Constructing a Boolean implication network to study the interactions between environmental factors and OTUs Supplementary materials

Congmin Zhu, Rui Jiang, Ting Chen

Supplementary Text

Simulation models

Let X and Y be two factors sampled from the real data and α and β be fractions of low abundant samples of X and Y, respectively. Let p be the probability that both X and Y are low abundant. We have the joint probabilities

$$P(X=0,Y=0)=p, \qquad P(X=1,Y=0)=\beta-p,$$
$$P(X=0,Y=1)=\alpha-p, \qquad P(X=1,Y=1)=1-\alpha-\beta+p.$$

The range of the parameter p should be restricted since in a Boolean implication either one or two of the above joint probabilities should be significantly smaller than the others.

To generate the $X_{low}\rightarrow Y_{high}$ relationship, we should guarantee the values of the above four joint probabilities as follows and get the constraints of p which are shown in the following formulas:

$$\begin{cases} P(X=0,Y=0) < \mathrm{T}_1 \\ P(X=0,Y=1) > \mathrm{T}_2 \\ P(X=1,Y=0) > \mathrm{T}_2 \\ P(X=1,Y=1) > \mathrm{T}_2 \end{cases} \Rightarrow \begin{cases} p < \mathrm{T}_1 \\ p < \alpha - \mathrm{T}_2 \\ p < \beta - \mathrm{T}_2 \\ p > \beta + \alpha + \mathrm{T}_2 - 1 \end{cases}$$

To generate the $X_{low}\rightarrow Y_{low}$ relationship, we should guarantee the relationships between the above four joint probabilities as follows and get the constraints of p which are shown in the following formulas:

$$\begin{cases} P(X=0,Y=0) > \mathrm{T}_2 \\ P(X=0,Y=1) < \mathrm{T}_1 \\ P(X=1,Y=0) > \mathrm{T}_2 \\ P(X=1,Y=1) > \mathrm{T}_2 \end{cases} \Rightarrow \begin{cases} p > \mathrm{T}_2 \\ p > \alpha - \mathrm{T}_1 \\ p < \beta - \mathrm{T}_2 \\ p > \beta + \alpha + \mathrm{T}_2 - 1 \end{cases}$$

To generate the $X_{high} \rightarrow Y_{high}$ relationship, we should guarantee the relationships between the above four joint probabilities as follows and get the constraints of p which are shown in the following formulas:

$$\begin{cases} P(X=0,Y=0) > T_2 \\ P(X=0,Y=1) > T_2 \\ P(X=1,Y=0) < T_1 \\ P(X=1,Y=1) > T_2 \end{cases} \Rightarrow \begin{cases} p > T_2 \\ p < \alpha - T_2 \\ p > \beta - T_1 \\ p > \beta + \alpha + T_2 - 1 \end{cases}$$

To generate the $X_{high} \rightarrow Y_{low}$ relationship, we should guarantee the relationships between the above four joint probabilities as follows and get the constraints of p which are shown in the following formulas:

$$\begin{cases} P(X=0,Y=0) > T_2 \\ P(X=0,Y=1) > T_2 \\ P(X=1,Y=0) > T_2 \\ P(X=1,Y=1) < T_1 \end{cases} \Rightarrow \begin{cases} p > T_2 \\ p < \alpha - T_2 \\ p < \beta - T_2 \\ p < \beta + \alpha + T_1 - 1 \end{cases}$$

To generate the X opposite to Y relationship, we should guarantee the relationships between the above four joint probabilities as follows and get the constraints of p which are shown in the following formulas:

$$\begin{cases} P(X=0,Y=0) < T_1 \\ P(X=0,Y=1) > T_2 \\ P(X=1,Y=0) > T_2 \\ P(X=1,Y=1) < T_1 \end{cases} \Rightarrow \begin{cases} p < T_1 \\ p < \alpha - T_2 \\ p < \beta - T_2 \\ p < \beta + \alpha + T_1 - 1 \end{cases}$$

To generate the X equivalent to Y relationship, we should guarantee the relationships between the above four joint probabilities as follows and get the constraints of p which are shown in the following formulas:

$$\begin{cases} P(X=0,Y=0) > T_2 \\ P(X=0,Y=1) < T_1 \\ P(X=1,Y=0) < T_1 \\ P(X=1,Y=1) > T_2 \end{cases} \Rightarrow \begin{cases} p > T_2 \\ p > \alpha - T_1 \\ p > \beta - T_1 \\ p > \beta + \alpha + T_2 - 1 \end{cases}$$

We can therefore sample a p uniformly from its range to simulate the joint probabilities of a positive case and further generate a number of n (the number of samples in the real data) points according to these probabilities. Furthermore, to get continuous abundance data, we sample from two normal distributions $N(\mu_1,1)$ and $N(\mu_2,1)$ ($\mu_1 < \mu_2$) for the low and high abundant points, respectively. Noting that the abundance values cannot be negative numbers, so during the sampling the negative numbers are ignored.

