A Computational Framework to Quantify Host-Microbiome Interactions in Clostridioides Difficile Infection

Shanlin Ke
Brigham and Women's Hospital Channing Division of Network Medicine https://orcid.org/0000-0003-4101-0574

Nira R. Pollock
Beth Israel Deaconess Medical Center

Xu-wen Wang
Brigham and Women's Hospital Channing Division of Network Medicine

Xinhua Chen
Beth Israel Deaconess Medical Center

Kaitlyn Daugherty
Beth Israel Deaconess Medical Center

Qianyun Lin
Beth Israel Deaconess Medical Center

Hua Xu
Beth Israel Deaconess Medical Center

Kevin W. Garey
University of Houston College of Pharmacy

Anne J. Gonzales-Luna
University of Houston College of Pharmacy

Yang-Yu Liu (yyl@channing.harvard.edu)
Brigham and Women's Hospital Channing Division of Network Medicine https://orcid.org/0000-0003-2728-4907

Ciarán P. Kelly
Beth Israel Deaconess Medical Center

Research

Keywords: C. difficile infection (CDI), gut microbiome, host immune markers, machine learning

Posted Date: August 27th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-65421/v1

A Computational Framework to Quantify Host-Microbiome Interactions in *Clostridioides difficile* Infection

Shanlin Ke[1,2], Nira R. Pollock[3,4], Xu-Wen Wang[1], Xinhua Chen[5], Kaitlyn Daugherty[5], Qianyun Lin[5], Hua Xu[5], Kevin W. Garey[6], Anne J. Gonzales-Luna[6], Ciarán P. Kelly[5]*, Yang-Yu Liu[1]*

[1]*Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.*

[2]*State Key Laboratory of Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University 330045, China.*

[3]*Division of Infectious Diseases, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02115, USA.*

[4]*Department of Laboratory Medicine, Boston Children's Hospital, Boston, Massachusetts 02115, USA.*

[5]*Division of Gastroenterology, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02115, USA.*

[6]*Department of Pharmacy Practice and Translation Research, University of Houston College of Pharmacy, Houston, Texas 77204, USA.*

Email addresses:

Shanlin Ke: spske@channing.harvard.edu

Nira R. Pollock: Nira.Pollock@childrens.harvard.edu

Xu-Wen Wang: xuwenwang86@gmail.com

Xinhua Chen: xchen1@bidmc.harvard.edu

Kaitlyn Daugherty: kdaugher@bidmc.harvard.edu

Qianyun Lin: qlin@bidmc.harvard.edu

Hua Xu: hxu@bidmc.harvard.edu

Kevin W. Garey: kgarey@central.uh.edu

Anne J. Gonzales-Luna: ajgonzales-luna@uh.edu

Ciarán P. Kelly: ckelly2@bidmc.harvard.edu

Yang-Yu Liu: yyl@channing.harvard.edu

*To whom correspondence should be addressed: Y.-Y.L. (yyl@channing.harvard.edu) or C.P.K. (ckelly2@bidmc.harvard.edu).

Abstract

Background: Clostridioides difficile infection (CDI) is the most common cause of healthcare–associated infection and an important cause of morbidity and mortality among hospitalized patients. A comprehensive understanding of C. difficile infection (CDI) pathogenesis is crucial for disease diagnosis, treatment and prevention. To achieve that, a quantitative study of host-microbiome interactions in CDI is a prerequisite. Yet, an effective computational framework to quantify host-microbiome interactions in CDI was lacking.

Methods: Here, we characterized gut microbial compositions and abroad panel of immunological markers in a comprehensive clinical cohort of 243 well-characterized human subjects with four different *C. difficile* infection/colonization statuses (CDI, Asymptomatic Carriage, Non-CDI Diarrhea, and Control). Based on microbial and immunological features, we developed a computational framework to detect CDI status using random forest and symbolic classification models.

Results: First, by calculating the correlations between microbial compositions and the circulating levels of host immune markers for each of the four phenotype groups, we found that the interactions between gut microbiota and host immune markers are very sensitive to the status of *C. difficile* colonization and infection. Second, we demonstrated that incorporating both gut microbiome and host immune marker data into random forest classifiers can better distinguish CDI from other groups than can either type of data alone. Finally, we performed symbolic classification using selected features from random forest classifiers to derive simple mathematic formulas that explicitly model the interactions between gut microbiome and host immune markers.

Conclusions: Overall, this study provides an effective computational framework to quantify the role of the intricate interactions between gut microbiota and host immune markers in CDI pathogenesis. This framework may inform the design of future diagnostic and therapeutic strategies.

Keywords: *C. difficile* infection (CDI); gut microbiome; host immune markers; machine learning

Introduction

Clostridioides difficile infection (CDI) is the most common cause of healthcare–associated infection and an important cause of morbidity and mortality among hospitalized patients[1-3]. Exposure to toxinogenic *C. difficile* can lead to a range of clinical outcomes ranging from asymptomatic colonization to mild diarrhea and more severe disease syndromes such as pseudomembranous colitis, toxic megacolon, bowel perforation, sepsis, and death[4,5]. Asymptomatic *C. difficile* carriage is characterized by *C. difficile* colonization in the absence of symptoms of infection. The diagnosis of CDI is based on clinical signs and symptoms in combination with laboratory testing, including enzyme immunoassays (EIA) for TcdA and TcdB, nucleic acid amplification tests (NAAT), selective toxinogenic culture, cell cytotoxicity neutralization assay, and glutamate dehydrogenase EIA[6-8]. However, currently available approaches do not accurately differentiate CDI from diarrhea with another cause in a patient colonized with toxinogenic *C. difficile*.

Current treatment strategies for CDI, including vancomycin, metronidazole and fidaxomicin, have inconsistent cure rates and treatment failure or CDI recurrence may occur in approximately one third of cases[9,10]. Antibiotic exposure is considered the most important factor predisposing patients to CDI[11,12]. In fact, treatments with antibiotics have a tremendous impact on the composition and functionality of the gut microbiota, and accordingly are associated with reduced colonization resistance against pathogens such as *C. difficile*[13-15]. It has been reported that several gut commensal bacteria may contribute to the prevention of *C. difficile* colonization and infection[16,17]. Once colonized, *C. difficile* can produce toxins that mediate a robust inflammatory response. Toxin A (TcdA) and toxin B (TcdB) are the primary virulence factors of *C. difficile*[18] and act on intestinal epithelial cells first, inducing pro-inflammatory cytokines, loss of tight junctions, cell detachment and an impaired mucosal barrier[19-21] leading to further exposure of immune cells to toxins. The innate and adaptive immune responses to CDI play crucial roles in disease onset, expression, severity, progression, and overall prognosis[22,23]. The innate immune defense mechanisms against *C. difficile* and its toxins include the commensal intestinal flora, mucosal barrier, intestinal epithelial cells, and mucosal immune system[24,25]. TcdA and TcdB have multiple effects on the innate immune system, including inducing expression of numerous pro-inflammatory mediators (e.g., cytokines, chemokines and neuroimmune peptides) and the recruitment and activation of a variety of innate immune cells[26,27]. Adaptive immunity is also sufficient to provide some protection from CDI, likely via antibody-mediated neutralization of TcdA and TcdB[28-31].

The role of the immune response combined with the knowledge that a balanced microbiota can prevent colonization and infection demonstrates the importance of combining both gut microbiota and host immune markers in understanding the pathogenesis of CDI. Yet, an effective computational framework to quantify the role of the intricate interactions between gut microbiota and host immune markers in human CDI pathogenesis has been lacking. Here we leverage tools from machine learning to develop such a framework. Machine learning has a great impact in many areas of medical research, as it offers a principled approach for developing

sophisticated, automatic, and objective algorithms for analysis of complex data. Indeed, previous studies indicate that supervised learning can be successfully employed for clinical disease assessment for diverse disorders[32-35]. In our previous work, we found that specific immune markers, particularly G-CSF, can be used to distinguish adults with CDI from other groups including asymptomatic carriers and NAAT-negative patients with and without diarrhea[36]. Here, we leverage machine learning tools to integrate the host immune marker data and newly obtained gut microbiome data from subjects of the same cohort to identify collections of bacteria and immune markers that can be associated with CDI. Our aim is to quantify the role of intricate interactions between gut microbiota and immune response in CDI pathogenesis, which can inform the design of future diagnostic and therapeutic strategies.

Results

Baseline demographic and clinical characteristics of participants

Our clinical cohort consists of 243 well-characterized recruited participants, who were divided into four groups (see Methods)[36]: (1) Control (n=47); (2) Non-CDI Diarrhea (n=44); (3) Asymptomatic Carriage (n=40); (4) CDI (n=112). The first three groups can be combined as the Non-CDI group. The entire clinical cohort had a mean ± SD age of 63.66 ± 14.85 year and was 48.15% female. Demographic data of the cohort are summarized in Table 1. In total, 187 participants (76.95%) had both gut microbiome and immune marker data available (see Table S1).

Microbial community structure

To compare the overall microbial community structure of the four groups, we first calculated the alpha diversity (i.e., the within-sample taxonomic diversity) of each sample at the genus level using four different measures: *taxa richness* (the observed number of different taxa present in the sample), *Chao1* (abundance-based estimator of taxa richness), *Evenness* (the uniformity of the population size of each taxa present in the sample), and *Shannon diversity index* (estimator of taxa richness and evenness: more weight on richness). As shown in Fig. 1:a-d, we found that taxa richness and Chao1 did not differ significantly among these groups. The gut microbiota of Non-CDI Diarrhea subjects showed lower evenness than that of the Control group. Shannon diversity was significantly lower in the Non-CDI Diarrhea and CDI groups than in the Control group.

To determine whether the gut microbial compositions of participants are affected by *C. difficile* infection/colonization status, we performed Principal Coordinates Analysis (PCoA) at the genus level using Bray-Curtis dissimilarity (which is a beta diversity measure to quantify the between-sample compositional dissimilarity). We found no distinct clusters corresponding to the four different phenotype groups, implying that the gut microbial compositions of participants from the four groups are not significantly different (Fig. 1e). Interestingly, by directly comparing the beta diversity of each group, we did find that the CDI group displays higher beta diversity than other groups (Fig. 1f), indicating that the microbial compositions of participants within the CDI group vary more prominently than other groups. Permutational multivariate analysis of

variance (PERMANOVA) showed that the overall bacterial composition differed significantly among different groups based on the CDI status ($P < 0.001$; Table S2), whereas other host factors such as age, sex, race and ethnicity had no significant effect on the microbiome composition.

To identify microbiome markers (i.e., certain taxa with very high discriminatory ability) to differentiate those different phenotype groups, we performed differential abundance analysis. In particular, we used ANCOM[37] (analysis of composition of microbiomes) with a Benjamini-Hochberg correction, and adjusted for age and sex. We found that the abundances of 15 genera were significantly different between CDI and Asymptomatic Carriage groups (Fig. 2a and Table S3). Among the 15 genera, 4 of them (*Veillonella*, *Enterobacter*, *Granulicatella* and *Dialister*) of these genera were enriched in the CDI group, while the other 11 genera (including *Lactococcus, Dorea*, *Moryella*, *Stenotrophomonas* and *Agathobacter*) were enriched in the Asymptomatic Carriage group. We also found 16 differentially abundant genera between the Non-CDI Diarrhea group and the CDI group (Fig. 2b and Table S4). Of these, 10 genera (including *Clostridioides*, *Enterobacter*, *Dialister*, and *Veillonella*) were enriched in the CDI group, and the other 6 genera (*[Eubacterium]_hallii_group*, *Collinsella*, *Agathobacter*, *Dorea*, *Stenotrophomonas and Streptococcus*) were enriched in the Non-CDI Diarrhea group. ANCOM analysis also enabled us to identify 40 genera (including *Clostridioides* and *Veillonella*) that have significant differential abundances between the CDI group and the whole Non-CDI group (Fig. 2c and Table S5). Note that a total of 6 differentially abundant genera were identified from all the three comparisons. Among them, *Veillonella*, *Enterobacter* and *Dialister* were enriched in the CDI group, while *Dorea, Stenotrophomonas* and *Agathobacter* were depleted in the CDI group.

Microbial correlation networks

To compare the microbial communities of the four groups at the network-level, we constructed the genus-level microbial correlation network for each group using SparCC[38]. We found that the microbial correlation network of the CDI group has quite different structure compared to other groups (Fig. 3). More precisely, it has fewer nodes and edges, lower average degree, but higher modularity (Table S6). These indicate that the overall microbial correlations in the CDI group are much weaker than those in other groups.

