bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Community-level signatures of ecological succession in natural bacterial communities

Alberto Pascual-García[1, 2,*] and Thomas Bell[1]

May 12, 2019

(1) Department of Life Sciences. Silwood Park Campus. Imperial College London, Ascot, United Kingdom
(2) Current address: Institute of Integrative Biology. ETH-Zürich, Zürich, Switzerland
(*) Correspondence: alberto.pascual.garcia@gmail.com

Abstract

A central goal in microbial ecology is to simplify the extraordinary biodiversity that inhabits natural environments into ecologically coherent units. We present an integrative top-down analysis of over 700 bacterial communities sampled from water-filled beech tree-holes that combines an analyses of community composition (16S rRNA sequencing) with assays of community functional capacity (exoenzymatic activities, ATP production, CO2 dissipation and yield). The composition of these communities have a strong relationship with the date of sampling and its location, and we identified six distinct community classes. Using structural equation modelling, we explored how functions are interrelated, and how their differences across communities explained the community classes, and found a representative functional signature associated with each community class. We obtained a more mechanistic understanding of the classes using metagenomic predictions. Notably, this approach allowed us to show that these classes contain distinct genetic repertoires reflecting community-level ecological strategies, likely representative of an ecological succession process. These results allowed us to formulate an over-arching ecological hypothesis about how local conditions drive succession in these habitats. The ecological strategies resemble the classical distinction between r and K-strategists, suggesting that a coarse grained picture of ecological succession in complex natural communities may be explained by relatively simple ecological mechanisms.

Introduction

The microbial communities inhabiting natural environments are unmanageably complex. It is therefore difficult to establish causal relationships between community composition, environmental conditions and ecosystem functions (such as rates of biogeochemical cycles) because of the large number of factors influencing these relationships. There is great interest in developing methods that reduce this complexity in order to understand whether there are predictable changes in community composition across space and time, and whether those differences alter microbe-associated ecosystem functioning. The most common approach has been to search for physical (e.g. disturbance) and chemical (e.g. pH) features that correlate with community structure and function. This approach has often been successful in identifying some major major differences among bacterial communities associated with different habitats [1] and some of the edapthic correlates [2]. However, we argue that alternatives are needed because this approach has haunted macroecological studies of larger organisms, often generating empirical relationships that are spurious, difficult to validate, or which could be produced by multiple underlying processes (e.g. neutral vs. niche explanations, [3]). In addition, in microbial communities, measured environmental variables often only explain a small proportion of the variance in community structure, and different studies at different locations often draw dramatically different conclusions of which variables are important. This "context dependence" is expected: community composition results from multiple interacting environmental drivers, so each individual parameter only contributes a small part of the explanation, which varies according to local conditions, and results in a series of case studies that are difficult to generalise. Finally, we lack even basic natural history information for most microbial taxanomic units. This makes it challenging to identify important environmental factors *a priori*,

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resulting in "fishing expeditions" that correlate dozens of measurements with community structure, resulting in many spurious correlations.

An alternative to searching for environmental correlates is to blindly classify communities, and then characterise the environmental properties and functional traits of the community classes. There is a rich array of analytic tools that identify clusters within multivariate datasets, such as the detection of communities in species co-occurrences networks [4] or the reduction of the dimensionality of $\beta-$diversity similarities [5], and that assess whether there are distinct community types. These approaches have been pervasive in the medical microbiome literature, for example in the search for "enterotypes"– i.e. whether individuals are characterised by diagnostic sets of species representing alternative community states [5, 6]. One disadvantage of the classification approach is that the communities may not exhibit distinct clustering (i.e. form a continuum of composition). However, the advantage is that classification simplifies complex datasets, providing direction by then searching for environmental correlates of the community classes. We suggest that this approximation is a necessary first step to explore the natural variability occuring in the communities.

After the communities have been classified, the next step would be to identify functional differences among community classes, focusing particularly on the genetic repertoires of the constituent taxa [7]. Investigating the dominant genes present in the different community classes allows for the determination of ecological strategies, apparent in differences in genes that may be related with environmental sensing, degradation of extracellular substrates, or metabolic preferences. Any difference in functional differentiation among community classes would reflect underlying differences in environmental conditions [8], and so point to specific environmental parameters that could be measured. The identification of these groups would allow us to reduce the complexity of the compositional data by showing which species are likely to have similar impacts on- and responses to environmental conditions. This approach also solves the problem of blindly measuring many environmental parameters in the hope that some will be significantly associated with community structure or ecosystem functioning. Lack of any clear functional differentiation among community classes is also informative, and would indicate alternative community states with redundant functions [9, 10, 11]. Such redundancy could arise in the absence of environmental variability, which could also help explain any lack of a dominant environmental axis that explains variation in composition.

The final step would be to obtain independent validation of whether the genetic correlations result in functional differences. Microbes inhabitating a host sometimes have a substantial impact on host performance, for example turning a "healthy" into a "diseased" host [12]. Such extreme impacts of individual taxa make it relatively simple to infer a direct link between community composition and function. In open, natural environments (e.g. soil, lakes, oceans), it is often difficult to envisage which functions to assess, the impact of individual taxa on those functions is often minor, and generalisations may depend on subjective choices and assumptions. An important step forward comes from manipulative experiments in natural environments, that have identified variables such as pH [13], salinity [14], sources of energy [15], the number of species [16], and environmental complexity [17] to be key players in the relationship between bacterial community structure and functioning. Despite the increasing number of controlled experiments, interpreting how changes to community composition result in functions to complex communities remains a topical question [18], because a link between the generality aimed at by large-scale compositional analysis and the precision pursued by controlled experiments is often missing. Combining community classification with metagenomic analysis can result in specific hypotheses about which functions would be expected to change in ecosystems dominated by each of the community classes, thereby allowing a more mechanistic understanding of the underlying processes.

To investigate these questions, in this work we used a large dataset consisting of more than 700 samples of rainwater filled puddles that can form at the base of beech [18]. This is an ideal system to investigate these questions given the relatively similar conditions across different locations is unique in terms of replicability of a natural aquatic environment [19, 20]. Indeed, while effects caused by environmental variation on phytotelmata ecosystems has been investigated in meio- and macrofaunal communities [20], the influence in microbial communities is largely unknown. Moreover, emphasis was made for understanding bottom-up drivers of tree-holes diversity like nutrients [21, 22, 23], but top-down approaches that may help us to understand other drivers of microbial composition like stochastic dispersion or interactions have received comparatevely less attention [20]. The large dataset allowed us to study natural variation in bacterial community composition through the top-down classification of communities into classes. In deciduous forests, bacterial life on the forest floor is characterised by seasonal and daily changes to temperature and resource availability. We therefore address whether differences in the communities are due to the historical processes at the different geographic locations, or if they are rather more influenced by contingent local conditions, a dichotomy difficult to resolve [11]. In addition, while tree-hole

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

macro-faunal communities have well known successional changes during leaf litter degradation, the micro-faunal communities are poorly understood, and we may expect high variability in short periods of time, as observed in compost ecosystems [24].

The bacterial communities present in the tree-holes are key players in the decomposition of detritus of leaves, and therefore of great interests more broadly for understanding decomposition in forest soils and riparian zones. To link the compositional analysis with bacterial functioning, we analysed a set of community-level functional profiles obtained from laboratory assays of the same communities [18] to investigate whether the communities differed in their functional capacity. Instead of focusing on the performance of any specific function, we looked for the relationships between functional variables, and whether their interplay differed across the community classes. Finally, we used metagenomic information to understand whether similar compositions and functions were translated into different classes of genetic repertoires, susceptible of a functional interpretation. Interestingly, we found signatures of community-level trade-offs between classes ressembling those found for single-cells, which we believe reflect an ecological succession driven by local environmental dynamics of the tree-holes.

Results

Microbial communities classes are determined by local conditions

We analysed 753 bacterial communities sampled from water-filled beech tree-holes in the South West of UK [18]. We classified the communities according to two different metrics, the Jensen-Shannon divergence (D_{JSD}) [25], and a transformation of the SparCC metric (D_{SparCC}, see Methods [26]), both of which revealed six distinct community classes (see Fig. 1). The whole set of communities are dominated by *Proteobacteria*, and the community classes were distinguished by being dominated by OTUs identified with the genus *Klebsiella* (class 1, red; class 3, pink), *Paenibacillus* (class 2, green), *Serratia* (class 5 blue), and *Pseudomonas,* (classes 4, yellow; 6 grey). *Rhizobium* were also characterisitic of class 1, *Escherichia/Shigella* were found mostly in class 5, and low abundant OTUs like *Acinetobacter* and *Herbaspirillum* were most apparent in class 6.

We investigated how the different classes were distributed in space using Principal Coordinate Analysis [27] (see Methods) and projecting the communities into the first three coordinates, together with the centroid of the locations where they were sampled (see Fig.1 and Suppl. Fig. 3). The first coordinate mostly separates *Klebsiella* classes (red and pink) from the others. The second coordinate clearly separates communities blue and grey, which have the highest and lowest presence of *Serratia*, respectively, while the third coordinate (see Suppl. Fig. 3), influences the green community only, likely representing the effect of *Paenibacillus*. Notably, the classes relate with the locations where they were collected, suggesting that strong environmental effects are in place. We used class 1 (red) as the reference class because it encompassed the largest number of communities, so we took this to be the archetypal community state. Later analysis confirmed the convenience of this choice since it exhibits intermediate functional and genetic features among those found for the different classes.

