Title: Boolean Implication Analysis Unveils Candidate Universal Relationships in Microbiome Data

Authors: Daniella Vo[1†]; Shayal Charisma Singh[2†]; Sara Safa[3†]; Debashis Sahoo[3,4,5]

Affiliations:

[1]Department of Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA 92093-083.

[2]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA 92093-083.

[3]Department of Computer Science & Engineering, Jacob's School of Engineering, University of California San Diego, La Jolla, CA 92093-083.

[4] Department of Pediatrics, University of California San Diego, La Jolla, CA 92093-083.

[5]Moores Cancer Center, University of California San Diego, La Jolla, CA 92093-083.

† Equal contribution

Corresponding author:

Debashis Sahoo, Ph.D.; Assistant Professor, Department of Pediatrics, University of California San Diego; 9500 Gilman Drive, MC 0730, Leichtag Building 132; La Jolla, CA 92093-0831. **Phone:** 858-246-1803: **Fax:** 858-246-0019: **Email:** dsahoo@ucsd.edu

Abstract

Background: Microbiomes consist of bacteria, viruses, and other microorganisms, and are responsible for many different functions in both organisms and the environment. Some previous analyses of microbiomes focus on the relationships between specific microbiomes and a particular disease. These typically use correlation which is fundamentally symmetric with respect to pairs of microbes. Correlation focuses on the symmetry of the data distribution, and asymmetric data is often discarded as having a weak correlation. With all the data available on the microbiome, there is a need for a method that comprehensively studies microbiomes and how they are related to each other.

Results: We collect publicly available datasets from human, environment, and animal samples to determine both symmetric and asymmetric Boolean relationships between a pair of microbes. We then find relationships that are potentially invariants, meaning they will hold in any microbe community. In other words, if we determine there is a relationship between two microbes, we expect the relationship to hold in almost all context. We discovered that certain pairs of microbes always exhibit the same relationship in almost all the datasets we studied, thus making them good candidates for universal relationships. Our results confirm known biological properties and seem promising in terms of disease diagnosis.

Conclusions: Since the relationships are likely universal, we expect that they will hold in a clinical setting as well as in the general population. Strong universal relationships may provide insight on prognostic, predictive, or therapeutic properties of a clinically relevant disease. These new

analyses may improve disease diagnosis and drug development in terms of accuracy and efficiency.

Keywords: Boolean analysis, microbiome, invariants, systems biology, microbes interaction

Background

In recent years, microbiome research has progressed rapidly and there is an abundance of data publicly available. It is important to understand these microbes that are found in organisms and the environment, along with the relationships between them. There is also a growing interest to find the connection between microbiomes and the healthy or diseased states of an organism. Therefore, it is vital to efficiently analyze publicly available microbiome data, which will highlight the microorganisms and their interactions with the host.

Current methods of analysis fundamentally use Pearson's correlation coefficient to determine relationships within a microbiome, such as in co-occurrence networks (1). Correlation only identifies relationships whose distribution is symmetric or linear. Therefore, this method may only work for certain data in which there is a linear relationship between the two variables, but not for asymmetric data. In some cases, past studies disregard asymmetric data by concluding a weak correlation due to its nonlinearity, resulting in an incomplete analysis of asymmetric data. In addition, there may be other types of relationships that cannot be identified through the standard methods of linear analysis currently in use.

Past research used Boolean analysis to find the connection between oral microbiome and HIV-associated periodontists (2) and to find a metabolic network of interactions in the gut microbiome (3). However, these methods tend to analyze smaller datasets, which may have reproducibility issues. Additionally, these studies focus on specific regions, such as the mouth and gut microbiomes.

Instead, we propose using Boolean analysis, a logical method to determine dependency between two variables in order to study large amounts of microbiome data, comprehensively. This method of Boolean implication analysis was previously used to analyze relationships between genes using publicly available microarray datasets for humans, mice, and fruit flies. Our research adapts the method of analysis using Boolean implications, which was successfully used to discover markers of blood stem cells (4), progenitors and a branch point in B-cell and T-cell differentiation (5), and has also been applied to study colon (6, 7), bladder (8) and prostate cancer (9). Since this method of Boolean analysis was previously used on gene expression data, we want to demonstrate the universality of this method by analyzing pairwise microbe relationships.