Supplementary

Figures

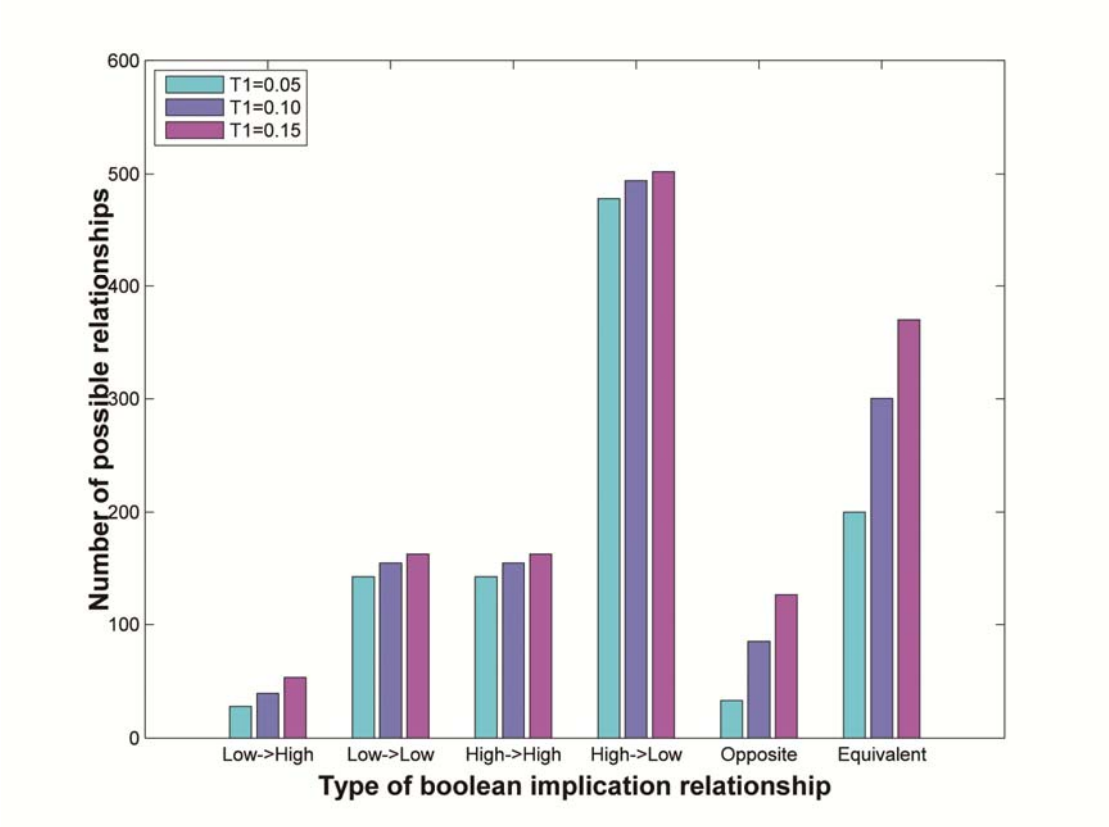

Figure S1. Number of possible α and β for generating each kind of Boolean implication relationships in the simulation studies. When given different value of T_1, we enumerated pairwise combination of these 209 factors and analyzed the number of possible Boolean implications of each type, based on the constraints of p derived in the Simulation models. We find that the numbers of possible low→high and opposite implications are relatively small, while that of the high→low implication is relatively large.