To quantify the effect of spatial autocorrelation, we hierarchically clustered together samples that were closer in space across 10 distance thresholds spanning 5 orders of magnitude (from 5 metres to 200km). We computed the ANOSIM statistics [28] for each of the 10 resultant classifications, and we found a strong distance-decay relationship that relates community similarity with the spatial distance threshold used (see Fig. 2). To identify possible biases due to the different number of clusters at each threshold, we perfomed two randomizations (see Methods). The corrected model the corrected model shows a consistent exponential distance decay in the similarity of the communities, consistent with the observed data. Nevertheless, it suggests that the observed data underestimate the communities similarity at distances of 50km, possibly due to the uneven sampling of the communities over space, with large gaps in distance between locations like Birchcleave Woods and the main area sampled around Silwood Park and Oxfordshire (see Suppl. Table 1 and Suppl. Fig. 1). The ANOSIM values would also overestimate the similarity of the communities for spatial distances lower than 50m, which could have been inflated due to the method used.

Notably, the corrected model supports a gain in the ANOSIM statistics between 50km and 10km. This gain matches the difference in the ANOSIM statistics computed clustering the tree-holes according with the sampling location (Site values in Fig. 2) or according to the sampling date (Day and Month values). Therefore, the specific date (Day) turned out to be more informative than the site and, notably, it was higher than the value obtained clustering samples collected within the same month, suggesting that seasonal environmental conditions were not the main drivers of the similarities, but rather conditions happening in a daily-basis. Strikingly, the value we obtain for the ANOSIM statistics when the classification considered are the community classes found, reaches

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the same value than the one found at 50m. Since this classification in classes considers only six clusters, and the members may be much further than 50m in space (see Suppl. Fig. 1), this result suggests that the classes capture intrinsic local conditions that are analogous even at distant locations.

Community classes reflect different functional performances

To explore whether compositional differences in the communities are translated into different community functional capacity, we analysed data quantifying the function performance of these communities [18]. The communities were isolated and cryo-conserved after sampling, and later revived in a medium made of beech leaves as substrate. Cells were grown for seven days while monitoring CO2 dissipation and, after this period, cells were counted, and there were also measured ATP stocks and the capacity of the communities to secrete four ecologically relevant exoenzymes [29] related with i) uptake of carbon: xylosidase (X) and $\beta-$glucosidase (G); ii) carbon and nitrogen: $\beta-$chitinase (N); and iii) phosphate: phosphatase (P) .

Functional capacities of the communities were clearly distinct in these measurements, see Suppl. Fig. 5. Therefore, we wondered if compositional differences among the community classes may result in different functional signatures. To tackle this question, we first addressed how the functional variables were entangled using structural equation models (SEM) [30]. We obtained a global model with an excellent fit to the data (RMSEA$<10^{-3}$, CI=[0–0.023], AIC=7493 see Methods). We hypothesized that a model with different parameters for each community class (i.e. up to six parameters per pathway) would better explain the data, considering penalizations for the larger number of parameters. After little respecification of the model's pathways (see Suppl. Methods) we determined a better model (RMSEA$<10^{-3}$, CI=[0–0.035], AIC=6658). Some pathways were required to be constrained to the same value across classes, while others depended on the specific class considered (see Fig. 3 and Suppl. Table 6). That different coefficients for the classes in some pathways brought a better fit than the model with a single parameter for all classes, supported the hypothesis that the classes had differentiated function performances.

The model showed that measurements related with uptake of nutrients were all exogenous, being ATP production, cells' yield and CO2 endogenous (see Fig. 3). In addition, ATP production influenced yield, which in turn influenced CO2. Among exoenzyme variables, N influenced ATP and, notably, only X affected yield, while G and P influenced both ATP and CO2. Analysis and clustering of the standarized partial regression coefficients (see Suppl. Fig. 24) established a classification of the classes in three groups according with their function, with classes 1 and 2 being the most similar, then classes 4 and 6, and 3 and 5. Therefore, the analysis shows the possibility of functionally classifying classes by identifying their most representative pathways. Nevertheless, to sharply identify these pathways and given the complexity of the SEM model found, we first ruled out the possibility that differences in pathways' coefficients across classes were due to the influence of other (confounding) variables. To control this possibility we found, for each pair of endogenous-exogenous variables, its set of confounding variables with dagitty [31]. We next performed, for each pair of variables involved in a pathway, a linear regression including its adjustment set of confounding factors, and an interaction term with a factor codifying for the different classes. Coefficients should be interpreted as deviations with respect to class 1, which was taken as reference (see Methods). Notably, the significant interaction terms obtained led to a classification of the classes through distinctive relationships, shown in Fig. 3.

The analysis revealed that yield was negatively influenced by $\beta-$chitinase activity for class 2 and for ATP production for class 5, while being positive related with $\beta-$glucosidase for classses 3 and 6. CO2 is negatively influenced by $\beta-$glucosidase and positively by xylosidase by class 4, and also positively by phosphatase activity for class 6. Finally, ATP is affected negatively by xylosidase activity only for class 5.

Community classes depict different genetic repertoires

To get a more mechanistic understanding of the above results, we analysed the communities' genetic repertoire performing metagenomic predictions with PiCRUST [32], and further statistical analysis with STAMP [33]. The fraction of exo-enzymatic genes with respect to the total pool of genes in communities belonging to classes 2, 4 and 6 was significantly larger than the fractions found for classes 1, 3 and 5, suggesting that the former classes are specialized in degrading a wider array of substrates (see Fig. 4). To simplify the complexity of the large pool of genes, we clustered the KO annotations into KEGG's pathways (see Methods). The analysis showed that the 6 community classes differed in their genetic repertoires. Furthermore, these divergent genetic repertoires suggested different ecological adaptations, schematically summarized in Fig. 5. Consistent with PCA analysis of KEGG's pathways (see Suppl. Figs. 10, 11 and 12), we divided the classes in two large groups: classes 1, 3 and 5 carried the genetic machinery compatible with fast growth, while classes 2, 4 and 6 carried the genetic

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

machinery for autonomous amino-acid biosynthesis. Evidence for fast growth in classes 1, 3 and 5 comes from the large fraction of genes related with genetic information processing (see Suppl. Fig. 15), mostly related with DNA replication (DNA replication proteins genes, transcription factors, mismatch repair or homologous recombination genes or ribosome biogenesis), which are known to be a good genetic predictor of fast growth [34]. Secondly, communities from classes 1, 3, and 5 also carry a larger fraction of genes related with intake of readily available extracelullar compounds (see Suppl. Fig. 16), including ABC transporters, phosphotransferase system, or peptidases, and environmental adaptation like bacterial motility proteins, synthesis of siderophores or the two-component systems. Rapid replication requires a more accurate control of protein folding and trafficking. Consistent with this hypothesis, we found a significantly inflated fraction of genes involved in folding stability, sorting and degradation, including chaperones, or the phosphorelay system (see Suppl. Fig. 17).

A second series of evidences pointing towards orthogonal ecological strategies came from differences in metabolic pathways genes. *Serratia*-dominated class (5) had an inflated fraction of genes related to carbohydrate degradation, including genes involved in the glycolysis and the trycaborxylic acid (TCA) cycle (Suppl. Fig. 18). In contrast, the *Pseudomonas* and *Paenibacillus* classes (2, 4, 6) were associated with genes involved in alternative pathways like nitrogen or methane metabolism, and in secondary metabolic pathways related with degradation of xenobiotics or chlorophyl metabolism. Notably, the genes codifying for the exo-enzymes assayed were higher for these communities, suggesting that were well adapted to environments with more recalcitrant nutrients (see Fig. 5) In addition, classes 2, 4, and 6 had a remarkable repertoire of genes for amino acids biosynthesis –possibly at odds of classes 1, 3 and 5 that would invest in proteases for amino acid uptake for the environment (see Suppl. Fig. 19 and 20). Indeed, the apparently low glycolitic capabilities of these communities may have as a consequence pyruvate depravation, needed to generate acetil-CoA and oxaloacetate to activate the TCA cycle. Consistent with this observation, we observed that these communities exhibited a significantly larger proportion of genes related with glyoxylate metabolism and degradation of benzoate, that may be used as substitutes to glycolysis (see Suppl. Fig. 21). Finally, we observed that *Serratia* (5) and *Klebsiella* (1,3) communities had a significantly lager repertoire of genes to synthezise amino acids for valine, leucine and isoleucine, which requires pyruvate that, according with our interpretation, they would generate through glycolysis (see Suppl. Fig. 18). Orthogonal to this observation, *Pseudomonas* and *Paenibacillus* classes (2, 4 and 6) had a significantly larger proportion of genes degrading these amino acids (see Suppl. Fig. 19) and hence, either they take these essential amino acids from the environment or they generate the from other pathways. Consistently, we observed that they were enriched of genes related with the glycine, serine and threonine metabolism (see Suppl. Fig. 19), through which it is possible to obtain valine, leucine and isoleucine which is, in turn, another alternative to generate acetil-CoA (see Fig. 5).