To analyze these patterns in more detail, we used NetShift[39] to identify potentially important "driver" taxa responsible for the change of microbial correlations. This analysis revealed 24 potential driver taxa linked with the change of microbial correlations between CDI and Asymptomatic Carriage groups (Fig. S1). The top driver taxa were *Alistipes*, *Clostridioides*, *Desulfovibrio*, *Eggerthella*, *Erysipelatoclostridium*, *Klebsiella*, *Odoribacter Proteus*, *[Ruminococcus]_torques_group*, *Streptococcus*, *Vagococcus* and *Veillonella*. We then identified 24 genera as potential driver taxa underlying the change of microbial correlations between CDI and Non-CDI Diarrhea groups (Fig. S2). The top driver taxa were *Alistipes*, *Buttiauxella*, *Citrobacter*, *Clostridium_sensu_stricto_13*, *Desulfovibrio*, *Klebsiella*, *Oscillibacter*, *Phascolarctobacterium*, *Streptococcus* and *Veillonella*. Finally, Netshift analysis revealed 38

potential driver taxa underlying the change of microbial correlations between CDI and Non-CDI groups. The top driver taxa were *Bifidobacterium*, *Clostridioides*, *Klebsiella*, *Oscillibacter*, *Streptococcus* and *Veillonella* (Fig. S3). Together, these results suggested that certain bacterial taxa (e.g., *Clostridioides*, *Klebsiella*, *Streptococcus* and *Veillonella*) could play an important role in driving the changes of microbial correlations in subjects with different *C. difficile* infection/colonization status.

Host immune markers and CDI

To determine the systemic levels of proinflammatory cytokines in CDI, we measured the circulating levels of granulocyte-colony stimulating factor (G-CSF), interleukin-1β (IL-1β), IL-2, IL-4, IL-6, IL-8, IL-10, IL-13, IL-15, monocyte chemoattractant protein-1 (MCP-1), vascular endothelial growth factor-A (VEGF-A), tumor necrosis factor-alpha (TNF-α), and serum concentrations of immunoglobulin A (IgA), IgG, and IgM antibodies against *C. difficile* toxin A and toxin B as previously reported[36]. We previously demonstrated specific markers of these innate and adaptive immunity that can distinguish CDI from each of the other three groups[36]. In the current study, we are particularly interested in comparing the CDI group and the combined Non-CDI group. Based on the Mann-Whitney U test, we identified in total 11 immune markers that displayed significantly different concentrations in these two groups, including G-CSF, IL-4, IL-6, IL-8, IL-10, IL-15, TNF-α, MCP1, IgA anti-toxin A and B, and IgG anti-toxin A in blood (Table S7). All of these immune markers had higher concentrations in the CDI group than in the Non-CDI group. Host immune marker variations between samples were evaluated using the Principal Component Analysis (PCA) (Fig. 1g). PCA plot showed no clear clustering of those subjects based on immune marker concentrations. However, boxplot of Euclidean distance of immune marker profiles from CDI patients showed higher within-group variation than that in all the other three groups (Fig. 1h). PERMANOVA analysis indicated that the immune homeostasis was significantly different among different groups based on the CDI status ($P = 0.016$; Table S2).

Interactions between gut microbiome and host immune markers

To reveal the interactions between the gut microbiome and the host immune system, we calculated the correlations between microbial compositions and the circulating levels of host immune markers for each of the four groups. The results are shown in Fig. 4 and Fig. S4. For the Control group, the most significant correlations were identified as *Christensenellaceae R-7 group* negatively correlated with TNF□, *Bifidobacterium* positively correlated with VEGFA and IL-13, *Rothia* positively correlated with IL-15, and *Veillonella* positively correlated with IL-4 (Fig. 4a and Fig. S4). For the Non-CDI Diarrhea group, *Ruminococcaceae UCG-011* was negatively correlated with IL-8 and IL-6, *Defluviitaleaceae UCG-011* was positively correlated with IL-1b, and *Blautia* was negatively correlated with MCP1 levels (Fig. 4b). For the Asymptomatic Carriage group, we found that *Lactobacillus* was negatively correlated with VEGFA, *Akkermansia* was positively correlated with IL-6, and *Enterococcus* was positively

correlated with TNF (Fig. 4c). For the CDI group, negative correlations involved *Akkermansia* and IL-10, *Lactococcus* and G-CSF, while positive correlations involved *Lactobacillus* and IgG and IgA anti-toxin B (Fig. 4d). Interestingly, none of these most significant correlations was universally present across different groups. This indicated that the interactions between gut microbiota and host immunological markers can be very sensitive to the status of *C. difficile* colonization and infection. Although the rudimentary correlation analysis cannot reveal any nonlinear interactions between gut microbiota and host immune markers, the result implies that the integration of gut microbiota and host immune markers might be quite useful for highly accurate classification of CDI.

Classification of CDI using host immune markers and gut microbiota

To determine whether host immune markers or gut microbiota could serve as biomarkers to classify subjects into different groups, we constructed a multi-class classifier based on random forests (RF). One of the most popular performance metrics of a classifier is the Area Under the receiver operating characteristic Curve (AUC). The performance of a multi-class classifier is measured by both micro-average and macro-average AUCs. We considered three different feature types: (1) host immune maker concentrations alone; (2) gut microbial compositions alone; and (3) the integration of (1) and (2) in our classification analysis. To eliminate confounding effects, we excluded the genus *Clostridioides* from our classification analysis. The immune marker-based classifier achieved macro-average AUC ~ 0.827 and micro-average AUC ~ 0.828 (Fig. S5a), which are quite comparable to the performance of microbiota-based classifier (Fig. S5b). Interestingly, integrating immune marker with gut microbiota showed much better classification performance (macro-average AUC ~ 0.926 and micro-average AUC ~ 0.869) (Fig. S5c).

We further performed binary classifications to distinguish CDI subjects from Asymptomatic Carriage, Non-CDI Diarrhea, and Non-CDI subjects, using different feature types (Fig. 5). The goal of this analysis was to assess whether any single taxon or immune marker could reliably differentiate CDI status. In the classification of CDI vs. Asymptomatic Carriage, we found that G-CSF and *Moryella* were the most important immune and microbial features, respectively (Fig. S6:a-b). But the classification based on G-CSF (or *Moryella*) alone did not yield very high performance: mean AUC ~ 0.817 (or 0.701), respectively (Fig. 5:a1-a2). When we used all the immune markers (or all the genera) as features, we achieved mean AUC ~ 0.867 (or 0.805), respectively (Fig. 5:a3-a4). Interestingly, when we integrated all the host immune markers and gut microbial composition data together, we achieved a much higher performance with mean AUC ~ 0.900 (Fig. 5:a5). In order to select a subset of features that is as discriminatory as the whole set of features, we followed the "1-SE" rule (i.e., one chooses the model with fewest features such that its classification performance is less than one standard error away from that of the model with all the features), and selected the following 4 features: 2 bacterial genera (*Moryella* and *Veillonella*) and 2 immune markers (G-CSF and IL-6) in classifying CDI and Asymptomatic Carriage groups (Fig. S6:g-j). The RF classifier with those

selected features displayed an outstanding classification performance, with mean AUC ~ 0.916 (Fig. 5:a6). Note that a significant negative correlation between *Moryella* and G-CSF was found in the Asymptomatic Carriage group (Fig. 4c), which might contribute to the outstanding performance of the RF classifier with *Moryella* and G-CSF as selected features.

In the classification of CDI vs. Non-CDI Diarrhea groups, we found that G-CSF and *[Eubacterium]_hallii_group* are the top immune and microbial features, respectively (Fig. S6:c-d). But the classification based on G-CSF (or *[Eubacterium]_hallii_group*) alone did not perform very well: mean AUC ~ 0.747 (or ~ 0.630), respectively (Fig. 5:b1-b2). When we used all the immune marker (or all the microbial genera) as features, we achieved mean AUC ~ 0.851 (or ~ 0.884), respectively (Fig. 5:b3-b4). By integrating all features from both host immune marker and gut microbial genera, we further improved the classification performance to mean AUC ~ 0.918 (Fig. 5:b5). Following the "1-SE" rule, we selected the following 5 features: 3 genera: *Enterococcus*, *Epulopiscium and [Eubacterium]_hallii_group*; and 2 immune markers: G-CSF and IgA anti-toxin A (Fig. S6:h-k). The RF classifier with those selected features achieved mean AUC ~ 0.917 (Fig. 5:b6), which is quite comparable to that of using all the features. Note that *Enterococcus* was found to be significantly associated with G-CSF in the Non-CDI Diarrhea group (Fig. 4b). This might partially explain the outstanding performance of the RF classifier with *Enterococcus* and G-CSF as selected features.

In the classification of CDI vs. Non-CDI groups, we found that G-CSF and *Curvibacter* are the top immune and microbial features, respectively (Fig. S6:e-f). Classification based on G-CSF (or *Curvibacter*) alone achieved mean AUC ~ 0.802 (or ~ 0.683), respectively (Fig. 5:c1-c2). When we used all the immune marker (or all the microbial genera) as features, we achieved mean AUC ~ 0.878 (or ~ 0.903), respectively (Fig. 5:c3-c4). Integrating all features from both host immune marker and gut microbial genera, we further improved the classification performance to mean AUC ~ 0.941 (Fig. 5:c5). Following the "1-SE" rule, we selected the following 10 features: 6 genera: *Stenotrophomonas*, *Curvibacter*, *Enterobacter*, *Anaerobacillus*, *Fusobacterium* and *Veillonella*; and 4 immune markers: G-CSF, IL-6, TNF-α and IgA anti-toxin B (Fig. S6:i-l). Classification with those well selected features achieved mean AUC ~ 0.929 (Fig. 5:c6).

Derive nonlinear interactions between gut microbiota and host immune markers using symbolic classification

As mentioned earlier, traditional correlation analysis cannot reveal any nonlinear interactions between gut microbiota and host immune markers. This fact and the outstanding classification results based on well-selected features prompt us to derive simple mathematical models to quantify the intricate interactions between gut microbiota and host immune markers. To achieve that, we leveraged symbolic classification (SC)[40,41], a genetic programming technique that automatically searches the space of mathematical expressions to find the model that best fits a given dataset. The fitness function in SC is a maximization function, and the number of generations is chosen based on the saturation of the fitness score (Fig. S7). Using the same set of

selected features and trained with the entire dataset, the SC model outperformed logistic regression (LR) in differentiating CDI (see Table 2).

Indeed, as shown in Table 2, we derived a simple SC model with selected features, reaching a very high accuracy (0.896) in distinguishing CDI subjects from Asymptomatic Carriage. Basically, for each subject i, we calculate the diagnostic score $f(i)$ that will be used for CDI diagnosis: the class of subject i is CDI if $f(i) > 0$; Asymptomatic Carriage, if $f(i) \leq 0$. Similarly, we derived a SC model with accuracy of 0.900 (or 0.882) in distinguishing CDI from Non-CDI Diarrhea (or Non-CDI) with the corresponding diagnostic score shown in Table 2. To ensure the SC models learned from the entire dataset are not overfitting, we performed cross-validation. With different training sets, SC will derive different mathematical formulas (i.e., diagnostic scores). However, those SC models learned from different training datasets demonstrated quite robust performance in terms of Accuracy, Precision, Recall and F1-score (see Table S8). More importantly, even trained with less data, the SC models still outperformed LR models learned from the entire dataset.

As shown in Table 2, in the formulas of the diagnostic score, we colored the gut microbiota (or host immune marker) features in red (or blue), respectively. It is clearly seen that any potential interactions between gut microbiota and host immune makers are completely ignored in the formulas derived from LR. But for the formulas derived from SC, nonlinear interactions between gut microbiota and host immune makers can be clearly seen. We emphasize that those nonlinear interactions are not always pairwise. Those explicit interaction terms could inform further mechanistic studies to further reveal the role of intricate interactions between gut microbiota and host immune markers in CDI pathogenesis.

Discussion

Consistent with previous studies[42-45], we found that the gut microbiota of CDI patients was characterized by lower Shannon diversity than that of the Control group. Interestingly, we observed an increased variation of both immune markers and gut microbial compositions in the CDI group with respective to other studied groups. This suggests that CDI is characterized by a significantly less stable microbiome and immune homeostasis. Our findings are in line with the Anna Karenina principle, which suggests that CDI linked changes in the microbiome and immune homeostasis are likely stochastic, leading to community instability[46-48].

