolome characteristics. To characterize fecal metabolomic variations in the study cohort, we detected and quantified trimethylsilyl-derivatized fecal extracts using GC-MS. GC-MS is sensitive to low-molecular-weight analytes and does not detect proteins, peptides, complex lipids, or other macromolecules. We detected ions produced by electron ionization (EI), which often-times provides sufficient structure information to chemically identify metabolites of interest. Fecal metabolites may originate from human cells, microbiome, and/or diet. To compare metabolomes between specimens in the study population, GC-MS profiles were aligned so that each analyte (hereafter called a feature) is defined by its characteristic EI mass spectrum and GC retention time. Within the 186 patient specimens, we detected 2540 distinct features, 77 of which were removed as contaminants because they were present at comparable levels in multiple blank controls, leaving 2463 features for metabolomic analyses. These features were sparsely distributed with a heavy tail (Figure 1A), with only 593 features appearing in at least 8 (5%) specimens. The number of molec-

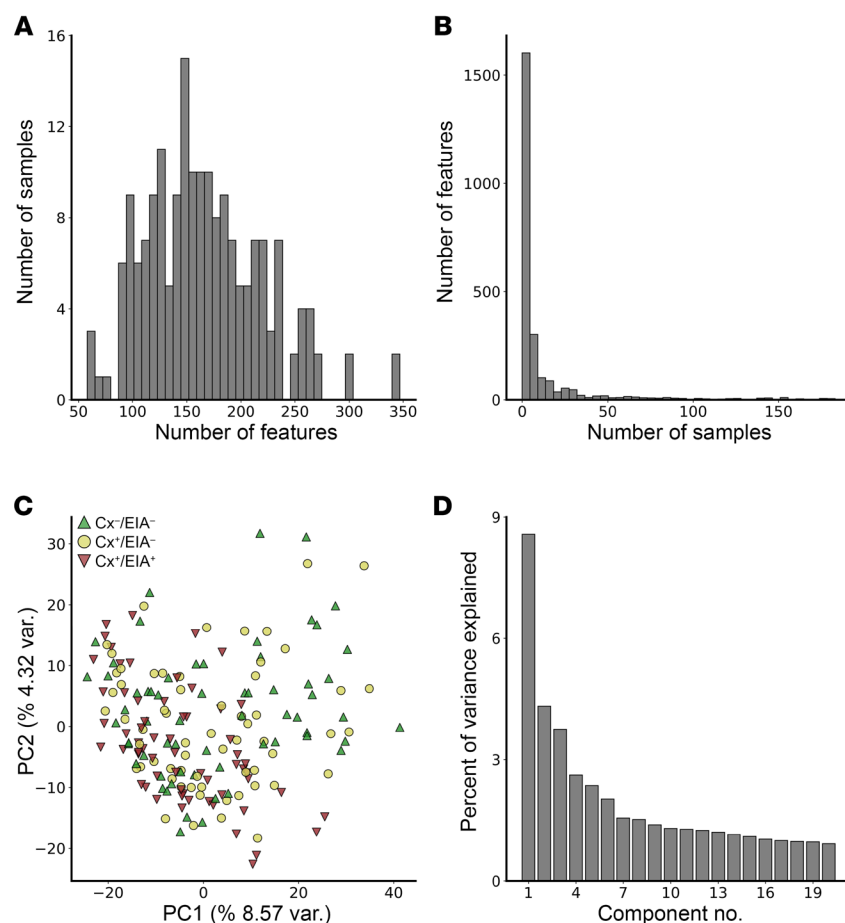


Figure 1. Metabolomic characteristics of the patient cohort. (A) Histogram showing the distribution of feature richness (number of features present per sample) across all patient specimens. (B) Histogram showing the number of samples within which each unique feature is present. Fecal metabolomes were highly individualistic: among the more than 2000 features detected, most were infrequent. While the resulting data are very sparse overall, the distribution has a relatively heavy tail with a few features present in many samples. (C) Principal component analysis (PCA) score plot across the first 2 components created using log-transformed feature intensities across all metabolomic features. (D) PCA does not appear to reveal dominant modes of variation, with no single component explaining more than 9% of the variance and a long tail of modes each explaining approximately 1% each.

ular features per sample was approximately normally distributed (Figure 1B; mean 164 features, standard deviation 54 features). Principal component analysis (PCA) of log-transformed feature intensities revealed no dominant modes of variation, with the first principal component explaining less than 10% of the overall variance in the data (Figure 1, C and D). Fecal metabolomes defined by GC-MS thus exhibit a high degree of individual variation, with only a small minority of metabolites common to all subjects.

Metabolomic differences between *C. difficile*-infected and uninfected controls. To identify CDI-associated fecal metabolites, we conducted a supervised multivariate comparison of Cx⁺/EIA⁺ and Cx⁻/EIA⁻ specimens. We used Cx⁺/EIA⁺ specimens to represent CDI because they harbor viable, toxigenic *C. difficile* alongside evidence of concurrent toxin production. Given the chemical complexity of fecal metabolomes (the >2000 resolved features greatly exceed the 124 samples), we employed multiple complementary measures to avoid overfitting the data, including repeated cross-validation (see Methods). Sparse partial least squares-discriminatory analysis (sPLS-DA) (Figure 2A) demonstrates good separation between metabolite profiles from the Cx⁺/EIA⁺ and Cx⁻/EIA⁻ groups, despite this model's use of an explicit penalty to prevent overfitting. To further assess this relationship, we conducted a separate logistic regression analysis on the Cx⁺/EIA⁺ and Cx⁻/EIA⁻ groups with a similar penalization parameter to avoid overfitting. Using repeated 5-fold cross-validation with random subsets to select an appropriate penalization level, we found that

relatively few molecular features yielded a large jump in average accuracy of the regression model (Figure 2B). We fixed the penalty parameter to the value yielding the maximum percent predicted, indicated by the star in Figure 2B, and again performed penalized logistic regression fit to the Cx⁺/EIA⁺ and Cx⁻/EIA⁻ groups with repeated randomized 5-fold cross-validation. The observed distributions of log-odds for the test folds (that is, excluding the training sets) for Cx⁺/EIA⁺ and Cx⁻/EIA⁻ again demonstrate good separation (Figure 2C). For comparison, Figure 2C also includes the distributions of the log-odds values for the Cx⁺/EIA⁻ cases. The 9 metabolite features most consistently associated with Cx⁺/EIA⁺ specimens (Table 2 and Supplemental Table 1) include both positive and negative associations. The features consist of 2 SCFAs, 1 amino acid, 1 bile acid, 1 lipid, 3 carbohydrates, and 1 aromatic alcohol. These results implicate biochemically diverse metabolites in human CDI pathogenesis. We then fit a logistic model using only the 6 features that were most frequently selected across the cross-validation runs. This model achieves a ROC AUC (area under the receiver-operator characteristic curve) of 96.7%, with a 95% confidence interval of 85.6%–100% obtained under repeated randomized 5-fold cross-validation (Figure 2D). These results are consistent with a strong, characteristic signal that distinguishes Cx⁺/EIA⁺ specimens from Cx⁻/EIA⁻ controls.

Stickland amino acid fermentation in CDI. Among the most highly CDI-associated metabolites (Table 2) is the SCFA 4-methylpentanoic acid (4-MPA/4-methylvaleric acid/iso-

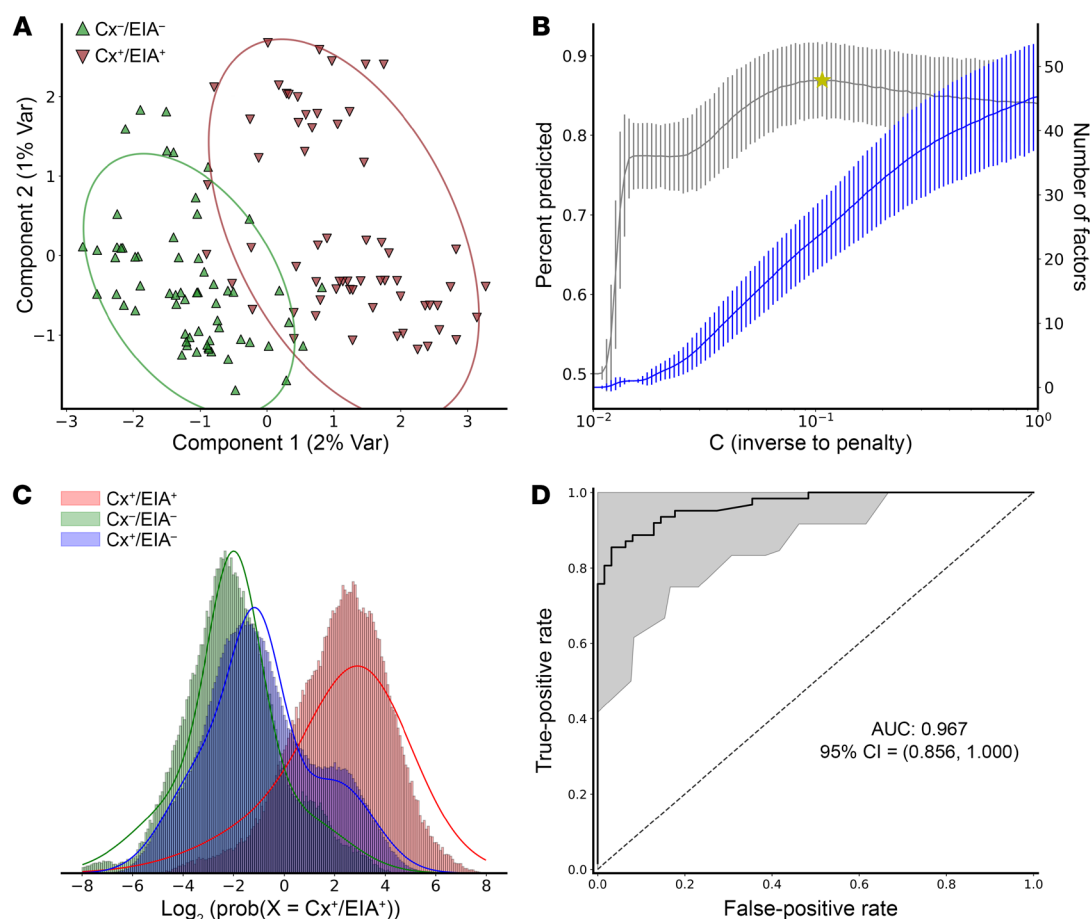


Figure 2. Supervised metabolomic analyses comparing Cx⁺/EIA⁺ with Cx⁻/EIA⁻ samples. (A) Observed separation of Cx⁺/EIA⁺ and Cx⁻/EIA⁻ samples under sparse partial least squares–discriminatory analysis (sPLS-DA). The data ellipses are drawn around each group of samples (at the 95% level). (B) Penalized logistic regression under repeated 5-fold cross-validation shows how the number of features used relates to the obtained accuracy, yielding high accuracy with a relatively small number of features. The maximum percent predicted is indicated by a star. (C) Using the penalty parameter associated with the maximum percent predicted, penalized logistic regression demonstrates good separation in the distribution of log-odds to be classified Cx⁺/EIA⁺ versus Cx⁻/EIA⁻. In the log-odds distribution shown here, only the test folds of Cx⁺/EIA⁺ and Cx⁻/EIA⁻ for each randomized cross-validated run are shown (that is, the corresponding distribution of the training set is not shown). For comparison, the corresponding log-odds of the Cx⁺/EIA⁻ samples are also shown. (D) Logistic regression (without penalty) to classify Cx⁺/EIA⁺ versus Cx⁻/EIA⁻ was performed using only the 6 features most frequently used in the penalized logistic regressions. Fitting to all samples gives 96.7% ROC AUC. The 95% CI of 85.6%–100% AUC was obtained under repeated randomized 5-fold cross-validation using the same 6 features.

