Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Boolean Implication Analysis Unveils Candidate Universal Relationships in Microbiome Data

Daniella Vo
University of California San Diego https://orcid.org/0000-0003-3432-2326

Shayal Charisma Singh
University of California San Diego

Sara Safa
University of California San Diego

Debashis Sahoo (dsahoo@ucsd.edu)
University of California San Diego https://orcid.org/0000-0003-2329-8228

Research article

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

Posted Date: December 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-65519/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published on February 5th, 2021. See the published version at https://doi.org/10.1186/s12859-020-03941-4.

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. Past analyses of microbiomes focused on using correlation to determine linear relationships between microbes and diseases. Weak correlations due to nonlinearity between microbe pairs may cause researchers to overlook critical components of the data. With the abundance of available microbiome, we need a method that comprehensively studies microbiomes and how they are related to each other.

Results: We collected publicly available datasets from human, environment, and animal samples to determine both symmetric and asymmetric Boolean implication relationships between a pair of microbes. We then found 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 contexts. We discovered that around 330,000 pairs of microbes universally exhibit the same relationship in almost all the datasets we studied, thus making them good candidates for invariants. Our results also confirm known biological properties and seem promising in terms of disease diagnosis.

Conclusions: Since the relationships are likely universal, we expect them to hold in clinical settings, as well as general populations. If these strong invariants are present in disease settings, it may provide insight into prognostic, predictive, or therapeutic properties of clinically relevant diseases. For example, our results indicate that there is a difference in the microbe distributions between patients who have or do not have IBD, eczema and psoriasis. These new analyses may improve disease diagnosis and drug development in terms of accuracy and efficiency.

Keywords: Boolean analysis, microbiome, invariants, systems biology, microbe interactions

Background

In recent years, microbiome research has progressed rapidly, and there is an abundance of publicly available data. It is important to learn more about these microbes that are found in organisms and the environment, and the relationships between these microbes. There is also a growing interest to find the connection between microbiomes and diseases. Therefore, it is vital to efficiently analyze publicly available microbiome data focusing on microorganisms and their interactions with their host or environment.

Current research and its limitations

Current methods of analysis fundamentally use Pearson's correlation coefficient to determine relationships within a microbiome, such as in co-occurrence networks (1). However, correlation only identifies relationships whose distribution is linear. This method may only work for data in which there is a linear relationship between the two variables bypassing data that is not fundamentally linear. Past studies may deduce weak correlations due to its nonlinearity thereby resulting in an incomplete analysis that overlooks critical components of the data. Additionally, there may be other types of relationships that cannot be identified through the standard methods of linear analysis currently in use. Correlation is also symmetric because corr(x, y) is same as corr(y, x). Therefore, asymmetric relationships are not captured well using correlation analysis.

Some forms of research used Boolean analysis to find links between oral microbiomes 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 areas such as the mouth and gut microbiomes.

Boolean analysis is a more comprehensive approach

Instead of correlation-based analyses, we propose using Boolean analysis, a logical method to comprehensively study large amounts of microbiome data in order to determine dependencies between two variables. An overview of this methodology is described in Figure 1. This method of Boolean implication analysis was successfully used to analyze relationships between genes to discover markers of blood stem cells (4), progenitors and a branch point in B-cell and T-cell differentiation (5), and has been applied in the study of 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.

Microbe normalized counts (in $\log_2$ scale) are first classified as either 'low' or 'high' using a threshold that is derived for each microbe species (Figure 2a). 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). A low → B low is asymmetric because it is different from B low → A low, and the same applies to other

asymmetric Boolean implication relationships. The convention used in this paper is to write Boolean implications with the microbe on the x-axis first; however, there exists a contrapositive relationship for each asymmetric Boolean implication, which is obtained by stating the microbe on the y-axis first and inverting low values to high values, and vice versa. For instance, A high → B low is identical to B high → A low.

Although our method of Boolean implication analysis has not been widely used in microbiology, a group of researchers attempted 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 a larger and 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.

Boolean methods have the potential to uncover universal invariants

Since Boolean implication analysis captures 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 and diverse datasets, we expect this relationship to be a promising candidate invariant. This research includes more diverse datasets

117 and a novel mathematical model compared to past studies, which helps produce stronger universal
118 candidates.
119
120 The goal of this research is to comprehensively identify Boolean relationships that are likely
121 universal. Through Boolean implication analysis, we will be able to determine candidate universal
122 invariants within diverse microbe communities. Since these universal rules are expected to be
123 robust, it can be translated to a clinical domain because fundamental rules should appear
124 irrespective of diseases. These candidate universal invariants provide a basis in which scientists
125 could use to determine how microbes are associated with diseases and make the process of
126 identifying diseases easier because we expect these relationships to hold in the general population.
127 Therapeutic use of microbes depends on their reproducibility in the general population, which
128 makes our approach more suitable for discovering appropriate microbes.
129

130 **Results**

131 Many universal Boolean relationships were uncovered using the proposed method of Boolean
132 implication analysis. These pooled datasets are comprised of a variety of environmental, animal,
133 and human samples. A main dataset was used to perform Boolean analysis for approximately 365
134 million microbe pairs and we discovered about 27 million relationships (Figure 2b) with a high
135 statistical significance (false discovery rate (FDR) of 2.3×10^{-4}). We also used three pooled
136 independent datasets to validate these relationships (with FDRs of 3.3×10^{-4}, 4.4×10^{-4} and $2.1 \times$
137 10^{-4} respectively), and found approximately 330,000 relationships that were consistent throughout
138 all four datasets. We present some relationships in Figures 2 and 3 that have a superior BooleanNet
139 statistics (S, p) with higher independence statistic (S), lower error rate (p) and an underlying