We were able to identify several candidate driver taxa (e.g., *Desulfovibrio, Klebsiella, Streptococcus* and *Veillonella*) that played a key role in driving the changes of microbial correlation networks between CDI and Asymptomatic Carriage (or Non-CDI Diarrhea, Non-CDI) groups. Among those driver taxa, *Streptococcus* has previously been shown to produce lactate thus impacting *C. difficile* TcdA expression to alleviate CDI[49]. Previous study indicated that *Desulfovibrio* has a pathogenic role in ulcerative colitis due to its ability to generate sulfides[50]. *Klebsiella* bacteria have been increasingly shown to develop antimicrobial resistance, most recently to the class of antibiotics known as carbapenems[51,52]. It is thus possible that the

CDI pathogenesis is further enforced by the enrichment of antagonistic bacteria present in the gut microbiome of CDI subjects.

We developed classification models aimed at differentiating CDI status based on host immune markers and gut microbiome data. We were able to identify specific immune and microbial features that could accurately distinguish CDI subjects. In addition, most of the selected features identified by feature selection were also differentially abundant genera and differentially expressed immune markers. From the classification of CDI and Asymptomatic Carriage, we were able to select a few features with outstanding discriminability, including *Veillonella* and *Moryella*. Interestingly, a positive relationship between *Veillonella* and CDI has been identified in recent studies[53-56]. An important role for *Veillonella* in CDI is supported by the fact that *Veillonella* species were associated with low coprostanol levels that correlated strongly with CDI[53]. A similar negative relationship between *Moryella* species and CDI has previously been observed[57]. *Enterococcus*, a feature selected from the classification of CDI vs. Non-CDI Diarrhea, has been reported to be associated with CDI due to vancomycin resistance[58]. Consistent with the findings from previous reports[59,60], *Epulopiscium* was significantly enriched in the CDI group and played an important role in differentiating this comparison. Among those features selected from the classification of CDI and Non-CDI groups, *Enterobacter* and *Fusobacterium* have been considered as opportunistic pathogens involved in multiple diseases[61,62].

Machine learning method has the potential to identify biomarkers and aid in the diagnosis of many diseases. However, the learnt relationships between predictors and outcome are typically non-transparent, especially non-linear methods (i.e., decision tree learning)[63]. Classical logistic regression is one of the most common machine learning models in medicine. Yet, it fails to solve non-linear problems where there are multiple or non-linear decision boundaries[64]. Furthermore, the log odds scale in LR is hard to interpret[65]. Symbolic classification based on genetic programming is an automated technique to derive formulas from features[66]. Using the selected features from the random forests model, we demonstrated that the mathematical formulas automatically derived from symbolic classification have robust diagnostic accuracy to differentiate CDI patients from Asymptomatic Carriage (or Non-CDI Diarrhea, and Non-CDI groups). Specifically, symbolic classification provides explicit mathematic formulas as its output, which significantly improves the transparency of the learned relationship between predictors and outcomes.

We previously demonstrated the potential clinical utility of a specific immunological biomarker (i.e., G-CSF) for diagnosis of CDI[36]. This study leverages the same unique and well-characterized study cohort, allowing us to study integrated host immune marker and microbial signatures associated with CDI. The fundamental differences between this study and the previous one are the clinical utilization of comprehensive immunological and microbial markers to explore the pathogenesis of CDI, and to generate clinical diagnostic models to detect CDI. We acknowledge the following limitations of this study. First, the 16S rRNA sequencing may not have capture additional insights associated with CDI at the species or strain level. Second,

observed association does not prove a causal relationship, and further studies are needed to validate the mechanism underlying the observed associations between these biomarkers and CDI. Finally, further external validation of the classification models and derived formulas need to be performed on an additional cohort with same inclusion criteria as the current cohort.

Conclusion

Utilizing this well-characterized cohort and leveraging machine learning tools, we proposed an effective computational framework to quantify the role of the intricate interactions between gut microbiota and host immune markers in CDI pathogenesis. We believe this framework can inform the design of future diagnostics of CDI, as well as therapeutic strategies for its prevention and treatment.

Methods

Study cohort

The background and design of this cohort has been described in detail previously[67]. Basically, our clinical cohort consists of 243 well-characterized recruited participants, who were divided into four groups associated with different *C. difficile* infection/colonization statuses: (1) CDI (n=112): Eligible patients were inpatients ≥18 years old with new-onset diarrhea, positive clinical stool NAAT result, and a decision to treat for CDI. The stool sample was captured as a discarded sample, and a discarded serum sample collected within 24 hours of that stool sample also captured. Patients were excluded if the diagnostic stool specimen was >72 hours old, if they had received CDI treatment for >24 hours prior to stool collection, or if they had a colostomy. Assessment for the presence of diarrhea included review of nursing input/output logs for number and consistency of stools, consultation with treating clinicians, and detailed chart review (requiring mention of "diarrhea", "loose stools", and/or increased frequency, in notes written by multiple providers). Patients for whom there was any doubt about the presence of diarrhea, or who had chronic diarrhea, were excluded. (2) Asymptomatic Carriage (n=40): Eligible patients were inpatients ≥18 years old, admitted for at least 72 hours, who had received at least one dose of an antibiotics within the past 7 days, and did not have diarrhea in the 48 hours prior to stool specimen submission. Patients with 2 or more loose stools within 24 hours were excluded; patients who had 1 loose stool were included only if they had recently received a laxative. Patients were excluded if they had a colostomy; received oral or intravenous metronidazole, oral vancomycin, oral rifaximin, and/or oral fidaxomicin for >24 hours within the prior 7 days; had been diagnosed with CDI in the past 6 months; or had tested negative for *C. difficle* within the past 7 days. Stool specimen were collected prospectively under verbal informed consent. A discarded serum sample from within 24 hours of the stool specimen were also captured. NAAT (Xpert C. difficile/Epi) was performed on all samples, and positive samples retained as the Asymptomatic Carriage cohort. (3) Non-CDI Diarrhea (n=44): patients with diarrhea (confirmed using the same definition used for the CDI cohort) but had NAAT-negative stool on clinical *C.*

difficile testing; (4) Control (n=47): patients without diarrhea who had screened as eligible for the Asymptomatic Carriage cohort but were NAAT-negative on research stool testing. In our previous study[36], the four groups were named as (1) CDI-NAAT = CDI; (2) Carrier-NAAT = Asymptomatic Carriage; (3) diarrhea NAAT-negative = Non-CDI Diarrhea and (4) no Diarrhea NAAT-Negative = Control. In this work, for simplicity we used the simpler and more clearly descriptive titles.

Serum immune marker measurement

The measurement of host serum cytokines concentrations of IL-2, IL-4, IL-6, IL-8, IL-10, IL-13, IL-15, IL-1β, G-CSF, IL-1β, MCP-1, VEGF-A, and TNF-α was performed using a Milliplex magnetic bead kit and Luminex analyzer (MAGPIX) (Millipore Sigma, Inc., Burlington, MA) as per the manufacturer's instructions. Purified toxin A and B were separately prepared from *C. difficile* strain VPI 10463 (American Type Culture Collection 43255-FZ, Manassas, VA). Serum antibody (IgA, IgG, and IgM) levels against *C. difficile* toxins A and B were measured by semi-quantitative enzyme-linked immunosorbent assay (ELISA). All the experimental details have been reported previously[36,67].

Fecal DNA extraction and bacterial 16S rRNA sequencing data analysis

Stool DNA was extracted using the DNeasy PowerSoil Pro Kit (Qiagen, cat# 12888-100) in a QiaCube automated DNA extraction system (Qiagen) according to instructions. Briefly, 250mg stool was transferred into a PowerBead Pro Tube provided with the kit and 200 ug RNaseA and 800 µl of CD1 solution were added. Tubes were vortexed briefly, transferred into an adapter, and then vortexed at maximum speed for 10 min. Tubes were centrifuged at 15,000 xg for 1 min and about 500–600 µl supernatant was used for DNA extraction according to instructions. DNA were eluted in 70 µl elution solution C6 and stored at -80^{0}C until use. 16S rRNA microbiome characterization was performed by sequencing the V4 region of the 16S rRNA gene using the Illumina MiSeq[68]. Each sample was amplified using a barcoded primer, which yielded a unique sequence identifier tagged onto each individual sample library. Illumina-based sequencing yielded greater than 15,000 reads per sample. CLC Genomics Workbench version 12 (Qiagen) was used for OTU clustering and generation of abundance tables. Analyses were performed using the tutorial "OTU Clustering Step by Step" updated September 2, 2019 and available at: https://resources.qiagenbioinformatics.com/tutorials/OTU_Clustering_Steps.pdf

Microbial diversity and differential abundance analysis

The diversity measures and permutational multivariate analysis of variance (PERMANOVA) were calculated using the vegan package in R (see Supplementary Methods for details). For differential abundance analysis, we used ANCOM[37] (analysis of composition of microbiomes), with a Benjamini–Hochberg correction at 5% level of significance, and adjusted for age and sex. The Mann–Whitney U test was used to compare the difference of immune marker levels between different groups.

Microbial correlation network and microbiome-immune association analysis

The microbial correlation networks were constructed using SparCC[38] (sparse correlations for compositional data, https://github.com/luispedro/sparcc) (see Supplementary Methods for details). We also used the NetShift[39] (https://web.rniapps.net/netshift) to identify potential "driver" taxa underlying the differences of microbial correlation networks associated with CDI and Asymptomatic Carriage (or Non-CDI Diarrhea, and Non-CDI). The key driver taxa were identified based on the neighbor shift (NESH) score, Jaccard Index and delta betweenness (□B). Associations between the gut microbiota and host immune markers were quantified by Spearman correlation coefficients in combination with Benjamini-Hochberg FDR correction to account for multiple hypothesis testing (significance threshold □ □ 0.05). All included genera were required to be detected in □50% of all samples in each group.

Classification with Random Forests model

To build a classification model capable of testing the overall contribution of immunological or microbial data in distinguishing the CDI status, we developed a multi-class random forests (RF) classifier. The data is split into a training set and a test set, with 70% of the data forming the training data and the remaining 30% forming the test set. The performance of the multi-class model was measured by micro-average and macro-average AUC. To determine whether more specific host immune markers or gut microbial taxa could differentiate CDI subjects from Asymptomatic Carriage, Non-CDI Diarrhea and Non-CDI groups, we constructed the binary classifiers based on RF models with integrated immune markers and microbiome data (see Supplementary Methods for details).

Symbolic classification with genetic programming

We employed Karoo GP[69], a genetic programming application suite written in Python that support both symbolic regression (SR) and symbolic classification (SC) analysis, to derive simple formulas for CDI diagnosis. Due to the different training sets, SC will derive different formulas, but their classification performances are quite comparable (Table S8). The formulas shown in Table 2 were derived based on the whole dataset (for details see supplementary methods). To demonstrate the advantage of SC, for each classification task (i.e., CDI vs. Asymptomatic Carriage, CDI vs. Non-CDI Diarrhea, and CDI vs. Non-CDI), we also performed logistic regression (LR) using the same set of selected features as used in SC (Table 2) (see Supplementary Methods for details).

Declarations

Ethics approval and consent to participate

Approval of this study was given by the Beth Israel Deaconess Medical Center. All human subjects provided informed consent for participation in the study and collection and analysis of data.

Consent for publication

Not applicable.

Availability of data and material

Data will be available from corresponding authors upon reasonable request.

Competing interests

C.P.K. has acted as a paid consultant to Artugen, Facile Therapeutics, First Light Biosciences, Finch, Matrivax, Merck, Seres Health, and Vedanta and has received grant support from Merck. X. C. has acted as a paid consultant to Artugen. All other authors report no potential conflicts of interest.

Funding

Y.-Y.L. acknowledged grants from National Institutes of Health (R01AI141529, R01HD093761, UH3OD023268, U19AI095219 and U01HL089856). N.R.P. and C.P.K. acknowledged grants from National Institutes of Health (R01AI116596) and Institut Mérieux. S.K. was supported by the China Scholarship Council.

Authors' contributions

Y.-Y.L, N.R.P., X.C., and C.P.K. conceived and designed the project. C.P.K., N.R.P., X.C. and K.D. performed the clinical study. X.C., H.X., and Q.L. contributed to the serum immune marker measurement. K.W.G. and A.J.G. performed fecal DNA extraction and bacterial 16S rRNA sequencing. S.K., X.-W.W., and Y.-Y.L. performed all the data analysis and wrote the manuscript. N.R.P., K.W.G., C.P.K., and K.D. edited the manuscript.

Acknowledgements

The authors thank all patients who participated in this study, as well as Carolyn Alonso, Javier Villafuerte Gálvez, and the technologists in the Beth Israel Deaconess Medical Center Clinical Microbiology Laboratory for their help with sample collection. The authors thank Zheng Sun for valuable discussion on the microbiome data analysis. S.K. would like to acknowledge the support and help from Professor Lusheng Huang (Jiangxi Agricultural University).