caproic acid). Unlike the SCFAs formate, acetate, and butyrate, which are produced during microbial carbohydrate fermentation, 4-MPA is produced from leucine through the Stickland reactions, amino acid fermentation pathways associated with *C. difficile* and other anaerobic bacteria (15–24). Ten established Stickland products were detected in the study cohort, representing both oxidative and reductive fermentation of 8 different amino acid precursors (Figure 3A and Supplemental Figure 2). These products exhibit varying degrees of association with CDI, with 8 of 10 products (80%) detected more frequently in CDI specimens than controls (Figure 3B and Supplemental Figure 5). Many Stickland products were present in Cx⁺/EIA⁻ specimens, consistent with production by bacteria other than toxigenic *C. difficile*. Bootstrapped logistic regression (fit on 2000 bootstrap samples, stratified on Cx/EIA status) of Stickland metabolites consistently assigns the highest odds ratios for CDI to 4-MPA, the end product of leucine reduction (Figure 3C). Although

other canonical Stickland products like 5-aminopentanoic acid (5-aminovaleric acid) are frequently present in CDI, they offer negligible discriminatory power beyond that of 4-MPA in the adjusted analysis.

To more precisely quantify the relationship between 4-MPA production and CDI, we devised a targeted GC-MS assay to quantify Stickland fermentation activity through product/precursor ratios. In addition to increasing assay sensitivity and precision, this targeted biomarker ratio is intrinsically insensitive to the variations in fecal dilution that characterize diarrheal specimens. In an arbitrary subset of matched specimens, the 4-MPA/leucine ratio varied significantly between groups ($P = 1.3 \times 10^{-8}$, Kruskal-Wallis test). This variation distinguishes Cx⁺/EIA⁺ specimens from Cx⁻/EIA⁻ specimens with an ROC AUC of 92.8% (95% CI: 86.8%–98.7%; Figure 4, A and B) that rivals the 6-feature regression model described above and in the Methods (Figure 2D; AUC = 96.7%; 95% CI: 85.6%–100%).

Table 2. Top CDI-associated metabolites selected during cross-validation of logistic regression model

Feature (mass@RT)	Frequency ^A	Median odds ratio (95% CI)	Metabolite	Class
173.0@4.71	100%	1.54 (1.41–1.69)	4-methylpentanoic acid, TMS derivative ^B	SCFA
117.0@4.70	99%	1.16 (1.04–1.30)	4-methylpentanoic acid, TMS derivative ^B	SCFA
159.0@7.55	99%	1.12 (1.02–1.25)	2-hydroxy-4-methylpentanoic acid	SCFA
86.0@6.755	98%	0.91 (0.84–0.98)	Isoleucine	Amino acid
215.0@24.80	96%	0.87 (0.78–0.98)	Choleonic acid	Bile acid
73.0@13.79	95%	0.92 (0.83–0.99)	Ribitol	Carbohydrate
67.0@20.17	94%	1.11 (1.01–1.24)	Eicosatrienoic acid	Lipid
179.0@12.06	94%	1.15 (1.02–1.27)	Tyrosol, 2TMS derivative	Aromatic alcohol
204.0@19.45	91%	0.87 (0.77–0.98)	Glyceryl glycoside	Carbohydrate
217.0@14.72	89%	0.93 (0.84–0.99)	Fructose	Carbohydrate

^AProportion of cross-validation models including the metabolite. ^BTwo ions from same metabolite were independently resolved.

Together, these results are consistent with a pathophysiologic role for Stickland fermentation in CDI. While the presence of these metabolites in Cx[−]/EIA[−] specimens suggests that intestinal Stickland metabolism in patients is not generally unique to CDI, the selective increase in 4-MPA in CDI specimens raises the possibility that leucine reduction is a selectively emphasized pathway in *C. difficile* during clinical infections.

The isomeric amino acid allo-isoleucine is associated with CDI. Among the metabolites that are positively associated with CDI is *allo*-isoleucine, an isoleucine diastereomer in which the beta carbon stereocenter is inverted from an *S* to an *R* configuration (Figure 5A). This noncanonical, nonproteinogenic amino acid has been identified as a biomarker of branched chain ketoaciduria (maple syrup urine disease, an inborn error of metabolism) but has not previously been associated with *C. difficile* or CDI. Its origins in feces are unclear, although a previously reported bacterial metabolic pathway producing it from L-isoleucine raises the possibility that it derives from the intestinal microbiome (25). To more carefully assess the relationship between *allo*-isoleucine and CDI, we devised a targeted GC-MS assay to quantify *allo*-isoleucine as a ratio to isoleucine, its putative precursor. The *allo*-isoleucine-to-isoleucine ratio varied significantly between groups ($P = 6.5 \times 10^{-5}$, Kruskal-Wallis test; Figure 5B and Supplemental Figures 3 and 4). ROC analysis (Figure 5C) (AUC = 79.7%; 95% CI: 68.2%–91.3%) suggested favorable diagnostic potential for distinguishing Cx⁺/EIA⁺ specimens from Cx[−]/EIA[−] specimens. These observations identify *allo*-isoleucine as a new and biochemically distinctive CDI correlate of unclear origin.

Bile acid metabolic pathways active in patients without CDI. Three negatively loaded bile acid features are among the most frequently detected Cx⁺/EIA⁺ correlates in our cross-validated analysis (Table 2 and Supplemental Table 1). This corresponds to previous scholarship, which has associated bile acid dehydroxylation by the intestinal microbiota with CDI susceptibility (6, 7, 26, 27). Canonical bile acid processing by the microbiome involves successive dehydroxylation of cholic acid (CA; a tri-hydroxylated primary bile acid) to deoxycholic acid (DCA, a di-hydroxylated secondary bile acid) and chenodeoxycholic acid (CDCA; a di-hydroxylated primary bile acid) to lithocholic acid (LCA, a mono-hydroxylated secondary bile acid). Unexpectedly, the 2 most highly CDI-associated bile acids in our cohort were identified

as choleonic acid and monohydroxycholeonic acid (CE and MHCE, respectively, Supplemental Figures 6–11), noncanonical unsaturated, dehydroxylated bile acids. As with DCA and LCA, these bile acids were more abundant in the non-CDI group, consistent with an alternative bile acid dehydroxylation pathway based on dehydration reactions (net loss of H₂O to yield a double bond).

Unsaturated, nonhydroxylated bile acids are seldom considered in the bile acid literature. Their absence from our metabolite database compelled us to identify them through manual interpretation of spectra and comparison to chemically related reference compounds (Supplemental Figures 6–11). CE, a nonhydroxylated, unsaturated bile acid, was previously identified by Robben et al. as a lithocholic acid sulfate (LCA-S) desulfation product generated by an intestinal isolate of the *Bacteroidaceae* family (28). Robben et al. noted 2 isomeric CE products of these bacteria that differ in double bond location. We similarly observed 2 closely eluting CE products, consistent with a similar product distribution in our patient cohort (Supplemental Figure 9). Human tissues are known to generate sulfated bile acids, including LCA-S, which may provide substrates for fecal CE production through enzymatic desulfation (29). These observations are consistent with diminished microbial bile acid desulfation activity in patients with CDI.

Identification of a CDI-associated human bile acid network. Based on the presence of CE and MHCE in patient specimens, we hypothesized that sulfated bile acids (the precursors of unsaturated bile acids) (28) are also present. We further hypothesized that the desulfation mechanism of unsaturated bile acid production is generalizable such that an extended series of bile acid sulfates and unsaturated bile acids are present in the human fecal metabolome (Figure 6B). Using the calculated molecular weights, MS/MS fragmentation patterns, and chromatographic elution ranges for these hypothesized bile acids, we constructed a liquid chromatography–tandem mass spectrometry (LC-MS/MS) assay (Supplemental Figures 12–14 and Supplemental Table 4) because sulfated bile acids are undetectable by GC-MS. This assay resulted in tentative detection of 14 sulfated bile acids, 6 of which were dehydrogenated (possessing either an alkene or ketone; Table 3 and ref. 30). Many of these bile acids are distinguishable only by retention time, consistent with isomers that differ in the position(s) of double bonds, hydroxyl groups, and/or sulfate.

