Methods to delineate membership in a 'core' community are inconsistent, rarely test the hypothesis of a 'core', and can mislead ecological analysis

Maya Gans[1], Gordon Custer[1], Linda van Diepen[1], and C. Alex Buerkle[1]

[1]University of Wyoming

June 8, 2021

Abstract

Community ecology includes linking variation in system functions to the distribution and abundance of taxa. In inferring processes, functions, and causal taxa, it is common practice to assume a core community can be defined and that attributes of the core are representative of the entire dataset. Assuming categorical thresholds in abundance exist has the potential to be misleading, especially if rare taxa are contributing to ecological processes. Additionally, there are no standard criteria for core membership, complicating comparisons across studies. Rather, the existence of a core set of taxa can be treated as a hypothesis that may or may not be supported. We considered four methods commonly used for defining a core in studies of microbiomes and applied them to two published microbial data sets and simulations covering a range of plausible communities. We evaluated the ability of each method to correctly categorize taxa. Assignment of core taxa varied substantially among methods and datasets. Additionally, the ability of evaluated methods to capture the simulated core was contingent on the distribution of taxon abundances. While able to correctly identify core taxa in select cases, the methods disagreed more often than not. Given the lack of agreement among core assignment methods, categorization of taxa into sets corresponding to core and non-core is questionable and requires testing and validation before use in any particular context. Our results do not support applying methods of dimension reduction for core taxa classification, but instead provide additional rationale to favor analyses that use abundance data in their entirety.

Introduction

Commonly occurring taxa within a particular habitat are thought to be critical to that habitat's and ecosystem's functions (Hamady & Knight, 2009; Shade & Handelsman, 2012; Turnbaugh & Gordon, 2009; Turnbaugh et al., 2009; Umaña, Zhang, Cao, Lin, & Swenson, 2017). A core set of taxa has been defined as the consistent assemblage of organisms associated with a certain niche space (Hamady & Knight, 2009), and cataloging core taxa has been the focus of many recent microbiome studies, due to the complex nature of high throughput sequence data. Separating taxa into a set of core and non-core or transient members serves the purpose of simplifying multidimensional data and is thought to be advantageous for performing statistical analyses. Support for this simplification stems from empirical consistency between patterns observed with only the core taxa and all taxa present (Delgado-Baquerizo et al., 2018), the idea that commonly occurring core taxa are responsible for community function (Saunders, Albertsen, Vollertsen, & Nielsen, 2016), and from the conservative practice of statistical testing for treatment effects by examining only the most commonly occurring taxa (Wirth et al., 2018). Examining the dynamics and patterns of variation of the core assemblage is often seen as an important step in analyzing and understanding complex community interactions.

The concept of a core community has been operationalized in various sets of criteria that can be applied to identify taxa that could belong to the core assemblage (Delgado-Baquerizo et al., 2018; Gray, Amjad, &

Gray, 1983; Lundberg et al., 2012; Shade & Handelsman, 2012; Shade & Stopnisek, 2019; Soliveres et al., 2016; Turnbaugh & Gordon, 2009; Turnbaugh et al., 2009). However, the assumption that a core set of taxa can be accurately identified underlies all core methods, and it is unclear to what extent the concept of a core assemblage is supported by data. Shade and Handelsman (2012) reviewed different criteria for defining the core microbiome including abundance, phylogeny, and function. However, they did not evaluate evidence or support for the concept of a core community. More recently, studies have indicated that some habitats are not occupied by a consistent, core, set of taxa and instead host transients (Hamady & Knight, 2009; Hammer, Janzen, Hallwachs, Jaffe, & Fierer, 2017).

Beyond methodological considerations, focusing on a core subset of taxa might overlook consequential effects of rare taxa. With attention shifting from "who is there?" (i.e. taxonomic composition) to "what are they doing?" (i.e. functionality), the contribution of rare taxa, especially those that serve as hub taxa in complex microbial networks, should not be disregarded simply due to lower abundances (Banerjee, Schlaeppi, & van der Heijden, 2018; Shi et al., 2020). Certain narrowly distributed microbial functions such as nitrification, denitrification, methanogenesis, or sulfate reduction are performed by relatively rare microbes (Jousset et al., 2017; Lynch & Neufeld, 2015). Use of community analyses that only examine abundant or commonly occurring microbes (i.e. core assignments), has the potential to overlook those taxa responsible for important ecosystem functions, like the ones listed above. In focusing solely on core taxa, the contributions of transient or rare taxa are discounted and attributed to commonly occurring ones, potentially overemphasizing the importance of common taxa while simultaneously underestimating the contribution of rare taxa.

Given the considerable and growing interest in using molecular data to characterize diverse communities across many samples and conditions (e.g. Ahrendt et al., 2018; Delgado-Baquerizo et al., 2018; Desnues et al., 2008; Geisen, Laros, Vizcaíno, Bonkowski, & de Groot, 2015; Porazinska et al., 2010; Stat et al., 2017; Tedersoo et al., 2014) and the frequent use of core community analyses, we evaluated the definition a core community and its consequences via multiple methods: First, we compared different methods for defining core membership. Next, we used the core assignments and the full datasets to determine whether the interpretation of differences in community diversity (beta-diversity) would be the same. And finally, we examined to what extent core assignment methods could identify significant hub taxa as determined by cooccurrence network analysis. Our study used microbial datasets from the human microbiome project (Turnbaugh et al., 2007) and soil rhizosphere samples from *Arabidopsis thaliana* (Lundberg et al., 2012) as well as simulations to examine the validity of splitting taxon count data into two sets (core and non-core), while also assessing the effects of varying criteria on core membership.

Materials and Methods

We conducted a non-exhaustive survey via Web of Science (April 2018) to understand what methodologies were being used for core membership assignment. We limited our survey to papers containing the terms 'core' and 'microbiome' within the title or abstract, resulting in 1034 papers. We selected papers from 2008-2018, then ordered the search results by the number of citations and considered the 200 most cited papers. Of the 200 papers, 45 publications sufficiently detailed their methods for assigning core membership. These remaining papers were then subdivided into five categories according to the varying methodologies and criteria that were used to identify core taxa (Table 1). While not an exhaustive review, the survey provided an adequate representation and summary of the core assignment methodologies used in contemporary analyses. The categories are as follows: cumulative proportion of sequence reads, proportion of replicates, proportion of sequence reads and replicates, hard cutoff, and the Venn diagram method. In practice, DNA sequence reads (hereafter referred to as "reads") within a given level of sequence similarity (e.g. 97%, 99%, or 100%) are counted as a measure of taxon abundance. The counts of all taxa found in the taxonomic table, as used here to describe the composition of microbiomes, are analogous to counts of plants, animals, or any other counts of organisms in community ecology. In microbial and other organismal ecology, the total sampling effort is finite and greater certainty of taxon relative abundance is achieved with increasing sampling effort (increasing read depth in microbiome research), with diminishing returns (Forcino, Leighton, Twerdy, & Cahill, 2015; Zaheer et al., 2018).

Here we use four out of the five methods described in our literature review on both simulated and real published datasets: cumulative proportion of sequence reads, proportion of replicates, proportion of sequence reads and replicates, and hard cutoffs (Table 1). We excluded the Venn Diagram method from our analysis because it can be considered as a more stringent version of the proportion of replicates method, in that a taxon must be present in every sample to be included in the core.

