SCNIC: Sparse Correlation Network Investigation for Compositional Data

Michael Shaffer[1,2,†], Kumar Thurimella[1,3,†] and Catherine A. Lozupone[1,*]

1 Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA 80045

2 Current: Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO 80523, USA

3 Current: Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK

†*These authors contributed equally to this work*

**To whom correspondence should be addressed*

ABSTRACT

Background

Microbiome studies are often limited by a lack of statistical power due to small sample sizes and a large number of features. This problem is exacerbated in correlative studies of multi-omic datasets. Statistical power can be increased by finding and summarizing modules of correlated observations. Additionally, modules provide biological insight as groups of microbes can have relationships among themselves.

Results

To address these challenges we developed SCNIC: Sparse Cooccurrence Network Investigation for Compositional data. SCNIC is open-source software that can generate correlation networks and detect and summarize modules of highly correlated features. We applied SCNIC to a published dataset comparing microbiome composition in men who have sex with men (MSM) who were at a high risk of contracting HIV to non-MSM. By applying SCNIC we achieved increased statistical power and identified microbes that not only differed with MSM-status, but also correlated strongly with each other, suggesting shared environmental drivers or cooperative relationships among them.

Conclusions

SCNIC provides an easy way to generate correlation networks, identify modules of correlated features and summarize them for downstream statistical analysis. Although SCNIC was designed considering properties of microbiome data, such as compositionality, it can be applied to a variety of data types including metabolomics data and used to integrate multiple data types. Using SCNIC allows for the identification of functional microbial relationships at scale while increasing statistical power.

BACKGROUND

Microbial communities play important roles in environmental and human health systems and can often reach great complexity. In these rich ecosystems, microbes interact with each other, forming relationships based on predator prey dynamics [1], competition for resources [2],cross-feeding of small compounds [3] and other factors. Identifying correlated pairs of microbes can suggest potential interactions or shared environmental preferences. Accordingly, studies have identified complex networks of co-occurring microbes in a variety of different environments ranging from the human mouth and gut [4] to soil [5] and stream ecosystems [6].

To detect correlations between microbes a variety of methods have been developed. While traditional correlation metrics are used by some [7–9], newer methods have been developed that take into account the properties of 16S rRNA sequencing data [10–12]. A recent review tested these methods on a variety of models and identified some methods that performed better than others in ways that can depend on underlying data characteristics [13]. Although these tools are useful for finding pairwise relationships between organisms, less attention has been given toward developing methods for finding correlations among groups of microbes.

One way to explore complex interactions is to form networks in which correlated organisms are joined with an edge, and highly correlated sets of microbes are defined, which we will refer to here as modules. There are two primary benefits of finding groups of correlated microbes. First, the combination of microbes in a module could be further explored to understand microbial interactions, such as cross-feeding relationships, or shared environmental niches [5,14–16]. Second, considering correlation structure among microbes can aid in statistical analysis aimed at uncovering relationships between microbes and other environmental factors. Specifically, by eliminating or summarizing highly correlated features, dependence between

features is decreased which will increase accuracy of methods that assume the independence of features such as false discovery rate technique (FDR) measurement [17], and statistical power is increased by reducing the number of feature comparisons.

One workflow for considering groups of correlated microbes in downstream statistical analyses requires three steps: First, correlations between microbes must be measured and used to form a network. Second, modules must be identified. Third, abundance of the microbes in modules must be summarized for use in subsequent statistical analyses. One software tool that has implemented this workflow, developed for application to gene expression data, is weighted gene correlation network analysis (WGCNA) [18]. WGCNA builds correlation networks based on a correlation coefficient (such as Pearson, Spearman, or biweight midcorrelation [19]), and detects modules as subtrees in a hierarchical cluster of features [20]. Modules are summarized by setting module abundance to that of network hubs or an eigenvector of the abundance of all module members [18].

Several groups have used WGCNA to analyze 16S rRNA sequencing data [21–24], but this approach may not be appropriate for several reasons [25]. First, the correlation metrics implemented in WGCNA do not account for compositionality. Only relative abundance of taxa is collected instead of true taxa abundance and this can lead to the detection of spurious correlations [26]. Second, the primary method WGCNA uses to pick modules assumes the correlation network will have a scale-free topology that may not be relevant to microbiome data [27]. Third, summarizing modules through identifying hub taxa works well in gene expression where a single transcription factor can control the expression of many genes, but may not be appropriate in microbial communities. Both the hub and eigenvector approaches to module summarization do not allow for output tables that maintain the total counts of microbial

abundance per sample and so cannot be used with tools developed for microbiome data analysis that make assumptions based on total sample counts such as ANCOM [28] or metagenomeSeq [29].

Optimal methods for identifying and summarizing modules of correlated features in 16S rRNA sequencing data have not been deeply explored. One study [25] recommended an ensemble approach for correlation detection [13], and the Louvain modularity maximization (LMM) method [30] to identify modules [31]. LULU is a tool that follows a binning approach towards OTUs that co-occur, but only does so if they're highly phylogenetically related [32]. Another tool, CoNet, uses an ensemble approach to build and visualize networks [33]. However, no implementation of module summarization was made available for downstream statistical analysis.

To address these concerns, we have developed a tool for sparse, compositional correlation network investigation for compositional data (SCNIC), which uses methods optimized for microbiome data analysis. SCNIC is available as standalone Python software, via Bioconda [34] and the Python package index (pip), and as a QIIME 2 plugin [35]. The source code for SCNIC and the QIIME 2 plugin is freely available on GitHub (https://github.com/lozuponelab/SCNIC, https://github.com/lozuponelab/q2-SCNIC) under the BSD-3-Clause License.

MATERIALS AND METHODS

The SCNIC method

SCNIC takes a feature table containing counts of each feature in all samples as input and performs three steps: 1) a correlation network is built, 2) modules are detected in the network

and 3) feature counts within a module are summed into a new single feature (identified as "module-*x*" where x is numbered consecutively starting at zero) (Figure 1). To summarize modules, SCNIC uses a sum of count data from all features in a module. The newly generated modules are included in a new feature table alongside all features not grouped into a module. Maintaining the total counts per sample allows for downstream analyses with tools that have assumptions related to total sample counts. SCNIC produces a graph modeling language (GML) format [36] file compatible with Cytoscape [37] for network visualization in which the edges in the correlation network represent the positive correlations which are stronger than a user specified R-value cutoff (between 0 and 1), a file describing which features compose each defined module, and a feature table in the Biological Observation Matrix (BIOM) format (McDonald et al., 2012) (Figure 1).