In our process, the frequency of the microbes is first classified as either ‘low’ or ‘high’ using a threshold that is derived for each microbe species. An example of a Boolean implication rule is “if there is a high number of microbe A, then there will almost always be a low number of microbe B”, or A high → (implies) B low. There are six possible Boolean implication relationships: four asymmetric (A low → B low, A low → B high, A high → B low, and A high → B high) and two symmetric relationships (A equivalent B, and A opposite B).

Although our particular method of Boolean implication analysis has not been widely used in the area of microbiology, a group of researchers did attempt to use the method to perform analysis on microbiome data (10). In their research, they analyzed environmental data from Visualization and Analysis of Microbial Population Structures (VAMPS) (http://vamps.mbl.edu/diversity/diversity.php), and focused specifically on marine microbes. Their research differs from ours in the aspect of diversity, as we want to determine relationships that are present in a broader range of microbiomes, including humans, animals, and the environment. While this research demonstrates the impact Boolean implication analysis will have on microbiome analysis, we want to incorporate more as well as a greater variety of samples. By having a wider array of samples, we want to find microbe relationships that not only exist in the environment, but also in humans and animals.

Since Boolean implication analysis is able to capture relationships that are often overlooked in the existing methods of analysis, we aim to uncover candidate invariants between pairs of microbes that are likely applicable to every microbe community. For example, if we find a recurring Boolean relationship between two microbe species in our large, diverse datasets, we expect this relationship to be a promising candidate invariant. This research includes more diverse datasets and a novel mathematical model compared to past studies, which helps produce stronger universal candidates.

This project's goal is to identify Boolean relationships that are comprehensive and likely universal. Through Boolean implication analysis, we will be able to determine candidate universal invariants within diverse microbe communities. If there are general rules that any microbe community follows, it would make the process of identifying diseases easier because we expect these

relationships to hold in the general population. These candidate universal invariants provide a basis which scientists could use to determine which microbes are associated with diseases. Therapeutic use of microbes depends on their reproducibility in the general population, which makes our approach more suitable for discovering appropriate microbes.

Results

Many universal Boolean relationships were uncovered using the proposed method of Boolean implication analysis. Out of the 29,743,824 relationships discovered, some of the stronger ones are presented in Figures 2 and 3, and are described below.

Boolean implication relationships are conserved across environments and species

The high → low Boolean implication shows a high frequency of *Akkermansia muciniphila* (Operational Taxonomic Unit (OTU) ID 4306262) implying a low frequency of Stramenopiles (OTU ID 4350498) (Figure 2c). *A. muciniphila* is a human gut bacteria linked to preventing obesity, diabetes, and inflammation (11) and Stramenopiles is found in aquatic environments, mostly made up of algae (12). Since these two microbes are rarely found in similar environments, it makes sense that when one microbe's frequency is high, the other is low. This logic is consistent with the high → low Boolean relationship found.

The graph in Figure 2e displays a strong low → low relationship, showing that when *Polynucleobacter* (OTU ID 145533) is low, *Candidatus Xiphinematobacter* (OTU ID 786420) is also low. This relationship is confirmed in other studies which found *Polynucleobacter* makes up a large portion of freshwater bacterioplankton (13) and *Candidatus Xiphinematobacter* is a known

nutrient supplier to nematodes, which are abundant in freshwater environments (14). It is presumable that a low frequency count of *Polynucleobacter* indicates an environment that does not contain freshwater; therefore, it is unlikely that the frequency of *Candidatus Xiphinematobacter* is high, further confirming this low → low Boolean implication.

