2016

Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota

Feng Wang
Boston University

Jess L. Kaplan
Harvard Medical School, Boston

Benjamin D. Gold
3Children's Healthcare of Atlanta

Manoj K. Bhasin
Beth Israel Deaconess Medical Center and Harvard Medical School

Naomi L. Ward
University of Wyoming, Laramie

See next page for additional authors

Authors

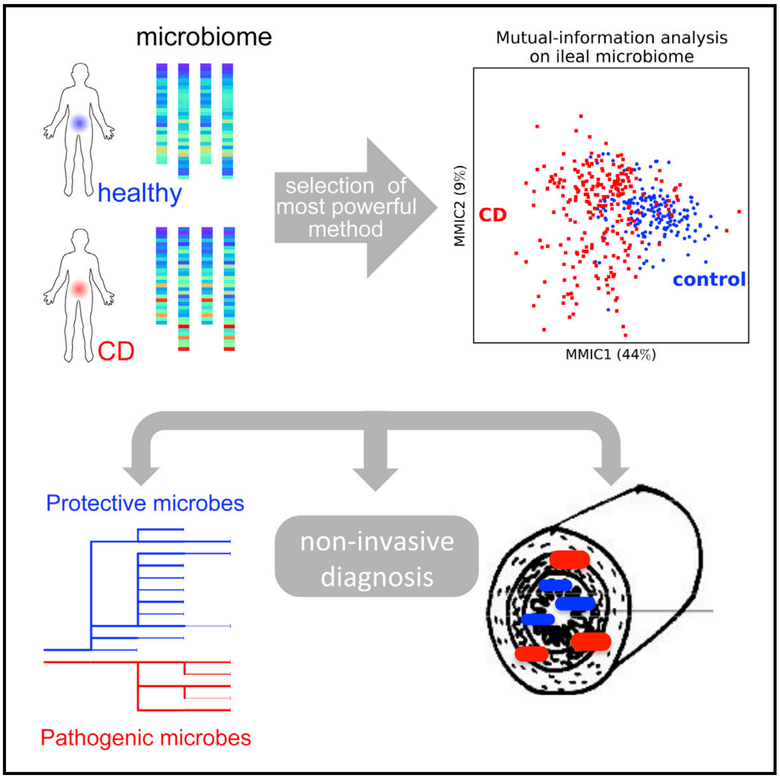
Feng Wang, Jess L. Kaplan, Benjamin D. Gold, Manoj K. Bhasin, Naomi L. Ward, Richard Kellermayer, Barbara S. Kirschner, Melvin B. Heyman, Scot E. Dowd, Stephen B. Cox, Haluk Dogan, Blaire Steven, George Ferry, Stanley A. Cohen, Robert N. Baldassano, Christopher J. Moran, Elizabeth A. Garnett, Lauren Drake, Hasan H. Otu, Leonid A. Mirny, Towia A. Libermann, Harland S. Winter, and Kirill S. Korolev

Cell Reports

Resource

Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota

Graphical Abstract



Highlights

- Mutual information distinguishes Crohn disease and controls by using the microbiome
- Stool and ileal microbiomes contain the same information about Crohn disease
- Microbes more abundant in ileum than in stool positively correlate with Crohn disease
- Statistical power to detect association varies greatly among commonly used methods

Authors

Feng Wang, Jess L. Kaplan, Benjamin D. Gold, ..., Towia A. Libermann, Harland S. Winter, Kirill S. Korolev

Correspondence

hwinter@partners.org (H.S.W.), korolev@bu.edu (K.S.K.)

In Brief

Wang et al. develop computational methods to detect and depict associations between microbes and disease. These methods improve diagnosis of Crohn disease based on the microbiome, especially from fecal samples.

Wang et al., 2016, Cell Reports *14*, 945–955
February 2, 2016 ©2016 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.12.088

Cell Reports
Resource

OPEN
ACCESS
CellPress

Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota

Feng Wang,[1,14] Jess L. Kaplan,[2,14] Benjamin D. Gold,[3] Manoj K. Bhasin,[4] Naomi L. Ward,[5] Richard Kellermayer,[6] Barbara S. Kirschner,[7] Melvin B. Heyman,[8] Scot E. Dowd,[9] Stephen B. Cox,[9] Haluk Dogan,[10] Blaire Steven,[5] George D. Ferry,[6] Stanley A. Cohen,[3] Robert N. Baldassano,[11] Christopher J. Moran,[2] Elizabeth A. Garnett,[8] Lauren Drake,[2] Hasan H. Otu,[10] Leonid A. Mirny,[12] Towia A. Libermann,[4] Harland S. Winter,[2,15,*] and Kirill S. Korolev[1,13,15,*]

[1] Bioinformatics Graduate Program, Boston University, Boston, MA 02215, USA
[2] Department of Pediatrics, MassGeneral Hospital for Children, Harvard Medical School, Boston, MA 02114, USA
[3] Children's Healthcare of Atlanta, LLC; GI Care for Kids, LLC; Atlanta, GA 30342, USA
[4] BIDMC Genomics, Proteomics, Bioinformatics and Systems Biology Center and Department of Medicine, Division of Interdisciplinary Medicine and Biotechnology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02115, USA
[5] Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA
[6] Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
[7] Department of Pediatrics, University of Chicago Comer Children's Hospital, Chicago, IL 60637, USA
[8] Department of Pediatrics, University of California, San Francisco, San Francisco, CA 94143, USA
[9] Molecular Research MR DNA, Shallowater, TX 79363, USA
[10] Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA
[11] Division of Gastroenterology, Hepatology, and Nutrition, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[12] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[13] Department of Physics, Boston University, Boston, MA 02215, USA
[14] Co-first author
[15] Co-senior author
*Correspondence: hwinter@partners.org (H.S.W.), korolev@bu.edu (K.S.K.)
http://dx.doi.org/10.1016/j.celrep.2015.12.088

SUMMARY

The relationship between the host and its microbiota is challenging to understand because both microbial communities and their environments are highly variable. We have developed a set of techniques based on population dynamics and information theory to address this challenge. These methods identify additional bacterial taxa associated with pediatric Crohn disease and can detect significant changes in microbial communities with fewer samples than previous statistical approaches required. We have also substantially improved the accuracy of the diagnosis based on the microbiota from stool samples, and we found that the ecological niche of a microbe predicts its role in Crohn disease. Bacteria typically residing in the lumen of healthy individuals decrease in disease, whereas bacteria typically residing on the mucosa of healthy individuals increase in disease. Our results also show that the associations with Crohn disease are evolutionarily conserved and provide a mutual information-based method to depict dysbiosis.

INTRODUCTION

Hosts rely on microbiota for the digestion of food (Breznak and Brune, 1994), vitamin biosynthesis (Turnbaugh et al., 2007), behavioral responses (Cryan and Dinan, 2012), protection from pathogens, (Buffie et al., 2012), and other functions (Stefka et al., 2014). The host-microbe relationship, however, can turn awry due to a simple infection, changes in nutrition, or a more nuanced dysbiosis. Microbial dysbiosis has been implicated in many human diseases including diabetes, autism, and obesity. A particularly strong relationship between disease and microbiota exists for Crohn disease (CD) and ulcerative colitis, the two major subtypes of inflammatory bowel disease (IBD) (Mazmanian et al., 2008; Greenblum et al., 2012; Manichanh et al., 2012), characterized by chronic inflammation of the gastrointestinal tract, which causes significant morbidity and can lead to colorectal cancer or death (Card et al., 2003). With more than 1.4 million people affected in the United States (CCFA, 2015), IBD poses an urgent challenge to understand the link between microbiota and human health.

The development of IBD depends on a diverse set of factors including lifestyle (Bernstein and Shanahan, 2008), environment (Danese et al., 2004), and genetic predisposition (Jostins et al., 2012). Gut microbes also contribute to IBD, and deviations from the microbial composition of the healthy human gut have been detected in patients with long-standing or newly diagnosed

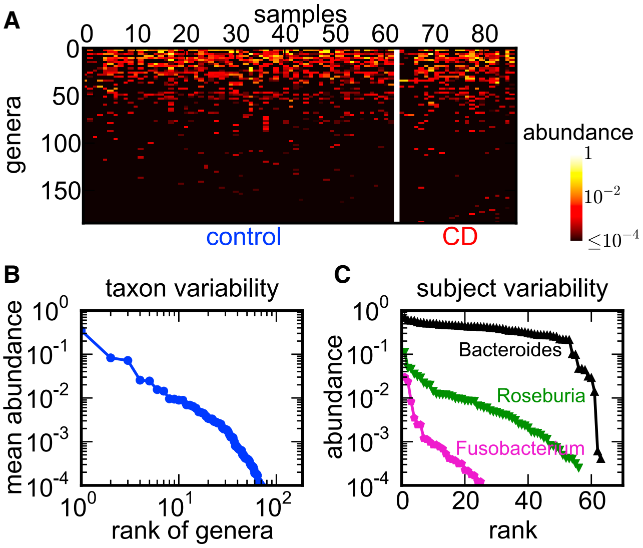


Figure 1. High Variability of Bacterial Abundances in the Human Gut Microbiota

(A) Abundance variation across all genera detected in PIBD-CC dataset. Genera are ranked by their mean relative abundance in controls.

(B) Mean genera abundances are distributed according to a power law.

(C) Rank-abundance distributions are shown for three typical genera in controls. The high subject-to-subject variability (two to three orders of magnitude) is typical for other genera, other phylogenetic levels, and in CD.