Discussion

In this article, we analyzed a large set of microbial communities isolated from tree-holes, whose function was further investigated under laboratory conditions. We found a striking distance-decay relationship that may be explained by different means. We observed that it is possible to arrange the communities into classes, and that the classes matched both the sites and the date of collection, which are tightly entangled. There was a geographic influence particularly apparent for the larger distances, which suggests the influence of broad environmental conditions and perhaps historical processes [11]. Nevertheless, the date of collection best explained the communities classes, as reflected by the ANOSIM statistics. The finding is consistent with the idea that environmental conditions on a particular day strongly influenced species composition, consistent with previous findings on macroinvertebrate tree-holes communities [35]. Moreover, that communities from the same class were found in different seasons suggests that factors like temperature are of secondary importance, despite results highlighting their importance in similar systems [10]. In addition, given that commmunities collected within the same day were more similar if they were closer in space, with values well above to those found when only the day of collection was considered, suggests the existence of significant autocorrelation, previously reported in soil sampling [36, 13]. Our results show that autocorrelation extends to scales above the short distances (<10m) previously reported [36] in a non-trivial way. The fact that the ANOSIM statistics for the classes is much higher than the one for date of collection, and the uneven distribution of the classes members in time and space (see Suppl. Figs. 4 and 1) suggests that analogous local conditions occur at distant locations, that we interpret below in terms of ecological succession.

The analysis of functions performed in laboratory experiments confirmed that these classes separated the communities into ecologically meaningful subgroups. First of all, we observe a strong negative relationship for class 2 between chitinase activity (N) and yield. Since chitin was not expected to be in a high amount in the growth media (made from beech leaves), investing in this exoenzyme in excess should lead to a lower yield with

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respect to the reference class number 1, due to protein allocation trade-offs [37]. A similar reasoning may be applied for classes 3 and 5 which had a negative relationship between X and ATP. For class 3, this may be due to the fact that investing in G leads to a more efficient pathway to grow through aerobic production of ATP. This is deduced from a significantly higher activity between G and Cells than class 1, combined with positive relationships in the SEM diagram between G and ATP and Yield and CO2 (which do not appear significant in this analysis possibly because are also high in class 1). However, it would not be the case for class 5 (who do not present this pattern), and hence we envisage the use of a glycolytic pathway to grow faster with lower ATP production. This is consistent with the fact that the relation between ATP and cells is significantly more negative than for class 1. A similar interpretation may be given for class 4, for which a lower relation between G investment and CO2 we interpret as a signature of more glycolytic activity than for class 1. Therefore, for both class 4 and 5, an increase in G suggests a preference in ATP production through glycolisis. A last notable observation is a positive relation between P and CO2 for class 6. As observed in the SEM model, this class had the lowest relationship between P and ATP, in turn leading to a low yield. Therefore, those communities from class 6 investing more in P are likely those deprived the most of phosphorous in the experiment compared with their natural habitat. Since these communities also have high levels of CO2 (see Suppl. Fig. 5), we are inclined to think that it probably reflects lack of adaptation to the growth media and cell's death accounts for an anomalous CO2 increase.

These observations are compatible with a scenario of ecological succession in which there was a transition from communities dominated by r-strategists to K-strategists [38]. We suggest that early successional stages were characterised by the *Serratia* class (5, blue). This class had a negative relationship between ATP and yield and its respiration values were not influenced by the number of cells. In addition, investing in xylosidase had a much lower transfer into ATP production than for the reference class 1, possibly pointing towards a preference for immediately available nutrients like sugar monomers. These observations pointed towards anaerobic activity, and the analysis of the metagenome consistently revealed that this class dominated pathways necessary to extracellularly degrade and uptake nutrients, metabolic processes such as glycolysis, and it had a wide variety of genes related with environmental processing, fast replication and accurate molecular control of protein folding and trafficking. The mean Shannon diversity of communities belonging to this class was almost the lowest (see Suppl. Table 3) which would be expected in a rich environment dominated by few well adapted fast growers, in line with the notion of r-strategists.

The next communities in the succession were the two *Klebsiella* classes (1 and 3, red and pink, respectively). Although still sharing some of the features of the *Serratia* class 5, they already depicted some distinctive features such as a higher transfer of ATP into yield and perhaps respiratory activity, although more apparent for classes 2, 4 and 6. Therefore, we interpreted these classes as belonging to intermediate successional stages. Later successional were characterised by the *Pseudomonas* classes, exhibiting high respiration values. These classes presented an inflated fraction of genes related with oxidative phosphorylation and synthesis of most amino acids, and a number of secondary metabolic pathways that may be valuable in environments in which resources are low but there are remaining by-products from former inhabitants. This is particularly apparent for class number 6 (grey) that dominated pathways orthogonal to those dominated by class number 5. Consistently, this class had the higher Shannon diversity, containing a larger number of rare OTUs and some notable abundances of strict aerobes (see below). Therefore, this class would be representative of communities dominated by K-strategists.

A last class to consider is the *Paenibacillus* class. While still close to *Pseudomonas* classes in many metagenomic features, it was strongly dominated by the *Paenibacillus* genus, and it was the class with lowest Shannon diversity. A possible explanation of this pattern comes from the fact that this class had a large fraction of sporulation and germination genes (see Suppl. Fig. 22) that suggests that these communities lived in an environment in which the amount of resources was particularly low. The laboratory resuls are consistent with this hypothesis: this community had the largest transfer of chitinase activity into yield, which may reflect its ability to take advantage of the remaining nutrients such as dead arthropod exoskeletons or fungi. Furthermore, it has been noted that low amount of water is possibly the main driver of fungi sporulation [39]. This observation would perfectly match our interpretation, locating this class at the last stage of the succession, where nutrients have been depleted to low levels.

There are several environmental conditions potentially driving this succession, that we aim to link with nutrients dynamics. A main source of carbon comes from leaf litter, which promotes the abundance, diversity and number of trophic levels of meio- and macrofaunal communities [40, 41]. Its degradation would be compatible with the succession described, because the dynamics of the different nutrients it brings to the tree-hole depends on how difficult it is to degrade them. Simple sugars decay 80% in two weeks, while starch decays at 50% and cellulose is hardly depleted [42]. Since simple sources of carbon like glucose are the main nutrients stimulating community

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

growth in one-week assays in lab conditions [43], we envision that succession occurs in periods of two weeks until simple sugars are exhausted. Nevertheless, if the quantity of leaf litter is the main driver of the succession, we would expect a strong seasonal signal, with a class dominating in autumm where the largest amount of leafs are collected. Our data do not provide evidence to support these observations, because the classification in classes provides much higher ANOSIM statistics than the one considering the months as classifiers, and members of the same class are often distant in space and time.

Other important source of variability comes from the rain and how it is collected into the tree-hole, depending on whether it's mainly through stemflow or throughfall. Significant effects for nitrogen-related ions and pH were reported with respect to the date of sampling, with significant statistical interaction terms with respect to both the content of leafs and the type of water collector, and hence presumibly with carbon sources [22]. These effects were particularly important for nitrate, ammonia and sulfate [21, 23]. Indeed, leafless holes have larger amounts of nitrate and nitrite, suggesting that nitrogen cycle is likely microbially mediated. Moreover, stemflow contains high amounts of nitrate and sulfate, with flushing events after heavy rain in tree-holes affected by stemflow bring phosphate to a minimum [22]. A progressive acidification in tree-holes which do not receive water inputs for long periods is also expected through nitrification, in particular for those filled by throughfall water [21, 23]. In addition, an increase in labile phosphate in the form of orthophosphate is expected at later successional stages [23]. Therefore, the rain pulses may explain the similarity of some samples collected at the same date (see Suppl. Fig. 1) and the properties of the tree-holes in terms of size (litter capability) and collection of water (stemflow vs. throughfall) would bring the variability needed to explain the lack of complete synchronization for all dates, or the similarities between distant locations.

These considerations lead us to envisage an ecological succession scenario in which rain events were the primary drivers of bacterial composition, illustrated in bottom-left corner of Fig. 5, followed by the tree-holes features. Rain would generate pulsed resources of different type and frequency [44], and tree-holes features would determine the rate of resource attenuation [45]. For instance, large tree-holes, those exposed through throughfall water or those with large leaf contents, would present slower attenuation. This would explain that in some specific dates all the tree-holes reflect the same communities (recent rain or long standing drough conditions, see Suppl. Fig. 1) while, beyond that, the classes are distributed across different dates and sites (due to the differential attenuation caused by diverse tree-hole features).