140 biological relevance. Figure 2 is mainly used to present the possible types of Boolean relationships
141 found in microbiomes while Figure 3 displays some of the biological properties associated with
142 certain microbes.
143
144 **Boolean implication relationships are conserved across environments and species**
145 The high → low Boolean implication shows a high count of *Akkermansia muciniphila* (Operational
146 Taxonomic Unit (OTU) ID 4306262) implying a low count of Stramenopiles (OTU ID 4350498)
147 (Figure 2c). The contrapositive relationship also holds (Stramenopiles high → *A. muciniphila* low).
148 *A. muciniphila* is a human gut bacterium linked to preventing obesity, diabetes, and inflammation
149 (11) and Stramenopiles is found in aquatic environments, mostly made up of algae (12). Since
150 these two microbes are rarely found in similar environments, it makes sense that when one
151 microbe's frequency is high, the other is low. This logic is consistent with the high → low Boolean
152 relationship found.
153
154 The graph in Figure 2d displays a strong low → low relationship, showing that when
155 *Polynucleobacter* (OTU ID 145533) is low, *Candidatus Xiphinematobacter* (OTU ID 786420) is
156 also low. This relationship is confirmed in other studies which found *Polynucleobacter* makes up
157 a large portion of freshwater bacterioplankton (13) and *Candidatus Xiphinematobacter* is a known
158 nutrient supplier to nematodes, which are abundant in freshwater environments (14). It is
159 presumable that a low count of *Polynucleobacter* indicates an environment that does not contain
160 freshwater; therefore, it is unlikely that the count of *Candidatus Xiphinematobacter* is high, further
161 confirming this low → low Boolean implication.
162

163 **Boolean relationships confirm some known biological properties**

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

171

172 An example of a symmetric relationship is shown in Figure 2f, where *Corynebacterium* (OTU ID
173 1062356) is equivalent to *Corynebacterium* (OTU ID 282360). *Corynebacterium* are a family of
174 Gram-positive bacteria with a large number of known species which are of interest in the medical
175 field (16). However, the specific species of these *Corynebacterium* are not stated in the
176 GreenGenes database. Further analysis could determine the specific *Corynebacterium* species
177 which would allow us to confirm this symmetric relationship. Although the specific species in the
178 database are unknown, there is a symmetric relationship between the two species such that as the
179 count of one species increases so does the other.

180

181 **Microbes yield different Boolean implications in environmental versus animal samples**

182 In the next two examples in Section 1 of Figure 3, green represents environmental samples and red
183 represents both human and animal samples. Figure 3a shows the relationship *Polynucleobacter*
184 (OTU ID 145533) high → *Candidatus Xiphinematobacter* (OTU ID 786420) low. The abundance
185 of green samples suggests that this relationship is mainly present in environmental microbiomes

186 and is not typically present in animal microbiomes. As stated previously, *Polynucleobacter* makes
187 up a large portion of freshwater bacterioplankton (13), while *Candidatus Xiphinematobacter* tends
188 to be found in soil samples (17), meaning they are both environmental microbes and are not typical
189 microbes found in humans or animals.
190
191 Figure 3b presents the relationship *Polynucleobacter* (OTU ID 3071019) high → *Bacteroides*
192 *uniformis* (OTU ID 197072) low. This Boolean relationship suggests that *Polynucleobacter* is
193 mainly present in the environmental microbiome while *Bacteroides uniformis* mostly exist in the
194 animal microbiome. Previous studies have shown that *Bacteroides uniformis* is one of the main
195 bacterial species of the human gut microbiome (18).
196
197 **Different body regions affect the presence of microbes relationships**
198 In Section 2 of Figure 3, both plots show the relationship between *Staphylococcus aureus* (OTU
199 ID 446058) and *Corynebacterium* (OTU ID 1000986) but in different regions of the human body.
200 While Figure 3c suggests that *S. aureus* and *Corynebacterium* have a low → low relationship in
201 the skin region, based on Figure 3d, there is no specific Boolean relationship between these two
202 microbes when they are present in human feces. Specific regions may have differing relationships
203 due to the proclivity microbes have towards one region versus another.
204
205 **Boolean implications using disease-specific microbes is promising in potential diagnosis**
206 Inflammatory bowel disease (IBD) is a gastrointestinal disorder that is currently difficult to treat,
207 but treatments using the gut microbiome have been proposed (19). The relationship *Actinomyces*
208 (OTU ID 12564) high → *Lachnospiraceae* (OTU ID 4469576) low (Figure 3e) specifically

