IMPERIAL

IMPERIAL COLLEGE LONDON
DEPARTMENT OF LIFE SCIENCES
SILWOOD PARK CAMPUS

The role of interactions in determining the impact of keystone taxa upon microbial community functions

---

Author: Matthew Shaun Grainger
CID: 01875008
matthew.grainger20@imperial.ac.uk

Supervisors:
Samraat Pawar &
Alberto Pascual-Garcia

5750 words

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Research at Imperial College London

Submitted for the MRes in Computational Methods in Ecology and Evolution August 2024

Declaration

I, Matthew Shaun Grainger, declare that the work done in this project is my own. The data was provided to me by one of my supervisors, Dr Alberto Pascual-Garcia, from his previously published literature. This literature has been cited appropriately. I was responsible for extensive data processing and cleaning, although this sometimes included the modification or use of a function extracted from the ”clean_ASV_table.R” script, from the GitHub repository of Dr. Pascual-Garcia. This script has been cited appropriately, and its GitHub of origin can be found via the Data Availability section of this dissertation. I did not develop any mathematical models. I received guidance from both of my supervisors, Prof. Samraat Pawar and Dr. Alberto Pascual-Garcia over the duration of my project.

Abstract

Microbial communities provide functions that facilitate all life on Earth, and the functions that each community provides is dependent upon its structure. The structure of a microbial community can be described by the abundances of the taxa within it and the interactions between them. Taxa may impact community functions via one of three methods: 1) providing isolated contributions that are independent of other taxa, 2) engaging in interactions that change the abundances of other taxa, and thus their associated functions, or 3) by forming subgroups in which taxa interact to provide functions that they would not be able to provide in isolation. Here, I investigate the role of interactions in determining the impact of keystone taxa upon microbial community functions, via the latter two methods. To do this, I identified whether some taxa were more important in determining community functions than others, and whether this related to their interactions or membership within a subgroup. This involved the use of machine learning techniques to infer interactions between taxa and to relate their interactions with their ability to determine community-level functions. Here, I found that interacting taxa were better able to explain variance in functions than non-interacting taxa, and that taxa that were involved in subgroups able to explain most of this variance. These findings were particularly robust, due to the use of a large data set of replicated microbial communities. They support previous studies upon the importance of highly interacting keystone taxa in determining microbial community structure and function, providing important implications upon the relationship between structure and function within other key microbial communities

1 Introduction

Microbial communities, groups of microbial taxa that share and interact within the same environment, provide functions that underpin all life on Earth (Konopka, 2009; Widder et al., 2016; Nemergut et al., 2013). In the natural world, these communities drive the biogeochemical cycles on Earth, sequester carbon, improve soil fertility, promote plant growth and suppress disease (Sokol et al., 2022; Falkowski et al., 2008). Furthermore, microbial communities that are associated with host organisms such as humans play a key role in maintaining their health, via their impact upon physiology, nutrition, behavior and development (Venturelli et al., 2018; Culp and Goodman, 2023). In addition to contributing these key functions within natural systems, microbial communities are also used within a range of industries for direct human-benefit. They supply functions as diverse as bioremediation to remove heavy metals from wastewater (Sharma et al., 2021), the biofertilisation of crops (Lopes et al., 2021), and the fermentation of a range of food and drink products (Wolfe and Dutton, 2015). Because microbial communities confer so many crucial functions, there has been considerable interest in the mechanisms by which they achieve this.

The structure of a microbial community determines its functions. This structure is characterised by two components: the abundances of taxa (taxa composition) and the interactions between them. Communities with different taxa compositions often provide different functions, or may provide the same functions to different extents (Strickland et al., 2009; Waldrop et al., 2000). This is partly because certain taxa provide individual, isolated contributions to functions (Bashan and de Bashan, 2010; Trapet et al., 2016), however sometimes the functions that are provided by microbial communities are different to what would be observed from the sum of these isolated functions (Korir et al., 2017; Morin et al., 2022). These non-additive effects upon functions are known as emergent properties, and they are the result of microbial interactions (Geesink et al., 2024; van den Berg et al., 2022; Kodera et al., 2022; Röttjers and Faust, 2018). Taxa that engage in these interactions are therefore of great relevance to microbial community functions, and are referred to as being 'keystone' (Banerjee et al., 2018).

Keystone taxa drive community structure and function, and their interactions often play a role in this. Here, I suggest three methods by which microbial taxa can provide a function of interest, shown in Fig.1. The first of these methods is via their isolated contributions to functions, as seen in the production of insecticidal toxins by *Bacillus thuringiensis*. These contributions do not involve interactions with other microbial taxa and could be provided in isolation (Fig.1a). The second method of contribution to function is via interactions with other taxa that impact their abundances or transcription, and thus their isolated contributions to function (Fig.1b). This can be seen in the reduction of pH as a byproduct of *Lactobacillus* metabolism, which inadvertently excludes the pathogen that causes bacterial vaginosis and facilitates acidophilic taxa (Pace et al., 2021; Breshears et al., 2015). The third method in which taxa provide functions is by forming subgroups, which include a minimum of two taxa (Fig.1c). The back-and-forth interactions within these subgroups can provide emergent properties that taxa could not contribute in isolation. In the gut microbiome, for instance, *Bifidobacterium adolescentis* engages in a cross-feeding interaction with *Eubacterium hallii* to produce butyrate, something that neither taxa produces readily in the absence of the other (Belenguer et al., 2006).