References

1 Lessa, F. C. *et al.* Burden of Clostridium difficile infection in the United States. *N Engl J Med* **372**, 825-834, doi:10.1056/NEJMoa1408913 (2015).

2 Depestel, D. D. & Aronoff, D. M. Epidemiology of Clostridium difficile infection. *J Pharm Pract* **26**, 464-475, doi:10.1177/0897190013499521 (2013).

3 McDonald, L. C. *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* **66**, e1-e48, doi:10.1093/cid/cix1085 (2018).

4 Rupnik, M., Wilcox, M. H. & Gerding, D. N. Clostridium difficile infection: new developments in epidemiology and pathogenesis. *Nat Rev Microbiol* **7**, 526-536, doi:10.1038/nrmicro2164 (2009).

5 Schaffler, H. & Breitruck, A. Clostridium difficile - From Colonization to Infection. *Front Microbiol* **9**, 646, doi:10.3389/fmicb.2018.00646 (2018).

6 Tenover, F. C., Baron, E. J., Peterson, L. R. & Persing, D. H. Laboratory diagnosis of Clostridium difficile infection can molecular amplification methods move us out of uncertainty? *J Mol Diagn* **13**, 573-582, doi:10.1016/j.jmoldx.2011.06.001 (2011).

7 Burnham, C. A. & Carroll, K. C. Diagnosis of Clostridium difficile infection: an ongoing conundrum for clinicians and for clinical laboratories. *Clin Microbiol Rev* **26**, 604-630, doi:10.1128/CMR.00016-13 (2013).

8 Musher, D. M. *et al.* Detection of Clostridium difficile toxin: comparison of enzyme immunoassay results with results obtained by cytotoxicity assay. *J Clin Microbiol* **45**, 2737-2739, doi:10.1128/JCM.00686-07 (2007).

9 Bagdasarian, N., Rao, K. & Malani, P. N. Diagnosis and treatment of Clostridium difficile in adults: a systematic review. *JAMA* **313**, 398-408, doi:10.1001/jama.2014.17103 (2015).

10 Rineh, A., Kelso, M. J., Vatansever, F., Tegos, G. P. & Hamblin, M. R. Clostridium difficile infection: molecular pathogenesis and novel therapeutics. *Expert Rev Anti Infect Ther* **12**, 131-150, doi:10.1586/14787210.2014.866515 (2014).

11 Stevens, V., Dumyati, G., Fine, L. S., Fisher, S. G. & van Wijngaarden, E. Cumulative antibiotic exposures over time and the risk of Clostridium difficile infection. *Clin Infect Dis* **53**, 42-48, doi:10.1093/cid/cir301 (2011).

12 Slimings, C. & Riley, T. V. Antibiotics and hospital-acquired Clostridium difficile infection: update of systematic review and meta-analysis. *J Antimicrob Chemother* **69**, 881-891, doi:10.1093/jac/dkt477 (2014).

13 Lewis, B. B. *et al.* Loss of Microbiota-Mediated Colonization Resistance to Clostridium difficile Infection With Oral Vancomycin Compared With Metronidazole. *J Infect Dis* **212**, 1656-1665, doi:10.1093/infdis/jiv256 (2015).

14 Becattini, S., Taur, Y. & Pamer, E. G. Antibiotic-Induced Changes in the Intestinal Microbiota and Disease. *Trends Mol Med* **22**, 458-478, doi:10.1016/j.molmed.2016.04.003 (2016).
15 Buffie, C. G. *et al.* Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. *Infect Immun* **80**, 62-73, doi:10.1128/IAI.05496-11 (2012).
16 Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517**, 205-208, doi:10.1038/nature13828 (2015).
17 Rea, M. C. *et al.* Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against Clostridium difficile. *Proc Natl Acad Sci U S A* **107**, 9352-9357, doi:10.1073/pnas.0913554107 (2010).
18 Leffler, D. A. & Lamont, J. T. Clostridium difficile Infection. *N Engl J Med* **373**, 287-288, doi:10.1056/NEJMc1506004 (2015).
19 Genth, H., Dreger, S. C., Huelsenbeck, J. & Just, I. Clostridium difficile toxins: more than mere inhibitors of Rho proteins. *Int J Biochem Cell Biol* **40**, 592-597, doi:10.1016/j.biocel.2007.12.014 (2008).
20 Sun, X., He, X., Tzipori, S., Gerhard, R. & Feng, H. Essential role of the glucosyltransferase activity in Clostridium difficile toxin-induced secretion of TNF-alpha by macrophages. *Microb Pathog* **46**, 298-305, doi:10.1016/j.micpath.2009.03.002 (2009).
21 Riegler, M. *et al.* Clostridium difficile toxin B is more potent than toxin A in damaging human colonic epithelium in vitro. *J Clin Invest* **95**, 2004-2011, doi:10.1172/JCI117885 (1995).
22 Sun, X. & Hirota, S. A. The roles of host and pathogen factors and the innate immune response in the pathogenesis of Clostridium difficile infection. *Mol Immunol* **63**, 193-202, doi:10.1016/j.molimm.2014.09.005 (2015).
23 Kelly, C. P. & Kyne, L. The host immune response to Clostridium difficile. *J Med Microbiol* **60**, 1070-1079, doi:10.1099/jmm.0.030015-0 (2011).
24 Bibbo, S. *et al.* Role of microbiota and innate immunity in recurrent Clostridium difficile infection. *J Immunol Res* **2014**, 462740, doi:10.1155/2014/462740 (2014).
25 Iacob, S., Iacob, D. G. & Luminos, L. M. Intestinal Microbiota as a Host Defense Mechanism to Infectious Threats. *Front Microbiol* **9**, 3328, doi:10.3389/fmicb.2018.03328 (2018).
26 Madan, R. & Petri, W. A., Jr. Immune responses to Clostridium difficile infection. *Trends Mol Med* **18**, 658-666, doi:10.1016/j.molmed.2012.09.005 (2012).
27 Sun, X., Savidge, T. & Feng, H. The enterotoxicity of Clostridium difficile toxins. *Toxins (Basel)* **2**, 1848-1880, doi:10.3390/toxins2071848 (2010).
28 Kyne, L., Warny, M., Qamar, A. & Kelly, C. P. Association between antibody response to toxin A and protection against recurrent Clostridium difficile diarrhoea. *Lancet* **357**, 189-193, doi:10.1016/S0140-6736(00)03592-3 (2001).

29 Wilcox, M. H. *et al.* Bezlotoxumab for Prevention of Recurrent Clostridium difficile Infection. *N Engl J Med* **376**, 305-317, doi:10.1056/NEJMoa1602615 (2017).
30 Giannasca, P. J. *et al.* Serum antitoxin antibodies mediate systemic and mucosal protection from Clostridium difficile disease in hamsters. *Infect Immun* **67**, 527-538 (1999).
31 Johnston, P. F., Gerding, D. N. & Knight, K. L. Protection from Clostridium difficile infection in CD4 T Cell- and polymeric immunoglobulin receptor-deficient mice. *Infect Immun* **82**, 522-531, doi:10.1128/IAI.01273-13 (2014).
32 Abos, A. *et al.* Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning. *Sci Rep* **7**, 45347, doi:10.1038/srep45347 (2017).
33 Dagliati, A. *et al.* Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol* **12**, 295-302, doi:10.1177/1932296817706375 (2018).
34 Mossotto, E. *et al.* Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci Rep* **7**, 2427, doi:10.1038/s41598-017-02606-2 (2017).
35 Schubert, A. M. *et al.* Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *mBio* **5**, e01021-01014, doi:10.1128/mBio.01021-14 (2014).
36 Kelly, C. P. *et al.* Host Immune Markers Distinguish Clostridioides difficile Infection From Asymptomatic Carriage and Non-C. difficile Diarrhea. *Clin Infect Dis*, doi:10.1093/cid/ciz330 (2019).
37 Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* **26**, 27663, doi:10.3402/mehd.v26.27663 (2015).
38 Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687, doi:10.1371/journal.pcbi.1002687 (2012).
39 Kuntal, B. K., Chandrakar, P., Sadhu, S. & Mande, S. S. 'NetShift': a methodology for understanding 'driver microbes' from healthy and disease microbiome datasets. *ISME J* **13**, 442-454, doi:10.1038/s41396-018-0291-x (2019).
40 Bannister, C. A., Halcox, J. P., Currie, C. J., Preece, A. & Spasic, I. A genetic programming approach to development of clinical prediction models: A case study in symptomatic cardiovascular disease. *PLoS One* **13**, e0202685, doi:10.1371/journal.pone.0202685 (2018).
41 Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81-85, doi:10.1126/science.1165893 (2009).
42 Song, Y. *et al.* Microbiota dynamics in patients treated with fecal microbiota transplantation for recurrent Clostridium difficile infection. *PLoS One* **8**, e81330, doi:10.1371/journal.pone.0081330 (2013).
43 Milani, C. *et al.* Gut microbiota composition and Clostridium difficile infection in hospitalized elderly individuals: a metagenomic study. *Sci Rep* **6**, 25945, doi:10.1038/srep25945 (2016).

44 Jiang, Z. D. *et al.* Randomised clinical trial: faecal microbiota transplantation for recurrent Clostridum difficile infection - fresh, or frozen, or lyophilised microbiota from a small pool of healthy donors delivered by colonoscopy. *Aliment Pharmacol Ther* **45**, 899-908, doi:10.1111/apt.13969 (2017).
45 Shankar, V. *et al.* Species and genus level resolution analysis of gut microbiota in Clostridium difficile patients following fecal microbiota transplantation. *Microbiome* **2**, 13, doi:10.1186/2049-2618-2-13 (2014).
46 Zaneveld, J. R., McMinds, R. & Vega Thurber, R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* **2**, 17121, doi:10.1038/nmicrobiol.2017.121 (2017).
47 Giongo, A. *et al.* Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J* **5**, 82-91, doi:10.1038/ismej.2010.92 (2011).
48 Caussy, C. *et al.* A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. *Nat Commun* **10**, 1406, doi:10.1038/s41467-019-09455-9 (2019).
49 Kolling, G. L. *et al.* Lactic acid production by Streptococcus thermophilus alters Clostridium difficile infection and in vitro Toxin A production. *Gut Microbes* **3**, 523-529, doi:10.4161/gmic.21757 (2012).
50 Rowan, F. *et al.* Desulfovibrio bacterial species are increased in ulcerative colitis. *Dis Colon Rectum* **53**, 1530-1536, doi:10.1007/DCR.0b013e3181f1e620 (2010).
51 Arnold, R. S. *et al.* Emergence of Klebsiella pneumoniae carbapenemase-producing bacteria. *South Med J* **104**, 40-45, doi:10.1097/SMJ.0b013e3181fd7d5a (2011).
52 Navon-Venezia, S., Kondratyeva, K. & Carattoli, A. Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol Rev* **41**, 252-275, doi:10.1093/femsre/fux013 (2017).
53 Antharam, V. C. *et al.* An Integrated Metabolomic and Microbiome Analysis Identified Specific Gut Microbiota Associated with Fecal Cholesterol and Coprostanol in Clostridium difficile Infection. *PLoS One* **11**, e0148824, doi:10.1371/journal.pone.0148824 (2016).
54 Khanna, S. *et al.* Gut microbiome predictors of treatment response and recurrence in primary Clostridium difficile infection. *Aliment Pharmacol Ther* **44**, 715-727, doi:10.1111/apt.13750 (2016).
55 Han, S. H., Yi, J., Kim, J. H., Lee, S. & Moon, H. W. Composition of gut microbiota in patients with toxigenic Clostridioides (Clostridium) difficile: Comparison between subgroups according to clinical criteria and toxin gene load. *PLoS One* **14**, e0212626, doi:10.1371/journal.pone.0212626 (2019).
56 Daquigan, N., Seekatz, A. M., Greathouse, K. L., Young, V. B. & White, J. R. High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden. *NPJ Biofilms Microbiomes* **3**, 35, doi:10.1038/s41522-017-0043-0 (2017).
57 Hudson, L. E., Anderson, S. E., Corbett, A. H. & Lamb, T. J. Gleaning Insights from Fecal Microbiota Transplantation and Probiotic Studies for the Rational Design of

Combination Microbial Therapies. *Clin Microbiol Rev* **30**, 191-231, doi:10.1128/CMR.00049-16 (2017).