SCNIC allows users to choose between multiple methods for detecting correlations and of defining modules of co-occurring microbes. For correlations, in addition to implementing traditional correlation metrics (including Pearson's r, Spearman's ρ and Kendall's τ), SCNIC also invokes SparCC [39,40] to correct for compositionality in microbiome data. SparCC has been shown to perform well in detecting correlations compared to other correlation measures [13], in communities with an inverse Simpson index above 13 [39,40] as so was chosen as the default metric. To define modules of co-correlated features, we implement two methods: 1) Louvain modularity maximization (LMM) and 2) a novel shared minimum distance (SMD) module detection algorithm; unlike WGCNA, neither of these algorithms make assumptions about network topology. LMM was previously proposed as a method for clustering correlation networks of microbes into modules [30]. LMM works by first assigning one module per feature. Each pair of adjacent modules are joined and the change in modularity (defined by the number of

edges within the module compared to outside) is calculated for each module. The pair which increases the mean modularity of the network the most is then joined. This process is repeated until the modularity of the network is not increased. LMM uses two parameters provided by the use: The first parameter, R-value, defines the minimum correlation coefficient for defining an edge between features. The second parameter, gamma, controls the size of modules detected, with large gamma values yielding larger modules. WGCNA and LMM have a potential weakness in that modules can contain pairs of taxa that are not strongly correlated (e.g. if they are several steps away from each other in the network). To address this weakness we also implement the SMD method to ensure that correlations between all pairs of features in the module have an R-value greater than the user provided minimum. Specifically, the SMD method defines modules by first applying complete linkage hierarchical clustering to correlation coefficients to make a tree of features. Next, SMD defines modules as subtrees where correlations between all pairs of tips have an R-value above the specified value. SMD has been set as the default method in SCNIC because of the desirable property of only producing modules where all features are correlated over a user-specified threshold.

A large proportion of microbiome studies sample highly uneven communities which leads to strong compositionality-driven artifacts [26,41,42]. Because of this, we use SparCC, specifically the implementation of FastSpar [40], as the default correlation measure. SparCC was used as the correlation metric based on analysis that suggested a high precision in the number of correct edges recovered when correlations calculated in synthetic data [13]. SCNIC additionally includes the option of using Pearson's r, Spearman's ρ and Kendall's $\boldsymbol{\tau}$ to evaluate non-compositional or dense data types.

Demonstrating the use of SCNIC

We demonstrate the use of SCNIC with a study that used 16S rRNA sequencing of fecal material to compare microbiome composition in individuals with and without HIV and in men who have sex with men (MSM) who were at a high risk of contracting HIV [43]. The Noguera-Julian et al. data set was retrieved from NCBI SRA accession number SRP068240, and samples from the BCN0 cohort were used for these analyses. Reads were error corrected, quality trimmed, and primers were removed using default parameters in BBTools [44]. DADA2 was used to find amplicon sequence variants (ASVs) with reads trimmed from the left by 30 base pairs and truncated at 269. ASVs were binned into operational taxonomic units (OTUs) using USEARCH [45] at 99% identity via QIIME 1 [46]. A phylogenetic tree was made using a single representative sequence from each OTU and the SEPP protocol [47,48] via QIIME 2 [35] for the 99% OTUs. We evaluated the average phylogenetic distance between OTUs in the same module using the *distance* method of Biopython [49,50]. Taxonomy was assigned using the Naive Bayes QIIME 2 feature classifier, version gg-13-8-99-515-806-nb-classifier.qza.

The original study describing these data showed a strong divergence in gut microbiome composition in MSM compared to non-MSM independent of HIV status and more subtle differences associated with HIV when controlling for MSM behavior. The goal of our analysis was to evaluate whether comparing gut microbiome composition between HIV negative MSM and non-MSM with SCNIC modules provide additional significant taxa compared to without, and additional insights as to which taxa that differ with MSM also are in turn demonstrating co-correlated structure with each other, indicating that they may be a part of a broader community type, interact with each other, or have shared environmental drivers of their prevalence. A further goal of this analysis is to examine the effects of using different R-value

thresholds on the results. The SMD method was specifically used with SparCC R-value thresholds between 0.20 and 1.0, with 0.05 increments.

Evaluating effects of applying SCNIC to discern microbes that differ with MSM status

OTUs/modules that differed with MSM were identified using ANCOM [28] for each feature. We chose ANCOM because it is also a tool designed specifically for working with compositional microbiome data. ANCOM was applied to the original feature table where SCNIC was not applied, as well as to feature tables output from SCNIC using SparCC at different R value thresholds with the SMD algorithm. To focus on evaluating differences in the microbiome between MSM and non-MSM without confounding by HIV infection status, we only used samples from HIV negative individuals.

While applying SparCC, SCNIC uses the recommended practice of the SparCC manuscript of filtering based on average relative abundance across samples [39]. The SparCC manuscript suggests this filter because removing features with high abundances, even in a few samples, will upset the ability of the method to control for the number of reads per sample in its compositionality adjustment. Because this method can retain OTUs that are highly abundant in only a single sample, we removed features that had 0 values in more than 20% (~ 29/146) of samples before applying ANCOM but after applying SparCC. Significant differences with MSM status were determined as those above the W value threshold determined by ANCOM.

RESULTS

R-value thresholds influence module size and phylogenetic relatedness of OTUs binned into a module

A key parameter to set when running SCNIC is the R value threshold to use when picking modules. Use of a high R-value threshold would be expected to bin only very tightly correlated microbes with strong relationships while less stringent thresholds may identify community-level patterns representing more loosely connected microbial pairs. To illustrate this concept, we binned OTUs into modules using the SMD method at R-value thresholds between 0.2 and 1.0. As expected, at lower R-value thresholds, more OTUs were binned into modules and lower numbers of modules of smaller average size were formed as the threshold increased (Figure 2). To illustrate the effects of R values thresholds on the nature of the identified modules, we compare SCNIC outputs using R-value thresholds of 0.2, 0.4, and 0.65 (Figure 3). As shown in Figure 3 which visualizes modules in Cytoscape using SCNIC output files, the R value threshold influences the size and connectivity of the network. We also illustrate the effect of using different thresholds by examining the correlations between OTUs that are included in the first module output by SCNIC, module-*0* (Figure 4). SCNIC orders its modules by size, with the first modules being the largest and the last modules being the smallest. All of the OTUs in module-*0* are positively correlated with each other, since SCNIC only captures positive correlations.

Microbes co-occurring in the same environmental niche have previously been observed to be phylogenetically closer on average [4]. This is likely because phylogenetic relatedness has been correlated with functional relatedness, such as through having more shared genome content, leading towards success in similar environments [51]. We show that increasing the R-value

threshold results in modules that contain OTUs that are more phylogenetically similar on average (Figure 4).