IBD (Gevers et al., 2014; Papa et al., 2012). Mouse studies have demonstrated that microbes are required for IBD, and microbial dysbiosis precedes IBD onset (Kim et al., 2007; Overstreet et al., 2010); moreover, microbiome-derived compounds can ameliorate chronic intestinal inflammation (Furusawa et al., 2013). Given the substantial role of microbes in the disease, we need to carefully characterize the changes in the microbiota that accompany IBD, particularly in early or new-onset disease This information can improve IBD diagnostics, identify disease subtypes, elucidate the mechanisms of IBD onset and progression, and uncover novel therapeutic strategies.

Although *16S rRNA* and metagenomic sequencing provide a detailed view of the gut microbiota, translating these data into clinical insights has been difficult (De Cruz et al., 2012). The analysis is often complicated by the extreme variability of the microbial abundances across both patients and species. As a result, commonly used statistical approaches may overlook important changes associated with IBD and fail to translate these changes into useful predictions. Here, we present a set of methods to identify changes in gut microbial composition associated with a disease and use them to diagnose CD based on an individual's microbiota. The performance of these methods was evaluated on two datasets: the previously interrogated RISK cohort, the most comprehensive dataset of treatment-naive pediatric CD (Gevers et al., 2014), and an independently obtained Pediatric Inflammatory Bowel Disease Consortium Cohort (PIBD-CC), which similarly includes only pediatric patients with treatment-naive IBD and controls (see Experimental Procedures and Tables S1 and S2). Our methods had a substantially higher statistical power and could find disease-associated microbes with fewer samples compared to more commonly used statistical approaches.

In addition to the development and validation of the improved approaches to the statistical analysis and depiction of microbial communities, we report several important biomedical findings. Both CD and healthy microbiota showed a power-law distribution of taxa abundance, indicating that the vast majority of taxa are rare, including those associated with the disease. The subject-to-subject variation of microbial abundance was also extreme and posed a significant challenge to standard statistical methods. Despite this high variation, we identified additional taxa associated with CD and found that the phylogenetic trees of CD-associated and health-associated bacteria do not overlap, suggesting that factors promoting health or disease have distinct evolutionary history. We also found that microbes preferentially associated with the ileal mucosa in healthy people proliferate in the stool of patients with CD, whereas bacteria more prevalent in the stool of healthy people tend to decrease in abundance in patients with CD. This observation allowed us to develop a diagnostic tool based on non-invasively collected stool samples. Contrary to the previous analysis of the RISK cohort (Gevers et al., 2014), we found that both stool and ileal mucosal samples have equal predictive power.

RESULTS AND DISCUSSION

Here we focus on two independent cohorts of patients with CD and non-IBD controls: PIBD-CC and RISK. Both cohorts were mostly pediatric (ages 2–20 years), balanced with respect to sex, race, and other factors (Table S1), and contained only subject with newly diagnosed and treatment-naive CD. The PIBD-CC contained only ileal mucosa samples, whereas RISK had samples from both stool and ileal mucosa. For both cohorts, the compositions of bacterial communities in mucosal and stool samples were obtained via DNA extraction followed by *16S rRNA* gene sequencing and processing with the Quantitative Insights Into Microbial Ecology (QIIME) software (Caporaso et al., 2010); see Experimental Procedures and Gevers et al. (2014) for further details. The sizes and sequencing depth of the two cohorts were very different. RISK is larger with over 700 patients and ~30,000 mean number of reads per sample. In contrast, PIBD-CC had only 87 patients with the mean number of reads per samples of only ~3,000. These order of magnitude variations in sample sizes and sequencing depths span the spectrum of microbiome research and illustrate the performance of our statistical approaches in different settings: from a pilot study to a large nation-wide effort.

High Variability in Microbial Abundances

Host-associated microbial communities are highly variable (Figure 1A). The first aspect of this variability is the power law distribution of relative abundances of different taxa (Figure 1B). This power-law variability is observed in both health and CD as well as in both microbiota obtained from a single subject and averaged across the cohort. Not only do different taxa have abundances that vary by orders of magnitude, but also the number of taxa grows as their abundance declines, so most taxa are rare. While the more abundant taxa are probably more important

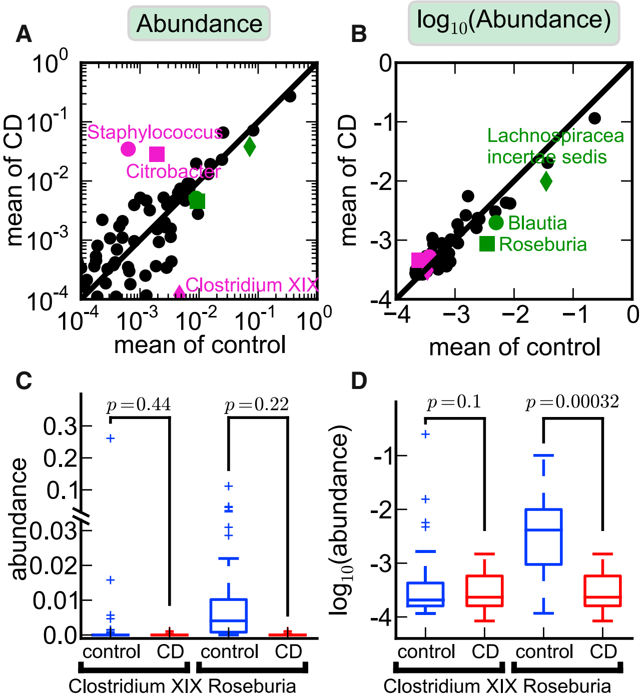


Figure 2. Log-Transformation Reduces the Variability and Helps Detect Significant Changes in Abundance between Control and CD

(A) The scatterplot shows the mean abundances of all genera in control versus CD in PIBD-CC. The purple symbols correspond to the largest changes in the mean abundance between control and CD.

(B) The same as (A) but for mean log-abundance. Note the dramatic reduction in the deviation from the diagonal compared to (A). The green symbols label the largest changes in the mean log-abundance.

(C and D) The statistical significance of outliers in (A) and (B) is evaluated in (C) and (D). The large difference in mean abundance for *Clostridium* cluster XIX is not statistically significant, whereas the highly significant association of *Roseburia* is only detected by the mean log-abundance.

for gut health, even a rare microbe can trigger chronic inflammation or dysbiosis (Powell et al., 2012) Thus, analysis should be able to handle many rare taxa and integrate changes in abundances across taxa with different prevalence. The second aspect of this high variability is the dramatic subject-to-subject variation in the abundance of a single taxon observed in both healthy subjects and patients (Figure 1C). The abundance of a given genus typically varies by more than two orders of magnitude among individuals, even for highly abundant microbes like *Bacteriodes* and *Roseburia*. Deep sequencing of a large number of samples is an expensive and time-consuming way to overcome the high variability in species abundance. Moreover, large sample sizes may not be available for rare or emergent diseases. Hence, methods that can manage with both small sample sizes and high variability are needed to analyze changes in microbial communities.

Log-Abundance Is a Less Variable Metric Than Abundance

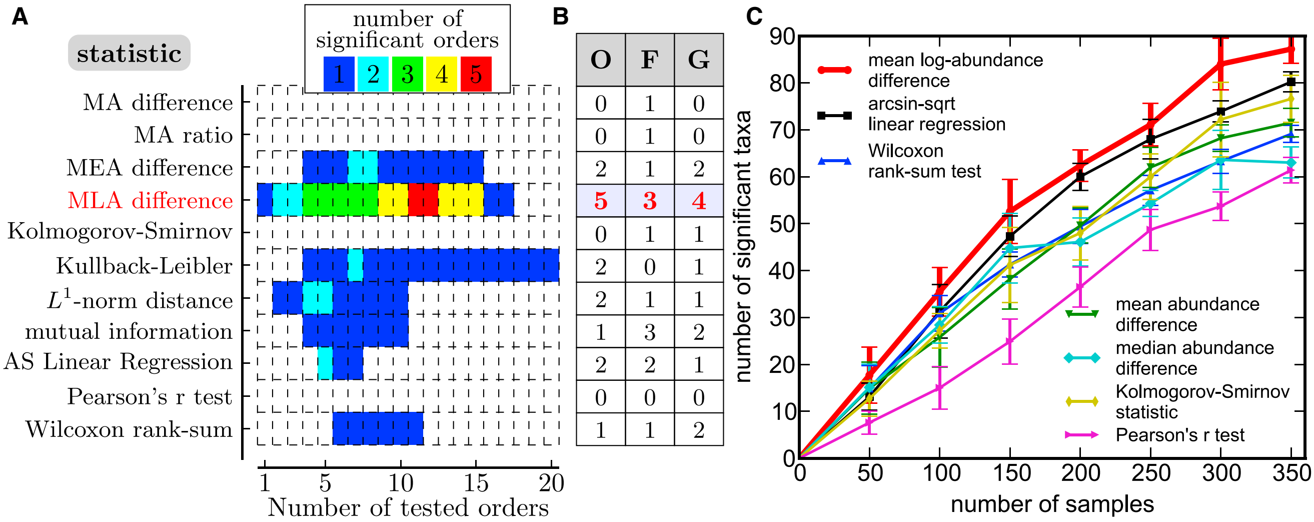
Our main observation in Figure 1 is that the variation of species abundances is better captured on a logarithmic rather than linear scale. Indeed, variations in diet, immune pressure, and other aspects of host-microbe interactions affect microbial composition, for example, by changing the growth rates of the different bacterial species (Caballero and Pamer, 2015). The randomness associated with growth rates is known to break the assumptions of the central limit theorem in statistics and even prevent the convergence of the sample mean to the true population mean (Redner, 1989). These difficulties can be resolved with a simple log-transformation of the data; instead of describing each taxonomic unit by its average abundance, as it is commonly done in the literature (Claesson et al., 2011; Wang et al., 2014), we first computed logarithm of the relative abundances in each sample and then averaged them over the samples (see Experimental Procedures).

