PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Interactions between species introduce spurious associations in microbiome studies

Rajita Menon[1], Vivek Ramanan[2,3], Kirill S. Korolev[1,4]*

1 Department of Physics, Boston University, Boston, Massachusetts, United States of America, **2** BRITE Bioinformatics REU Program, Boston University, Boston, Massachusetts, United States of America, **3** Department of Biology and Computer Science, Swarthmore College, Swarthmore, Pennsylvania, United States of America, **4** Graduate Program in Bioinformatics, Boston University, Boston, Massachusetts, United States of America

* korolev@bu.edu

Citation: Menon R, Ramanan V, Korolev KS (2018) Interactions between species introduce spurious associations in microbiome studies. PLoS Comput Biol 14(1): e1005939. https://doi.org/10.1371/journal.pcbi.1005939

Editor: Stefano Allesina, University of Chicago, UNITED STATES

Received: September 27, 2017

Accepted: December 21, 2017

Published: January 16, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by a grant from the Simons Foundation (#409704, KSK) and by the startup fund from Boston University to KSK. VR was also supported by NSF grant DBI-1559829 through BRITE Bioinformatics REU Program at Boston University. RM was partly supported by the Graduate Fellowship from the Rafik B. Hariri Institute for Computing and Computational Sciences & Engineering. The funders had no role in

Abstract

Microbiota contribute to many dimensions of host phenotype, including disease. To link specific microbes to specific phenotypes, microbiome-wide association studies compare microbial abundances between two groups of samples. Abundance differences, however, reflect not only direct associations with the phenotype, but also indirect effects due to microbial interactions. We found that microbial interactions could easily generate a large number of spurious associations that provide no mechanistic insight. Using techniques from statistical physics, we developed a method to remove indirect associations and applied it to the largest dataset on pediatric inflammatory bowel disease. Our method corrected the inflation of p-values in standard association tests and showed that only a small subset of associations is directly linked to the disease. Direct associations had a much higher accuracy in separating cases from controls and pointed to immunomodulation, butyrate production, and the brain-gut axis as important factors in the inflammatory bowel disease.

Author summary

Microbiomes are important for better health, sustainable agriculture, and climate management. Since experimental studies of natural microbial communities are often challenging, microbiome wide association studies (MWAS) have been the primary method to reveal the connection between specific microbes and host phenotype. MWAS have established that many diseases are associated with a complex dysbiosis driven by a large number of microbes. We show that many of these associations may not reflect a mechanistic link with the disease, but arise instead due to interspecific interactions such as cross-feeding and competition for resources. We also propose a method grounded in the maximum entropy models of statistical physics to separate direct from indirect associations. Using both synthetic and real microbiome data, we show that this method detects only direct associations, identifies most predictive features of microbiomes, and avoids p-value inflation. We demonstrate the power of this method on one of the largest microbiome data sets on inflammatory bowel disease.

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Microbes are essential to any ecosystem be it the ocean or the human gut. The sheer impact of microbial processes has however been underappreciated until the advent of culture-independent methods to assess entire communities *in situ*. Metagenomics and 16S rRNA sequencing identified significant differences in microbiota among hosts, and experimental manipulations established that microbes could dramatically alter host phenotype [1–8]. Indeed, anxiety, obesity, colitis, and other phenotypes can be transmitted between hosts simply by transplanting their intestinal flora [9–13].

New tools and greater awareness of microbiota triggered a wave of association studies between microbiomes and host phenotypes. Microbiome wide association studies (MWAS) have been carried out for diabetes, arthritis, cancer, autism and many other disorders [14–23]. MWAS clearly established that each disease is associated with a distinct state of intestinal dysbiosis, but they often produced conflicting results and identified a very large number of associations both within and across studies [14, 19, 21, 23–26]. For example, a recent study on inflammatory bowel disease (IBD) reported close to 100 taxa associated with IBD [25], a number that is fairly typical [14]. Such long lists of associations defy simple interpretation and complicate mechanistic follow-up studies because one needs to examine the role of almost every species in the microbiota. In fact, one can argue that MWAS are most useful when they can identify a small network of taxa driving the disease.

Although extensive dysbiosis might reflect the multifactorial nature of the disease, it is also possible that MWAS detect spurious associations because their statistical methods fail to account for some important aspects of microbiome dynamics. One such aspect is the pervasive nature of microbial interactions: species compete for similar resources, rely on cross-feeding for survival, and even produce their own antibiotics [27–37]. Hence, microbial abundances must be correlated with each other, and even a simple change in host phenotype could manifest as collective responses by the microbiota. Traditional MWAS, however, completely neglect this possibility because they treat each species as an independent manifestation of host phenotype. As a result, MWAS cannot distinguish taxa directly linked to disease from taxa that are affected only through their interactions with other species.

The main conclusion of this paper is that realistic microbial interactions produce a large number of spurious associations between particular members of the microbiome and phenotypes. Many of these indirect associations can be removed by a simple procedure based on maximum entropy models from statistical physics [38, 39]. We dubbed this approach Direct Association Analysis, or DAA for short.

When applied to the largest MWAS on IBD, DAA shows that many of the previously reported associations could be explained by interspecific interactions rather than the disease. At the genus and species level, the direct associations include only *Roseburia*, *Faecalibacterium prausnitzii*, *Bifidobacterium adolescentis*, *Blautia producta*, *Turicibacter*, *Oscillospira*, *Eubacterium dolichum*, *Aggregatibacter segnis*, and *Sutterella*. Some of these associations are well-known [40–47], while others have received little attention in IBD research. The phenotypes of the taxa directly linked to disease suggest that immunomodulation, butyrate production, and the brain-gut interactions play an important role in the etiology of IBD.

Compared to traditional MWAS, DAA corrected the inflation of p-values responsible for the large number of spurious associations and identified taxa most informative of the diagnosis. We found that directly associated taxa are much better at discriminating between cases and controls than an equally-sized subset of indirect associations. In fact, direct associations have the same potential to discriminate between health and disease as the entire set of almost a hundred associations detected by conventional methods.

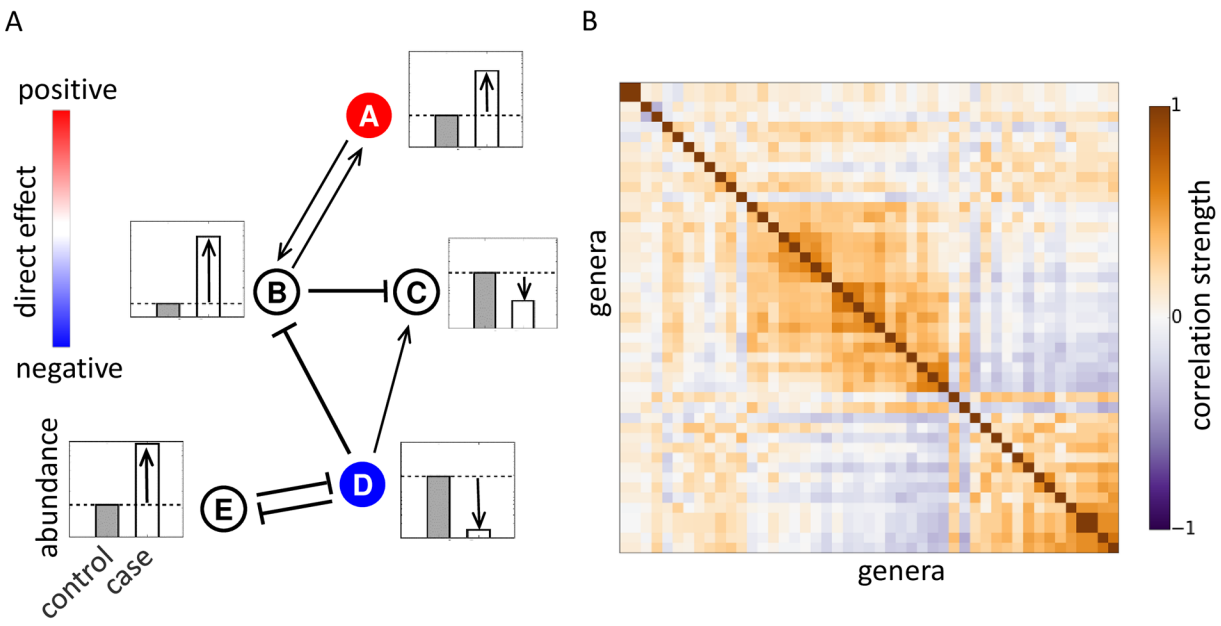

Fig 1. Microbial interactions generate spurious associations. (**A**) A hypothetical interaction network of five species together with their dynamics in disease. Only two species (shown in color) are directly linked to host phenotype. These directly-linked species inhibit or promote the growth of the other members of the community (shown with arrows). As a result, all five species have different abundances between case and control groups. (**B**) Microbial interactions are visualized via a hierarchically-clustered correlation matrix computed from the data in Ref. [21]. We used Pearson's correlation coefficient between log-transformed abundances to quantify the strength of co-occurrence for each genus pair. Dark regions reflect strong interspecific interactions that could potentially generate spurious associations. See S1 Text for the list of 47 most prevalent genera included in the plot.

https://doi.org/10.1371/journal.pcbi.1005939.g001

Results

Traditional MWAS detect species with significantly different abundances between case and control groups. Some changes in the abundances are directly associated with the disease while others are due to microbial interactions. The emergence of indirect changes in abundance is illustrated in Fig 1A for a hypothetical network of five species. Only two species A and D are directly linked to the disease. However, strong interactions make the abundances of all five species differ between control and disease groups. For example, the mutualistic interaction between A and B helps B grow to a higher density following the increase in the abundance of A. The expansion of B in turn inhibits the growth of C and reduces its abundance in disease. Strong mutualistic, competitive, commensal, and parasitic interactions have been demonstrated in microbiota [27–37], and Fig 1B shows that almost every species present in the human gut participates in a strong interaction. Thus, the propagation of abundance changes from directly-linked to other species could pose a significant challenge for MWAS. To test this hypothesis, we turned to a minimal mathematical model of microbiota composition.

