Tool Details

CoNet(Faust et al 2012):

All OTUs occurring in less than one third of the samples were discarded (except for table set 3, where minimum occurrence across samples was set to 350 for tables 0-22, 28, 30 and 32 – a more lenient threshold because these tables had a large number of samples – and to 50 for table 23, to yield more initial edges). If counts were provided, they were converted into relative abundances by dividing each entry by the total read count of its corresponding sample. For table set 3, a minimum sample sum of 800 was imposed to avoid zero count samples (except for tables 23 and 34-42, which were much less sparse). If lineages were available, higher-level taxa were assigned up to phylum level by summing relative abundances of lower-level member taxa. Parent-child relationships between taxa were prevented from occurring in all subsequent computations.

RMT(Deng et al 2012):

Initially proposed by Wigner and Dyson(Mehta 1990, Wigner 1967) for studying the spectrum of complex nuclei, RMT is a powerful approach for identifying and modeling phase transitions associated with disorder and noise in statistical physics and materials science. It has been successfully used for studying the behavior of different complex systems, such as the spectra of large atoms(Mehta 1990, Wigner 1967), metal insulator transitions in disordered systems(Altshuler and Shklovskii 1986, Hofstetter and Schreiber 1993), spectra of quasi-periodic systems(Zhong and Geisel 1999), chaotic systems(Bohigas et al 1984), brain response(Seba 2003), and the stock market(Plerou et al 1999). It was first adopted for delineating gene expression networks(Luo et al 2007, Zhou et al 2013).

All OTUs occurring in fewer than half of the samples were discarded except in table set 3 where minimum occurrence across samples was 50 in the 2000 total samples. Since RMT requires more than 80 OTUs to remain after removing the above OTUs, a few of the tested tables were not analyzed.

LSA(Ruan et al 2006, Xia et al 2013):

The theoretical p-values approximate the statistical significance of local similarity analysis based on the tail distribution of the excursion range of a random walk(Xia et al 2013). The approximation works reasonably well (starting at time points $t>10$ with no delay) and provides p-values comparable to those from permutations. One significant advantage of theoretical p-values is that it enables constant time calculation of statistical significance for pairwise local similarity analysis, making possible all-to-all comparisons for high-throughput data which are otherwise prohibitive.

Model Details

Copula

The particular copula method used was the Gaussian copula, which is founded on the fact that applying the normal cumulative distribution function (CDF) to a standard normal random variable results in a uniform random variable between 0 and 1. Inverse transform sampling then enables the creation of any distribution by applying that distribution's inverse CDF to a uniform random variable between 0 and 1 (Supplementary Fig. 16)

(Trivedi and Zimmer 2007). The copula function controls the joint distribution of the random variables and their rank correlations. Real covariance matrices are symmetric and positive definite; therefore the Cholesky decomposition was used to test for positive definiteness and so ensure meaningful OTU generation.

Table Set Construction Details:

Methods for table set 1

Tables 1 and 2 were created with the copula method with margins from the lognormal (u = 3, s.d. = 0) and gamma (shape parameter = 1, location = 0, lambda = 100) distributions, and with rho matrix entries ranging from [-0.01, 0.02]. Table 4 was created with the null model and no compositionality. It was created by random calls to the lognormal, gamma, nakagami, uniform, and chi-squared functions – again, distributions that could mimic bacterial growth and real OTU table sparsity, although the overall sparsity was still lower than in reality. Table 5 was created with OTUs from a Dirichlet distribution where the prior counts were given by random variables with a lognormal distribution. Tables 6 and 7 were ecological tables, having competitive, mutual, commensal, amensal, parasitic, obligate, and partial oligate syntrophic relationships of various strengths (2, 3, and 5) as well as two-species (OTU1 acts on OTU2) and three-species (OTU1 and OTU2 together act on OTU3) interactions.