Log-transformation resolved many of the complications due to the high variability of the gut microbiota (Figure 2) and the artifacts of the compositional bias due to the conversion of sequence counts into relative abundances (Friedman and Alm, 2012). We found that when abundances are used to detect diseases-associated taxa, large variation made it hard to detect any significant association between a taxon and the disease. When log-abundances were used instead, several significant associations were detected as determined by low p values for permutation tests of association (Figure 2).

Log-transformations have become standard in other areas of bioinformatics (Quackenbush, 2002), but they are not universally used in microbiome research (Gevers et al., 2014; White et al., 2009; Huse et al., 2014). Although a few microbiome studies have incorporated log-transformations in their analysis (Hong et al., 2006), untransformed data or non-parametric statistical tests are predominantly used to detect associations (Le Chatelier et al., 2013; David et al., 2014; Qin et al., 2014; Wang et al., 2014). Analysis of untransformed data suffers from the extreme variation in the microbiome abundances whereas rank-based, non-parametric methods discard some of the available information and lose statistical power. Indeed, small changes in abundance typically result in large changes in rank when relative abundances have a fat-tailed distribution (Huse et al., 2014).

Comparison of Methods to Detect Associations

We compared the performance of log-transforms and other commonly used techniques, which can be divided into four classes. The first class includes mean abundance, mean log-abundance, and median. These statistics represent the distribution of relative abundances found in a particular group of subjects by a single number. The second class contains methods aiming to estimate the actual distributions of relative abundances in each of the subject groups and then quantify the differences between these distributions. Such methods include Kolmogorov-Smirnov statistic, Kullback-Leibler divergence, L^2-norm distance, and mutual information between the distribution and diagnosis (Reza, 2012; Experimental Procedures). The third class is based on the regression of the diagnosis on the abundance of a given taxon, and we examined the linear regression on arcsine-square-root transformed abundances (Gevers et al., 2014). Finally, the fourth class consists of the non-parametric methods based on the differences in the ranks of taxa across



statistic	O	F	G
MA difference	0	1	0
MA ratio	0	1	0
MEA difference	2	1	2
MLA difference	5	3	4
Kolmogorov-Smirnov	0	1	1
Kullback-Leibler	2	0	1
L^1-norm distance	2	1	1
mutual information	1	3	2
AS Linear Regression	2	2	1
Pearson's r test	0	0	0
Wilcoxon rank-sum	1	1	2

Figure 3. Statistical Methods Differ in Their Ability to Detect Association

(A) The number of significantly associated orders (FDR = 0.05; permutation test) is shown for different statistical tests and different number of orders tested. Orders were tested in the order of their abundance. Initially, the number of detected orders increases with the number of tests because true positive results are more likely to be included in the set of orders tested, but it eventually declines because the threshold for statistical significance increases with the number of tests. MA, mean abundance; MEA, median abundance; MLA, mean log-abundance; AS, arcsine-square-root.

(B) The maximal number of associations detected by a procedure illustrated in (A) is shown for three phylogenetic levels (O, order; F, family; G, genus); mean log-abundance outperforms other methods.

(C) The procedure illustrated in (A) was applied to subsamples from the RISK data. The mean log abundance outperforms other methods for all sample sizes; see Figure S1 for statistical significance. Error bars are SDs from ten sub-samplings.

The power of the detected associations at discriminating control from CD samples and further details are shown in Figure S1.

subject groups. Here, we examined the Wilcoxon rank-sum statistic (Le Chatelier et al., 2013) commonly used in ecological literature (Hoekstra et al., 2001). The UniFrac statistic (Lozupone and Knight, 2005) was not considered because it is primarily based on the presence and absence of evolutionary distant taxa, whereas no major taxa losses or gains have been observed in the data (Figure 1A).

For many of the aforementioned statistics, there are approximate methods to estimate their significance. These methods, however, rely on assumptions that may not hold for the highly variable data on microbiota composition. To avoid the associated biases in comparing the different methods, we subjected all statistics to the permutation test, which is an exact statistical test of association (Experimental Procedures). Moreover, we analyzed different phylogenetic levels separately as not to bias false discovery rate (FDR) estimates by the correlations among higher and lower phylogenetic ranks. The results of our comparison of different methods to detect associations in the PIBD-CC dataset are shown in Figure 3A. Surprisingly, the mean log-difference—one of the simplest tests—detected more orders, families, and genera associated with CD than any other method. Because all of the evaluated methods relied on the same assumptions about the data and had the same FDR, the higher number of detected associations faithfully represents the higher statistical power of a test to distinguish signal from noise.

To test if the advantage of the mean log-abundance method is robust, we repeatedly subsampled the ileal mucosa samples from the larger RISK cohort at various sample sizes ranging from 50 to 350 with equal number of control and CD and obtained the average number of associations detected with the FDR corrected q value lower than 0.05. This analysis (Figure 3C) shows that the mean-log difference identified more taxa across all sample sizes. A comparable number of associations was also detected using the arcsine-square-root transform, which is similar to the log-transform because both discount the contribution of samples with exceedingly large relative abundance. Nevertheless, the number of associations detected by our method was significantly higher for most sample sizes tested (Figure S1). Importantly, none of the methods approached saturation in the number of identified taxa as the amount of data increased. This indicates that larger studies may uncover additional taxa associated with CD and provide deeper understanding of this disease. We also note that the association of taxa with the disease status is highly heterogeneous across subjects (Figure S1), a finding that parallels the recent discovery of two microbiome clusters in patients with CD: one similar to and one different from the typical control microbiome (Lewis et al., 2015).

The associations detected by the mean log-abundance method in RISK cohort are shown in Figure 4. In total, we identified 15 orders, 26 families, 31 genera, and 20 species associated with CD; many of them not found by previously used methods (Gevers et al., 2014). In agreement with the previous studies, some of the strongest associations were with the *Lachnospiraceae* family, a core and ancient member of the commensal microbiota and *Pasteurellaceae* family, which contains many human pathogens (De Cruz et al., 2012). The additional associations found in RISK data are also unlikely to be spurious because

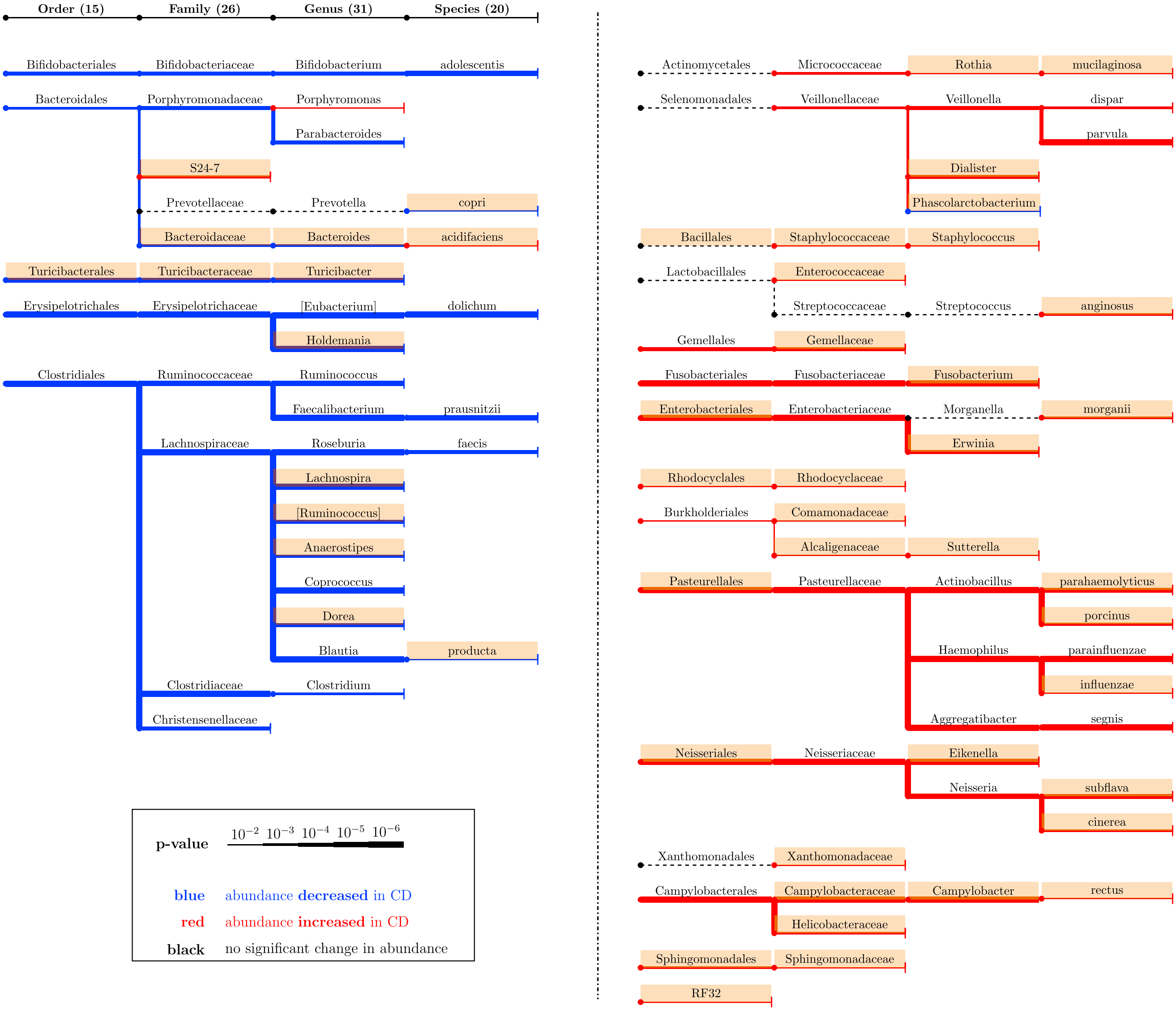


Figure 4. Significant Associations in the RISK Cohort

The phylogenetic tree of associated taxa (detected by the mean log-abundance). Note that the phylogenetic trees formed by the health-associated and disease-associated bacteria have little overlap, suggesting deep evolutionary roots of traits related to health and disease. New findings compared to Gevers et al., 2014 are shaded. The findings from stool samples are shown in Figure S2.

