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

Shanlin Ke, Nira R. Pollock, Xu-Wen Wang, Xinhua Chen, Kaitlyn Daugherty, Qianyun Lin, Hua Xu, Kevin W. Garey, Anne J. Gonzales-Luna, Ciarán P. Kelly\*, Yang-Yu Liu\*

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

**Supplementary Materials**

Fig. S1. The “driver” taxa responsible for the change of microbial correlations between CDI and Asymptomatic Carriage

Fig. S2. The “driver” taxa responsible for the change of microbial correlations between CDI and Non-CDI Diarrhea.

Fig. S3. The “driver” taxa responsible for the change of microbial correlations between CDI and Non-CDI.

Fig. S4. Significant correlations between gut microbial abundances and host immune markers in the Control group.

Fig. S5. Gut microbiota and host immune markers can accurately differentiate different groups in multi-class classification models.

Fig. S6. Using the mean decrease accuracy (MDA) ranking and the 1-SE rule to select features to distinguish CDI from other groups.

Fig. S7. The fitness evolution during the genetic programming.

Table S1. Sample sizes of different data types in different groups.

Table S2. Permutational multivariate analysis of variance (PERMANOVA) in microbial compositions and immune markers.

Table S3. Differentially abundant genera between CDI and Asymptomatic Carriage groups detected by ANCOM, adjusted for age and sex.

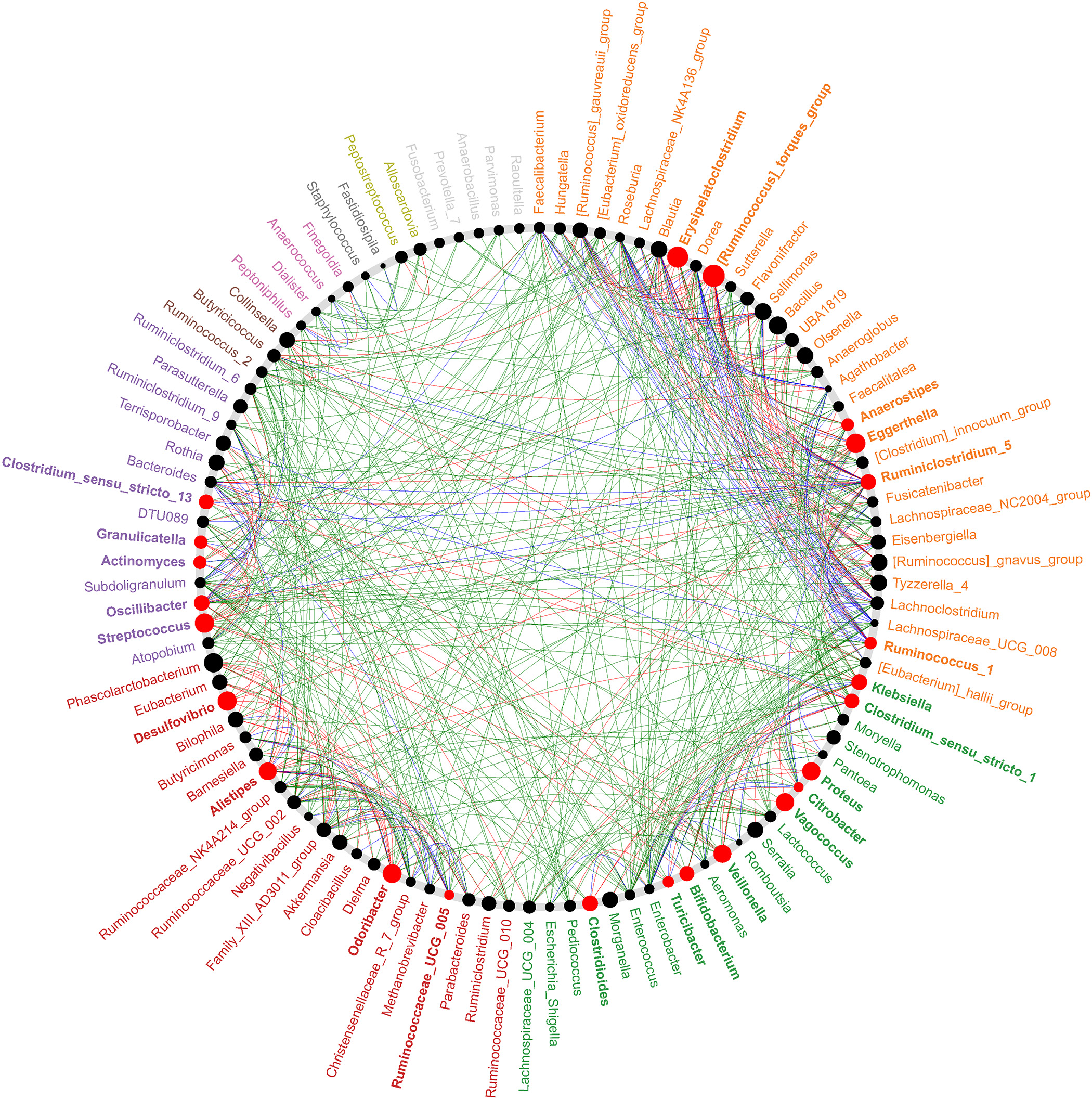
Table S4. Differentially abundant genera between CDI and Non-CDI Diarrhea groups detected by ANCOM, adjusted for age and sex.

Table S5. Differentially abundant genera between CDI and Non-CDI groups detected by ANCOM, adjusted for age and sex.