Use of SCNIC results in the detection of novel MSM associated taxa

We next evaluated the effects of applying SCNIC with default SparCC and SMD parameters and varying R value thresholds on downstream statistical analysis results. To investigate differential abundance based on MSM status in the Noguera-Julian et al. dataset, we use ANCOM [28]. We found that 12 OTUs were significantly different between MSM and non-MSM without using SCNIC using ANCOM (Table 1). Using SCNIC at R-values of 0.2, 0.5, and 0.65 and running ANCOM on the filtered output OTU table, we found that most significant features were modules (Table 1), which is interesting because the vast majority of OTUs were not a part of modules (Figure 2). The majority of 12 of the OTUs that were significant without running SCNIC, were grouped into modules with each other and with OTUs that were not individually significant without running SCNIC (Table 1). These significant modules contained 74, 26, and 1 new OTU at R-values of 0.2, 0.4 and 0.65 respectively. Using SCNIC also resulted in the identification of 1, 5 and 25 (at R-values of 0.2, 0.4 and 0.65) OTUs that were individually significant that were not significant without running SCNIC, indicating an increase in statistical power resulting from running a test like ANCOM that controls the FDR.

Considering correlation structure of significant features can help in understanding the broader community context of bacteria that differ with MSM status. In module-*0* for each of the R-values *Prevotella*, which significantly differed by MSM status in all cases, was the dominant genus. At an R-value of 0.65, all of the OTUs in the module were assigned to the genus *Prevotella*. However, at an R-value of 0.4 the module included seven *Prevotella* OTUs, one

Dialister, and an unidentified member of the *Bacteroidetes* phylum. At the R-value of 0.2, *Prevotella* accounted for 13 of the 25 OTUs and 11 of the 12 Pre-SCNIC significant OTUs were all found in this module. This suggests that individual OTUs that differ with MSM status may in some cases be a part of a consortium of diverse members that collectively display features that may contribute to differences in microbiome function.

To further explore this concept, we investigated the results generated with an R-value of 0.4, as the significant features maintain a strong level of correlation while being phylogenetically diverse. When running ANCOM on this feature table, we found that these individually significant OTUs tended to be joined into modules with other highly co-correlated microbes and that these modules significantly differed with MSM (Figure 5). Of particular note, we observe that the modules and taxa that are significantly related to MSM do not all correlate with each other. At the R-value of 0.4, module-*36* contains two taxa, *Erysipelotrichaceae* and *Clostridium* that are negatively correlated with the other significant taxa and modules (Figure 5). Module-*2* contains *Eubacterium, Catenibacterium* and *Prevotella* which are phylogenetically heterogenous but mutually co-occurring. A follow up experiment, which leverages insights that SCNIC generates, may combine different strains of microbes to assemble a community type to test for functional correlates of disease.

DISCUSSION

SCNIC provides a method to measure correlations, find and visualize modules of correlated features, and summarize modules by summing their counts for use in downstream statistical analysis. Using SCNIC with the SMD algorithm for module detection aids in dimensionality reduction in 16S rRNA sequencing data while ensuring a minimum strength of

association within modules. As expected, our workflow identified modules in which OTUs tended to be phylogenetically related, especially at relatively high values of R. Using SCNIC, we are able to detect previously insignificant features by grouping them into modules which are significant. In this analysis, we used OTUs as features however, other microbiome features can be used with SCNIC, such as ASVs, genera or species defined with a taxonomic classifier, as well as other data types such as metabolome data.

SCNIC complements existing methods because these either: 1) form correlation networks of microbes for visualization but do not have functionality for selecting and summarizing modules for downstread statistical analysis [33], 2) can select and summarize modules for downstream statistical analysis but are designed for gene expression and not microbiome data [18], only summarize features if they are phylogenetically related [32], or suggest methods for finding modules of correlated microbes but do not provide a convenient implementation [30]. SCNIC is available both as a stand-alone application and as a QIIME2 plugin for easy integration with existing microbiome workflows.

We illustrate here that varying the R-value threshold input by the user has a great impact on the results. However, we have avoided giving specific R-value threshold recommendations here, because optimal R-values may vary across datasets and data types. Using higher R-values thresholds was more likely to identify highly phylogenetically related microbes that likely share overlapping functionality, and in principle could also identify diverse organisms with overlapping niches or highly complementary metabolic functions. Using a lower R-value threshold bins a broader community of more loosely correlated features with the risk of bringing together features which should not be grouped. By summarizing correlated features, SCNIC mitigates overcorrection in multiple test adjustments by reducing the number of taxa and false

discovery rate for downstream analysis. When these organisms are grouped into a broader module that is truly independent from other modules, any penalties on two highly similar features may be avoided in statistical analysis.

The results of our analysis of the Noguera-Julian et al (2016) data set yielded findings that not only confirm what was found in their original analysis as well as another study [52], but included many new significantly associated taxa. At differing R-values of 0.2, 0.4 and 0.65 there were 74, 26 and 1 new OTU that were included in significant modules that were not individually significant. Additionally, at R-values of 0.2, 0.4 and 0.65 there were 1, 5 and 25 OTUs that became individually significant (Table 1). This primary result describes the many strong microbial associations with MSM status. The associations in the Noguera-Julian et al. study are done at the genus level which obscures some of the complexity in the data.

SCNIC assists in interpretation of microbiome data by finding new significant features and investigating correlations among these features. At an R-value of 0.2, 13 were of the *Prevotella* in significant modules and 1 was individually significant, while 3 OTUs of the *Bacteroides* genus were in significant modules. At 0.4, 12 were of the *Prevotella* genus in significant modules and 2 were individually significant while 1, respectively, were of the *Bacteroides* genus in a significant module. At the R-value of 0.65, 1 was of *Prevotella* genus through significant modules and 10 were individually significant while 1 was of the *Bacteroides* genus that was individually significant. Several previous HIV microbiome studies all found these genera most associated strongly with MSM status [43,52–54]. In module-*0*, which was more abundant in MSM samples, *Prevotella* species are correlated with two OTUs identified as *Eubacterium biforme* (which has recently been renamed *Holdemanella biformis* [55]). *Prevotella copri* has previously been associated with increased inflammation [53] while *in vitro*

stimulations of human immune cells have found that *P. copri* did not induce particularly high levels of inflammation but *E. biforme* did [54]. This strong correlation between *P. copri* and *E. biforme* in MSM could explain the increased inflammation seen in individuals with higher levels of *P. copri,* with *E. biforme* being the true driver. Indeed MSM status has previously been associated with increased inflammation [56,57]. With the use of SCNIC, this correlation highlighted a route of mechanistic understanding which could be functionally followed up on in further experimental studies.

SCNIC detected multiple significant modules, of which none of the OTUs within were significant when analyzed separately. Module-*20*, which was associated with MSM status, is the fourth most significant feature at R-value of 0.2, and is made up of *Acidaminococcus*, *Megasphaera*, and *Mitsuokella* species. These are all from the Veillonellaceae family which is likely the explanation to their correlation. Members of the Veillonellaceae family have been linked with inflammation [58].