209 highlights samples from patients that either have Crohn's Disease (CD), Ulcerative Colitis (UC),
210 or neither (No IBD). There tends to be a higher proportion of *Lachnospiraceae* in patients with
211 CD than UC, and a higher proportion of *Actinomyces* in patients of UC versus CD. The relationship
212 *Streptococcus* (OTU ID 4467992) high → *Lachnospiraceae* (OTU ID 4469576) low (Figure 3f)
213 also highlights differences between samples of IBD patients. A similar trend of a higher proportion
214 of *Lachnospiraceae* in CD patients and a higher proportion of *Streptococcus* in UC patients
215 appears with these microbes.
216
217 Certain microbes seem to be related to skin conditions, such as eczema and psoriasis. The
218 relationship *Acinetobacter johnsonii* (OTU ID 4482374) low → *Corynebacterium* (OTU ID
219 361600) low shows how patients with psoriasis tend to have higher counts of both *A. johnsonii*
220 and *Corynebacterium* than patients with eczema and patients with neither skin condition. In
221 looking at another relationship, *Ruminococcaceae* (OTU ID 4346675) high → *Anaerococcus*
222 (OTU ID 927089) low, it is clear that patients with psoriasis have higher counts of *Anaerococcus,*
223 while patients with eczema have higher counts of *Ruminococcaceae*, with both having minimal
224 amounts of the other microbes. Our method of Boolean implication analysis attempts to provide a
225 mathematical model of identifying candidate universal invariants. This will enable the determined
226 microbiome properties to apply in almost all states, and hopefully provide treatments for such
227 diseases to be universally successful.
228

229 **Discussion**

230 Boolean implication analysis can be used to comprehensively determine microbial relationships,
231 which can then be used to build abstract versions of biological systems. Understanding and

232 simulating biological systems has always been the goal of researchers, but current analysis has not
233 met that objective with simplified symmetric analysis and smaller volume of datasets.
234 Relationships between microbes and diseases have always been evident, so our research intends
235 to build the foundation of the biological system.
236
237 **Correlation versus Boolean implication analysis**
238 In a correlation-based analysis, we found that most of the microbe pairs that exhibited an
239 asymmetric Boolean relationship had a weak correlation. As expected, equivalent Boolean
240 relationships (which are symmetric relationships) typically contain highly correlated microbes.
241 Figure 2 (g-j) depicts the distribution of correlation values for each of the four Boolean
242 relationships found in our dataset. It was rare to find asymmetric microbe pairs that yielded a
243 strong correlation which further highlights the shortcomings of analyses based on correlations.
244 This reinforces the observation that studying these relationships may result in a more
245 comprehensive and complete analysis.
246
247 Our results reveal that Boolean analysis is a promising method for analyzing different
248 microbiomes. After analyzing more than 400 diverse datasets consisting of over 100,000 samples,
249 we uncovered candidate invariants that held in all our datasets, which is consistent with our
250 hypothesis. However, only four of the six Boolean implication relationships were found in the
251 datasets: low → low, high → low, high → high, and equivalent. The other two relationships,
252 opposite and low → high, did not appear in the datasets because in these two relationships, the
253 low, low quadrant is sparse. A sparse low, low quadrant means both microbes cannot exist in low
254 amounts at the same time. Due to the diverse nature of microbes, we believe that it should be rare

255 for the low, low quadrant to be sparse which justifies the lack of low → high and opposite
256 relationships in our results.
257
258 **Boolean analysis unveils differences in microbe interactions due to environment, body site,**
259 **and disease**
260 Environmentally distinct microbes exhibit relationships such as *A. muciniphila* high →
261 Stramenopiles low because *A. muciniphila* is present in the human gut (11) while Stramenopiles
262 is present in aquatic environments (12). Boolean analysis also highlighted the dichotomies in
263 different body sites, as some relationships are present in certain regions of the body, but not in
264 others. The *Corynebacterium* high → *S. aureus* high relationship observed in skin samples is
265 supported by biological research (15) but is not present in fecal samples. Boolean analysis allows
266 us to see where a clinically relevant relationship might occur in the body. The relationships *A.*
267 *johnsonii* low → *Corynebacterium* low and *Ruminococcaceae* high → *Anaerococcus* low reveal
268 that samples from eczema and psoriasis were clustered in certain regions of the graph. Although
269 there has been research to detect microbial diversity on the skin, experts still cannot agree on a
270 universal method of using microbes to diagnose psoriasis in patients (20). Using the clustered
271 regions from eczema and psoriasis-specific relationships, scientists would be able to define healthy
272 and diseased ranges for microbial frequencies which is promising for disease diagnosis.
273
274 Scientists can develop therapeutics targeting diseases like IBD using relationships such as
275 *Actinomyces* high → *Lachnospiraceae* low and *Streptococcus* high → *Lachnospiraceae* low.
276 Numerous studies show that there is a decreased amount of Clostridiales (*Lachnospiraceae* is in
277 the class Clostridiales) in patients with Irritable Bowel Syndrome (IBS) (19, 21-23), but there is

limited information about IBD. Although there have been studies showing a connection between *Actinomyces* and infections (24), studies have not compared the amount of *Actinomyces* to IBD. However, our results suggest that Clostridiales and *Actinomyces* are connected to UC and CD. There is no previous research suggesting there is a connection between *Streptococcus* and IBD, indicating a need for more research to determine if this relationship has any disease-identifying properties. Relationships like these might have only been discovered using methods that consider asymmetry like Boolean implication analysis which might explain why no studies can confirm such relationships.

Limitations of Boolean implication analysis

One of the limitations of Boolean analysis is that the data focuses on the stronger relationships making the analysis less noisy and weaker relationships are lost in the process. Further analysis might prove whether these weaker relationships have significance, but this method focuses 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 limiting the scope to microbiomes found in a specific region, but focusing on microbes found universally in humans, animals, and the environment, we could be excluding compelling relationships that are only found in specific regions. Future studies using this method of Boolean analysis can be done to focus on these specific regions to provide 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 incorporates 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, and on a smaller scope, such as various body sites. Boolean implication analysis is promising in terms of potential disease diagnosis including IBD. We found that higher frequencies of certain microbes seem to be associated with either CD or UC.