The first method, cumulative proportion of sequence reads, ranks taxa based on relative abundance and includes the most abundant taxa in the core. Briefly, this is calculated by ranking all taxa by their relative abundance, from highest to lowest. Starting with the most abundant, the percentage of total reads is summed cumulatively until 75% of the total reads are accounted for. Taxa that account for a portion of the first 75% of total reads are assigned core membership. This method was adopted from vegetation counts in classic ecology (Hanski, 1982) and accounts for the relative abundance of individual taxa and the total sampling effort. The second method, proportion of replicates, assigns a taxon to the core when that taxon is present in at least 50% of samples within a given treatment (or observational category; our study included a single treatment). This method assigns taxa to the core that are found in the majority of samples and accounts for sample size alone. The third method, proportion of sequence reads and replicates, assigns a taxon to the core if it is present in a pre-determined proportion of the total number of samples from a given treatment (or observational category) as well a predetermined proportion of the total reads (Delgado-Baquerizo et al., 2018). This third method accounts for both sample size and sampling effort simultaneously. Though various proportions of samples and reads were used in publications, for our analysis, we set these thresholds to include taxa present in at least 50% of samples within a given treatment (we only included a single treatment) and account for at least 0.02% of the total reads across all samples (adapted from Callahan et al. (2016)). For the fourth method, hard cutoffs, we implemented a cutoff similar to Lundberg et al. (2012), such that for a taxon to be considered part of the core microbiome, it must be present in at least some number of samples and with at least some number of reads. This method differs from the proportion of sequence reads and replicates method in that the hard-cutoff method *a priori* assigns an absolute value for the cutoff that is not based on the size of the experiment or sequencing depth achieved. In our analysis, we used the following cutoffs: the taxon must be present in five samples with at least 25 total reads across all samples (Lundberg et al., 2012).

To test each of the four core methods, we used two published and 6,250 simulated datasets. For the two published datasets (Table 2), we used 1) the rhizosphere and site M21 subset of the final rarified operational taxon table from the *Arabidopsis thaliana* root microbiome project ((Lundberg et al., 2012); dataset name "arabidopsis_R_M21.rda") and 2) the fecal sample subset from The Human Microbiome Project ((Consortium, 2012); dataset name "human_stool.rda"). The two published datasets were plotted by the log-transformed taxon mean abundance and the coefficient of variance (CV) to examine the grouping of taxa assigned to the core. This was done to assess whether a clear threshold in abundance and coefficient of variance existed between core and non-core taxa for any of the examined core assignment methods, as an obvious threshold may provide support for that core assignment methodology. Additionally, we used 250 simulations for each of 25 possible combinations of 1) five levels of magnitude of difference in abundance (π) of core versus non-core taxa (represented as the $\pi_{\text{core}}/\pi_{\text{non-core}}$, ranging from 1 to 25), and 2) five levels of variance of the abundances $\pi_{\text{core}}$to $\pi_{\text{non-core}}$ among replicates (quantified by an intensity parameter ϑ, ranging from 1 to 50). This resulted in a total of 6,250 unique simulations to assess each assignment method. Each simulation of taxon relative abundances involved random draws from a Dirichlet distribution parameterized by the expected frequencies of all taxa ($\Sigma\pi_{\text{i}} = 1$, with 25 taxa parameterized by π_{core} and 975 by $\pi_{\text{non-core}}$) and a single intensity parameter (ϑ) that affects the precision of taxon abundances (i.e. scales the variance around expected taxon abundance defined by π_{core} and $\pi_{\text{non-core}}$), using R v3.4.2 (R Development Core Team, 2020). Across sets of simulations, we varied the relative abundance of core and non-core taxa ($\pi_{\text{core}}/\pi_{\text{non-core}}$), with $\pi_{\text{core}}/\pi_{\text{non-core}} = 1$ corresponding to a community that lacks a true core, because all taxa have equal expected abundances. This ratio acts as a control, in that core taxa should not be recovered from this dataset. On the other end, a $\pi_{\text{core}/}\pi_{\text{non-core}} = 25$ simulated a dataset in which core taxa had an abundance 25 times greater than non-core taxa. Further, we utilized the intensity parameter (ϑ) to set the precision of taxa abundances

across replicates for a given set of expected frequencies (π), with ϑ of 50 corresponding to high precision and low variance in taxon relative abundances among replicates and a ϑ of 1 leading to low precision and a large variance in taxon relative abundances. All simulations with $\pi_{\text{core}}/\pi_{\text{non-core}} > 1$ were of 25 taxa as core taxa and the remainder 975 as non-core. The 25 core taxa were simulated to receive the expected abundances and precision of core taxa, while the non-core received the abundances and precision expected for non-core members. Consequently, up to 25 taxa could be detected as true core taxa (true positives), and 975 taxa as false core members (false positives) or true non-core members (true negatives). Additionally, simulations of $\pi_{\text{core}}/\pi_{\text{non-core}} = 1$ were useful in quantifying the false positive rate, as these communities did not include any core taxa. A simulation's random draw from the Dirichlet distribution yielded a vector of sample proportions for each of 1000 taxa ($P(p_1, p_2, \ldots, p_{1000}| \pi_{\text{core}},\pi_{\text{non-core}}, \vartheta)$, to which the four criteria for core membership were applied directly.

The ability of each method to accurately recover the known core was assessed using simulated taxon tables by the following metrics: true positive rate (signal), false positive rate (noise), and net assignment value (signal-noise). The true positive rate represents the proportion of known core taxa that were classified as such, regardless of the number of false negatives. This is expressed as the probability of a true core taxon being assigned as such. The false positive rate represents the proportion of non-core taxa classified as core and is represented as the probability of a non-core taxon being assigned as a member of the core. The net assignment value represents the differences between the absolute number of true positives and the number of false positives. A net assignment value of 25 represents perfect classification. A net assignment value of -975 would indicate all non-core taxa were ill-assigned to the core with no true positive classifications. The more negative the number, the more highly inflated the core is. This metric can be interpreted as the difference in signal and noise.

Next, to examine to what extent core assignments produce the same differences in community diversity (beta-diversity) as the entire dataset, we created dissimilarity matrices with two different distance metrics (Bray-Curtis and Jaccard) using each set of core assignments and the full dataset as independent inputs. We then used PermANOVA testing (adonis) in the vegan package (Oksanen et al., 2018) to examine the variance explained by categorical predictors in each of the core and full datasets. We examined three categorical predictors for the human microbiome project dataset: patient visit number (1^{st}, 2^{nd}, or 3^{rd}), sex of subject (male vs. female), and where the sequencing was performed (12 different sequencing centers). In the *Arabidopsis* dataset we were able to examine two categorical predictors: developmental stage (young vs. old) and genotype (nine different genotypes). Results from the core subsets and total dataset were compared and differing statistical significance for categorical variables was noted.

In addition to examining differences in beta-diversity between the full dataset and core subsets, we also used cooccurrence networks to identify significant taxa. Cooccurrence networks are used in microbial ecology to determine microbial taxa that occur together in a statistically significant manner, with nodes and edges representing significant microbes and the connections between them respectively. Network analyses can be used to mine for keystone or hub taxa and can be used to identify interactions between groups of microbial taxa (Banerjee et al., 2016, 2018; Shi et al., 2020). The *Arabidopsis* rhizosphere microbiome graph consisted of 14,890 agents (taxa) and 288 artifacts (samples), and the human microbiome project graph consisted of 11,752 agents (taxa) and 319 artifacts (samples). From each bipartite graph we obtained the weighted bipartite projection, then extracted its signed backbone using the backbone package (Domagalski, Neal, & Sagan, 2019). Edges were retained in the backbone if their weights were statistically significant (alpha = 1e-04, Bonferroni corrected) by comparison to a null Hypergeometric Model (Neal, 2013). The corresponding nodes IDs (taxa IDs) with significant edges were then extracted and compared to the core assignments core assignment methods. Assuming core taxa drive abundance and occurrence patterns, and that rare taxa do not largely contribute to community structure, one could expect to find core taxa serving as important nodes with many significant edges (higher degree centrality) and the opposite to be true for non-core assignments.