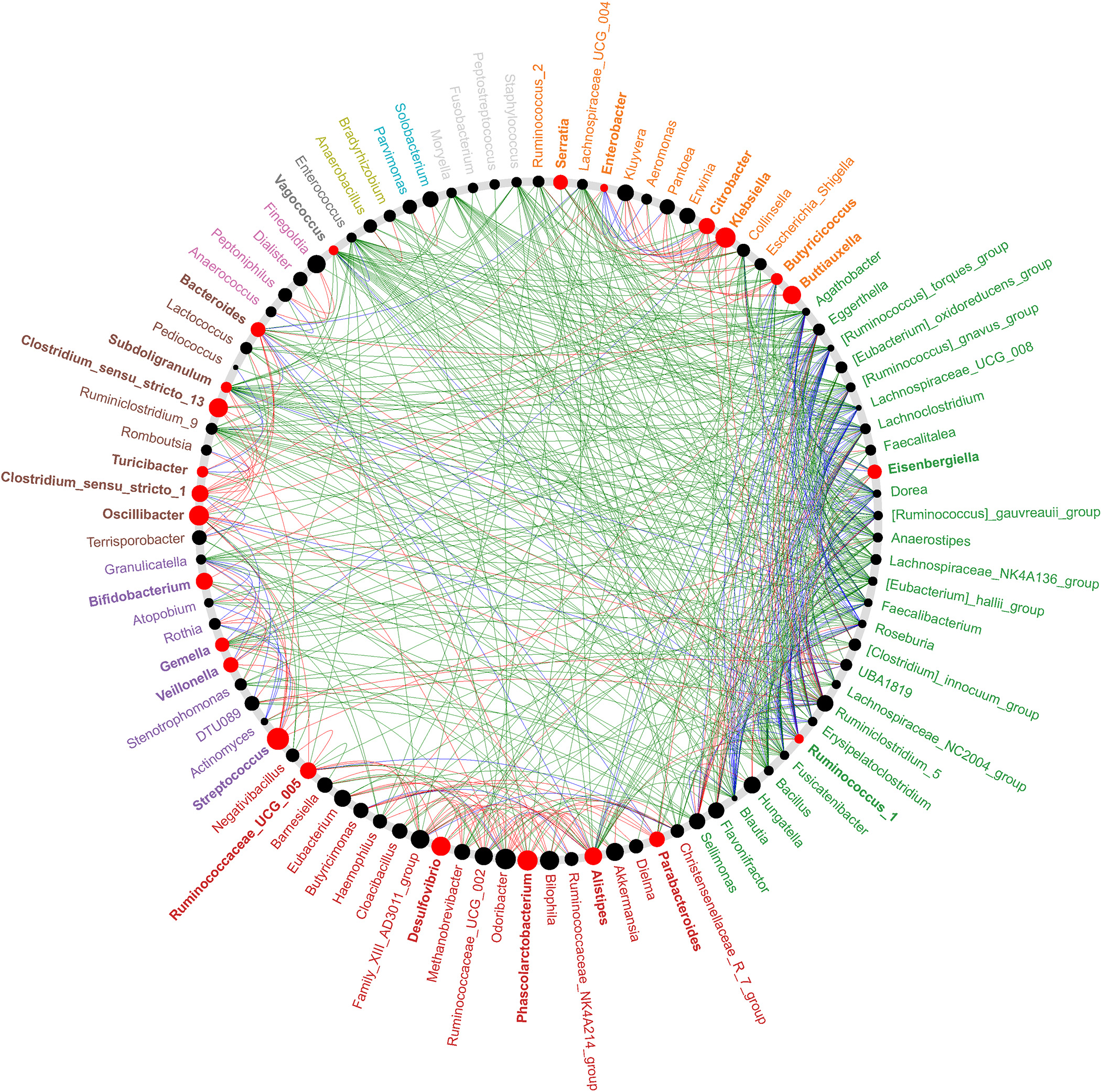
Table S6. Characteristics of microbial correlation networks associated with different groups.

Table S7. Comparison of host immune markers in different groups.

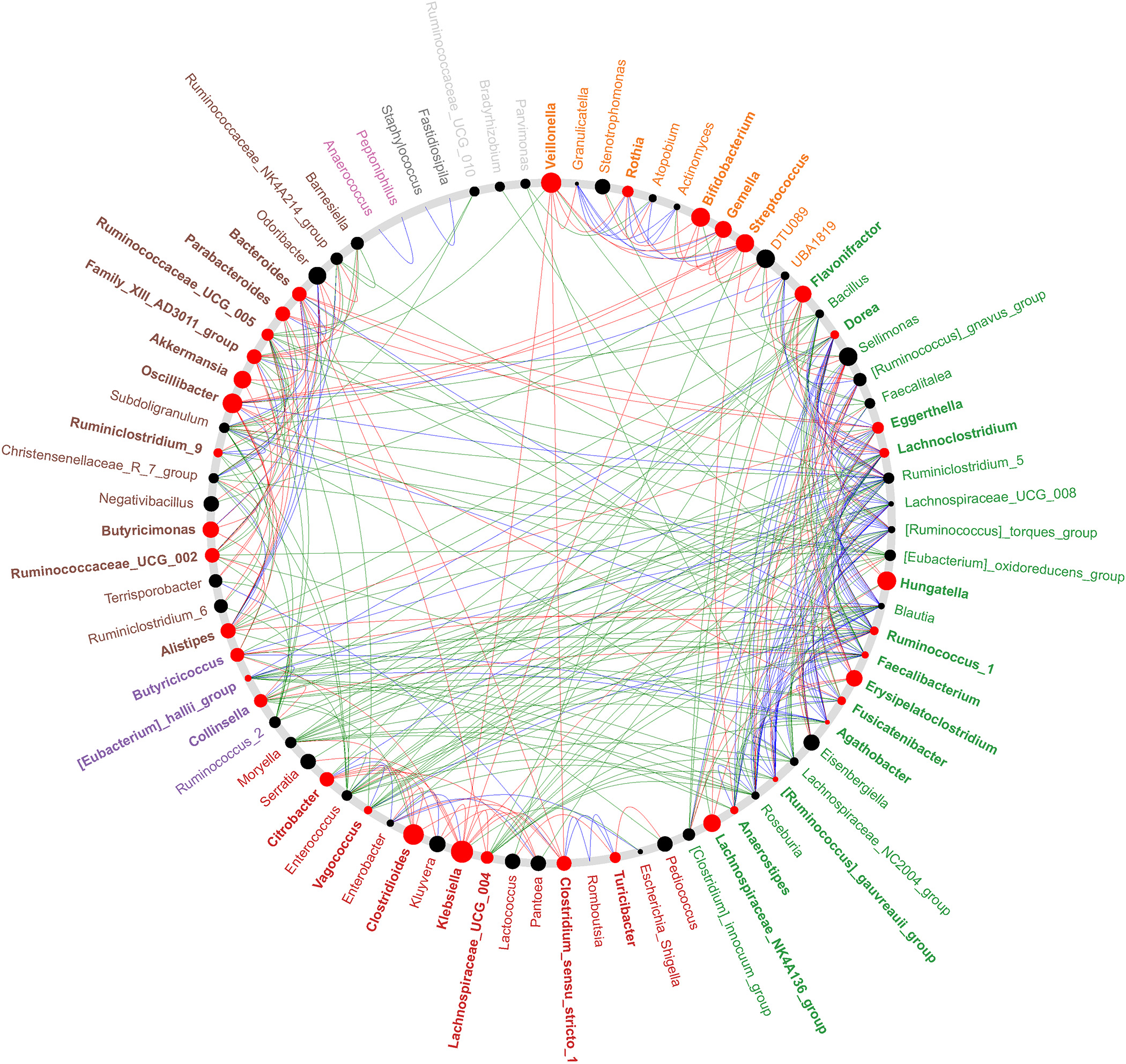
Table S8. Accuracy, Precision, Recall and F1-score of symbolic classification in CDI diagnosis.