58 Fujitani, S., George, W. L., Morgan, M. A., Nichols, S. & Murthy, A. R. Implications for vancomycin-resistant Enterococcus colonization associated with Clostridium difficile infections. *Am J Infect Control* **39**, 188-193, doi:10.1016/j.ajic.2010.10.024 (2011).

59 Antharam, V. C. *et al.* Intestinal dysbiosis and depletion of butyrogenic bacteria in Clostridium difficile infection and nosocomial diarrhea. *J Clin Microbiol* **51**, 2884-2892, doi:10.1128/JCM.00845-13 (2013).

60 Sokol, H. *et al.* Specificities of the intestinal microbiota in patients with inflammatory bowel disease and Clostridium difficile infection. *Gut Microbes* **9**, 55-60, doi:10.1080/19490976.2017.1361092 (2018).

61 Mezzatesta, M. L., Gona, F. & Stefani, S. Enterobacter cloacae complex: clinical impact and emerging antibiotic resistance. *Future Microbiol* **7**, 887-902, doi:10.2217/fmb.12.61 (2012).

62 Umana, A. *et al.* Utilizing Whole Fusobacterium Genomes To Identify, Correct, and Characterize Potential Virulence Protein Families. *J Bacteriol* **201**, doi:10.1128/JB.00273-19 (2019).

63 Deng, H. Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics* **7**, 277-287, doi:10.1007/s41060-018-0144-8 (2019).

64 Tollenaar, N. & van der Heijden, P. G. M. Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLoS One* **14**, e0213245, doi:10.1371/journal.pone.0213245 (2019).

65 Norton, E. C. & Dowd, B. E. Log Odds and the Interpretation of Logit Models. *Health Serv Res* **53**, 859-878, doi:10.1111/1475-6773.12712 (2018).

66 Liu, K. H. & Xu, C. G. A genetic programming-based approach to the classification of multiclass microarray datasets. *Bioinformatics* **25**, 331-337, doi:10.1093/bioinformatics/btn644 (2009).

67 Pollock, N. R. *et al.* Comparison of Clostridioides difficile Stool Toxin Concentrations in Adults With Symptomatic Infection and Asymptomatic Carriage Using an Ultrasensitive Quantitative Immunoassay. *Clin Infect Dis* **68**, 78-86, doi:10.1093/cid/ciy415 (2019).

68 Fadrosh, D. W. *et al.* An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* **2**, 6, doi:10.1186/2049-2618-2-6 (2014).

69 Staats, K., Pantridge, E., Cavaglia, M., Milovanov, I. & Aniyan, A. TensorFlow Enabled Genetic Programming. *arXiv e-prints*, arXiv:1708.03157 (2017). <https://ui.adsabs.harvard.edu/abs/2017arXiv170803157S>.

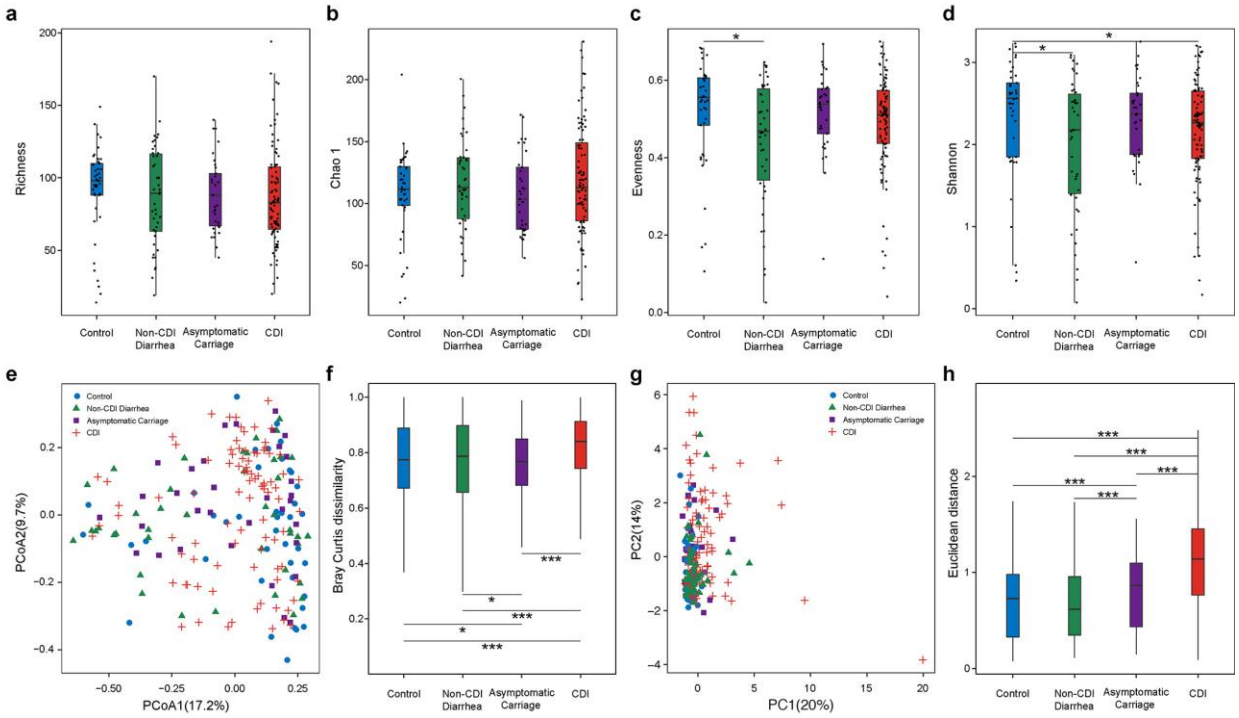

Fig. 1. Comparing the diversity of the gut microbiota (and host immune markers) of subjects with different *C. difficile* infection/colonization statuses (Control, Non-CDI Diarrhea, Asymptomatic Carriage, and CDI). **(a)** Taxa richness. **(b)** Chao1. **(c)** Evenness. **(d)** Shannon index. **(e)** Principal Coordinates Analysis (PCoA) plot based on Bray–Curtis dissimilarities of microbial compositions. **(f)** Boxplot of the gut microbiome Bray–Curtis dissimilarity between subjects within each group. **(g)** Principle component analysis (PCA) plot of host immune marker concentrations. **(h)** Boxplot of the Euclidean distance for the host immune markers of subjects within each group. Statistical significance was determined by Mann–Whitney test, $*P < 0.05$.

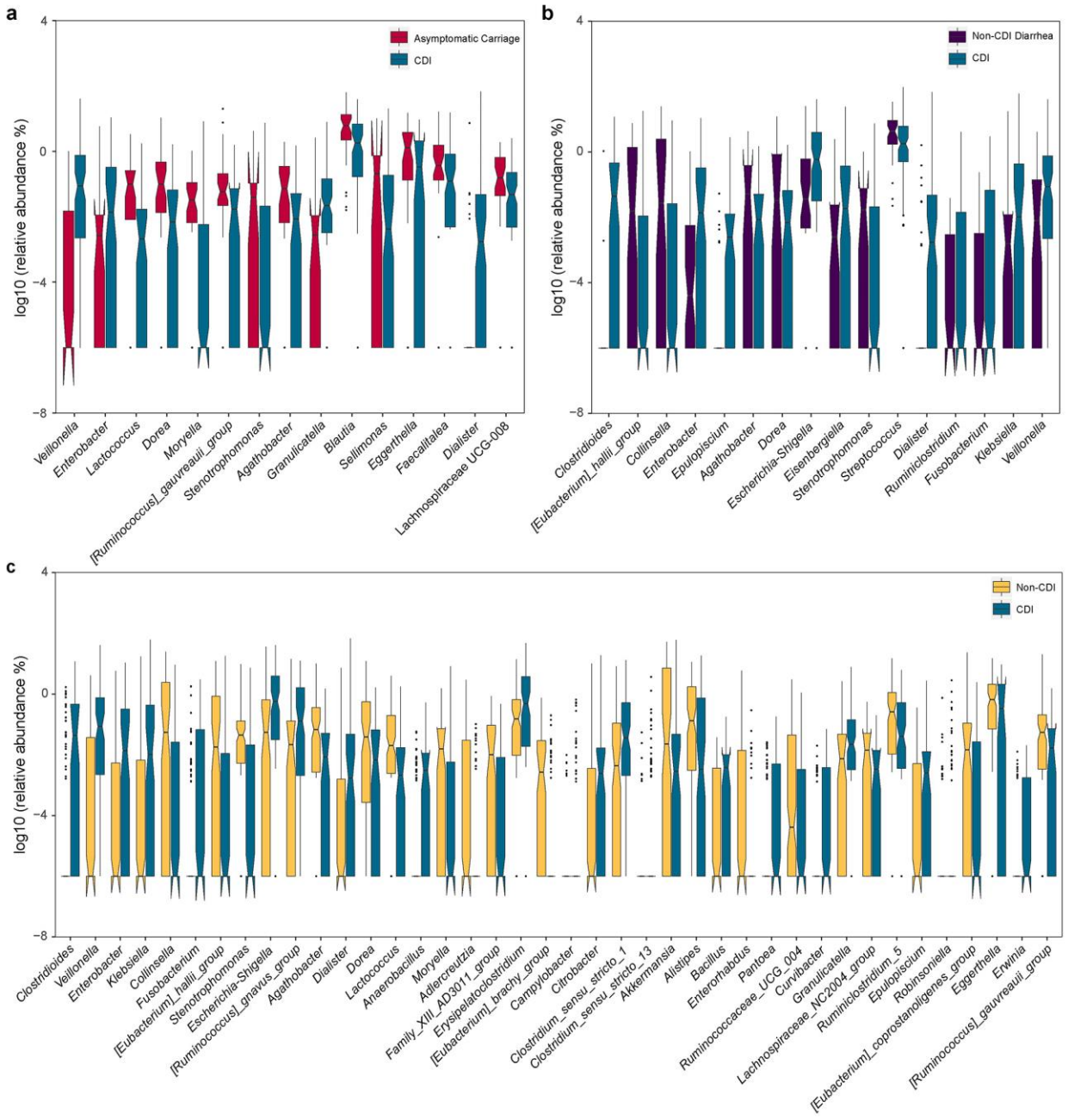

Fig. 2. Relative abundances of differentially abundant genera identified by ANCOM in comparing different groups. (**a**) CDI vs. Asymptomatic Carriage. (**b**) CDI vs. Non-CDI Diarrhea. (**c**) CDI vs. Non-CDI. The top differentially abundant taxa were ranked based on their W statistics (from left to right). The relative abundance (%) are plotted on log10 scale. The notches in the boxplots show the 95% confidence interval around the median.