To facilitate the use of our analytical methods by researchers curious about the validity of core microbiome assignment, we wrote an R package, CoreMicro, that can be installed through github

(github.com/mayagans/coremicro). The package includes functions that accept a taxon table as an input and can be used to generate plots and tables of core inclusion by method. This functionalized approach facilitates the comparison of methods and provides a means to check for the existence of the hypothesized core:non-core divide. In addition, the package includes all data, including the full used within this study as well as the code for all simulations. Full OTU tables and metadata files of the*Arabidopsis* and Human Microbiome Project datasets can be found at 10.5281/zenodo.4909346.

Results

Our analysis of simulated datasets showed the rate of true positives (probability of a core taxon assigned as such or signal) is close to one in many cases and appears to provide support for the ability of those methods to correctly assign core taxa (Figure 1a, denoted in blue). Furthermore, the rate of false positives (probability of a non-core taxon assigned as core member or noise) is close to zero in many cases seemingly providing additional support for core assignment methods (Figure 2b, denoted in blue). However, when examined individually these two metrics only tell half the story, as we are concerned with the ability of a given method to accurately identify the core taxa (i.e. true positives), while not over inflating membership through inclusion of non-core taxa (i.e. false positives); thus, being able to discern signal from noise.

The net assignment scores for simulations revealed the inability of the methods to accurately assign core membership (Figure 1c). The net assignment value quantifies the absolute difference in true positives (signal) and false positives (noise), with a net assignment value of 25 meaning the method assigned all of the correct taxa to the core with no erroneous assignments and smaller values indicating poorer performance in accurate core assignment. Our results show that rarely did the methods accurately recover the correct number and identity of core taxa (those simulated to be included in the core). In general, a large difference in the abundance of core and non-core taxa ($\pi_{\text{core}}/\pi_{\text{non-core}}$, with varying degrees of precision), led to the greatest success in accurate identification of the correct 25 core taxa (Figure 1c, right side x-axis, success denoted by dark blue squares, white and red indicate poor performance). When comparing results of the four core assignment criteria, the proportion of sequence replicates and proportion of sequence reads and replicates methods most often accurately assigned the 25 core taxa, with multiple instances of a net assignment value >24 (Figure 1c). The two methods tha0t utilized the proportion of replicates produced similar results in our simulations. They were followed by the hard cutoff method and then the cumulative proportion of sequence reads method (Figure 1c). All methods, with the exception of the cumulative proportion of sequence reads, were able to accurately recover the known core in some circumstances (net assignment value >24). However, they did so for different ranges of parameter combinations, suggesting each method may better suited to different taxon distributions.

Even though core methods accurately assigned core membership in some circumstances, the same methods produced negative net assignment values in other situations, consistent with overestimation of core membership. Core inclusion was most severely overestimated in the cumulative proportion of sequence reads and hard cutoffs methods in simulations with low π_{core} to $\pi_{\text{non-core}}$ ratio and high precision (parameterized by ϑ). This overestimation manifested as a high false positive rate (noise) in certain simulated communities. In general, the methods based on proportionality tended to assign the smallest set of core taxa and possessed the best net assignment value (i.e. correct assignment of known core taxa and limited erroneous assignment of non-core taxa to the core) and as such could be considered the most conservative.

For the two published datasets, the four core methods led to different conclusions, with the inferred core corresponding to 1.21%-15.74% of total taxa (Table 2). All methods assigned taxa with high abundance to the core, though methods differed in their assignments with respect to CV among replicates (Figure 2). More specifically, the cumulative proportion of sequence reads method and the proportion of sequence replicates method included highly abundant taxa regardless CV in both datasets. The method based on proportionality of replicates and sequence reads selected only abundant taxa with a relatively low CV in the human microbiome dataset (Figure 2a) and selected abundant taxa regardless of CV in the *Arabidopsis* dataset (Figure 2b). Finally, the core method that uses both the proportion of reads and replicates appear to arbitrarily exclude taxa with relatively high mean abundance and low CV, taxa that fit multiple criteria

for core membership. This is especially evident in the Human Microbiome Project dataset. These exclusions highlight problems associated with assigning continuously distributed count data into core and non-core groups.

Examination of core assignments in the published datasets showed that co-assignment (i.e. common core assignment by multiple methods) varied depending on the dataset (Figure 3). The Human Microbiome Project dataset yielded 176 core assignments that were assigned by all four methods (9.5% of total unique core assignments). The *Arabidopsis*dataset produced 165 core assignments that were shared among all four methods (8.1% of total unique core assignments). These common core assignments equivalate to 1.49% and 1.1% of the total number of taxa in each taxon table, respectively. For the *Arabidopsis* dataset, 758 taxa (37.2% of total unique core assignments) were assigned to the core by two methods and 322 taxa (15.8% of total unique core assignments) by three methods. As for the Human Microbiome Project dataset, 404 taxa (21.8% of total unique core assignments) were assigned to the core by three methods, and 530 taxa (28.6% of total unique core assignments) were assigned by two methods.

Comparisons of differences in beta-diversity between assigned cores and the full datasets, showed that in some cases the core datasets matched the entire dataset, but this was not always true. The entire*Arabidopsis* dataset showed both developmental stage and genotype to be significant in structuring the community (p=0.001); this was true for both the Bray-Curtis and Jaccard dissimilarity indices. The taxon table including only taxa assigned by all four methods matched these results (p=0.001) when using Bray-Curtis dissimilarity, but the Jaccard index only resulted in a significant effect of developmental stage (p=0.001) with genotype not significant predictor (p=0.132). Beta-diversity analysis of the core communities based on each of the four core-assignment methods separately mostly produced the same effects on beta-diversity as observed for the entire dataset, except the hard cutoff method. However, this method had comparable results to the taxon table created from taxa co-assigned by all four methods, with developmental stage being significant for both the Bray-Curtis and Jaccard dissimilarity indices (p=0.001), and genotype being significant for the Bray-Curtis index (p=0.001) but not Jaccard (p=0.148).

As for the Human Microbiome Project dataset, estimates of beta-diversity were affected by the use of core taxa or all taxa, raising concern for interpretation and the validity of core assignments. The full Human Microbiome Project stool dataset showed both sex and sequencing center to be significant ($p<0.01$), while visit number was shown to be statistically insignificant ($p>0.05$). These results were true for both Bray-Curtis and Jaccard dissimilarity indices. When examining only taxa assigned by all four core assignment methods, visit number, sex, and sequencing center were all significant ($p<0.05$) with Bray-Curtis dissimilarity, but only sequencing center was significant ($p<0.001$) with the Jaccard index. Results of beta-diversity analysis based on the core communities determined by each of the four core-assignment methods were similar to the results from the full dataset for both dissimilarity indices, except for the proportion of replicates reps and reads method. While the proportion of replicates and reads agreed with the others on the significance of sequencing center, this core assignment method showed visit number to be significant ($p<0.05$) for Bray-Curtis dissimilarity and insignificant for Jaccard dissimilarity ($p>0.05$). As for sex, Jaccard dissimilarity was insignificant ($p>0.05$) and Bray-Curtis dissimilarity was significant ($p< 0.01$).

