A

sample size	t-test statistic	p-value
50	2.54	0.032
100	2.48	0.035
150	2.26	0.05
200	1.35	0.21
250	1.86	0.095
300	4.42	0.0017
350	6.51	0.00011

B

mainly detected by regression methods	mainly detected by mutual information, Kullback-Leibler divergence
Prevotellaceae Bifidobacterium longum	Lactobacillales Streptococcaceae Streptococcus Bilophila Oscillospira Paraprevotella

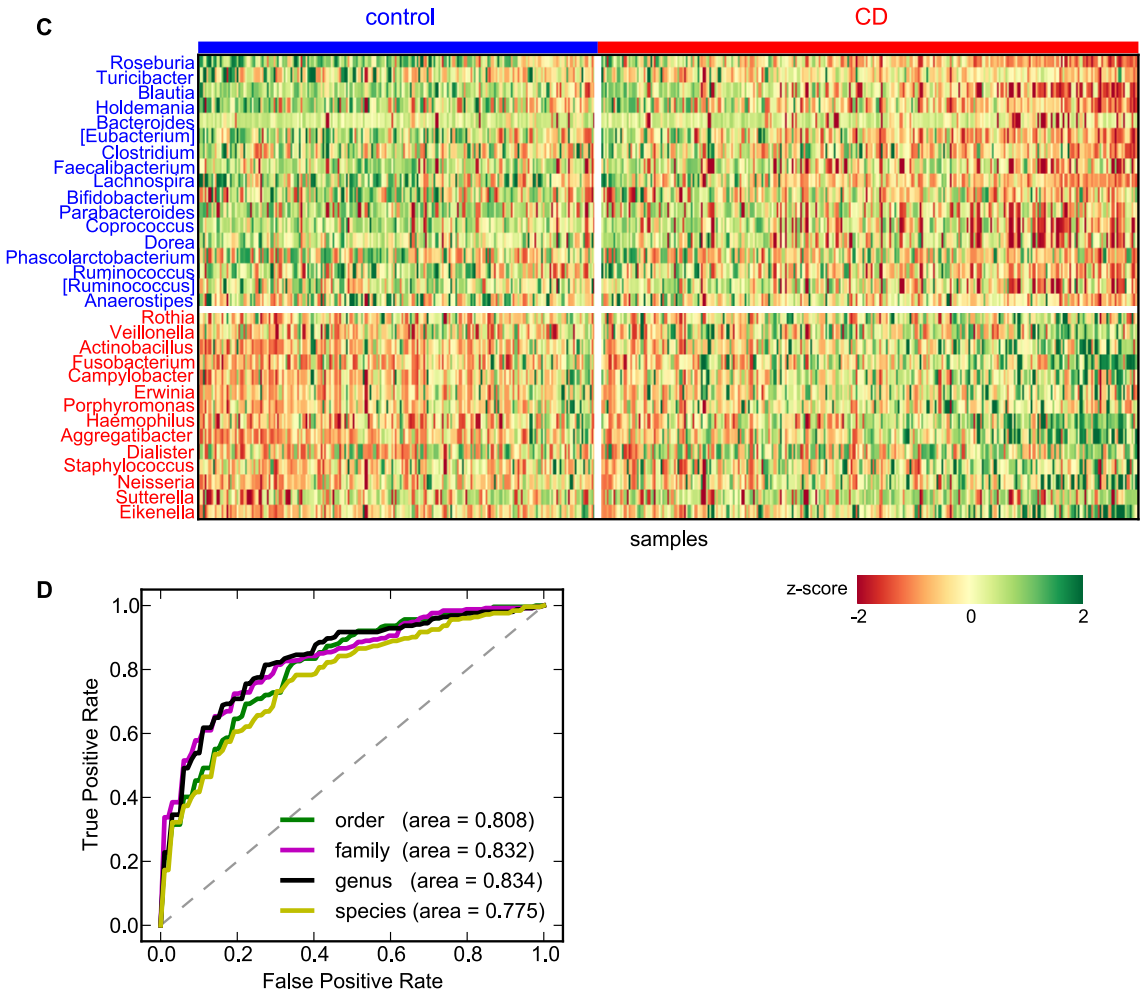

Figure S1. Related to Figure 3.

(A) Comparison of the two best methods from Figure 3C: mean log-abundance difference and arcsine-square-root regression. The log-transform-based method detected more associations at all sample sizes (positive statistic). This greater performance was statistically significant so for both small and large sample sizes; p-values were computed using the paired t-test.