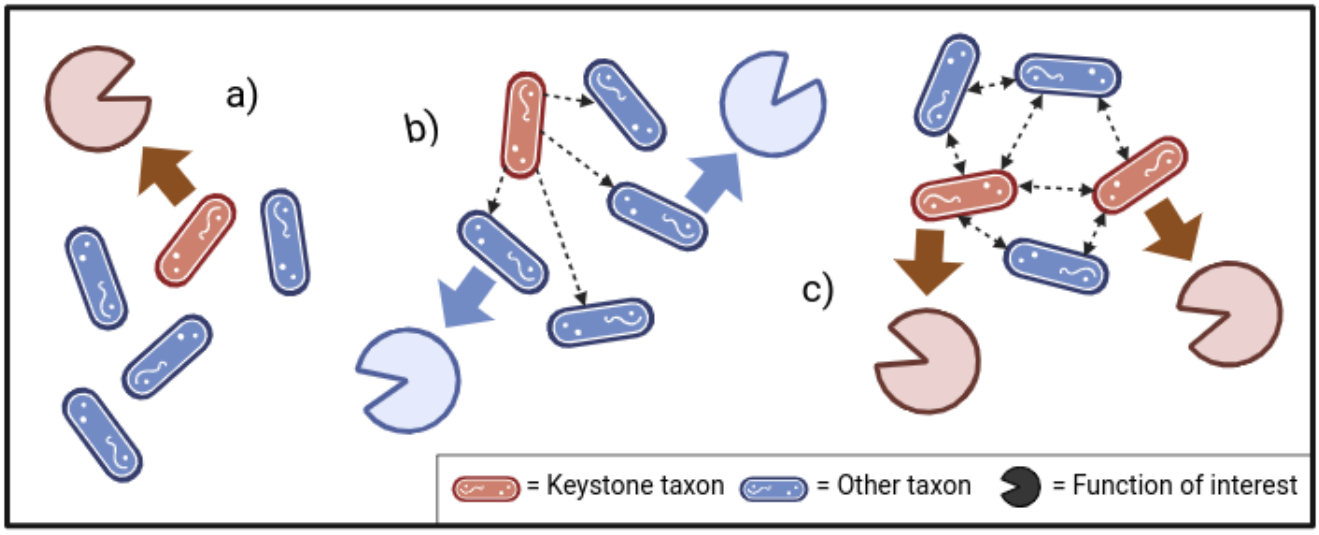

Figure 1: Keystone taxa can determine a function of interest, represented here as an enzyme, via a) direct individual contributions that are unrelated to other taxa, b) interactions that impact taxa which confer functions, or c) the formation of subgroups that provide shared emergent properties. In b) and c) the interactions themselves can be provided by other functions, such as the production of intermediate metabolites.

While a number of studies have begun to investigate the role of keystone taxa in determining microbial community structure and function, few have directly compared these three methods of action. In this study, I investigated the role of interactions in determining the impact of bacteria upon microbial community functions, involving a comparison between the contributions of these three methods. To do this, I aimed to answer the following questions using a large data set of relatively understudied bacterial communities from freshwater tree hole pools.

1. Were some taxa more important in determining community functions than others, or did taxa have equal importance?

2. Were taxa that were involved in interactions more important in determining community functions than non-interacting taxa?

3. Were lone taxa or subgroups of taxa more important in determining community functions?

By answering these questions, I further intended to elucidate whether the importance of taxa in determining function was due to their isolated contributions, their impact upon other taxa, or their involvement in cooperative subgroups. This unique data set of replicate microbial communities allowed me to infer robust microbial interactions, so that I could relate the importance of taxa in network structure to their importance in determining functions.

2 Methods

To investigate my three questions, I compared the ability of three sets of random forest regression models to explain variance in six functions between communities. The first set of models used the abundances of all of the taxa in the communities, the second set used only the most interactive taxa, and the third set used only subgroups of taxa, to the exclusion of lone taxa. For the latter two models, I constructed a co-occurrence network to determine the number of interactions that taxa were involved in, and I used an algorithm to group taxa into subgroups based upon these interactions.

2.1 Dataset

Bacterial amplicon sequence variant (ASVs) abundance table sourced from Pascual-García et al. (2023). In brief, this table contains the abundances of ASVs from 275 bacterial communities within the rainwater pools of beech tree roots (*Fagus sylvatica*). The communities were extracted, grown in a sterile beech leaf medium, and incubated at 22°C under static conditions for 1 week so that they could reach stationary phase. The communities were then cryopreserved. Following cryopreservation, each of the 275 communities were revived four times and at the same time to provide four replicates, and incubated under static conditions at 22°C for 7 days. Samples of the 275 stationary state communities were taken prior to cryopreservation (starting time-point communities), and samples of all 4 of the replicates of these 275 communities were taken at the end of the 7 day growth period that followed their revival (final time-point communities). In my analysis, I consider all of these samples to be their own communities. The composition of each of the samples was determined via sequencing the V4 region of the 16S rRNA gene, and by quality filtering using the DADA2 pipeline (Callahan et al., 2016). Functional measurements from a subset of the final time-point communities were sourced from (Pascual-García and Bell, 2020). These included the cell count, the mass of carbon dioxide and the concentrations of ATP, xylosidase, glucosidase, chitinase and phosphatase. All functional measurements were normalised by the cell count and log-transformed prior to use. The ASV abundance table contains the abundances of 1468 ASVs across the 275 starting time-point samples and 1100 final time-point samples. Out of the 1100 final time-point samples, 932 samples were associated functional measurements, and there were 1458 ASVs across these samples.

The metadata table that accompanied the ASV abundance table was also sourced from Pascual-García et al. (2023). This metadata table was used to identify starting time-point and final time-point communities, as well as to provide information upon the community class of each sample. In short, these community classes are sets of communities obtained after clustering communities according to their beta diversity dissimilarity, computed using the Jensen-Shannon distance. The bacterial communities clustered in compositional space according to their collection location and date, which means that the different community classes reflect the slightly different environmental histories of the communities. In my investigation, therefore, I treat community classes as a proxy for the environment.

2.2 Determining whether some taxa were more important in determining community functions than others

I fit 6 random forest regression models, 1 for each function, using differences in the abundances of ASVs between communities as predictors to explain variance in the functions across communities. The logarithm of the number

of reads was used as an additional predictor, to control for differences in the number of reads between samples. I cleaned and processed the ASV table and function data in R (R Core Team, 2021), and used the clean_ASV_table function from Pascual-Garcia (2023) to remove data from unrelated experiments or with fewer than 10,000 reads. I used the randomForest regression package to fit each model (Liaw and Wiener, 2002), and I used the randomForestExplainer package to analyse the randomForest models, such as to extract and visualise the importance of different ASVs (Paluszynska et al., 2020).

Prior to fitting the random forest regression models, I first optimised the parameters that were used in these fits. To optimise the number of decision trees that were used in fitting the random forest regression models, I fit each of the models using different numbers of trees with a controlled number of variables, and identified the number of trees at which the mean squared error of the model no longer meaningfully decreased with the addition of further trees. For all of the models, this was at approximately 10,000 trees. I performed the same optimisation upon the number of variables, using a fixed number of trees. The optimal number of variables was found to be 750. These parameters were used to fit each of the 6 models. Optimisation was carried out using the HPC systems that are provided by the Research Computing Service (RCS) at Imperial College London.

To assess the robustness of the models, I trained 6 other models (1 for each function) upon only the first set of replicates of each community. I then used these models to predict each of the functions in the second, third and fourth sets of replicates. I compared the predicted values with the observed values to evaluate the fit of the models by calculating R-squared and mean-squared error values.

2.3 Determining whether taxa that were involved in interactions were more important in determining community functions than non-interacting taxa

I constructed a co-occurrence network first by splitting the ASV table into starting time-point and final time-point communities. I then converted them into the correct format, and input them separately into FlashWeave (Tackmann et al., 2019b). Briefly, FlashWeave is a recently developed machine-learning approach for inferring interactions between ASVs without being impacted by compositionality effects or indirect associations between taxa based upon their co-abundance. It has shown a strong ability to reconstruct both expert-curated interactions and those from synthetic data sets (Tackmann et al., 2019a). Here, I first compared two modes of FlashWeave: FlashWeave-S and FlashWeaveHE-S. The former is for use upon homogeneous data, whereas the latter takes environmental heterogeneity into account by incorporating metavariables as nodes within the network. Here, I used community classes as a proxy for the environment, and selected them as metavariables for FlashWeaveHE-S. When comparing the interaction tables, I found that FlashWeaveHE-S inferred far fewer interaction taxa, and far fewer interactions. I took this conservative approach, and merged the starting time-point and final time-point interactions tables that were produced by FlashWEaveHE-S. To do this, I removed the weights from interactions such that they were either '1' for positive or '-1' for negative. There were no discrepancies in the type of interaction between a given pair of ASVs between time-points. I imported the combined interaction table into Cytoscape, which provided me with network metrics (Shannon et al., 2003).