By increasing statistical power and providing context for the relationships between significant taxa, SCNIC modules open new opportunities for analysis. When a module is associated with a variable of interest, the correlations within the module may imply functional relationships. These can be further investigated with *in vitro* and *in vivo* experiments. Studies which aim to test hypotheses generated commonly will use culture or gnotobiotic mouse studies to test the effects of single significantly associated microbes on a condition. These studies do not adequately represent *in vivo* systems because microbes in isolation often do not affect a disease state or their environment. SCNIC can enhance these confirmatory studies by identifying groups of microbes that may grow better than individual microbes and may better elicit relevant phenotypes than when grown separately.

ACKNOWLEDGEMENTS

We would like to thank Elmar Pruesse for input on the design of SCNIC. We thank Abigail Armstrong and Casey Martin for beta testing SCNIC. Funding for KT came from the University of Colorado School of Medicine Research Track. Funding for MS came from NIH NLM 4 T15 LM009451-10.

REFERENCES

1. Corno, G., Villiger, J., and Pernthaler, J. (2013) Coaggregation in a microbial predator–prey system affects competition and trophic transfer efficiency. *Ecology*, **94** (4), 870–881.
2. Burkepile, D.E., Parker, J.D., Woodson, C.B., Mills, H.J., Kubanek, J., Sobecky, P.A., and Hay, M.E. (2006) Chemically mediated competition between microbes and animals: microbes as consumers in food webs. *Ecology*, **87** (11), 2821–2831.
3. LaSarre, B., McCully, A.L., Lennon, J.T., and McKinlay, J.B. (2017) Microbial mutualism dynamics governed by dose-dependent toxicity of cross-fed nutrients. *ISME J.*, **11** (2), 337–348.
4. Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput. Biol.*, **8** (7), e1002606.
5. Barberán, A., Bates, S.T., Casamayor, E.O., and Fierer, N. (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.*, **6** (2), 343–351.
6. Widder, S., Besemer, K., Singer, G.A., Ceola, S., Bertuzzo, E., Quince, C., Sloan, W.T., Rinaldo, A., and Battin, T.J. (2014) Fluvial network organization imprints on microbial co-occurrence networks. *Proc. Natl. Acad. Sci.*, **111** (35), 12799–12804.
7. Bray, J.R., and Curtis, J.T. (1957) An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.*, **27** (4), 325–349.
8. Pearson, K. (1909) Determination of the coefficient of correlation. *Science*, **30** (757), 23–25.
9. Spearman, C. (1904) Measurement of association, Part II. Correction of 'systematic deviations.' *Am J Psychol*, **15**, 88–101.
10. Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methé, B., DeSantis, T.Z., Human Microbiome Consortium, Petrosino, J.F., Knight, R., and Birren, B.W. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.*, **21** (3), 494–504.
11. Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, **12** (7), 1889–1898.
12. Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods*, **7** (10), 813–819.
13. Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., and Knight, R. (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J*, **10** (7), 1669–1681.
14. Ban, Y., An, L., and Jiang, H. (2015) Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, **31** (20), 3322–3329.
15. Dugas, L.R., Bernabé, B.P., Priyadarshini, M., Fei, N., Park, S.J., Brown, L., Plange-Rhule, J., Nelson, D., Toh, E.C., Gao, X., Dong, Q., Sun, J., Kliethermes, S., Gottel, N., Luke, A., Gilbert, J.A., and Layden, B.T. (2018) Decreased microbial co-occurrence network stability and SCFA receptor level correlates with obesity in African-origin women. *Sci Rep*, **8** (1), 17135.
16. Lozupone, C., Faust, K., Raes, J., Faith, J.J., Frank, D.N., Zaneveld, J., Gordon, J.I., and

Knight, R. (2012) Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Res.*, **22** (10), 1974–1984.
17. Benjamini, Y., and Hochberg, Y. (2000) On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. *J Educ Behav Stat*, **25** (1), 60–83.
18. Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
19. Wilcox, R.R. (2011) *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press.
20. Barabási, A.-L., and Albert, R. (1999) Emergence of Scaling in Random Networks. *Science*, **286** (5439), 509–512.
21. Castillo, J.D., Vivanco, J.M., and Manter, D.K. (2017) Bacterial Microbiome and Nematode Occurrence in Different Potato Agricultural Soils. *Microb Ecol*, **74** (4), 888–900.
22. Tong, M., McHardy, I., Ruegger, P., Goudarzi, M., Kashyap, P.C., Haritunians, T., Li, X., Graeber, T.G., Schwager, E., Huttenhower, C., Fornace, A.J., Sonnenburg, J.L., McGovern, D.P.B., Borneman, J., and Braun, J. (2014) Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *ISME J.*, **8** (11), 2193–2206.
23. Yin, J., Han, H., Li, Y., Liu, Z., Zhao, Y., Fang, R., Huang, X., Zheng, J., Ren, W., Wu, F., Liu, G., Wu, X., Wang, K., Sun, L., Li, C., Li, T., and Yin, Y. (2017) Lysine Restriction Affects Feed Intake and Amino Acid Metabolism via Gut Microbiome in Piglets. *Cell Physiol Biochem*, **44** (5), 1749–1761.
24. Younge, N., Yang, Q., and Seed, P.C. (2017) Enteral High Fat-Polyunsaturated Fatty Acid Blend Alters the Pathogen Composition of the Intestinal Microbiome in Premature Infants with an Enterostomy. *J Pediatr*, **181**, 93-101.e6.
25. Jackson, M.A., Bonder, M.J., Kuncheva, Z., Zierer, J., Fu, J., Kurilshikov, A., Wijmenga, C., Zhernakova, A., Bell, J.T., Spector, T.D., and Steves, C.J. (2018) Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ*, **6**, e4303.
26. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.*, **8**.
27. Broido, A.D., and Clauset, A. (2019) Scale-free networks are rare. *Nat Commun*, **10** (1), 1017.
28. Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., and Peddada, S.D. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis*, **26**, 27663.
29. Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat Methods*, **10** (12), 1200–1202.
30. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008** (10), P10008.
31. Baldassano, S.N., and Bassett, D.S. (2016) Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease. *Sci Rep*, **6**, 26087.
32. Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A.K., Pietroni, C., and Hansen, A.J. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.*, **8** (1), 1188.
33. Faust, K., and Raes, J. (2016) CoNet app: inference of biological association networks using

