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

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1016/j.celrep.2015.12.088">http://dx.doi.org/10.1016/j.celrep.2015.12.088</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Elsevier</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Fri Feb 01 04:18:52 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/101707">http://hdl.handle.net/1721.1/101707</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-nd/4.0/">http://creativecommons.org/licenses/by-nc-nd/4.0/</a></td>
</tr>
</tbody>
</table>
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.
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 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. Korolev1,13,15,*

1Bioinformatics Graduate Program, Boston University, Boston, MA 02215, USA
2Department of Pediatrics, Massachusetts General Hospital for Children, Harvard Medical School, Boston, MA 02114, USA
3Children’s Healthcare of Atlanta, LLC; GI Care for Kids, LLC; Atlanta, GA 30342, USA
4BIDMC 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
5Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA
6Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
7Department of Pediatrics, University of Chicago Comer Children’s Hospital, Chicago, IL 60637, USA
8Department of Pediatrics, University of California, San Francisco, San Francisco, CA 94143, USA
9Molecular Research MR DNA, Shallowater, TX 79363, USA
10Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA
11Division of Gastroenterology, Hepatology, and Nutrition, The Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA
12Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
13Department of Physics, Boston University, Boston, MA 02215, USA
14Co-first author
15Co-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
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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...
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-naïve 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-naïve 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-naïve 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 1. High Variability of Bacterial Abundances in the Human Gut Microbiota](image-url)
Our main observation in Figure 1 is that the variation of species abundance, and high variability are needed to analyze changes in microbial communities. Hence, methods that can manage with both small sample sizes and high variability are needed to analyze changes in microbial communities. The large difference in mean 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 Bacteroides.

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).

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, Kulback-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
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...
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 (Sudo 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*, *Bacteroides*, and *F. prausnitzii* produce butyrate, which acts as an energy source for epithelial cells in the gut (Duncan et al., 2002). Supple-

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.
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.
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. (2013) 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 0.66–0.77), 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).

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 MMIC, 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,
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 (Alenberg 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, 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 Egliston 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.

** риск**
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’ GTNTTACNGC GGKGGCTG, 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.

** риск**
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, Kulback-Leibler divergence, L2-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 rule of thumb (Silverman, 1987).

**Statistical Significance**
Statistical significance was evaluated by a permutation test with 10⁶ 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**

**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