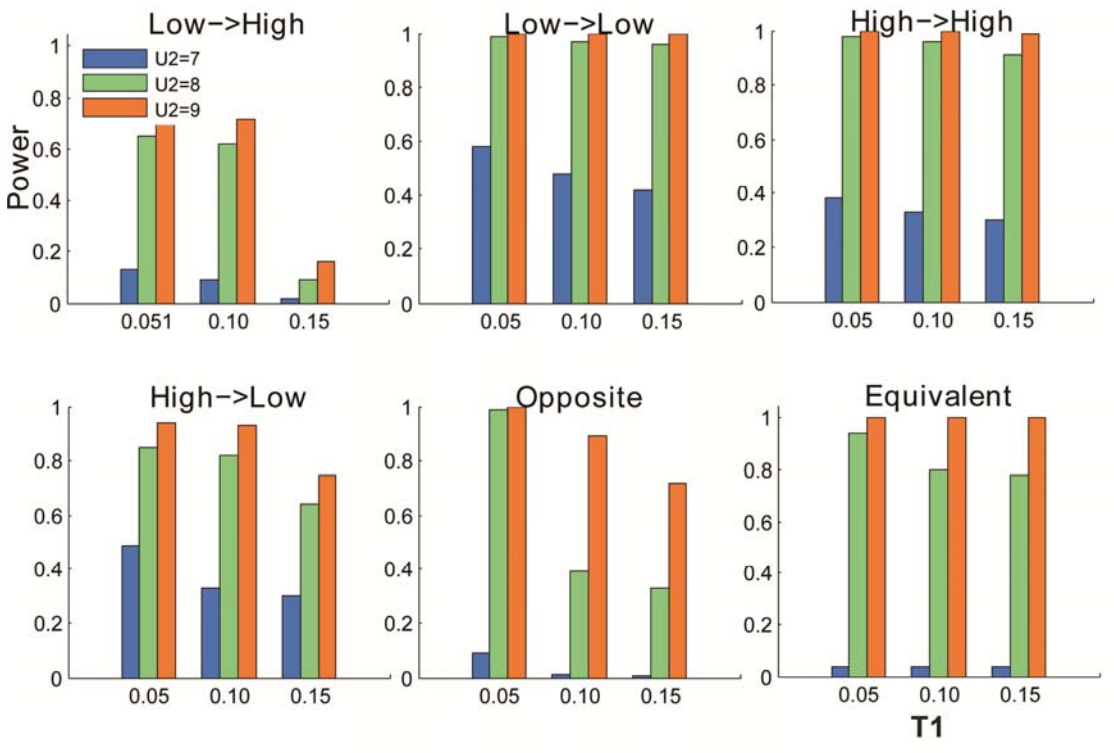

Figure S2. The power changes with μ_2 and T_1 in simulation studies. Setting different values of T_1 and μ_2, we generate 100 positive relationships for each type of Boolean implication and respectively mixing them with 100 negative relationships. We then apply our method with Fisher's exact test as the first stage on the each simulated data to detect Boolean implication relationships. Bars here show that power of our method is apparently related to the parameters T_1 and μ_2 which are used to produce simulated data. With the increase of μ_2 the power of our method increases, while decrease with the increase of T_1.

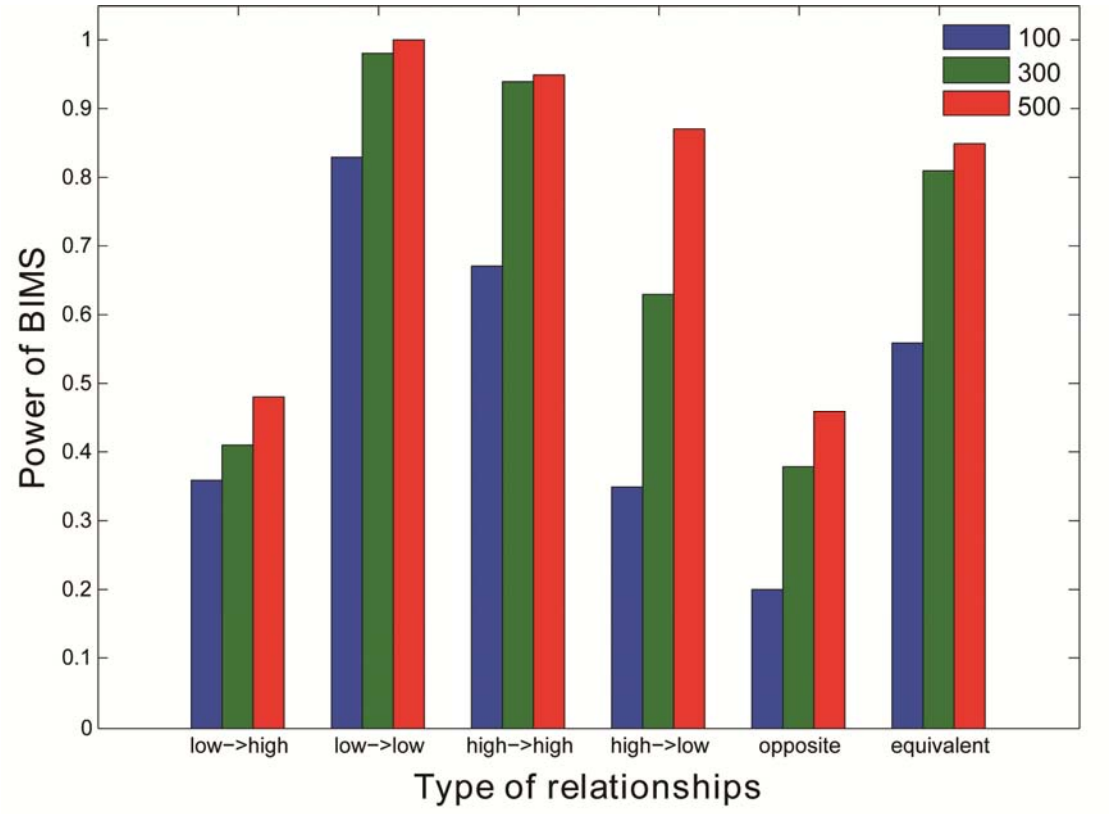

Figure S3. The power of BIMS changes with the sample size. Setting setting T_1=0.1 and μ_2=10 , we generate 100 positive relationships for each type of Boolean implication and respectively mixing them with 600 negative relationships. To explore the power of BIMS to the sample size, we further vary the sample number from 500 to 100 with step 200 and identify Boolean implications at the false discovery rate of 0.01 in each situation. Bars here show that power of our method is apparently related to the sample size. With the increase of sample size, the power of our method increases.