I determined the percentage of ASVs that were involved in interactions (ASVs with a Degree value of 1 or greater) and whether there was a power law distribution of the number of interactions amongst interacting ASVs. I then calculated the percentage of the taxa that were important in determining functions within the random forest regression models that were also involved in interactions. I also compared the median number of interactions that these taxa were involved in to the median number of interactions that all interacting taxa were involved in.

Following this, I identified the Pearson's correlation between the logarithm of the number of interactions that ASVs were involved in and the logarithm of their mean-square error increase, node purity increase, and mean minimum depth within the random forest regression models. Finally, I performed another set of 6 random forest regression models using the abundances of only taxa that were involved in the top 5% highest number of interactions to explain variance in function. I then compared the explanatory ability of these models to the first set of models. I fit these models in the same way as the first set of models, except for the reduction of the number of variables to the number of taxa that were involved in the highest number of interactions.

2.4 Determining whether lone taxa or subgroups of taxa were more important in determining community functions

I identified subgroups of taxa from their interaction using functionInk (Pascual-García and Bell, 2020). To summarise, this algorithm groups ASVs by their shared types of interactions with other ASVs. If two ASVs share the same neighbours with the same type of interaction, then they will be calculated as having a high similarity. ASVs with a high similarity are then placed in the same subgroup, up until the point at which the maximum density of shared interactions between ASVs in a subgroup or shared between ASVs in a subgroup and external ASVs is reached. Results in the formation of subgroups that are either dominated by interactions between taxa within the subgroup, or subgroups in which taxa do not themselves interact, but that have the same type of interactions with the same types of other taxa. In my investigation, the types of interaction were defined both by whether the interaction was between an ASV and a metavariable node or between two ASVs, and by whether the interaction was positive or negative (e.g. type 1 could be a positive interaction between two ASVs.) One of the output tables provided by functionInk contains the subgroup that each ASV was found in. I used this to identify the percentage of all taxa that were involved in subgroups, and the percentage of important taxa from the first set of random forest regressions that were involved in subgroups. I then combined the subgroup information with the ASV table to get the abundances of subgroups. This allowed me to perform a last set of random forest regression models using the abundances of subgroups to explain variance in functions. These models were fit in the same way as before, except that the number of variables was optimised to 50.

3 Results

3.1 Some taxa were more important in determining community functions than others

The random forest regression models that used differences in the abundances of taxa to explain differences in functions between communities indicated that some taxa were more important than others in explaining the differences in function between communities, as shown in Fig.2. Overall, the models suggested that differences in the abundances of taxa between communities were able to explain differences in all of their functions, as shown in Table 1. The models explained greater than 50% of the variation in functions between communities, with the exception of the model that explained the variation in ATP concentration per cell. In all of the models, some taxa appeared to play a much greater role than others in determining functions, to different extents. The taxa that were important in determining functions were also often important in determining multiple functions. When looking at the 10 most important taxa in determining function for each function, only 24 unique taxa were identified out of a possible 60. This suggests that approximately 2% of the 1458 taxa explained the majority of the differences in functions between communities.

The random forest models that were trained with only one replicate of the final communities, and tested upon the other three replicates in their ability to predict function were moderately effective. They explained little of the variance according to the R-squared values shown in Table 1, however the plots of their predicted values compared to the observed values in Fig.3 show that higher predicted values approximately matched higher observed values. Interestingly, for each function there were approximately two clusters of data points for which the predicted and observed values were similar.

Importance of different taxa in explaining the differences in functions between microbial communities

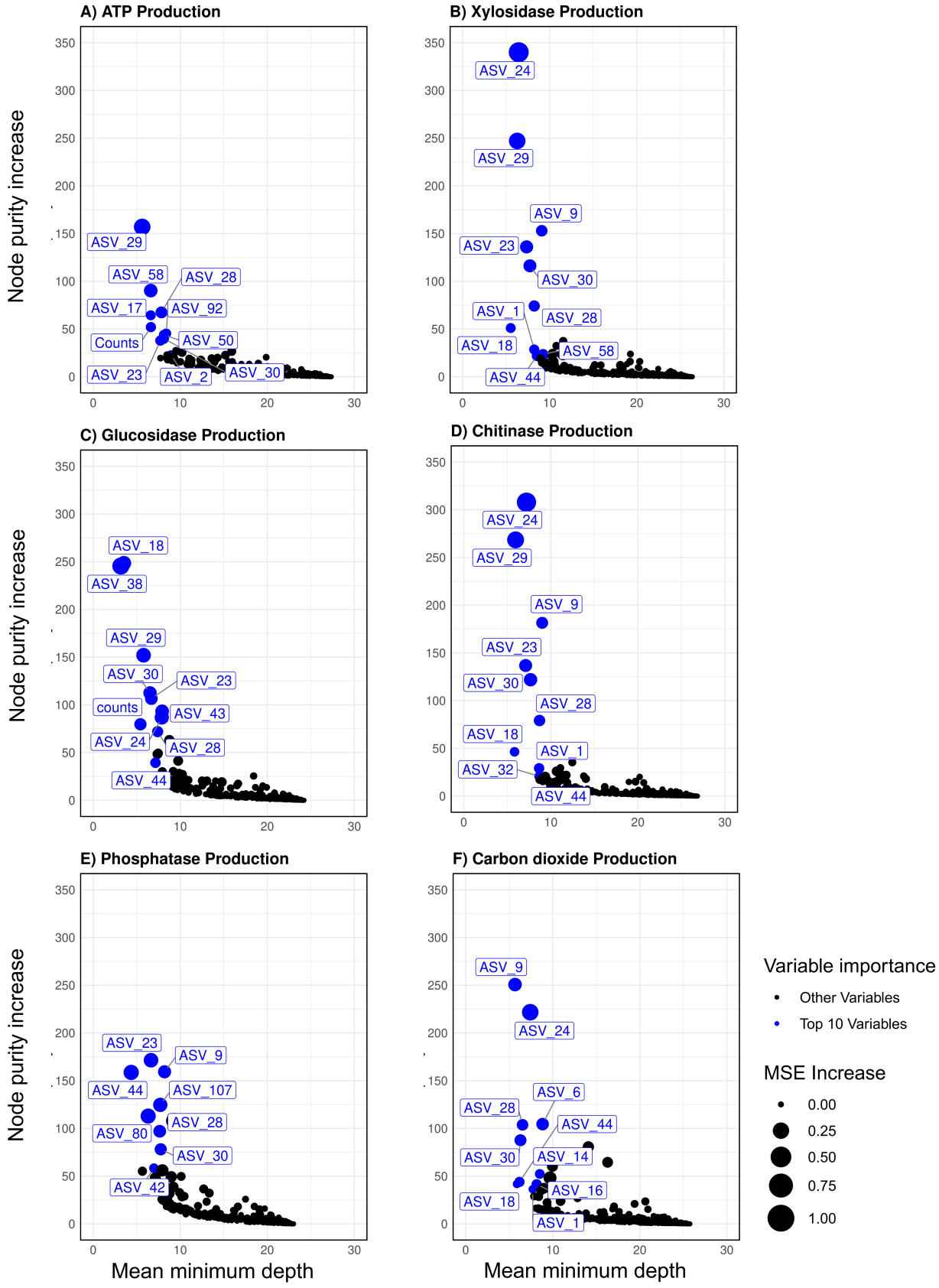

Figure 2: These plots are for the random forest regression models that use the abundances of all taxa to explain functions.