Boolean relationships confirm some known biological properties

A strong high → high relationship that was found (Figure 2d) is between *Corynebacterium* (OTU ID 1062356) and *Staphylococcus aureus* (OTU ID 4446058). *Corynebacterium* and *S. aureus* species both reside in the nose trail and skin microbiota of humans. *S. aureus* can be pathogenic and cause infections. Studies have shown that *Corynebacterium* spp. and *S. aureus* reside together, meaning they are positively correlated (15). In addition to being positively correlated, the *Corynebacterium* high → *S. aureus* high relationship reveals that it is also possible to have a low frequency of *Corynebacterium* and a high frequency of *S. aureus.*

An example of a symmetric relationship is shown in Figure 2f, where *Corynebacterium* (OTU ID 1062356) is equivalent to *Corynebacterium* (OTU ID 282360). *Corynebacterium* are a family of Gram-positive bacteria with a large number of known species due to interest in the medical field (16). However, the specific species of these *Corynebacterium* are not stated in the GreenGenes database. Further analysis could determine the species for these *Corynebacterium*, which would allow us to confirm this relationship. Although the species are not known, there is a symmetric relationship between the two species in that as the count of one increases, so does the other.

Microbes yield different Boolean implications in environmental versus animal samples

In the next two examples in Section 1 of Figure 3, green represents environmental samples and red represents the animal samples, including both humans and animals. Figure 3a shows the relationship *Polynucleobacter* (OTU ID 145533) high → *Candidatus Xiphinematobacter* (OTU ID 786420) low. The abundance of green and the absence of red samples suggest that this relationship is only present in the environmental microbiomes, and is not present in animal microbiomes. As stated previously, *Polynucleobacter* makes up a large portion of freshwater bacterioplankton (13), while *Candidatus Xiphinematobacter* tends to be found in soil samples (17), meaning they are both environmental microbes.

Figure 3b presents the relationship *Polynucleobacter* (OTU ID 3071019) high → *Bacteroides uniformis* (OTU ID 197072) low. This Boolean relationship suggests that *Polynucleobacter* is mainly present in the environmental microbiome while *Bacteroides uniformis* mostly exist in the animal microbiome. Previous studies have shown that *Bacteroides uniformis* is one of the main bacterial species of the human gut microbiome (18).

Different body sites affect the presence of microbes relationships

In Section 2 of Figure 3, both plots show the relationship between *Staphylococcus aureus* (OTU ID 446058) and *Corynebacterium* (OTU ID 1000986) but in different regions of the human body. While Figure 3c suggests that *S. aureus* and *Corynebacterium* have a low → low relationship in the skin region, based on Figure 3d, there is no specific Boolean relationship between these two microbes when they are present in human feces. This emphasizes the importance of different factors, like body region, in the way two microbes behave in the presence of each other.

Boolean implications using disease-specific microbes is promising in potential diagnosis

Irritable bowel disease (IBD) is a gastrointestinal disorder that currently is difficult to treat, but treatments using the gut microbiome have been proposed (19). The relationship *Actinomyces* (OTU ID 12564) high → *Lachnospiraceae* (OTU ID 4469576) low (Figure 3e) specifically highlights samples from patients that either have Crohn's Disease (CD), Ulcerative Colitis (UC), or neither (No IBD). Numerous studies show that there is a decreased amount of Clostridiales (*Lachnospiraceae* is in the class Clostridiales) in patients with IBS (19-22), but not much information about IBD. Although there have been studies that show there is a connection between *Actinomyces* and infections (23), studies have not compared the amount of *Actinomyces* to IBD. This indicates that more research needs to be done to determine if Clostridiales and *Actinomyces* are connected to certain types of IBD. There tends to be a higher proportion of *Lachnospiraceae* in patients with CD than UC, and a higher proportion of *Actinomyces* in patients of UC versus CD. The relationship *Streptococcus* (OTU ID 4467992) high → *Lachnospiraceae* (OTU ID 4469576) low (Figure 3f) also highlights differences between samples of IBD patients. A similar trend of a higher proportion of *Lachnospiraceae* in CD patients and a higher proportion of *Streptococcus* in UC patients appears with these microbes. There is no previous research suggesting there is a connection between *Streptococcus* and IBD, but more research needs to be done to see if this relationship has any disease-identifying properties. Relationships like these might have only been discovered using methods taking into account asymmetry like Boolean implication analysis which may be why there are no studies that can confirm these relationships.