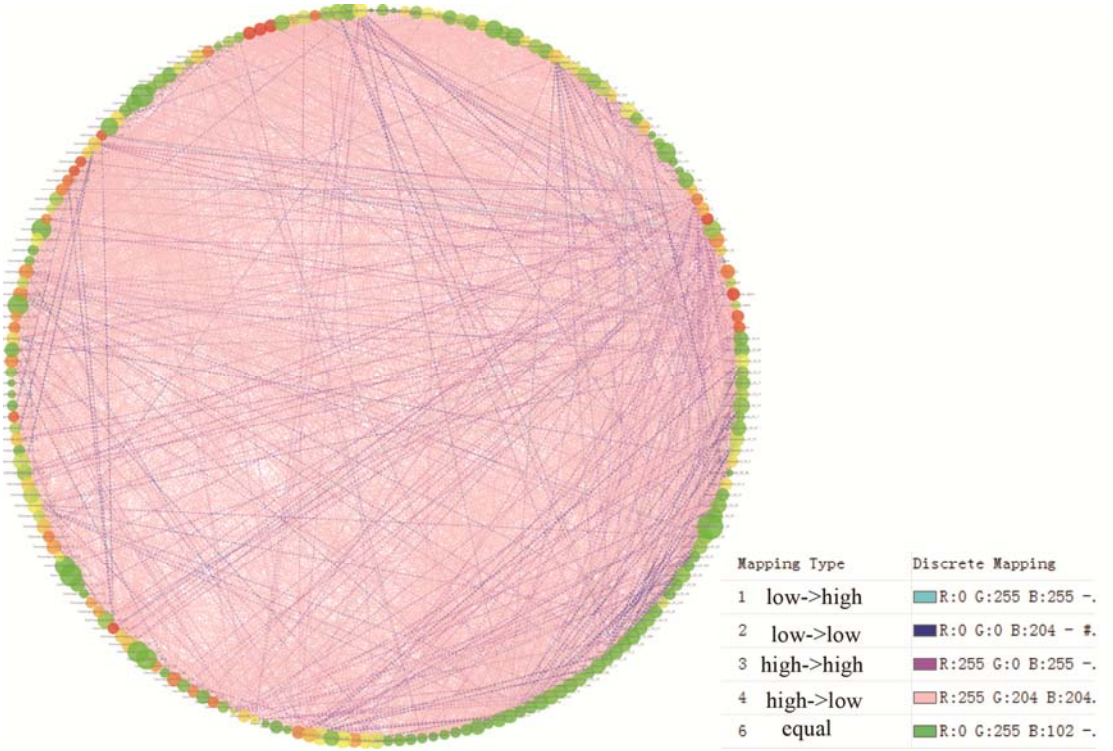

Figure S4. Circular network showing all significant Boolean implication relationships (FDR<0.01) of marine microbes and environmental factors. Nodes are either OTUs or EFs, and directed edges with different colors represent different types of Boolean implication relationships. Nodes with bigger size represent factors with more neighbors connected. Node with brighter colours represent factors with bigger out-degree.