The community classes may be linked to their oxygen requirements as well. Although dissolved oxygen is lower in the deeper water column, it may be high when the tree-hole has low levels of water, and only intersticial water remains. This lead us to conjecture that this may be the environment to which the *Pseudomonas* (6) communities were adapted, with the following observations supporting adaptation to aerobic respiration and high levels of phosphate. We observed an increase in abundances of strict aerobes, including *Acinetobacter*, *Paucimonas* and *Phenylobacterium*. There was also an increase in genes related with metabolism of nitrate, methane, degradation of benzoate (likely associated with the presence of resines), or chlorophyl (which indicates an increase in photoheterotrophs). This class might also be able to run the TCA cycle generating acetyl-CoA from acetate, and from the degradation of valine, leucine and isoleucine, further complemented with glyoxylate metabolism and the degradation of benzoate to generate oxaloacetate. Finally, the class was found in summer and winter, and clustered in specific areas. This would discard the temperature as an important variable or the light (which in the UK have large differences between summer and winter) and also the tree-holes size (given the high packing in space of some samples it is unlikely that many tree-holes in the same area have the same size) thereby pointing towards the amount of water and oxygen as key variables. This observation would also hold for the *Paenibacillus* class for which particularly drought periods would lead to lack of water independently of the tree-holes features, and we observed locations in which all the tree-holes belonged to this class (see Suppl. Fig. 4).

We cannot rule out other site-based conditions like the type of forest management. A study analysing this factor, however, did not found substantial differences in enzymatic activities despite of reporting compositionally different communities [46], perhaps because the low number of samples did not bring sufficient resolution. Another possible local influence for the composition are trophic ecological interactions, like the prevalence of invertebrates in certain areas such as the frequently-studied presence of mosquito larvae, with known effects on taxa like *Sphingomonadaceae* and *Flavobacteria* [47]. Or the relationship between high xylosidase activity and sporulation genes of the *Paenibacillus* class, perhaps being a defensive mechanism against fungi. There are other processes like passive dispersal due to the movement of insects or animals between tree-holes that may make the tree-holes composition to be more similar [48], ressembling the existence of metacommunities [35]. The interplay between stochastic and deterministic processes would be needed to investigate these questions.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The large collection of bacterial communities collected here provides a detailed insight into the community ecology of this environment. Notably, by identifying community classes a priori, we were able to piece together the natural history of this environment from the perspective of the bacterial communities. The picture presented here is one in which the metagenomes of sets of complex communities under different environmental conditions reflect the metabolic specialisations of the dominant members. In this way, we were able to identify classes resembling the classical r- vs. K-strategists classification [38]. Despite of criticisms of this classification due to its oversimplification, the main arguments against it hold for macroscopic systems, raising questions such as the importance of age-structured populations [49]. However, we still find suggestive this conceptual framework for microbes [50], since this ecological dichotomy may well be supported by energetic [51] and protein-allocation trade-offs [37]], which possibly underlie differentiations such as olitgotroph vs. copiotroph strategies [52]. Therefore, our results would support approximations aiming to predict community-level functions such as the metabolome, from imperfect community-level information like metagenome predictions [53, 54]. We believe these results provide great promise for the capacity of top-down approaches for reducing the complexity of microbial community datasets, and help developing a bottom-up synthetic ecology that can be predictive in the wild.

Methods

Dataset

We analyzed 753 bacterial communities characterized by sequencing 16S rRNA amplicon libraries from Ref. [18]. These communities were sampled from rainwater-filled beech tree-holes (Fagus spp.) collected from different locations in the South West of United Kingdom, see Suppl. Table 1. 95% of the samples were collected between 28 of August and 03 of December 2013, being the remaining 5% collected in April 2014 . Spatial distances between samples span five orders of magnitude (from <1m to >100km, see Suppl. Fig. 1). We considered only samples with more than 10K reads, and were removed species with less than 100 reads across all samples. This leaded to a final dataset comprising 680 samples and 2874 Operative Taxonomic Units (OTUs) at the 97% of 16 rRNA sequence similarity. In previous work [18], four replicates of each of these communities were regrown in standard laboratory conditions using a tea of beech leaves as a substrate, and eventually further supplemented with low quantities of 4 substrates labelled with 4-methylumbelliferon. The experiments quantified the capacity of the communities to degrade xylosidase (abreviated X in the text, cleaves the labile substrate xylose, a monomer prevalent in hemicellulose), of β-chitinase (N, breaks down chitin, which is the major component of arthropod exoskeletons and fungal cell walls), β-glucosidase (G, break down cellulose, the structural component of plants), and phosphatase (P, breaks down organic monoesters for the mineralisation and acquisition of phosphorus). Full experimental details can be found in [18].

Determination of classes

We computed all-against-all communities dissimilarities with Jensen-Shannon divergence [25], D_{JSD}, and a transformation of the SparCC metric [26], D_{SparCC} (see Suppl. Methods), and then clustered the samples following a similar prospect to the one proposed in Ref. [5] to identify enterotypes. In the text, we call these clusters community classes. The method consists of a Partition Around Medoids (PAM) clustering for both metrics, with the function PAM implemented in the R package CLUSTER [55]. This clustering requires as input the number of output clusters desired k. We performed the clustering considering a wide range of k values and computing alongside the Calinski-Harabasz index (CH) that quantifies the quality of the classification obtained, and selecting as optimal classification $k_{\mathrm{opt}} = \arg\max_k(CH)$. We further performed an alternative dimensionality reduction through Principal Coordinate Analysis (R function DUDI.PCO, package ADE4), and projected the samples into the first three coordinates, coloured by the classes found in the previous analysis. The classes found with both measures were significantly similar and visual inspection of the PCO projections are qualitatively equivalent (see Suppl. Methods and Results). Processing of data and taxa summaries provided as Supplementary Material were generated with Qiime [56].

Relation between community similarity and the sampling date and location

To investigate the relationship between the sampling location, the sampling date and the similarity in composition of bacterial communities, we performed analysis of similarities tests (ANOSIM) of the communities [28], using both D_{JSD} and D_{SparCC}.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We considered as grouping units one automatic spatial classification and two temporal classifications in which samples are joined in clusters depending on whether they were collected in the same day, or in the same month. Details for the spatial automatic classification and results for two other definitions of sampling sites (one given by the field-researchers and another by the names popularly given to the sites) are found in Suppl. Methods and Results. For each of these classifications, an ANOSIM test was run with the R function ANOSIM of package VEGAN, with 10^4 permutations. To get further insight into potential spatial autocorrelation, we clustered the communities considering different spatial distance thresholds every half order of magnitude, from 5 m. to 100 km, and computed the ANOSIM test for each of the resultant classifications R^2_{obs}. We then performed two kinds of randomizations (see Suppl. Methods) intended to explore i) if the increasing number of clusters for decreasing distance thresholds biases the ANOSIM statistics and ii) to detect distances where the community similarity significantly increases.

The first randomization completely shuffled the GPS coordinates of the samples. In this way, when the spatial clustering was performed, communities coming from very different locations were joined into the same cluster. This procedure, labelled as "unconstrained" in Fig. 2, completely lost any signal, and discarded trivial inflating bias in the ANOSIM statistics coming from the splitting of the samples into different number of clusters. Then we perfomed a second randomization procedure (R^2_{rnd}, labelled as "constrained" in Fig. 2) which tooks the classifications obtained from the original data at a given distance threshold (reference clasification), and shuffled the GPS coordinates of the samples constraining the shuffling within the clusters obtained in the classification at a threshold immediately higher (parent classification). The value of the parent classification should be the expected value, R^2_{exp}, in the absence of biases. Therefore the value $\delta R^2 = R^2_{\text{rnd}} - R^2_{\text{exp}}$ brings an estimation of possible biases due to the partitioning of communities into samples, and a corrected estimator is given by $R^2_{\text{corr}} = R^2_{\text{obs}} - \delta R^2$.

Structural equation modelling

Structural equation models [30] were built and analysed with LAVAAN (version 0.523) and visualized with SEMPLOT R package [57, 58]. The modelling procedure was split into different stages detailed in Suppl. Methods. First, a global model considering all data was investigated following several theoretical assumptions about the relationship between the functions, until a final model was achieved. Then, we looked for a second series of models in which it was possible to fit a different coefficient for each of the parameters in the global model, constraining the data into subsets corresponding to the community classes found (therefore, six possible coefficients for each SEM pathway). Minor reespecification of the model was performed (see Suppl. Methods and Results). We investigated if models considering more or less parameters provided better fits, penalizing by the number of degrees of freedom. The main criteria to accept a change was to require that the AIC of the modified model to be smaller than the original one [59], while additionally verifying that a number of estimators were consistently improved after the modification, such as the RMSEA, the Comparative Fit Index or the Tucker-Lewis Index [60, 61].

To investigate causal relationships between endogenous and exogenous variables within the final specified model required the control for confounding factors. We identified, for each pathway involving a regression in the SEM model, its adjustment set with dagitty [31]. Then we performed a linear regression of each pathway adjusted by the confounding factors, and adding a factor codifying for the different classes. The coefficients obtained from the regression were estimated with respect to class number 1, which was the class with the largest set of samples, and it appeared in an outgroup cluster with class 2 (see Suppl. Fig. 24). Later inspection of the metagenomics information showed that the genes of this class are evenly distributed among the KEGG pathways analysed, hence suggesting that its communities are representative of the whole dataset. We identified significant interaction terms between classes and the exogenous variable under exam in the pathway. A significant interaction coefficient involving a given class was safely interpreted as a different performance of that class with respect to class number 1, and we identify it as a distinctive functional feature of the class.