Certain microbes seem to be related to different skin conditions, such as eczema and psoriasis. The relationship *Acinetobacter johnsonii* (OTU ID 4482374) low → *Corynebacterium* (OTU ID

361600) low shows how patients with psoriasis tend to have higher counts of both *Acinetobacter johnsonii* and *Corynebacterium* than patients with eczema and patients with neither skin condition. In looking at another relationship, *Ruminococcaceae* (OTU ID 4346675) high → *Anaerococcus* (OTU ID 927089) low, it is clear that patients with psoriasis have higher counts of *Anaerococcus*, while patients with eczema have higher counts of *Ruminococcaceae*, with both having minimal amounts of the other microbes. Although there has been research to detect microbial diversity on the skin, experts still cannot agree on a universal method of using microbes to diagnose psoriasis in patients (24). Our method of Boolean implication analysis attempts to provide a mathematical model of identifying candidate universal invariants. This will allow the properties determined within these microbiomes to apply in almost all situations, and hopefully provide treatments for diseases like these that will be universally successful.

Discussion

There is potential in using Boolean implication analysis to comprehensively determine microbial relationships, which can then be used to build abstract versions of biological systems. Understanding and simulating biological systems has always been the goal of researchers, but current analysis has not met that objective due to the lack of complexity that symmetric analysis has, and the smaller datasets researchers tend to use. Relationships between microbes and diseases have always been evident, so our research intends to build the foundation of the biological system. Specified research will provide more concrete evidence of the connection between microbes and diseases.

Our results reveal that Boolean analysis is a promising method for analysis of different microbiomes. After analyzing more than 400 diverse datasets consisting of over 100,000 samples, we uncovered candidate invariants that held in all our datasets, which is consistent with our hypothesis. However, only four of the six Boolean implication relationships were found in the datasets: low → low, high → low, high → high, and equivalent. The other two relationships, opposite and low → high, did not appear in the datasets. In both of these cases, the low, low quadrant is considered sparse, meaning there will be a high number of at least one microbe in most of the samples. Due to the diverse nature of microbes, we know that certain microbes are not present in every type of sample. Therefore, it should be rare for the low, low quadrant to be sparse, which is justified by the lack of opposite and low → high relationships in our results. Boolean analysis also unveils differences in microbe interactions due to factors such as body site, disease, and environment. Each implication is believed to be a universal candidate because it holds in all the datasets we analyzed. These microbe relationships can be validated in the lab by generating and sequencing additional samples to confirm these relationships.

One of the limitations in using Boolean analysis in which the data focuses on the stronger relationships, making the analysis less noisy. However, the weaker relationships are lost in the process. Further analysis might prove these weaker relationships have significance, but this method will only focus on the stronger relationships. A second limitation is that we only analyzed datasets downloaded from Qiita that were processed using the GreenGenes database (25). However, the latest version of GreenGenes was published in 2013, which may not include the most up to date information involving microbiome taxonomy. Additionally, since we are not focusing on microbiomes found in a specific region, we are limiting our scope to focus on microbes that are

found universally in humans, animals, and the environment. We may be excluding relationships that are compelling due to the fact that they might not be found everywhere. Future studies using this method of Boolean analysis should be done to focus on specific regions, which will give better insight into particular microbiomes, such as the gut microbiome.

Conclusions

The lack of comprehensive analysis of microbiome data created a need for more extensive approaches. Boolean implication analysis presents a solution that takes into account both symmetric and asymmetric relationships. Our results show that some biological properties were confirmed by Boolean analysis. For example, it is proven that the *Corynebacterium* and *S. aureus* species reside together and are positively correlated, which is consistent with the high → high relationship found. Our results also show that different microbiomes affect the presence of microbe relationships, on a broader scope such as environmental versus animal samples, or on a smaller scope such as various body sites. Boolean implication analysis is promising in terms of potential disease diagnosis, as well such as IBD. We found that higher frequencies of certain microbes seem to be associated with either CD or UC.