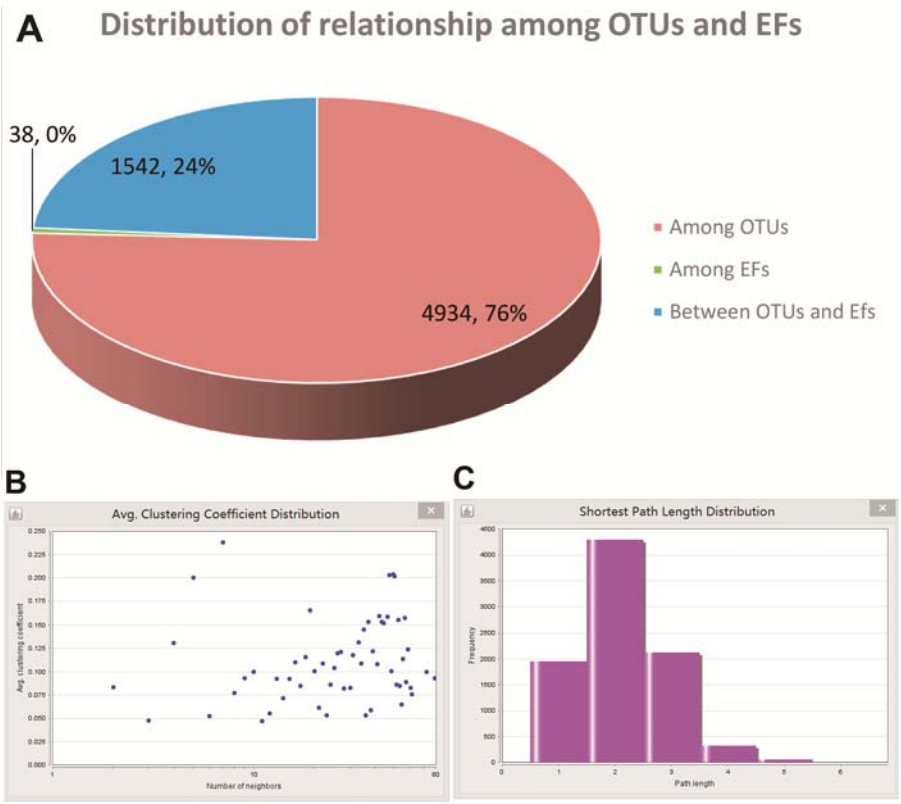

Figure S5. Property analysis of pairwise Boolean implication network of all factors. (A) Distribution of all discovered relationship. From the pie we find that most of Boolean implication relationships found exist among OTUs. (B) The distribution of clustering coefficient. The average clustering coefficient is 0.120. (C) The distribution of shortest path length. The shortest path length are among 1 to 3, indicating most microbes in the resulted network are closely linked with high level of dependence. The characteristic path length of this network is close to the random characteristic path length of 2.516 for a random network with same scale while the clustering coefficient of this network is larger than the random clustering coefficient of 0.044 for a random network with same scale, which shows that the network has 'small world' properties.

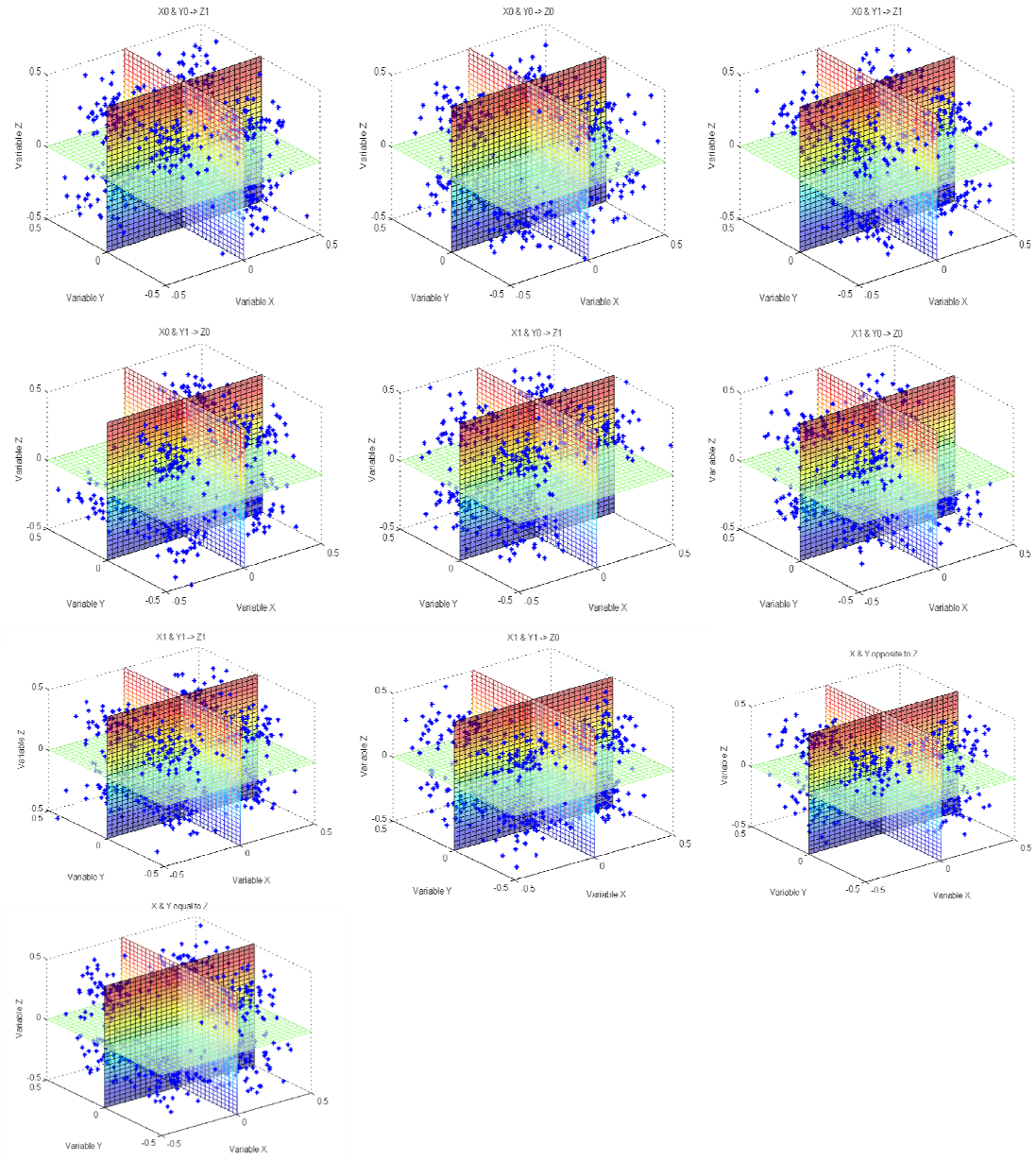

Figure S6. **Ten types of Boolean implication relationship in three-dimensional space illustrated with simulated data.** Each point in the scatter plot corresponds to a sample condition, where the three axes correspond to the abundance levels of three factors. From these figures we can get the intuitively corresponding relationships between the sparse situations of eight quadrants and the types of Boolean implication.