(B) CD-associated taxa detected by methods other than log-abundance difference, which could also be of interest in further biological investigations.

(C) Discriminating ability and heterogeneity of z-scores for the associated genera are shown across the RISK cohort. The genera with decreased abundance in CD are colored in blue, and those with increased abundance are colored in red. Although on average red genera are increased in CD, i.e. have high z-scores (shown in green), there is large patient-to-patient variation, and some control samples show very high levels of these bacteria. Similar pattern of variation is observed for the blue genera. The samples were

arranged according to MMIC1, so a gradual transition of microbiome from control-like to CD-like is observed and the microbiomes near the control-CD boundary are quite similar. This observation parallels that made by Lewis et al. (2015), who observed two clusters of CD microbiomes: one similar to control samples and one very different. Lewis et al. also reported similar levels of heterogeneity in the patterns of microbial abundances as shown here.
(D) Grouping bacteria by either genus or family results in most accurate patient classification. ROCs are shown for SVM classifiers trained on the log-abundances of significantly associated taxa at different phylogenetic levels. Family and genus levels have comparable classification performance and are better than order and species levels.

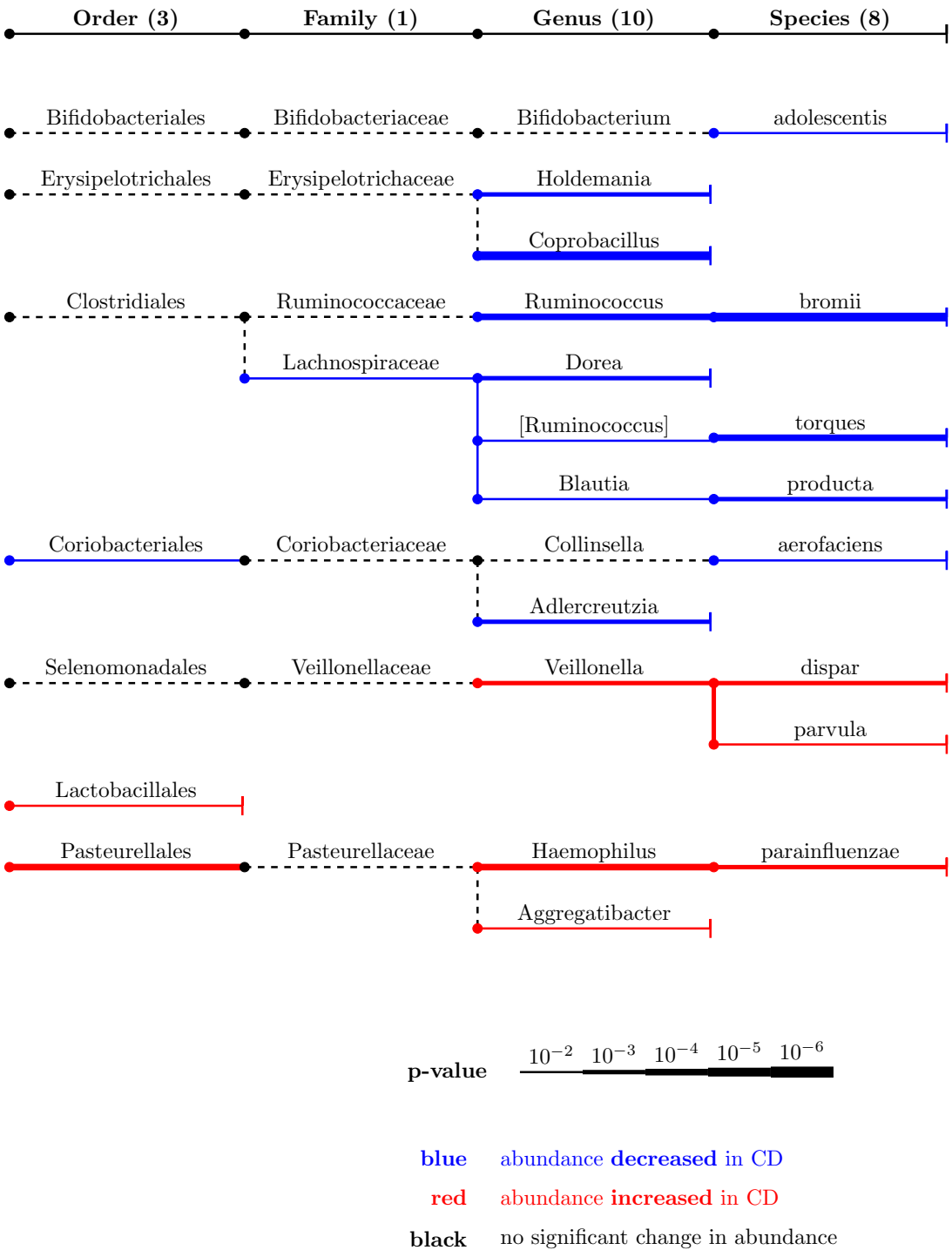

Figure S2. Related to Figure 4.
Associations with health and CD discovered in stool samples from the RISK cohort are shown across phylogenetic levels. There are both similarities and differences compared to the Figure 4, which is based on ileal samples.

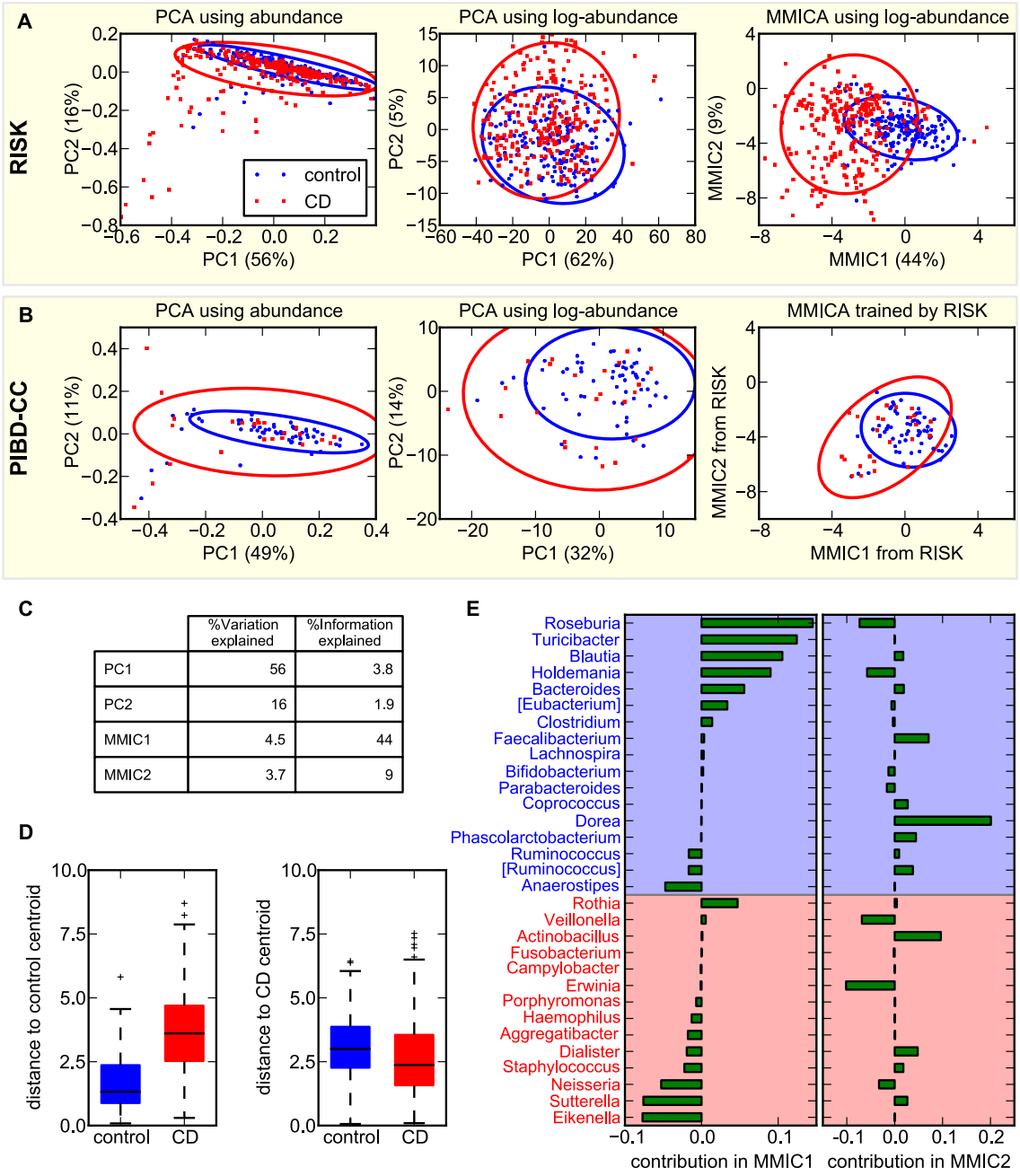

Figure S3. Related to Figure 5.