Future work includes building a Boolean implication network to further analyze how microbe implications are connected to each other. A Boolean implication network with the candidate microbe invariants may help in developing better models for biological systems. Our research also helps determine strong properties of biological systems, and future research on this topic provide novel directions in understanding biological system. Invariants help formulate new theories that may provide more effective diagnostic and therapeutic applications.

Materials and Methods

Data Collection

We extracted pre-processed OTU tables along with the corresponding metadata from Qiita (26), a microbiome database and study management platform. Qiita uses third party plugins including QIIME (Quantitative Insights Into Microbial Ecology) (http://qiime.org/) or QIIME 2 (https://qiime2.org/) to process microbial 16S rRNA sequences of each study, which are contributed by users on this platform. Qiita classifies the microbes using the GreenGenes database, and generates into OTU tables. OTU tables display the frequencies of all the microbes species present in each of the samples. Each study and the corresponding raw data comes from different individuals and institutes, which makes this database comprehensive. The metadata includes information about the samples, such as location and sample identification. For easier analysis of the collected data, we separated the downloaded studies from QIITA and pooled them into 4 datasets. These OTU tables and metadata were uploaded onto a web-based tool for analyzing big data called Hegemon (7, 8, 27-29).

Boolean Implication Analysis

To classify a relationship, thresholds are first determined for each microbe using the StepMiner algorithm. The StepMiner algorithm (30) is a tool that helps identify step-wise transitions (either step-up or step-down transitions) calculated using sum-of-square errors. Steps are defined as the sharpest change between low microbe frequency count and high microbe frequency count. In order to fit a step function, the StepMiner algorithm computes the average of the values on both sides of the step for all possible step positions. The midpoint of the step position that minimizes the square

299 error is chosen as the threshold for each respective microbe. The step is placed at the largest jump
300 from low values to high values, and the sets the threshold at the point where the step crosses the
301 original data. If the microbe frequencies are evenly distributed from low to high, the mean
302 frequency level tends to be used. Microbe frequencies are further classified as either 'high', 'low',
303 or 'intermediate'. If t is the microbe frequency threshold, frequency levels above $t + 0.5$ are 'high',
304 levels below $t - 0.5$ are 'low', and levels between $t - 0.5$ and $t + 0.5$ are 'intermediate'. Points in
305 the intermediate region are ignored because these points might appear on either side of the
306 threshold due to noise.
307
308 The OTU tables are uploaded onto Hegemon to visualize the samples on scatter plots, comparing
309 two microbes against each other. All microbe pairs are analyzed to determine if any Boolean
310 relationships are present. In each graph, one microbe species' frequency (using OTU ID A) is
311 plotted on the x-axis, and another microbe species' frequency (OTU ID B) is plotted on the y-axis.
312 Each data point represents a sample, and the frequencies are plotted on a log-log scale. From here,
313 Boolean implication analysis is applied to the scatter plots. Using the graphs constructed on
314 Hegemon, microbe pairs are analyzed to determine if a Boolean relationship is present. There are
315 six possible Boolean implications: symmetric (opposite and equivalent) or asymmetric (low →
316 low, low → high, high → low, high → high).
317
318 The asymmetric are determined by checking if one of the four quadrants in the scatter plot is
319 significantly sparse compared with other quadrants. If A low → B low and A high → B high are
320 both sparsely populated, then A is equivalent to B. If A high → B low and A low → B high are
321 both sparsely populated, then A is opposite to B. The BooleanNet statistics tests (27) determine

whether there is a Boolean relationship between A and B. Consider the relationship A low → B high. First, test if the microbe frequencies in the sparse quadrant is significantly less than the expected frequencies in an independence model. Let a_{00}, a_{01}, a_{10}, and a_{11} represent the quadrants in which the microbe frequencies of A and B are low and low, low and high, high and low, and high and high, respectively.