Cytoscape. *F1000Research*, **5**, 1519.
34. Grüning, B., The Bioconda Team, Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., and Köster, J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15** (7), 475–476.
35. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., 2nd, Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., and Caporaso, J.G. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*, **37** (8), 852–857.
36. Himsolt, M. (1997) GML: A portable graph file format.
37. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13** (11), 2498–2504.
38. McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., and Caporaso, J.G. (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience*, **1** (1), 7.
39. Friedman, J., and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, **8** (9), e1002687.
40. Watts, S.C., Ritchie, S.C., Inouye, M., and Holt, K.E. (2019) FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*, **35** (6), 1064–1066.
41. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2** (1), 15.
42. Tsilimigras, M.C.B., and Fodor, A.A. (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.*, **26** (5), 330–335.
43. Noguera-Julian, M., Rocafort, M., Guillén, Y., Rivera, J., Casadellà, M., Nowak, P., Hildebrand, F., Zeller, G., Parera, M., Bellido, R., Rodríguez, C., Carrillo, J., Mothe, B., Coll, J., Bravo, I., Estany, C., Herrero, C., Saz, J., Sirera, G., Torrela, A., Navarro, J., Crespo, M., Brander, C., Negredo, E., Blanco, J., Guarner, F., Calle, M.L., Bork, P., Sönnerborg, A., Clotet, B., and Paredes, R. (2016) Gut Microbiota Linked to Sexual

Preference and HIV Infection. *EBioMedicine*, **5**, 135–146.
44. Bushnell, B., Rood, J., and Singer, E. (2017) BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE*, **12** (10), e0185056.
45. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26** (19), 2460–2461.
46. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, **7** (5), 335–336.
47. Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J.A., Jiang, L., Xu, Z.Z., Winker, K., Kado, D.M., Orwoll, E., Manary, M., Mirarab, S., and Knight, R. (2018) Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems*, **3** (3).
48. Mirarab, S., Nguyen, N., and Warnow, T. (2012) SEPP: SATé-enabled phylogenetic placement. *Pac Symp Biocomput*, 247–258.
49. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25** (11), 1422–1423.
50. Talevich, E., Invergo, B.M., Cock, P.J., and Chapman, B.A. (2012) Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, **13** (1), 209.
51. Zaneveld, J.R., Lozupone, C., Gordon, J.I., and Knight, R. (2010) Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.*, **38** (12), 3869–3879.
52. Armstrong, A.J.S., Shaffer, M., Nusbacher, N.M., Griesmer, C., Fiorillo, S., Schneider, J.M., Preston Neff, C., Li, S.X., Fontenot, A.P., Campbell, T., Palmer, B.E., and Lozupone, C.A. (2018) An exploration of Prevotella-rich microbiomes in HIV and men who have sex with men. *Microbiome*, **6** (1), 198.
53. Dillon, S.M., Lee, E.J., Kotter, C.V., Austin, G.L., Dong, Z., Hecht, D.K., Gianella, S., Siewe, B., Smith, D.M., Landay, A.L., Robertson, C.E., Frank, D.N., and Wilson, C.C. (2014) An altered intestinal mucosal microbiome in HIV-1 infection is associated with mucosal and systemic immune activation and endotoxemia. *Mucosal Immunol.*, **7** (4), 983–994.
54. Lozupone, C.A., Li, M., Campbell, T.B., Flores, S.C., Linderman, D., Gebert, M.J., Knight, R., Fontenot, A.P., and Palmer, B.E. (2013) Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host Microbe*, **14** (3), 329–339.
55. De Maesschalck, C., Van Immerseel, F., Eeckhaut, V., De Baere, S., Cnockaert, M., Croubels, S., Haesebrouck, F., Ducatelle, R., and Vandamme, P. (2014) Faecalicoccus acidiformans gen. nov., sp. nov., isolated from the chicken caecum, and reclassification of Streptococcus pleomorphus (Barnes et al. 1977), Eubacterium biforme (Eggerth 1935) and Eubacterium cylindroides (Cato et al. 1974) as Faecalicoccus pleomorphus comb. nov., Holdemanella biformis gen. nov., comb. nov. and Faecalitalea cylindroides gen. nov., comb. nov., respectively, within the family Erysipelotrichaceae. *Int. J. Syst. Evol. Microbiol.*, **64**

(Pt_11), 3877–3884.
56. Gianella, S., Strain, M.C., Rought, S.E., Vargas, M.V., Little, S.J., Richman, D.D., Spina, C.A., and Smith, D.M. (2012) Associations between virologic and immunologic dynamics in blood and in the male genital tract. *J Virol*, **86** (3), 1307–1315.
57. Palmer, C.D., Tomassilli, J., Sirignano, M., Romero-Tejeda, M., Arnold, K.B., Che, D., Lauffenburger, D.A., Jost, S., Allen, T., Mayer, K.H., and Altfeld, M. (2014) Enhanced immune activation linked to endotoxemia in HIV-1 seronegative MSM. *AIDS*, **28** (14), 2162–2166.
58. Bajaj, J.S., Ridlon, J.M., Hylemon, P.B., Thacker, L.R., Heuman, D.M., Smith, S., Sikaroodi, M., and Gillevet, P.M. (2012) Linkage of gut microbiome with cognition in hepatic encephalopathy. *Am. J. Physiol.-Gastrointest. Liver Physiol.*, **302** (1), G168–G175.

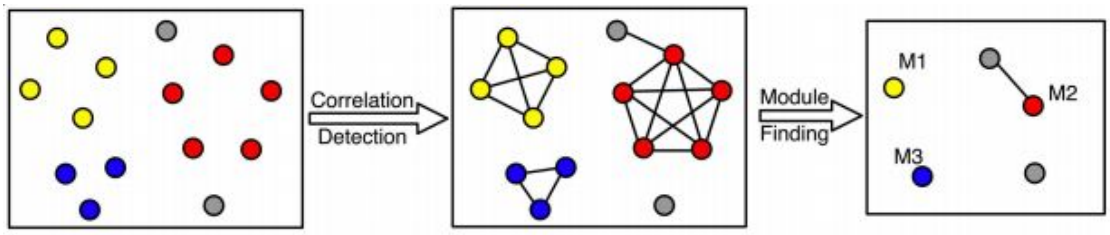
DATA ACCESSIBILITY

The Noguera-Julian et al. data set is available from NCBI SRA accession number SRP068240.

AUTHOR CONTRIBUTIONS

M.S. coded the initial implementation of SCNIC and made major contributions to its conceptualization and design. K.T. improved the SCNIC implementation and performed the case study. C.L. conceptualized SCNIC and guided its implementation and design. K.T., M.S. and C.L. all wrote the manuscript together.