Ability of random forest regression model trained on a subset of the microbial communities to predict functions in the other communities

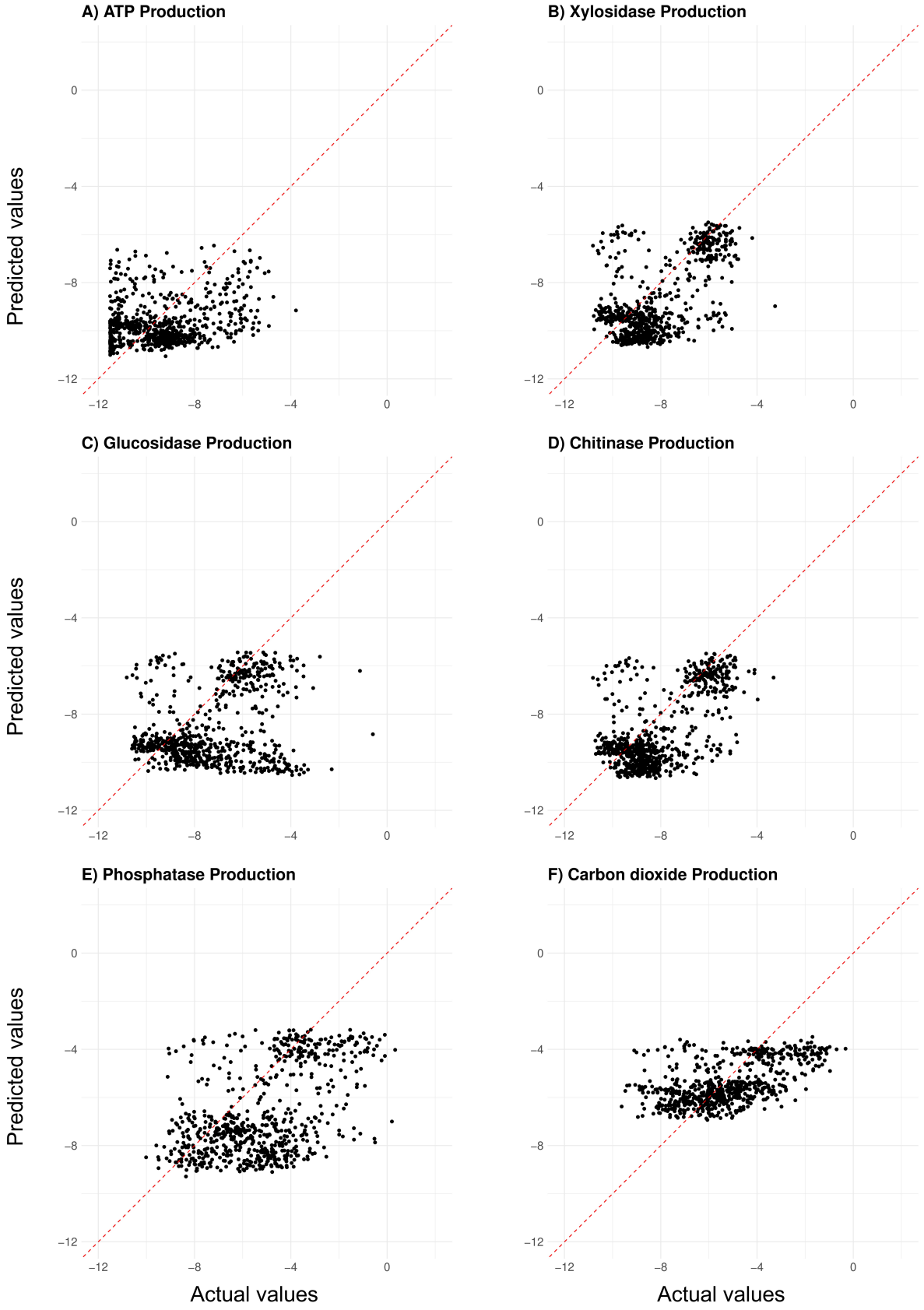

Figure 3: These prediction plots are for the random forest regression models that used the abundances of all taxa to explain differences in function.

3.2 Taxa that were involved in interactions were more important in determining community functions

The most important taxa in explaining differences in functions between communities were highly interactive. Out of these 24 taxa, 92% were involved in interactions. This was despite the fact that only 43% of the taxa in the community were involved in interactions. The most important taxa were also involved in more interactions than other taxa. The median number of interactions that the most important taxa were involved in was 20, whereas even when only considering taxa that were involved in one or more interactions, the median number of interactions in the community was 3 (as shown in Fig.). Over 80% of all interactions in the network were positive, and 74% of the interactions that important taxa were involved in were positive.