$$total = a_{00} + a_{01} + a_{10} + a_{11}$$

number of A low expression values = $nA_{low} = (a_{00} + a_{01})$

number of B low expression values = $nB_{low} = (a_{00} + a_{10})$

$$expected = \left(\frac{nA_{low}}{total} \times \frac{nB_{low}}{total}\right) \times\ total = (a_{00} + a_{01}) \times \frac{(a_{00} + a_{10})}{total}$$

$$observed = a_{00}$$

$$S\ statistic = \frac{expected - observed}{\sqrt{expected}}$$

Second, the observed values in the sparse quadrant are considered erroneous points, so a sparse quadrant should have a small number of these erroneous points. A maximum likelihood estimate of the error rate is then computed:

$$errorrate = \frac{1}{2}\left(\frac{a_{00}}{a_{00}+a_{01}} + \frac{a_{00}}{a_{00}+a_{10}}\right)$$

If both tests succeed, the low-low quadrant is considered sparse, so the implication A low → B high is true. An implication is considered significant if the S statistic is greater than 3 and the error rate is less than 0.1. All the microbe relationships that pass both of these tests are now considered candidate invariants.

Analysis of metadata

We used a Hegemon function that calculates the differential analysis for specific factors in the metadata using t-tests in R software framework (R version 3.4.4 - 2018-03-15). We then selected the OTU IDs that had higher mean differential values and higher $-\log_{10}(p)$ values. After the list of potential IDs were generated, we visually confirmed on Hegemon whether certain factors such as environment versus animal or body site affects the microbe counts.

List of Abbreviations

CD: Crohn's Disease

IBD: Irritable Bowel Disease

QIIME: Quantitative Insights Into Microbial Ecology

OTU: Operational Taxonomic Units

UC: Ulcerative Colitis

VAMPS: Visualization and Analysis of Microbial Population Structures

Declarations

Ethics approval and consent to participant

Publicly available data were generated after appropriate ethics approval and consent to participate.

Consent for publication

Not Applicable.

Availability of data and materials

All dataset are publicly available from https://qiita.ucsd.edu, under the Study, and View Studies tab. We selected a subset of all the available studies. The specific list of studies used for this research is included as an additional file.

Competing interests

The authors declare that they have no competing interests.

Funding

DS is supported by NIH UG3TR003355, UG3TR002968, R01AI155696, R01GM138385 and R00CA151673, Hyundai Hope On Wheels, Padres Pedal the Cause/Rady Children's Hospital Translational PEDIATRIC Cancer Research Award (Padres Pedal the Cause/RADY #PTC2017), 2017, Padres Pedal the Cause/C3 Collaborative Translational Cancer Research Award (San Diego NCI Cancer Centers Council [C3] #PTC2017).

Authors' contributions

DV, SCS, SS, DS conceived and designed the analysis, collected the data, contributed analysis tools, performed the analysis, wrote the paper. DS supervised, acquired funding, reviewed, and edited the paper.

Acknowledgements

The authors thank Dr. Christine Alvarado for her guidance and the members of UCSD's Boolean Lab for reviewing our work and providing constructive criticism.

References

1. Bauer E, Thiele I. From Network Analysis to Functional Metabolic Modeling of the Human Gut Microbiota. mSystems. 2018;3(3).

2. Noguera-Julian M, Guillén Y, Peterson J, Reznik D, Harris EV, Joseph SJ, et al. Oral microbiome in HIV-associated periodontitis. Medicine (Baltimore). 2017;96(12):e5821.

3. Steinway SN, Biggs MB, Loughran TP, Papin JA, Albert R. Inference of Network Dynamics and Metabolic Interactions in the Gut Microbiome. PLoS Comput Biol. 2015;11(5):e1004338.

4. Chen JY, Miyanishi M, Wang SK, Yamazaki S, Sinha R, Kao KS, et al. Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. Nature. 2016;530(7589):223-7.

5. Inlay MA, Bhattacharya D, Sahoo D, Serwold T, Seita J, Karsunky H, et al. Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. Genes Dev. 2009;23(20):2376-81.

6. Dalerba P, Sahoo D, Clarke MF. CDX2 as a Prognostic Biomarker in Colon Cancer. N Engl J Med. 2016;374(22):2184.

7. Dalerba P, Sahoo D, Paik S, Guo X, Yothers G, Song N, et al. CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. N Engl J Med. 2016;374(3):211-22.