****

**Fig. S1.** **The “driver” taxa responsible for the change of microbial correlations** **between** **CDI and Asymptomatic Carriage.** Node sizes are proportional to their scaled neighbor shift (NESH) score (i.e., a score identifying important microbial taxa of microbial association networks). A node is colored red if it is a “driver” node/taxon: its betweenness increases when comparing microbial correlation networks of CDI with that of Asymptomatic Carriage. All taxa belonging to same community (common sub-network) are randomly assigned a color to their labels. Red (or green) edges represent microbial correlations that are only present in the CDI (or Asymptomatic Carriage) network, respectively. Blue edges present common microbial correlations that are present in both networks.

****

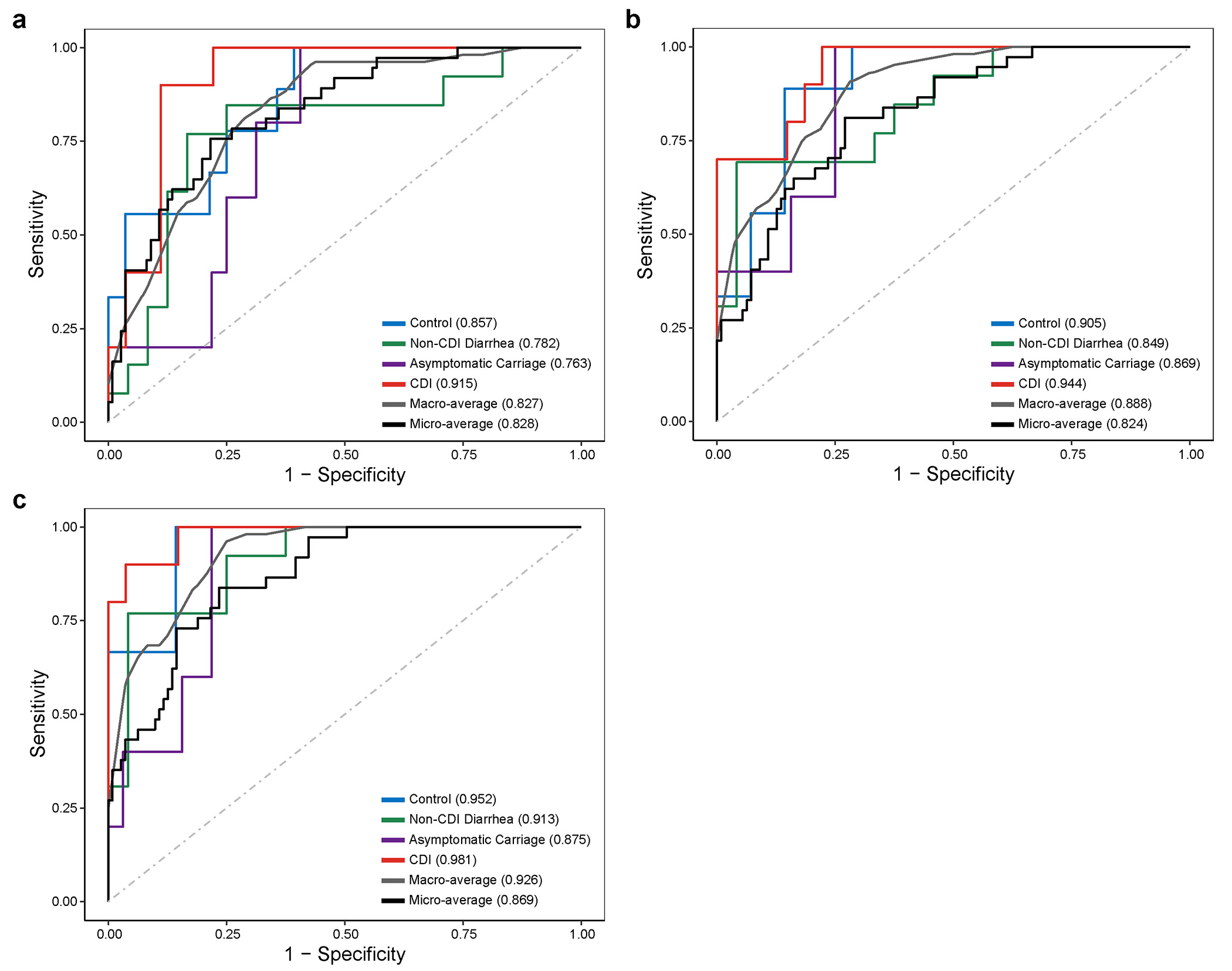
**Fig. S2. The “driver” taxa responsible for the change of microbial correlations between** **CDI and Non-CDI Diarrhea.** Node sizes are proportional to their scaled NESH score. A node is colored red if it is a “driver” node/taxon: its betweenness increases when comparing microbial correlation networks of CDI with that of Non-CDI Diarrhea. All taxa belonging to same community (common sub-network) are randomly assigned a color to their labels. Red (or green) edges represent microbial correlations that are only present in the CDI (or Non-CDI Diarrhea) network, respectively. Blue edges present common microbial correlations that are present in both networks.

