SUPPLEMENTARY NOTES:

Supplementary note for technology choice - Different underlying distributions significantly alter edge inferences

We tested the direct impact of different sequencing technologies on OTU distributions themselves in similarly processed sample replicates. Using technical replicates from an arthropod microbiome(Ponnusamy et al 2014), and Illumina HiSeq vs. MiSeq sample replicates from a gut microbiome(Caporaso et al 2012), we tested each HiSeq feature compared to the same MiSeq feature by the Kolmogorov-Smirnov (KS) test(Kolmogorov 1933, N 1948) and found no significant differences in feature count distributions. However, in contrast, using samples generated from 454 and replicates generated with Illumina(Fisher 1921, Yatsunenko et al 2012) we found on average 17% of shared features had significantly different count distributions (Benjamini-Hochberg $p \leq 0.05$). To investigate further, we processed the 454 and Illumina datasets from Yatsunenko et al. (Yatsunenko et al 2012) with the same protocol, removed OTUs that were not shared between the technologies, and calculated the fraction of correlated OTU pairs in common between the technologies for all co-occurrence techniques. We found that most techniques (with the exception of Bray-Curtis, which is more robust to differences in the fixed sum of sequences (Legendre et al 1998)) found < 15% of the same correlated pairs (Supplementary Fig. 15a). Given the poor agreement between the networks constructed with 454 and Illumina technologies, we tested the impact of distribution alone on the tool performance using the copula model(Trivedi and Zimmer 2007) (Methods). This model simulates contingency tables with the same covariance structure but different marginal distributions so that feature pairs have the same ranked correlation. Using distributions with many zeros that are often used to model microbiome data(McDonald et al 2012, Paulson et al 2013), such as lognormal, as well as ones mimicking bacterial growth, such as exponential, we found that distribution has less of an impact for those tools that use a rank-based correlation measure like MIC, and Spearman (Supplementary Fig. 18b, Supplementary Fig. 19), in agreement with Supplementary Fig. 18a.

Supplementary note for normalization

After sequencing and assembling a table of OTU sequence counts (OTU table), the next analysis step is 'normalization' of the data to account for differences in sample sequencing efforts, data sparsity, the limited number of rRNA sequences per sample (compositionality), and extremely rare features whose counts are especially uncertain(Anders and Huber 2010, Friedman and Alm 2012, Paulson et al 2013). Depending on the technique employed, normalization can address some but not all of the first three challenges. The most common normalization approach is 'rarefying'. Rarefying's strength as a normalization technique lies in addressing different column sums and sparse data well. However, it throws out data, does not attempt to correct for the compositional nature of the data, and due to the random nature of the subsampling, a small amount of variance is introduced into the rarefied data table on different trials(McMurdie and Holmes 2014). Therefore, we conducted 20 rarefactions (10 at 1000 and 10 at 2000 sequences/sample) and compared the detection profiles of the tools using data from Ridaura et al.(Ridaura et al 2013), who discovered a causal link between gut microbial community composition and the obesity phenotype. The fraction of edges inferred in all 10 rarefactions (at a given depth, for a given tool) was under 0.6,

suggesting that most tools are very sensitive to small count perturbations, with MIC and CoNet being notable exceptions (Supplementary Figs. 2 and 3).

Also, since rarefaction reduces the number of species present by subsampling, this suggests that rarefying at a lower depth may intensify compositionality effects on correlation networks. However, the main effect is a decrease in the number of edges found (Fig. 2, Supplementary Fig. 4).