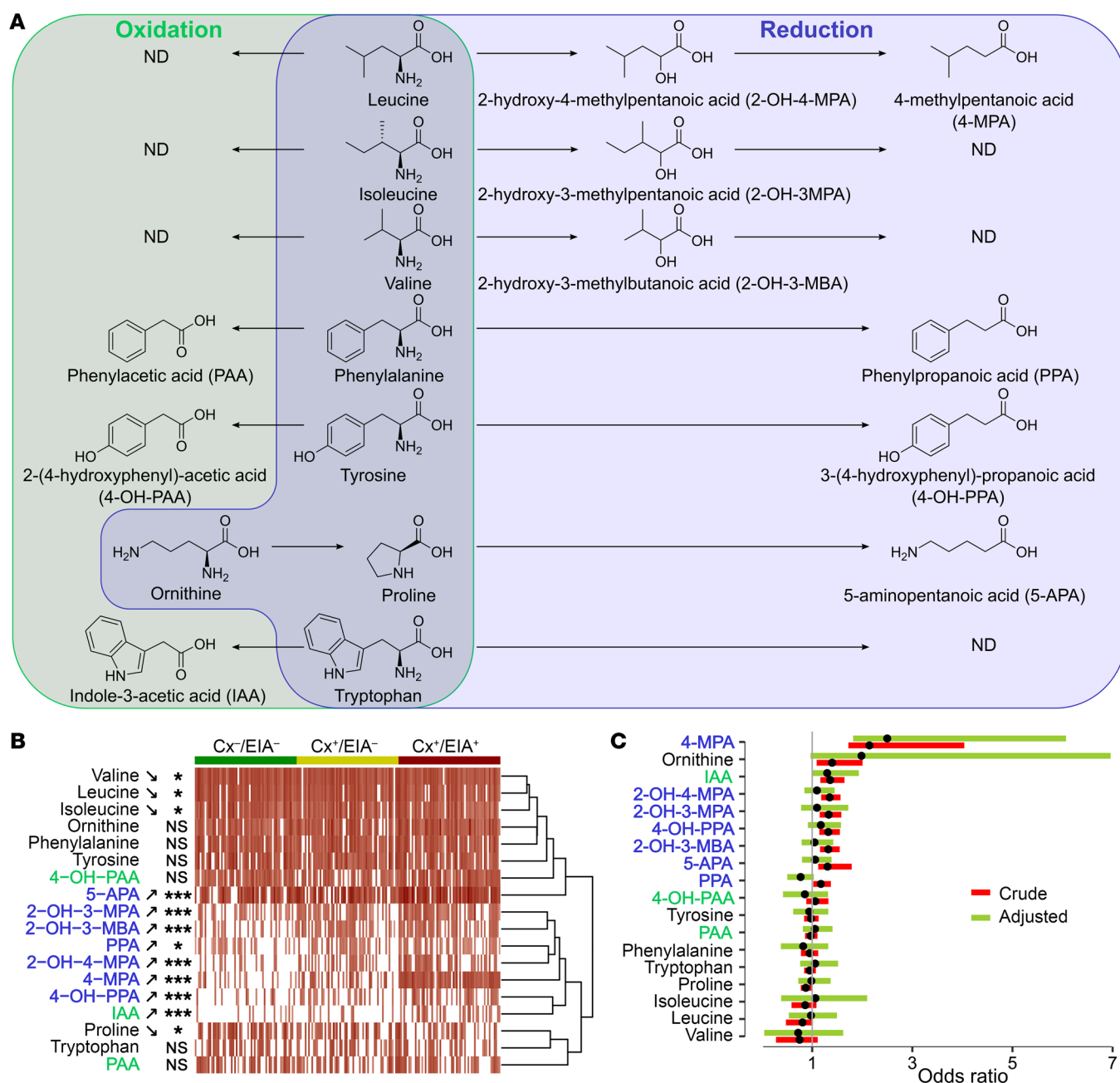


Figure 3. Amino acid metabolism in *C. difficile*. (A) Stickland metabolism consists of anaerobic amino acid fermentation through coupled oxidation and reduction pathways. In the reductive pathway, amino acids are first deaminated to form 2-hydroxy acids and then reduced to carboxylic acids. In the oxidative pathway, amino acids are deaminated and oxidized with loss of CO_2 to yield a distinct set of carboxylic acids. Depicted here are established Stickland substrates and products identified within patient fecal metabolomes. Stickland substrates include the nonproteinogenic amino acid ornithine. ND, not determined. (B) Heatmap of Stickland precursor and product abundances corresponding to patient fecal metabolomes from the 3 diagnostic groups. Metabolites were organized using unsupervised hierarchical clustering. Metabolites differing significantly (Mann-Whitney U test; $*P < 0.05$, $***P < 0.001$) between Cx-/EIA- and Cx+/EIA+ groups are labeled, along with the direction of the difference relative to the Cx-/EIA- control group. Stickland products are labeled according to the color scheme in A. (C) Adjusted and unadjusted (crude) CDI odds ratios and confidence intervals (95%) for Stickland precursors and products. Odds ratios were estimate by fitting logistic regression models to each of 2000 bootstrap samples stratified on Cx-/EIA status (Cx-/EIA- vs. Cx+/EIA+). Logistic models containing a single metabolite were fit to obtain crude odds ratios (red). A single logistic model including all metabolites was fit to obtain the adjusted odds ratios (green). Bars represent 95% bootstrap percentile confidence intervals and black dots represent median odds ratios across all bootstrap samples. Stickland products are labeled according to the color scheme in A.

Although fecal bile acids largely originate from 2 primary bile acids (CA and CDCA), subsequent host conjugation, divergent microbiome cometabolism, and enterohepatic circulation create a complex, nonlinear bile acid physiology. To characterize bile acid interrelationships, we therefore performed community detection

(31) on the weighted network of positive correlations among the 14 noncanonical bile acids described above and 17 canonical conjugated and nonconjugated primary and secondary bile acids. Seven bile acid communities emerged from this unbiased network community detection analysis, many of which could be rationalized by

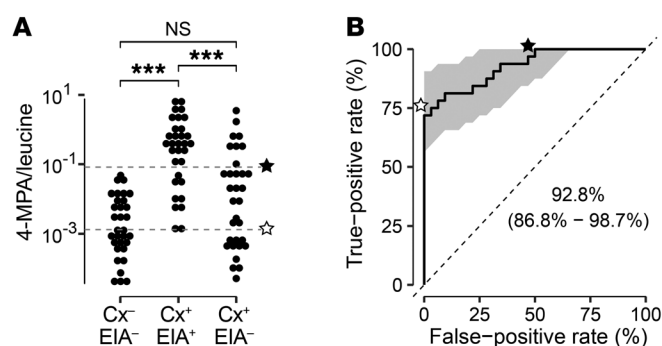


Figure 4. 4-MPA/leucine ratio elevated in CDI. (A) Dot plots of 4-MPA/leucine product/precursor ratios measured by targeted (SIM) reanalysis of fecal specimens ($n = 32$ for each group). Patient groups were compared using the Kruskal-Wallis test ($P = 1.3 \times 10^{-8}$). To further characterize pair-wise differences between groups, Bonferroni-corrected Mann-Whitney U test P values are indicated (3 comparisons; NS: $P \geq 0.05$, *** $P < 0.001$). Ratio thresholds giving perfect specificity (0.0825, black star) or sensitivity (0.00132, white star) for CDI⁺/EIA⁺ are marked as gray dashed lines. (B) Receiver-operator characteristic (ROC) plot distinguishing Cx⁺/EIA⁺ patients from Cx⁻/EIA⁻ patients. The gray region represents the bootstrapped 95% confidence interval for the true-positive rate at each false-positive rate. Thresholds with perfect specificity or sensitivity are marked by stars, as in A.

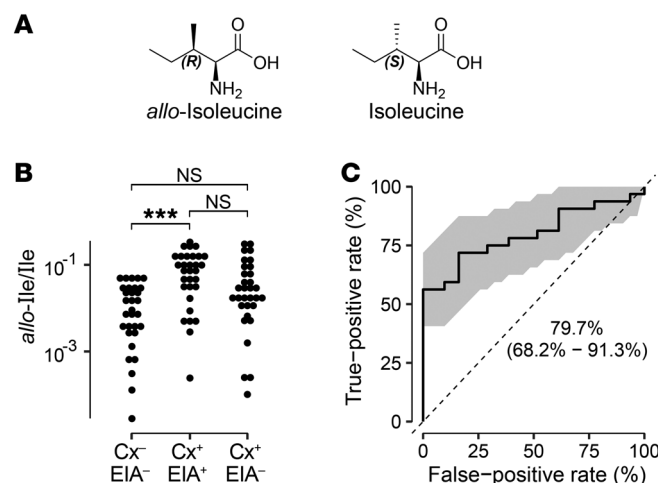
shared chemical features (Table 3 and Figure 6A). Where unavailability of authentic internal standards prevents identification of hydroxylation sites (e.g., the 3, 7, and 12 carbon positions) or epimers, bile acids are designated with general names. Communities 1 to 3 are composed exclusively of canonical primary and secondary bile acids. Community 1 consists of classic primary bile acids while community 2 consists of their glycine or taurine conjugates. Community 3 consists of conjugated secondary (dehydroxylated) bile acids. Community 4 includes secondary bile acids, secondary bile acid sulfates, and 1 candidate di-hydroxylated cholenic acid sulfate. Communities 5 and 6 consist entirely of sulfated bile acids, with a single sulfated cholenic acid candidate. The 5 bile acids in community 7 are all sulfated, with 4 cholenic acid sulfate candidates. The 5 candidate dehydroxylated cholenic acid sulfates may plausibly include sulfated keto bile acids, secondary bile acids of identical mass. In a force-directed layout depicting this network (Figure 6A), the primary bile acids (CA, CDCA) are located centrally, consistent with their recognized roles as precursors to conjugated and secondary bile acids. Clockwise progression moves from bile acid communities defined by host glycine and taurine conjugation, to classical microbial dehydroxylation, to sulfation, to desaturation or ketone formation (Figure 6B). The community organization emerging from this analysis reflects the distinctive metabolic transformations identified in the present study and in previous work.

Bile acid metabolomic associations with CDI. Disruption of microbiome-mediated bile acid metabolism has long been regarded to increase CDI risk. In our inpatient cohort, we hypothesized that the Cx⁻/EIA⁻ group includes a subset of patients with disrupted, CDI-susceptible microbiomes. To test this hypothesis, we used PCA to graphically summarize bile acid metabolomic variation in culture-negative specimens (Figure 7, A and B). Next, we projected Cx⁺/EIA⁺ bile acid profiles onto these principal components. Consistent with the hypothesis, Cx⁺/EIA⁺ specimens preferential-

ly occupied a restricted portion of the Cx⁻/EIA⁻ patient bile acid profile distribution. Specifically, Cx⁺/EIA⁺ specimens preferentially exhibit elevated values along the first PCA-derived principal component (PC1). High PC1 scores correspond to higher primary (cholic and chenodeoxycholic) and low secondary (deoxycholic and lithocholic) bile acids (Figure 7D), similar to previous studies (26, 27). Low PC1 scores correspond to higher levels of sulfated and dehydroxylated cholenic and cholanic acids (DHCA-S3, DHCE-S3, LCA from community 4). ROC analysis using PC1 as the discriminator revealed an AUC of 61.3% (Figure 7C). These results are consistent with a negative association between CDI and bile acid sulfation, dehydroxylation, and unsaturation. While we cannot conclude a causative role from these correlative data, these metabolic processes may indicate the presence of a CDI-resistant intestinal microbiome.