Comparison of core assignments to taxa deemed important by their degree centrality revealed further disagreement. The *Arabidopsis* dataset produced 2,258 taxa that were deemed important, either by any of the four core assignment methods or by the cooccurrence network (Figure 4a,c). Of these 2,258 taxa, 1655 (73.3%) were uniquely assigned by the core methodologies, while 222 (9.8%) were assigned by the network alone. A small number of taxa, 381 (16.9%), was identified by both core assignment methods and the network analysis. The average degree centrality of taxa assigned as core by any method was 9.4, while the average degree centrality of non-core taxa was 0.25. This dataset produced large number of taxa deemed important by solely core assignment, with 1655 core taxa possessing zero significant edges in the network. The top 62 taxa, determined by degree centrality, were identified by the core assignment methods as well. The taxa with the highest degree centrality not picked up by core methods had a degree centrality of 118. On the other hand, the human microbiome project dataset produced very different results, with 3,181 taxa being

identified as important by either the any of the core assignment methods or the network (Figure 4b,d), and almost half, 1586 (49.86%), identified by both core assignment methods and the network analysis. Of these 3,181 taxa, only a small portion of 264 (8.3%) were uniquely assigned by the core methodologies, while 1331 (41.84%) were assigned by the network alone. The average degree centrality of taxa assigned as core by any method was 22.9, while the average degree centrality of non-core taxa was 2.2. In the human microbiome network, the top 61 taxa, in terms of degree centrality, were identified by both the cooccurrence network and core assignment methods. The taxa with the highest degree centrality not picked up by core methods had a degree centrality of 98.

Discussion

Though the concept of a core community is prevalent in ecology, specifically microbial ecology, our findings draw attention to the inconsistency of examined methods and the potential analyses based on a core set of taxa to be misleading. Application and comparison of four commonly used core assignment methods to both simulated and empirical datasets yielded conflicting results, with no clear threshold in abundance or commonness defining core membership inclusion. Our results show that in some situations core assignment methods agree, but in many others, the methods do not reach consensus. We also reveal that this variation in core assignments can produce different statistical results and lead to different ecological interpretations. Furthermore, core co-assignment (i.e. assignment by multiple methods) was limited by the most conservative assignment method, which varied between datasets. Rather than consistent assignments of taxa to a core, assignments were not robust and instead differed among methods and their criteria. This was also highlighted in our comparison of assigned core taxa to statistically significant nodes from our cooccurrence network analyses, with many taxa possessing high degree centrality being absent from core assignments, and in the *Arabidopsis* dataset, many taxa assigned as core possessing zero significant edges. Our finding highlights the statistical nature of core assignments as opposed to an underlying biological phenomenon and demonstrate the importance of the underlying data structure for assigning a core. Researchers could evaluate the statistical support for a distinction between core and non-core taxa and thereby justify focusing on a subset of taxa for convenience. However, use of a subset of the taxa will in all cases lead to some loss of information about the communities. Instead, researchers could use all of the available abundance data including rare taxa in statistical procedures and rely on model-based methods to recognize groups of taxa that differ in abundance (e.g. Harrison, Calder, Shastry, & Buerkle, 2020; Martin, Witten, & Willis, 2020). Beyond the statistical considerations, the contribution of these less common individuals is becoming increasingly recognized (Amor, Ratzke, & Gore, 2020; Jousset et al., 2017).

Community ecology has a long-standing interest in distinguishing between taxa that drive ecological functions and patterns, and taxa that are less obviously important in ecological systems. Dubos et al. (1965) ascribed importance to 'indigenous flora', defined as those microorganisms present during the development of an animal that are so ubiquitous they establish in all its members. These autochthonous microbiota are similar to the concept of a climax community within a habitat, where nonindigenous or transient taxa were compared to flowing streams just passing through, not contributing to the functionality of the system (Savage, 1977). Furthermore, it has been hypothesized that core taxa drive and are disproportionately responsible for the functions of ecological systems (Grime, 1998). These viewpoints of classic ecology have been invoked as rationale to consider only the 'indigenous flora' or core taxa, in an attempt to reduce the noise arising when considering mixtures of taxa with different levels of association with their host environments (Astudillo-García et al., 2017). Thus, it is common, current practice in studies of microbiomes to consider only taxa that are common enough to meet criteria for membership in the core community (e.g. Turnbaugh & Gordon, 2009; Turnbaugh et al., 2009; Wirth et al., 2018). However, our results indicate this practice may lack statistical support in some natural systems and may even go as far as to exclude taxa that are important for microbial network structure.

Given that analyses of microbial communities are frequently based on taxon counts for thousands of taxa in a single sample (high dimensionality), the appeal of focusing our attention on fewer dimensions and taxa is understandable. Dimension reduction can reduce noise or variation among samples and resolve the strongest

patterns in data sets (Nguyen & Holmes, 2019). Because many statistical methods lack power when applied to highly dimensional data (Nguyen & Holmes, 2019), scientists rarely analyze an entire ecological dataset and instead focus on a subset of the most common individuals (Hawinkel, Kerckhof, Bijnens, & Thas, 2019). From a conceptual standpoint, the practice of focusing on a core set of taxa, and discarding variation in the remaining taxa, comes at little cost if the ecological functions of interest are associated with variation in the core set of taxa. There are certainly examples (Winfree, Fox, Williams, Reilly, & Cariveau, 2015) and even theory (mass-ratio hypothesis; (Grime, 1998)) of ecological processes being tied to variation in the abundance of a small number of relatively common taxa. Yet, variation in ecological and especially microbially-driven processes is sometimes associated with rare taxa (e.g. sulfur reduction, nitrification, or methanogenesis) and thus demonstrates the risk of discarding uncommon members from the community of interest (Harrison et al., 2021; Jousset et al., 2017; Mikkelson, Bokman, & Sharp, 2016; Shade et al., 2014). Our results support this latter viewpoint and show that trends observed in the common taxa (i.e. core assignments) do not always accurately represent the entire taxon assemblage.

Similarly, in initial genomic studies of human trait variation that considered millions of variable nucleotides (analogous to the high number of taxa in microbiomes), researchers focused on those nucleotides that had a particular minor allele frequency (commonly at least 0.05). The conceptual rationale for the focus on a subset of genomic sites was the hypothesis that common conditions (e.g., disease) should be associated with common nucleotide variants (Lohmueller, Pearce, Pike, Lander, & Hirschhorn, 2003; Pritchard & Cox, 2002). The statistical rationale included the difficulty of estimating the effect of rarely observed variants, as is true for rare taxa in a microbiome. The hypothesis of *common disease-common variant* received poor support for some traits of interest (Cirulli & Goldstein, 2010). Instead there was a growing recognition of the potential contribution of rare alleles and the potential exchangeability of neighboring rare variants that, in aggregate, could explain trait variation (Zhou & Stephens, 2012). Likewise, variation in ecological processes could be associated with variation of the common taxa in communities, or through variation among any of their members. As for genomics, this is a hypothesis to be tested in ecology and for some systems discarding rare taxa will preclude understanding ecological processes.