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. Future work also 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 how these systems work. Invariants help formulate new theories that may provide more effective diagnostic and therapeutic applications.

324 # Methods

325 **Data Collection**

326 We extracted pre-processed OTU tables along with the corresponding metadata from Qiita (26), a
327 microbiome database and study management platform. Qiita uses third party plugins including
328 QIIME (Quantitative Insights Into Microbial Ecology) (http://qiime.org/) or QIIME 2
329 (https://qiime2.org/) to process microbial 16S rRNA sequences of each study, which are
330 contributed by users on this platform. Qiita classifies the microbes using the GreenGenes database
331 and generates OTU tables. OTU tables display the counts of all the microbe species present in
332 every sample. Each study and the corresponding raw count data comes from different individuals
333 and institutions, which makes our analysis comprehensive. The metadata includes information
334 about the samples, such as location and sample identification. For easier analysis of the collected
335 data, we separated the downloaded studies from QIITA and pooled them into four independent
336 datasets. We performed our main analysis on one of the pooled datasets and used the other three
337 datasets for validation. In Supplementary Table 1, we specify all the studies used and which pooled
338 dataset it belongs to (main, validation datasets 1, 2 or 3). Data from the main dataset is presented
339 in the figures unless explicitly stated otherwise. The data from the OTU tables and metadata are
340 transformed using a $\log_2$ scale and were subsequently uploaded onto a web-based tool for
341 analyzing big data called Hegemon (7, 8, 27-29).

342

343 **StepMiner Algorithm**

344 To classify a relationship, thresholds are first determined for each microbe using the StepMiner
345 algorithm. The StepMiner algorithm (30) is a tool that helps identify stepwise transitions (either
346 step-up or step-down transitions) calculated using sum-of-square errors. After the data is

347 normalized, steps are defined as the sharpest change between low microbe count and high microbe
348 count. In order to fit a step function, the StepMiner algorithm computes the average of the values
349 on both sides of the step for all possible step positions. The midpoint of the step position that
350 minimizes the square error is chosen as the threshold for each respective microbe. The step is
351 placed at the largest jump from low values to high values and sets the threshold at the point where
352 the step crosses the original data. The microbe counts are normalized and transformed into $\log_2$
353 scale before Boolean analysis. Microbe counts (in $\log_2$ scale) are further classified as either 'high',
354 'low', or 'intermediate'. If t is the microbe count threshold, levels above $t + 0.5$ are 'high', levels
355 below $t - 0.5$ are 'low', and levels between $t - 0.5$ and $t + 0.5$ are 'intermediate'. Points in the
356 intermediate region are ignored because these points might appear on either side of the threshold
357 due to noise.
358
359 **Boolean Implication Analysis**
360 There are six possible Boolean implications: symmetric (opposite and equivalent) or asymmetric
361 (low → low, low → high, high → low, high → high). The asymmetric relationships are determined
362 by checking if one of the four quadrants in the scatter plot is significantly sparse compared with
363 other quadrants. If A low → B low and A high → B high are both sparsely populated, then A is
364 equivalent to B. If A high → B low and A low → B high are both sparsely populated, then A is
365 opposite to B. The BooleanNet statistic tests (27) determine whether there is a Boolean relationship
366 between A and B. Consider the relationship A low → B high. First, test if the microbe counts in
367 the sparse quadrant are significantly less than the expected counts in an independence model. Let
368 a_{00}, a_{01}, a_{10}, and a_{11} represent the quadrants in which the microbe counts of A and B are low and
369 low, low and high, high and low, and high and high, respectively.

370

371 $total = a_{00} + a_{01} + a_{10} + a_{11}$

372 number of A low counts = $nA_{low} = (a_{00} + a_{01})$

373 number of B low counts = $nB_{low} = (a_{00} + a_{10})$

374 $expected = \left(\frac{nA_{low}}{total} \times \frac{nB_{low}}{total}\right) \times\ total = (a_{00} + a_{01}) \times \frac{(a_{00} + a_{10})}{total}$

375 $observed = a_{00}$

376 $S\ statistic = \frac{expected - observed}{\sqrt{expected}}$

377

378 Second, the observed values in the sparse quadrant are not ideal for Boolean implication formula.
379 They are assumed to be erroneous points for the purpose of analysis only as was described
380 previously in the context of gene expression analyses (27). However, these points may or may not
381 be erroneous from a real biological point of view. We wanted to discover the general trends of
382 Boolean implication relationships with such a strong assumption. A maximum likelihood estimate
383 of this error rate is then computed:

384

385 $errorrate = \frac{1}{2}\left(\frac{a_{00}}{a_{00}+a_{01}} + \frac{a_{00}}{a_{00}+a_{10}}\right)$