****

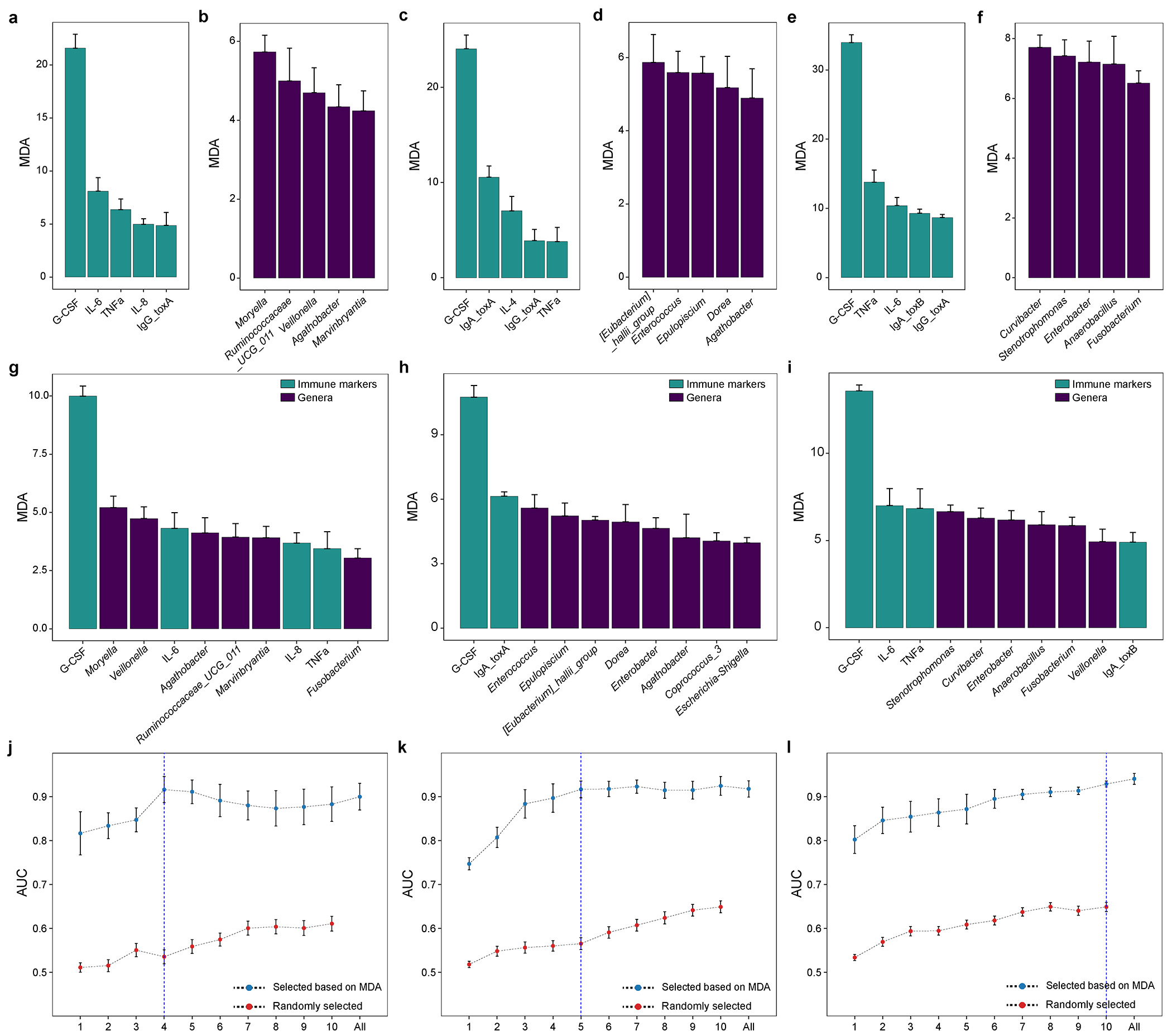
**Fig. S3. The potential “driver taxa” responsible for the change of microbial correlations between** **CDI and Non-CDI.** Node sizes are proportional to their scaled NESH score. A node is colored red if it is a “driver” node/taxon: its betweenness increases when comparing microbial correlation networks of CDI with that of Non-CDI. All taxa belonging to same community (common sub-network) are randomly assigned a color to their labels. Red (or green) edges represent microbial correlations that are only present in the CDI (or Non-CDI) network, respectively. Blue edges present common microbial correlations that are present in both networks.

****

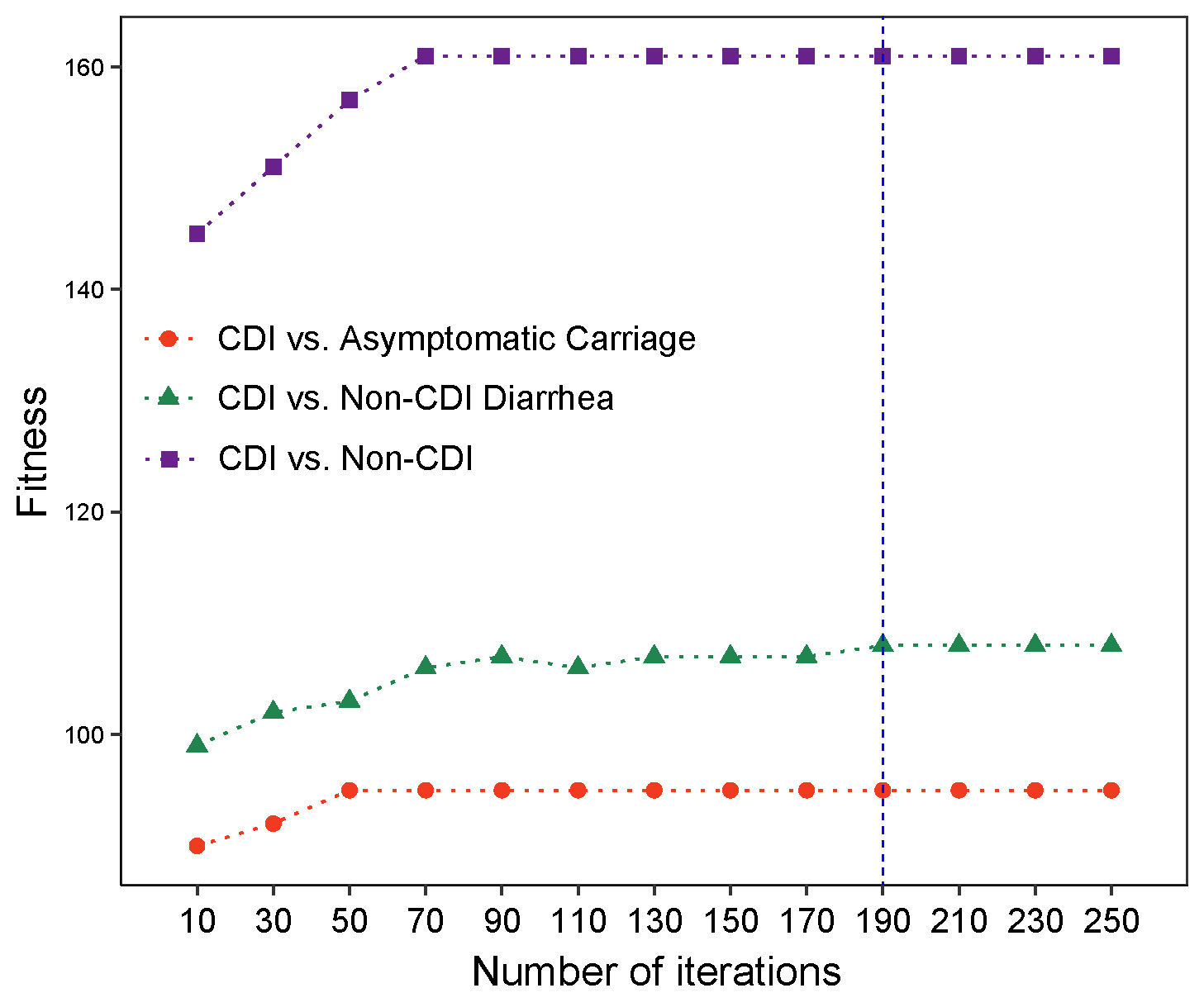
**Fig. S4. Significant correlations between gut microbial abundances and host immune markers in the Control group.** Gut microbial compositions and host immune markerswere clustered through hierarchical clustering. Rows correspond to bacterial taxa at genus level; columns correspond to host immune markers. Red/blue represents positive/negative association, respectively. The intensity of the colors denotes the strength of correlation between the genus abundance and the immunological expression level.