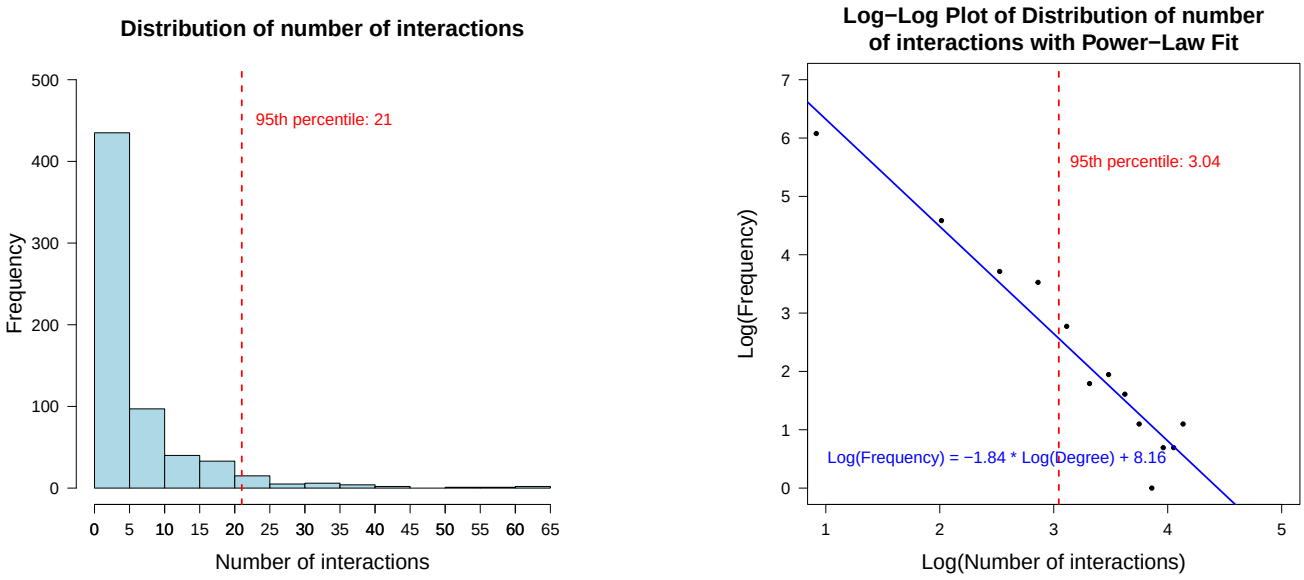

Figure 4: Out of the taxa that were involved in interactions, the majority of these taxa were involved in only very few interactions. The number of interactions that taxa were involved in follow a power-law, such that taxa that were involved in a high number of interactions were rare.

The number of interactions that taxa were involved in was positively correlated with their importance in determining functions. A Pearson correlation analysis revealed a significant negative correlation between the number of interactions that taxa were involved in and their mean minimum depth in the random forest regression models for ATP (r = -0.29, p ¡ 0.05, xylosidase (r = -0.35, p ¡ 0.01), glucosidase (r = -0.44, p ¡ 0.001) and chitinase (r = -0.50, p ¡ 0.001) per cell, as well as the mass of carbon dioxide per cell (r = - 0.31, p ¡ 0.001). There were also significant, slightly weaker negative correlations between the betweenness centrality of taxa and their ability in explaining each of these functions. For the models that explained variation in glucosidase and chitinase per cell, there was an additional positive correlation between the number of interactions that taxa were involved in and their node purity increase measure (r = 0.29, p ¡ 0.05 for the former and r = 0.25, p ¡ 0.05 for the latter).

The taxa that were involved in the highest 5% number of interactions determined the majority of community functions. When considering only taxa that were involved in interactions, the 95th percentile for the number of interactions was 21. Another set of random forest regression models that used the abundances of only taxa that

were involved in more than 21 interactions (the taxa involved in the top 5% number of interactions) to explain variation in the functions showed a similar explanatory ability to the original models that used the abundances of all taxa. While these models had a slightly lower explanatory power than the original models, as shown in Table 1, they used only 30 taxa compared to the 750 taxa that were used in the original models. Out of these 30 taxa, 11 were also the most important taxa in the original set of random forests. This is such that 40% of the most important taxa in the original models were involved in the top 5% highest number of interactions.

Table 1: Explanatory ability of random forest regression models using either the abundances of all taxa, taxa with the highest number of associations, all subgroups of taxa or the largest subgroups of taxa to explain differences in community functions.

Variables	Function	R^2	MSE
Models that were fit across all end-point communities			
All taxa	ATP	0.24	2.33
High association	ATP	0.17	2.41
Subgroups	ATP	0.20	2.32
All taxa	Xylosidase	0.58	1.15
High association	Xylosidase	0.51	1.32
Subgroups	Xylosidase	0.54	1.27
All taxa	Glucosidase	0.52	1.80
High association	Glucosidase	0.39	2.27
Subgroups	Glucosidase	0.46	2.00
All taxa	Chitinase	0.56	1.23
High association	Chitinase	0.50	1.39
Subgroups	Chitinase	0.52	1.33
All taxa	Phosphatase	0.53	2.33
High association	Phosphatase	0.46	2.66
Subgroups	Phosphatase	0.50	2.45
All taxa	Carbon dioxide	0.51	1.52
High association	Carbon dioxide	0.45	1.72
Subgroups	Carbon dioxide	0.47	1.66
Models that were trained upon one set of replicates and tested upon the others			
All taxa	ATP	-0.06	3.05
All taxa	Xylosidase	0.03	2.36
All taxa	Glucosidase	-0.72	5.95
All taxa	Chitinase	0.04	2.36
All taxa	Phosphatase	-0.37	6.12
All taxa	Carbon dioxide	0.28	2.54

3.3 Subgroups of taxa were more important in determining community functions than lone taxa

A third set of random forest regression models that used the differences in the abundances of subgroups to explain differences in functions between communities had a similar ability to explain these differences in function to the original models (as shown in Table 1). This was despite the removal of lone taxa, leaving only the 38% of taxa that were in subgroups to explain differences in function. These models used the abundances of 50 subgroups, which included the abundances of only 100 to 318 taxa compared to the 750 taxa that were used in the initial models. As with taxa in the original models, some subgroups were more important than others in determining community functions, and these subgroups were often important in determining multiple functions. When identifying the top 10 most important subgroups across the functions, 35 important subgroups out of a total possible of 60 were identified. There were 155 subgroups in total, so only 23% of them were considered to be important. While many of these important subgroups contained important taxa, many others did not. Of the important subgroups that were identified, 56% contained important taxa and 44% were composed of taxa that were individually less important.

4 Discussion

Microbial communities contribute a wide arrange of key functions, both in the natural world and in human industry. The structure of a microbial community determines its ability to provide these functions. This structure can be described by its composition of taxa and the interactions between them. In many communities, certain taxa play a greater role in determining function. These keystone taxa may contribute to functions in isolation, without the involvement of other taxa, however they may also contribute to functions via their interactions. They might impact the composition of other functionally-relevant taxa by changing the environment, or they might form subgroups in which different members cooperate towards a shared function. In this study I investigated the role of interactions in the ability of keystone taxa to determine community functions. To do this, I used a large data set containing multiple replicates of bacterial communities from freshwater tree hole pools. This allowed me to reveal novel insights into the relationship between structure and function in these relatively understudied communities, and to gain a uniquely robust inference of interactions. Here, I found that certain taxa were more important in determining community functions than others within these communities, that taxa that were involved in interactions were more important in determining community functions, and that subgroups of taxa were more important in determining community functions than lone taxa.

My findings suggest that keystone taxa drive the relationship between structure and function in freshwater tree hole pool bacterial communities, something that has not widely been established. The random forest regression models that used differences in the abundance of taxa between communities to explain differences in their functions explained over 50% of the variance in their concentrations of xylosidase, glucosidase, chitinase and phosphatase per cell, and in their mass of carbon dioxide per cell. The low explanatory ability of the model for ATP concentration per cell and the low predictability of the models for each function may partly have been due to the inclusion of failing communities that produced low concentrations of ATP. Overall, these results recapitulated the relationship between taxa composition and these functions that were previously observed in these communities (Pascual-García and Bell, 2020). The observation that certain taxa played a much greater role in explaining variance in these functions between communities than others suggested that they were keystone. This provides further evidence for the importance of keystone taxa in microbial community structure and function, as has been shown in systems ranging from the human gut to Arctic ice caps (Trosvik and de Muinck, 2015; Gokul et al., 2016). By using this unique, large data set of replicate communities, this further evidence that I provided is particularly robust.