386

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

391

392 The OTU tables are then uploaded onto Hegemon to visualize the Boolean relationships on scatter
393 plots for comparing two microbes against each other. In each graph, one microbe species' counts
394 (using OTU ID A) is plotted on the x-axis, and another microbe species' counts (OTU ID B) is
395 plotted on the y-axis. Each data point represents a sample and the counts are plotted on a log-log
396 scale. Using the graphs constructed on Hegemon, we can visually confirm if a Boolean implication
397 relationship, determined using the BooleanNet statistics, is present.
398
399 **Metadata analysis**
400 We used a Hegemon function that calculates the differential analysis for specific factors in the
401 metadata using t-tests in R software framework (R version 3.4.4 - 2018-03-15). We then selected
402 the OTU IDs that had higher mean differential values and higher $-\log_{10}(p)$ values. After the list of
403 potential IDs were generated, we visually confirmed on Hegemon whether certain factors such as
404 environment versus animal or body site affects the microbe counts.
405
406 **Computation of False Discovery Rate**
407 To evaluate the significance of the Boolean implication relationships found, we computed the FDR
408 for each of the pooled datasets by randomly permuting the counts for each microbe independently
409 10 times and determining the Boolean relationships for this randomized dataset using the method
410 described above. The FDR is the ratio of the average number of Boolean relationships in the
411 randomized dataset to the original dataset.
412
413 **Correlation analysis**

414 Correlation analysis was conducted on a subset of 500 microbes from the main dataset. The
415 Pearson correlation coefficients and Boolean relationships for each pair of microbes were
416 calculated to compare these two methods of analysis. The package "pearsonr" from the python
417 library "scipy.stats" was used to perform the calculations.
418

419 ## List of Abbreviations

420 CD: Crohn's Disease
421 FDR: False Discovery Rate
422 IBD: Inflammatory Bowel Disease
423 IBS: Irritable Bowel Syndrome
424 QIIME: Quantitative Insights Into Microbial Ecology
425 OTU: Operational Taxonomic Units
426 UC: Ulcerative Colitis
427 VAMPS: Visualization and Analysis of Microbial Population Structures
428

429 ## Declarations

430 **Ethics approval and consent to participant**

431 Publicly available data were generated after appropriate ethics approval and consent to participate.
432

433 **Consent for publication**

434 Not Applicable.
435

436 **Availability of data and materials**

437 All datasets are publicly available from https://qiita.ucsd.edu, under the Study, and View Studies
438 tab. We selected a subset of all the available studies. The specific list of studies used for this
439 research is included as an additional file.
440

441 **Competing interests**

442 The authors declare that they have no competing interests.
443

444 **Funding**

445 DS is supported by NIH UG3TR003355, UG3TR002968, R01AI155696, R01GM138385 and
446 R00CA151673, Hyundai Hope On Wheels, Padres Pedal the Cause/Rady Children's Hospital
447 Translational PEDIATRIC Cancer Research Award (Padres Pedal the Cause/RADY #PTC2017),
448 2017, Padres Pedal the Cause/C3 Collaborative Translational Cancer Research Award (San Diego
449 NCI Cancer Centers Council [C3] #PTC2017). DS is the corresponding author in this manuscript
450 and the grant organizations did not play a direct role in this study.

451 **Authors' contributions**

452 DV, SCS, and SS collected the data, performed the analysis, and wrote the paper. DS conceived
453 and designed the analysis, contributed analysis tools, supervised, acquired funding, reviewed, and
454 edited the paper. All authors have read and approved the manuscript.

455 **Acknowledgements**

456 The authors thank Dr. Christine Alvarado for her guidance in the early stages of this project and
457 the members of UCSD's Boolean Lab for reviewing our work and providing constructive
458 feedback.
459

460 **References**

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

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

465 3. Steinway SN, Biggs MB, Loughran TP, Papin JA, Albert R. Inference of Network
466 Dynamics and Metabolic Interactions in the Gut Microbiome. PLoS Comput Biol.
467 2015;11(5):e1004338.

468 4. Chen JY, Miyanishi M, Wang SK, Yamazaki S, Sinha R, Kao KS, et al. Hoxb5 marks long-
469 term haematopoietic stem cells and reveals a homogenous perivascular niche. Nature.
470 2016;530(7589):223-7.

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

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

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

478 8. Volkmer JP, Sahoo D, Chin RK, Ho PL, Tang C, Kurtova AV, et al. Three differentiation
479 states risk-stratify bladder cancer into distinct subtypes. Proc Natl Acad Sci U S A.
480 2012;109(6):2078-83.

481 9. Sahoo D, Wei W, Auman H, Hurtado-Coll A, Carroll PR, Fazli L, et al. Boolean analysis
482 identifies CD38 as a biomarker of aggressive localized prostate cancer. Oncotarget.
483 2018;9(5):6550-61.

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

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

489 12. Derelle R, López-García P, Timpano H, Moreira D. A Phylogenomic Framework to Study
490 the Diversity and Evolution of Stramenopiles (=Heterokonts). Mol Biol Evol. 2016;33(11):2890-
491 8.

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

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

498 15. Ramsey MM, Freire MO, Gabrilska RA, Rumbaugh KP, Lemon KP. Staphylococcus
499 aureus Shifts toward Commensalism in Response to Corynebacterium Species. Front Microbiol.
500 2016;7:1230.

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

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

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

508 19. Pittayanon R, Lau JT, Yuan Y, Leontiadis GI, Tse F, Surette M, et al. Gut Microbiota in
509 Patients With Irritable Bowel Syndrome-A Systematic Review. Gastroenterology.
510 2019;157(1):97-108.