If one accepts the necessity of considering only common and prevalent taxa, our comparison of core assignment methods should raise concern about their subjectivity and inconsistency. In our simulations, we considered core taxa to be those that were 2-25 times more common than non-core taxa from the same samples, representing a range of plausible taxon abundances. Analyses of simulated and two empirical datasets highlight the inconsistency among the four common methods considered for defining a core community and call into question the validity of dichotomizing taxon abundances into core and non-core assignments. The lack of a clear threshold in taxon abundance and coefficient of variation across real datasets suggest that this divide may not be supported in some cases. Furthermore, core taxa were more or less associated with study variables than the entire taxon assemblage and could lead to different ecological interpretation. The categorization of taxa into core and non-core groups is contingent on both the criteria used for identification and the underlying structure of the taxon table. Our results demonstrate the statistical nature of core assignment criteria and the forced dichotomization as opposed to an underlying biological difference between core and non-core taxa. Our comparisons along with other empirical studies (Caporaso et al., 2011; Clooney et al., 2016; Pollock, Glendinning, Wisedchanwet, & Watson, 2018; Shafer et al., 2017) indicate that the size and membership of the core community are not robust to differences in bioinformatic methods and core classification criteria.

The lack of consistency across core assignment methods should concern researchers as inferences drawn from core assignment can change drastically even when using the same dataset. Our simulations covered a range of plausible community structures, providing core assignment methods opportunity to potentially accurately assign core taxa. The individual samples in our simulated datasets contained nearly identical taxon distributions and offer core assignment methods a best case scenario. The lack of agreement and consistency in core assignment even under best case scenarios calls into question whether using these methods in ecological studies is fruitful or misleading. For example, researchers have related energy acquisition to variance in core gut microbiota, and the role this may play in obesity (Ley, 2010; Turnbaugh & Gordon, 2009; Turnbaugh et

al., 2009). The observed relationship between the energy acquisition and microbial community composition is likely to depend on which criterion was used for identifying core taxa, thus potentially leading to different interpretations.

While a focus on a core set of taxa has been common, research suggests it may not be entirely warranted (Engel & Moran, 2013; Hammer, Sanders, & Fierer, 2019; Martinson, Moy, & Moran, 2012) and that all taxa could instead be used for analysis. More attention is now being paid to the "rare biosphere" and the contribution these less abundant taxa make to ecosystem processes (reviewed in Jousset et al., 2017), community structure (Mikkelson et al., 2016; Shade et al., 2014), and as a reservoir of metabolic diversity (Mikkelson et al., 2016). In ignoring these rare taxa and focusing solely on common ones, researchers may wrongfully attribute the functions of rare taxa to common ones. This is especially concerning in human microbiome and agricultural studies where environments are scanned for beneficial microbes, e.g. if the functions of rare taxa are attributed to common ones, researchers may be chasing the wrong taxa for biotechnological applications. In addition, this dichotomous assignment of core and non-core ignores situations in which taxa, both common and rare, form networks and function through interactions. Our network analysis highlights many taxa that are characterized by high degree centrality but were excluded from any core assignment method, again demonstrating the danger of focusing solely on taxa assigned to a core.

To proceed with the analysis of a core set of taxa, researchers can either investigate the effects of focal taxa that were chosen based on other information (analogous to candidate gene analysis), or closely examine evidence for categorical differences in the abundance of taxa. Statistical evidence for a distinction between core and non-core taxa could come from consistent categorization by different core assignment methods. Our analyses demonstrate that while a common core can be assigned by multiple methods, co-assignments are limited by the most conservative method, which can change based upon the underlying structure of the taxon table.

Alternatively, researchers can rely on established statistical approaches that incorporate variation in the abundance of all taxa. These include standard methods for multivariate analysis, including dimension reduction, including those implemented in statistical packages such *vegan* (Oksanen et al., 2018) and *phyloseq* (McMurdie & Holmes, 2013). Additionally, specialized methods for differential abundance analysis exist, including Dirichlet multinomial models (Grantham, Guan, Reich, Borer, & Gross, 2019; Harrison et al., 2020; La Rosa et al., 2012; Shafiei et al., 2015) and related methods that model the relative abundance of all taxa (Fernandes et al., 2014; Love, Huber, & Anders, 2014; Mandal et al., 2015; Robinson, McCarthy, & Smyth, 2009; Wang et al., 2015).

In summary, our application of core assignment methods to simulated and published data sets demonstrated the inconsistent classifications that resulted from commonly applied criteria for determining membership in the set of core taxa. Changes in the set of taxa assigned to the core could lead to drastically different conclusions regarding statistical associations and ecological consequences. These findings suggest that analyses that rely on the identification of core taxa should be disfavored in many cases and instead researchers can rely on multivariate analyses that make use of all of the abundance data.

Data accessibility: Taxon tables, simulated data and all code used for analysis are available online in the CoreMicro R package found at https://github.com/MayaGans/CoreMicro.git and zenodo repository (10.5281/zenodo.4909346) .

Acknowledgments: This research was supported by the Microbial Ecology Collaborative with funding from NSF award #EPS-1655726.

Author contributions:

M.G. and G.C. conceived the ideas presented. M.G., G.C., and C.A.B. wrote code for testing core hypothesis and simulations. M.G., G.C., L.v.D., and C.A.B. developed and edited manuscript. CoreMicro R package was developed by M.G. and G.C.

Table List:

List of core methods from Web of Science literature review

Summary of published data sets and core inclusion by method

Figure List:

1. Heat map of true positive rate, false positive rate, and net assignment value for assigning core taxa by method for simulated data
2. Bivariate plot of core inclusion by method and data set
3. Venn diagram of core co-assignments for each of the datasets used.
4. Top panels (A & B): Venn diagrams of all core assignments compared to significant nodes identified from cooccurrence network analysis. Bottom panels (C & D): Venn diagrams of core assignments by individual methods shared with significant nodes from cooccurrence networks.

Supplementary tables – available at 10.5281/zenodo.4909346.

Literature review reference table

Ahrendt, S. R., Quandt, C. A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., ... Grigoriev, I. V. (2018). Leveraging single-cell genomics to expand the fungal tree of life. *Nature Microbiology* , *3* (12), 1417–1428. https://doi.org/10.1038/s41564-018-0261-0

Amor, D. R., Ratzke, C., & Gore, J. (2020). Transient invaders can induce shifts between alternative stable states of microbial communities. *Science Advances* , *6* (8), eaay8676. https://doi.org/10.1126/sciadv.aay8676

Astudillo-García, C., Bell, J. J., Webster, N. S., Glasl, B., Jompa, J., Montoya, J. M., & Taylor, M. W. (2017). Evaluating the core microbiota in complex communities: A systematic investigation. *Environmental Microbiology* , *19* (4), 1450–1462. https://doi.org/10.1111/1462-2920.13647

Banerjee, S., Kirkby, C. A., Schmutter, D., Bissett, A., Kirkegaard, J. A., & Richardson, A. E. (2016). Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry* , *97* , 188–198. https://doi.org/10.1016/j.soilbio.2016.03.017

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning.*Nature Reviews Microbiology* , *16* (9), 567–576. https://doi.org/10.1038/s41579-018-0024-1

Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J., & Holmes, S. P. (2016). Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. *F1000Research* ,*5* (3), 1492. https://doi.org/10.12688/f1000research.8986.2

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., ... Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biology* , *12* (5). https://doi.org/10.1186/gb-2011-12-5-r50

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* , *11* (6), 415–425. https://doi.org/10.1038/nrg2779

Clooney, A. G., Fouhy, F., Sleator, R. D., O'Driscoll, A., Stanton, C., Cotter, P. D., & Claesson, M. J. (2016). Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis.*PLoS ONE* , *11* (2), 1–16. https://doi.org/10.1371/journal.pone.0148028

Consortium, T. H. M. P. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* , *486* (7402), 207–214. https://doi.org/10.1038/nature11234

[Dataset] Gans, M., Custer, G.F., Buerkle, C.A., van Diepen, L.T.A. 2021. Zenodo. 1.4.0. 10.5281/zenodo.4909346.