Methods for table set 2

Tables 1 - 5 are time-series tables with changing frequency, amplitude, phase, noise, and subsampling routine. Table 1 is OTUs with sine wave variations, while table 2 is OTUs with a square wave for half the samples and a cosine wave for the other half of the samples. Table 3 is a half-sampling of the table 2 OTUs, table 4 is OTUs composed of sawtooth/cosine summations, and table 5 is OTUs made of a significantly undersampled sawtooth wave added to a low-frequency wave. Tables 6-10 are two-species Lotka-Volterra relationships, and tables 11-15 are six-species Lotka-Volterra relationships(Idema 2005, McMurdie and Holmes 2014). All Lotka-Volterra relationships are n species abundances described by n systems of differential equations mimicking interesting ecological relationships, such as predator-prey (Supplementary Fig. 15). The Lotka-Volterra relationships in tables 6-15 were padded and confounded with random OTUs from lognormal and gamma distributions. For the values generated with the Lotka-Volterra equations and confounding OTUs, tables 6 and 11 were made into relative abundance tables with points taken at equal intervals, while tables 7 and 12 were the same as 6 and 11 except the values were counts instead. Tables 8 and 13 were relative abundance tables with points taken at random indices, and tables 9 and 14 were the same as tables 8 and 13 except the values were counts instead. Tables 10 and 15 were generated from the same system of differential equations as tables 6-9 and 11-14 respectively, except 60% of the values were randomly set to zero. Tables 16-18 were again ecological tables, except with one-dimensional linear relationships only. The values were relative abundance, 50% sparsity, and relative abundance of the 50% sparsity table relatively. Tables 19-21 were copula tables drawn from lognormal (u = 3, s.d. = 0), gamma (shape parameter = 1, location = 0, lambda = 100), and exponential (u = 0, lambda = 1000) distributions using the same generating rho matrix as tables 1 and 2 from table set 1.

Methods for table set 3

Sequences for the study "Cultured gut bacterial consortia from twins discordant for obesity modulate adiposity and metabolic phenotypes in gnotobiotic mice" by Ridaura et al.(Ridaura et al 2013) were retrieved from the QIIME database(Caporaso et al 2010) and picked with default closed reference settings (QIIME 1.7-dev, GreenGenes(Paulson et al 2013) reference database v. 13_5) at 97 percent similarity. Briefly, 10 independent rarefactions were conducted at 1000 seqs/sample and 10 at 2000 seqs/sample using the QIIME script 'multiple_rarefactions_even_depth.py'(Caporaso et al 2010). These formed tables 0-9 and 10-19 (respectively) of table set 3. Tables 20-23 were created by taking table 0 (described above) and filtering out OTUs that did not occur in some percentage of samples (table 20, 21, 22, 23; 5, 10, 20, 50%). Tables 24-26 were created by filtering the unprocessed OTU table (described above) to eliminate OTUs whose overall sequence count was below a percentage threshold (a suggested step in Bokulich et al(Berry and Widder 2014)) and then rarefying at 1000 seqs/sample (table 24, 25, 26; 0.00005, 0.00010, 0.000025%). Table 27 was created by taking table 24 and performing the additional step of removing OTUs found in less than 20 percent of the samples. Table 28 was created by summarizing OTUs from the raw unprocessed table at L6 (genus level) using the QIIME script 'summarize_taxa.py'(Caporaso et al 2010). The table was then rarefied to 1000 seqs/sample, and OTUs not found in at least 20% of samples were removed. Table 29 was created by picking from the Ridaura et. al.(Ridaura et al 2013) sequences (described above) using the same parameters except that the similarity threshold for OTU clustering was reduced to 94% (the genus level). The resulting table then underwent the same processing steps as table 28. Tables 30 and 31 were the same as 28 and 29 except the summary was conducted at L5 (family level), and the similarity threshold was reduced to 91 percent (respectively). Tables 32 and 33 were again the same, but with summary at L4 (order level) and similarity threshold of 88% (order level).

Tables 34-43 were created with the generator methods described in the main text. All of these tables have periodic signals that are composed of sine, cosine, and square waves (superimposed, in some cases) as well a logistic growth curve and a Gaussian pulse and envelope. There are 6 parameters that are varied in these tables (other than the signal function): frequency, amplitude, phase, noise, sampling routine and sparsity. The sampling routine is either to evenly space the points in time, to randomly draw an ordered subset, or to draw an evenly spaced subset and then randomly select a fraction of those samples to be zeroed (abundance = 0). For table 34, frequency is varied from 0.25 to 200 (arbitrary units), phase is varied between 0 and pi/2, and the subsampling routine is varied between even, random, and even with zeroing, while other parameters are held constant. The OTUs may consist of sin, square, sin for half into square for half, and logistic growth signals. Table 35 is the same as table 34 in all respects except the pseudo-random number generator is set to a different seed and the percentage of subsampling is doubled (50 samples instead of 26). Table 36 is again the same but with subsampling again increased from 50 to 74 samples. Table 37 is a half-sampling (evenly) of table 34, table 38 is a half-sampling of table 35, and table 39 is a half-sampling of table 35. Tables 40-42 have OTUs that are constructed as Gaussian pulses and their envelopes. The frequency of the pulse is varied (table 40, 41, 42; 1, 10, .1hz).