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

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

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

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

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

525 25. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a
526 chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ
527 Microbiol. 2006;72(7):5069-72.

528 26. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann
529 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 were found in our dataset. (a) Describes the StepMiner algorithm that creates thresholds for each microbe. Among all samples, the normalized counts (log2 scale) of a particular microbe are sorted, and a step function is fitted where the sharpest

557 change between low microbe count and high microbe count takes place. The midpoint of the step
558 position that minimizes the square error is chosen as the threshold (red line) for each respective
559 microbe. A noise margin of +/- 0.5 is considered around the threshold which are ignored for
560 Boolean analysis. (b) Depicts a log-log plot of the number of each type of relationship in the main
561 dataset and the corresponding number of microbes that exhibit that specific relationship. Each of
562 the four types of relationships found in our datasets are shown. Each point in the scatter plot
563 corresponds to a sample, where the two axes represent the counts of each microbe. (c) high → low
564 (d) low → low (e) high → high (f) equivalent. The remaining diagrams (g-j) show the correlation
565 distributions according to the four Boolean implication relationships found in our datasets, with
566 correlation on the x-axis and number of microbe pairs on the y-axis.
567

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

570 Analysis of scatter plots with various experimental conditions using metadata files that provided
571 additional information about the samples. *Section 1*: green represents environmental samples
572 (plants, water, soil, etc.) and red indicates animal samples (humans, animals). (a) *Polynucleobacter*
573 (145533) low → *Candidatus Xiphinematobacter* (786420) low; this relationship is only present in
574 environmental microbiomes due to the lack of red samples in the plot. (b) *Polynucleobacter*
575 (3071019) high → *Bacteroides uniformis* (197072) low*;* this relationship suggests
576 *Polynucleobacter* is mainly present in the environment, and *Bacteroides uniformis* is mainly
577 present in animals. *Section 2*: (c) and (d) have the same microbes on the axes *Staphylococcus*
578 *aureus* (446058) and *Corynebacterium* (1000986), but different regions of the body plotted: skin
579 (dark blue) and feces (green). (c) shows the relationship *S. aureus* low → *Corynebacterium* low

580 holds for the skin region. (d) shows the relationship using fecal samples, and there is no clear
581 relationship that can be determined from this. *Section 3*: Pink represents Crohn's Disease (CD),
582 teal represents Ulcerative Colitis (UC), and light gray represents neither disease. (e) The
583 relationship *Actinomyces* (12574) high → *Lachnospiraceae* (4469576) low is shown, with higher
584 counts of *Lachnospiraceae* in CD, and higher counts of *Actinomyces* in UC. (f) shows the
585 relationship *Streptococcus* (4467992) high → *Lachnospiraceae* (4469576) low, with higher counts
586 of *Lachnospiraceae* in CD, while higher counts of Streptococcus in UC. *Section 4*: Magenta
587 represents eczema, blue represents psoriasis, and beige represents neither skin condition. (g) The
588 relationship *Acinetobacter johnsonii* (4482374) low → *Corynebacterium* (361600) low is shown.
589 *Patients with psoriasis tend to have higher counts of Corynebacterium* than patients with eczema.
590 (h) The relationship *Ruminococcaceae* (4346675) high → *Anaerococcus* (927089) low is shown.
591 Patients with psoriasis tend to have higher counts of *Anaerococcus*, while patients with eczema
592 tend to have higher counts of *Ruminococcaceae*.