While many studies have identified keystone taxa, few have specifically compared the roles of isolated contributions, interactions that impact the composition of other taxa, and the formation subgroups in the ability of taxa to determine function. My findings suggest that keystone taxa in these tree hole communities determined function via their interactions, rather than via isolated contributions. The 92% majority of the most important taxa in determining functions were involved in interactions. This was despite the fact that only 43% of the taxa in the entire community were involved in interactions. Moreover, the number of interactions that taxa were involved in positively correlated with their importance in determining the concentrations of ATP, xylosidase, glucosidase and chitinase per cell and the mass of carbon dioxide per cell. The random forest regression models that only included taxa that were involved in the top 5% highest number of interactions were able to explain almost the same amount of variance in the functions as the models that used all taxa, and the median number of interactions that important taxa were involved in was 20 compared to a median of 3 for all of the taxa that were involved in interactions within the community. This aligns with previous studies that have identified keystone taxa as 'hubs' within networks, with

a high mean degree (Berry and Widder, 2014).

The relation between a higher number of interactions and importance in determining community function could suggest that keystone taxa determined function via their impacts upon taxa composition, and thus the functions that were provided by other taxa. Over 80% of all interactions within the network were positive, and 74% of the interactions that taxa were involved in were positive. It is possible that keystone taxa influenced structure and function via a beneficial impact upon the environment. Perhaps the most parsimonious explanation is that one of them released chitinase or xylosidase, directly providing these functions whilst facilitating taxa that subsist upon secondary metabolites. If this were the case, then this would still be an isolated contribution with relation to these functions of interest, even if it may have resulted in other unmeasured functions that were conferred by the facilitated taxa. It is also possible, however that keystone taxa changed the pH such that is was favourable for other taxa that produced chitinase or xylosidase. This would especially be the case if the taxa providing the functions were dependent upon this change in pH, as they would not be able to occur without their interaction with the keystone taxa.

Alternatively, some taxa may have provided functions via subgroup-derived emergent properties. Of the most important taxa, 25% were involved in both the top 5% number of interactions and in subgroups, and the random forest regression models that used subgroups to explain variance in function (to the exclusion of lone taxa) were able to explain almost the same variance as the models that used all of the taxa. This suggests that lone taxa that did not form subgroups were not as important in determining function. Cooperation between taxa within subgroups may have played a greater role in determining function via their emergent properties. This would explain why 71% of the most important taxa were in important subgroups, with 50% of the most important taxa being involved in important subgroups but not the top 5% highest number of interactions. Perhaps by mostly interacting amongst themselves, these taxa exchanged enzymes such as chitinase or xylosidase within extracellular space. This has previously been observed in the gut microbiome, where *Bacteroides ovatus* releases extracellular enzymes to digest polysaccharides. It gains no direct benefit from this, but this facilitates other taxa that then provide reciprocal benefits to *B. ovatus* (Rakoff-Nahoum et al., 2016). Moreover, 44% of the important subgroups did not contain important taxa. Perhaps the taxa in these groups were co-dependent, only providing functions when they were all present. This has previously been observed in the human gut microbiome, where subgroups of interacting bacteria form based upon their vitamin con-dependencies (Molina Ortiz et al., 2022).

The implication that interactions are related to the way in which keystone taxa determine function is relevant to both understanding natural microbial communities and to the use of communities for human industry. Probiotics and whole-community transplants, two ways in which microbial communities can be modified for human benefit, depend entirely upon the relationship between structure and function (Pandey et al., 2015; Sanders et al., 2019; Jiang et al., 2022). My findings imply that keystone taxa, key subgroups of taxa and interactions between taxa will influence the effectiveness of these techniques. My findings into the dynamics of tree hole bacterial communities also have more specific importance. The taxa in tree hole bacterial communities subsist by degrading leaf litter and insect carcasses, and they are therefore of interest for understanding decomposition in forest soils and riparian soils (Pascual-García and Bell, 2020). These processes are also relevant to both carbon and nitrogen cycles (Pascual-García et al., 2023). Understanding the roles of keystone taxa and their interactions in determining these cycles within a variety of microbial ecosystems helps to reveal underlying patterns that may be relevant to future environmental change. This is especially the case because tree hole communities serve as potential indicators of pollution (Ager et al., 2010; Petermann and Gossner, 2022). Tree hole bacterial communities are also tangentially

relevant to human health. Mosquito larvae within tree holes primarily feed upon bacteria, so understanding how interactions determine which taxa are present may be relevant to the spread of disease by these vectors (Xu et al., 2008; Walker et al., 1991; Verdonschot et al., 2008; Petermann and Gossner, 2022).

4.1 Limitations and Future Directions

While this study suggested that certain taxa drive the relationship between structure and function in these tree hole microbial communities, it is unclear whether these taxa are truly "keystone". Keystone taxa have a disproportionate impact upon structure and function relative to their abundance (Banerjee et al., 2018), however all of the taxa that were important in determining function in this study had a relatively high abundance. Both the most abundant and second most abundant ASVs were important in determining function, and the median abundance position for the important taxa was the 29.5th highest position out of 1458. It is possible that these taxa may have had a high raw impact due to their abundance, despite my attempt to control for this by normalising the functions by cell count. It is also possible that higher numbers of interactions were related to higher abundance. This is a problem with inferring interactions from co-occurrence that are inherently based upon abundance. Future work would need to establish the relationship between abundance importance in determining function and interactions more clearly. Other caveats that impacted this study, such as the impact of sampling depth, the different abilities of taxa to be sampled and the classification precision of taxa are also typical of this co-occurrence-based approach (Banerjee et al., 2018).

Perhaps the greatest caveat of this study was that I was unable to determine exact nature of interactions or subgroups. If a keystone taxon was involved in many interactions but not a subgroup, it was assumed to have done so by changing the taxa composition. However, this change in taxa composition may not have changed function, which may instead have originated directly from the keystone taxon. This limitation could be solved by measuring a greater range of functions. That way, if a keystone taxon provided a function that changed the composition of other taxa, the functions that these taxa then provide could be elucidated. This would also help to identify whether subgroups were engaging in cross-feeding, because different taxa within a subgroup could then be linked to different stages of such a pathway. Ultimately, co-culturing experiments and metagenomics to confirm the interactions that I inferred would be the key focus of future studies (Lozupone et al., 2012), and assaying functions that are more closely related to one another would be an important part of this.