A)



B)

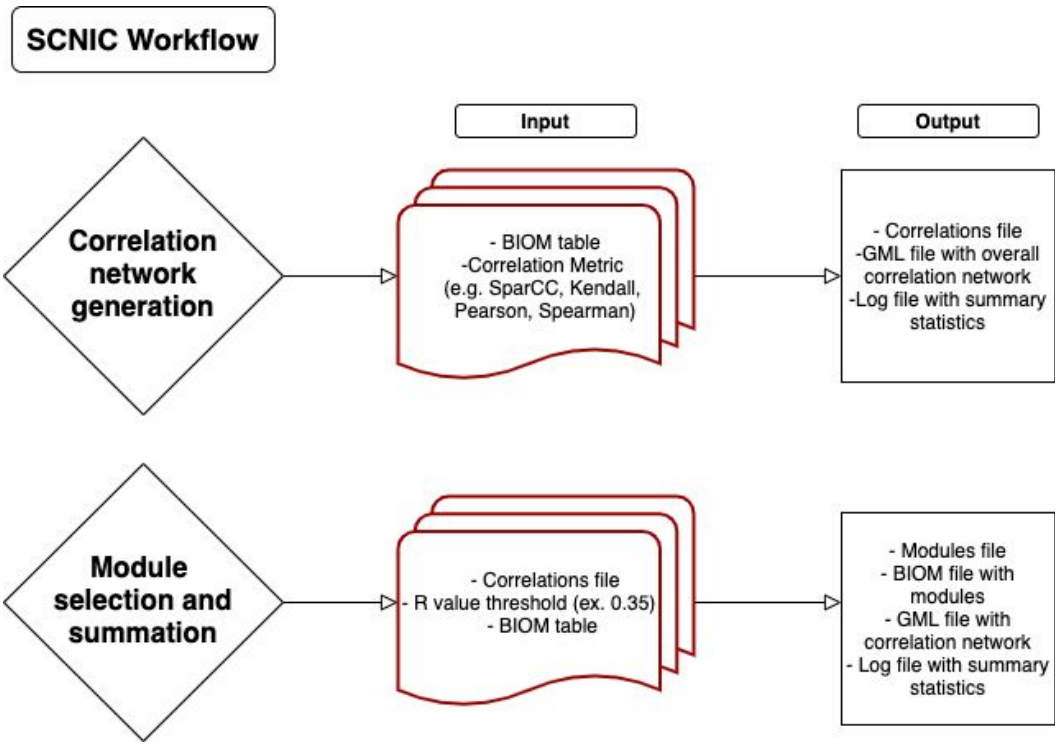


Figure 1 SCNIC Schematic and Data Flow

A. The basic process of SCNIC involves first identifying pairwise correlations between species and using them to build a correlation network. Modules of correlated features are identified and then summarized so that there is a decrease in the total number of features. Resulting modules are then summarized for downstream statistical analysis, or multi-omic analysis between modules of microbes and other feature types. B. The input to SCNIC comes in the form of a count table in BIOM format. The first step takes the table and generates a correlation table and network. The table is in a tab separated format and the network is in GML format and can be used to visualize the network in Cytoscape. Modules are detected and summarized in the final step which generates a module membership file indicating which features are in each module. The collapsed BIOM table contains the same total counts per sample as the original table, but with less features. All features not included in modules are retained with their original counts and all modules have a total count per sample of the sum of all counts of all features in that module.

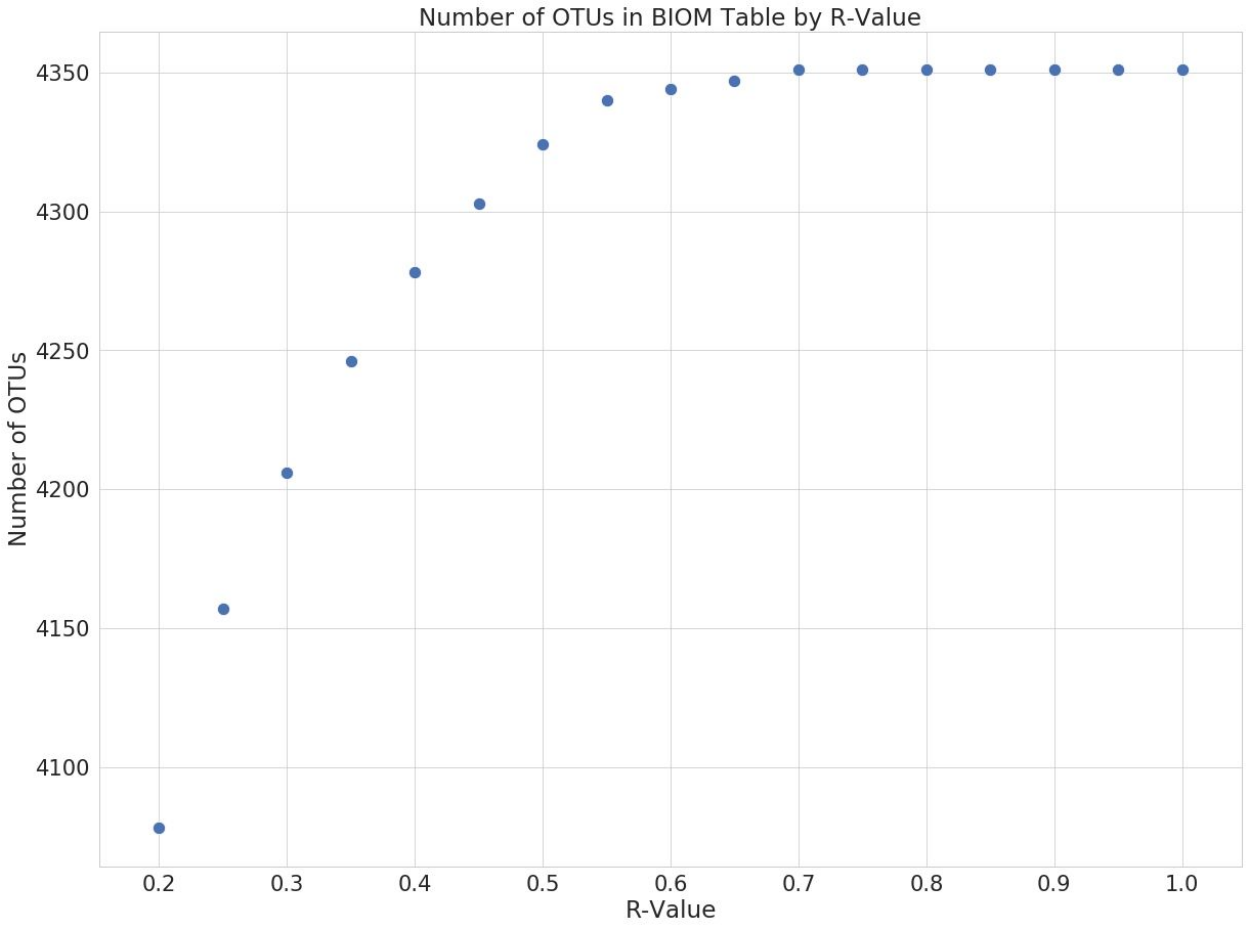


Figure 2 SCNIC Feature Reduction

We used SCNIC to select modules using the same OTU table, the SMD module selection algorithm, and SparCC R-values ranging from 0.2 to 1.0, in increments of 0.05. The R-Value is plotted against the number of features in the resulting BIOM table produced by SCNIC. As the R-value increases the number of modules decreases and the number of single features (modules + OTUs not included in modules) increases. After the R-value of 0.65, the number of features in the resulting file remained the same at 4351 features which was the same size as the input OTU table, because there were no modules that were created past a SparCC R of 0.65.