Maximum entropy model of microbiota composition

A quantitative description of interspecific interactions and their effect on MWAS requires a statistical model of host-associated microbial communities. Ideally, such a model would describe the probability to observe any microbial composition, but the amount of data even in large studies is only sufficient to determine the means and covariances of microbial abundances. This situation is common in the analysis of biological data and has been successfully managed with the use of maximum entropy distributions [38]. These distributions are chosen

to be as random as possible under the constraints imposed by the first and second moments. Maximum entropy models introduce the least amount of bias and reflect the tendency of natural systems to maximize their entropy [48, 49]. In other contexts, these models have successfully described the dynamics of neurons, forests, flocks, and even predicted protein structure and function [50–54]. In the context of microbiomes, a recent work derived a maximum entropy distribution for microbial abundances using the principle of maximum diversity [55].

We show in S1 Text that the maximum entropy distribution of microbial abundances $P(\{l_i\})$ takes the following form

$$P(\{l_i\}) = \frac{1}{Z} e^{\sum_i h_i l_i + \frac{1}{2}\sum_{ij} J_{ij} l_i l_j} \quad (1)$$

where l_i is the log-transformed abundance of species i, h_i represents the direct effect of the host phenotype on species i, and J_{ij} describes the interaction between species i and j; the factor of $1/Z$ is the normalization constant. The log-transformation of relative abundances alleviates two common difficulties with the analysis of the microbiome data. The first difficulty is the large subject-to-subject variation, which is much better captured by a log-normal rather than a Gaussian distribution; see S1 Fig, and Ref. [25]. The second difficulty arises from the fact that the relative abundances must add up to one. This constraint is commonly known as the compositional bias because it leads to artifacts in the statistical analysis [56–58]. The log-transformation is an essential first step in most methods that account for the compositional bias including the widely advocated log-ratio transformation [56–59], which includes additional steps that are not relevant in the context of Eq (1). In S1 Text, we generalize Eq (1) to account for the constraint imposed by data normalization and show that our conclusions are not affected by the compositional bias.

The key prediction of Eq (1), see S1 Text, is that h and mean microbial abundances $m_i = \langle l_i \rangle$ are related by $m = J^{-1}h$. Because of interspecific interactions, J is not diagonal, and, therefore, a change in one component of h affects the abundances of many species. We show below that this nontrivial cause-effect relationship gives rise to spurious associations in both synthetic and real microbiome data.

Testing for spurious associations in synthetic data

We obtained realistic model parameters from one of the largest case-control studies previously reported in Ref. [21]. The samples were obtained from mucosal biopsies of 275 newly diagnosed, treatment-naive children with Crohn's disease (a subtype of IBD) and 189 matched controls. Microbiota composition was determined by 16S rRNA sequencing with about 30,000 reads per sample. From this data, we inferred the interaction matrix J and the typical changes in microbial abundances associated with the disease for 47 most prevalent genera (Methods and S1 Text). Even though the number of data points significantly exceeds the number of free parameters in the model, overfitting could still be a potential concern. However, overfitting is unlikely to affect our main conclusions because they depend only on the overall statistical properties of J rather than on the precise knowledge of every interaction. In fact, none of our results changed when we analyzed only about half of the data set (Fig 2 and S12 Fig). To improve the quality and robustness of the inference procedure, we also used the spectral decomposition of J to remove any interaction patterns that were not strongly supported by the data; see Methods and S1 Text for further details.

To determine the effect of microbial interactions on conventional MWAS analysis, we generated synthetic data with a known number of direct associations. The data for the control group was used without modification from Ref. [21]. The disease group was generated using

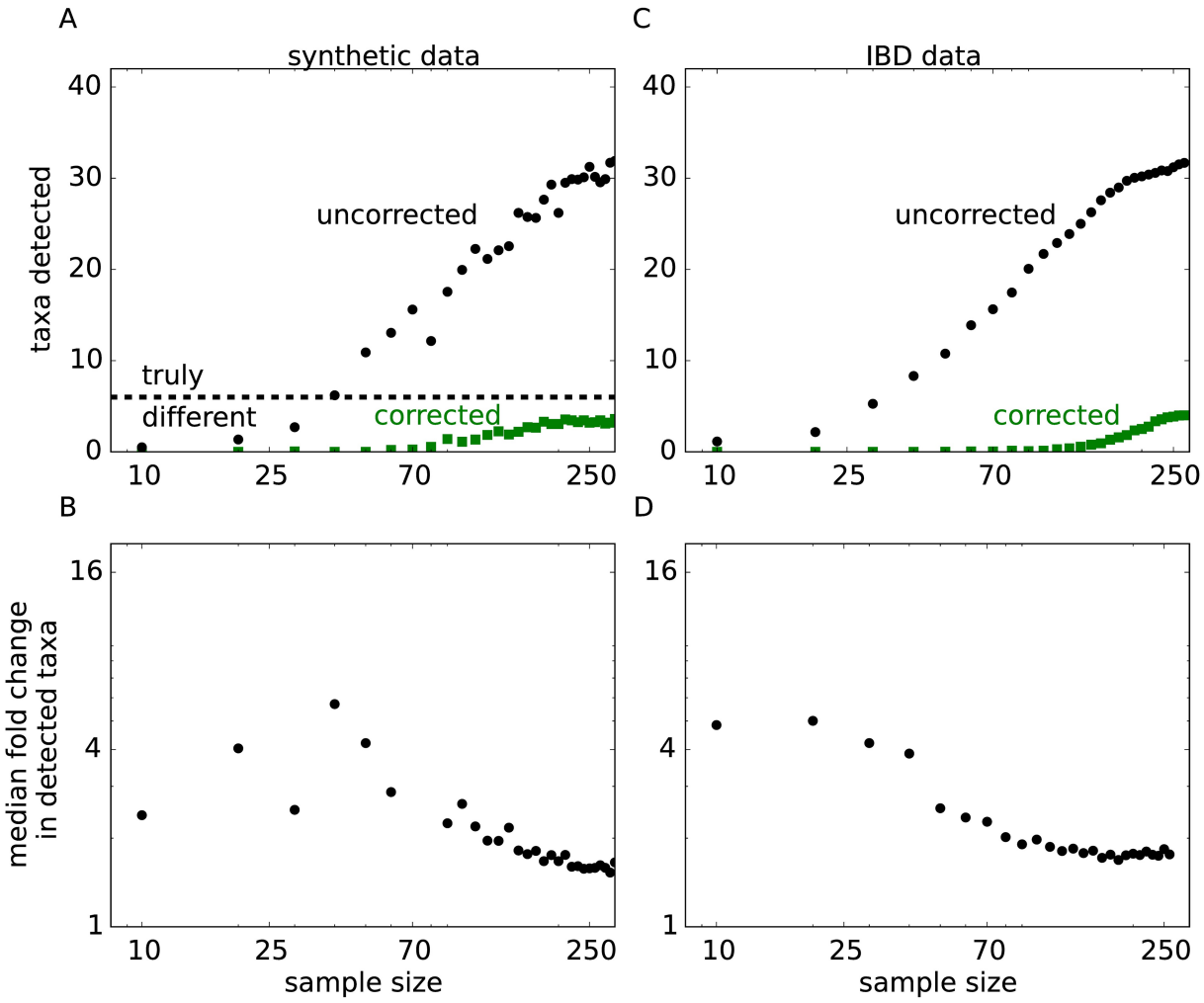

Fig 2. Signatures of indirect associations in synthetic and IBD data sets. The synthetic data set was generated to match the statistical properties of the IBD data set from Ref. [21], but with a predefined number of 6 directly associated taxa (See S1 Text). (**A**) In synthetic data, DAA identifies no spurious association and detects 4 out of 6 directly associated genera. All 6 genera and no false positives are detected when the sample size is increased further (S9 Fig). In sharp contrast, a large number of spurious associations is observed for metrics that rely on changes in abundance between cases and controls and do not correct for microbial interactions. The number of false positives grows rapidly with statistical power until all taxa are reported as significantly associated with the disease. (**B**) All spurious associations show substantial differences between cases and controls and, therefore, cannot be discarded based on their effect sizes. To quantify the effect size, we estimated the magnitude of the fold change for each genus. Specifically, we first computed the difference in the mean log-abundance between cases and controls and then exponentiated the absolute value of this difference. The plot shows how the median effect size for significantly associated genera depends on the sample size. Larger samples sizes result in much higher number of associations, but only a small drop in the typical effect size. (**C**) and (**D**) are the same as (A) and (B), but for the IBD data set. The results are consistent between the two data sets suggesting that most associations detected by traditional MWAS are spurious. The complete list of indirect associations inferred from the IBD data set is shown in S1 Text, and the results for different synthetic data sets are shown in S14 Fig.

https://doi.org/10.1371/journal.pcbi.1005939.g002

Eq (1) with the same values of h and J as in the control group, except we modified the values of h for 6 representative genera (see S1 Text). We also generated two other synthetic data sets with smaller and larger effect sizes. The results for all three data sets were very similar (S1 Text).