****

**Fig. S5. Gut microbiota and host immune markers can accurately differentiate different groups in multi-class classification models.** **(a)** Use host immune markers alone. **(b)** Use gut microbiota data (at genus level) alone. **(c)** The integration of host immune markers and microbial data. The performance of each classifier is measured by the macro-average and micro-average AUCs.

****

**Fig. S6. Using the mean decrease accuracy (MDA) ranking and the 1-SE rule to select features to distinguish CDI from other groups.** The most important features of cytokine data, microbiome data, and the integration of cytokines and microbiome data in classifying CDI vs. Asymptomatic Carriage (**a, b and g**), CDI vs. Non-CDI Diarrhea (**c, d and h**) and CDI vs. Non-CDI (**e, f and i**). The performance of classifiers using different sets of integrated features: selected based on MDA or randomly selected in CDI vs Asymptomatic Carriage (**j)**, CDI vs Non-CDI Diarrhea (**k)** and CDI vs Non-CDI (**l)**. The minimum set of features selected based on the MDA ranking and the 1-SE rule is highlighted by a vertical blue dashed line. Error bars represent the standard errors of the means (SEM).

****

**Fig. S7.** **The fitness evolution during the symbolic classification based on genetic programming.** The fitness function is a maximization function, and the tree with highest fitness score in each iteration were plotted. The final selected number of generations is highlighted with a vertical blue dashed line.

**Table S1. Sample sizes of different data types in different groups.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **NAAT negative** | | **NAAT positive** | |  |
| **Characteristics** | **Control** | **Non-CDI Diarrhea** | **Asymptomatic Carriage** | **CDI** | **Total** |
| Immunological data | 45 | 44 | 35 | 99 | 223 |
| Microbial data | 41 | 42 | 33 | 91 | 207 |
| Immunological & microbial data | 39 | 42 | 28 | 78 | 187 |

**Table S2. Permutational multivariate analysis of variance (PERMANOVA) in microbial compositions and immune markers.** CDI statuses: Control, Non-CDI Diarrhea, Asymptomatic Carriage, and CDI. Race: White, Native American, Asian, African American, Pacific Islander and mixed origin. Ethnicity: Hispanic and Not Hispanic. Here F represents the F-statistic: a larger F value indicate that the between-group variation is greater than within-group variation. R2 represents the variation explained by the model. P represents the *P-*value calculated from permutation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Microbiome** | | | **Cytokines** | | |
| Test factors | F | R2 | P | F | R2 | P |
| CDI status | 2.285 | 0.0388 | 0.001 | 3.351 | 0.052 | 0.016 |
| Age | 1.605 | 0.009 | 0.081 | 0.541 | 0.003 | 0.516 |
| Sex | 1.557 | 0.009 | 0.095 | 0.916 | 0.005 | 0.372 |
| Race | 0.881 | 0.031 | 0.832 | 1.595 | 0.050 | 0.153 |
| Ethnicity | 0.476 | 0.003 | 0.961 | 0.206 | 0.001 | 0.771 |

**Table S3. Differentially abundant genera between CDI and Asymptomatic Carriage groups detected by ANCOM, adjusted for age and sex.** For each genus, the first column represents its W statistic, and subsequent four columns represent logical indicators of whether it is differentially abundant under a series of cutoffs (0.9, 0.8, 0.7 and 0.6). The last two columns represent its relative abundance (mean ± standard deviation) in the two groups.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genera** | **W\_stat** | **Cutoff**  **0.9** | **Cutoff**  **0.8** | **Cutoff**  **0.7** | **Cutoff**  **0.6** | **Relative abundance (%) in CDI** | **Relative abundance (%) in Asymptomatic Carriage** |
| ***Veillonella*** | 204 | TRUE | TRUE | TRUE | TRUE | 1.86 ± 5.55 | 0.06 ± 0.20 |
| ***Enterobacter*** | 182 | FALSE | TRUE | TRUE | TRUE | 0.79 ± 1.81 | 0.20 ± 1.01 |
| ***Lactococcus*** | 179 | FALSE | TRUE | TRUE | TRUE | 0.10 ± 0.30 | 0.36 ± 0.69 |
| ***Dorea*** | 177 | FALSE | TRUE | TRUE | TRUE | 0.10 ± 0.26 | 0.79 ± 2.00 |
| ***Moryella*** | 174 | FALSE | TRUE | TRUE | TRUE | 0.20 ± 1.19 | 0.13 ± 0.24 |
| ***[Ruminococcus]\_gauvreauii\_group*** | 173 | FALSE | TRUE | TRUE | TRUE | 0.09 ± 0.26 | 1.10± 3.68 |
| ***Stenotrophomonas*** | 167 | FALSE | TRUE | TRUE | TRUE | 0.13 ± 0.79 | 0.28 ± 0.76 |
| ***Agathobacter*** | 158 | FALSE | FALSE | TRUE | TRUE | 0.07 ± 0.19 | 0.25 ± 0.42 |
| ***Granulicatella*** | 157 | FALSE | FALSE | TRUE | TRUE | 0.31 ± 1.10 | 0.10 ± 0.46 |
| ***Blautia*** | 154 | FALSE | FALSE | TRUE | TRUE | 5.30 ± 7.99 | 10.18 ± 14.05 |
| ***Sellimonas*** | 150 | FALSE | FALSE | TRUE | TRUE | 0.46 ± 2.18 | 1.20 ± 2.50 |
| ***Eggerthella*** | 147 | FALSE | FALSE | TRUE | TRUE | 1.39 ± 2.16 | 2.98 ± 4.06 |
| ***Faecalitalea*** | 145 | FALSE | FALSE | TRUE | TRUE | 0.91 ± 2.07 | 1.41 ± 3.05 |
| ***Dialister*** | 141 | FALSE | FALSE | FALSE | TRUE | 1.10 ± 7.07 | 0.23 ± 1.27 |
| ***Lachnospiraceae\_UCG\_008*** | 135 | FALSE | FALSE | FALSE | TRUE | 0.20 ± 0.39 | 0.37 ± 0.45 |