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., ... Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science* , *359* (6373), 320–325. https://doi.org/10.1126/science.aap9516

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., ... Rohwer, F. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* ,*452* (7185), 340–343. https://doi.org/10.1038/nature06735

Domagalski, R., Neal, Z., & Sagan, B. (2019). backbone: An R Package for extracting the backbone of bipartite projections, 1–17. Retrieved from http://arxiv.org/abs/1912.12779

Dubos, R., Schaedler, R. W., Costello, R., & Hoet, P. (1965). Indigenous, normal, and autochthonous flora of the gastrointestinal tract. *The Journal of Experimental Medicine* , *122* (1), 67–76. https://doi.org/10.1084/jem.122.1.67

Engel, P., & Moran, N. A. (2013). The gut microbiota of insects - diversity in structure and function. *FEMS Microbiology Reviews* ,*37* (5), 699–735. https://doi.org/10.1111/1574-6976.12025

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & 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. https://doi.org/10.1186/2049-2618-2-15

Forcino, F. L., Leighton, L. R., Twerdy, P., & Cahill, J. F. (2015). Reexamining Sample Size Requirements for Multivariate, Abundance-Based Community Research: When Resources are Limited, the Research Does Not Have to Be. *PloS One* , *10* (6), e0128379–e0128379. https://doi.org/10.1371/journal.pone.0128379

Geisen, S., Laros, I., Vizcaíno, A., Bonkowski, M., & de Groot, G. A. (2015). Not all are free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa.*Molecular Ecology* , *24* (17), 4556–4569. https://doi.org/10.1111/mec.13238

Grantham, N. S., Guan, Y., Reich, B. J., Borer, E. T., & Gross, K. (2019). MIMIX: A Bayesian Mixed-Effects Model for Microbiome Data From Designed Experiments. *Journal of the American Statistical Association* . https://doi.org/10.1080/01621459.2019.1626242

Gray, K. I. . U. and J. S. ., Amjad, S., & Gray, J. S. (1983). Lognormal Distributions and the Concept of Community Equilibrium.*Marine Pollution Bulletin.* , *39* (2), 178–181. https://doi.org/10.2307/3544482

Grime, J. P. (1998). Benefits of plant diversity to ecosystems: immediate, filter and founder effects. *Journal of Ecology* ,*86* (6), 902–910. https://doi.org/10.1046/j.1365-2745.1998.00306.x

Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges.*Genome Research* . https://doi.org/10.1101/gr.085464.108

Hammer, T. J., Janzen, D. H., Hallwachs, W., Jaffe, S. P., & Fierer, N. (2017). Caterpillars lack a resident gut microbiome. *Proceedings of the National Academy of Sciences* , *114* (36), 9641–9646. https://doi.org/10.1073/pnas.1707186114

Hammer, T. J., Sanders, J. G., & Fierer, N. (2019). Not all animals need a microbiome. *FEMS Microbiology Letters* , *366* (10), 1–11. https://doi.org/10.1093/femsle/fnz117

Hanski, I. (1982). Dynamics of Regional Distribution: The Core and Satellite Species Hypothesis. *Oikos* , *38* (2), 210. https://doi.org/10.2307/3544021

Harrison, J. G., Beltran, L. P., Buerkle, C. A., Cook, D., Gardner, D. R., Parchman, T. L., ... Forister, M. L. (2021). A suite of rare microbes interacts with a dominant, heritable, fungal endophyte to influence plant trait expression. *The ISME Journal* , 608729. https://doi.org/10.1038/s41396-021-00964-4

Harrison, J. G., Calder, W. J., Shastry, V., & Buerkle, C. A. (2020). Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular Ecology Resources* , *20* (2), 481–497. https://doi.org/10.1111/1755-0998.13128

Hawinkel, S., Kerckhof, F. M., Bijnens, L., & Thas, O. (2019). A unified approach to unconstrained and constrained ordination of microbiome count data. *Under Review* , 1–20.

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., ... Hol, G. W. H. (2017). Where less may be more: How the rare biosphere pulls ecosystems strings. *ISME Journal* ,*11* (4), 853–862. https://doi.org/10.1038/ismej.2016.174

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., ... Shannon, W. D. (2012). Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. *PLoS ONE* ,*7* (12), e52078. https://doi.org/10.1371/journal.pone.0052078

Ley, R. E. (2010). Obesity and the human microbiome. *Current Opinion in Gastroenterology* , *26* (1), 5–11. https://doi.org/10.1097/MOG.0b013e328333d751

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* , *33* (2), 177–182. https://doi.org/10.1038/ng1071

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* , *15* (12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., ... Dangl, J. L. (2012). Defining the core <i>Arabidopsis thaliana<i> root microbiome. *Nature* , *488* (7409), 86–90. https://doi.org/10.1038/nature11237

Lynch, M. D. J., & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology* , *13* (4), 217–229. https://doi.org/10.1038/nrmicro3400

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease* , *26* (0), 1–8. https://doi.org/10.3402/mehd.v26.27663

Martin, B. D., Witten, D., & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics* , *14* (1). https://doi.org/10.1214/19-AOAS1283

Martinson, V. G., Moy, J., & Moran, N. A. (2012). Establishment of Characteristic Gut Bacteria during Development of the Honeybee Worker.*Applied and Environmental Microbiology* , *78* (8), 2830–2840. https://doi.org/10.1128/aem.07810-11

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* , *8* (4), e61217. https://doi.org/10.1371/journal.pone.0061217

Mikkelson, K. M., Bokman, C. M., & Sharp, J. O. (2016). Rare Taxa Maintain Microbial Diversity and Contribute to Terrestrial Community Dynamics throughout Bark Beetle Infestation. *Applied and Environmental Microbiology* , *82* (23), 6912–6919. https://doi.org/10.1128/AEM.02245-16

Neal, Z. (2013). Identifying statistically significant edges in one-mode projections. *Social Network Analysis and Mining* , *3* (4), 915–924. https://doi.org/10.1007/s13278-013-0107-y

Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology* ,*15* (6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2018). vegan: Community Ecology Package.

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology* , *84* (7), 1–12. https://doi.org/10.1128/AEM.02627-17

Porazinska, D. L., Giblin-Davis, R. M., Esquivel, A., Powers, T. O., Sung, W., & Thomas, W. K. (2010). Ecometagenetics confirm high tropical rainforest nematode diversity. *Molecular Ecology* , *19* (24), 5521–5530. https://doi.org/10.1111/j.1365-294X.2010.04891.x

Pritchard, J. K., & Cox, N. (2002). The allelic architecture of human disease genes: common disease-common variant... or not? *Human Molecular Genetics* , *11* (20), 2417–2423. https://doi.org/10.1093/hmg/11.20.2417

R Development Core Team. (2020). A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* . Vienna, Austria. Retrieved from http://www.r-project.org

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26* (1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Saunders, A. M., Albertsen, M., Vollertsen, J., & Nielsen, P. H. (2016). The activated sludge ecosystem contains a core community of abundant organisms. *The ISME Journal* , *10* (1), 11–20. https://doi.org/10.1038/ismej.2015.117

Savage, D. C. (1977). Microbial Ecology of the Gastrointestinal Tract.*Annual Review of Microbiology* , *31* (1), 107–133. https://doi.org/10.1146/annurev.mi.31.100177.000543

Shade, A., & Handelsman, J. (2012). Beyond the Venn diagram: the hunt for a core microbiome. *Environmental Microbiology* , *14* (1), 4–12. https://doi.org/10.1111/j.1462-2920.2011.02585.x

Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., & Gilbert, J. A. (2014). Conditionally Rare Taxa Disproportionately Contribute to Temporal Changes in Microbial Diversity. *MBio* , *5* (4), 3–11. https://doi.org/10.1128/mBio.01371-14

Shade, A., & Stopnisek, N. (2019). Abundance-occupancy distributions to prioritize plant core microbiome membership. *Current Opinion in Microbiology* , *49* , 50–58. https://doi.org/10.1016/j.mib.2019.09.008

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution* , *8* (8), 907–917. https://doi.org/10.1111/2041-210X.12700

Shafiei, M., Dunn, K. A., Boon, E., MacDonald, S. M., Walsh, D. A., Gu, H., & Bielawski, J. P. (2015). BioMiCo: A supervised Bayesian model for inference of microbial community structure. *Microbiome* ,*3* (1), 8. https://doi.org/10.1186/s40168-015-0073-x

Shi, Y., Delgado-Baquerizo, M., Li, Y., Yang, Y., Zhu, Y. G., Peñuelas, J., & Chu, H. (2020). Abundance of kinless hubs within soil microbial networks are associated with high functional potential in agricultural ecosystems. *Environment International* , *142* (April), 105869. https://doi.org/10.1016/j.envint.2020.105869

Soliveres, S., van der Plas, F., Manning, P., Prati, D., Gossner, M. M., Renner, S. C., ... Allan, E. (2016). Biodiversity at multiple trophic levels is needed for ecosystem multifunctionality.*Nature* , *536* (7617), 456–459. https://doi.org/10.1038/nature19092

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., ... Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports* , *7* (1), 12240. https://doi.org/10.1038/s41598-017-12501-5

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., ... Abarenkov, K. (2014). Global diversity and geography of soil fungi. *Science* , *346* (6213), 1256688–1256688. https://doi.org/10.1126/science.1256688

Turnbaugh, P. J., & Gordon, J. I. (2009). The core gut microbiome, energy balance and obesity. *Journal of Physiology* ,*587* (17), 4153–4158. https://doi.org/10.1113/jphysiol.2009.174136

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., ... Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature* , *457* (7228), 480–484. https://doi.org/10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project.*Nature* , *449* (7164), 804–810. https://doi.org/10.1038/nature06244

Umaña, M. N., Zhang, C., Cao, M., Lin, L., & Swenson, N. G. (2017). A core-transient framework for trait-based community ecology: an example from a tropical tree seedling community. *Ecology Letters* ,*20* (5), 619–628. https://doi.org/10.1111/ele.12760

Wang, A. Y., Naumann, U., Wright, S., Eddelbuettel, D., Warton, D., & Davidwartonunsweduau, M. D. W. (2015). mvabund: Statistical Methods for Analysing Multivariate Abundance Data. CRAN. Retrieved from https://cran.r-project.org/web/packages/mvabund/index.html

Winfree, R., Fox, J. W., Williams, N. M., Reilly, J. R., & Cariveau, D. P. (2015). Abundance of common species, not species richness, drives delivery of a real-world ecosystem service. *Ecology Letters* ,*18* (7), 626–635. https://doi.org/10.1111/ele.12424

Wirth, R., Kádár, G., Kakuk, B., Maróti, G., Bagi, Z., Szilágyi, Á., ... Kovács, K. L. (2018). The Planktonic Core Microbiome and Core Functions in the Cattle Rumen by Next Generation Sequencing.*Frontiers in Microbiology* , *9* (SEP), 2285. https://doi.org/10.3389/fmicb.2018.02285

Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S. R., Marinier, E., Van Domselaar, G., ... McAllister, T. A. (2018). Impact of sequencing depth on the characterization of the microbiome and resistome.*Scientific Reports* , *8* (1), 5890. https://doi.org/10.1038/s41598-018-24280-8

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* , *44* (7), 821–824. https://doi.org/10.1038/ng.2310

Table 1. Five commonly employed core methods along with descriptions and number of publications found using these methods within Web of Science, accessed April 2018. The Venn Diagram method was not utilized in our analysis due to its similarity to the proportion of replicates method.

Core method	Method description
Cumulative proportion of sequence reads	Accounts for read depth (accounting for a portion of top 75% of total reads)
Proportion of replicates	Accounts for sample size (present in >50% of samples)
Proportion of sequence reads and replicates	Accounts for both sample size and read depth (present in >50% of samples w
Hard cutoff	Hard cutoffs i.e. specific number of samples or reads (present in 5 samples wi
Venn Diagram	Present in all subsets of samples

Table 2. Summary information for the two selected published datasets, including a summary of the number and percent of operational taxa assigned to the core by each of the four methods. The *Arabidopsis thaliana* data set was generated by Lundberg et al. (2012) and only utilizes rhizosphere samples from the M21 site. Human microbiome data was generated by The Human Microbiome Consortium (2012) and includes only fecal samples.

	Arabidopsis thaliana Microbiome	Human Microbiome Project
Total taxa	14,890	11,752
Total reads	1,770,731	1,893,867
Total samples	288	319
NCBI accession number	ERP001384	HM16STR

Sequencing platform	454	Illumina
Method	*Number of taxa assigned to core*	*Number of taxa assigned to core*
Cumulative proportion of sequence reads	1245 (8.36%)	1108 (9.42%)
Proportion of replicates	2036 (13.67)	1850 (15.74%)
Proportion of sequence reads and replicates	907 (6.09%)	204 (1.73%)
Hard cutoff	181 (1.21%)	554 (4.71%)
Not assigned to core by any method	12854 (86.32%)	9902 (84.3%)
Unique taxa assigned to core by any method	2,036	1,850

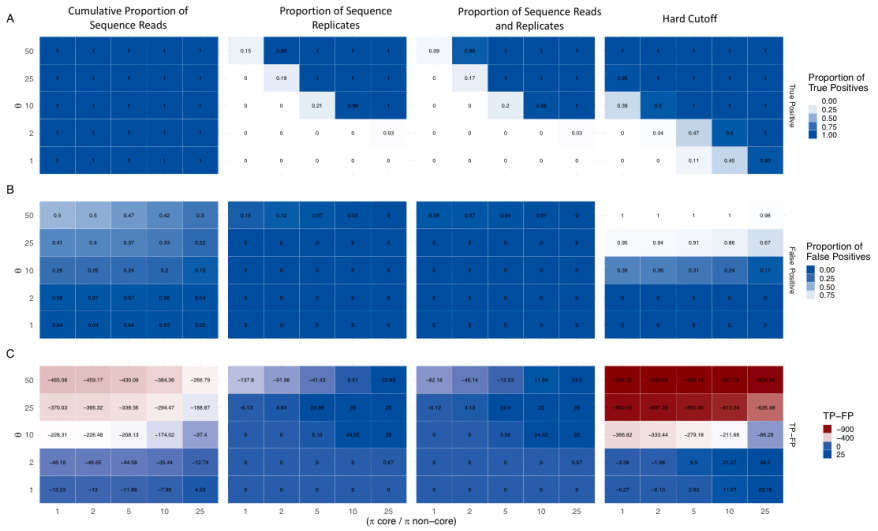

Figure 1. a) True positive rate, b) false positive rate, and c) net assignment value for assignment of taxa to the core by four different core assignment methods. Individual 5x5 heatmaps represent 6,250 simulations each. Each square within a heatmap represents 250 simulations at one of the possible 25 combinations of intensity ($\vartheta = 1$ corresponding to low precision in taxon abundance, and 2, 10, 25, 50 corresponding to increasing precision in taxon frequency), and ratio of the abundance of core to non-core taxa (1, which is a simulation with no difference in the expected taxon frequencies, and 2, 5, 10, 25 that corresponding to greater differences in the frequency of core and non-core taxa). The top row a) presents true positive rates, giving the probability that taxa simulated with core characteristics were recovered as such. A true positive rate of 1 means all 25 true core taxa were assigned as such. The middle row b) presents false positive rates or the probability that non-core taxa were assigned as core. A false positive rate of zero represents simulations when no non-core taxa were assigned to the core, whereas a rate of 1 indicates all non-core taxa were incorrectly assigned to the core. The bottom row c) presents the net assignment rate, an absolute value of true positives – false positives for each of the four methods. A positive net assignment rate indicates better performance, while larger negative numbers indicate poorer assignment of core membership.