Methods for table set 4:
An OTU table was generated with the copula model and lognormal distribution, with the rank correlation matrix specified as having all OTU correlations close to zero. Then six positively correlated OTUs were added, having rank correlations greater than 0.2. Six negatively correlated OTUs were added as well, with rank correlation less than -0.2. The effective number of species (n_{eff}), calculated with the inverse Simpson alpha diversity measure, in this table (table 0) was 36. Four more tables (tables 1-4) were created by replicating table 0 but multiplying one OTU by a constant factor such that the n_{eff} of the resulting tables was 25, 19, 10, and 4, respectively. Tables 0-4 were taken to be the absolute abundances, reflective of the microbial correlations in the natural environment. Compositionality was then induced, reflecting the sampling/sequencing process, by rarefying tables 0-4 at a depth of 2000 sequences per sample to create tables 5-9. To test the effect of rarefying at a lower depth, tables at 1000 sequences per sample were created. To test the effect of alternate normalization techniques, we also created CSS-normalized(Anders and Huber 2010) (tables 10-14) and DESeq-normalized(Lovell D 2010, Pearson 1897) (setting the negatives to zero as in McMurdie and Holmes(McMurdie and Holmes 2014), tables 15-19) versions of tables 5-9.

Altshuler BL, Shklovskii BI (1986). Repulsion of Energy-Levels and the Conductance of Small Metallic Samples. *Zh Eksp Teor Fiz+* **91:** 220-234.

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

Berry D, Widder S (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology* **5:** 219.

Bohigas O, Giannoni MJ, Schmit C (1984). Spectral Properties of the Laplacian and Random Matrix Theories. *J Phys Lett-Paris* **45:** 1015-1022.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7:** 335-336.

Deng Y, Jiang YH, Yang Y, He Z, Luo F, Zhou J (2012). Molecular ecological network analyses. *BMC bioinformatics* **13:** 113.

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J *et al* (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8:** e1002606.

Hofstetter E, Schreiber M (1993). Statistical properties of the eigenvalue spectrum of the three-dimensional Anderson Hamiltonian. *Physical review B, Condensed matter* **48:** 16979-16985.

Idema T (2005). The behaviour and attractiveness of the Lotka-Volterra equations. Doctorate thesis, Leiden University.

Lovell D MW, Taylor J, Zwart A, Helliwell C (2010). Caution! compositions! can constraints on omics data lead analyses astray? *CSIRO***:** 1-44.

Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK *et al* (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC bioinformatics* **8**.

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

Mehta ML (1990). *Random Matrices, 2nd edition.* Academic Press.

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

Pearson K (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* **60:** 489-502.

Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Stanley HE (1999). Universal and nonuniversal properties of cross correlations in financial time series. *Physical review letters* **83:** 1471-1474.

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.

Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F (2006). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22:** 2532-2538.

Seba P (2003). Random matrix analysis of human EEG data. *Physical review letters* **91:** 198104.

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

Wigner EP (1967). Random Matrices in Physics. *Siam Rev* **9:** 1-&.

Xia LC, Ai D, Cram J, Fuhrman JA, Sun F (2013). Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* **29:** 230-237.

Zhong JX, Geisel T (1999). Level fluctuations in quantum systems with multifractal eigenstates. *Phys Rev E* **59:** 4071-4074.

Zhou AF, Baidoo E, He ZL, Mukhopadhyay A, Baumohl JK, Benke P *et al* (2013). Characterization of NaCl tolerance in Desulfovibrio vulgaris Hildenborough through experimental evolution. *Isme Journal* **7:** 1790-1802.