many of these taxa were previously identified in other IBD cohorts or are otherwise known to contribute to disease. For example, the association of *Staphylococcus* with CD is known from a different IBD cohort (Nguyen et al., 2010), *Turicibacter* is less abundant in the fecal samples from dogs with IBD (Suchodolski et al., 2012), *Eikenella* can cause periodontitis and other infections in the oral cavity (Aas et al., 2005), and some strains of *Enterobacteriaceae* thrive in the presence of inflammation and outcompete the healthy microbiota in mice (Garrett et al., 2010). Of special interest are the two commensal species *Bacteroides fragilis* and *Fecalibacterium prausnitzii* that suppress inflammation and have received a lot of attention in IBD literature (De Cruz et al., 2012). Although we found a strong association between health and *F. prausnitzii*, no significant effect of *B. fragilis* could be inferred from the data.

Many of the taxa enriched in controls are known to provide important functions for gut health. For example, *Roseburia*, *Blautia*, and *F. prausnitzii* produce butyrate, which acts as an energy source for epithelial cells in the gut (Duncan et al., 2002). Supplementing patients with such bacteria or the metabolites they produce could then be a potential intervention strategy (Furusawa et al., 2013). In contrast, a majority of the taxa enriched in CD are known to be opportunistic pathogens (Mukhopadhya et al., 2012). Collectively, these previous studies suggest that some of the CD-associated bacteria are deleterious to the host, whereas some of the health-associated bacteria are beneficial

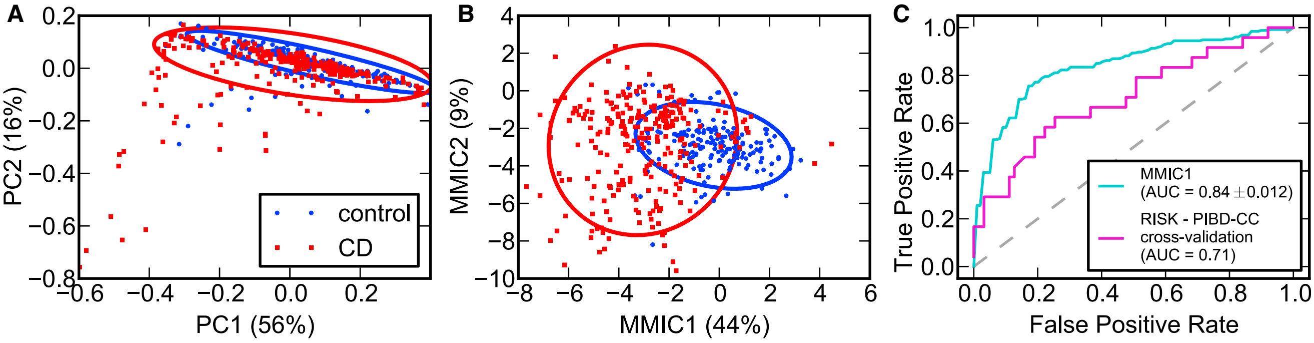


Figure 5. Microbiota Composition Distinguishes Health from Disease

(A) Principle component analysis (PCA) on abundance data yields poor separation of CD and control samples; the ellipses contain 95% of the probabilities for control and CD samples, centering at the corresponding centroids.

(B) Maximal mutual information component analysis (MMICA) on log-abundance data yields a much better separation of CD and control samples; the ellipses contain 95% of the probabilities for control and CD samples, centering at the corresponding centroids. The difference between the distances to the centroids is statistically significant; see Figure S3.

(C) The first MMIC trained on RISK cohort can classify both RISK and PIBD-CC samples. For RISK data, the curve is the averages over 5-fold cross validation. See also Figure S3.

to the host. However, further experimental studies that can establish causality are required to determine the specific roles of the associations reported in Figure 4.

The increase in the number of pathogenic bacteria in CD may reflect the overall decline of gut health. We tested this hypothesis by looking for a link between active inflammation and bacterial composition using the mean log-abundance method, but failed to detect any significant associations in agreement with other studies (De Cruz et al., 2012). The types of pathogens established in the gut could affect disease progression beyond inflammation, a hypothesis worthy of further investigation.

Association Congruence across Phylogenetic Ranks in Ileal Microbiome

Phylogenetic structure of associations is rarely discussed in microbiome research. Most studies focus on order or family-level associations because the large number of genera and species increases the number of hypotheses tested and reduces the statistical power. However, to link compositional changes in microbiota to human health, we need to understand the relationship between ecological functions and phylogenetic distance better.

Typically, related organisms have similar genomes and occupy similar ecological niches. Yet, studies of parasites, symbionts, and commensals have shown that closely related species could have dramatically different life styles and effects on the host (Siddall et al., 1993; Moran et al., 2008). We found that, when a higher phylogenetic level is more abundant in CD, then lower level associations are more abundant in CD as well (Figure 4). A similar pattern of phylogenetic congruence is also observed for the taxa decreased in CD. In fact, only 4 out of 92 associations do not follow this rule. Although these exceptions can be attributable to the 5% FDR, they might also indicate particularly interesting microbial dynamics associated with the disease.

Phylogenetic congruence could be driven by conserved ecological traits or by a single lower-level taxon contributing to the association at the higher level. Consistent with the first mechanism, the strong association between CD and *Enterobacteriaceae* results from the weak associations of several genera in this family (Figure 4). The second mechanism could explain the association between CD and *Bifidobacteriaceae* family, which seems to be primarily driven by a single species—*Bifidobacterium adolescentis*.

Given these patterns of associations, there is likely an optimal phylogenetic level for microbiome analysis. Strain and species level may be too idiosyncratic since the behavior of genetically very similar organisms could be very different, whereas order and family level may be too coarse and miss important ecological functions present only in a specific genus. Here we found that the genus level yields not only the largest number of associations, but also better patient classification compared to order and species levels (Figure S1).

Classification of Ileal Samples

Diagnostics is an important application of microbiota association studies in CD; so we attempted to classify patients in RISK and PIBD-CC as CD or controls based only on the composition of their gut microbial communities. None of the commonly used unsupervised clustering methods could provide an acceptable classification of the data (see Experimental Procedures). Similarly, principle component analysis (PCA) based on microbial abundances of either all or only significantly associated genera could not differentiate patients with CD and controls (Figure 5A), consistent with the earlier analysis by Gevers et al. (2014). To improve on PCA performance, we implemented a supervised projection method by maximizing the mutual information between the linear combination of log-abundances of significantly associated bacterial taxa and diagnosis (see Experimental Procedures). This technique, termed maximal mutual information component analysis (MMICA), effectively filters out large patient-to-patient variation in microbiome composition due to numerous factors unrelated to the disease, and focuses on the few variables indicative of CD.

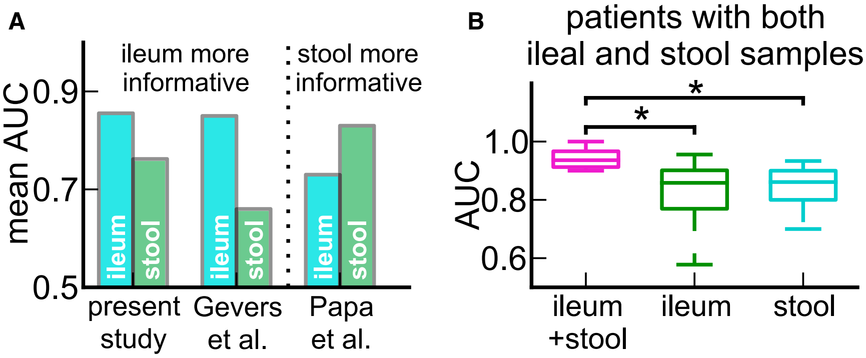


Figure 6. The Classification Power of Stool and Ileal Samples

(A) Mean AUCs of classifiers developed in this study, Gevers et al. (2014), and Papa et al. (2012), based on all available ileal or stool samples.

(B) SVM classifiers based on a subgroup of subjects with both ileal and stool samples in the RISK cohort. Ileal and stool samples had similar discriminating power, whereas their combination further increased performance.

MMICA showed dramatic improvement over PCA for both RISK (Figure 5B) and PIBD-CC (Figure S3). In particular, the first two components contained more than half of the information on the diagnosis (44% and 9% of maximally possible 0.98 bits). MMICA was also a significant improvement over a single genus analysis. *Roseburia*, the most informative genus, explained only 16.5% of the information compared to the 44% explained by the first maximal mutual information component (MMIC). Thus, MMICA significantly increases the diagnostic information contained in a single metric, capturing the major difference between controls and CD.