The synthetic data was further subsampled to several sample sizes in order to simulate variation in statistical power between different studies. For an ideal method, the number of detected associations should increase with the cohort size, but eventually saturate once all 6

directly associated genera are discovered. In contrast to this expectation, the number of associations detected by the conventional approach increased rapidly with the sample size until almost all genera were found to be statistically associated with the disease in our synthetic data. At this point, traditional MWAS completely lost the power to identify the link between the phenotype and microbiota. Unbounded growth in the number of detections was also observed for the real data (Fig 2C) suggesting that many previously reported associations between microbiota and IBD could be indirect.

Are spurious associations simply an artifact of our ability to detect even minute differences between cases and controls? Fig 2B and 2D show that this was not the case. The median effect size declined only moderately with the number of associations, and most associations corresponded to about a factor of two difference in the taxon abundance. Thus, spurious associations are not weak and could not be discarded based on their effect size.

Direct association analysis (DAA)

Fortunately, the maximum entropy model provides a straightforward way to separate direct from indirect associations. Since direct effects are encoded in h, MWAS should be performed on h rather than on l. This simple change in the statistical analysis correctly recovered 4 out of 6 directly associated taxa in the synthetic data and yielded no indirect associations even for large cohorts (Fig 2A and S9 Fig). Similarly good performance was found for the two other synthetic data sets (S14 Fig). For the IBD data, DAA also identified a much smaller number of associations compared to traditional MWAS analysis and showed clear saturation at large sample sizes (Fig 2B). Direct associations with IBD are summarized in Fig 3 at the genus and species levels, and the entire phylogenetic tree of direct associations is shown in S4 Fig and in S1 Text.

In addition to associations, DAA also infers the network of direct microbial interactions (Fig 3, S5 and S6 Figs). While the sample size is insufficient to accurately infer the interactions between every pair of microbes, strong interactions and the overall properties of the interaction network can nevertheless be determined from the data. The interactions inferred by DAA describe only direct effects of the species on each other and do not include induced correlations present in the correlation matrix. That is, DAA controls for the fact that species A and C could be correlated because both interact with species B, but not with each other (Fig 1A). The ability of maximum entropy models to separate direct from indirect interactions has been the primary reason for their applications to biological data [50–54]. Similar to these previous studies, many direct interactions reported in Fig 3 are also present in the correlation-based network, but DAA removes some induced interactions and identifies a few interactions that are not evident in the correlation data; see S5 and S6 Figs. Overall, the interaction network is much sparser than the correlation network in Fig 1B. In S1 Text, we also compare the results from DAA and SparCC [56], a widely used package to infer correlation networks from microbiome data (S6 Fig).

To demonstrate that DAA isolates direct effects from collective changes in the microbiota, we examined the p-value distribution in this method. The distribution of p-values is commonly used as a diagnostic tool to test whether a statistical method is appropriate for the data. In the absence of any associations, p-values must follow a uniform distribution because the null hypothesis is true [60]. A few strong deviations from the uniform distribution signal true associations [61]. In contrast, large departures from the uniform distribution typically indicate that the statistical method does not account for some properties of the data, for example, population stratification in the context of genome wide association studies [62, 63]. Fig 4A compares the distribution of p-values for DAA and a conventional method in MWAS. Consistent

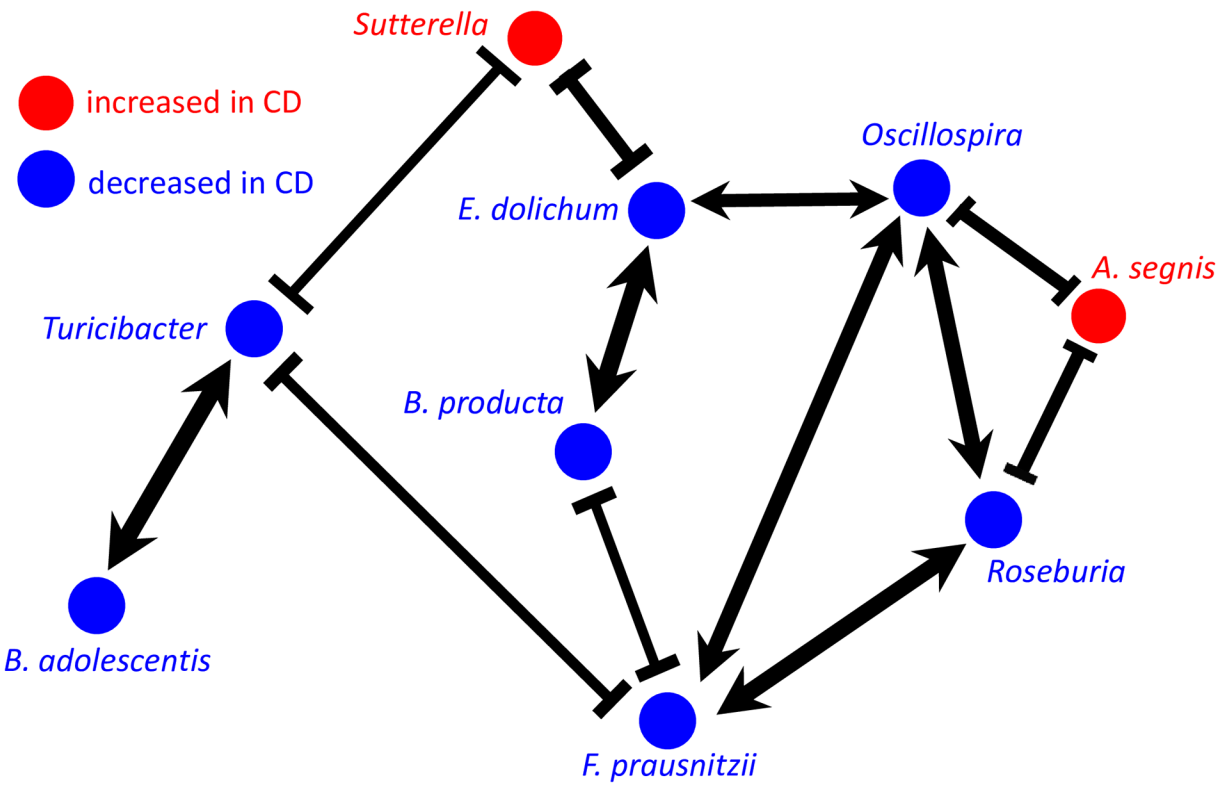

Fig 3. Network of direct associations with Crohn's Disease. Five species and four genera were found to be significantly associated with Crohn's Disease ($q < 0.05$) after correcting for microbial interactions (S1 and S4 Figs). The links correspond to significant interactions ($q < 0.05$) between the taxa with $J_{ij} > 0.27$ or $J_{ij} < -0.15$; the width of the arrows reflects the strength of the interactions. For comparison, the correlation-based network for directly associated taxa is shown in S7 and S5 Figs, and a complete summary of correlations and interactions for all species pairs is provided in S1 Text.

https://doi.org/10.1371/journal.pcbi.1005939.g003

with our hypothesis that interspecific interactions cannot be neglected, conventional analysis generates an excess of low p-values and, as a result, a large number of potentially indirect associations. In contrast, the distribution of p-values from DAA matches the expected uniform distribution and, thus, provides strong support for our method.

Discussion

The primary goal of MWAS is to guide the study of disease etiology by detecting microbes that have a direct effect on the host. These direct effects could be very diverse and include secretion of toxins, production of nutrients, stimulation of the immune system, and changes in mucus and bile [67, 68]. In addition to the host-microbe interactions, the composition of microbiota is also influenced by the interspecific interactions among the microbes such as competition for resources, cross-feeding, and production of antibiotics [27–37]. In the context of MWAS, microbial interactions contribute to indirect changes in microbial abundances, which are less informative of the disease mechanism and are less likely to be valuable for follow-up studies or in interventions. Here, we estimated the relative contribution of indirect associations to MWAS and showed how to isolate direct from indirect associations.