Fecal carbohydrate associations with CDI. We next hypothesized that the Cx⁻/EIA⁻ group includes patients with CDI-susceptible intestinal metabolites other than bile acids. To test this hypothesis, we used PCA to graphically summarize total GC-MS detectable metabolomic variation in culture-negative specimens. Next, we projected CDI patient metabolomes onto these principal components. Consistent with the hypothesis, CDI patient fecal metabolomes occupy a restricted portion of the uncolonized patient distribution, characterized by a high PC1 score (Figure 8, A and B). ROC analyses of PC1 scores yielded a modest AUC of 61.1% when distinguishing Cx⁺/EIA⁺ from Cx⁻/EIA⁻ specimens (Figure 8C). These metabolites are not clearly related to bile acid compo-

Figure 5. Isoleucine isomer correlated with *C. difficile*. (A) Chemical structures of isoleucine and its diastereomer, *allo*-isoleucine. (B) Dot plot of *allo*-isoleucine/isoleucine ratios as measured by SIM ($n = 32$ for each group). Patient groups were compared using the Kruskal-Wallis test ($P = 6.5 \times 10^{-5}$). To further characterize pair-wise differences between groups, Bonferroni-corrected Mann-Whitney U test P values are indicated (3 comparisons; NS: $P \geq 0.05$, *** $P < 0.001$). (C) ROC plot showing ability to distinguish Cx⁺/EIA⁺ patients from Cx⁻/EIA⁻ patients. The gray region represents the bootstrapped 95% confidence interval for the true-positive rate at each false-positive rate.



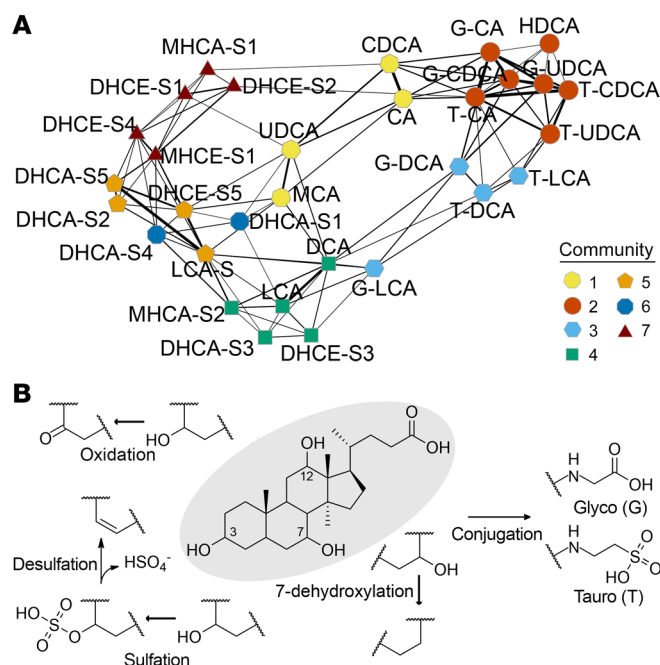


Figure 6. Bile acid transformations in the clinical cohort. (A) A force-directed network layout illustrates associations between bile acids in the study cohort. Each node represents a bile acid and each connecting line (edge) represents an association between 2 bile acids as 1 of the 5 highest correlations for at least 1 of the corresponding nodes. Edge lengths are determined by the level of correlation between connected bile acids. Nodes are colored by community assignment. (B) Scheme showing metabolic transformations producing bile acids in the network analysis. The central structure highlighted in gray represents a tri-hydroxylated primary bile acid (e.g., cholic acid). Taurine or glycine conjugation forms peptide bonds to the carboxylic acid group (right). Alcohol groups are removed from the bile acid nucleus (dehydroxylation, bottom right) or oxidized to a ketone (top left). Bile acid sulfation involves substitution of an alcohol group with a sulfate ($R = SO_4^-$) group (bottom left). Desulfation of bile acid sulfates yields unsaturated bile acids (left).

sition, since the total metabolome PC1 exhibits a low degree of association with the bile acid PC1 determined above ($r^2 < 0.007$; Supplemental Figure 17). Instead, high PC1 scores are primarily characterized by diminished monosaccharides, disaccharides, and sugar alcohols with uncertain relationships to CDI (Figure 8D and Supplemental Figure 16). While these metabolite classes can be reasonably identified by GC-MS, identifying specific isomers is often unreliable (e.g., sorbitol and mannitol are both $C_6H_{14}O_6$ and differ only in the orientation of 1 hydroxyl group and yield comparable spectra). The monosaccharide fructose, a favored *C. difficile* carbon substrate (32), emerged as a negative CDI correlate in the logistic regression analysis above (Table 2), raising the possibility that some carbohydrates may be consumed by metabolically active *C. difficile*. Trehalose, a disaccharide recently reported to be a favored substrate of epidemic *C. difficile* ribotypes 027 and 078, was not identified in our differential analysis (33). To more carefully assess the relationship between trehalose and CDI, we quantified fecal trehalose using a targeted GC-MS analysis based on stable isotope dilution with a $^{13}C_6$ -labeled internal standard (Supplemental Figure 15). It was detectable in 61% (115/189) of specimens but did not distinguish Cx^+/EIA^+ from Cx^-/EIA^- specimens (35/63 vs. 41/63, $P = 0.36$, 2-tailed Fisher's exact test). In 027-positive specimens, trehalose also did not distinguish toxin-positive from toxin-negative specimens (6/8 vs. 12/23, $P = 0.41$, 2-tailed Fisher's exact test). A subset of fecal carbohydrates thus has some potential to distinguish CDI and possibly CDI-susceptible patients, though the basis for this remains unclear.

A metabolomic model of CDI. To determine whether fecal Stickland metabolites and bile acids can be used to construct a metabolomic definition of CDI, we conducted logistic regression using the 4-MPA/leucine ratio ($\log_{10}$ -transformed) and the bile acid PC1 (Table 4 and Figure 9A). Each parameter alone exhibited significant ($P < 0.05$) independent associations with Cx^+/EIA^+ status when compared with Cx^-/EIA^- specimens. When the logistic model criterion is

applied (corresponding to $>50\%$ probability), Cx^+/EIA^+ specimens clustered in the high 4-MPA/leucine and high bile acid PC1 quadrant (Figure 9, A and B). ROC analysis of this model yields an AUC of 98.2%, out-performing the original 6-feature model described above (Figure 9C). Each parameter contributed independently—adding a term for interaction between 4-MPA/leucine ratio and bile acid PC1 did not significantly improve the logistic model ($P = 0.53$, analysis of deviance). These results are consistent with distinctive host and microbial metabolic processes in human CDI.

Metabolomic differences in colonized patients with and without detectable fecal toxin. To determine whether Cx^+/EIA^+ specimens possess distinctive metabolomes, we compared 4-MPA/leucine and bile acid composition profiles from Cx^+/EIA^+ specimens to those of Cx^+/EIA^- or Cx^-/EIA^- specimens. In the logistic regression model, only 38% (20/32) resembled Cx^+/EIA^+ specimens, with the remainder exhibiting low 4-MPA/leucine ratios in specimens with or without susceptible bile acid profiles (Figure 9, A and B). These observations are consistent with low *C. difficile* metabolic activity and a protective bile acid profile in many patients with undetectable fecal toxin. Using the logistic regression parameter compared with toxigenic culture or toxin EIA results alone defines a positive test group that is smaller than (but almost entirely encompassed by) toxigenic culture-positive specimens but greater than the number of toxin EIA-positive specimens (Figure 9D). If the metabolic criterion is highly accurate, it may restrict false-positive results from toxigenic *C. difficile* detection alone and also restrict false-negative results from the toxin EIA test. Further study is necessary to determine whether this possibility can be realized.

Discussion

In this study, we compared the fecal metabolomic profiles from 186 hospitalized patients to investigate relationships between fecal metabolites, the presence of toxigenic *C. difficile*, and the presence of detectable *C. difficile* toxins. Untargeted metabolomic profiling in the context of uncontrolled patient dietary and microbiome contributions yielded extremely diverse fecal metabolomes. Nevertheless, numerous CDI-associated metabolites were resolved. Among the 2463 features detected in this cohort, 43 had some ability to resolve CDI from uncolonized controls. Many of these discriminatory molecules are associated with Stickland and bile acid metabolism, processes previously implicated in CDI pathogenesis (6–9, 18, 19, 23, 26, 27, 34, 35). The specific molecular signatures best able to resolve CDI from controls exhibit only

Table 3. Fecal bile acids monitored by LC-MS/MS

Community ^a	Abbreviation	Name ^b	Primary ^c	Hydroxylations, no.	Conjugation, type	Sulfate	Unsaturated
1	CA	Cholic acid	Yes	3			
	MCA	α -Muricholic acid	Yes	3			
	CDCA	Chenodeoxycholic acid	Yes	2			
	UDCA	Ursodeoxycholic acid	Yes	2			
2	G-CA	Glycocholic acid	Yes	3	Glyco		
	T-CA	Taurocholic acid	Yes	3	Tauro		
	HDCA	Hyodeoxycholic acid	Yes	2			
	G-CDCA	Glycochenodeoxycholic acid	Yes	2	Glyco		
	G-UDCA	Glycoursodeoxycholic acid	Yes	2	Glyco		
	T-CDCA	Taurochenodeoxycholic acid	Yes	2	Tauro		
	T-UDCA	Tauroursodeoxycholic acid	Yes	2	Tauro		
3	G-DCA	Glycodeoxycholic acid		2	Glyco		
	T-DCA	Taurodeoxycholic acid		2	Tauro		
	T-LCA	Tauroolithocholic acid		1	Tauro		
	G-LCA	Glycolithocholic acid		1	Glyco		
4	DCA	Deoxycholic acid		2			
	LCA	Lithocholic acid		1			
	DHCA-S3	<i>Dihydroxycholanolic acid sulfate-3</i>		2		Yes	
	DHCE-S3	<i>Dihydroxycholenic acid sulfate-3^o</i>		2		Yes	Yes
	MHCA-S2	<i>Monohydroxycholanolic acid sulfate-2</i>		1		Yes	
5	LCA-S	Lithocholic acid sulfate		1		Yes	
	DHCE-S5	<i>Dihydroxycholenic acid sulfate-5^o</i>		2		Yes	Yes
	DHCA-S2	<i>Dihydroxycholanolic acid sulfate-2</i>		2		Yes	
	DHCA-S5	<i>Dihydroxycholanolic acid sulfate-5</i>		2		Yes	
6	DHCA-S1	<i>Dihydroxycholanolic acid sulfate-1</i>		2		Yes	
	DHCA-S4	<i>Dihydroxycholanolic acid sulfate-4</i>		2		Yes	
7	DHCE-S1	<i>Dihydroxycholenic acid sulfate-1^o</i>		2		Yes	Yes
	DHCE-S2	<i>Dihydroxycholenic acid sulfate-2^o</i>		2		Yes	Yes
	DHCE-S4	<i>Dihydroxycholenic acid sulfate-4^o</i>		2		Yes	Yes
	MHCA-S1	<i>Monohydroxycholanolic acid sulfate-1</i>		1		Yes	
	MHCE-S1	<i>Monohydroxycholenic acid sulfate-1</i>		1		Yes	Yes