Metagenomic analysis

Metagenomics predictions were performed using PiCRUST v1.1.2 [32]. A subset of genes appearing at intermediate frequencies was selected (see Suppl. Methods) and aggregated into KEGG pathways [62]. The mean proportion of genes assigned to a specific pathway was computed across communities belonging to the same class. Then we tested if the differences in mean proportions between classes were statistically significant with post-hoc tests with STAMP [63] (see Suppl. Methods). To create Fig. 5 we visually inspected each post-hoc test and ranked the classes according with the number of pairwise tests in which they appeared significantly inflated. We qualitatively

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

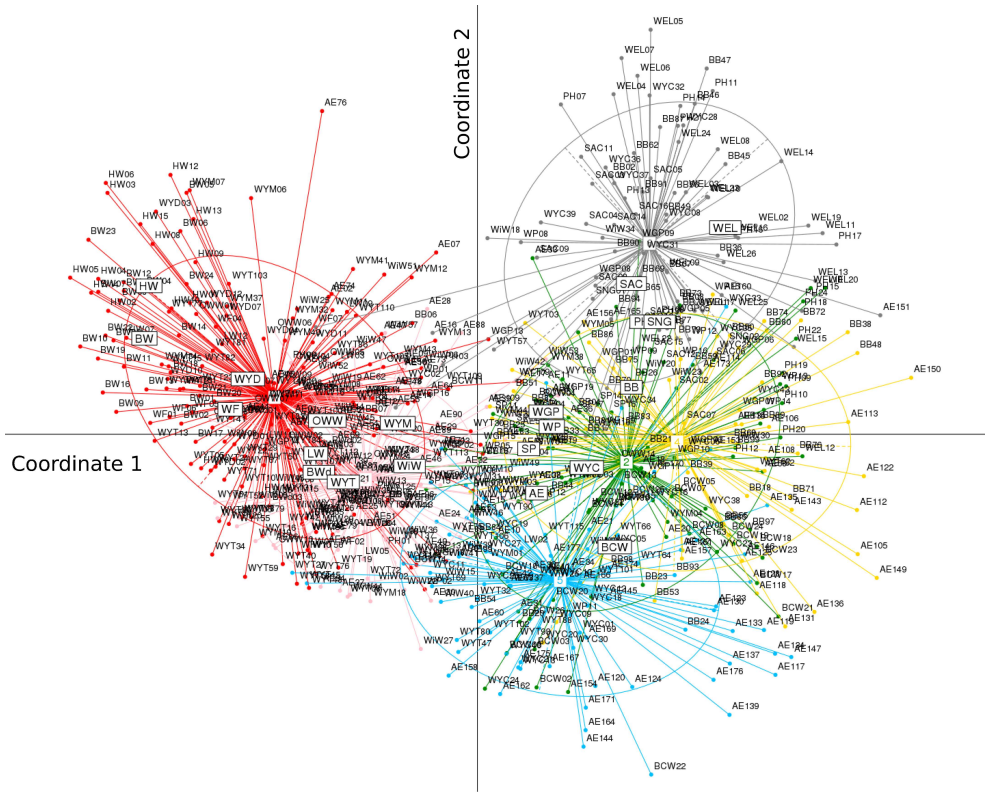
represent this ranking with circles of different sizes. Classes that do not appear inflated in any pairwise test in the pathway are not represented.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

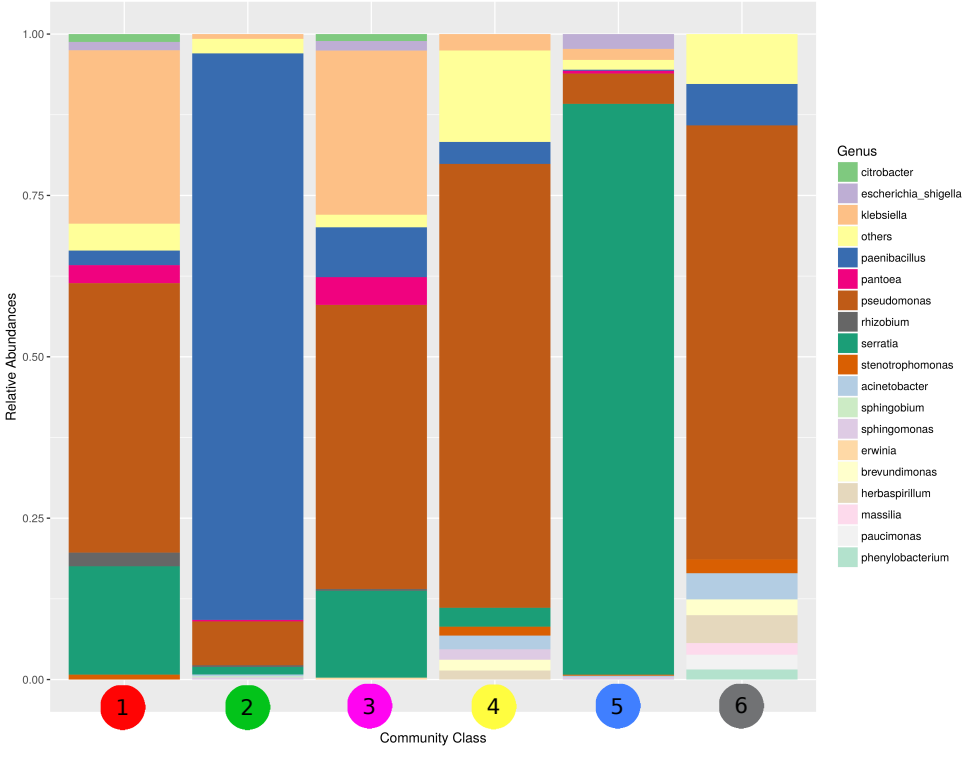
Acknowledgements

We acknowledge Damian Rivett for explanations about the experimental methods, and to Matt Jones, Lara Durán-Trío and Yonathan Friedman for helpful discussions. We thank Andreas Steingötter from the ETH Seminar in Statistics for the support discussing the statistical methods. The research was funded by a European Research Council starting grant (311399-Redundancy) awarded to T.B. T.B. was also funded by a Royal Society University Research Fellowship. APG was also funded by the Simons Collaboration: Principles of Microbial Ecosystems (PriME), award number 542381.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



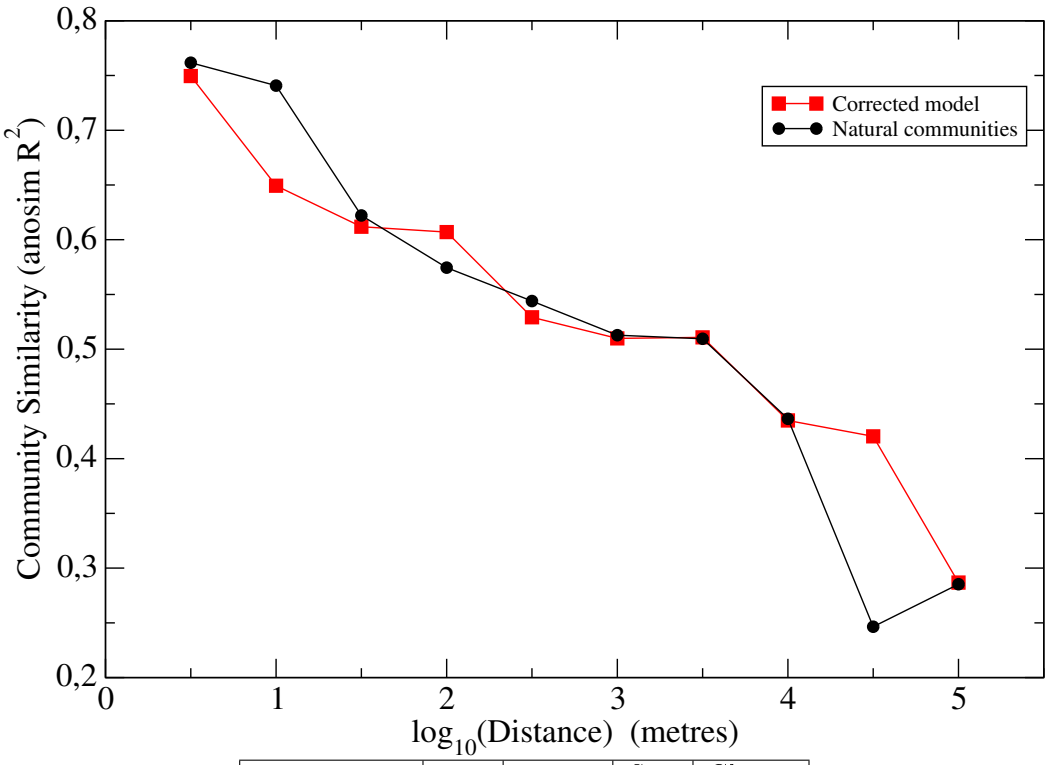
A.