Supplementary note for ecological data

We assessed the tools on their ability to detect simple two-species ecological relationships (two features, one edge) when the data were presented as unaltered (tables 1.6 and 1.7), compositional (table 2.16), sparse (table 2.17), or sparse and compositional (table 2.18) (Supplementary Table 3), to maximally confound the tools. In general, the tools performed reasonably well but precision was low—on average 0.25 for tables with 40% sparsity, and 0.01 for tables with 70% sparsity (tables 2.17 and 2.18). For instance, using the common p-value threshold of 0.05 for p-values calculated from Spearman correlation with Fisher z-transformation (SZ) resulted in a precision of 0.021 (table 2.16): for every correctly detected edge in this network there would be 50 incorrect edges. For unaltered or compositional data, LSA and MIC were the most precise by far (with precision 0.54, and 0.79 respectively), but this degraded when sparsity was added. A combination of tools was the most precise for tables with realistic sparsity levels (tables 2.17 and 2.18). Specificity was fairly high with an average close to one across all tools and ecological tables. Sensitivity was relatively low, with an average of 0.22 for tables with approximately 40% sparsity (tables 1.6, 1.7 and 2.16), and 0.03 for tables with 70% sparsity (tables 2.17 and 2.18).

In these ecological comparisons, we also assessed the performance of the tools on different types of ecological relationships. The detection profiles for the different ecological relationships were striking, with amensal and partial obligate-syntrophic relationships virtually undetectable by any tool and mutual relationships detectable by all tools (Fig. 3b, Fig. 4b and c, Supplementary Fig. 12b and c). To determine if the 'strength' of a relationship played a role in its detectability, our unaltered data (mentioned above) contained 90 relationships for each of the ecological relationships (e.g. 90 different OTU pairs related in a mutualistic way) split into 3 groups of 30 that were each generated with different strengths (higher strength corresponded to more change from the background distribution and a cleaner signal). For amensal edges, only SparCC and SZ with permissive p-value thresholds detected more than ~10% of all available edges. Furthermore, in contrast to the other relationships types, there was no correspondence between the strength of the edge relationship and the detection probability. For competitive edges, SparCC, LSA, and SZ all performed well, and detected more edges as the strength of relationship increased. CoNet, RMT, and Pearson with Fisher-Z transform (PZ) were unable to detect competitive relationships. For commensal or mutually related edges, SparCC, LSA, SZ, and PZ performed well, with CoNet performing at an intermediate level and RMT finding no edges. Parasitic edges were best detected by PZ and SZ, and had intermediate detectability with the other tools. RMT did not find many edges, perhaps due to setting a stricter threshold on the PZ values.

We also tested detection profiles of the tools for more complex (but still linear) ecological relationships (Fig. 3b, Fig. 4a, Supplementary Table 3). In these relationships, we required two or more OTUs to be present to cause an interaction and a modification to a third OTU (or more). Ecological literature suggests that there are likely important relationships mediated by more than two members(Shade et al 2012) and we tested a simple case of this. In general, the detection profiles of the three-species relationships were similar to those in the two-species case. SparCC, LSA, and SZ more easily identified the three-species competitive relationships than their two-species counterparts (the same was true of PZ, but it had minimal detection of either). Parasitic three-species edges were identified well, but the correlation patterns were hard to interpret; edges which we *a priori* assumed would be assigned as negatively co-occurring were positive and vice versa. This suggests that the non-linearity of multiple OTUs interacting in a network can confound assignment. Mutual three-species OTUs were discovered with high efficiency by most tools. However, detection of any of the above two and three-species ecological relationship types decayed to little better than random guessing when the sparsity in the OTU table was raised to realistic levels.

The importance of deciding which tool is best at finding which relationships is clear when one considers the post-hoc way in which correlation networks are used. For example, given the knowledge that SparCC can detect competitive relationships more easily than amensal relationships, negative edges (mutual exclusions) in a SparCC-generated correlation network are more likely to result from competitive interactions between taxa rather than amensal ones.

Supplementary note for time-series discussion

Here we modeled simple temporal relationships between OTUs as signals with varying noise, amplitude offset, phase shift, frequency, and coupling. In these mixture model tables, composed of sine, cosine, saw-tooth, and logarithmic patterns, none of the tools exhibited a marked preference for a given signal type. For all tools, the most frequently detected co-occurring pairs stemmed from mixed signal types (e.g. co-occurring sine and square wave signals). This, together with the evidence next presented, indicates that the correlation techniques are not immune to temporal aliasing, or distinct signals beginning to look similar due to insufficient sampling. For example, sine waves with different periods will become indistinguishable if not sampled frequently enough in time (Gerber 2014).