We inspected the contribution of each genus to the MMICs and found that the first MMIC was primarily comprised of *Roseburia*, *Turicibacter*, *Blautia*, and *Holdemania*. All four genera were decreased in CD, so the first MMIC was primarily negatively correlated to CD (Figure 5B; Figure S3). The second MMIC contained bacteria that both decrease and increase in CD and was mainly in the direction of *Dorea*, *Erwinia*, and *Actinobacillus*. The probability distribution along this second MMIC was bimodal for CD and unimodal for control patients. Most variance-based methods would not be able to take advantage of such differences in the distribution, suggesting that information-based approaches such as MMICA could have an important advantage for microbiota studies.

In addition to being a powerful depiction tool, MMICA was also able to classify samples as CD or control based on their microbial composition. We found that a simple classifier that uses the projection on first MMIC as a microbial dysbiosis index yields an area under the curve (AUC) of 0.84 (Figure 5C). To test the power of this method, the MMIC-based classifier was trained on the RISK cohort and applied to the PIBD-CC dataset. Despite the fact that PIBD-CC samples were independently collected, processed using different protocols, and sequenced at much lower depth, the classifier achieved a high AUC of 0.71, suggesting that the first MMIC is a robust indicator of dysbiosis that could reach a sufficient power for clinical applications.

Classification on Stool Samples

Stool samples hold the key to non-invasive diagnostics of IBD, but two recent studies reached opposite conclusions on the feasibility of predicting IBD from stool microbiota. On the one hand, Papa et al. (2012) found stool samples to be very predictive with an AUC of 0.83, but their cohort contained patients undergoing treatment for established disease and thus could have a much larger difference in the microbial composition between patients with CD and controls due to past and current medication use and prolonged inflammation. On the other hand, Gevers et al. (2014) found that, for treatment-naive children, stool samples were a poor predictor of the diagnosis, yielding an AUC of only 0.66.

The stool samples from the RISK cohort were reanalyzed by performing a log-transformation and identifying genera significantly associated with CD (Figure S2). We then trained an SVM classifier on the log-abundances of the significant genera and found a substantially higher mean AUC of 0.72 (95% CI 0.663–0.770), which is approaching clinically useful values (Figure 6A). This higher value of AUC (0.72 versus 0.66) underscores the advantage of our methods for small-to-medium datasets with high variability. Although our method improved the classification power of stool samples in RISK, we still found that ileal samples were more informative (AUC of 0.84 versus 0.72), in agreement with the RISK analysis (Gevers et al., 2014).

The above comparison, however, is not entirely valid because there were fewer stool than ileal samples in the RISK cohort and, unlike the ileal samples (CD to controls ratio of 254 to 187), the stool samples are very imbalanced (CD to controls ratio of 187 to 31). One way to make the comparison more fair is to focus only on patients who have both stool and ileal samples because that would make the training and test data the same for both stool and ileal based classifiers. In addition, this approach allows one to test whether stool and ileal samples contain equivalent or complementary information.

The RISK cohort had 74 patients (14 controls and 60 CDs) with both stool and ileal samples. Because the sample size was too small to detect a sufficient number of significant associations and find MMICs, we used all 31 significant genera found in ileal samples (shown in Figure 4) as features in an SVM classifier (Experimental Procedures). In contrast to the RISK-wide analysis above (Gevers et al., 2014), we found that stool samples contain comparable information to ileal samples (AUC of 0.84, 95% CI 0.747–0.911 versus 0.81 95% CI 0.654–0.925); see Figure 6B. This observation agrees with previous findings in patients with established and treated CD (Papa et al., 2012) and raises the possibility that stool samples may actually aid in the initial diagnosis of CD. We also found that an SVM classifier trained on both ileal and stool samples had an AUC of 0.94 (95% CI 0.908–0.979) (Figure 6B); however,

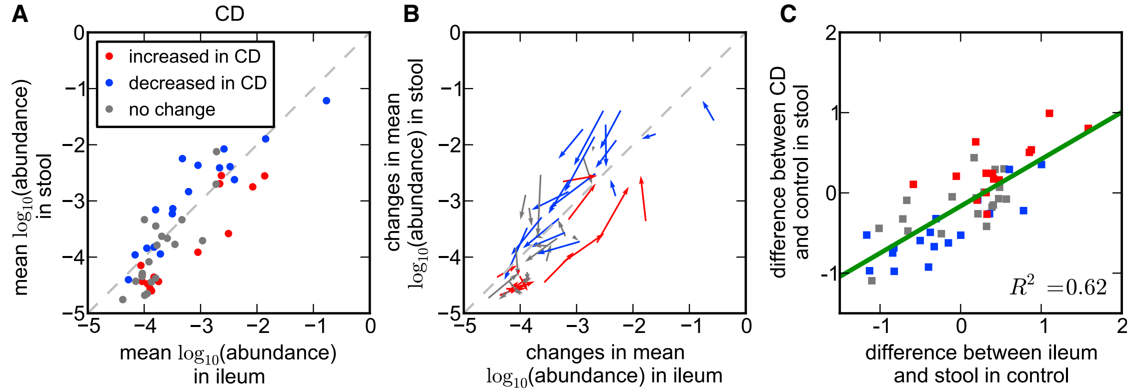


Figure 7. Stool- and Ileum-Dwelling Bacteria Have Distinct Contributions to IBD

(A) Mean log-abundance of top 50 abundant genera in ileal and stool microbiota for patients with CD. Note that health-associated bacteria are more abundant in stool than ileum (above the diagonal), whereas CD-associated bacteria are more abundant in ileum (below the diagonal).

(B) The same as (A), but the beginning and end of the arrows show the mean log-abundance in control and CD respectively; thus, the arrows show the shift in the community associated with CD. Note that stool and ileal abundances change almost equally.

(C) The preference of a genus for ileal versus stool habitat is strongly correlated with the change in its abundance between the stool of patients with CD and controls.

with the limited number of samples, it is premature to say whether stool and ileal microbiota contain complementary information or the increase in AUC simply came from doubling the number of features in the training data.

Differential Enrichments of Health and CD-Associated Microbes in Ileum and Stool

The subset of patients with both ileal and stool samples described above allowed us to examine the relationship between ileal and stool microbiota further. Specifically, we asked whether bacterial genera have similar abundances in ileal mucosa and the stool of patients with CD (Figures 7A and 7B). On average, the relative abundances of bacteria in stool and ileal samples were similar, but there was a striking difference between bacteria increased and decreased in CD (Figure 4). Bacteria associated with health were more abundant in the stool than in the ileum in patients with CD, whereas bacteria associated with CD were preferentially found in the ileum.

One possible interpretation of these data is that some of the bacteria decreased in CD perform essential digestive functions in the gut and therefore are primarily found in the stool whereas opportunistic pathogens primarily colonize the mucosa and trigger IBD when not controlled by the immune system. Microbial dysbiosis may then be driven further by mucosal oxygenation, which disrupts the normal intestinal oxygen gradient (Albenberg et al., 2014). This hypothesis would suggest that bacteria's role in CD can be predicted by its abundance in the ileal mucosa and stool of healthy patients: Bacteria more abundant in stool should decrease in CD and bacteria abundant in the mucosa should increase. Indeed, we found a very strong correlation ($R^2 = 0.62$, $p = 1.3 \times 10^{-11}$) between the difference of stool and mucosal abundances in controls and the changes in the stool of patients with CD relative to controls (Figure 7C). Consistent with our findings, previous studies have found that stool microbiota consist of two distinct components: one shed from the mucosa and a separate non-adherent luminal population (Eckburg et al., 2005). Our results further suggest that these different components may play a distinct role in IBD.

Conclusions

Broad abundance distribution of different taxa and high patient-to-patient variability challenge existing statistical tools to detect microbial associations with disease. We found that performing statistical analysis on the logarithm of relative abundances makes patterns of microbiota changes more clear and robust. For both the RISK and PIBD-CC datasets, our technique required fewer samples for similar statistical power and identified additional taxa associated with CD. Discovered associations could distinguish patients with CD versus non-IBD using a new classifier and depiction tool that identifies directions in the multi-dimensional space of microbial abundances with maximal information on the diagnosis. This maximal mutual information components analysis was superior to the commonly used principle component analysis and remained informative when validated on the independently obtained PIBD-CC dataset.

Our analysis indicates that health- and disease-associated bacteria have distinct ecology and evolutionary history. We found that bacteria increased in CD and bacteria decreased in CD formed two largely non-overlapping phylogenetic trees, suggesting that factors promoting health or disease have deep evolutionary roots and are not frequently exchanged between gut bacteria. Moreover, bacteria that proliferate in CD are preferentially associated with ileal mucosa, whereas bacteria decreased in CD reside mostly in the stool. The connection between lumen and mucosa compartments enabled patient classification using either ileal biopsies or stool samples with about equal accuracy.

Collectively, our results provide a set of statistical tools for the analysis of microbiome data, refine the link between shifts in microbial abundances and disease, and show the relevance of microbiota to the diagnosis and management of pediatric CD.

EXPERIMENTAL PROCEDURES

Study Populations

Clinical characteristics of patients from the RISK cohort have been previously described (Gevers et al., 2014). The PIBD-CC is a previously unreported cohort of children (ages 1–17 years) who underwent endoscopic evaluation for IBD at seven centers in the United States (MassGeneral Hospital for Children, University of California San Francisco, Children's Hospital of Philadelphia, University of Chicago, Texas Children's Hospital, Children's Center for Digestive Healthcare-Atlanta, GA, Children's Healthcare of Atlanta [Scottish Rite and Egleston Children's Hospital campuses]) from September 2005 until January 2008 (Emory University HIC IRB number 060-2002 and additional approval by local internal review boards of all participating institutions). PIBD-CC patients with terminal ileal biopsies and available clinical data (24 children with newly diagnosed, treatment-naive CD and 63 non-IBD control subjects) were included in this study. Clinical characteristics can be found in Tables S1 and S2. The study was supported by the following grant: Role of Infectious Agents in Pediatric Crohn's Disease, NIH R03 DK064544 (principal investigator, B.D. Gold).