Figures

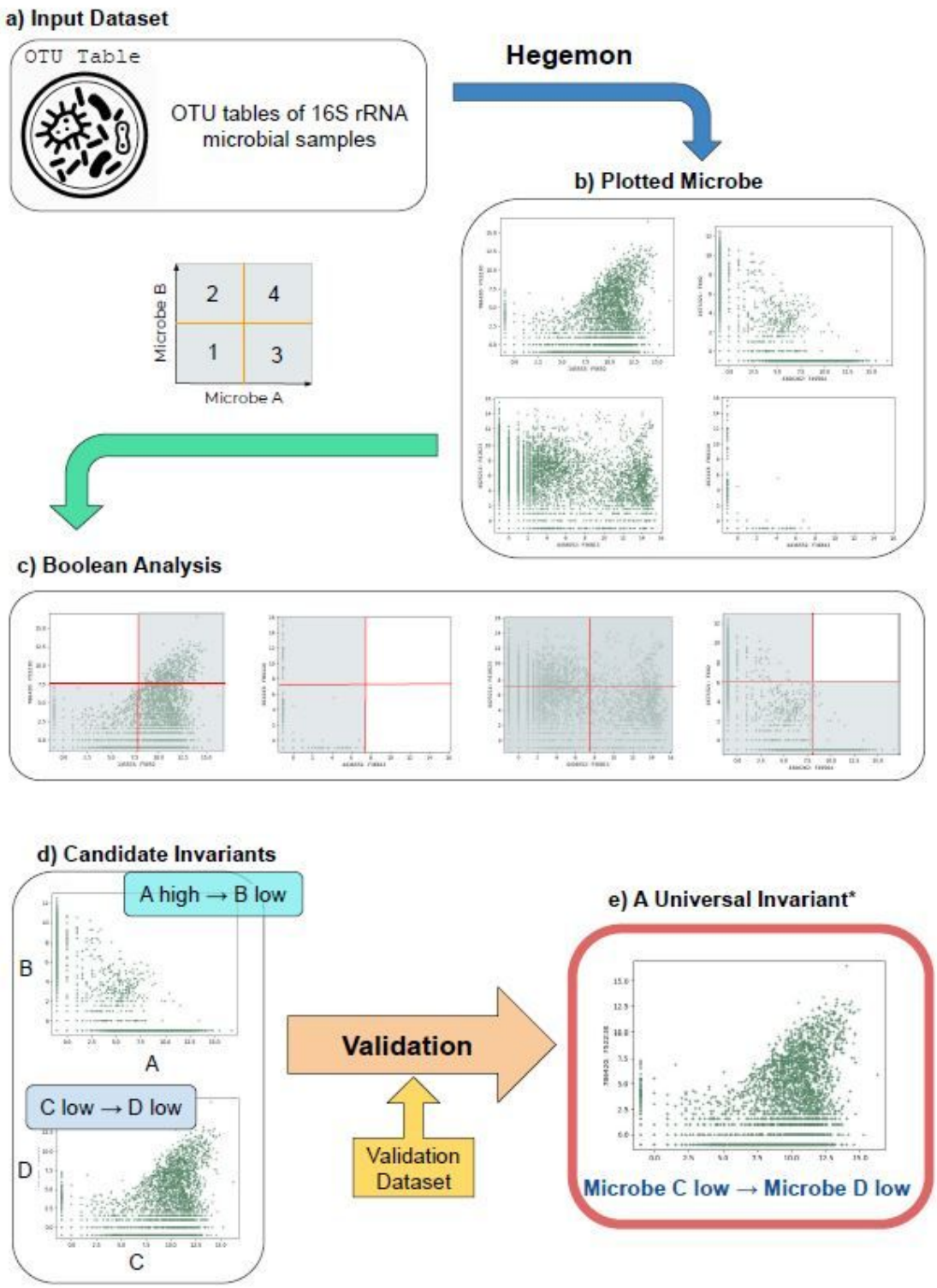

Figure 1

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.