^aBile acids monitored by LC-MS are listed in order of community assignment (Figure 6). ^bWhere a reference standard is available to confirm chemical identity, the common name is used. Where lack of a reference standard prevents isomeric assignment, a candidate chemical structure is provided in italics.

^cBile acids are marked "primary" if they have a known host biosynthetic origin. ^oAlternatively, structure may include a keto group.

partial overlap with those identified in prior metabolomic studies using mouse models, which may reflect species differences, the presence of a variable host microbiome background, and the specific mass spectrometric approach. Toxin-negative, toxigenic *C. difficile*-positive (Cx⁺/EIA⁻) specimen metabolomes span a metabolomic continuum ranging from control-like to CDI-like. Among Cx⁺ specimens, fecal metabolites have the potential to distinguish infected from colonized patients.

Identification of 4-MPA as the most prominent CDI correlate is consistent with its production by *C. difficile* from leucine during Stickland metabolism. Other Stickland products were also detected and observed to be elevated in patients with CDI, although their abundance among the control specimens (Cx⁺/EIA⁻) diminished some of their associations (low positive predictive value), especially that of 5-aminopentanoic acid. This contrasts with previously reported murine studies in which multiple Stickland metabolites are highly CDI-associated. The discrepancy between patient and mouse studies likely arises from Stickland-metabolizing organisms

in Cx⁺/EIA⁻ patient microbiomes, which may be limited or absent in the antibiotic-treated mice used in experimental CDI models. 4-MPA has not been uniformly identified as a CDI correlate in metabolomic studies of murine CDI. This may reflect host-associated substrate selection of leucine for Stickland metabolism by toxin-producing *C. difficile* but may also reflect lack of detection due to the apparent insensitivity of typical untargeted LC-MS approaches to 4-MPA (unpublished observations). Indeed, GC-MS remains a favored modality for SCFA analyses by many investigators. Nevertheless, the implication of Stickland fermentation in CDI is generally consistent with previous human and animal model studies.

The association between Stickland fermentation and CDI is consistent with the hypothesis that fecal amino acid availability enhances CDI susceptibility. Our data do not rule out an important role for carbohydrate metabolism, the *C. difficile* fermentation products of which (pyruvate, formate, acetate, butyrate) are less distinctive than Stickland metabolites (18). Although we observed no association between CDI and fecal trehalose, a glucose disac-

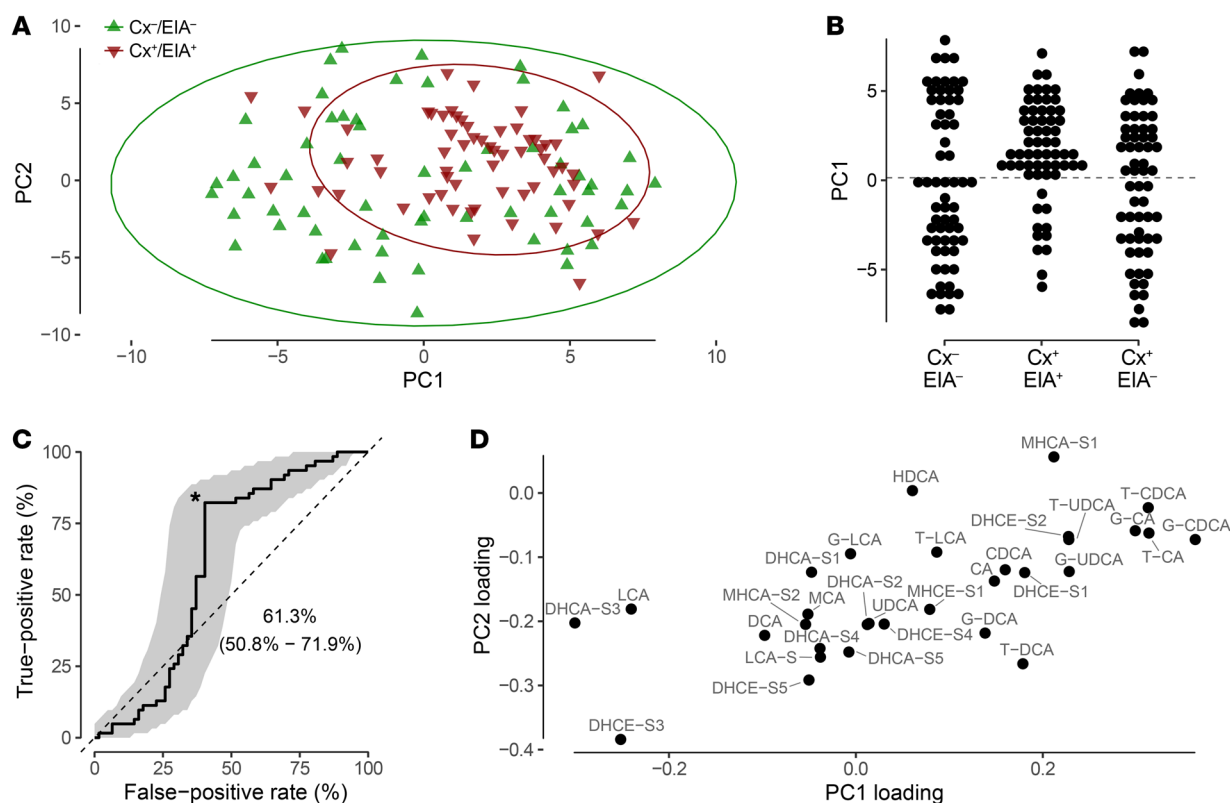


Figure 7. The bile acid distribution in patients with CDI resembles that of a characteristic subgroup of uninfected, hospitalized patients. (A) Depicted here is a PCA plot of uninfected patients' bile acid profiles (green, $n = 62$). Onto this space, we projected the bile acid metabolome of patients with CDI (red, $n = 62$). Data ellipses are drawn around each group of samples (95% level). Clustering of CDI specimens at high PC1 values is consistent with a favored bile acid distribution among patients with CDI. (B) Dot plot of PC1 scores for each patient sample ($n = 62$ in each group). Gray dashed line represents optimal PC1 threshold for distinguishing Cx⁻/EIA⁻ from Cx⁺/EIA⁺ samples. This threshold was chosen by maximizing the sum of percent sensitivity and specificity. (C) ROC plot evaluating the ability of PC1 to distinguish CDI patients from controls. The gray region represents the bootstrapped 95% confidence interval for the true-positive rate at each false-positive rate. An asterisk marks the point corresponding to the optimal PC1 threshold depicted in B. (D) PCA loading plot depicting the relative contributions of each bile acid to the distribution of Cx⁺/EIA⁺ samples in A. Abbreviations are indicated in Table 3.

charide generated during bacterial stress responses, utilized as a food additive, and proposed as a dietary risk factor for CDI caused by hypervirulent strains (ribotype O27; Table 1), the other fecal carbohydrates detected in this study may plausibly serve as metabolic substrates (33). A recent study by Battaglioli et al. observed a broad spectrum increase in fecal amino acid concentrations in gnotobiotic mice colonized with dysbiotic human gut microbiota. This increase corresponded to high fecal *C. difficile* colonization after experimental challenge (35). In the present study, amino acids tend to be diminished in CDI specimens compared with controls (Figure 3, B and C, and Supplemental Figure 18). This apparent contradiction may be reconciled by interpreting the decrease in amino acids during CDI as evidence of consumption by metabolically active *C. difficile*, which yields the aforementioned Stickland products. The importance of amino acid substrate selection by *C. difficile* during clinical CDI remains unclear. The present data are consistent with a preference for branched chain amino acids (leucine, isoleucine, and valine) relative to other intestinal microbes, though it is possible that other Stickland substrates, such as proline, tyrosine, phenylalanine, and ornithine, could substitute for branched chain amino acid deficiencies. If so, gut microbiota that deplete a broad range of fecal amino acids may help hosts resist CDI.