8. Volkmer JP, Sahoo D, Chin RK, Ho PL, Tang C, Kurtova AV, et al. Three differentiation states risk-stratify bladder cancer into distinct subtypes. Proc Natl Acad Sci U S A. 2012;109(6):2078-83.

9. Sahoo D, Wei W, Auman H, Hurtado-Coll A, Carroll PR, Fazli L, et al. Boolean analysis identifies CD38 as a biomarker of aggressive localized prostate cancer. Oncotarget. 2018;9(5):6550-61.

10. Zhu C, Jiang R, Chen T. Constructing a Boolean implication network to study the interactions between environmental factors and OTUs. Quantitative Biology. 2014;2(4):127-41.

11. Depommier C, Everard A, Druart C, Plovier H, Van Hul M, Vieira-Silva S, et al. Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: a proof-of-concept exploratory study. Nat Med. 2019;25(7):1096-103.

12. Derelle R, López-García P, Timpano H, Moreira D. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). Mol Biol Evol. 2016;33(11):2890-8.

13. Hahn MW, Jezberová J, Koll U, Saueressig-Beck T, Schmidt J. Complete ecological isolation and cryptic diversity in Polynucleobacter bacteria not resolved by 16S rRNA gene sequences. ISME J. 2016;10(7):1642-55.

14. Brown AM, Howe DK, Wasala SK, Peetz AB, Zasada IA, Denver DR. Comparative Genomics of a Plant-Parasitic Nematode Endosymbiont Suggest a Role in Nutritional Symbiosis. Genome Biol Evol. 2015;7(9):2727-46.

15. Ramsey MM, Freire MO, Gabrilska RA, Rumbaugh KP, Lemon KP. Staphylococcus aureus Shifts toward Commensalism in Response to Corynebacterium Species. Front Microbiol. 2016;7:1230.

16. Liebl W. Corynebacterium taxonomy. Handbook of Corynebacterium glutamicum CRC Press, Boca Raton, FL. 2005:9-34.

17. Lazarova S, Peneva V, Kumari S. Morphological and molecular characterisation, and phylogenetic position of X. browni sp. n., X. penevi sp. n. and two known species of Xiphinema americanum-group (Nematoda, Longidoridae). Zookeys. 2016(574):1-42.

18. Renouf M, Hendrich S. Bacteroides uniformis is a putative bacterial species associated with the degradation of the isoflavone genistein in human feces. J Nutr. 2011;141(6):1120-6.

19. Pittayanon R, Lau JT, Yuan Y, Leontiadis GI, Tse F, Surette M, et al. Gut Microbiota in Patients With Irritable Bowel Syndrome-A Systematic Review. Gastroenterology. 2019;157(1):97-108.

20. Jalanka-Tuovinen J, Salojärvi J, Salonen A, Immonen O, Garsed K, Kelly FM, et al. Faecal microbiota composition and host-microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. Gut. 2014;63(11):1737-45.

21. Carroll IM, Ringel-Kulka T, Siddle JP, Ringel Y. Alterations in composition and diversity of the intestinal microbiota in patients with diarrhea-predominant irritable bowel syndrome. Neurogastroenterol Motil. 2012;24(6):521-30, e248.

22. Li G, Yang M, Jin Y, Li Y, Qian W, Xiong H, et al. Involvement of shared mucosal-associated microbiota in the duodenum and rectum in diarrhea-predominant irritable bowel syndrome. J Gastroenterol Hepatol. 2018;33(6):1220-6.

23. Könönen E, Wade WG. Actinomyces and related organisms in human infections. Clin Microbiol Rev. 2015;28(2):419-42.

24. Langan EA, Griffiths CEM, Solbach W, Knobloch JK, Zillikens D, Thaçi D. The role of the microbiome in psoriasis: moving from disease description to treatment selection? Br J Dermatol. 2018;178(5):1020-7.

25. 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. Appl Environ Microbiol. 2006;72(7):5069-72.

26. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods. 2018;15(10):796-8.

27. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK. Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biol. 2008;9(10):R157.