B.

Figure 1: **Communities classes description and relation with sampling sites.** (A) Principal Coordinate Analysis of communities with D_{SparCC}. The centroids of the sampling sites are indicated with their label within rectangles. (B) Proportion of genus into the six community classes. Genus representing less than 1% are classified as Others. Colours to identify the classes are in correspondence in both figures, and are consistently used alongthrough the article.

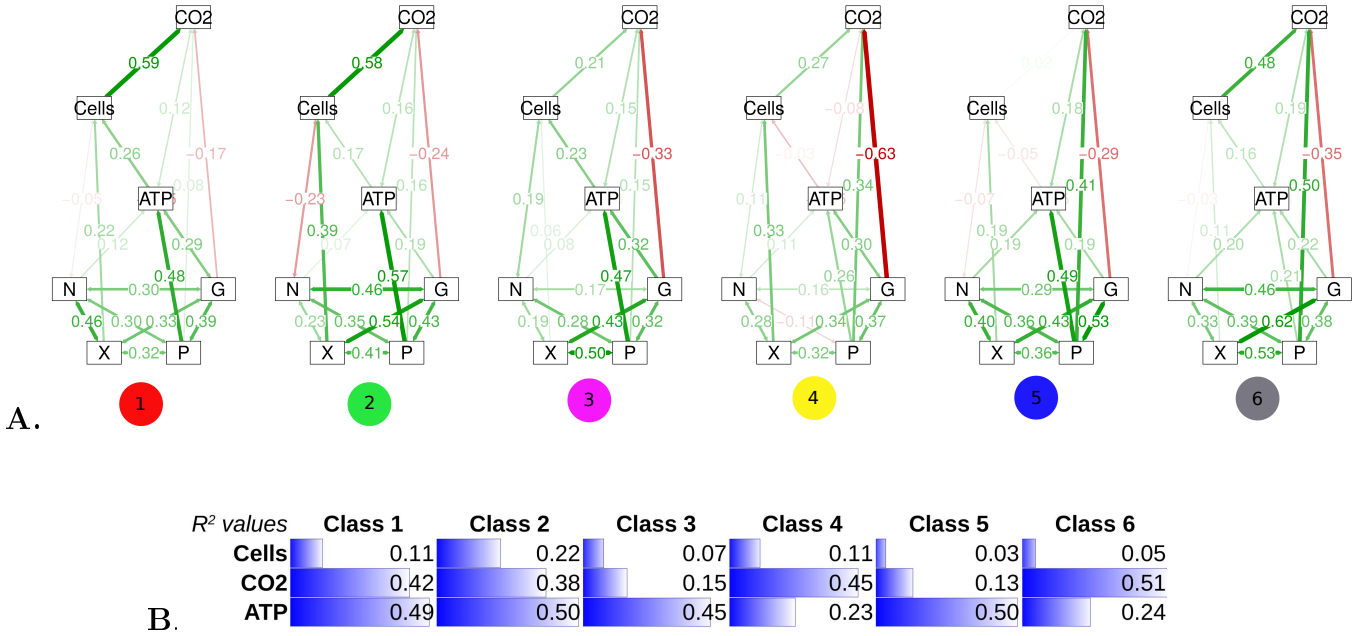
bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



	Day	Month	Site	Classes
ANOSIM R^2	0.42	0.34	0.26	0.60

Figure 2: **Distance-decay of communities similarity.** (Top) ANOSIM statistics of communities clustered at different spatial distances exhibit a distance-decay exponential trend in their similarity measured with D_{SparCC} (Natural communities). Each of these classifications was randomized with two different procedures (see Methods and Suppl. Fig. 2) to generate a corrected model. The corrected model suggests that the data underestimates the similarity of the communities at 50km and overestimate it at distances lower than 50m, being apparent an otherwise consistent exponenial trend. (Bottom) As a reference, the ANOSIM statistics obtained when samples are clustered according to the date, month and site where they were sampled, and also according to the community class, are also shown in the Table.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



C.

Measurement (units)	Abbr.
CO2 dissipation (ug)	CO2
Cells Count (CPM)	Cells
ATP production (uM)	ATP
phosphatase (uM)	P
β-chitinase (uM)	N
xylosidase (uM)	X
$\beta-$glucosidase (uM)	G

D.

Formula	Var	Class	Var * Class
$Cells \sim (N + X + ATP) * Class$	N	(2)	-1.29 (*)
$CO2 \sim G * Class$	G	(4)	-0.19 (**)
$CO2 \sim (X + G + P) * Class$	X		0.86 (*)
$Cells \sim ATP * Class$	ATP	(5)	-0.57 (*)
$ATP \sim (X + N + G + P) * Class$	X		-1.38 (**)
		(3)	-0.71 (*)
$Cells \sim (G + X + ATP) * Class$	G		0.31 (*)
		(6)	0.44 (*)
$CO2 \sim P * Class$	P		0.18 (*)

Figure 3: **Communities classes functional performance.** (A) Diagrams of the final SEM. The standarized coefficients for each of the six classes are shown on the correspondent pathway. (B) R^2 values for the three endogenous variables in the SEM model. (C) Experimental measurements used in this study and their abbreviation. (D) Causal analysis of the influence of the exogenous variable (Var) for each SEM pathway (determined by the response variable and Var in Formula) when the identity of the classes is included as a factor. Confounding factors involved in the pathway are also included (remainder variables in Formula). The column Class reflects the class identity of any significant coefficient found with its value shown in the column Var * Class (significance code: * = p-val<0.01, ** = pval<0.001).

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

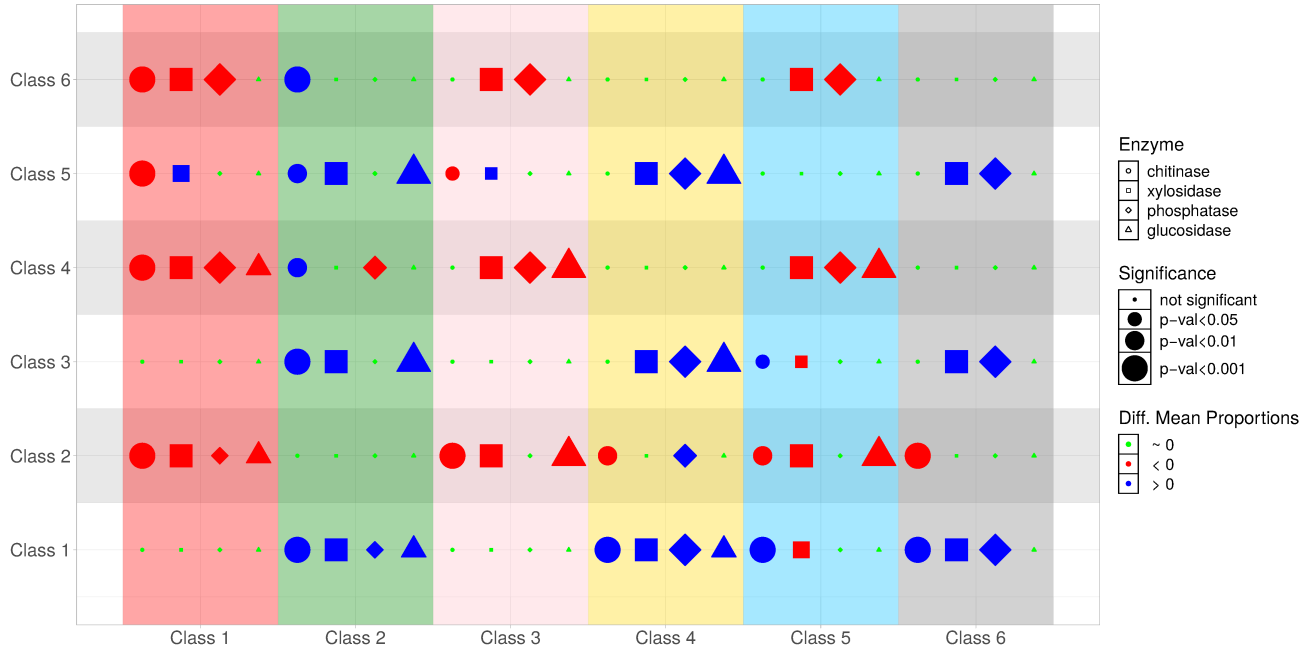


Figure 4: **Summary of post-hoc analysis of exo-enzymatic genes predictions.** The significance of differences in mean proportions of chitinase (circles), β-xylosidase (squares), β-glucosidase (triangles) and phosphatase (diamonds) were tested across all-against-all pairwise combinations of community classes. Red (blue) symbols denote that the class in the correspondent column has a significantly lower (higher) mean proportion of the enzyme, with respect to the class shown in the row. Size of the symbol is larger for more significant differences. A higher proportion of these enzymes for classes 2 (green), 4 (yellow) and 6 (grey) is apparent. All tests are provided in Suppl. Fig. 4.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