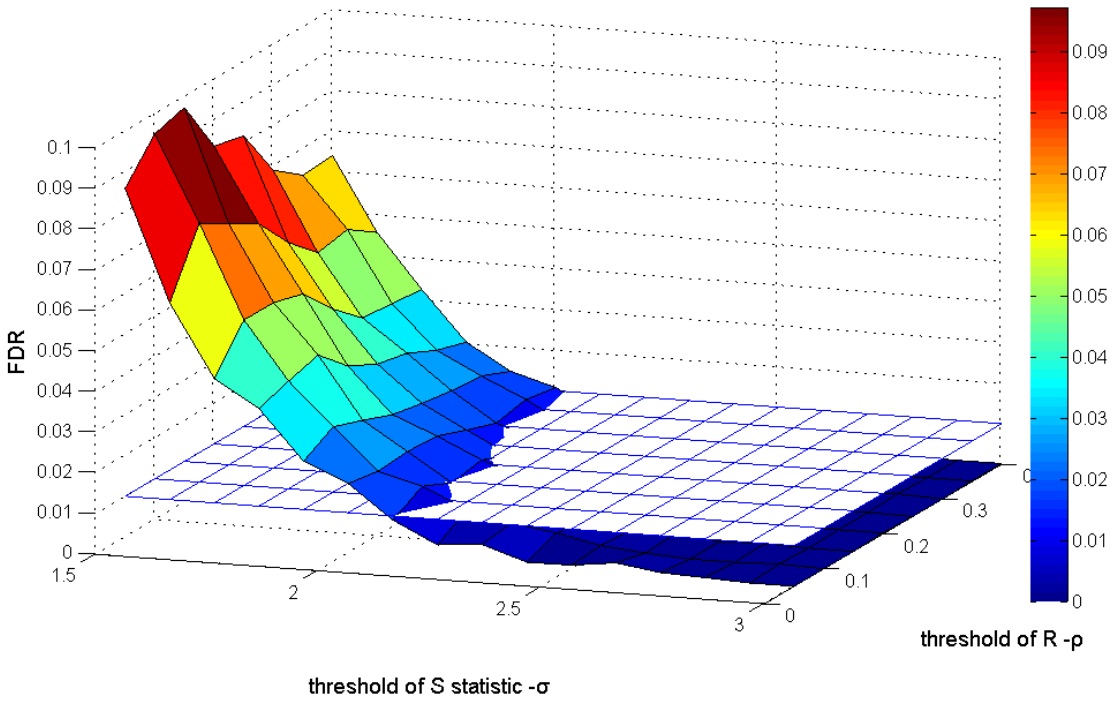

Figure S7 – An empirical relationship between FDR and thresholds σ and ρ. The surface in the figure shows an empirical relationship between FDR (z-axis) and thresholds σ (x-axis) and ρ (y-axis). From the figure we can see that when σ is between 2.0 and 3.0 while ρ is chosen as 0.1, the FDR of a network is typically small (< 0.01).