In addition to implicating *C. difficile* metabolic pathways in CDI patients, the present study also identifies a series of CDI-associated bile acids. Previous mouse model studies have identified associations between diminished fecal secondary bile acids and increased *C. difficile* fecal colonization, which agrees with the general findings of the current study (6–9). Differences in specific bile acids between this study and murine studies likely reflect both species differences (murine bile acids exhibit substantial 6-hydroxylation compared with humans) and different analytical approaches. Human cells synthesize and chemically conjugate bile acids, whereas intestinal microbes have been shown to modify them through dehydroxylation at the 7-carbon position to yield deoxycholic and lithocholic acids (from cholic and chenodeoxycholic acids, respectively). Here, unbiased detection of 2 cholenic acids (cholenic and hydroxycholenic acids, Supplemental Figures 6–11 and Supplemental Table 3) by GC-MS profiling as the most highly CDI-associated bile acids raises the possibility that beneficial microbes can also dehydroxylate bile acids at the 3-carbon position, leaving behind unsaturated, nonhydroxylated bile acids. Detection of monohydroxycholenic acid sulfate (MHCE-S1, Table 3) provides additional evidence of this pathway. Five additional bile acid sulfate candidates may represent either cholenic acids

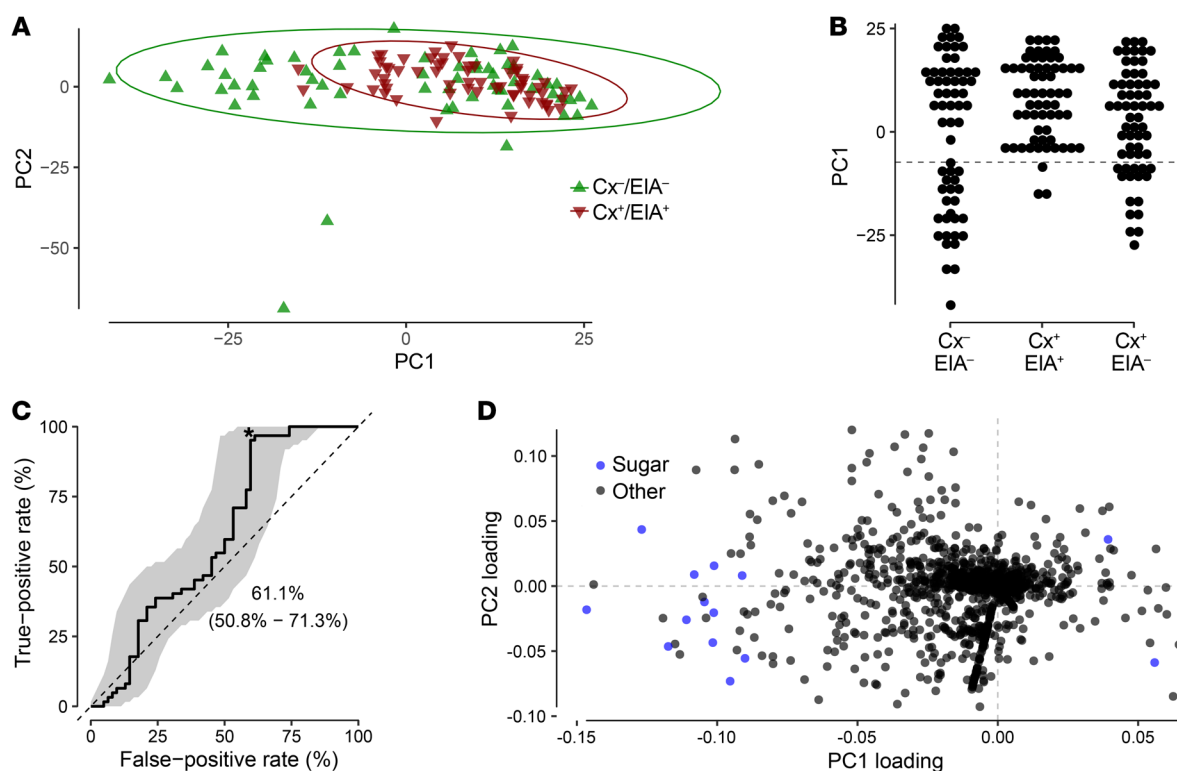


Figure 8. Principal component analysis of GC-MS-defined metabolome in the clinical cohort. (A) Depicted here is a PCA plot of uninfected patients' GC-MS metabolomes (green, $n = 62$), onto which is projected the GC-MS metabolomes of patients with CDI (red, $n = 62$). Data ellipses are drawn around each group of samples (95% level). The clustering of CDI specimens at high PC1 values is consistent with a favored metabolomic profile among patients with CDI. (B) Dot plot of PC1 scores for each patient ($n = 62$ in each group). Gray dashed line depicts the PC1 threshold that maximizes the sum of percent sensitivity and specificity for distinguishing Cx⁻/EIA⁻ from Cx⁺/EIA⁺ samples. (C) ROC plot evaluating the ability of PC1 to distinguish between CDI patients and controls. The gray region represents 95% confidence intervals bootstrapped for the true-positive rate at each possible false-positive rate. An asterisk marks the point corresponding to the optimal PC1 threshold depicted in panel B. (D) Plot of PC1 and PC2 loadings for all 2539 GC-MS features. It depicts the relative contributions of each GC-MS feature to the distribution of Cx⁻/EIA⁻ samples in the PCA projection in A. Features in the top or bottom 1% of PC1 loadings tentatively identified as sugars or sugar alcohols are highlighted in blue.

or *keto*-bile acids, both of which would exhibit ions 2 mass units below their canonical counterparts. It remains unclear whether cholenic acids are solely CDI-negative patient biomarkers or whether their formation protects patients from CDI (34). Production of these bile acids might confer CDI protection through consumption of progermination bile acids or by direct inhibition of *C. difficile* spore germination. Additional experimental work is necessary to evaluate these possibilities and could help identify desirable microbiome constituents for future therapeutic strategies.

The biochemical signatures resolved in this study suggest a metabolomic model of human CDI. In addition to identifying therapeutic strategies, such a model may also identify new or refined diagnostic approaches to appropriately identify patients who would benefit from treatment. Current diagnostic approaches are based on nucleic acid-based detection of toxigenic *C. difficile* and immunoassay-based detection of fecal toxin, each of which raise valid concerns over their associated false-positive and false-negative rates (3, 10, 36). The metabolomic profiles identified in the current work are biochemically distinct from existing tests and, in a multistep diagnostic approach with existing tests, could improve diagnostic accuracy. Detection of Stickland metabolites would be consistent with the presence of antibiotic-responsive, vegetatively

growing *C. difficile*. Moreover, individualized metabolomic information on whether an unfavorable bile acid profile is present could guide microbiome-directed interventions such as fecal transplant or probiotic administration. The signatures identified here may aid larger patient studies aimed at assessing the value of this approach.

In summary, this metabolomic study suggests specific host, pathogen, and microbiome factors associated with CDI pathogenesis. Strengths of this study include use of a valid clinical study population with relevant control specimens and comparison to clinically accessible test results, use of an unbiased screening approach, use of multiple mass spectrometric methods, and use of strategies to

Table 4. Logistic regression model of CDI metabolome

Parameter	Crude odds ratio (95% CI)	Adjusted	P value ^a
Log ₁₀ (4-MPA index)	2.09 (2.01–6.06)	4.06 (2.15–7.53)	<0.0001
Bile acid PC1	0.15 (0.01–0.30)	0.71 (0.31–1.42)	0.033

^aAnalysis of deviance comparison between single-parameter logistic regression model and the null logistic regression model.

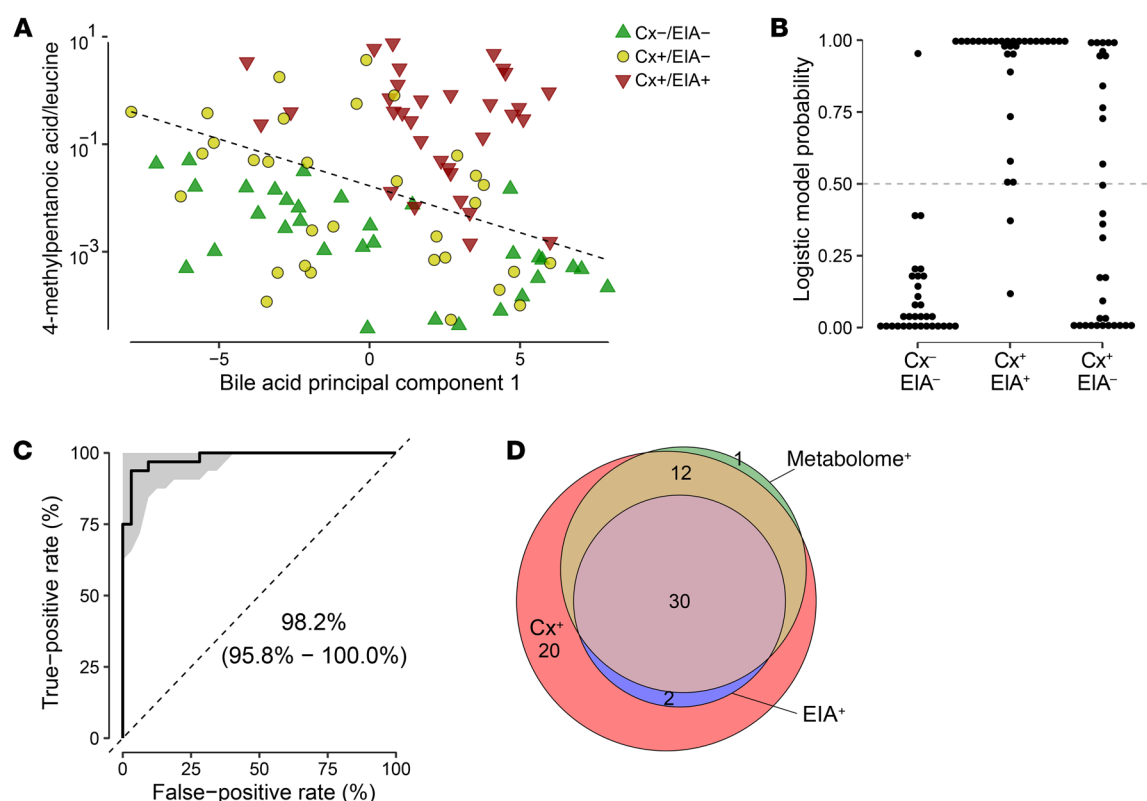


Figure 9. Interrelationships between host- and *C. difficile*-associated metabolites. (A) Plotting bile acid PC1 (Figure 7) versus 4-methylpentanoic acid index (Figure 4) reveals that high PC1 score and high 4-methylpentanoic acid index values coincide in patients with CDI compared with control patients ($n = 32$ for each group). The dashed line marks the dividing line assigned 50% probability of being Cx+/EIA+ by a logistic regression model incorporating both PC1 and 4-methylpentanoic acid index. (B) Probabilities assigned to each patient by the logistic regression model ($n = 32$ per group). Higher values indicate higher certainty of Cx+/EIA+ status. The gray line marks the 50% probability cutoff above which samples are considered Cx+/EIA+. (C) ROC curve showing the performance of the logistic regression model in discriminating Cx-/EIA- patients from Cx+/EIA+ patients. The gray region represents 95% confidence intervals bootstrapped for the true-positive rate at each possible false-positive rate. The AUC and its 95% confidence interval are also reported. (D) Euler diagram showing the overlap between culture, EIA, and metabolome status. Samples were considered metabolome-positive if assigned a probability above 50% by the logistic regression model.

avoid the overfitting issues inherent in many comparative metabolomic approaches. The uniquely high chromatographic resolution and informative electron ionization spectra of GC-MS analysis was likely essential to our detection of 4-MPA, *allo*-isoleucine, and cholenoic acid, analytes that are poorly detected or resolved under typical LC-MS conditions. Moreover, the ability to identify metabolites using spectrally rich EI fragmentation spectra in GC-MS allowed us to place our analytic findings within a broader biological context. GC-MS is, however, restricted to small thermostable analytes, a notable limitation of this modality when compared with LC-MS. Other limitations of this study include its observational nature, lack of longitudinal data, lack of nondiarrheal control specimens, and insensitivity to host and pathogen-derived macromolecules (proteins, complex lipids, etc.). This work identifies a potential diagnostic approach to CDI as well as new hypotheses for future evaluation regarding host bile acid networks' interaction with CDI.