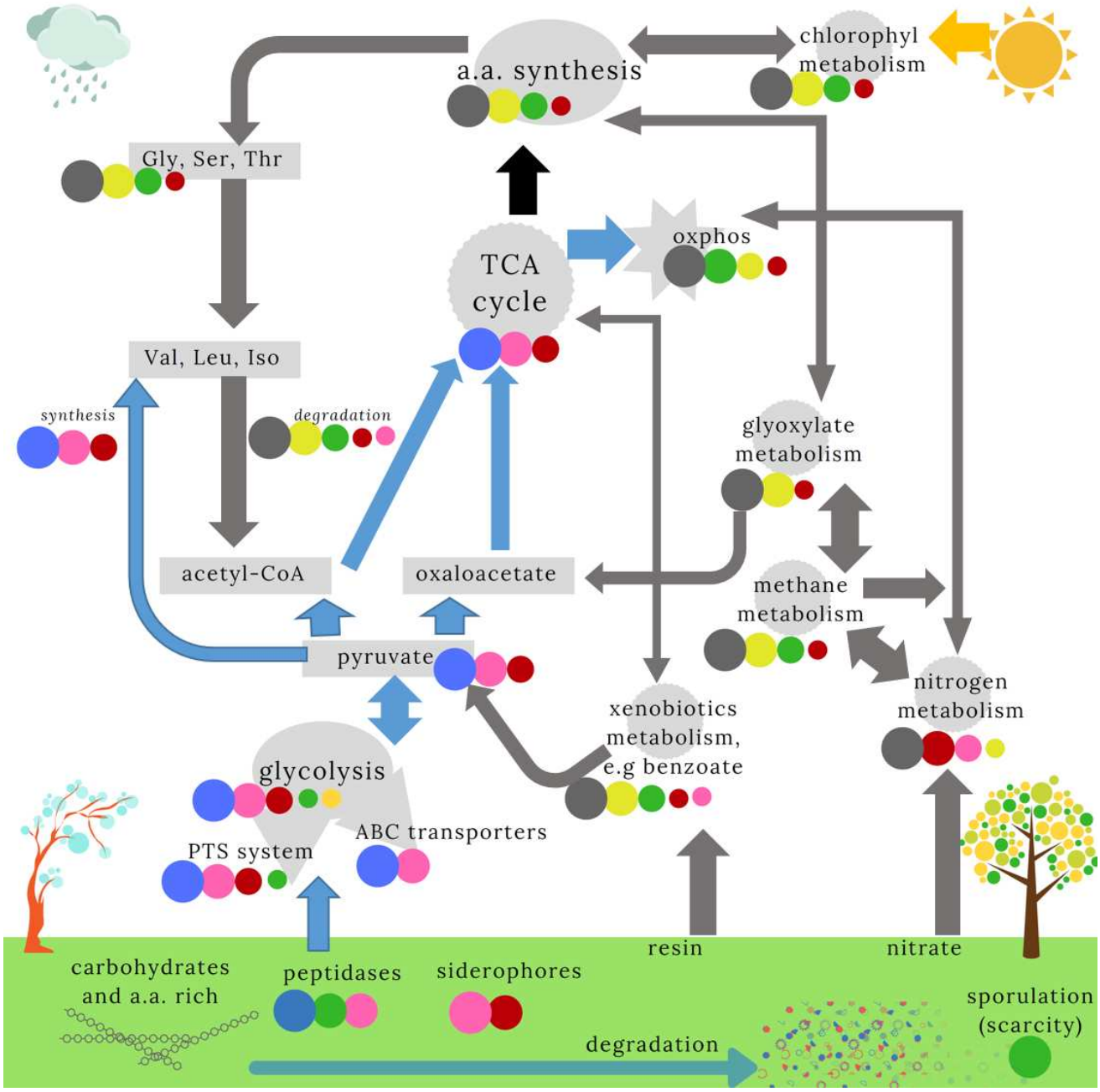


Figure 5: **Scheme describing the genetic repertoires of the different classes.** Pathways from the KEGG's database found to be most relevant to describe the different classes are shown. The mean proportion of genes of each pathway between the classes is compared (see Suppl. Results) and a rank generated, which is indicated by circles with bigger sizes meaning a larger proportion, as explained in Fig. 4. Only classes getting a significant pairwise post-hoc comparison are shown. It is apparent that the grey (*Paenibacillus*) and blue (*Serratia*) classes dominate orthogonal pathways. Accordingly, we plotted some important relations between these pathways with the colour of the dominating community, except for the link between TCA cycle and amino acid synthesis, which remains unclear. We further illustrate the substrates and hypothetical environmental conditions expected for *Serratia* (after rain) and *Pseudomonas* (drough and sunny) communities. We expect the other communities to be intermediate states between these two classes, being *Klebsiella* communities (red and pink) closer to *Serratia* class, and the yellow (*Pseudomonas*) and green (*Paenibacillus*) closer to the grey one. The *Paenibacillus* community, with a large number of sporulation and germination genes, may reflect conditions of very low nutrient availability.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] L. R. Thompson, J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, *et al.*, "A communal catalogue reveals Earth's multiscale microbial diversity," *Nature*, vol. 551, no. 7681, p. 457, 2017.

[2] R. I. Griffiths, B. C. Thomson, P. James, T. Bell, M. Bailey, and A. S. Whiteley, "The bacterial biogeography of british soils," *Environmental microbiology*, vol. 13, no. 6, pp. 1642–1654, 2011.

[3] G. Bell, "Neutral macroecology," *Science*, vol. 293, no. 5539, pp. 2413–2418, 2001.

[4] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Microbial co-occurrence relationships in the human microbiome," *PLoS computational biology*, vol. 8, no. 7, p. e1002606, 2012.

[5] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, *et al.*, "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, 2011.

[6] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, and R. E. Ley, "A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets," *PLOS Computational Biology*, vol. 9, p. e1002863, Jan. 2013.

[7] J. E. Koenig, A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, and R. E. Ley, "Succession of microbial consortia in the developing infant gut microbiome," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4578–4585, 2011.

[8] S. Louca, L. W. Parfrey, and M. Doebeli, "Decoupling function and taxonomy in the global ocean microbiome," *Science*, vol. 353, no. 6305, pp. 1272–1277, 2016.

[9] J. M. Bhatnagar, K. G. Peay, and K. K. Treseder, "Litter chemistry influences decomposition through activity of specific microbial functional guilds," *Ecological Monographs*, vol. 88, pp. 429–444, Aug. 2018.

[10] X. Zhang, E. R. Johnston, A. Barberán, Y. Ren, X. Lü, and X. Han, "Decreased plant productivity resulting from plant group removal experiment constrains soil microbial functional diversity," *Global Change Biology*, vol. 23, pp. 4318–4332, Oct. 2017.

[11] Y. Shi, P. Grogan, H. Sun, J. Xiong, Y. Yang, J. Zhou, and H. Chu, "Multi-scale variability analysis reveals the importance of spatial distance in shaping Arctic soil microbial functional communities," *Soil Biology and Biochemistry*, vol. 86, pp. 126–134, July 2015.

[12] V. K. Ridaura, J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkayeva, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. V. Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath, and J. I. Gordon, "Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice," *Science*, vol. 341, p. 1241214, June 2013.

[13] D. Bru, A. Ramette, N. P. A. Saby, S. Dequiedt, L. Ranjard, C. Jolivet, D. Arrouays, and L. Philippot, "Determinants of the distribution of nitrogen-cycling microbial communities at the landscape scale," *The ISME Journal*, vol. 5, pp. 532–542, Mar. 2011.

[14] M. Berga, Y. Zha, A. J. Székely, and S. Langenheder, "Functional and compositional stability of bacterial metacommunities in response to salinity changes," *Frontiers in microbiology*, vol. 8, p. 948, 2017.

[15] B. Jenkins, R. Kitching, and S. Pimm, "Productivity, disturbance and food web structure at a local spatial scale in experimental container habitats," *Oikos*, pp. 249–255, 1992.

[16] D. Gravel, T. Bell, C. Barbera, T. Bouvier, T. Pommier, P. Venail, and N. Mouquet, "Experimental niche evolution alters the strength of the diversity–productivity relationship," *Nature*, vol. 469, no. 7328, p. 89, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[17] S. Langenheder, M. T. Bulling, M. Solan, and J. I. Prosser, "Bacterial biodiversity-ecosystem functioning relations are modified by environmental complexity," *PloS one*, vol. 5, no. 5, p. e10834, 2010.

[18] D. W. Rivett and T. Bell, "Abundance determines the functional role of bacterial phylotypes in complex communities," *Nature microbiology*, p. 1, 2018.

[19] K. A. R. Nishadh, A. DAS, and K. Sakthidas, "Tree-hole aquatic habitats: inhabitants, processes and experiments. A review.," *International Journal of Conservation Science*, vol. 5, no. 2, 2014.

[20] R. Kitching, "Food webs in phytotelmata: bottom-up and top-down explanations for community structure," *Annual Review of Entomology*, vol. 46, no. 1, pp. 729–760, 2001.