Also, detected edges varied depending upon at which point in time/how many samples were taken of the fluctuating OTUs (Fig. 5). With fewer time points taken for a given 100 time-unit signal (76, 50 and 26 points respectively), the tools generally found fewer edges. The main exceptions to this were CoNet, RMT, Bray-Curtis, and MIC. Bray-Curtis and MIC found very few edges, suggesting that they are not very sensitive to time series relationships. This implies that different signals are construed from the actual signal depending on the sampling frequency, greatly affecting OTU pairs deemed to be co-occurring. CoNet and RMT were relatively stable across sampling frequencies.

Simple time-shifted OTU relationships were also tested. These data sets were composed of OTUs exhibiting a pulse (sharp increase in abundance for some of the time points) or envelope centered at different times (from 1 to 200). Most measures only considered OTUs displaying pulses at similar times as correlated (Supplementary Figs. 16-17). This is important insofar as two pulse signals that peak at day zero might be more easily detected as correlated than two signals with the same pulse shape but offset in phase. These OTU pulse tables were too sparse and/or noisy (due to high cutoffs used by the default criteria) for RMT to evaluate, even though in one table the pulse was sustained over 50 samples. Bray-Curtis and MIC did not detect OTUs exhibiting high frequency pulses, and did not distinguish between lower frequency time-shifted signals.

Supplementary note for tool timings

Rough estimates of correlation technique run time on local 64-bit machine, 1053 feature x 257 column OTU matrix.(Yatsunenko et al 2012) Pearson with fisher z-transform: 96s*, Spearman with fisher z-transform: 267s*, Bray-Curtis: 10s*, LSA: 6153s**, CoNet: 3826s(including 1000 permutations, 1000 bootstraps)**, MIC: 457s**, RMT: 284s***. SparCC correlations calculation: 107s*. Note that SparCC p-value calculations scale with the number of permutations, so for 100 permutations it takes 100*107 = 10700s*, or about 3 hours. Thus, it is much faster to threshold only by correlation value of 0.35, with no significant difference in results at a 0.01 or 0.001 p-value threshold (data not shown).

Timings were done with * 8GB, **16GB, ***64GB memory, respectively.

Anders S, Huber W (2010). Differential expression analysis for sequence count data. *Genome biology* **11:** R106.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N *et al* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal* **6:** 1621-1624.

Fisher RA (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* **1:** 3-32.

Friedman J, Alm EJ (2012). Inferring correlation networks from genomic survey data. *PLoS computational biology* **8:** e1002687.

Gerber GK (2014). The dynamic microbiome. *FEBS letters* **588:** 4131-4139.

Kolmogorov AN (1933). Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell' Instituto Italiano delgi Attuari* **4:** 83-91.

Legendre P, Legendre L, Legendre L, Legendre L (1998). *Numerical ecology*, 2nd English edn. Elsevier: Amsterdam ; New York.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A *et al* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *Isme Journal* **6:** 610-618.

McMurdie PJ, Holmes S (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS computational biology* **10**.

N S (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19:** 279-281.

Paulson JN, Stine OC, Bravo HC, Pop M (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10:** 1200-1202.

Ponnusamy L, Gonzalez A, Van Treuren W, Weiss S, Parobek CM, Juliano JJ *et al* (2014). Diversity of Rickettsiales in the microbiome of the lone star tick, Amblyomma americanum. *Applied and environmental microbiology* **80:** 354-359.

Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL *et al* (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341:** 1241214.

Shade A, Peter H, Allison SD, Baho DL, Berga M, Burgmann H *et al* (2012). Fundamentals of microbial community resistance and resilience. *Frontiers in microbiology* **3:** 417.

Trivedi PK, Zimmer DM (2007). *Copula modeling : an introduction for practitioners*. Now: Boston.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M *et al* (2012). Human gut microbiome viewed across age and geography. *Nature* **486:** 222-227.