**Table S4. Differentially abundant genera between CDI and Non-CDI Diarrhea groups detected by ANCOM, adjusted for age and sex.** For each genus, the first column represents its W statistic, and subsequent four columns represent logical indicators of whether it is differentially abundant under a series of cutoffs (0.9, 0.8, 0.7 and 0.6). The last two columns represent its relative abundance (mean ± standard deviation) in the two groups.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genera** | **W\_stat** | **detected\_0.9** | **detected\_0.8** | **detected\_0.7** | **detected\_0.6** | **Relative abundance (%) in CDI** | **Relative abundance (%) in Non-CDI Diarrhea** |
| ***Clostridioides*** | 206 | TRUE | TRUE | TRUE | TRUE | 0.67 ± 1.80 | 0.03 ± 0.16 |
| ***[Eubacterium]\_hallii\_group*** | 199 | TRUE | TRUE | TRUE | TRUE | 0.40 ± 2.12 | 1.14 ± 2.05 |
| ***Collinsella*** | 195 | TRUE | TRUE | TRUE | TRUE | 0.57 ± 1.59 | 2.57 ± 5.33 |
| ***Enterobacter*** | 189 | TRUE | TRUE | TRUE | TRUE | 0.79 ± 1.81 | 0.05 ± 0.20 |
| ***Epulopiscium*** | 166 | FALSE | TRUE | TRUE | TRUE | 0.06 ± 0.30 | 0.00 ± 0.01 |
| ***Agathobacter*** | 165 | FALSE | TRUE | TRUE | TRUE | 0.07 ± 0.19 | 0.42 ± 0.93 |
| ***Dorea*** | 165 | FALSE | TRUE | TRUE | TRUE | 0.10 ± 0.26 | 0.96 ± 2.24 |
| ***Escherichia\_Shigella*** | 163 | FALSE | FALSE | TRUE | TRUE | 3.54 ± 6.46 | 1.84 ± 5.01 |
| ***Eisenbergiella*** | 149 | FALSE | FALSE | TRUE | TRUE | 1.03 ± 3.36 | 0.10 ± 0.40 |
| ***Stenotrophomonas*** | 147 | FALSE | FALSE | TRUE | TRUE | 0.13 ± 0.79 | 0.09 ± 0.17 |
| ***Streptococcus*** | 147 | FALSE | FALSE | TRUE | TRUE | 6.16 ± 13.27 | 7.00 ± 7.39 |
| ***Dialister*** | 138 | FALSE | FALSE | FALSE | TRUE | 1.10 ± 7.07 | 0.06 ± 0.25 |
| ***Ruminiclostridium*** | 137 | FALSE | FALSE | FALSE | TRUE | 0.08 ± 0.44 | 0.00 ± 0.01 |
| ***Fusobacterium*** | 131 | FALSE | FALSE | FALSE | TRUE | 0.18 ± 0.51 | 0.01 ± 0.04 |
| ***Klebsiella*** | 131 | FALSE | FALSE | FALSE | TRUE | 1.75 ± 6.94 | 0.58 ± 2.74 |
| ***Veillonella*** | 125 | FALSE | FALSE | FALSE | TRUE | 1.86 ± 5.55 | 0.27 ± 0.78 |

**Table S5. Differentially abundant genera between CDI and Non-CDI groups detected by ANCOM, adjusted for age and sex.** For each genus, the first column represents its W statistic, and subsequent four columns represent logical indicators of whether it is differentially abundant under a series of cutoffs (0.9, 0.8, 0.7 and 0.6). The last two columns represent its relative abundance (mean ± standard deviation) in the two groups.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genera** | **W\_stat** | **detected\_0.9** | **detected\_0.8** | **detected\_0.7** | **detected\_0.6** | **Relative abundance (%) in CDI** | **Relative abundance (%) in Non-CDI** |
| ***Clostridioides*** | 201 | TRUE | TRUE | TRUE | TRUE | 0.67 ± 1.81 | 0.08 ± 0.26 |
| ***Veillonella*** | 201 | TRUE | TRUE | TRUE | TRUE | 1.86 ± 5.58 | 0.14 ± 0.50 |
| ***Enterobacter*** | 200 | TRUE | TRUE | TRUE | TRUE | 0.79 ± 1.82 | 0.08 ± 0.55 |
| ***Klebsiella*** | 196 | TRUE | TRUE | TRUE | TRUE | 1.75 ± 6.98 | 0.49 ± 2.25 |
| ***Collinsella*** | 194 | TRUE | TRUE | TRUE | TRUE | 0.57 ± 1.60 | 2.29 ± 4.64 |
| ***Fusobacterium*** | 194 | TRUE | TRUE | TRUE | TRUE | 0.18 ± 0.51 | 0.04 ± 0.25 |
| ***[Eubacterium]\_hallii\_group*** | 193 | TRUE | TRUE | TRUE | TRUE | 0.40 ± 2.13 | 1.21 ± 2.55 |
| ***Stenotrophomonas*** | 193 | TRUE | TRUE | TRUE | TRUE | 0.13 ± 0.79 | 0.30 ± 1.06 |
| ***Escherichia\_Shigella*** | 191 | TRUE | TRUE | TRUE | TRUE | 3.54 ± 6.50 | 1.96 ± 5.31 |
| ***[Ruminococcus]\_gnavus\_group*** | 185 | FALSE | TRUE | TRUE | TRUE | 1.73 ± 3.15 | 0.50 ± 1.81 |
| ***Agathobacter*** | 179 | FALSE | TRUE | TRUE | TRUE | 0.07 ± 0.19 | 0.50 ± 1.37 |
| ***Dialister*** | 176 | FALSE | TRUE | TRUE | TRUE | 1.10 ± 7.11 | 0.10 ± 0.69 |
| ***Dorea*** | 170 | FALSE | TRUE | TRUE | TRUE | 0.10 ± 0.26 | 0.86 ± 2.17 |
| ***Lactococcus*** | 169 | FALSE | TRUE | TRUE | TRUE | 0.10 ± 0.30 | 0.23 ± 0.56 |
| ***Anaerobacillus*** | 164 | FALSE | FALSE | TRUE | TRUE | 0.02 ± 0.06 | 0.00 ± 0.01 |
| ***Moryella*** | 164 | FALSE | FALSE | TRUE | TRUE | 0.20 ± 1.19 | 0.11 ± 0.26 |
| ***Adlercreutzia*** | 162 | FALSE | FALSE | TRUE | TRUE | 0.00 ± 0.02 | 0.08 ± 0.33 |
| ***Family\_XIII\_AD3011\_group*** | 161 | FALSE | FALSE | TRUE | TRUE | 0.03 ± 0.07 | 0.09 ± 0.16 |
| ***Erysipelatoclostridium*** | 160 | FALSE | FALSE | TRUE | TRUE | 3.56 ± 7.56 | 0.80 ± 1.79 |
| ***[Eubacterium]\_brachy\_group*** | 154 | FALSE | FALSE | TRUE | TRUE | 0.01 ± 0.03 | 0.04 ± 0.10 |
| ***Campylobacter*** | 154 | FALSE | FALSE | TRUE | TRUE | 0.03 ± 0.11 | 0.00 ± 0.00 |
| ***Citrobacter*** | 154 | FALSE | FALSE | TRUE | TRUE | 0.41 ± 2.11 | 0.20 ± 1.09 |
| ***Clostridium\_sensu\_stricto\_1*** | 151 | FALSE | FALSE | TRUE | TRUE | 0.72 ± 1.76 | 0.38 ± 1.19 |
| ***Clostridium\_sensu\_stricto\_13*** | 151 | FALSE | FALSE | TRUE | TRUE | 0.07 ± 0.41 | 0.00 ± 0.01 |
| ***Akkermansia*** | 149 | FALSE | FALSE | TRUE | TRUE | 3.3 ± 9.68 | 6.41 ± 12.03 |
| ***Alistipes*** | 142 | FALSE | FALSE | FALSE | TRUE | 1.28 ± 3.05 | 1.30 ± 2.34 |
| ***Bacillus*** | 139 | FALSE | FALSE | FALSE | TRUE | 0.01 ± 0.03 | 0.00 ± 0.01 |
| ***Enterorhabdus*** | 137 | FALSE | FALSE | FALSE | TRUE | 0.01 ± 0.03 | 0.08 ± 0.55 |
| ***Pantoea*** | 137 | FALSE | FALSE | FALSE | TRUE | 0.01 ± 0.03 | 0.00 ± 0.00 |
| ***Ruminococcaceae\_UCG\_004*** | 134 | FALSE | FALSE | FALSE | TRUE | 0.05 ± 0.17 | 0.15 ± 0.41 |
| ***Curvibacter*** | 132 | FALSE | FALSE | FALSE | TRUE | 0.01 ± 0.01 | 0.00 ± 0.00 |
| ***Granulicatella*** | 131 | FALSE | FALSE | FALSE | TRUE | 0.31 ± 1.11 | 0.08 ± 0.27 |
| ***Lachnospiraceae\_NC2004\_group*** | 131 | FALSE | FALSE | FALSE | TRUE | 0.02 ± 0.04 | 0.05 ± 0.09 |
| ***Ruminiclostridium\_5*** | 131 | FALSE | FALSE | FALSE | TRUE | 0.50 ± 1.11 | 1.33 ± 2.79 |
| ***Epulopiscium*** | 130 | FALSE | FALSE | FALSE | TRUE | 0.06 ± 0.30 | 0.01 ± 0.04 |
| ***Robinsoniella*** | 130 | FALSE | FALSE | FALSE | TRUE | 0.06 ± 0.33 | 0.01 ± 0.07 |
| ***[Eubacterium]\_coprostanoligenes\_group*** | 128 | FALSE | FALSE | FALSE | TRUE | 0.13 ± 0.39 | 0.39 ± 2.19 |
| ***Eggerthella*** | 128 | FALSE | FALSE | FALSE | TRUE | 1.39 ± 2.17 | 1.86 ± 2.96 |
| ***Erwinia*** | 126 | FALSE | FALSE | FALSE | TRUE | 0.00 ± 0.00 | 0.00 ± 0.00 |
| ***[Ruminococcus]\_gauvreauii\_group*** | 124 | FALSE | FALSE | FALSE | TRUE | 0.09 ± 0.26 | 0.45 ± 2.04 |