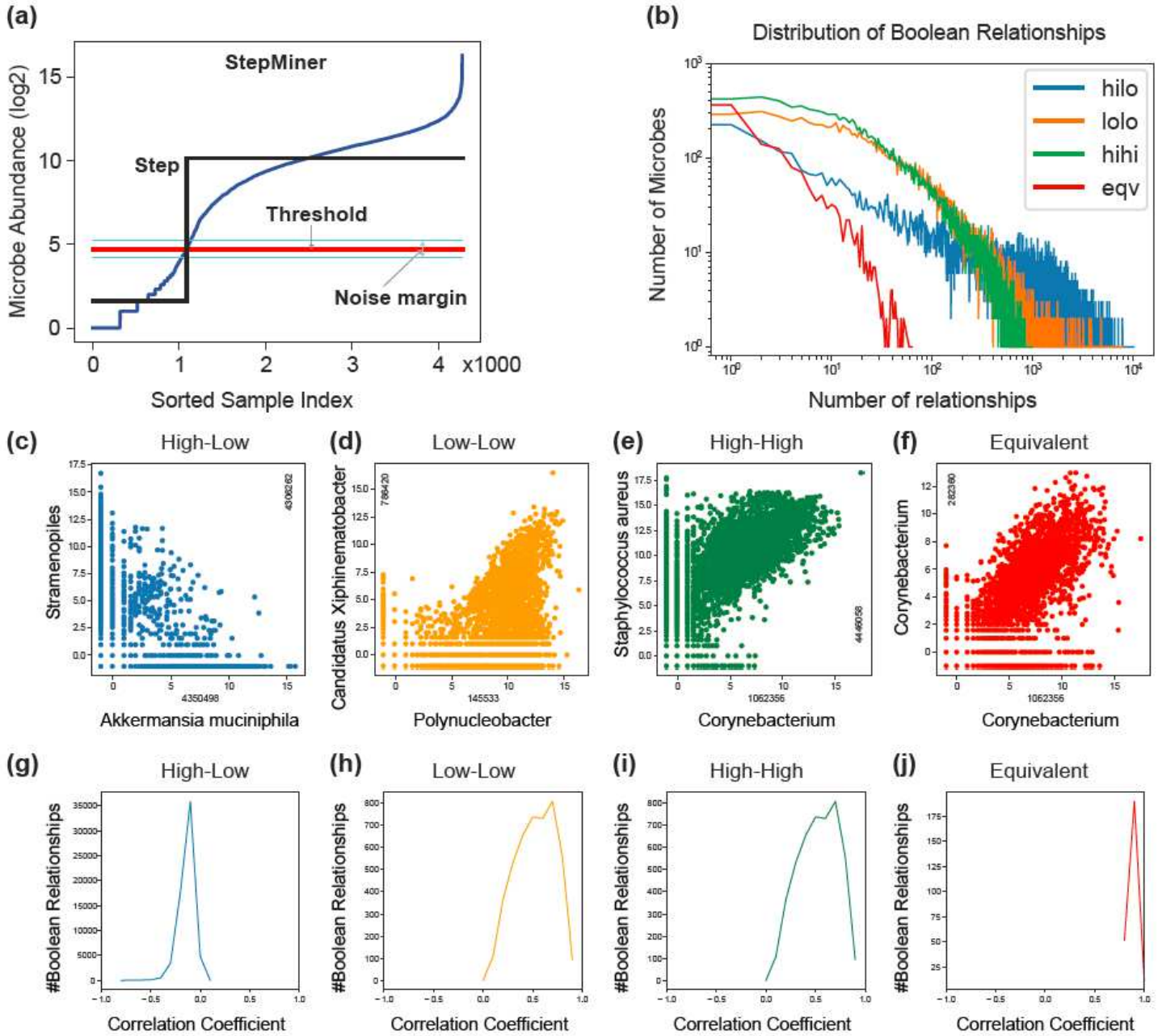

Figure 2

Boolean implication relationships represent diversity in microbiome data All four types of Boolean relationships were found in our dataset. (a) Describes the StepMiner algorithm that creates thresholds for each microbe. Among all samples, the normalized counts (log2 scale) 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 (red line) for each respective microbe. A noise margin of +/- 0.5 is considered around the threshold which are ignored for Boolean analysis. (b) Depicts a log-log plot of the number of each type of relationship in the main dataset and the corresponding number of microbes that exhibit that specific relationship. Each of the four types of relationships found in our datasets are shown. Each point in the scatter plot corresponds to a sample, where the two axes represent the counts of each microbe. (c) high ⊠ low (d) low ⊠ low (e) high ⊠ high (f) equivalent. The remaining diagrams (g-j) show the correlation distributions according to the four Boolean implication relationships found in our datasets, with correlation on the x-axis and number of microbe pairs on the y-axis.

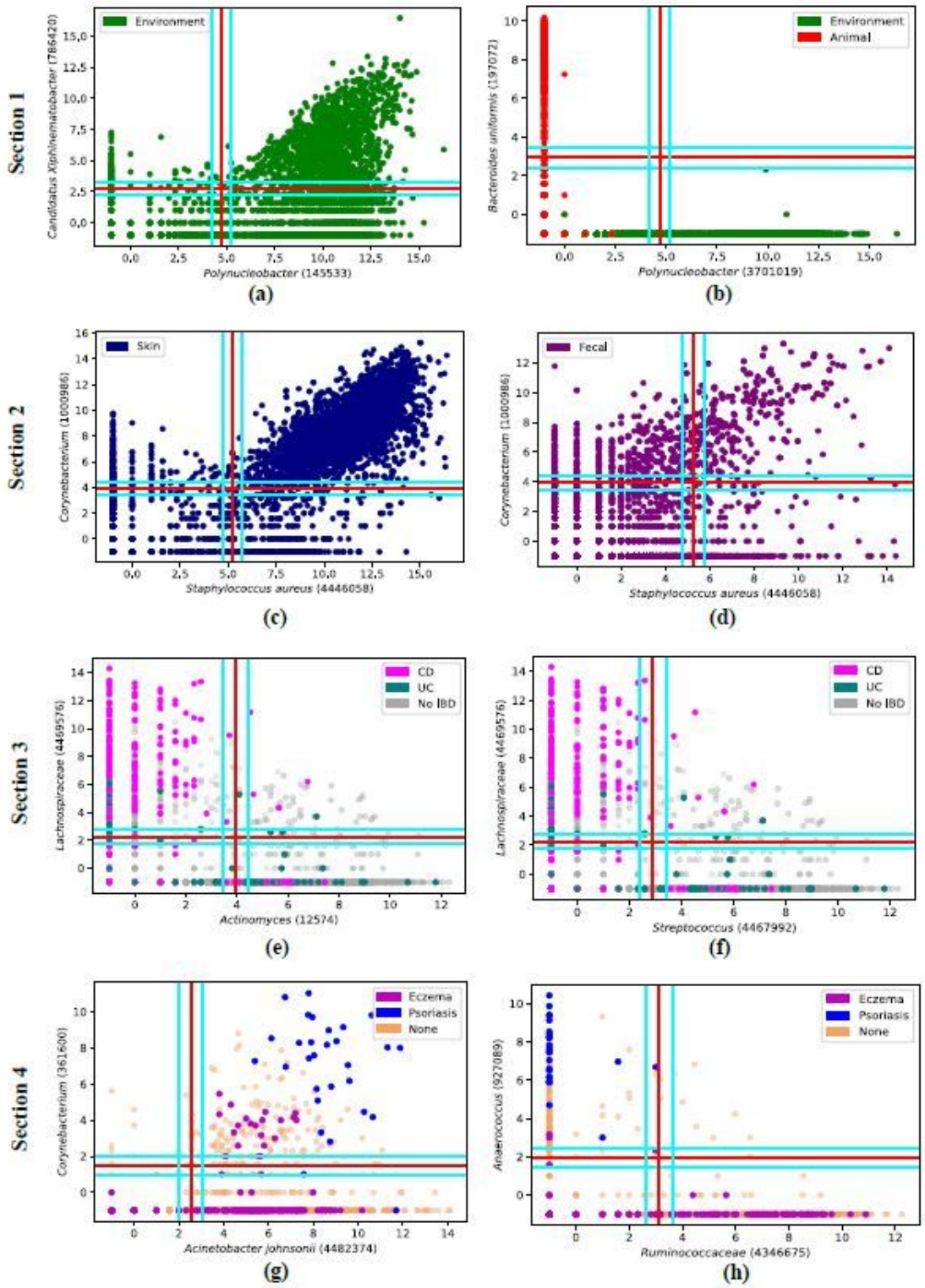

Figure 3

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 no clear relationship 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Qiitastudies.csv