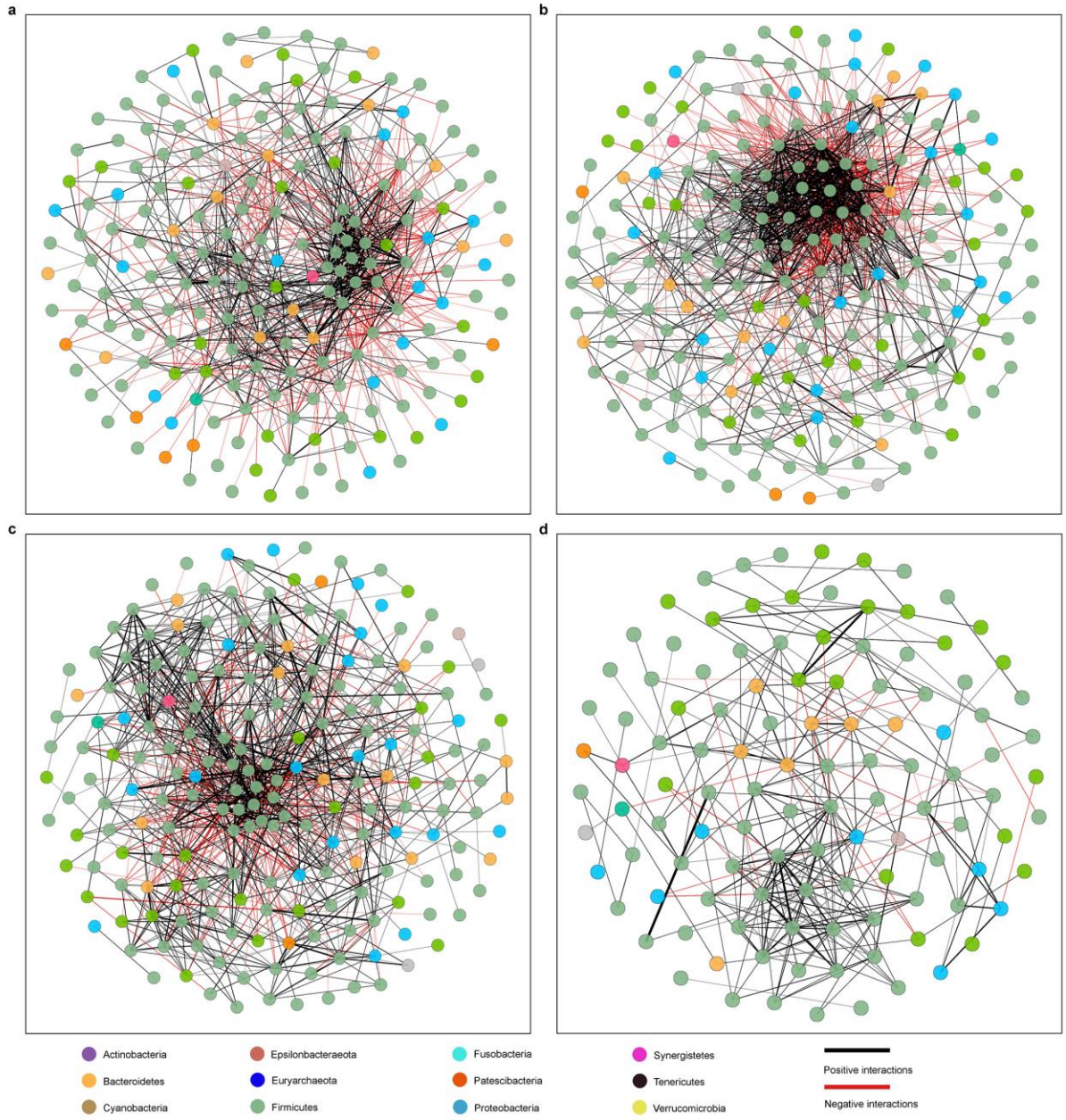

Fig. 3. Microbial correlation networks of different groups. (a) Control. **(b)** Non-CDI Diarrhea. **(c)** Asymptomatic Carriage. **(d)** CDI. Nodes represent genera and are colored based on their phylum. Edges represent microbial correlations: green/red means positive/negative correlations, respectively. Edge thickness indicates correlation strength, and only the high-confidence interactions (p-value < 0.05) with high absolute correlation coefficients (> 0.3) were presented. For each group, we further identified the top-three most connected genera/nodes. They are *Ruminococcus_1*, *Roseburia* and *Lachnospiraceae_UCG-008* for the Control group, *[Ruminococcus]_torques_group*, *[Eubacterium]_hallii_group* and *Blautia* for the Non-CDI Diarrhea group, *Ruminiclostridium_5*, *Enterococcus* and *Lachnospiraceae_UCG_008* for the Asymptomatic Carriage group, and *Alistipes*, *Ruminiclostridium_5* and *Lachnoclostridium* for the CDI group.

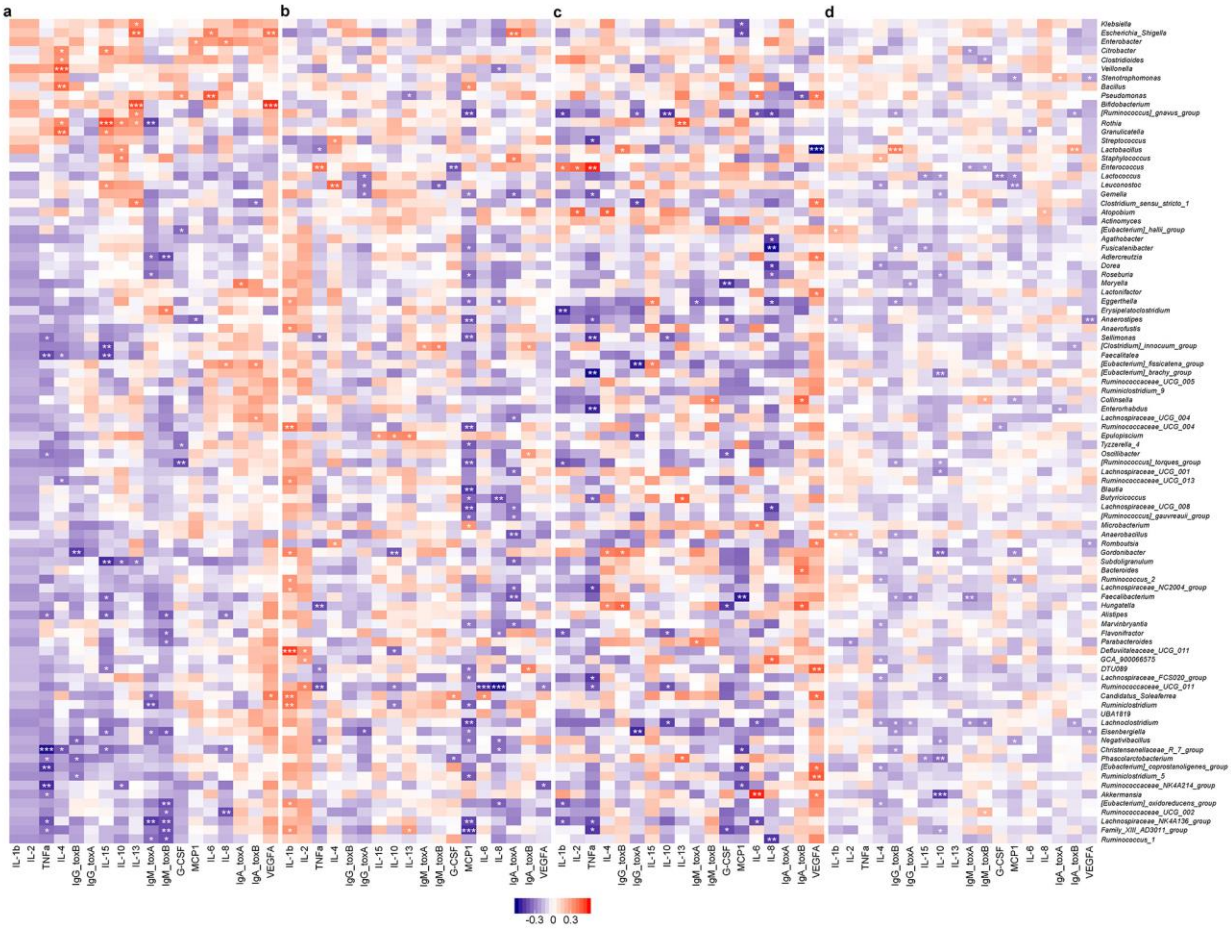

Fig. 4. Correlations between gut microbial abundances and host immune markers in different groups, quantified by Spearman correlation with Benjamini-Hochberg correction. (**a**) Control. (**b**) Non-CDI Diarrhea. **(c)** Asymptomatic Carriage. **(d)** CDI. Rows represent genera; columns represent immune markers. The layout of the heatmap is followed the hierarchical clustering results of Control cohort (see Fig.S4). Red/blue represents positive/negative correlation, respectively. The intensity of the colors denotes the strength of the correlation. *□ < 0.05, **□ < 0.01, ***□ < 0.001.

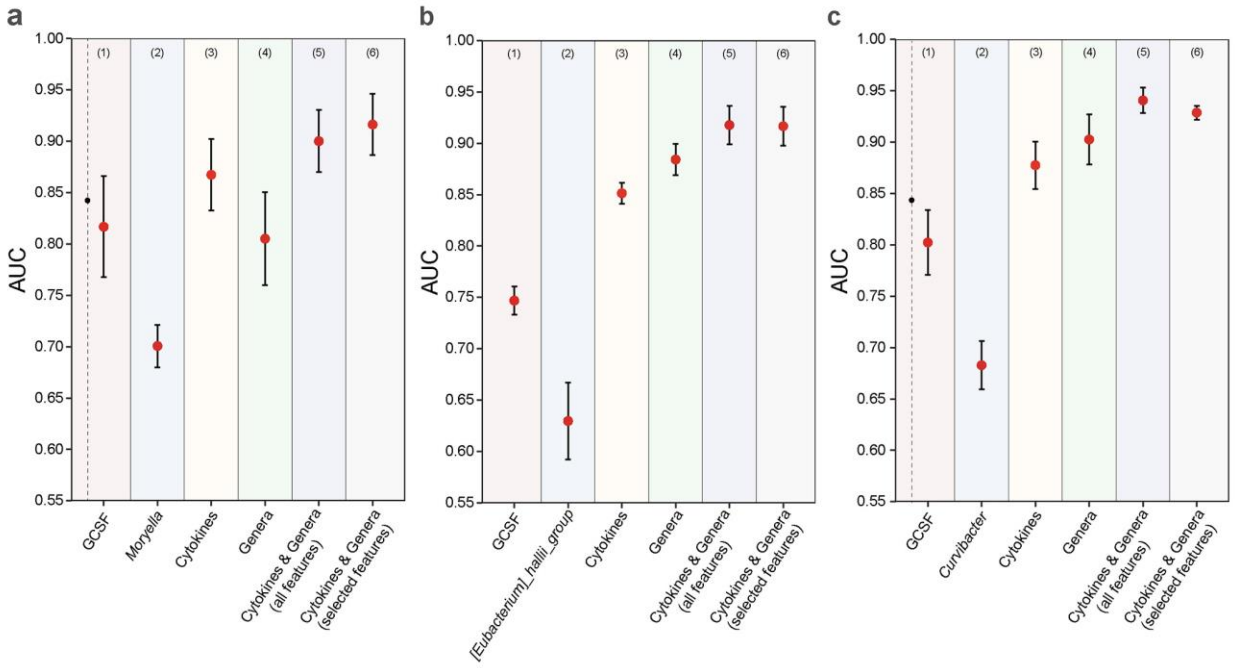

Fig. 5. The performance of RF-based classification models based on various types of features in differentiating CDI from other groups. (a) CDI vs. Asymptomatic Carriage. **(b)** CDI vs. Non-CDI Diarrhea. **(c)** CDI vs. Non-CDI. For each classification task, we used different types of features: (1) the top-1 immune marker feature (based on mean decrease accuracy); (2) the top-1 genus feature; (3) all immune markers; (4) all genera; (5) integration of all immune markers and genera; (6) selected features from the set of all immune markers and genera. Error bars represent the standard errors of the means (SEM).

Table 1. Demographic characteristics of the enrolled subjects.

	NAAT negative		NAAT positive	
Characteristics	**Control (n=47)**	**Non-CDI Diarrhea (n=44)**	**Asymptomatic Carriage (n=40)**	**CDI (n=112)**
Sex				
Female	14 (29.79%)	22 (50.00%)	20 (50.00%)	61 (54.46%)
Male	33 (70.21%)	22 (50.00%)	20 (50.00%)	51 (45.54%)
Age, Avg □ SD	62.40 □ 12.33	63.07 □ 13.15	62.15 □ 17.25	64.99 □ 15.62
Ethnicity				
Hispanic	1 (2.13%)	3 (6.82%)	1 (2.50%)	6 (5.36%)
Non-Hispanic	38 (80.85%)	37 (84.09%)	31 (77.50%)	96 (85.71%)
Unknown	8 (17.02%)	4 (9.09%)	8 (20.00%)	10 (8.93%)
Race				
White	33 (70.21%)	28 (63.64%)	28 (70.00%)	89 (79.46%)
Other	4 (8.51%)	10 (22.73%)	3 (7.50%)	23 (20.54%)
Unknown	10 (21.28%)	6 (13.64%)	9 (22.50%)	0 (0.00%)

Table 2. Diagnostic scores derived from symbolic classification (SC) and logistic regression (LR). For each subject i, we calculate his/her diagnostic score $f(i)$ (or $p(i)$) based on one of the following formulas derived from SC (or LR), respectively. For SC, the class of subject i is CDI if $f(i) > 0$; or Asymptomatic Carriage (or Non-CDI Diarrhea, Non-CDI) if $f(i) \leq 0$. For LR, the class of subject i is CDI if $p(i)$ □ 0.5; or Asymptomatic Carriage (or Non-CDI Diarrhea, Non-CDI) if $p(i)$ □ 0.5. Here, both $f(i)$ and $p(i)$ were learned from the entire dataset. Features used here include: x_1: GCSF; x_2: IgA_toxA; x_3: IgA_toxB; x_4: IL6; x_5: TNF□; x_6: *Anaerobacillus*; x_7: *Curvibacter*; x_8: *Enterobacter*; x_9: *Enterococcus*; x_{10}: *Epulopiscium*; x_{11}: *[Eubacterium]_haillii_group*; x_{12}: *Fusobacterium*; x_{13}: *Moryella*; x_{14}: *Stenotrophomonas*; x_{15}: *Veillonella*. In particular, for each classification task (regardless of using SC or LR), the following selected features were: (1) CDI vs. Asymptomatic Carriage: x_1, x_4, x_{13} and x_{15}; (2) CDI vs. Non-CDI Diarrhea: x_1, x_2, x_9, x_{10}, and x_{11}; (3) CDI vs. Non-CDI: x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_{12}, x_{14} and x_{15}. Note that in the calculation of precision, recall and F1-score, we can treat either CDI (or Asymptomatic Carriage, Non-CDI Diarrhea, Non-CDI) as the true positive. Results shown in the parenthesis represent the latter case.