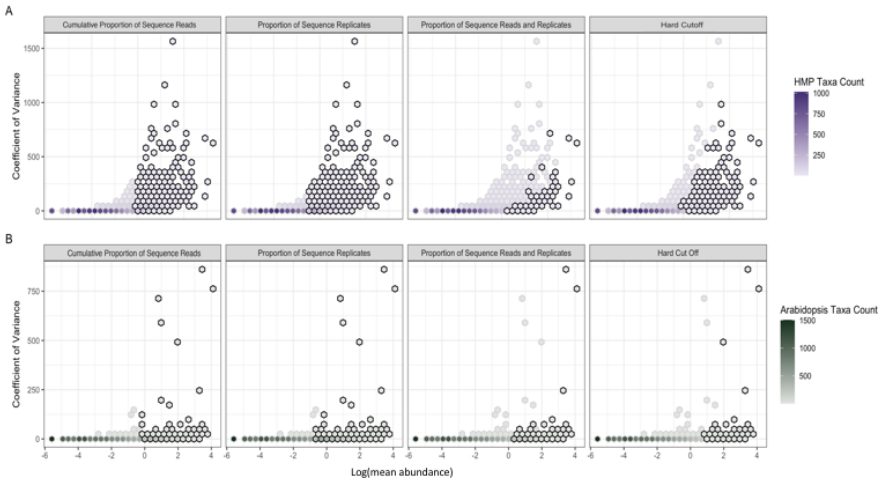

Figure 2. Four core methods identify different sets of core taxa from abundance data in a study of human microbiomes (a), HMP, (Consortium, 2012); and b) *Arabidopsis thaliana* , (Lundberg et al., 2012). Additionally, the sets of core and non-core taxa do not exhibit categorically distinguishable abundances. The outlined bins denote whether taxa in the bin were included within the core. Fill color corresponds to taxa counts. The hard cutoff and proportion of reads methods categorize taxa on the basis of an abundance threshold (horizontal axis) without accounting for variance (vertical axis), whereas methods based on the proportion of replicates incorporates both abundance and variance criteria.

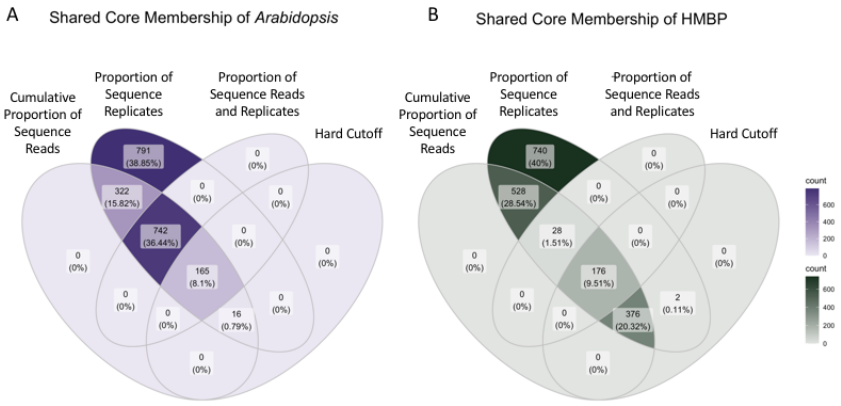

Figure 3. Venn diagrams of co-assigned core taxa. a) Left, core taxa co-assignments for the *Arabidopsis* dataset, b) Right, core taxa co-assignments for the human microbiome project dataset. Numbers in the overlapping regions represent the number of core assignments shared by that pair or combination of core assignment methods. Percentages represent the percent of total core assignments. For example, in the*Arabidopsis* dataset, the center of the Venn diagram shows there were 165 core assignments shared by all four methods. This 165 core assignments represents 8.1% of the total core assignments.

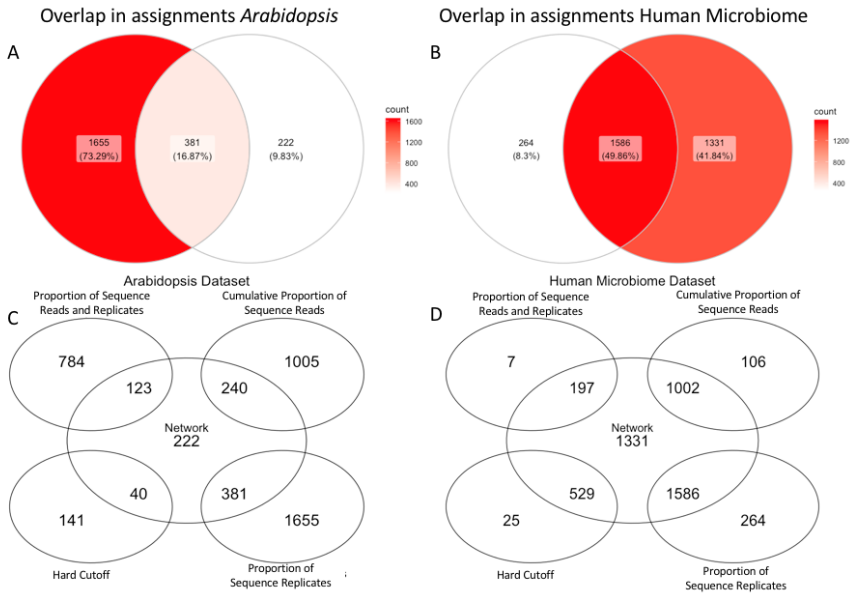

Figure 4. Venn diagrams of overlapping assignments by core methods and cooccurrence networks. Top: Each figure shows the overlap in assignments by the cooccurrence networks and at least a one core assignment method for the (a) *Arabidopsis* dataset and (b) Human Microbiome Project dataset. The left side of each figure shows the number of core assignments that were not deemed significant nodes by the cooccurrence network. The center shows the number of taxa that were assigned core membership and significance by the cooccurrence network, and the right portion of each Venn diagram shows those taxa that were deemed significant by the cooccurrence network but were not assigned core membership by any core assignment method. Bottom: Venn diagrams of shared taxa between individual core assignment methods and the cooccurrence networks c) *Arabidopsis* dataset d) Human Microbiome Project dataset. The center of the Venn diagrams show the number of taxa that were deemed significant by the cooccurrence network but not assigned core membership by any of the core assignment methods. Each of the outer circles is an individual Venn diagram showing those taxa that were assigned core membership by that method but not assigned significance by the cooccurrence network (outer) and those taxa that were assigned by both (overlap).

Hosted file

`Gans_et_al_2021_MarkedUpManuscript.docx` available at https://authorea.com/users/418784/articles/525431-methods-to-delineate-membership-in-a-core-community-are-inconsistent-rarely-test-the-hypothesis-of-a-core-and-can-mislead-ecological-analysis