**Table S6. Characteristics of microbial correlation networks associated with different groups.** In order to quantify the difference of the network structure, we calculated the number of nodes, number of edges, average degree (the average number of connections per node), graph density (measure of how close the network is to a complete graph), clustering coefficient (measure of how complete the neighborhood of a node is) and modularity (measure of how well a network decomposes into modular communities).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Groups** | **Average degree** | **Clustering coefficient** | **Edges** | **Graph density** | **Modularity** | **Nodes** |
| **Control** | 9.475 | 0.368 | 938 | 0.048 | 0.377 | 198 |
| **Non-CDI Diarrhea** | 11.314 | 0.474 | 1171 | 0.055 | 0.271 | 207 |
| **Asymptomatic Carriage** | 9.730 | 0.349 | 973 | 0.049 | 0.442 | 200 |
| **CDI** | 5.200 | 0.502 | 299 | 0.046 | 0.568 | 115 |

**Table S7.** **Comparison of host immune markers in different groups.** Mean (Q1, Q3); p-value calculated with Mann-Whitney U test.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Immune markers** | **Control (n=45)** | **Non-CDI Diarrhea (n=44)** | **Asymptomatic Carriage (n=35)** | **CDI (n=99)** | **P-value (CDI vs. Asymptomatic Carriage)** | **P-value (CDI vs. Non-CDI Diarrhea)** | **P-value (CDI vs. Non-CDI)** |
| **IgA\_toxA** | 33.63 (7.24, 62.54) | 16.94 (8.44, 16.96) | 44.59 (10.21, 102.50) | 48.41 (12.10, 104) | 0.543 | < 0.001 | 0.001 |
| **IgG\_toxA** | 19.87 (9.75, 22.43) | 24.20 (11.11, 27.77) | 22.20 (11.48, 24.86) | 40.09 (14.77, 59.18) | 0.009 | 0.002 | < 0.001 |
| **IgM\_toxA** | 2.04 (0.00, 2.87) | 2.45 (0.00, 3.36) | 2.15 (0.00, 2.98) | 2.09 (0.00, 2.84) | 0.316 | 0.643 | 0.172 |
| **IgA\_toxB** | 9.73 (2.68, 8.61) | 18.78 (4.07, 19.18) | 23.02 (5.80, 20.12) | 35.07 (4.76, 67.79) | 0.469 | 0.102 | 0.002 |
| **IgG\_toxB** | 9.66 (4.30, 9.80) | 11.97 (5.92, 13.46) | 13.57 (3.98, 15.98) | 14.61 (4.92, 17.56) | 0.980 | 0.870 | 0.379 |
| **IgM\_toxB** | 12.12 (2.28, 9.24) | 13.72 (2.78, 12.77) | 11.08 (2.74, 9.04) | 14.99 (1.87, 10.47) | 0.261 | 0.192 | 0.256 |
| **GCSF** | 11.27 (0.46, 14.17) | 49.37 (2.18, 29.36) | 20.01 (2.18, 20.49) | 386.64 (22.56, 159.95) | < 0.001 | < 0.001 | < 0.001 |
| **IL-10** | 8.29 (0.00, 3.15) | 33.22 (0.00, 14.06) | 9.05 (0.00, 9.78) | 35.17 (1.63, 27.99) | 0.002 | 0.021 | < 0.001 |
| **IL-13** | 1.38 (0.00, 0.00) | 1.97 (0.00, 0.34) | 5.22 (0.00, 0.00) | 9.35 (0.00, 1.09) | 0.529 | 0.593 | 0.167 |
| **IL-15** | 1.56 (0.00, 0.24) | 2.82 (0.00, 3.28) | 2.03 (0.00, 1.46) | 5.22 (0.19, 5.33) | 0.007 | 0.037 | < 0.001 |
| **IL-1b** | 0.05 (0.00, 0.00) | 0.11 (0.00, 0.00) | 0.44 (0.00, 0.00) | 0.7 (0.00, 0.00) | 0.920 | 0.387 | 0.159 |
| **IL-2** | 0.04 (0.00, 0.00) | 0.05 (0.00, 0.00) | 0.45 (0.00, 0.00) | 1.4 (0.00, 0.00) | 0.630 | 0.151 | 0.051 |
| **IL-4** | 1.88 (0.00, 0.00) | 9.58 (2.57, 12.44) | 5.73 (0.00, 0.00) | 11.54 (0.00, 9.12) | 0.002 | 0.011 | 0.032 |
| **IL-6** | 15.78 (0.00, 3.77) | 24.97 (0.00, 10.71) | 9.52 (0.00, 5.46) | 47.09 (2.52, 37.71) | < 0.001 | < 0.001 | < 0.001 |
| **IL-8** | 100.51 (12.37, 59.19) | 80.85 (15.22, 73.45) | 59.78 (10.33, 44.76) | 128.05 (27.20, 122.24) | < 0.001 | 0.004 | < 0.001 |
| **MCP1** | 477.00 (397.19, 545.00) | 591.61 (399.55, 779.32) | 613.28 (430.85, 791.37) | 844.96 (522.41, 990.98) | 0.053 | 0.020 | < 0.001 |
| **TNFa** | 8.61 (6.20, 10.95) | 21.76 (8.93, 21.22) | 12.45 (4.91, 14.84) | 26.68 (13.88, 28.94) | < 0.001 | 0.006 | < 0.001 |
| **VEGFA** | 102.78 (35.76, 118.27) | 109.85 (34.53, 127.69) | 118.44 (27.54, 188.60) | 125.93 (19.54, 140.49) | 0.499 | 0.603 | 0.390 |