Our main result is that interspecific interactions are sufficiently strong to generate detectable changes in the abundance of many microbes that are not directly linked to host phenotype. As a result, conventional approaches to MWAS detect a large number of spurious associations and produce inflated p-values that do not match their expected distribution

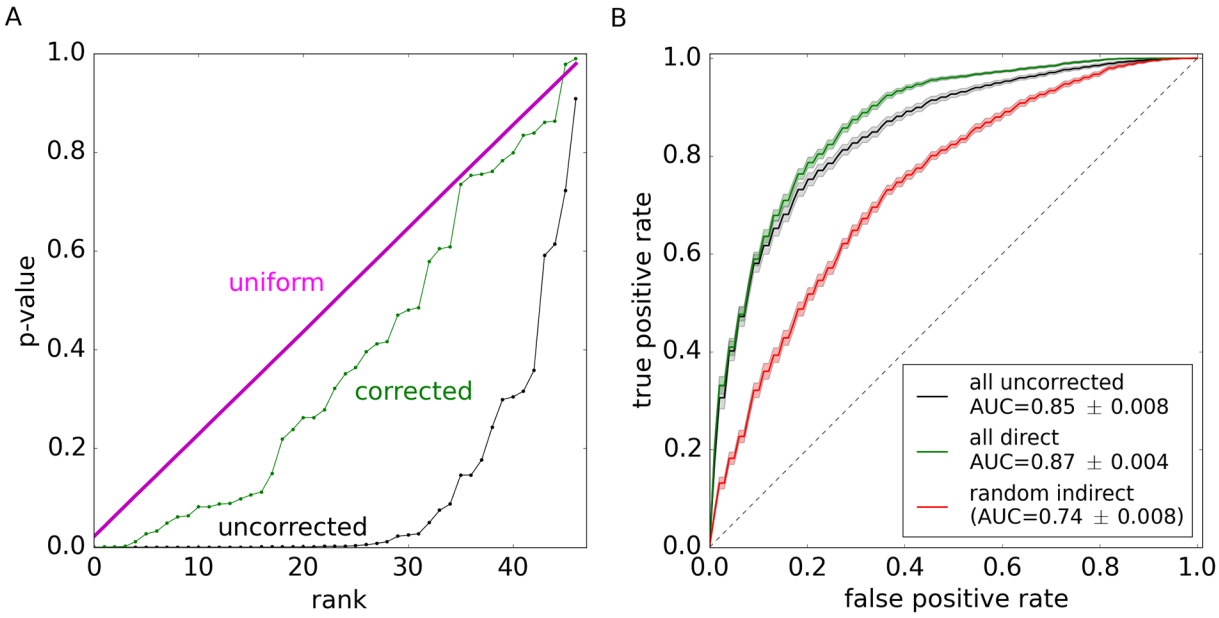

Fig 4. Direct associations analysis corrects p-value inflation and retains diagnostic accuracy. (**A**) The distribution of p-values in DAA closely follows the expected uniform distribution. Because conventional MWAS does not correct for microbial interactions, it yields an excess of low p-values, which is a strong signature of indirect associations. For both methods, p-values were computed using a permutation test. The expected uniform distribution was obtained by sampling from a generator of uniform random numbers. The ranked plot of p-values visualizes their cumulative distribution functions; this is a variant of a Q-Q plot. (**B**) Direct associations are a small subset of all associations with IBD (see S4 Fig), yet they retain full power in classifying samples as cases or controls. In contrast, the classification power is substantially reduced for an equally-sized subset of randomly-chosen indirect associations. In each case, we used sparse logistic regression to train a classifier on 80% of the data and tested its performance on the remaining 20% (Methods). The shaded regions show one standard deviation obtained by repeated partitioning the data into training and validation sets. Identical results were obtained with a random forest [64, 65] and support vector machine [66] classifiers (S8 Fig)

https://doi.org/10.1371/journal.pcbi.1005939.g004

(Fig 4A). These challenges are resolved by Direct Association Analysis (DAA), which uses maximum entropy models to explicitly account for interspecific interactions. We applied DAA to a large data set of pediatric Crohn's disease and found that it restores the distribution of p-values and substantially simplifies the pattern of dysbiosis while retaining full classification power of a conventional MWAS.

The relatively simple dysbiosis identified by DAA in IBD has strong support in the literature and offers interesting insights into disease etiology. Four of the taxa identified by our method have a well-established role in IBD: *B. adolescentis*, *F. prausnitzii*, *B. producta*, and *Roseburia*. They have been repeatedly found to have lower abundance in both Crohn's disease and ulcerative colitis [40–47], and several studies have demonstrated their ability to suppress inflammation and alleviate colitis [43, 69–73]. *Bifidobacterium* species occupy a low trophic level in the gut and ferment complex polysaccharides such as fiber [74, 75]. Fermentation products include lactic acid, which promotes barrier function, and maintains a healthy, slightly acidic environment in the colon [76]. Due to these properties *Bifidobacterium* species are commonly used as probiotics [74]. *F. prausnitzii*, *Blautia producta* and *Roseburia* occupy a higher trophic level and ferment the byproducts of polysaccharides digestion into short-chain fatty acids (SCFA), which are an important energy source for the host [42, 43, 77, 78].

The ability of DAA to detect taxa strongly associated with IBD is reassuring, but not surprising. What is surprising is that many strong associations are classified as indirect by our method. For example, *Roseburia* and *Blautia* are the only genera of *Lachnospiraceae* that DAA finds to be directly linked to the disease. In sharp contrast, traditional MWAS report seven

genera in this family that are strongly associated with IBD [25]. All seven genera are involved in SCFA metabolism, but their specializations differ. Species in *Blautia* genus are major producers of acetate, a SCFA that is commonly involved in microbial crossfeeding [79, 80]. In particular, many species extract energy from acetate by converting it into butyrate, another SCFA that plays a major role in gut health by nourishing colonocytes and regulating the immune function [77, 80]. *Roseburia* genus specializes almost exclusively in the production of butyrate and acts as a major source of butyrate for the host [77, 81]. Thus, our findings suggest that butyrate production plays an important role in IBD etiology and that the dysregulation of this process is directly linked to the depletion of *Roseburia* and possibly *Blautia*.

The important role of butyrate is further supported by our detection of *E. dolichum* and *Oscillospira*, which are known to produce butyrate [82–84]. The latter taxon has not been detected in three independent analyses of this IBD data set [21, 25, 85] presumably because its involvement is masked by indirect associations and interactions with other microbes. Several other studies support this DAA finding and confirm that *Oscillospira* is suppressed in IBD [86, 87]. *Oscillospira* was also found to be positively associated with leanness and negatively associated with the inflammatory liver disease [88–90]. The interactions between *Oscillospira* and the host appears to be quite complex and involve the consumption of host-derived glycoproteins including mucin, production of SCFA, and modulation of bile-acid metabolism [84, 91]. The latter interaction was suggested to be a major factor in the protective role of *Oscillospira* against infections with *Clostridium difficile* [91–93].

The final taxon that was suppressed in IBD is *Turicibacter*. This genus is not very well characterized, and few MWAS studies point to its involvement in IBD [21, 25, 94]. Two studies in animal models, however, directly looked into the connection between IBD and *Turicibacter* [95, 96]. The first study found that iron limitation eliminates colitis in mice while at the same time restoring the abundance of *Turicibacter*, *Bifidobacterium*, and four other genera [95]. The second study identified *Turicibacter* as the only genus that is fully correlated with immunological differences between mice resistant and susceptible to colitis: high abundance of *Turicibacter* in the colon predicted high levels of MZ B and iNK T cells, which are potent regulators of the immune response [96]. Moreover, *Turicibacter* was the only genus positively affected by the reduction in $CD8^+$ T cells. Thus, our method identified a taxon that is potentially directly linked to IBD via the modulation of the immune system.

Perhaps the most unexpected finding was our detection of *A. segnis* and *Sutterella* as the only species and genus increased in disease compared to 26 positive associations detected by the previous analysis [25]. All other associations were classified as indirect even though they often corresponded to much more significant changes in abundance between IBD and control groups. Thus, our results indicate that expansion of many taxa including opportunistic pathogens is driven by their interactions with the core IBD network shown in Fig 3. One possibility is that the dysbiosis of the symbiotic microbiota makes it less competitive against other bacteria and opens up niches that can be colonized by opportunistic pathogens. The other, less explored possibility, is that commensal microbiota can not only protect from pathogens, but also facilitate their invasion, a phenomenon that has been recently demonstrated in bees [97].

Little is known about the specific roles that *A. segnis* and *Sutterella* play in IBD, and more generally in gut health. *Aggregatibacter* is a common member of the oral microbiota that thrives in local infections such as periodontal disease and bacterial vaginosis [98–100]. The high abundance of *A. segnis* is also associated with an increased risk of IBD recurrence [101]. *Sutterella*, on the other hand lacks overt pathogenicity, and MWAS produced inconsistent findings [102–108] on its involvement in IBD. Some studies reported that *Sutterella* is increased in patients with good outcomes [21, 105] while other studies found positive or no association between *Sutterella* and IBD [25, 103, 106–108]. Experimental investigations

showed that *Sutterella* lacks many pathogenic properties; in particular, it does not induce a strong immune-response and has only moderate ability to adhere to mucus [107, 108]. Further, *Sutterella* strains from IBD and control patients showed no phenotypic differences in metabolomic, proteomic, and immune response assays [108]. Nevertheless, *Sutterella* is strongly associated with worse behavioral scores in children with autism spectrum disorder and Down syndrome [19, 20, 109]. Therefore, the direct link between *Sutterella* and IBD could involve the gut-brain axis.

In summary, we found a small number of taxa can explain extensive dysbiosis in IBD and accurately predict disease status. Directly associated taxa have strains with dramatically different abilities to trigger colitis and are specifically targeted by the immune system of patients and animals with IBD [12]. Previous studies of these taxa point to facilitated colonization by pathogens, butyrate production, immunomodulation, bile metabolism, and the gut-brain axis as the primary factors in the etiology of IBD.

Many disorders are accompanied by substantial changes in host microbiota, but our work shows that only a small subset of these changes could be directly related to the disease. Similarly, only a handful of taxa could drive the dynamics of the ecosystem-level changes in the environment. To untangle the complexity of such dysbioses, it is important to account for microbial interactions using mechanistic or statistical methods. Direct association analysis proposed in this paper is a simple statistical approach based on the principle of maximum entropy. DAA can be applied to any microbiome data set that is sufficiently large to infer interspecific interactions.