Sample Collection

PIBD-CC

During diagnostic ileocolonoscopy, terminal ileal mucosal biopsy specimens were obtained using standard biopsy forceps and immediately placed in liquid nitrogen or dry ice and stored at -80°C until use.

RISK

The collection of the RISK cohort is presented in Gevers et al., 2014; 441 ileal samples (184 control and 245 CD) and 218 stool samples (31 control and 187 CD) were extracted from the RISK cohort dataset after filtering out the ones with antibiotic exposure for detection of taxa associated with CD. A subgroup of these patients (14 control and 60 CD) with both ileal and stool samples were used for the analysis in Figures 6 and 7.

DNA Extraction and *16S rRNA* Gene Sequencing

PIBD-CC

Bulk DNA was extracted from samples using the QIAGEN Stool DNA kit. Tag-encoded FLX amplicon pyrosequencing was performed as described (Bailey et al., 2011; Callaway et al., 2010; Finegold et al., 2010; Handl et al., 2011) using Gray28F 5′TTTGATCNTGGCTCAG and Gray519r 5′ GTNTTACNGC GGCKGCTG, with primers numbered in relation to the primary sequence of *E. coli 16S rRNA* (Brosius et al., 1978). Initial generation of the sequencing library used one-step PCR with 30 cycles, generating amplicons extending from the 28F primer with average read length of 400 base pairs (bp). Tag-encoded FLX amplicon pyrosequencing analyses used a Roche 454 FLX instrument with titanium reagents, and titanium procedures performed at the Research and Testing Laboratory (Lubbock, TX).

Following sequencing, all failed sequence reads, low-quality sequence ends, tags, and primers were removed, and sequence collections depleted of non-bacterial ribosome sequences and those with degenerate base calls, homopolymers >5 bp in length, reads <200 bp, and chimeras (Gontcharova et al., 2010), as described (Bailey et al., 2011; Callaway et al., 2010; Finegold et al., 2010; Handl et al., 2011).

OTU Picking

PIBD-CC

We used a naive Bayes classifier with confidence cutoff = 0.5 and RDP database (Cole et al., 2014) for OTUs assignments.

RISK

OTU picking was described in Gevers et al., 2014. Briefly, OTUs were picked using closed reference OTU picking by QIIME software (Caporaso et al., 2010) and at 97% similarity against the Greengenes database (DeSantis et al., 2006).

Logarithmic Transformation of Relative Abundance

Tables of OTU counts were transformed into relative abundances by adding a pseudocount of 1 and normalization. These were transformed into mean log abundances by first taking the natural logarithm and then averaging over the samples.

Probability Distribution Estimation

Kolmogorov-Smirnov statistic, Kullback-Leibler divergence, L^2-norm distance, and mutual information are defined on probability distribution functions, which were obtained by kernel density estimation methods (Khan et al., 2007). We used Gaussian kernels and chose their bandwidths according to Silverman's thumb rule (Silverman, 1987).

Statistical Significance

Statistical significance was evaluated by a permutation test with 10^6 permutations. The FDR correction at the level 5% was performed using Benjamini-Hochberg method. To avoid imposing an arbitrary abundance cutoff on taxa, we analyzed all possible cutoffs and reported the maximal number of associations, as illustrated in Figure 3A. Concretely, for each phylogenetic level, the taxa were first ranked by their mean log-abundances for both control and CD separately, and then merged in a single list according to their minimal rank in the two lists. We then performed association tests for species with ranks between 1 and k for all possible values of k and reported the maximal number of association. This procedure was applied uniformly to all methods presented in Figure 3.

Maximal Mutual Information Component Analysis

To find MMIC1, we obtained a linear combination of taxa log-abundances, which maximizes the mutual information about the diagnosis. The second component was also found as a linear combination maximizing the mutual information on the diagnosis, but subject to the constraint that the correlation coefficient between MMIC1 and MMIC2 equals zero. Our approach is related to a recent method developed in neuroscience (Faivishevsky and Goldberger, 2012).

Software Packages and Classification

Kernel density estimation of the probability distribution function, Kolmogorov-Smirnov statistic, Pearson r, and Wilcoxon rank-sum statistic were computed by Python package SciPy 0.14.0. Mean abundance difference/ratio and median abundance difference were computed using their definitions. PCA, all unsupervised clustering methods, and all supervised classifiers were performed using Python machine learning package scikit-learn 0.15.2 (Pedregosa et al., 2011). The supervised classifiers included logistic regression with L1 penalty, support vector machine, and random forest. The best parameters for the classifiers were found by a 5-fold cross-validation. Their performances were then measured by the area under the ROC curve, which was obtained by averaging results from 5-fold cross-validations.

ACCESSION NUMBERS

The accession number for the raw PIBD-CC *16S rRNA* sequencing data reported in this paper is NCBI: PRJNA297124.

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.12.088.

AUTHOR CONTRIBUTIONS

B.D.G., H.S.W., M.B.H., B.S.K., G.D.F., S.A.C., R.B., M.K.B., T.A.L., L.D., E.A.G., and J.L.K. designed the study; M.K.B., H.H.O., K.S.K., F.W., N.L.W., B.S., S.E.D., S.B.C., and H.D. analyzed the data; T.A.L., K.S.K., L.A.M., F.W., J.L.K., and H.S.W. interpreted the data; F.W., J.L.K., H.S.W., and K.S.K. wrote the first draft; and all authors reviewed and contributed to the final draft of the manuscript.

ACKNOWLEDGMENTS

Support was provided by the startup fund from Boston University (to K.S.K.; supported K.S.K. and F.W.). Some of the analysis was carried out on the Shared Computing Cluster at BU. Martin Schlaff (philanthropic support) and the B. Hasso Family Foundation provided support (to H.S.W.). The Houston Men of Distinction and the Gutsy Kids Fund was provided and supported by the Karen and Brock Wagner family (to R.K.). The authors would also like to acknowledge Antone Opekun for his work on this project. The PIBD-CC study was supported by NIH—R03 DK064544 (principal investigator, B.D. Gold; Role of Infectious Agents in Pediatric Crohn's Disease).

Received: October 9, 2015
Revised: December 3, 2015
Accepted: December 18, 2015
Published: January 21, 2016

REFERENCES

Aas, J.A., Paster, B.J., Stokes, L.N., Olsen, I., and Dewhirst, F.E. (2005). Defining the normal bacterial flora of the oral cavity. J. Clin. Microbiol. *43*, 5721–5732.

Albenberg, L., Esipova, T.V., Judge, C.P., Bittinger, K., Chen, J., Laughlin, A., Grunberg, S., Baldassano, R.N., Lewis, J.D., Li, H., et al. (2014). Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. Gastroenterology *147*, 1055–63.e8.

Bailey, M.T., Dowd, S.E., Galley, J.D., Hufnagle, A.R., Allen, R.G., and Lyte, M. (2011). Exposure to a social stressor alters the structure of the intestinal microbiota: implications for stressor-induced immunomodulation. Brain Behav. Immun. *25*, 397–407.

Bernstein, C.N., and Shanahan, F. (2008). Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. Gut *57*, 1185–1191.

Breznak, J.A., and Brune, A. (1994). Role of microorganisms in the digestion of lignocellulose by termites. Annu. Rev. Entomol. *39*, 453–487.

Brosius, J., Palmer, M.L., Kennedy, P.J., and Noller, H.F. (1978). Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. Proc. Natl. Acad. Sci. USA *75*, 4801–4805.

Buffie, C.G., Jarchum, I., Equinda, M., Lipuma, L., Gobourne, A., Viale, A., Ubeda, C., Xavier, J., and Pamer, E.G. (2012). 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.

Caballero, S., and Pamer, E.G. (2015). Microbiota-mediated inflammation and antimicrobial defense in the intestine. Ann. Rev. Immunol. *33*, 227–256.

Callaway, T.R., Dowd, S.E., Edrington, T.S., Anderson, R.C., Krueger, N., Bauer, N., Kononoff, P.J., and Nisbet, D.J. (2010). Evaluation of bacterial diversity in the rumen and feces of cattle fed different levels of dried distillers grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing. J. Anim. Sci. *88*, 3977–3983.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Card, T., Hubbard, R., and Logan, R.F. (2003). Mortality in inflammatory bowel disease: a population-based cohort study. Gastroenterology *125*, 1583–1590.

CCFA (2015). Crohn's & Colitis Foundation of America. Facts about inflammatory bowel disease. Published online June 2015. http://www.ccfa.org.

Claesson, M.J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J.R., Falush, D., Dinan, T., Fitzgerald, G., et al. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proc. Natl. Acad. Sci. USA *108* (*Suppl 1*), 4586–4591.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642.

Cryan, J.F., and Dinan, T.G. (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. Nat. Rev. Neurosci. *13*, 701–712.

Danese, S., Sans, M., and Fiocchi, C. (2004). Inflammatory bowel disease: the role of environmental factors. Autoimmun. Rev. *3*, 394–400.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563.

De Cruz, P., Prideaux, L., Wagner, J., Ng, S.C., McSweeney, C., Kirkwood, C., Morrison, M., and Kamm, M.A. (2012). Characterization of the gastrointestinal microbiota in health and inflammatory bowel disease. Inflamm. Bowel Dis. *18*, 372–390.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

Duncan, S.H., Hold, G.L., Barcenilla, A., Stewart, C.S., and Flint, H.J. (2002). Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. Int. J. Syst. Evol. Microbiol. *52*, 1615–1620.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635–1638.