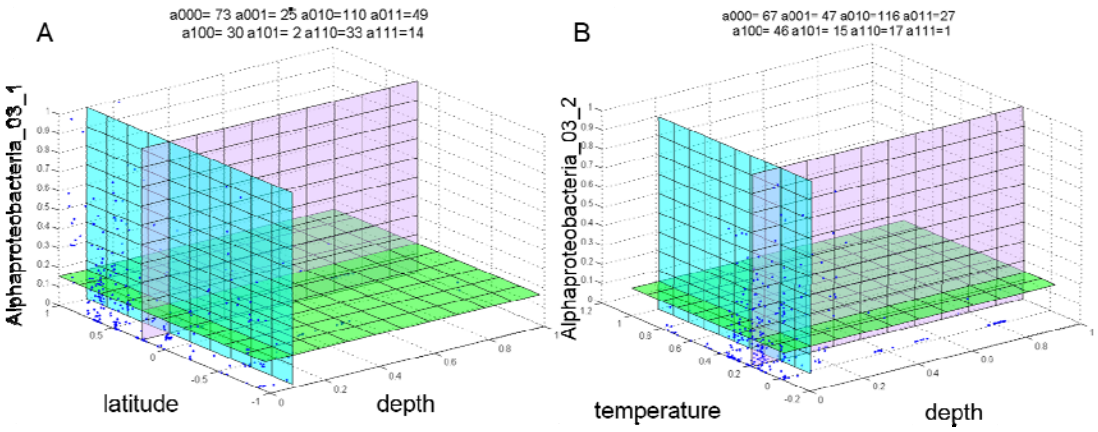

Figure S8 - Verify relationships found by Boolean implication method in three-dimensional space. A visual examination of the scatter plots is a straightforward way to check the quality of the results. (A) Depth high & latitude low →Alphaproteobacteria_03_1 low. The points in the high-low-high quadrant is less than other quadrants, so we can intuitively verify that when the depth is high and the latitude is low, the abundance of Alphaproteobacteria_03_1 is relatively low. (B) Depth high & temperature high→Alphaproteobacteria_03_2 low. Points in the high-high-high quadrant is less than other quadrant, so we can intuitively verify that when the depth is high and the temperature is relatively high, the abundance of Alphaproteobacteria_03_2 is relatively high too.

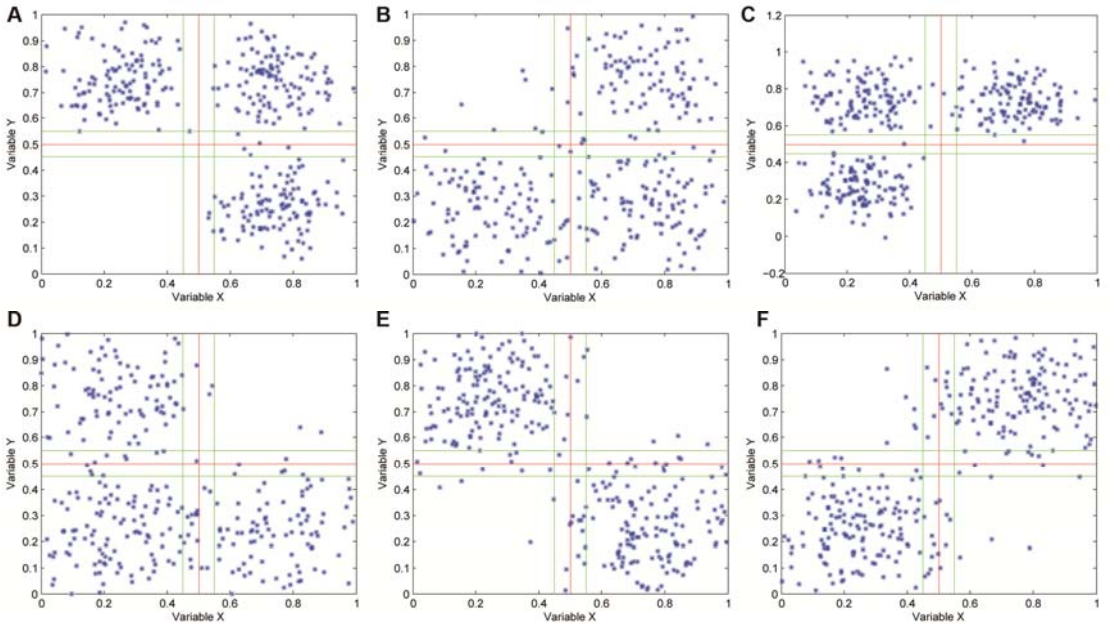

Figure S9. Six types of Boolean implications illustrated using simulated data. Each point in the scatter plot corresponds to a sample condition, where the two axes correspond to the normalized abundance levels of two factors (OTUs or EFs). Four asymmetric Boolean implication relationships (low→high, low→low, high→high, high→low) are shown (left-to-right and top-to-bottom) in (A-D). Two symmetric linear-shaped relationships (Boolean opposite and Boolean equivalent) are shown in (E-F).