Methods

The data used in this study was obtained from Ref. [21], which reported changes in the microbiome of newly-diagnosed, treatment-naive children with IBD compared to controls. This data was recently analyzed in Ref. [25], and we followed all the statistical procedures adopted in that study to enable direct comparison of the results. Specifically, we used a permutation test on mean log-transformed abundances to determine the statistical significance of an association.

To fit the maximum entropy model to the data, we first computed the mean log-abundance for each genus m_i and the covariance in the log-transformed abundances C_{ij}. The interaction matrix was computed as $J = C^{-1}$ by performing singular value decomposition [110] and removing all singular values that were comparable to the amount of noise present in the data. The host effects were computed as $h = Jm$. See S1 Text for further details.

All computation was carried out in Python environment. We used `scikit-learn 0.15.2` [111] for hierarchical clustering and to build the supervised classifiers used in Fig 4B of the main text and S8 Fig. The variance in the accuracy of classification was evaluated through 5-fold stratified cross-validation with 100 random partitions of the data into the training and validation sets. For all findings, statistical significance was evaluated with Fisher's exact test (permutation test) with 10^6 permutations. False discovery rate was controlled to be below 5% following Benjamini-Hochberg procedure [60].

For sparse logistic regression, we confirmed that the penalty parameter was in the range where the results are insensitive to its specific value. The features selected by this classifier in Fig 4 are as follows: *Erysipelotrichales*, *Pasteurellales*, *Turicibacterales* (also significant in DAA), and *Enterobacteriales* (not significant in DAA) at the order level; *Clostridiaceae* and *Pasteurellaceae* (also significant in DAA) and *Enterobacteriaceae* and *Erysipelotrichaceae* (not significant in DAA) at the family level; *Roseburia* (also significant in DAA) and *Dialister*, *Aggregatibacter*, and *Haemophilus* (not significant in DAA) at the genus level; and *B. adolescentis*, *F. prausnitzii*,

and *E. dolichum* (also significant in DAA) and *Prevotella copri* and *Haemophilus parainfluenzae* (not significant in DAA) at the species level. In total, both DAA and the sparse logistic regression relied on 17 features with 9 of them being the same. Thus, DAA identified many features that were also selected by the machine learning algorithm for their predictive value. At the same time, the results of DAA and the sparse logistic regression were not exactly the same and, therefore, could be complementary to each other.

Supporting information

S1 Text. Supplementary text and tables. Derivation of mathematical model of community composition and inference of model parameters, discussion of assumptions and limitations of DAA.
(PDF)

S1 Fig. Microbial abundances follow the log-normal distribution. The histograms show probability distributions of the relative log-abundance for the species and genera detected by DAA (summarized in Fig 3). The best fit of a Gaussian distribution is shown in green.
(TIF)

S2 Fig. Pairwise interactions are sufficient to explain the patterns of microbial co-occurence. The parameters in our maximum entropy model were chosen to fit only the first and the second moments of the multivariate distribution of microbial abundances. Nevertheless, the model captures most of the higher-order correlations in the data suggesting pairwise interactions are sufficient to accurately describe the patterns of microbial co-occurences. (**A**) For each choice of three genera, the third order moment was computed by averaging the product of the log-abundances over all the samples in the IBD data ("observed") or from Eq. (17) ("predicted"), which states the predictions of the maximum entropy model. The plot shows excellent agreement between the two quantities. (**B**) For each choice of three genera ("index"), we plot the third-order central moment computed from the IBD data ("observed") and from an equally-sized sample drawn from our maximum entropy model ("Gaussian distribution"). The latter quantifies the expected deviations between the observations and predictions due to the finite size of the sample. (**C**) Same as (A), but for the fourth-order central moment. The expected level of noise is quantified via a sample from the maximum entropy model that obeys Eq. (17) exactly in the limit of infinite sample size. The correlation coefficient between "observed" and "predicted" values from this sample sets the upper bound on the expected correlation coefficient in IBD data.
(TIF)

S3 Fig. Microbial interactions are only weakly affected by host phenotype. To determine whether Crohn's disease drastically alters the pattern of microbial interactions, we computed and compared the covariance matrixes C^{CD} and C^{control} for CD and control groups respectively. The results of this calculation for IBD data are shown in blue. Each dot corresponds to a matrix element of C_{ij}, which is the covariance between the log-abundances of genera i and j. The x-coordinate is the covariance computed in the control group and the y-coordinate is the covariance computed in the CD group. To estimate the expected level of noise, we carried out the same analysis on two random partitions of the data that contain both controls and subjects with CD (shown in magenta). Since the groups are drawn from the same distribution, their covariance matrices must be identical on average. The spread of the magenta data points, therefore, sets the upper limit on the correlation coefficient between C^{CD} and C^{control}. We note, however, that this upper bound is unlikely to be reached for IBD data because some taxa have different noise levels in CD and control groups: eg. the taxa depleted in CD have a low

abundance in this group and, therefore, higher error in the estimates of the correlation coefficients with other taxa. Overall, both IBD and partitioned data lie close to the diagonal and exhibit similar levels of variation. Thus, using the same covariance matrix for both CD and control groups is a reasonable first approximation. This approximation is valuable because it reduces the uncertainty in C_{ij} by allowing us to use the entire data to compute covariances and because it improves the stability of DAA to errors in C (see S12 Fig).
(TIF)

S4 Fig. Taxa directly associated with Crohn's disease. Note that the Green Genes database [112] used in QIIME [113] places Turicibacter under Erysipelotrichales and has a unique order of Turicibacterales. This apparent inconsistency may reflect insufficient understanding of Turicibacter phylogeny. The effect sizes and statistical significance are summarized and results for DAA and conventional MWAS are compared in S1 Text.
(TIF)

S5 Fig. Comparison between correlations and direct interactions. The matrix of microbial interactions J is shown in (**A**) and the correlation matrix C is shown in (**B**), which is the same as Fig 1B of the main text. Both matrices are inferred from the IBD data set. Note that J is sparser than C. For greater clarity, the matrices are hierarchically clustered; therefore, the order of species in (A) and (B) is not the same.
(TIF)

S6 Fig. Comparison of networks inferred by Pearson correlation, SparCC, and DAA at the genus level. Three networks quantifying microbial co-occurrence or interactions have been inferred: one based on the Pearson correlation coefficient between log-abundances (which is closely related to the covariance matrix C), one using SparCC package from Ref. [56] that attempts to reduce compositional bias, and one based on the direct interactions J from DAA. In each network, we kept only links that were statistically different from 0 under a permutation test with 5% false discovery rate. The panels display Venn diagrams showing unique and overlapping links in these networks. All links are included in (**A**), and the comparison is done irrespective of the sign of the link, i.e. agreement is reported even if one method reports a positive link and another method reports a negative link. In contrast, (**B**) and (**C**) show only positive and negative links respectively. Three conclusions can be drawn from these comparisons. First, the high overlap between SparCC and Pearson networks shows that log-transforms have largely accounted for the compositional bias. Second, all three methods agree on a large number of links suggesting that all methods are sensitive to some strong interactions. Third, DAA reports fewer links and identifies a few links not detected by other methods. This reflect the different nature of DAA links. While both Pearson correlation and SparCC infer correlation, which could be either direct or indirect (i.e. induced; see main text). DAA removes indirect correlations, thus reducing the total number of links, but also reveals pairwise interactions that could have been masked by strong correlations with a third species.
(TIF)

S7 Fig. The network based on the correlation coefficient between log-transformed abundances. We plotted the correlation-based network for the species detected by DAA. Note the similarities and differences with the interaction network shown in Fig 3 of the main text. Only the links with the correlation coefficient greater than 0.27 or lower than -0.15 are shown, and all links are statistically significant ($q < 0.05$). All correlation coefficients and direct interactions are summarized in S1 Text for the genera and species detected by DAA.
(TIF)

S8 Fig. Direct associations retain full diagnostic power. The same as Fig 4B of the main text, but for two other classifiers: random forest [64, 65] in (**A**) and support vector machine [66] in (**B**).
(TIF)

S9 Fig. DAA detects all directly associated taxa in synthetic data, provided the sample size is sufficiently large. The same as Fig 2A in the main text, but with the x-axis extended to larger sample sizes. Note that DAA recovers all 6 directly associated taxa when the sample size is greater than about 1200.
(TIF)

S10 Fig. Compositional bias has a negligible effect on DAA performance. All panels are the same as Fig 2C in the main text, but with different normalization of the data prior to the analysis. (**A**) No normalization: the analysis is done on the counts from the OTU table, which do not add up to a constant number. (**B**) Total-sum scaling: The counts are converted into relative abundances by dividing by the total number of counts (reads) per sample. This plot is the same as Fig 2C. (**C**) Centered-log ratio: First log-abundances were computed from unnormalized counts with a pseudocount of 1. Then, the mean log-abundances of the taxa was computed by averaging over the samples. Finally, the mean-log abundance of every taxon was subtracted from the log-abundances of this taxon in all samples. This procedure corresponds to normalizing by the geometric mean of the counts because it ensures that the mean log-abundance of a taxon is zero [55]. (**D**) Cumulative sum scaling: A normalization scheme proposed specifically for microbiome analyses was implemented following Ref. [114]. The results of the analyses in (A)-(D) are very similar suggesting that compositional bias does not lead to major artifacts. In particular, the number of associations in (A) grows at the same rate with the sample size as in (B)-(D). This would not be the case if the compositional bias was strong because spurious associations due to normalization would lead to a greater number of detected taxa. Thus, we conclude that interspecific interactions rather than compositional effects are the primary source of spurious associations.
(TIF)