Faivishevsky, L., and Goldberger, J. (2012). Dimensionality reduction based on non-parametric mutual information. Neurocomputing *80*, 31–37.

Finegold, S.M., Dowd, S.E., Gontcharova, V., Liu, C., Henley, K.E., Wolcott, R.D., Youn, E., Summanen, P.H., Granpeesheh, D., Dixon, D., et al. (2010). Pyrosequencing study of fecal microflora of autistic and control children. Anaerobe *16*, 444–453.

Friedman, J., and Alm, E.J. (2012). Inferring correlation networks from genomic survey data. PLoS Comput. Biol. *8*, e1002687.

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T.A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature *504*, 446–450.

Garrett, W.S., Gallini, C.A., Yatsunenko, T., Michaud, M., DuBois, A., Delaney, M.L., Punit, S., Karlsson, M., Bry, L., Glickman, J.N., et al. (2010). Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. Cell Host Microbe *8*, 292–300.

Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe *15*, 382–392.

Gontcharova, V., Youn, E., Wolcott, R.D., Hollister, E.B., Gentry, T.J., and Dowd, S.E. (2010). Black Box Chimera Check (B2C2): a Windows-Based Software for Batch Depletion of Chimeras from Bacterial 16S rRNA Gene Datasets. Open Microbiol. J. *4*, 47–52.

Greenblum, S., Turnbaugh, P.J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc. Natl. Acad. Sci. USA *109*, 594–599.

Handl, S., Dowd, S.E., Garcia-Mazcorro, J.F., Steiner, J.M., and Suchodolski, J.S. (2011). Massive parallel 16S rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats. FEMS Microbiol. Ecol. *76*, 301–310.

Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hoang, A., Hill, C.E., Beerli, P., and Kingsolver, J.G. (2001). Strength and tempo of directional selection in the wild. Proc. Natl. Acad. Sci. USA *98*, 9157–9160.

Hong, S.H., Bunge, J., Jeon, S.O., and Epstein, S.S. (2006). Predicting microbial species richness. Proc. Natl. Acad. Sci. USA *103*, 117–122.

Huse, S.M., Young, V.B., Morrison, H.G., Antonopoulos, D.A., Kwon, J., Dalal, S., Arrieta, R., Hubert, N.A., Shen, L., Vineis, J.H., et al. (2014). Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. Microbiome *2*, 5.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

Khan, S., Bandyopadhyay, S., Ganguly, A.R., Saigal, S., Erickson, D.J., 3rd, Protopopescu, V., and Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *76*, 026209.

Kim, S.C., Tonkonogy, S.L., Karrasch, T., Jobin, C., and Sartor, R.B. (2007). Dual-association of gnotobiotic IL-10-/- mice with 2 nonpathogenic commensal bacteria induces aggressive pancolitis. Inflamm. Bowel Dis. *13*, 1457–1466.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.M., Kennedy, S., et al.; MetaHIT consortium (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546.

Lewis, J.D., Chen, E.Z., Baldassano, R.N., Otley, A.R., Griffiths, A.M., Lee, D., Bittinger, K., Bailey, A., Friedman, E.S., Hoffmann, C., et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. Cell Host Microbe *18*, 489–500.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. *71*, 8228–8235.

Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. (2012). The gut microbiota in IBD. Nat. Rev. Gastroenterol. Hepatol. *9*, 599–608.

Mazmanian, S.K., Round, J.L., and Kasper, D.L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. Nature *453*, 620–625.

Moran, N.A., McCutcheon, J.P., and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. *42*, 165–190.

Mukhopadhya, I., Hansen, R., El-Omar, E.M., and Hold, G.L. (2012). IBD-what role do Proteobacteria play? Nat. Rev. Gastroenterol. Hepatol. *9*, 219–230.

Nguyen, G.C., Patel, H., and Chong, R.Y. (2010). Increased prevalence of and associated mortality with methicillin-resistant Staphylococcus aureus among hospitalized IBD patients. Am. J. Gastroenterol. *105*, 371–377.

Overstreet, A.M.C., Ramer-Tait, A.E., Atherly, T.A., Phillips, G.J., Hostetter, J., Ziemer, C.J., Wannemuehler, M.J., and Jergens, A. (2010). W1829 changes in composition of the intestinal microbiota precede onset of colitis in genetically-susceptible (IL-10−/−) mice. Gastroenterology *138*, 748–749.

Papa, E., Docktor, M., Smillie, C., Weber, S., Preheim, S.P., Gevers, D., Giannoukos, G., Ciulla, D., Tabbaa, D., Ingram, J., et al. (2012). Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PLoS ONE *7*, e39242.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Powell, N., Walker, A.W., Stolarczyk, E., Canavan, J.B., Gökmen, M.R., Marks, E., Jackson, I., Hashim, A., Curtis, M.A., Jenner, R.G., et al. (2012). The transcription factor T-bet regulates intestinal inflammation mediated by interleukin-7 receptor+ innate lymphoid cells. Immunity *37*, 674–684.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature *513*, 59–64.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nat. Genet. *32* (*Suppl*), 496–501.

Redner, S. (1989). Random multiplicative processes: An elementary tutorial. Am. J. Physiol. *58*, 267–273.

Reza, F.M. (2012). An Introduction to Information Theory (Dover Publications).

Siddall, M.E., Brooks, D.R., and Desser, S.S. (1993). Phylogeny and the reversibility of parasitism. Evolution *47*, 308–313.

Silverman, B.W. (1987). Density Estimation for Statistics and Data Analysis (Chapman and Hall/CRC).

Stefka, A.T., Feehley, T., Tripathi, P., Qiu, J., McCoy, K., Mazmanian, S.K., Tjota, M.Y., Seo, G.-Y., Cao, S., Theriault, B.R., et al. (2014). Commensal bacteria protect against food allergen sensitization. Proc. Natl. Acad. Sci. USA *111*, 13145–13150.

Suchodolski, J.S., Markel, M.E., Garcia-Mazcorro, J.F., Unterer, S., Heilmann, R.M., Dowd, S.E., Kachroo, P., Ivanov, I., Minamoto, Y., Dillman, E.M., et al. (2012). The fecal microbiome in dogs with acute diarrhea and idiopathic inflammatory bowel disease. PLoS ONE *7*, e51907.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.

Wang, J., Linnenbrink, M., Künzel, S., Fernandes, R., Nadeau, M.J., Rosenstiel, P., and Baines, J.F. (2014). Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. Proc. Natl. Acad. Sci. USA *111*, E2703–E2710.

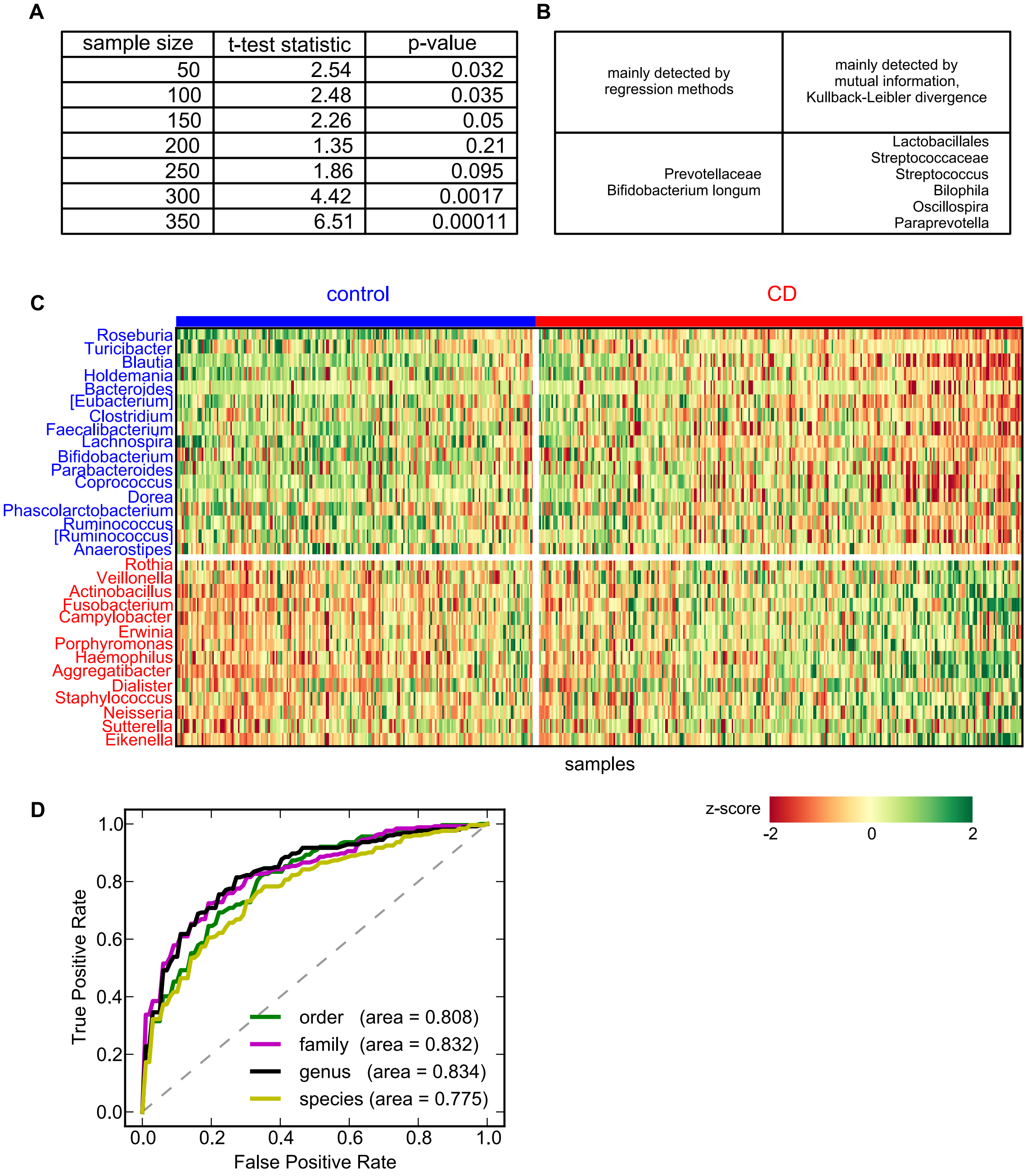
White, J.R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. *5*, e1000352.