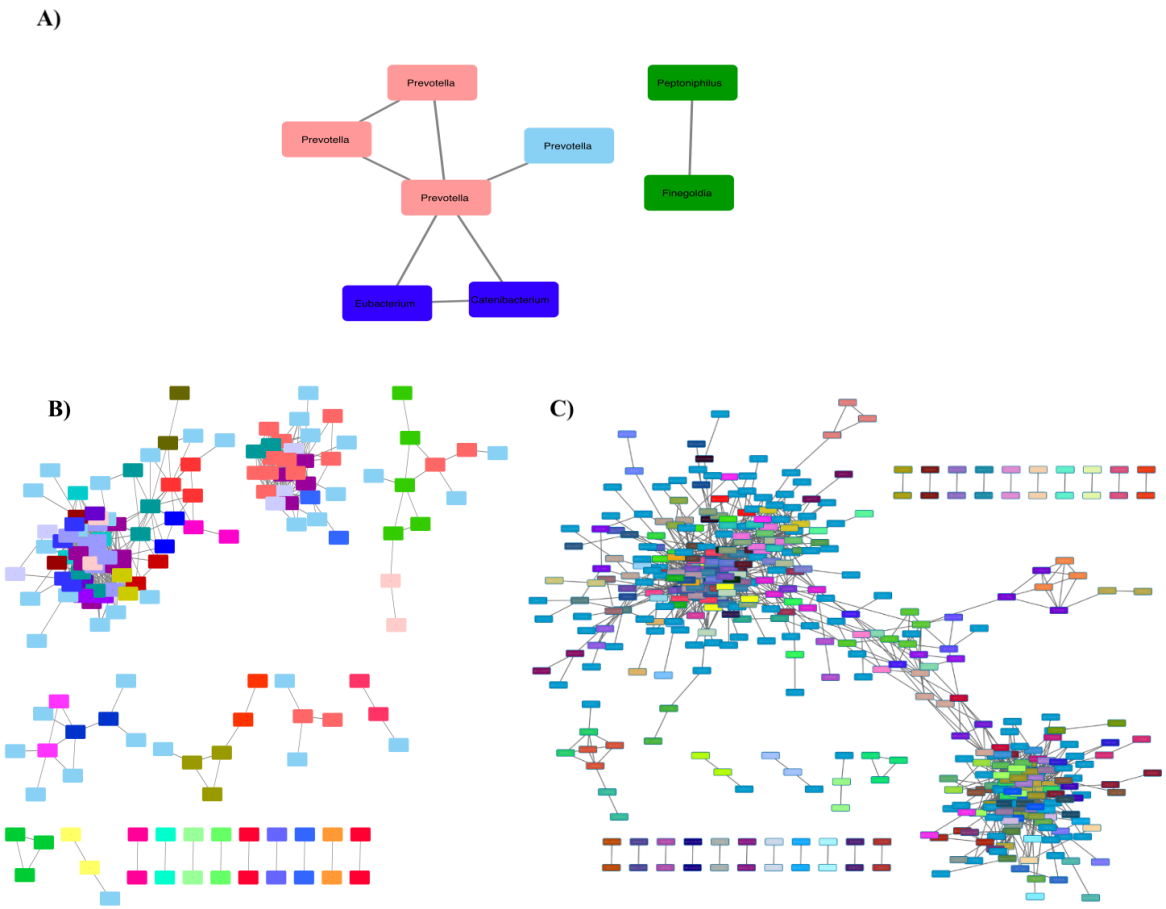


Figure 3 Visualization of SCNIC Networks

The Cytoscape output allows for easy exploration and visualization of relationships between different OTUS/taxa in an interactive interface. A) R = 0.65 B) R = 0.4 C) R = 0.2. As the R-value increases, the size of the network decreases as SCNIC does not include singletons (features with no significant positive correlations) in the resulting network file. Correlation network where edges are correlations with a R value greater than the threshold set. Nodes are OTUs and node color represents module membership (i.e. module-*0* is pink in Panel A).