S11 Fig. The inference of the eigenvalues of the covariance matrix is robust to variation in sample size and bootstrapping. We repeatedly subsampled the IBD data set to half of its size and computed the eigenvalues of the covariance matrix C. The means and standard deviations from this bootstrap procedure are shown in green, and the eigenvalue inferred from the entire data are shown in black. The agreement between the different sample sizes and the small variation due to subsampling indicate that the spectral properties of C can be inferred quite accurately.
(TIF)

S12 Fig. Results of DAA are robust to variation in sample size and bootstrapping. Similar to S11 Fig, we repeatedly subsampled the IBD data set to half of its size and carried out DAA on each of the subsamples. (**A**) shows that there is a modest variation in inferred h. To a large extent, this variation is driven by the uncertainty in C and its inverse J. (**B**) shows a much smaller variation in Δh between control and CD groups (green symbols). The noise is reduced because, even though C changes from subsample to subsample, the same C is used to infer h for control and disease groups. Therefore, the variability in C has a much weaker effect on Δh. For comparison, we also show Δh obtained by bootstrapping the entire data set without preserving the diagnosis labels (black symbols). These data show the expected distribution of Δh under the null hypothesis of no associations. For genera detected by DAA, the black and the green error bars do not overlap suggesting that the results of DAA are not affected by the

uncertainty in C and are robust to variation in sample size and bootstrapping.
(TIF)

S13 Fig. Results of DAA are not significantly affected by compositional effects. The quantity Δh between control and CD groups is the test statistic used to infer direct associations, and the variation of Δh due to sampling shows whether the statistical analysis is robust to small changes in the data set. To quantify these variations in Δh, we consider a sample drawn from the maximum entropy model fitted to the IBD data set and define two $\delta\Delta h$: one between normalized and not normalized sample and the other between the not normalized sample and the values of h in the maximum entropy model. The first $\delta\Delta h$ quantifies the variability due to normalization, while the second $\delta\Delta h$ quantifies the variability due to sampling. The plot shows the distribution of the absolute values of the difference between the absolute values of these $\delta\Delta h$ across genera for three normalization schemes: total-sum scaling (TSS), centered-log ratio (CLR) and cumulative sum scaling (CSS). The absolute Δh values of significant taxa in IBD RISK data (red rectangles) lie well outside of the distributions shown.
(TIF)

S14 Fig. Spurious associations in synthetic data with small and large effect sizes. The same analysis as in Fig 2A and 2B of the main text, but for synthetic data with smaller (A, B, C) and larger (D, E, F) effect sizes. (**A**) and (**D**) show the number of associations detected by traditional MWAS and DAA. (**B**) and (**E**) show the median effect sizes (median fold change) for the taxa detected by conventional MWAS. (**C**) and (**E**) show the effect sizes in both h and l for the taxa detected by DAA. The effect size for h was quantified by the relative difference in h between cases and controls. The effect size for l was quantified as in (B) and (E). Overall the results are similar to those in Fig 2. In addition, (A) and (B) show that DAA can recover all directly associated taxa given a large number of samples without any false positives. For sample sizes exceeding 5000, DAA starts to detect indirect associations due to compositional effects.
(TIF)

S15 Fig. Sensitivity of DAA to eigenvalue threshold λ_{min}. Large λ_{min} retains only a few eigenvalues and imposes an artificially strong correlation structure on the data. As a result, DAA detects a large number of associations because it cannot distinguish direct from indirect effects. The performance of DAA improves as more eigenvalues are included and reaches a plateau. The dashed lines show the number of eigenvalues included for $\lambda_{min} = 0.01$ used throughout our analysis. The insets show the eigenvalues of Λ in decreasing order. The four panels show the results for different taxonomic levels: from species to order.
(TIF)

Author Contributions

Conceptualization: Rajita Menon, Kirill S. Korolev.

Data curation: Rajita Menon.

Formal analysis: Rajita Menon, Vivek Ramanan.

Funding acquisition: Kirill S. Korolev.

Methodology: Rajita Menon, Kirill S. Korolev.

Project administration: Kirill S. Korolev.

Software: Rajita Menon.

Validation: Rajita Menon.

Visualization: Rajita Menon.

Writing – original draft: Rajita Menon, Kirill S. Korolev.

Writing – review & editing: Rajita Menon, Kirill S. Korolev.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature. 2007; 449(7164):804. https://doi.org/10.1038/nature06244 PMID: 17943116
2. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012; 486(7402):222–227. https://doi.org/10.1038/nature11053 PMID: 22699611
3. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nature Reviews Genetics. 2012; 13(4):260–270. https://doi.org/10.1038/nrg3182 PMID: 22411464
4. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014; 509(7500):357–360. https://doi.org/10.1038/nature13178 PMID: 24739969
5. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC biology. 2014; 12(1):69. https://doi.org/10.1186/s12915-014-0069-1 PMID: 25184604
6. Bakken JS, Borody T, Brandt LJ, Brill JV, Demarco DC, Franzos MA, et al. Treating Clostridium difficile infection with fecal microbiota transplantation. Clinical Gastroenterology and Hepatology. 2011; 9 (12):1044–1049. https://doi.org/10.1016/j.cgh.2011.08.014 PMID: 21871249
7. Suez J, Korem T, Zeevi D, Zilberman-Schapira G, Thaiss CA, Maza O, et al. Artificial sweeteners induce glucose intolerance by altering the gut microbiota. Nature. 2014; 514(7521):181–186. https://doi.org/10.1038/nature13793 PMID: 25231862
8. Jumpertz R, Le DS, Turnbaugh PJ, Trinidad C, Bogardus C, Gordon JI, et al. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. The American journal of clinical nutrition. 2011; 94(1):58–65. https://doi.org/10.3945/ajcn.110.010132 PMID: 21543530
9. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. nature. 2006; 444(7122):1027–131. https://doi.org/10.1038/nature05414 PMID: 17183312
10. Messaoudi M, Lalonde R, Violle N, Javelot H, Desor D, Nejdi A, et al. Assessment of psychotropic-like properties of a probiotic formulation (Lactobacillus helveticus R0052 and Bifidobacterium longum R0175) in rats and human subjects. British Journal of Nutrition. 2011; 105(05):755–764. https://doi.org/10.1017/S0007114510004319 PMID: 20974015
11. Cryan JF, O'Mahony S. The microbiome-gut-brain axis: from bowel to behavior. Neurogastroenterology & Motility. 2011; 23(3):187–192. https://doi.org/10.1111/j.1365-2982.2010.01664.x
12. Palm NW, De Zoete MR, Cullen TW, Barry NA, Stefanowski J, Hao L, et al. Immunoglobulin A coating identifies colitogenic bacteria in inflammatory bowel disease. Cell. 2014; 158(5):1000–1010. https://doi.org/10.1016/j.cell.2014.08.006 PMID: 25171403
13. Sampson TR, Debelius JW, Thron T, Janssen S, Shastri GG, Ilhan ZE, et al. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. Cell. 2016; 167(6):1469–1480. https://doi.org/10.1016/j.cell.2016.11.018 PMID: 27912057
14. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012; 490(7418):55–60. https://doi.org/10.1038/nature11450 PMID: 23023125
15. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen AM, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell host & microbe. 2015; 17(2):260–273. https://doi.org/10.1016/j.chom.2015.01.001
16. Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, et al. Toward defining the autoimmune microbiome for type 1 diabetes. The ISME journal. 2011; 5(1):82–91. https://doi.org/10.1038/ismej.2010.92 PMID: 20613793
17. Brusca SB, Abramson SB, Scher JU. Microbiome and mucosal inflammation as extra-articular triggers for rheumatoid arthritis and autoimmunity. Current opinion in rheumatology. 2014; 26(1):101. https://doi.org/10.1097/BOR.0000000000000008 PMID: 24247114

18. Taneja V. Arthritis susceptibility and the gut microbiome. FEBS letters. 2014; 588(22):4244–4249. https://doi.org/10.1016/j.febslet.2014.05.034 PMID: 24873878

19. Williams BL, Hornig M, Parekh T, Lipkin WI. Application of novel PCR-based methods for detection, quantitation, and phylogenetic characterization of Sutterella species in intestinal biopsy samples from children with autism and gastrointestinal disturbances. MBio. 2012; 3(1):e00261–11. https://doi.org/10.1128/mBio.00261-11 PMID: 22233678

20. Wang L, Christophersen CT, Sorich MJ, Gerber JP, Angley MT, Conlon MA. Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder. Molecular autism. 2013; 4(1):1. https://doi.org/10.1186/2040-2392-4-42

21. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host 'I&' Microbe. 2014; 15(3):382–392. https://doi.org/10.1016/j.chom.2014.02.005

22. El Mouzan M, Wang F, Al Mofarreh M, Menon R, Al Barrag A, Korolev KS, et al. Fungal Microbiota Profile in Newly Diagnosed Treatment-naïve Children with Crohn's disease. Journal of Crohn's and Colitis. 2017; p. 1–7.

23. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016; 535(7610):94–103. https://doi.org/10.1038/nature18850 PMID: 27383984

24. Son JS, Zheng LJ, Rowehl LM, Tian X, Zhang Y, Zhu W, et al. Comparison of fecal microbiota in children with autism spectrum disorders and neurotypical siblings in the Simons Simplex Collection. PloS ONE. 2015; 10(10):e0137725. https://doi.org/10.1371/journal.pone.0137725 PMID: 26427004

25. Wang F, Kaplan JL, Gold BD, Bhasin MK, Ward NL, Kellermayer R, et al. Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota. Cell Reports. 2016; 14(4):945–955. https://doi.org/10.1016/j.celrep.2015.12.088 PMID: 26804920

26. De Cruz P, Prideaux L, Wagner J, Ng SC, McSweeney C, Kirkwood C, et al. Characterization of the gastrointestinal microbiota in health and inflammatory bowel disease. Inflammatory bowel diseases. 2012; 18(2):372–390 PMID: 21604329

27. Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: networks, competition, and stability. Science. 2015; 350(6261):663–666. https://doi.org/10.1126/science.aad2602 PMID: 26542567

28. Rakoff-Nahoum S, Foster KR, Comstock LE. The evolution of cooperation within the gut microbiota. Nature. 2016; 533(7602):255–259. https://doi.org/10.1038/nature17626 PMID: 27111508

29. Flint HJ, Duncan SH, Scott KP, Louis P. Interactions and competition within the microbial community of the human colon: links between diet and health. Environmental microbiology. 2007; 9(5):1101–1111. https://doi.org/10.1111/j.1462-2920.2007.01281.x PMID: 17472627

30. Bashan A, Gibson TE, Friedman J, Carey VJ, Weiss ST, Hohmann EL, et al. Universality of human microbial dynamics. Nature. 2016; 534(7606):259–262. https://doi.org/10.1038/nature18301 PMID: 27279224

31. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol. 2012; 8(7):e1002606. https://doi.org/10.1371/journal.pcbi.1002606 PMID: 22807668

32. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature Biotechnology. 2017; 35:81–89. PMID: 27893703

33. Chu J, Vila-Farres X, Inoyama D, Ternei M, Cohen LJ, Gordon EA, et al. Discovery of MRSA active antibiotics using primary sequence from the human microbiome. Nature Chemical Biology. 2016; 12 (12):1004–1006. https://doi.org/10.1038/nchembio.2207 PMID: 27748750

34. Riley MA, Goldstone C, Wertz J, Gordon D. A phylogenetic approach to assessing the targets of microbial warfare. Journal of evolutionary biology. 2003; 16(4):690–697. https://doi.org/10.1046/j.1420-9101.2003.00575.x PMID: 14632232

35. Czárán TL, Hoekstra RF, Pagie L. Chemical warfare between microbes promotes biodiversity. Proceedings of the National Academy of Sciences. 2002; 99(2):786–790. https://doi.org/10.1073/pnas.012399899

36. Dethlefsen L, Eckburg PB, Bik EM, Relman DA. Assembly of the human intestinal microbiota. Trends in ecology & evolution. 2006; 21(9):517–523. https://doi.org/10.1016/j.tree.2006.06.013

37. Mackie RI. Gut environment and evolution of mutualistic fermentative digestion. In: Gastrointestinal microbiology. Springer; 1997. p. 13–35.

38. Stein RR, Marks DS, Sander C. Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol. 2015; 11(7):e1004182. https://doi.org/10.1371/journal.pcbi.1004182 PMID: 26225866

39. Bialek W. Biophysics: searching for principles. Princeton University Press; 2012.
40. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome biology. 2012; 13(9):1. https://doi.org/10.1186/gb-2012-13-9-r79
41. Machiels K, Joossens M, Sabino J, De Preter V, Arijs I, Eeckhaut V, et al. A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. Gut. 2013; p. gutjnl–2013.
42. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome biology. 2012; 13(9):1. https://doi.org/10.1186/gb-2012-13-9-r79
43. Travis AJ, Kelly D, Flint HJ, Aminov RI. Complete genome sequence of the human gut symbiont Roseburia hominis. Genome announcements. 2015; 3(6):e01286–15. https://doi.org/10.1128/genomeA.01286-15 PMID: 26543119
44. Forbes JD, Van Domselaar G, Bernstein CN. The gut microbiota in immune-mediated inflammatory diseases. Frontiers in Microbiology. 2016; 7:1081. https://doi.org/10.3389/fmicb.2016.01081 PMID: 27462309
45. Joossens M, Huys G, Cnockaert M, De Preter V, Verbeke K, Rutgeerts P, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut. 2011; 60(5):631–637. https://doi.org/10.1136/gut.2010.223263 PMID: 21209126
46. Sokol H, Seksik P, Furet J, Firmesse O, Nion-Larmurier I, Beaugerie L, et al. Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflammatory bowel diseases. 2009; 15(8):1183–1189 PMID: 19235886
47. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, et al. Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in Crohn's disease. Digestion. 2016; 93(1):59–65. https://doi.org/10.1159/000441768 PMID: 26789999
48. Plischke M, Bergersen B. Equilibrium statistical physics. World Scientific Publishing Co Inc; 1994.
49. Harte J. Maximum entropy and ecology: a theory of abundance, distribution, and energetics. Oxford University Press; 2011.
50. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006; 440(7087):1007–1012. https://doi.org/10.1038/nature04701 PMID: 16625187
51. Volkov I, Banavar JR, Hubbell SP, Maritan A. Inferring species interactions in tropical forests. Proceedings of the National Academy of Sciences. 2009; 106(33):13854–13859. https://doi.org/10.1073/pnas.0903244106
52. Mora T, Walczak AM, Del Castello L, Ginelli F, Melillo S, Parisi L, et al. Local equilibrium in bird flocks. Nature Physics. 2016; 12(12):1153–1157. https://doi.org/10.1038/nphys3846 PMID: 27917230
53. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108
54. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proceedings of the National Academy of Sciences. 2011; 108(28):11530–11535. https://doi.org/10.1073/pnas.1105315108
55. Fisher CK, Mora T, Walczak AM. Variable habitat conditions drive species covariation in the human microbiota. PLOS Computational Biology. 2017; 13(4):e1005435. https://doi.org/10.1371/journal.pcbi.1005435 PMID: 28448493
56. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS computational biology. 2012; 8(9):e1002687. https://doi.org/10.1371/journal.pcbi.1002687 PMID: 23028285
57. Aitchison J. The statistical analysis of compositional data. Chapman and Hall London; 1986.
58. Pawlowsky-Glahn V, Buccianti A. Compositional data analysis: Theory and applications. John Wiley & Sons; 2011.
59. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric logratio transformations for compositional data analysis. Mathematical Geology. 2003; 35(3):279–300. https://doi.org/10.1023/A:1023818214614
60. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300.
61. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003; 100(16):9440–9445. https://doi.org/10.1073/pnas.1530509100

62. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nature Reviews Genetics. 2016;. PMID: 27840430

63. Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. PloS one. 2011; 6(5):e19416. https://doi.org/10.1371/journal.pone.0019416 PMID: 21589913

64. Ho TK. Random decision forests. In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. vol. 1. IEEE; 1995. p. 278–282.

65. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

66. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–297. https://doi.org/10.1007/BF00994018

67. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491(7422):119–124. https://doi.org/10.1038/nature11582 PMID: 23128233

68. Xavier R, Podolsky D. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007; 448(7152):427–434. https://doi.org/10.1038/nature06005 PMID: 17653185

69. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermúdez-Humarán LG, Gratadoux JJ, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proceedings of the National Academy of Sciences. 2008; 105(43):16731–16736. https://doi.org/10.1073/pnas.0804812105

70. Zhang M, Qiu X, Zhang H, Yang X, Hong N, Yang Y, et al. Faecalibacterium prausnitzii inhibits interleukin-17 to ameliorate colorectal colitis in rats. PloS one. 2014; 9(10):e109146. https://doi.org/10.1371/journal.pone.0109146 PMID: 25275569

71. Qiu X, Zhang M, Yang X, Hong N, Yu C. Faecalibacterium prausnitzii upregulates regulatory T cells and anti-inflammatory cytokines in treating TNBS-induced colitis. Journal of Crohn's and Colitis. 2013; 7(11):e558–e568. https://doi.org/10.1016/j.crohns.2013.04.002 PMID: 23643066

72. Forbes JD, Van Domselaar G, Bernstein CN. The gut microbiota in immune-mediated inflammatory diseases. Frontiers in Microbiology. 2016; 7. https://doi.org/10.3389/fmicb.2016.01081 PMID: 27462309

73. Scharek L, Hartmann L, Heinevetter L, Blaut M. Bifidobacterium adolescentis modulates the specific immune response to another human gut bacterium, Bacteroides thetaiotaomicron, in gnotobiotic rats. Immunobiology. 2000; 202(5):429–441. https://doi.org/10.1016/S0171-2985(00)80102-3 PMID: 11205373

74. Oyetayo VO, Oyetayo FL. Review-Potential of probiotics as biotherapeutic agents targeting the innate immune system. African Journal of Biotechnology. 2005; 4(2):123–127.

75. Duranti S, Milani C, Lugli GA, Mancabelli L, Turroni F, Ferrario C, et al. Evaluation of genetic diversity among strains of the human gut commensal Bifidobacterium adolescentis. Scientific reports. 2016; 6. https://doi.org/10.1038/srep23971

76. Sonomoto K, Yokota A. Lactic acid bacteria and bifidobacteria: current progress in advanced research. Horizon Scientific Press; 2011.