A final caveat of this study was that it only investigated the role of bacteria in determining function. Bacteria are some of the most important taxa within microbial communities, and they are abundant. There are more than 10^{30} bacterial cells globally (Prosser et al., 2007). However, there are a range of other microbial taxa within microbial communities, spanning across entire kingdoms of life. Fungi, Archaea and protists, and even non-living viruses all interact within microbial communities, contributing to their functions (Dai et al., 2023; Lee et al., 2022; Caron et al., 2009). In tree hole communities in particular, there are a huge diversity of microbial fungal taxa that are involved in the decomposition of organic matter. These microfungi are likely to be involved in important parts of nutrient cycles within these systems, but very little information exists upon this (Magyar et al., 2017; Gönczöl and Révay, 2003; Petermann and Gossner, 2022). By including cross-kingdom interactions, future investigations might provide a greater picture of the role of interactions in determining microbial community functions.

4.2 Conclusion

In conclusion, this study provided novel insights into the role of interactions in determining the impact of keystone taxa upon microbial community functions. Here, I showed that some taxa were more important in determining community functions than others, that taxa which were involved in interactions were more important in determining community functions that non-interacting taxa, and that subgroups of taxa were more important than lone taxa in determining community functions. These findings provide further evidence for the role of keystone taxa in the relationship between microbial community structure and function, and this evidence is particularly robust due to the use of a unique, large data set of replicated bacterial communities. The importance of interactions in determining microbial community functions that was shown here provides a broad insight into the dynamics of both tree hole communities specifically and Earth's many other crucial microbial communities more generally. It also provides implications for the use of probiotics and whole community transplants, which aim to change the function of microbial communities via changes in structure. Future research could take the findings of my study further via co-culturing for the experimental validation of interactions, and via the inclusion of microbial taxa beyond only bacteria in this analysis.

Data and Code Availability

The code, and data required for this analysis are available from my GitHub research project repository , and the original source of the data that was used in this project can be found from the ReplayEcology GitHub repository, alongside the function from "clean_ASV_table.R".

Acknowledgements

I am sincerely grateful to my supervisors, Prof. Samraat Pawar and Dr. Alberto Pascual Garcia for their guidance, passion for science and thorough patience with me throughout my completion of this project. I am especially thankful for the specialised expertise that Dr. Alberto Pascual-Garcia provided me with in the developing area of microbial community dynamics. His provision of a reliable data set for inferring reactions, and his insights into possible methods of isolating their impacts from those of other factors were of invaluable importance. Furthermore, I am especially thankful to Prof. Samraat Pawar for his broader knowledge upon microbial community dynamics, and his ability to distill complex ideas down to their key elements - something that greatly helped me in my writing. I greatly appreciated his encouragement over the duration of my project.

I would also like to greatly thank friends and family for their support during the completion of my project. Completing our dissertations in tandem up until the last day is an experience that I am sure that I will grow to cherish.

References

Ager, D., Evans, S., Li, H., Lilley, A. K., and Van Der Gast, C. J. (2010). Anthropogenic disturbance affects the structure of bacterial communities. *Environmental Microbiology*, 12(3):670–678.

Banerjee, S., Schlaeppi, K., and van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*, 16(9):567–576.

Bashan, Y. and de Bashan, L. E. (2010). Chapter two - how the plant growth-promoting bacterium azospirillum promotes plant growth—a critical assessment. volume 108 of *Advances in Agronomy*, pages 77–136. Academic Press.

Belenguer, A., Duncan, S. H., Calder, A. G., Holtrop, G., Louis, P., Lobley, G. E., and Flint, H. J. (2006). Two routes of metabolic cross-feeding between bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut. *Appl. Environ. Microbiol.*, 72(5):3593–3599.

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

Breshears, L. M., Edwards, V. L., Ravel, J., and Peterson, M. L. (2015). Lactobacillus crispatus inhibits growth of gardnerella vaginalis and neisseria gonorrhoeae on a porcine vaginal mucosa model. *BMC Microbiology*, 15(1):276.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nat. Methods*, 13(7):581–583.

Caron, D. A., Worden, A. Z., Countway, P. D., Demir, E., and Heidelberg, K. B. (2009). Protists are microbes too: a perspective. *The ISME Journal*, 3(1):4–12.

Culp, E. J. and Goodman, A. L. (2023). Cross-feeding in the gut microbiome: Ecology and mechanisms. *Cell Host Microbe*, 31(4):485–499.

Dai, Q., Ding, J., Cui, X., Zhu, Y., Chen, H., and Zhu, L. (2023). Beyond bacteria: Reconstructing microorganism connections and deciphering the predicted mutualisms in mammalian gut metagenomes. *Ecol. Evol.*, 13(2):e9829.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879):1034–1039.

Geesink, P., Ter Horst, J., and Ettema, T. J. G. (2024). More than the sum of its parts: uncovering emerging effects of microbial interactions in complex communities. *FEMS Microbiol. Ecol.*, 100(4).

Gokul, J. K., Hodson, A. J., Saetnan, E. R., Irvine-Fynn, T. D. L., Westall, P. J., Detheridge, A. P., Takeuchi, N., Bussell, J., Mur, L. A. J., and Edwards, A. (2016). Taxon interactions control the distributions of cryoconite bacteria colonizing a high arctic ice cap. *Mol. Ecol.*, 25(15):3752–3767.

Gönczöl, J. and Révay, (2003). Treehole fungal communities: Aquatic, aero-aquatic and dematiaceous hyphomycetes. *Fungal Diversity*, 12.

Jiang, G., Zhang, Y., Gan, G., Li, W., Wan, W., Jiang, Y., Yang, T., Zhang, Y., Xu, Y., Wang, Y., Shen, Q., Wei, Z., and Dini-Andreote, F. (2022). Exploring rhizo-microbiome transplants as a tool for protective plant-microbiome manipulation. *ISME Communications*, 2(1):10.

Kodera, S. M., Das, P., Gilbert, J. A., and Lutz, H. L. (2022). Conceptual strategies for characterizing interactions in microbial communities. *iScience*, 25(2):103775.

Konopka, A. (2009). What is microbial community ecology? *The ISME Journal*, 3(11):1223–1230.

Korir, H., Mungai, N. W., Thuita, M., Hamba, Y., and Masso, C. (2017). Co-inoculation effect of rhizobia and plant growth promoting rhizobacteria on common bean growth in a low phosphorus soil. *Frontiers in Plant Science*, 8.

Lee, K. K., Kim, H., and Lee, Y.-H. (2022). Cross-kingdom co-occurrence networks in the plant microbiome: Importance and ecological interpretations. *Frontiers in Microbiology*, 13.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lopes, M., Dias-Filho, M., and Gurgel, E. (2021). Successful plant growth-promoting microbes: Inoculation methods and abiotic factors. *Frontiers in Sustainable Food Systems*, 5.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230.

Magyar, D., Vass, M., and Oros, G. (2017). Dendrotelmata (Water-Filled Tree Holes) as Fungal Hotspots - A Long Term Study. *Cryptogamie, Mycologie*, 38(1):55 – 66.

Molina Ortiz, J. P., Read, M. N., McClure, D. D., Holmes, A., Dehghani, F., and Shanahan, E. R. (2022). High throughput genome scale modeling predicts microbial vitamin requirements contribute to gut microbiome community structure. *Gut Microbes*, 14(1):2118831.