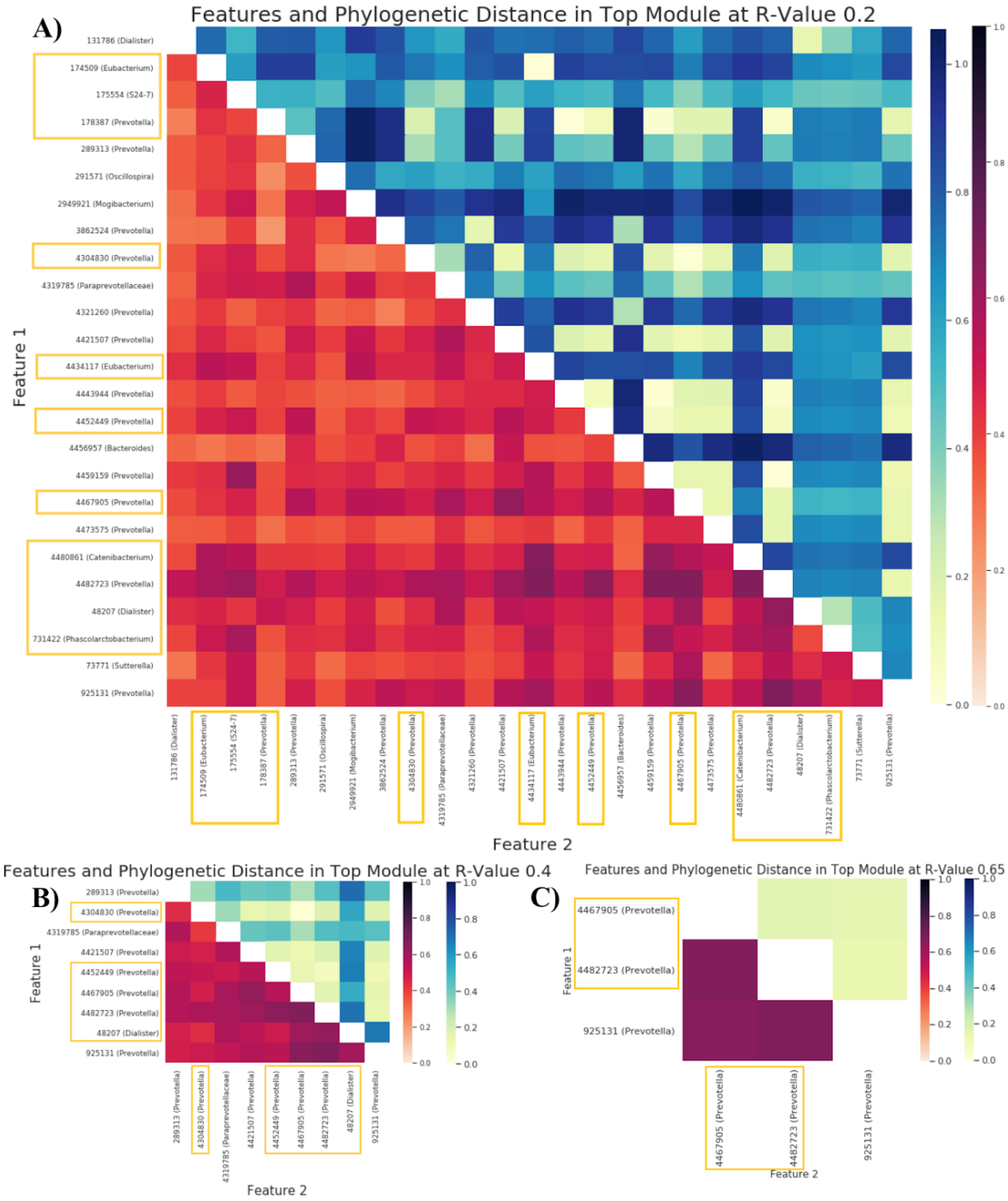


Figure 4 Module-*0* Across Different R-Values

Module-*0* expanded to view taxonomy and correlations amongst them at R-values of 0.2 (A), 0.4 (B), and 0.65 (C). As the R-value increases, the species in module-*0* become more phylogenetically similar. Module-*0* has 11, 5 and 2 of the significant Pre-SCNIC OTUs at R-values of 0.2, 0.4 and 0.65, and are highlighted in a yellow border.

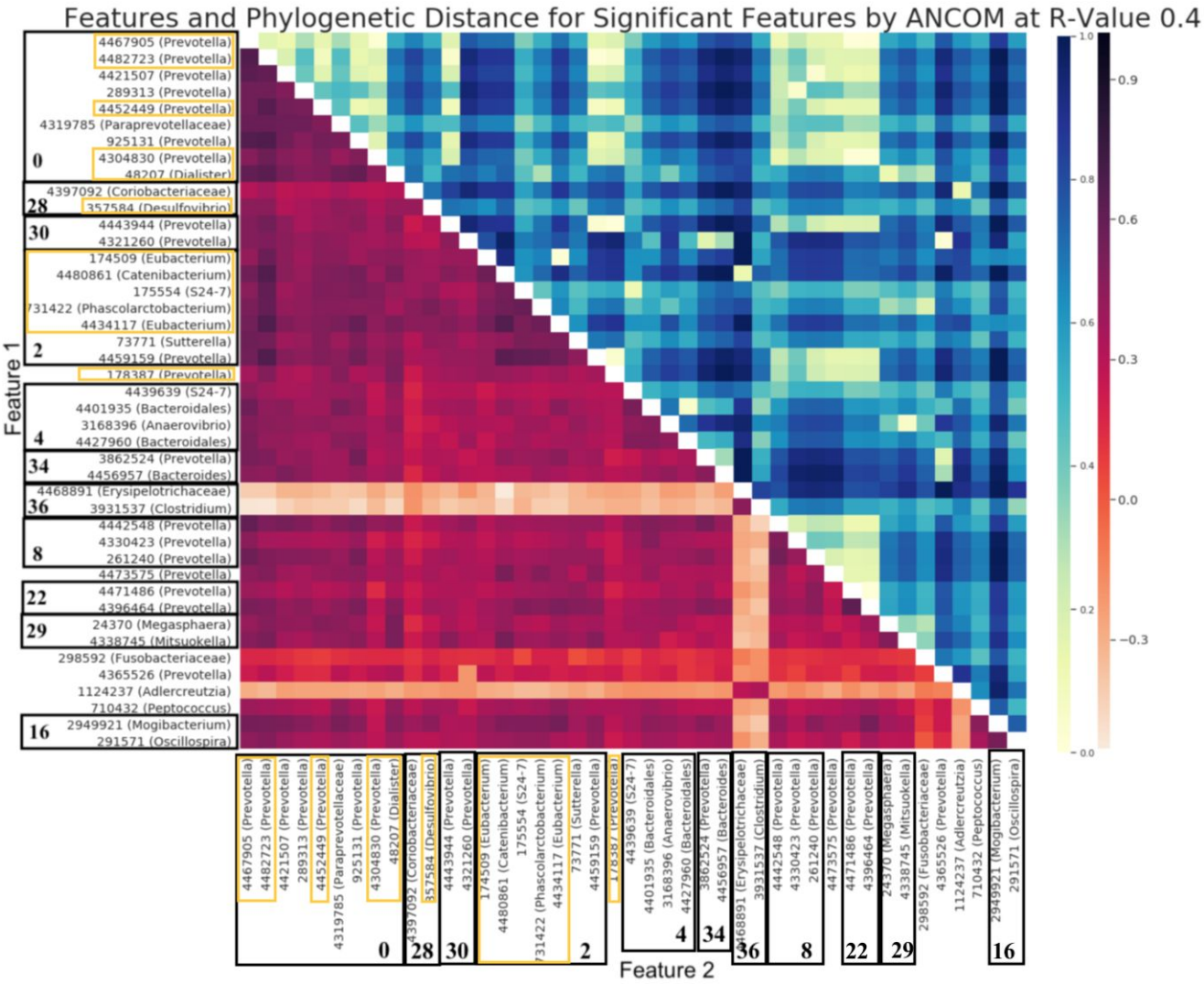


Figure 5 All Significant Features at R-value 0.4 Found by ANCOM

Each of the borders in the y-axis represents the different modules, with the module number bolded. The Pre-SCNIC OTUs that were significant are highlighted in a yellow border. The heatmap in the lower triangle corresponds to the correlation found by SparCC. The heatmap in the upper triangle represents the phylogenetic distance between organism pairs. The negative correlations in the lower triangle correspond to OTUs in relation to one another. However by design, no module contains any taxa that are negatively correlated with each other.

R-Value	New OTUs in Significant Modules	New Significant OTUs	Lost Significant OTUs
0.2	74	1	0
0.4	26	5	0
0.65	1	25	0

R-Value	Number of Significant Modules	Total Significant Features
0.2	14	15
0.4	11	17
0.65	2	35

Table 1 Significant SCNIC Modules and Features Across R-Values with

The ANCOM analysis of SCNIC with MSM as the categorical variable for differential abundance. Before running SCNIC, there were 12 OTUs found to be significant. Each R-value we tested yielded new OTUs in modules that were found significant, with the largest number of OTUs at R-value of 0.2. As the R-value increases, the ratio of number of significant modules to the number of significant features decreases.