Model	Diagnostic	Formula	Accuracy	Precision	Recall	F1-score
SC	CDI vs. Asymptomatic Carriage	$f(i) = x_1x_{15}(x_1^3 - 0.2x_{13} + 0.4) + 1.1x_1 - 0.1x_4 - 18.25$	0.896	0.914 (0.840)	0.949 (0.75)	0.931 (0.792)
	CDI vs. Non-CDI Diarrhea	$f(i) = x_9x_2(0.5x_{10} - 1) + x_{11}(0.02x_{11} - x_1) + x_2\left(1 - \frac{10}{x_1}\right) - \frac{0.003}{x_9}$	0.900	0.946 (0.826)	0.897 (0.905)	0.921 (0.864)
	CDI vs. Non-CDI	$f(i) = x_1x_3(0.2x_1x_5x_6x_{14} + 0.04x_1x_7 + 0.3x_1x_{15}x_8^4 + x_1x_{12}(0.5x_7 + x_1x_{14}) + x_7(0.1x_4 - x_6) + x_{14}(x_{14} - 2)$	0.882	0.889 (0.878)	0.821 (0.927)	0.853 (0.902)
LR	CDI vs. Asymptomatic Carriage	$\log\left(\frac{p(i)}{1-p(i)}\right) = 0.66725 - 0.04442x_1 + 0.01022x_4 + 7.51484x_{13} - 85.00213x_{15}$	0.830	0.895 (0.667)	0.872 (0.714)	0.883 (0.690)
	CDI vs. Non-CDI Diarrhea	$\log\left(\frac{p(i)}{1-p(i)}\right) = 0.01974 - 0.002084x_1 - 0.02391x_2 + 1.895x_9 - 12740x_{10} + 163.9x_{11}$	0.800	0.814 (0.765)	0.897 (0.619)	0.854 (0.684)
	CDI vs. Non-CDI	$\log\left(\frac{p(i)}{1-p(i)}\right) = 2.122 - 0.01002x_1 + 0.01833x_3 - 0.006334x_4 - 0.009566x_5 - 4609x_6 - 8576x_7 - 40.75x_8 - 101.1x_{12} + 32.84x_{14} - 43.4x_{15}$	0.813	0.841 (0.798)	0.679 (0.908)	0.752 (0.850)

Figures

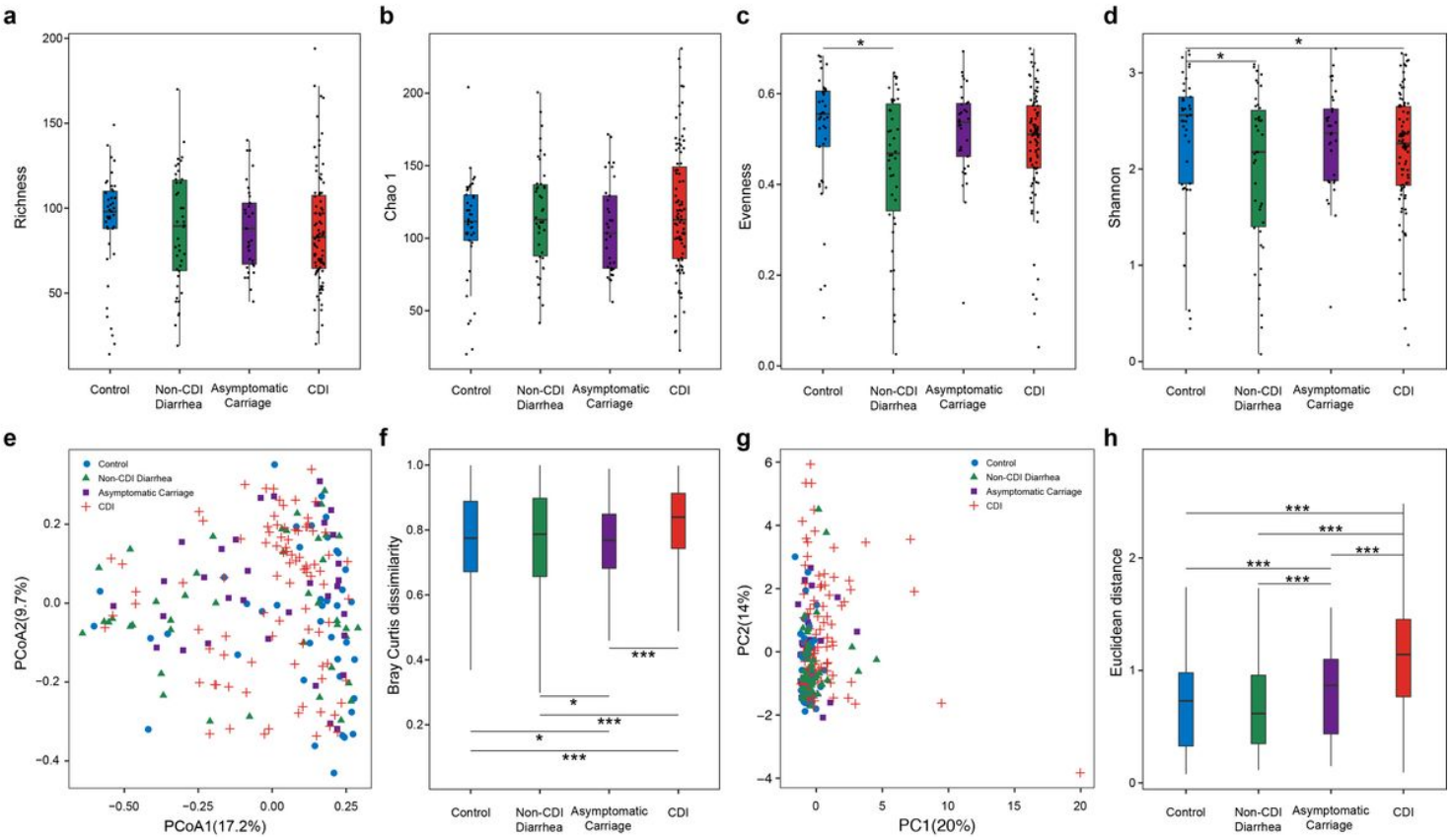

Figure 1

Comparing the diversity of the gut microbiota (and host immune markers) of subjects with different C. difficile infection/colonization statuses (Control, Non-CDI Diarrhea, Asymptomatic Carriage, and CDI). (a) Taxa richness. (b) Chao1. (c) Evenness. (d) Shannon index. (e) Principal Coordinates Analysis (PCoA) plot based on Bray–Curtis dissimilarities of microbial compositions. (f) Boxplot of the gut microbiome Bray–Curtis dissimilarity between subjects within each group. (g) Principle component analysis (PCA) plot of host immune marker concentrations. (h) Boxplot of the Euclidean distance for the host immune markers of subjects within each group. Statistical significance was determined by Mann–Whitney test, *$P < 0.05$.

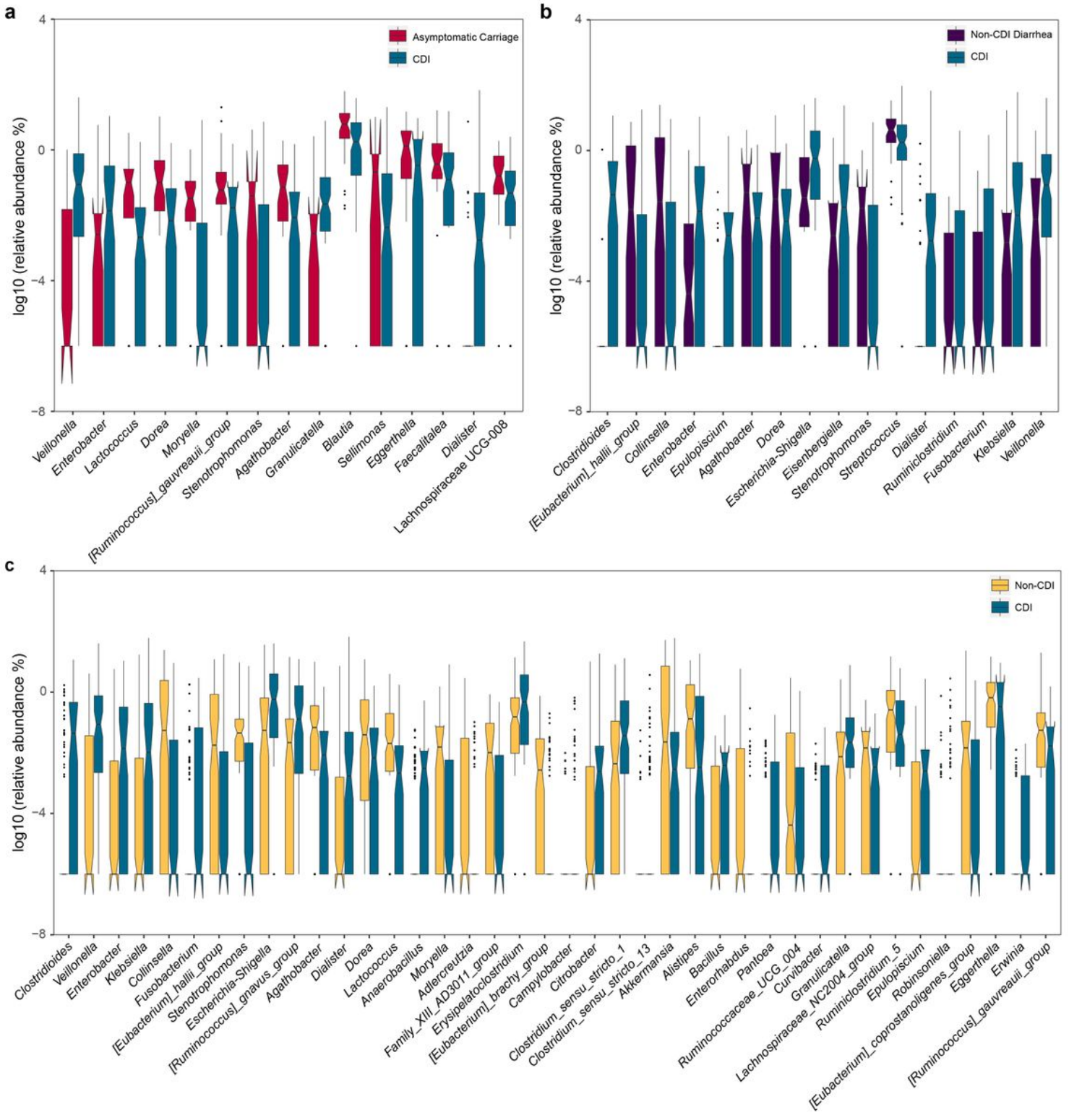

Figure 2

Relative abundances of differentially abundant genera identified by ANCOM in comparing different groups. (a) CDI vs. Asymptomatic Carriage. (b) CDI vs. Non-CDI Diarrhea. (c) CDI vs. Non-CDI. The top differentially abundant taxa were ranked based on their W statistics (from left to right). The relative abundance (%) are plotted on log10 scale. The notches in the boxplots show the 95% confidence interval around the median.