77. Louis P, Flint HJ. Formation of propionate and butyrate by the human colonic microbiota. Environmental Microbiology. 2016; 19:29–41. https://doi.org/10.1111/1462-2920.13589 PMID: 27928878

78. Jeraldo P, Hernandez A, Nielsen HB, Chen X, White BA, Goldenfeld N, et al. Capturing One of the Human Gut Microbiome's Most Wanted: Reconstructing the Genome of a Novel Butyrate-Producing, Clostridial Scavenger from Metagenomic Sequence Data. Frontiers in Microbiology. 2016; 7. https://doi.org/10.3389/fmicb.2016.00783 PMID: 27303377

79. Carbonero F, Benefiel AC, Gaskins HR. Contributions of the microbial hydrogen economy to colonic homeostasis. Nature Reviews Gastroenterology and Hepatology. 2012; 9(9):504–518. https://doi.org/10.1038/nrgastro.2012.85 PMID: 22585131

80. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nature reviews Microbiology. 2014; 12(10):661. https://doi.org/10.1038/nrmicro3344 PMID: 25198138

81. Kettle H, Louis P, Holtrop G, Duncan SH, Flint HJ. Modelling the emergent dynamics and major metabolites of the human colonic microbiota. Environmental microbiology. 2015; 17(5):1615–1630. https://doi.org/10.1111/1462-2920.12599 PMID: 25142831

82. Eeckhaut V, Van Immerseel F, Croubels S, De Baere S, Haesebrouck F, Ducatelle R, et al. Butyrate production in phylogenetically diverse Firmicutes isolated from the chicken caecum. Microbial biotechnology. 2011; 4(4):503–512. https://doi.org/10.1111/j.1751-7915.2010.00244.x PMID: 21375722

83. Louis P, Flint HJ. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS microbiology letters. 2009; 294(1):1–8. https://doi.org/10.1111/j.1574-6968.2009.01514.x PMID: 19222573

84. Gophna U, Konikoff T, Nielsen HB. Oscillospira and related bacteria—From metagenomic species to metabolic features. Environmental microbiology. 2017; 19(3):835–841. https://doi.org/10.1111/1462-2920.13658 PMID: 28028921

85. Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. The Journal of clinical investigation. 2014; 124(8):3617. https://doi.org/10.1172/JCI75436 PMID: 25003194

86. Kaakoush NO, Day AS, Huinao KD, Leach ST, Lemberg DA, Dowd SE, et al. Microbial dysbiosis in pediatric patients with Crohn's disease. Journal of clinical microbiology. 2012; 50(10):3258–3266. https://doi.org/10.1128/JCM.01396-12 PMID: 22837318

87. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS letters. 2014; 588(22):4223–4233. https://doi.org/10.1016/j.febslet.2014.09.039 PMID: 25307765

88. Verdam FJ, Fuentes S, de Jonge C, Zoetendal EG, Erbil R, Greve JW, et al. Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. Obesity. 2013; 21 (12). https://doi.org/10.1002/oby.20466 PMID: 23526699

89. Tims S, Derom C, Jonkers DM, Vlietinck R, Saris WH, Kleerebezem M, et al. Microbiota conservation and BMI signatures in adult monozygotic twins. The ISME journal. 2013; 7(4):707. https://doi.org/10.1038/ismej.2012.146 PMID: 23190729

90. Zhu L, Baker SS, Gill C, Liu W, Alkhouri R, Baker RD, et al. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. Hepatology. 2013; 57(2):601–609. https://doi.org/10.1002/hep.26093 PMID: 23055155

91. Keren N, Konikoff FM, Paitan Y, Gabay G, Reshef L, Naftali T, et al. Interactions between the intestinal microbiota and bile acids in gallstones patients. Environmental microbiology reports. 2015; 7(6):874–880. https://doi.org/10.1111/1758-2229.12319 PMID: 26149537

92. Milani C, Ticinesi A, Gerritsen J, Nouvenne A, Lugli GA, Mancabelli L, et al. Gut microbiota composition and Clostridium difficile infection in hospitalized elderly individuals: a metagenomic study. Scientific reports. 2016; 6. https://doi.org/10.1038/srep25945

93. Gu S, Chen Y, Zhang X, Lu H, Lv T, Shen P, et al. Identification of key taxa that favor intestinal colonization of Clostridium difficile in an adult Chinese population. Microbes and infection. 2016; 18(1):30–38. https://doi.org/10.1016/j.micinf.2015.09.008 PMID: 26383014

94. Minamoto Y, Otoni CC, Steelman SM, Büyükleblebici O, Steiner JM, Jergens AE, et al. Alteration of the fecal microbiota and serum metabolite profiles in dogs with idiopathic inflammatory bowel disease. Gut microbes. 2015; 6(1):33–47. https://doi.org/10.1080/19490976.2014.997612 PMID: 25531678

95. Werner T, Wagner SJ, Martínez I, Walter J, Chang JS, Clavel T, et al. Depletion of luminal iron alters the gut microbiota and prevents Crohn's disease-like ileitis. Gut. 2010; p. gut–2010.

96. Presley LL, Wei B, Braun J, Borneman J. Bacteria associated with immunoregulatory cells in mice. Applied and environmental microbiology. 2010; 76(3):936–941. https://doi.org/10.1128/AEM.01561-09 PMID: 20008175

97. Schwarz RS, Moran NA, Evans JD. Early gut colonizers shape parasite susceptibility and microbiota composition in honey bee workers. Proceedings of the National Academy of Sciences. 2016; 113 (33):9345–9350. https://doi.org/10.1073/pnas.1606631113

98. Raja M, Fajar Ummer C. Aggregatibacter actinomycetemcomitans—A tooth killer? Journal of clinical and diagnostic research: JCDR. 2014; 8(8):ZE13. https://doi.org/10.7860/JCDR/2014/9845.4766 PMID: 25302290

99. Kamma J, Nakou M, Manti F. Predominant microflora of severe, moderate and minimal periodontal lesions in young adults with rapidly progressive periodontitis. Journal of periodontal research. 1995; 30(1):66–72. https://doi.org/10.1111/j.1600-0765.1995.tb01254.x PMID: 7722848

100. Cassini M, Pilloni A, Condo S, Vitali L, Pasquantonio G, Cerroni L. Periodontal bacteria in the genital tract: are they related to adverse pregnancy outcome? International journal of immunopathology and pharmacology. 2013; 26(4):931–939. https://doi.org/10.1177/039463201302600411 PMID: 24355228

101. Sokol H, Leducq V, Aschard H, Pham HP, Jegou S, Landman C, et al. Fungal microbiota dysbiosis in IBD. Gut. 2016; p. gutjnl–2015.

102. Lavelle A, Lennon G, O'sullivan O, Docherty N, Balfe A, Maguire A, et al. Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. Gut. 2015; p. gutjnl–2014. https://doi.org/10.1136/gutjnl-2014-307873 PMID: 25596182

103. Mangin I, Bonnet R, Seksik P, Rigottier-Gois L, Sutren M, Bouhnik Y, et al. Molecular inventory of faecal microflora in patients with Crohn's disease. FEMS microbiology ecology. 2004; 50(1):25–36. https://doi.org/10.1016/j.femsec.2004.05.005 PMID: 19712374

104. Gophna U, Sommerfeld K, Gophna S, Doolittle WF, van Zanten SJV. Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis. Journal of clinical microbiology. 2006; 44(11):4136–4141. https://doi.org/10.1128/JCM.01004-06 PMID: 16988016

105. Tyler AD, Knox N, Kabakchiev B, Milgrom R, Kirsch R, Cohen Z, et al. Characterization of the gut-associated microbiome in inflammatory pouch complications following ileal pouch-anal anastomosis. PloS one. 2013; 8(9):e66934. https://doi.org/10.1371/journal.pone.0066934 PMID: 24086242

106. Hansen R, Berry SH, Mukhopadhya I, Thomson JM, Saunders KA, Nicholl CE, et al. The microaerophilic microbiota of de-novo paediatric inflammatory bowel disease: the BISCUIT study. PLoS One. 2013; 8(3):e58825. https://doi.org/10.1371/journal.pone.0058825 PMID: 23554935

107. Hiippala K, Kainulainen V, Kalliomäki M, Arkkila P, Satokari R. Mucosal prevalence and interactions with the epithelium indicate commensalism of Sutterella spp. Frontiers in microbiology. 2016; 7. https://doi.org/10.3389/fmicb.2016.01706 PMID: 27833600

108. Mukhopadhya I, Hansen R, Nicholl CE, Alhaidan YA, Thomson JM, Berry SH, et al. A comprehensive evaluation of colonic mucosal isolates of Sutterella wadsworthensis from inflammatory bowel disease. PLoS One. 2011; 6(10):e27076. https://doi.org/10.1371/journal.pone.0027076 PMID: 22073125

109. Biagi E, Candela M, Centanni M, Consolandi C, Rampelli S, Turroni S, et al. Gut microbiome in Down syndrome. PLoS one. 2014; 9(11):e112023. https://doi.org/10.1371/journal.pone.0112023 PMID: 25386941

110. Stewart GW. On the early history of the singular value decomposition. SIAM review. 1993; 35(4):551–566. https://doi.org/10.1137/1035134

111. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12(Oct):2825–2830.

112. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology. 2006; 72(7):5069–5072. https://doi.org/10.1128/AEM.03006-05 PMID: 16820507

113. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010; 1(May):335–336. https://doi.org/10.1038/nmeth.f.303

114. Paulson JN, Stine O Colin, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature methods. 2013; 10(12):1200–1202. https://doi.org/10.1038/nmeth.2658 PMID: 24076764