**Table S8. Accuracy, Precision, Recall and F1-score of symbolic classification in CDI diagnosis.** CDI subjects were considered as either true positive or true negative. Results shown in the parenthesis represents the latter case. The performance of the symbolic classification model evaluated by cross-validation. We randomly split the dataset to form a training set (80% of the data) and a test set (20% of the data) in 10 different ways. Each time, for each classification task (diagnostic goal), we learned the SC model from the training dataset and evaluated it on the test dataset. Data represents as mean ± standard deviation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Diagnostic goal** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| CDI vs. Asymptomatic Carriage | 0.873 ± 0.085 | 0.878 ± 0.101 (0.857 ± 0.180) | 0.963 ± 0.043 (0.639 ± 0.229) | 0.915 ± 0.061 (0.718 ± 0.185) |
| CDI vs. Non-CDI Diarrhea | 0.883 ± 0.036 | 0.899 ± 0.078 (0.865 ±0.098) | 0.912 ± 0.070 (0.847 ± 0.110) | 0.901 ± 0.039 (0.846 ± 0.055) |
| CDI vs. Non-CDI | 0.818 ± 0.040 | 0.770 ± 0.102 (0.861 ±0.059) | 0.784 ± 0.120 (0.839 ± 0.091) | 0.769 ± 0.074 (0.845 ± 0.036) |

**Supplementary Methods**

**Microbial diversity and differential abundance analysis**

The alpha and beta diversity measures were calculated at the genus level using the vegan: Community Ecology Package in R (https://CRAN.R-project.org/package=vegan). Measures of alpha diversity included: the richness S (the number of taxa present in the community/sample), Chao 1 index , Shannon index, and evenness . Here, and are the count of singletons and doubletons, respectively, and is the relative abundance of taxon-i in the community. For beta diversity, we used the Bray-Curtis dissimilarity measure, which was also used in the Principal Coordinates Analysis (PCoA). We applied principal component analysis (PCA) on the expression levels of all immune markers based on the Euclidean distance.

Difference in microbiome compositions and immune expression levels by CDI status (i.e., different groups) and other covariates (i.e., age, sex, race and ethnicity) were tested by the permutational multivariate analysis of variance (PERMANOVA) using the “adonis” function in the vegan R package. All PERMANOVA tests were performed with the default 999 permutations based on the Bray-Curtis dissimilarity and Euclidean distance for microbial composition and immune marker data, respectively. Note that in the PERMANOVA tests, we only included subjects with known information of age, sex, race and ethnicity.

**Microbial correlation network analysis**

The microbial correlation networks were constructed using SparCC1. Significant interactions were determined by the bootstrapped results (N=100) using the script PseudoPvals in SparCC. Significant correlations with absolute sparse correlations ≥ 0.3 were visualized using Gephi (https://gephi.org/).

**Classification with Random Forests model**

In the multi-class model, a macro-average score computed the metric independently for each group and then was averaged across all levels regardless of the number of samples in each group, whereas a micro-average will aggregate the contributions of all groups to compute the average metric. In the binary classifiers based on RF models, the performance of the classifiers was evaluated by a 5-fold cross validation. In order to reduce computation complexity and feature redundancy, a feature selection procedure was performed as follows. The importance of each feature was quantified by the Mean Decrease in Accuracy (MDA) of the classifier due to the exclusion (or permutation) of this feature. The more the accuracy of the classifier decreases due to the exclusion (or permutation) of a single feature, the more important that feature is deemed for classification of the data. We first ranked all the features based on their mean MDA. Then we followed the “1-SE strategy” to select the minimum set of top features whose mean AUC is within one standard error of the mean AUC from the model with all of the features.

**Symbolic classification and logistic regression**

Genetic programming (GP) is a genetic algorithm that searches the space of mathematical equations without any constraints on their forms2. GP involves reproduction, random mutation, crossover, a fitness function, and multiple generations of evolution of a population of computer programs to resolve a given task. GP is commonly used to investigate a functional relationship (i.e., a mathematical formula) between features in data (symbolic regression: SR) or to group data into categories (symbolic classification: SC). We performed a random data-split to create a training set (80% of the data) and a held-out test set (20% of the data) for ten times, which were used to evaluate the SC performance. The Karoo GP was used with the following settings: (1) the fitness function (Kernel) is c (representing “classification”); (2) the type of tree is r (ramped half/half); (3) the maximum tree depth for the initial population is 6; (4) the number of trees per generation is 100; (5) the maximum number of generations is 190 (based on the converging results shown in Fig. S7); (6) constants include 0.1, 0.2, 0.3, 0.4 and 0.5; and (7) all other parameters are set as default values. The fitness function in SC is a maximization function, which will seek the highest fitness score among the trees in each generation. The sign of the final formula will be used for CDI diagnosis: the class of subject is CDI if ; or Asymptomatic Carriage (or Non-CDI Diarrhea, Non-CDI) if . The LR models were constructed using the glm() function in R. The class of subject is CDI if ; or Asymptomatic Carriage (or Non-CDI Diarrhea, Non-CDI) if .

**References**

1. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol 2012;8:e1002687.

2. Biesheuvel CJ, Siccama I, Grobbee DE, et al. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. J Clin Epidemiol 2004;57:551-60.