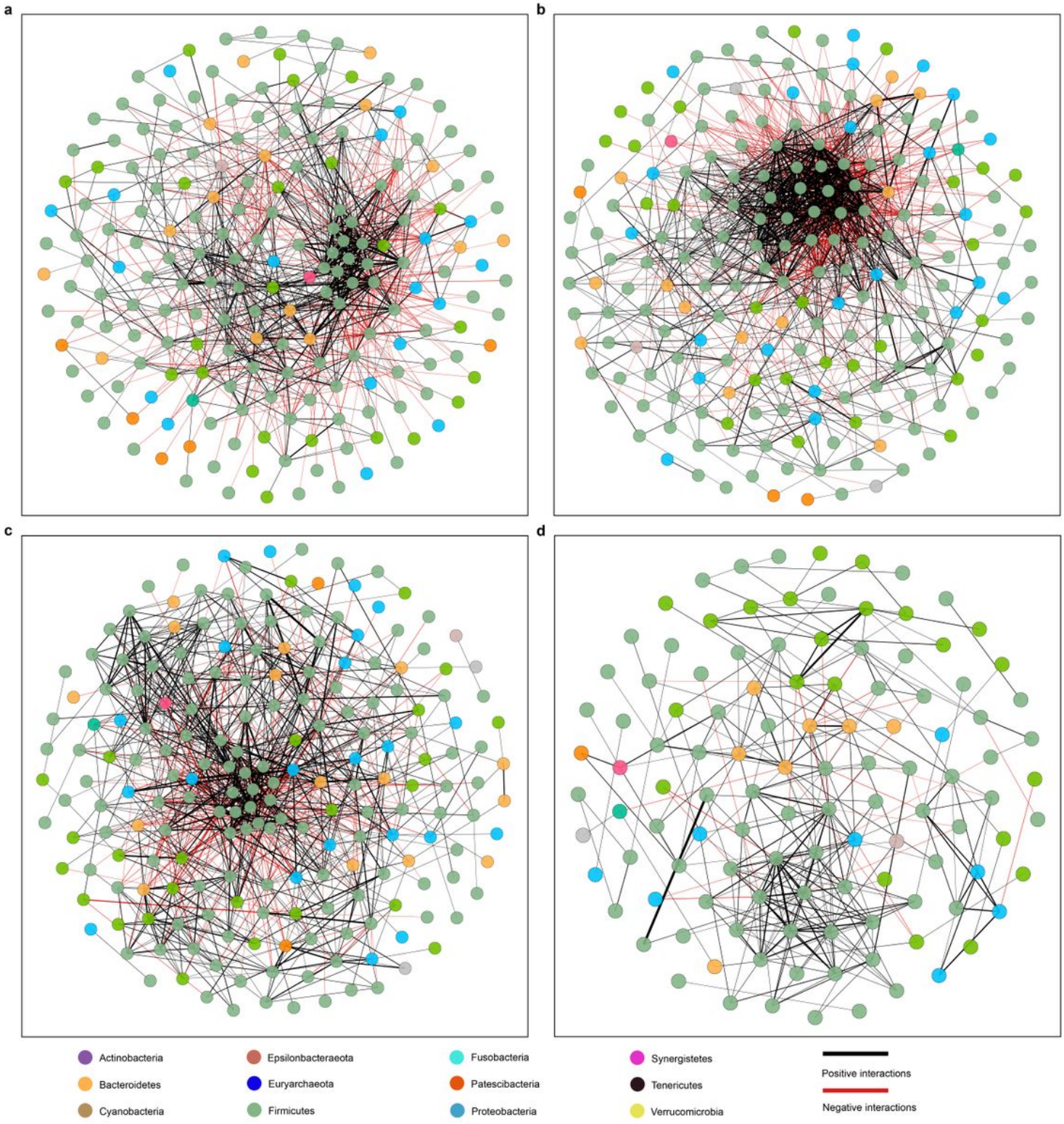

Figure 3

Microbial correlation networks of different groups. (a) Control. (b) Non-CDI Diarrhea. (c) Asymptomatic Carriage. (d) CDI. Nodes represent genera and are colored based on their phylum. Edges represent microbial correlations: green/red means positive/negative correlations, respectively. Edge thickness indicates correlation strength, and only the high-confidence interactions (p-value < 0.05) with high absolute correlation coefficients (> 0.3) were presented. For each group, we further identified the top-three

most connected genera/nodes. They are Ruminococcus_1, Roseburia and Lachnospiraceae_UCG-008 for the Control group, [Ruminococcus]_torques_group, [Eubacterium]_hallii_group and Blautia for the Non-CDI Diarrhea group, Ruminiclostridium_5, Enterococcus and Lachnospiraceae_UCG_008 for the Asymptomatic Carriage group, and Alistipes, Ruminiclostridium_5 and Lachnoclostridium for the CDI group.

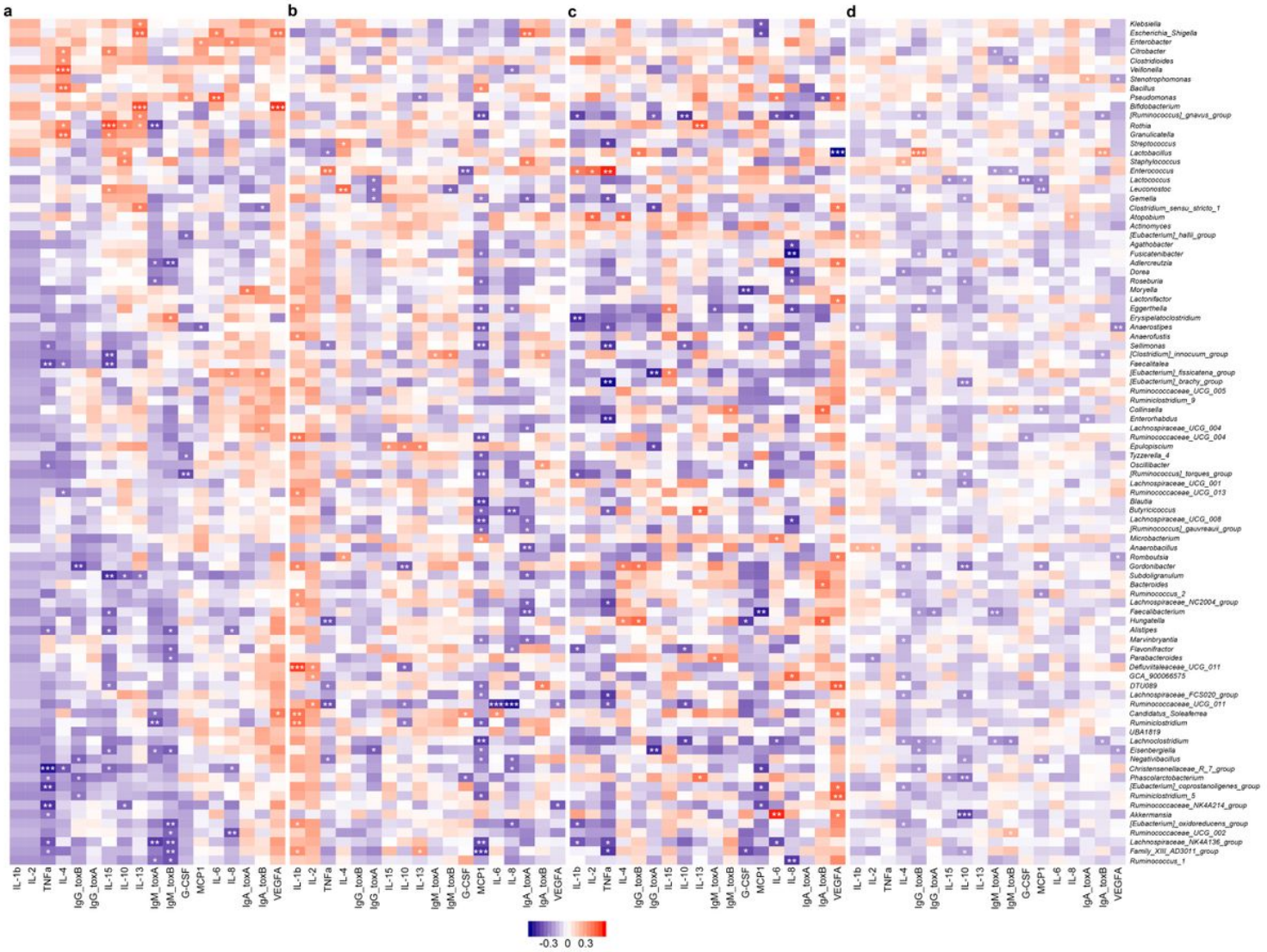

Figure 4

Correlations between gut microbial abundances and host immune markers in different groups, quantified by Spearman correlation with Benjamini-Hochberg correction. (a) Control. (b) Non-CDI Diarrhea. (c) Asymptomatic Carriage. (d) CDI. Rows represent genera; columns represent immune markers. The layout of the heatmap is followed the hierarchical clustering results of Control cohort (see Fig.S4). Red/blue represents positive/negative correlation, respectively. The intensity of the colors denotes the strength of the correlation. *□ < 0.05, **□ < 0.01, ***□ < 0.001.

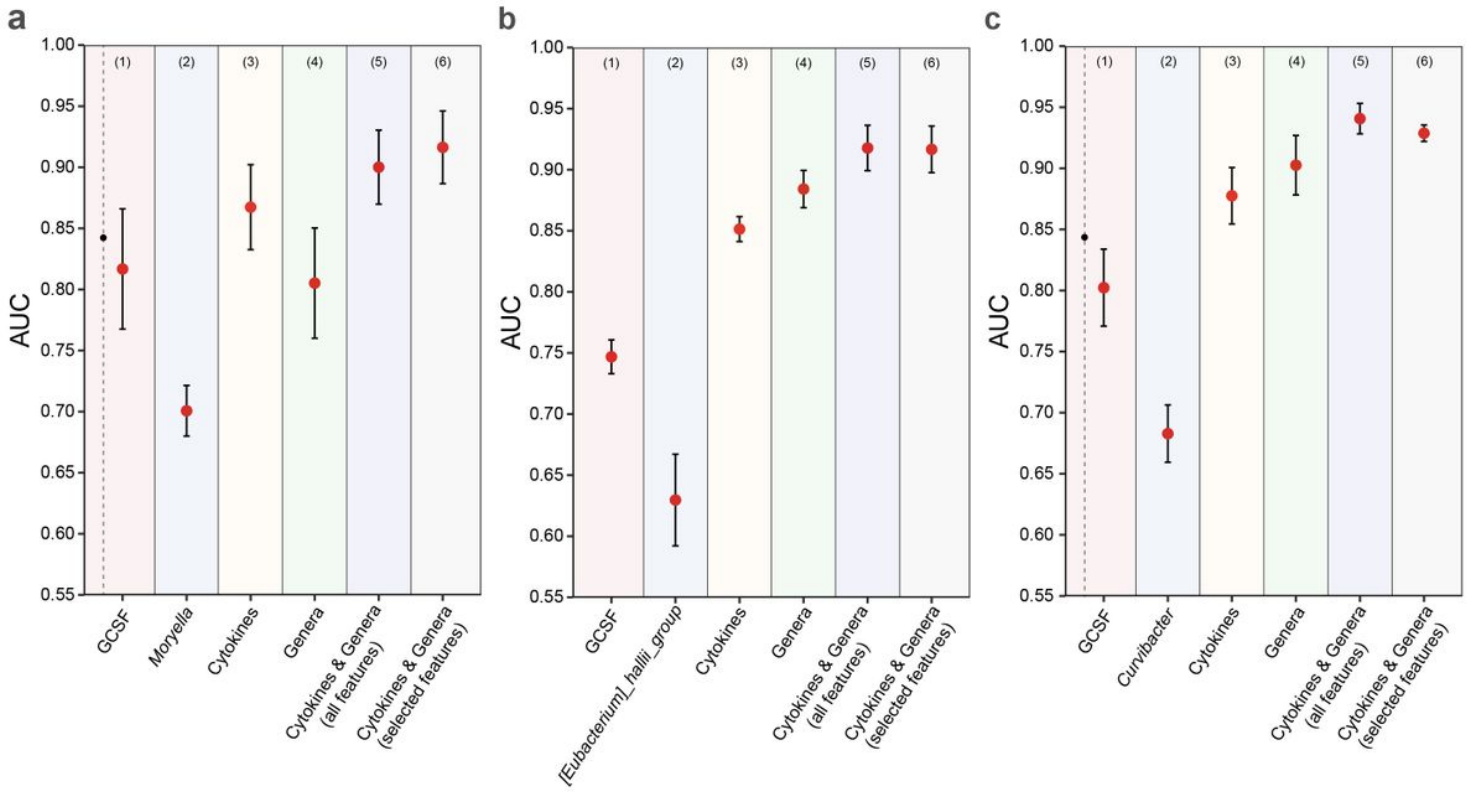

Figure 5

The performance of RF-based classification models based on various types of features in differentiating CDI from other groups. (a) CDI vs. Asymptomatic Carriage. (b) CDI vs. Non-CDI Diarrhea. (c) CDI vs. Non-CDI. For each classification task, we used different types of features: (1) the top-1 immune marker feature (based on mean decrease accuracy); (2) the top-1 genus feature; (3) all immune markers; (4) all genera; (5) integration of all immune markers and genera; (6) selected features from the set of all immune markers and genera. Error bars represent the standard errors of the means (SEM).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI08252020.docx