Methods

Patient specimen collection. This cohort was derived from samples submitted for physician-ordered *C. difficile* toxin testing as part of routine clinical care. Remnant specimens that would have been otherwise discarded were frozen at -80°C by the laboratory for future use. Approval was obtained from the Washington University Institutional Review

Board with a waiver of informed consent to use specimens for this study. Patient and specimen evaluation for this cohort has recently been described (14). From August 2014 through September 2016, the Barnes-Jewish Hospital (BJH) Microbiology Laboratory detected the presence of toxin A and B in these specimens using the Alere TOX A/B II toxin enzyme immunoassay (EIA).

Inclusion and exclusion criteria. To identify and exclude patients with a potential alternate cause of diarrhea, BJH medical informatics databases were queried to identify patients with these conditions and medications. Patient charts lacking an identifiable alternate cause of diarrhea were reviewed to determine whether the patient had clinically significant diarrhea, and to confirm that there were no other known causes of diarrhea. If it was not possible to determine whether the patient had clinically significant diarrhea based on the medical records, the specimen was excluded. Specimens that were toxin negative (EIA-) were also excluded if the patient received treatment for CDI within 14 days of stool specimen collection. Due to these rigorous criteria, patients that were toxin positive (EIA+) were considered to have CDI and patients who were EIA- but positive for a toxigenic strain of *C. difficile* (defined by the presence of *tcdA* and/or *tcdB* by PCR; Cx+) were considered colonized with toxigenic *C. difficile* but with diarrhea due to other reasons. EIA- stools were also excluded if the patient was receiving antibiotics that could treat CDI to better ensure that patients with Cx+/EIA- stool did not have CDI.

C. difficile culture and characterization. Briefly, 1 g stool was heat shocked at 80°C for 10 minutes. The specimen was then placed into cycloserine, cefoxitin, mannitol broth with taurocholate and lysozyme (Anaerobe Systems) and incubated anaerobically at 35°C. When turbid, broth was streaked onto prereduced blood agar (BAP, Becton, Dickinson and Company). *C. difficile* was identified by matrix-assisted laser desorption/ionization time of flight (MALDI-TOF MS). Isolates were evaluated for the presence of *tcdA*, *tcdB*, and binary toxin genes (*cdtA/cdtB*) by multiplex PCR. PCR ribotyping was then performed. The ribotyping banding patterns were analyzed using DiversiLab Bacterial Barcodes software. Similarity of at least 95% was required for isolates to be considered identical. All unique strains were compared with the Cardiff-ECDC collection of *C. difficile* strains for name assignment. Isolates that did not match to a strain in the Cardiff-ECDC collection were compared with unique strains in the Washington University collection for name assignments. Isolates that did not match strains in the Cardiff-ECDC collection or Washington University collection were assigned a unique name.

Fecal extracts. Stool specimens were thawed on ice and approximately 0.1 mg of each was transferred to a microfuge tube and weighed. MeOH (1.25 mL, 70%) was added to each stool sample. The samples were sealed with parafilm, vortexed for 10 seconds, and rotated in a cold room for 2 hours. The samples were vortexed, decanted into a microcentrifuge tube and centrifuged at 20817 × g in a desktop centrifuge for 15 minutes at 4°C. The supernatant was decanted into a tube and stored at -80°C until analysis.

Gas chromatography-mass spectrometry (GC-MS). Stool extract (30 µL) was pipetted into a glass vial, dried under N₂, and derivatized with 100 µL MSTFA (*N*-Methyl-*N*-trimethylsilyltrifluoroacetamide)/CH₃CN/pyridine(1:2.6:0.4), heated at 70°C for 30 minutes, then cooled at room temperature overnight. Derivatized samples were analyzed using an Agilent 7890A gas chromatograph interfaced to an Agilent 5975C mass spectrometer and equipped with an HP-5MS column (30 m, 0.25 mm i.d., 0.25 µm film coating). For GC, an initial temperature of 80°C for 2 minutes was followed by a linear gradient to 300°C at 10°C/minute followed by a 5-minute elution at 300°C. EI was conducted with source temperature, electron energy, and emission current of 250°C, 70 eV, and 300 µA, respectively. The injector and transfer line temperatures were 250°C. For metabolite profiling and spectral analysis, the quadrupole was scanned from 50 to 650 *m/z* units. Structural information about GC-EI-MS features was obtained through spectral matching with the NIST 14 spectral library.

For targeted analyses of specific metabolites, the mass spectrometer monitored specific diagnostic ions for each compound. Each targeted metabolite was quantified in the selected ion monitoring mode in which ion chromatogram peak areas were determined at their corresponding retention times (Supplemental Tables 2 and 3). For trehalose, stable isotope-labeled ¹³C-trehalose was added to each specimen as an internal standard before derivatization (37). The peak areas of trehalose and ¹³C internal standards were calculated as a ratio (Supplemental Figure 15).

Measurement of bile acids by liquid chromatography-mass spectrometry. LC-ESI-MS/MS detection of each bile acid in fecal specimens or reference standards was performed with a Shimadzu UFLC coupled to a BetaSil C18 HPLC column (50 mm × 2.1 mm × 3 µm; Thermo Fisher Scientific) and an AB Sciex API 4000 QTrap mass spectrometer (AB Sciex) running in negative-ion electrospray ionization mode (ESI)

using a Turbo V ESI ion source. Authentic bile acid standards (Table 3) were purchased and used to prepare 1 µM samples in 80% methanol. HPLC was conducted with a 0.4 mL/min flow rate using the following gradient: Solvent A (0.1% formic acid) and Solvent B (90% acetonitrile with 0.1% formic acid) were held constant at 95% and 5%, respectively, for 1 minute. Solvent B was increased to 98% by 8 minutes, held at 98% for 1 minute, and then reduced again to 5% in 1 minute. The column was equilibrated in 5% Solvent B for 3 minutes between runs. Optimized instrument settings are reported in Supplemental Table 4.

Data preparation. Ion chromatograms were used to align peaks and determine peak areas using Mass Profiler Professional software (Agilent) for GC-MS data and Analyst (AB Sciex) for LC-MS/MS data from the QTrap. Because of the large dynamic range and strong skew of feature intensities, we transformed observed signals at level x to $\log_{10}(1+x)$ values prior to multivariate analyses.

Sparse logistic regression. We use the framework of logistic regression models to classify samples using their measured metabolomic features. Since there are many more metabolomic features than there are samples, we employed multiple measures to avoid overfitting the data. First, we enforced sparsity with an L1 penalty on the number of parameters selected as shown in Equation 1.

$$\min_{w,c} \sum_i |w_i| + C \sum_i^n (\exp(-y_i(X_i^T w + c)) + 1)$$

(Equation 1)

Equation 1. This analysis is incorporated as part of the python module scikit-learn (38). The L1 penalty introduces a trade-off between model goodness of fit and the number of incorporated features that is tunable by an additional penalization parameter, C , in Equation 1. Second, we evaluated model performance using repeated 5-fold cross-validation with random subsets to optimize the sparsity penalty, as well as to identify which features were used most frequently and were consistently predictive on the hold-out (testing) sets. Because overfitting on training data is generally expected to reduce performance on a hold-out set, this procedure allowed us to identify the penalization level that maximizes expected performance on the testing set.

Finally, we obtained the 6-feature logistic regression described in the main text though combination of the results from our repeated 5-fold cross-validated L1-penalized regressions, selecting the 6 metabolomic features most frequently obtained in these sparse logistic regressions. Using only those 6 features, we performed logistic regression (not L1-penalized) on the 124 samples, obtaining the 96.7% AUC in Figure 2D. We established 95% confidence intervals by further 5-fold cross-validation, keeping the 6 features fixed but varying their coefficient contributions according to each training subset, yielding the 95% CI: 85.6%–100%.

By way of contrast, we compared these results to a logistic regression performed on the full set of features without any sparsity criteria. As expected, since there are an order of magnitude more features than samples, it was possible to select regression coefficients that perfectly separate (AUC = 1) the 2 classes in this case. However, this separation is potentially meaningless because of overfitting. Similarly establishing a 95% confidence interval by 5-fold cross-validation, using all metabolomic features, yields the 95% CI: 84.6%–99.5%. While this is still very high—and indeed, is comparable to the CI for our regres-

sion using only 6 features—the outlier nature of the artificial perfectly separated result trained using all of the data is a warning of possible overfitting. At the same time, we noted that even with this potential for overfitting using all of the data, it performed no better than our 6-feature regression in terms of CI, while the 6-feature model of course provided much greater ease of interpretation.

We noted that this regression analysis on log-transformed signals does not normalize across samples, nor employ methods to treat the data in a compositional framework, despite the fact that relative abundances of metabolites are the biologically meaningful quantity. Nevertheless, this analysis successfully identifies features whose ratios are informative in predicting classes (see main text).

Sparse partial least squares–discriminatory analysis (sPLS-DA). To further assess the consistency of our data analysis results, we employed sPLS-DA to find a low-rank approximation of the feature data set that aims to maximally preserve the covariance between the dependent variable (EIA status) and the independent variables (the features) (39, 40). This technique identifies a matrix decomposition similar to PCA that best explains the relationship between the variables of interest using the fewest number of features possible. This analysis was conducted using the R package mixOmics (41). We conducted both single- and multivariable prediction using sPLS-DA. PLS-DA attempts to find a single decomposition of both the observations and the variable of interest such that the covariance between the projected observations and the projected variables is maximized in the projected space. In this setting with many more features than observations ($p \gg n$), there are typically many low-dimensional combinations of features that can capture variation in the variable of interest; moreover, these combinations will be typically dense in the sense that most features will appear with small but nonzero contributions to prediction. In contrast, the sparse version of PLS-DA, sPLS-DA, simultaneously models the observations while performing feature selection by maximizing the original objective function under conditions to minimize the number of features incorporated.