28. Sahoo D, Seita J, Bhattacharya D, Inlay MA, Weissman IL, Plevritis SK, et al. MiDReG: a method of mining developmentally regulated genes using Boolean implications. Proc Natl Acad Sci U S A. 2010;107(13):5732-7.

29. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol. 2011;29(12):1120-7.

30. Sahoo D, Dill DL, Tibshirani R, Plevritis SK. Extracting binary signals from microarray time-course data. Nucleic Acids Res. 2007;35(11):3705-12.

Figure Legends:

Figure 1. Study design

The overview of the research process: (a) OTU tables were collected from publicly available microbiome datasets. (b) Tables were uploaded to Hegemon and all possible microbe pairs were plotted (Using 4 plots as examples - The number of total plots is larger). (c) Boolean analysis was performed on all the plots. (d) The plots that passed the BooleanNet statistics tests were marked as candidate invariants for further analysis and validation. (e) Any of the determined candidates that can be validated in other datasets, represents a likely universal invariant (a rule between two microbes that holds between them, any time the pair are present together in any environment).
*Note that this is just an example. A universal invariant can be any of the 6 possible Boolean relationship.

Figure 2. Boolean implication relationships represent diversity in microbiome data

All four types of Boolean relationships found in our dataset. (a) Describes the StepMiner algorithm that creates thresholds for each microbe. Among all samples, the frequencies of a particular microbe are sorted and a step function is fitted where the sharpest change between low microbe

count and high microbe count takes place. The midpoint of the step position that minimizes the square error is chosen as the threshold for each respective microbe. (b) Depicts a log-log plot of the number of each type of relationship and the corresponding number of microbes that exhibit that specific relationship. The remaining diagrams display each of the four types of relationships found. Each point in the scatter plot corresponds to a sample, where the two axes represent the frequency counts of each microbes. (c) high → low (d) high → high (e) low → low (f) equivalent

Figure 3. Boolean implication reveals strong patterns in diverse biological and environmental conditions

Analysis of scatter plots with various experimental conditions using metadata files that provided additional information about the samples. *Section 1*: green represents environmental samples (plants, water, soil, etc.) and red indicates animal samples (humans, animals). (a) *Polynucleobacter* (145533) low → *Candidatus Xiphinematobacter* (786420) low; this relationship is only present in environmental microbiomes due to the lack of red samples in the plot. (b) *Polynucleobacter* (3071019) high → *Bacteroides uniformis* (197072) low*;* this relationship suggests *Polynucleobacter* is mainly present in the environment, and *Bacteroides uniformis* is mainly present in animals. *Section 2*: (c) and (d) have the same microbes on the axes *Staphylococcus aureus* (446058) and *Corynebacterium* (1000986), but different regions of the body plotted: skin (dark blue) and feces (green). (c) shows the relationship *S. aureus* low → *Corynebacterium* low holds for the skin region. (d) shows the relationship using fecal samples, and there is not a clear relationships that can be determined from this. *Section 3*: Pink represents Crohn's Disease (CD), teal represents Ulcerative Colitis (UC), and light gray represents neither disease. (e) The relationship *Actinomyces* (12574) high → *Lachnospiraceae* (4469576) low is shown, with higher

counts of *Lachnospiraceae* in CD, and higher counts of *Actinomyces* in UC. (f) shows the relationship *Streptococcus* (4467992) high → *Lachnospiraceae* (4469576) low, with higher counts of *Lachnospiraceae* in CD, while higher counts of Streptococcus in UC. *Section 4*: Magenta represents eczema, blue represents psoriasis, and beige represents neither skin condition. (g) The relationship *Acinetobacter johnsonii* (4482374) low → *Corynebacterium* (361600) low is shown. *Patients with psoriasis tend to have higher counts of Corynebacterium* than patients with eczema. (h) The relationship *Ruminococcaceae* (4346675) high → *Anaerococcus* (927089) low is shown. Patients with psoriasis tend to have higher counts of *Anaerococcus*, while patients with eczema tend to have higher counts of *Ruminococcaceae*.