Morin, M. A., Morrison, A. J., Harms, M. J., and Dutton, R. J. (2022). Higher-order interactions shape microbial interactions as microbial community complexity increases. *Scientific Reports*, 12(1):22640.

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., Knelman, J. E., Darcy, J. L., Lynch, R. C., Wickey, P., and Ferrenberg, S. (2013). Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.*, 77(3):342–356.

Pace, R. M., Chu, D. M., Prince, A. L., Ma, J., Seferovic, M. D., and Aagaard, K. M. (2021). Complex species and strain ecology of the vaginal microbiome from pregnancy to postpartum and association with preterm birth. *Med (N. Y.)*, 2(9):1027–1049.

Paluszynska, A., Biecek, P., and Jiang, Y. (2020). *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*. R package version 0.10.1.

Pandey, K. R., Naik, S. R., and Vakil, B. V. (2015). Probiotics, prebiotics and synbiotics- a review. *J. Food Sci. Technol.*, 52(12):7577–7587.

Pascual-Garcia, A. (2023). Replayecology. Accessed: 2024-08-21. The function ‘clean$_A$SV$_t$able‘fromthescript‘clean$_A$SV$_t$able.R‘wasused.

Pascual-García, A. and Bell, T. (2020). Community-level signatures of ecological succession in natural bacterial communities. *Nature Communications*, 11(1):2386.

Pascual-García, A., Rivett, D., Jones, M. L., and Bell, T. (2023). Replaying the tape of ecology to domesticate wild microbiota. *bioRxiv*.

Pascual-García, A. and Bell, T. (2020). functionink: An efficient method to detect functional groups in multidimensional networks reveals the hidden structure of ecological communities. *Methods in Ecology and Evolution*, 11.

Petermann, J. S. and Gossner, M. M. (2022). Aquatic islands in the sky: 100 years of research on water-filled tree holes. *Ecol. Evol.*, 12(8):e9206.

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., Green, J. L., Green, L. E., Killham, K., Lennon, J. J., Osborn, A. M., Solan, M., van der Gast, C. J., and Young, J. P. W. (2007). The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, 5(5):384–392.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rakoff-Nahoum, S., Foster, K. R., and Comstock, L. E. (2016). The evolution of cooperation within the gut microbiota. *Nature*, 533(7602):255–259.

Röttjers, L. and Faust, K. (2018). From hairballs to hypotheses-biological insights from microbial networks. *FEMS Microbiol. Rev.*, 42(6):761–780.

Sanders, M. E., Merenstein, D. J., Reid, G., Gibson, G. R., and Rastall, R. A. (2019). Probiotics and prebiotics in intestinal health and disease: from biology to the clinic. *Nature Reviews Gastroenterology & Hepatology*, 16(10):605–616.

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.

Sharma, P., Pandey, A. K., Kim, S.-H., Singh, S. P., Chaturvedi, P., and Varjani, S. (2021). Critical review on microbial community during in-situ bioremediation of heavy metals from industrial wastewater. *Environmental Technology Innovation*, 24:101826.

Sokol, N. W., Slessarev, E., Marschmann, G. L., Nicolas, A., Blazewicz, S. J., Brodie, E. L., Firestone, M. K., Foley, M. M., Hestrin, R., Hungate, B. A., Koch, B. J., Stone, B. W., Sullivan, M. B., Zablocki, O., Trubl, G., McFarlane, K., Stuart, R., Nuccio, E., Weber, P., Jiao, Y., Zavarin, M., Kimbrel, J., Morrison, K., Adhikari, D., Bhattacharaya, A., Nico, P., Tang, J., Didonato, N., Paša-Tolić, L., Greenlon, A., Sieradzki, E. T., Dijkstra, P., Schwartz, E., Sachdeva, R., Banfield, J., Pett-Ridge, J., and Consortium, L. S. M. (2022). Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nature Reviews Microbiology*, 20(7):415–430.

Strickland, M. S., Lauber, C., Fierer, N., and Bradford, M. A. (2009). Testing the functional significance of microbial community composition. *Ecology*, 90(2):441–451.

Tackmann, J., Matias Rodrigues, J. F., and von Mering, C. (2019a). Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Systems*, 9(3):286–296.e8.

Tackmann, J., Rodrigues, J. F. M., and von Mering, C. (2019b). Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Systems*.

Trapet, P., Avoscan, L., Klinguer, A., Pateyron, S., Citerne, S., Chervin, C., Mazurier, S., Lemanceau, P., Wendehenne, D., and Besson-Bard, A. (2016). The pseudomonas fluorescens siderophore pyoverdine weakens arabidopsis thaliana defense in favor of growth in Iron-Deficient conditions. *Plant Physiol.*, 171(1):675–693.

Trosvik, P. and de Muinck, E. J. (2015). Ecology of bacteria in the human gastrointestinal tract–identification of keystone and foundation taxa. *Microbiome*, 3(1):44.

van den Berg, N. I., Machado, D., Santos, S., Rocha, I., Chacón, J., Harcombe, W., Mitri, S., and Patil, K. R. (2022). Ecological modelling approaches for predicting emergent properties in microbial communities. *Nature Ecology & Evolution*, 6(7):855–865.

Venturelli, O. S., Carr, A. C., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., Hromada, S., Northen, T., and Arkin, A. P. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.*, 14(6):e8157.

Verdonschot, R. C. M., Febria, C. M., and Williams, D. D. (2008). Fluxes of dissolved organic carbon, other nutrients and microbial communities in a water-filled treehole ecosystem. *Hydrobiologia*, 596(1):17–30.

Waldrop, M., Balser, T., and Firestone, M. (2000). Linking microbial community composition to function in a tropical soil. *Soil Biology and Biochemistry*, 32(13):1837–1846.

Walker, E. D., Lawson, D. L., Merritt, R. W., Morgan, W. T., and Klug, M. J. (1991). Nutrient dynamics, bacterial populations, and mosquito productivity in tree hole ecosystems and microcosms. *Ecology*, 72(5):1529–1546.

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., Cordero, O. X., Brown, S. P., Momeni, B., Shou, W., Kettle, H., Flint, H. J., Haas, A. F., Laroche, B., Kreft, J.-U., Rainey, P. B., Freilich, S., Schuster, S., Milferstedt, K., van der Meer, J. R., Groβkopf, T., Huisman, J., Free, A., Picioreanu, C., Quince, C., Klapper, I., Labarthe, S., Smets, B. F., Wang, H., Soyer, O. S., and Fellows, I. N. I. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME Journal*, 10(11):2557–2568.

Wolfe, B. and Dutton, R. (2015). Fermented foods as experimentally tractable microbial ecosystems. *Cell*, 161(1):49–55.

Xu, Y., Chen, S., Kaufman, M. G., Maknojia, S., Bagdasarian, M., and Walker, E. D. (2008). Bacterial community structure in tree hole habitats of ochlerotatus triseriatus: influences of larval feeding. *J. Am. Mosq. Control Assoc.*, 24(2):219–227.