Network-based analysis of bile acids. The network representation of detected bile acids was defined here using the correlations across all 186 samples as edge weights, keeping the 5 highest positive cor-

relations associated with each bile acid (5 nearest neighbors). Communities were detected from this network using the GenLouvain and CHAMP packages (31, 42, 43). We selected the obtained 7-community partition for visualization in Figure 6A, with the network layout produced by the ForceAtlas2 algorithm in Gephi (<http://gephi.org>) (44).

Study approval. Approval was obtained from the Washington University Institutional Review Board with a waiver of informed consent to use specimens for this study. Patient and specimen evaluation for this cohort has recently been described (14).

Author contributions

JPH, ERD, JHK, CDB, and PJM originally developed the concept and designed the overall study approach. TH and CDB conducted fecal specimen cultures, ribotyping, and toxin EIA analyses. ERD, JHK, KAR, and JPH selected human specimens. JRC, JIR, and JPH conducted mass spectrometric analyses, prepared the data, and interpreted spectra. JIR, WHW, PJM, and JPH analyzed the mass spectrometric data. JPH, JIR, WHW, PJM, ERD, and JHK wrote the manuscript.

Acknowledgments

The authors acknowledge funding from the Centers for Disease Control and Prevention's investments to combat antibiotic resistance under award number 200-2016-91939 and a CDC Prevention Epicenters Program Grant (CU54 CK 000162). JHK acknowledges support by the Washington University Institute of Clinical and Translational Sciences (grant no. UL1TR000448) and a subaward (grant no. KL2TR000450) from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health. We are grateful for helpful conversations with Clifford McDonald and Alison Laufer from the United States Centers for Disease Control.

Address correspondence to: Jeffrey P. Henderson, Box 8051, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, Missouri 63110, USA. Phone: 314.362.7250; Email: hendersonj@wustl.edu.

1. Lessa FC, et al. Burden of *Clostridium difficile* infection in the United States. *N Engl J Med*. 2015;372(9):825–834.
2. Magill SS, et al. Changes in prevalence of health care-associated infections in U.S. hospitals. *N Engl J Med*. 2018;379(18):1732–1744.
3. McDonald LC, et al. Clinical practice guidelines for *Clostridium difficile* infection in adults and children: 2017 update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). *Clin Infect Dis*. 2018;66(7):987–994.
4. Kwon JH, Reske KA, Hink T, Burnham CA, Dubberke ER. Evaluation of correlation between pretest probability for *Clostridium difficile* infection and *Clostridium difficile* enzyme immunoassay results. *J Clin Microbiol*. 2017;55(2):596–605.
5. Casadevall A, Fang FC, Pirofski LA. Microbial virulence as an emergent property: consequences and opportunities. *PLoS Pathog*. 2011;7(7):e1002136.
6. Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. *Nat Rev Gastroenterol Hepatol*. 2016;13(9):508–516.
7. Palmieri LJ, et al. Inhibitory effect of ursodeoxycholic acid on *Clostridium difficile* germination is insufficient to prevent colitis: a study in hamsters and humans. *Front Microbiol*. 2018;9:2849.
8. Theriot CM, Bowman AA, Young VB. Antibiotic-induced alterations of the gut microbiota alter secondary bile acid production and allow for *Clostridium difficile* spore germination and outgrowth in the large intestine. *mSphere*. 2016;1(1):e00045-15.
9. Thanissery R, Winston JA, Theriot CM. Inhibition of spore germination, growth, and toxin activity of clinically relevant *C. difficile* strains by gut microbiota derived secondary bile acids. *Anaerobe*. 2017;45:86–100.
10. Fang FC, Polage CR, Wilcox MH. Point-counterpoint: what is the optimal approach for detection of *Clostridium difficile* infection? *J Clin Microbiol*. 2017;55(3):670–680.
11. Hink T, Burnham CA, Dubberke ER. A systematic evaluation of methods to optimize culture-based recovery of *Clostridium difficile* from stool specimens. *Anaerobe*. 2013;19:39–43.
12. Westblade LF, et al. Development and evaluation of a novel, semiautomated *Clostridium difficile* typing platform. *J Clin Microbiol*. 2013;51(2):621–624.
13. Alasmari F, Seiler SM, Hink T, Burnham CA, Dubberke ER. Prevalence and risk factors for asymptomatic *Clostridium difficile* carriage. *Clin Infect Dis*. 2014;59(2):216–222.
14. Dubberke ER, et al. *Clostridium difficile* colonization among patients with clinically significant diarrhea and no identifiable cause of diarrhea. *Infect Control Hosp Epidemiol*. 2018;39(11):1330–1333.
15. Stickland LH. Studies in the metabolism of the strict anaerobes (genus *Clostridium*): the chemical reactions by which *Cl. sporogenes* obtains its energy. *Biochem J*. 1934;28(5):1746–1759.
16. Stickland LH. Studies in the metabolism of the strict anaerobes (genus *Clostridium*): The oxidation of alanine by *Cl. sporogenes*. IV. The

- reduction of glycine by *Cl. sporogenes*. *Biochem J*. 1935;29(4):889–898.
17. Nisman B. The Stickland reaction. *Bacteriol Rev*. 1954;18(1):16–42.
 18. Neumann-Schaal M, Jahn D, Schmidt-Hohagen K. Metabolism the difficile way: The Key to the success of the pathogen *Clostridioides difficile*. *Front Microbiol*. 2019;10:219.
 19. Kim J, Darley D, Selmer T, Buckel W. Characterization of (R)-2-hydroxyisocaproate dehydrogenase and a family III coenzyme A transferase involved in reduction of L-leucine to isocaproate by *Clostridium difficile*. *Appl Environ Microbiol*. 2006;72(9):6062–6069.
 20. Britz ML, Wilkinson RG. Leucine dissimilation to isovaleric and isocaproic acids by cell suspensions of amino acid fermenting anaerobes: the Stickland reaction revisited. *Can J Microbiol*. 1982;28(3):291–300.
 21. Elsden SR, Hilton MG. Volatile acid production from threonine, valine, leucine and isoleucine by clostridia. *Arch Microbiol*. 1978;117(2):165–172.
 22. Elsden SR, Hilton MG, Waller JM. The end products of the metabolism of aromatic amino acids by Clostridia. *Arch Microbiol*. 1976;107(3):283–288.
 23. Bouillaut L, Self WT, Sonenshein AL. Proline-dependent regulation of *Clostridium difficile* Stickland metabolism. *J Bacteriol*. 2013;195(4):844–854.
 24. Dyer JK, Costilow RN. Fermentation of ornithine by *Clostridium sticklandii*. *J Bacteriol*. 1968;96(5):1617–1622.
 25. Li Q, et al. Deciphering the biosynthetic origin of L-allo-isoleucine. *J Am Chem Soc*. 2016;138(1):408–415.
 26. Brown JR, et al. Changes in microbiota composition, bile and fatty acid metabolism, in successful faecal microbiota transplantation for *Clostridioides difficile* infection. *BMC Gastroenterol*. 2018;18(1):131.
 27. Seekatz AM, et al. Restoration of short chain fatty acid and bile acid metabolism following fecal microbiota transplantation in patients with recurrent *Clostridium difficile* infection. *Anaerobe*. 2018;53:64–73.
 28. Robben J, Janssen G, Merckx R, Eysen H. Formation of delta 2- and delta 3-cholenic acids from bile acid 3-sulfates by a human intestinal *Fusobacterium* strain. *Appl Environ Microbiol*. 1989;55(11):2954–2959.
 29. Hofmann AF. Detoxification of lithocholic acid, a toxic bile acid: relevance to drug hepatotoxicity. *Drug Metab Rev*. 2004;36(3–4):703–722.
 30. Franco P, et al. Identification and quantification of oxo-bile acids in human faeces with liquid chromatography-mass spectrometry: A potent tool for human gut acidic sterolbiome studies. *J Chromatogr A*. 2019;1585:70–81.
 31. Weir WH, Gibson R, Mucha PJ. CHAMP package: Convex Hull of Admissible Modularity Partitions. <http://champ.readthedocs.io/en/latest>. Updated May 22, 2019. Accessed June 24, 2019.
 32. Edwards AN, McBride SM. Isolating and purifying *Clostridium difficile* spores. *Methods Mol Biol*. 2016;1476:117–128.
 33. Collins J, et al. Dietary trehalose enhances virulence of epidemic *Clostridium difficile*. *Nature*. 2018;553(7688):291–294.
 34. Wilson KH. Efficiency of various bile salt preparations for stimulation of *Clostridium difficile* spore germination. *J Clin Microbiol*. 1983;18(4):1017–1019.
 35. Battaglioli EJ, et al. *Clostridioides difficile* uses amino acids associated with gut microbial dysbiosis in a subset of patients with diarrhea. *Sci Transl Med*. 2018;10(464):eaam7019.
 36. McDonald LC, Diekema DJ. Point-counterpoint: active surveillance for carriers of toxigenic *Clostridium difficile* should be performed to guide prevention efforts. *J Clin Microbiol*. 2018;56(8):e00782–18.
 37. Sergin I, et al. Exploiting macrophage autophagy-lysosomal biogenesis as a therapy for atherosclerosis. *Nat Commun*. 2017;8:15750.
 38. Pedregosa F, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
 39. Lê Cao KA, Martin PG, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*. 2009;10:34.
 40. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*. 2008;7(1):Article 35.
 41. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13(11):e1005752.
 42. Weir WH, Emmons S, Gibson R, Taylor D, Mucha PJ. Post-processing partitions to identify domains of modularity optimization. *Algorithms*. 2017;10(3):93.
 43. Jeub LGS, Bazzi M, Jutla IS, Mucha PJ. A generalized Louvain method for community detection implemented in MATLAB. <http://netwiki.amath.unc.edu/GenLouvain>. Updated August 29, 2017. Accessed June 24, 2019.
 44. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. <http://gephi.org>. Accessed June 24, 2019.