(A, B) PCA using abundance (left), PCA using log-abundance (middle) and MMICA (right) of ileal samples in RISK (A) and PIBD-CC (B). The ellipses contain 95% of the probabilities for control and CD samples, centering at the corresponding centroids. Using log-abundance instead of abundance only improves PCA marginally, and the visual separation between control and CD is strongest in MMICA.

(C) The percentages of explained variations and mutual information with diagnosis for the first two PCA and MMICA components. PCA explains larger portion the variance in community composition, but contains little information on the diagnosis. In contrast, MMICA explains only a small fraction of the variance in community composition, but contains a lot of information on the diagnosis.

(D) The distances of control samples to the control centroid in MMICA are significantly smaller than those of CD samples (p-value < 2.2e-16, t-test). The distances to CD centroid in MMICA are also significantly different for control and CD samples (p-value = 6e-4, t-test). These statistics illustrate that MMICA successfully discriminates CD and control samples.

(E) MMICs show the contribution of different genera to dysbiosis. The genera with decreased abundance in CD are shown on a blue background, and the genera with increased abundance are shown on a red background. The contribution of a genus is defined as $\text{sign}(\boldsymbol{e}_k)\boldsymbol{e}_k^2$, where $\boldsymbol{e}_k$ is the corresponding component of the kth genus in the normalized direction $\boldsymbol{e}$ of the MMIC.

Table S1. PIBD-CC: Patient Characteristics, Related to Experimental Procedures.

	Entire Cohort (n = 87)	**CD** (n = 24)	**Non-IBD Controls** (n = 63)	*P* [b]
Age, mean ± SD, years	**12.0 ± 3.5**	**12.6 ± 2.2**	**11.8 ± 3.9**	**0.35**
Male, N (%)	**45 (52%)**	**15 (63%)**	**30 (48%)**	**0.21**
Race, N (%)				**NS**
White	**71 (82%)**	**19 (79%)**	**52 (83%)**	
Black	**10 (11%)**	**4 (17%)**	**6 (10%)**	
Native American	**1 (1%)**		**1 (2%)**	
Pacific Islander	**1 (1%)**		**1 (2%)**	
Mixed Race	**2 (2%)**	**1 (4%)**	**1 (2%)**	
Unknown	**2 (2%)**		**2 (3%)**	
Hispanic, N (%)	**8 (9%)**	**3 (13%)**	**5 (8%)**	**0.43**
Montreal Classification, N (%)				
L1 (ileal)	**N/A**	**2 (8%)**	**N/A**	**N/A**
L2 (colonic)		**5 (21%)**		
L3 (ileocolonic)		**15 (63%)**		
L4 (isolated upper disease)		**2 (8%)**		
ESR, mm/h	(n = 39)	(n = 17)	(n = 22)	**0.02**[a]
Mean ± SD	**23.7 ± 21.9**	**34.2 ± 27.9**	**15.5 ± 10.7**	
Median	**17**	**25**	**15**	
Hematocrit, %	(n = 47)	(n = 19)	(n = 28)	**<0.01**[a]
Mean ± SD	**37.1 ± 3.9**	**34.5 ± 3.8**	**38.8 ± 3.1**	
Median	**37**	**35.8**	**38.4**	
Albumin, g/dL	(n = 38)	(n = 16)	(n = 22)	**<0.01**[a]
Mean ± SD	**4.1 ± 0.6**	**3.6 ± 0.5**	**4.5 ± 0.5**	
Median (IQR)	**4.3**	**3.65**	**4.5**	

CD = Crohn's disease
[a] Statistically significant
[b] Crohn's disease Vs. Non-IBD Controls
NA: Not applicable
NS: Not Significant

Table S2. PIBD-CC: Diagnosis in Non-IBD Controls, Related to Experimental Procedures.

Diagnosis	N (%)
Gastrointestinal Polyp(s)	11 (17%)
Irritable Bowel Syndrome	9 (14%)
Gastroesophageal Reflux Disease	7 (11%)
Eosinophilic Colitis	3 (5%)
Helicobacter pylori Disease	2 (3%)
Gastritis (not otherwise specified)	1 (2%)
Constipation	1 (2%)
Intussusception	1 (2%)
Immune Deficiency	1 (2%)
Unknown	27 (43%)