[21] S. R. Carpenter, "Stemflow chemistry: effects on population dynamics of detritivorous mosquitoes in tree-hole ecosystems," *Oecologia*, vol. 53, no. 1, pp. 1–6, 1982.

[22] E. D. Walker, D. L. Lawson, R. W. Merritt, W. T. Morgan, and M. J. Klug, "Nutrient dynamics, bacterial populations, and mosquito productivity in tree hole ecosystems and microcosms," *Ecology*, vol. 72, pp. 1529–1546, Oct. 1991.

[23] R. C. Verdonschot, C. M. Febria, and D. D. Williams, "Fluxes of dissolved organic carbon, other nutrients and microbial communities in a water-filled treehole ecosystem," *Hydrobiologia*, vol. 596, no. 1, pp. 17–30, 2008.

[24] K. Nakasaki, K. Nag, and S. Karita, "Microbial succession associated with organic matter decomposition during thermophilic composting of organic waste," *Waste Management & Research*, vol. 23, pp. 48–56, Feb. 2005.

[25] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[26] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS computational biology*, vol. 8, no. 9, p. e1002687, 2012.

[27] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[28] K. R. Clarke, "Non-parametric multivariate analyses of changes in community structure," *Austral Ecology*, vol. 18, no. 1, pp. 117–143, 1993.

[29] R. L. Sinsabaugh, B. H. Hill, and J. J. F. Shah, "Ecoenzymatic stoichiometry of microbial organic nutrient acquisition in soil and sediment," *Nature*, vol. 462, no. 7274, p. 795, 2009.

[30] J. B. Grace, *Structural equation modeling and natural systems*. Cambridge University Press, 2006.

[31] J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison, "Robust causal inference using directed acyclic graphs: The R package dagitty," *International journal of epidemiology*, vol. 45, no. 6, pp. 1887–1894, 2016.

[32] M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, *et al.*, "Predictive functional profiling of microbial communities using 16s rRNA marker gene sequences," *Nature biotechnology*, vol. 31, no. 9, pp. 814–821, 2013.

[33] D. H. Parks, G. W. Tyson, P. Hugenholtz, and R. G. Beiko, "Stamp: statistical analysis of taxonomic and functional profiles," *Bioinformatics*, vol. 30, no. 21, pp. 3123–3124, 2014.

[34] B. R. Roller, S. F. Stoddard, and T. M. Schmidt, "Exploiting rRNA operon copy number to investigate bacterial reproductive strategies," *Nature microbiology*, vol. 1, p. 16160, 2016.

[35] C. J. Paradise, J. D. Blue, J. Q. Burkhart, J. Goldberg, L. Harshaw, K. D. Hawkins, B. Kegan, T. Krentz, L. Smith, and S. Villalpando, "Local and regional factors influence the structure of treehole metacommunities," *BMC ecology*, vol. 8, no. 1, p. 22, 2008.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[36] R. B. Franklin and A. L. Mills, "Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field," *FEMS Microbiology Ecology*, vol. 44, pp. 335–346, June 2003.

[37] D. W. Erickson, S. J. Schink, V. Patsalo, J. R. Williamson, U. Gerland, and T. Hwa, "A global resource allocation strategy governs growth transition kinetics of escherichia coli," *Nature*, vol. 551, no. 7678, p. 119, 2017.

[38] M. Gadgil and O. T. Solbrig, "The concept of r-and k-selection: evidence from wild flowers and some theoretical considerations," *The American Naturalist*, vol. 106, no. 947, pp. 14–31, 1972.

[39] J. Gönczöl and Á. Révay, "Treehole fungal communities: aquatic, aero-aquatic and dematiaceous hyphomycetes," *Fungal Diversity*, vol. 12, pp. 19–34, 2003.

[40] C. Ptatscheck and W. Traunspurger, "Meio-and macrofaunal communities in artificial water-filled tree holes: Effects of seasonality, physical and chemical parameters, and availability of food resources," *PloS one*, vol. 10, no. 8, p. e0133447, 2015.

[41] C. Ptatscheck and W. Traunspurger, "The meiofauna of artificial water-filled tree holes: colonization and bottom-up effects," *Aquatic ecology*, vol. 48, no. 3, pp. 285–295, 2014.

[42] M. K. Aneja, S. Sharma, F. Fleischmann, S. Stich, W. Heller, G. Bahnweg, J. C. Munch, and M. Schloter, "Microbial colonization of beech and spruce litter–influence of decomposition site and plant litter species on the diversity of microbial community," *Microbial Ecology*, vol. 52, pp. 127–135, July 2006.

[43] M. Kaufman, W. Goodfriend, A. Kohler-Garrigan, E. Walker, and M. Klug, "Soluble nutrient effects on microbial communities and mosquito production in Ochlerotatus triseriatus habitats," *Aquatic Microbial Ecology*, vol. 29, pp. 73–88, 2002.

[44] D. A. Yee and S. A. Juliano, "Concurrent effects of resource pulse amount, type, and frequency on community and population properties of consumers in detritus-based systems," *Oecologia*, vol. 169, no. 2, pp. 511–522, 2012.

[45] D. W. Rivett, T. Scheuerl, C. T. Culbert, S. B. Mombrikotb, E. Johnstone, T. G. Barraclough, and T. Bell, "Resource-dependent attenuation of species interactions during bacterial succession," *The ISME journal*, vol. 10, no. 9, p. 2259, 2016.

[46] W. Purahong, D. Kapturska, M. J. Pecyna, E. Schulz, M. Schloter, F. Buscot, M. Hofrichter, and D. Krüger, "Influence of different forest system management practices on leaf litter decomposition rates, nutrient dynamics and the activity of ligninolytic enzymes: A case study from Central European forests," *PLoS ONE*, vol. 9, p. e93700, Apr. 2014.

[47] Y. Xu, S. Chen, M. G. Kaufman, S. Maknojia, M. Bagdasarian, and E. D. Walker, "Bacterial community structure in tree hole habitats of ochlerotatus triseriatus: Influences of larval feeding," *Journal of the American Mosquito Control Association*, vol. 24, pp. 219–227, June 2008.

[48] B. Maguire Jr, "The passive dispersal of small aquatic organisms and their colonization of isolated bodies of water," *Ecological Monographs*, vol. 33, no. 2, pp. 161–185, 1963.

[49] D. Reznick, M. J. Bryant, and F. Bashey, "r-and k-selection revisited: the role of population regulation in life-history evolution," *Ecology*, vol. 83, no. 6, pp. 1509–1520, 2002.

[50] J. H. Andrews and R. F. Harris, "r-and k-selection and microbial ecology," in *Advances in microbial ecology*, pp. 99–147, Springer, 1986.

[51] T. Pfeiffer, S. Schuster, and S. Bonhoeffer, "Cooperation and competition in the evolution of ATP-producing pathways," *Science*, vol. 292, pp. 504–507, Apr. 2001.

[52] F. M. Lauro, D. McDougald, T. Thomas, T. J. Williams, S. Egan, S. Rice, M. Z. DeMaere, L. Ting, H. Ertan, J. Johnson, *et al.*, "The genomic basis of trophic strategy in marine bacteria," *Proceedings of the National Academy of Sciences*, vol. 106, no. 37, pp. 15527–15533, 2009.

bioRxiv preprint doi: https://doi.org/10.1101/636233; this version posted May 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[53] C. Noecker, A. Eng, S. Srinivasan, C. M. Theriot, V. B. Young, J. K. Jansson, D. N. Fredricks, and E. Borenstein, "Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation," *MSystems*, vol. 1, no. 1, pp. e00013–15, 2016.

[54] P. E. Larsen, F. R. Collart, D. Field, F. Meyer, K. P. Keegan, C. S. Henry, J. McGrath, J. Quinn, and J. A. Gilbert, "Predicted relative metabolomic turnover (prmt): determining metabolic turnover from a coastal marine metagenomic dataset," *Microbial informatics and experimentation*, vol. 1, no. 1, p. 4, 2011.

[55] A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith, "Clustering rules: a comparison of partitioning and hierarchical clustering algorithms," *Journal of Mathematical Modelling and Algorithms*, vol. 5, no. 4, pp. 475–504, 2006.

[56] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[57] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.

[58] S. Epskamp, "semplot: Unified visualizations of structural equation models," *Structural Equation Modeling*, vol. 22, no. 3, pp. 474–483, 2015.

[59] L.-C. Lin, P.-H. Huang, and L.-J. Weng, "Selecting path models in sem: A comparison of model selection criteria," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 24, no. 6, pp. 855–869, 2017.

[60] P. Barrett, "Structural equation modelling: Adjudging model fit," *Personality and Individual differences*, vol. 42, no. 5, pp. 815–824, 2007.

[61] J. C. Loehlin, *Latent variable models: An introduction to factor, path, and structural analysis*. Lawrence Erlbaum Associates Publishers, 1998.

[62] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[63] J. R. White, N. Nagarajan, and M. Pop, "Statistical methods for detecting differentially abundant features in clinical metagenomic samples," *PLoS computational biology*, vol. 5, no. 4, p. e1000352, 2009.