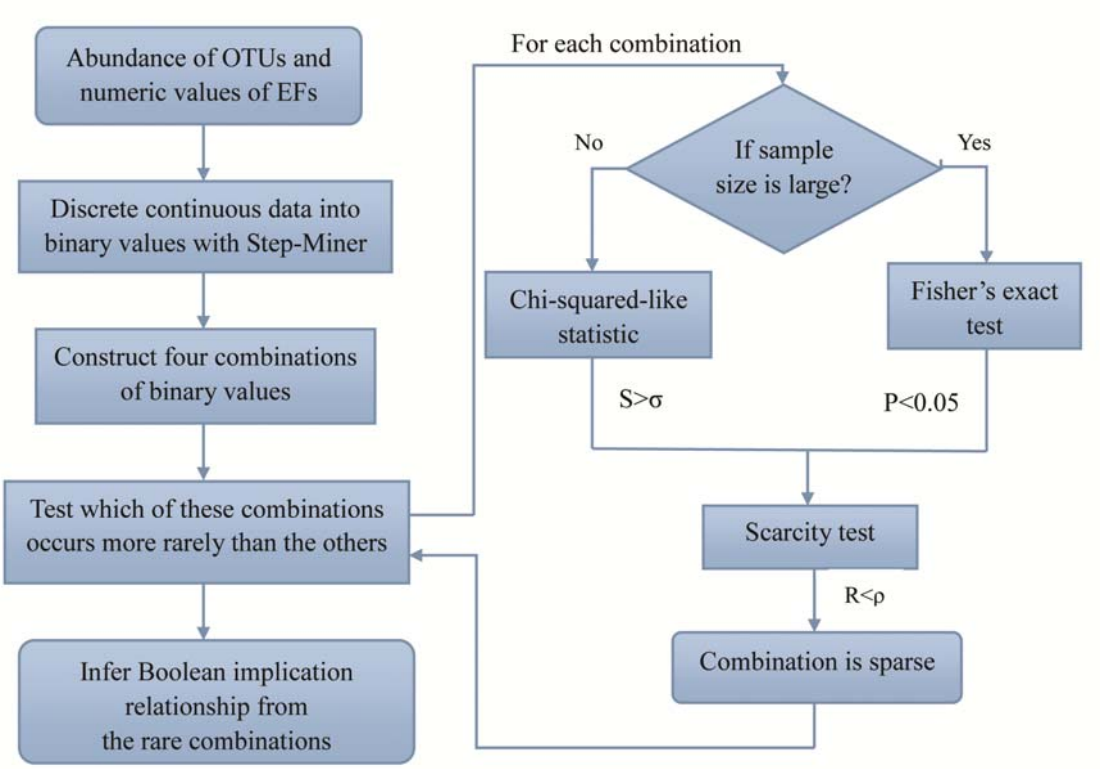

Figure S10. The process of detecting Boolean implication relationship between two factors. Firstly the abundance values of each factor (OTUs or EFs) are sorted and then a step function fitted (using StepMiner) to the sorted values minimizes the square error between the original and the fitted data. Given the discretized binary abundance levels of two factors, we construct a contingency table to represent occurrence frequencies of combinations of the discretized abundance levels for the two factors. We then adopt a two-phase hypothesis testing procedure for each combination to detect the Boolean implication between two factors by checking which combinations are sparse.

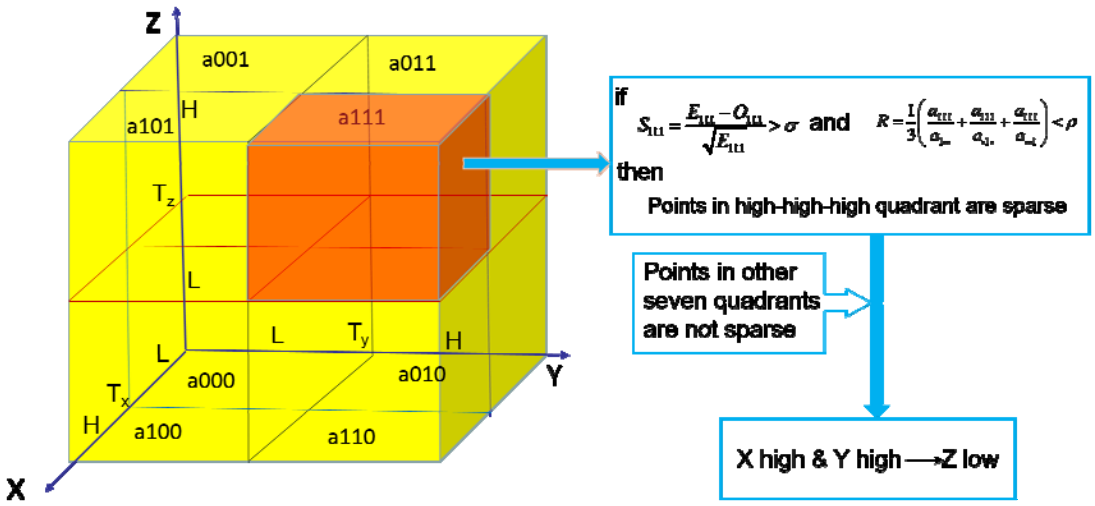

Figure S11. The process of detecting Boolean implication relationship among three factors. Given the discretized binary abundance levels of each factors using StepMiner, we firstly get eight combinations of the discretized abundance levels for the three factors, say low-low-low, low-low-high, low-high-low, low-high-high, high-low-low, high-low-high, high-high-low, high-high-high and respectively denote the numbers of samples that belong to the eight combinations as $a_{000}, a_{001}, a_{010}, a_{011}, a_{100}, a_{101}, a_{110}, a_{111}$. We then adopt a two-phase hypothesis testing procedure to detect the Boolean implication among three factors by checking which combinations are sparse. Then we adopt chi-squared-like statistic to detect whether there exists a non-random association between the three factors and resort to a sparsity test to detect which type of Boolean implication can best describe the relationship between the three factors.