Cell Reports

Supplemental Information

Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota

Feng Wang, Jess L. Kaplan, Benjamin D. Gold, Manoj K. Bhasin, Naomi L. Ward, Richard Kellermayer, Barbara S. Kirschner, Melvin B. Heyman, Scot E. Dowd, Stephen B. Cox, Haluk Dogan, Blaire Steven, George D. Ferry, Stanley A. Cohen, Robert N. Baldassano, Christopher J. Moran, Elizabeth A. Garnett, Lauren Drake, Hasan H. Otu, Leonid A. Mirny, Towia A. Libermann, Harland S. Winter, and Kirill S. Korolev



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 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.

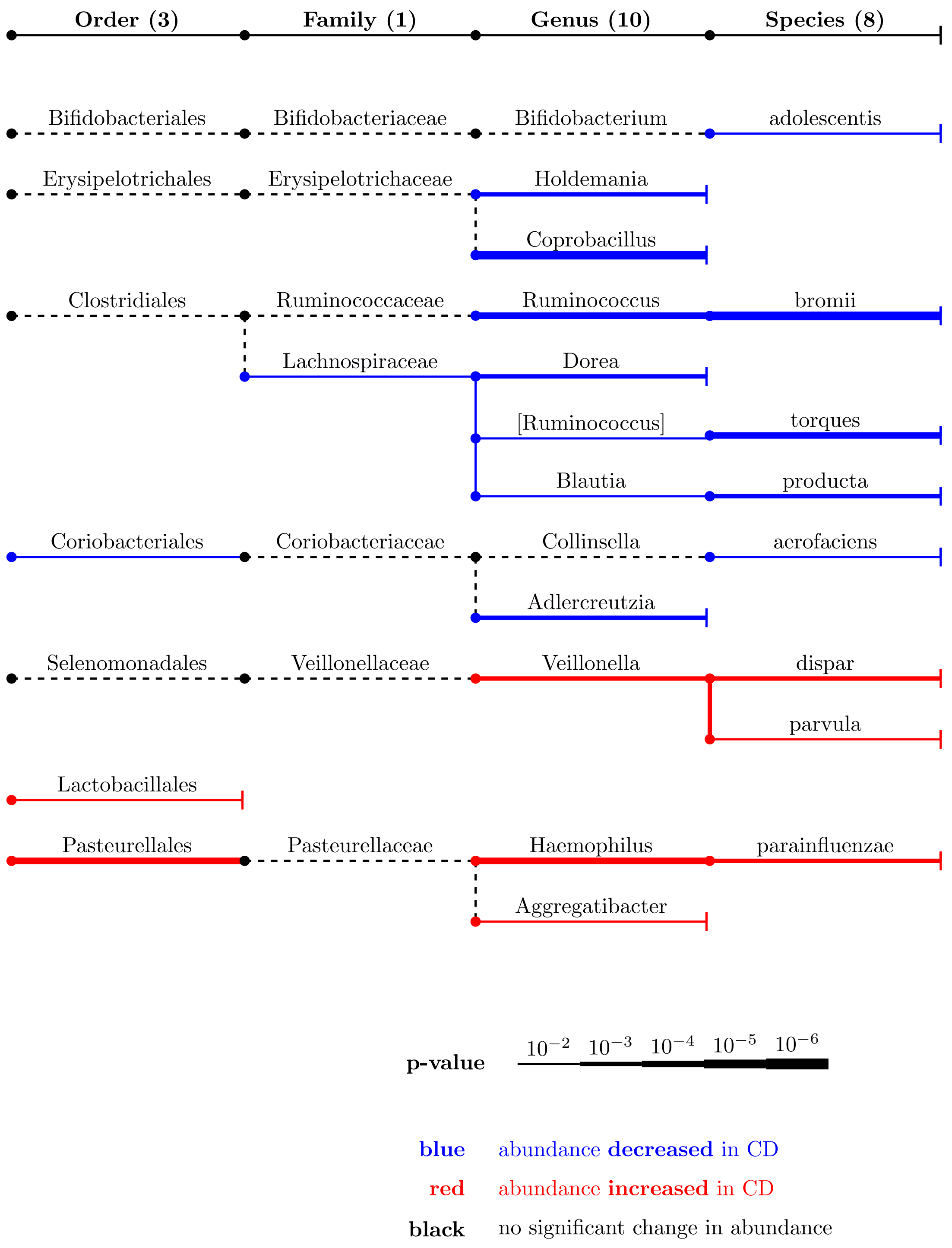
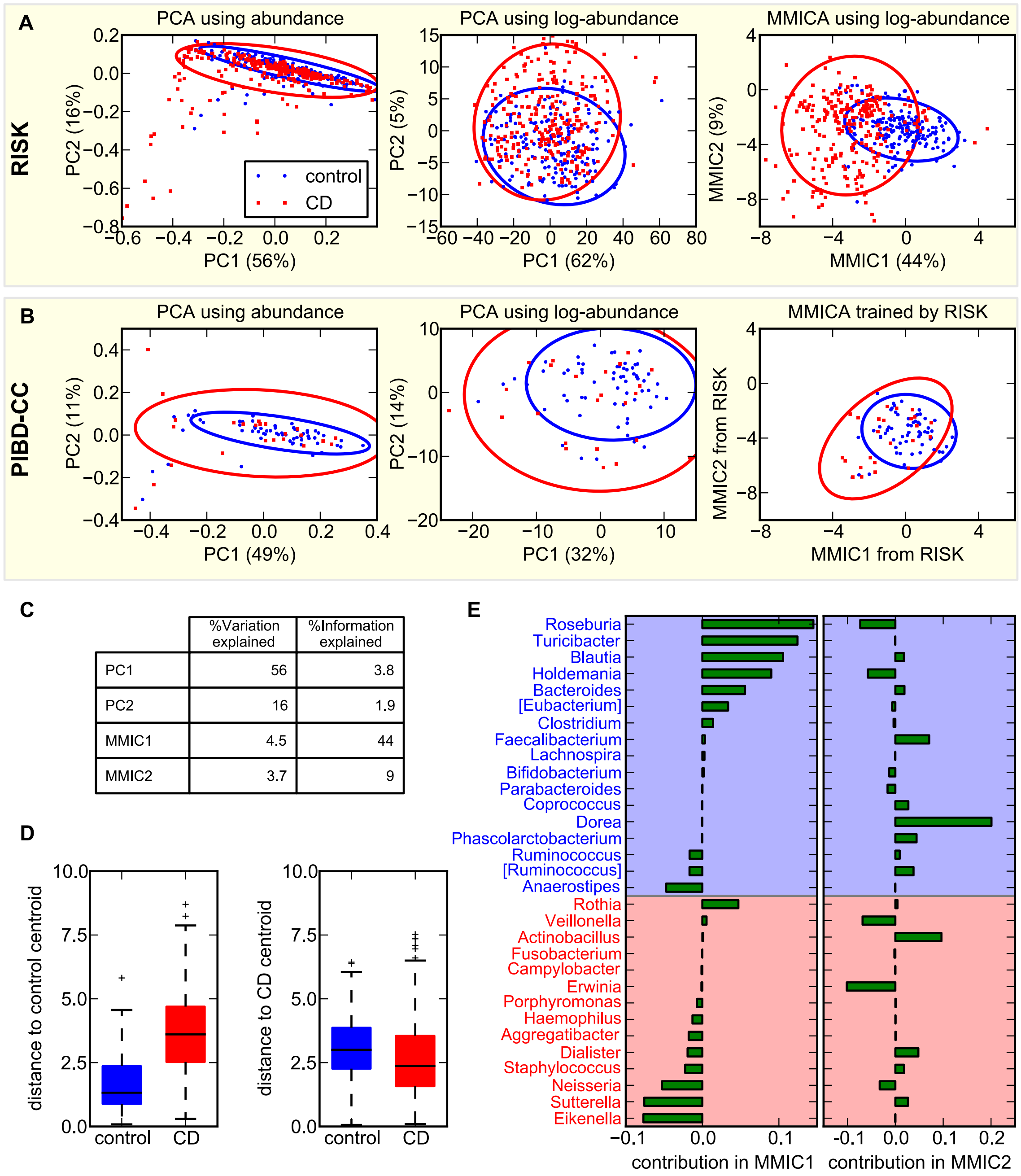


Figure S2. Related to Figure 4.
Associations with health and CD discovered in stool samples from the RISK cohort are shown across phylogenetic levels. There are both similarities and differences compared to the Figure 4, which is based on ileal samples.



	%Variation explained	%Information explained
PC1	56	3.8
PC2	16	1.9
MMIC1	4.5	44
MMIC2	3.7	9

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 in the RISK cohort. 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 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%)