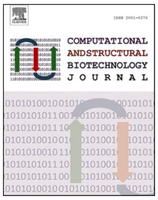
Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj



Research article

Leveraging phylogenetic signal to unravel microbiome function and assembly rules

Silvia Talavera-Marcos, Marcos Parras-Moltó, Daniel Aguirre de Cárcer *

Departamento de Biología, Universidad Autónoma de Madrid, Madrid, Spain

A R T I C L E I N F O

Keywords:
Community assembly
Microbiome
Phylogenetic clustering
16S rRNA gene
Metabolic modeling
Microbial ecology

A B S T R A C T

Clarifying the general rules behind microbial community assembly will foster the development of microbiome-based technological solutions. Here, we study microbial community assembly through a computational analysis of phylogenetic core groups (PCGs): discrete portions of the bacterial phylogeny with high prevalence in the ecosystem under study. We first show that the existence of PCGs was a predominant feature of the varied set of microbial ecosystems studied. Then, we re-analyzed an in vitro experimental dataset using a PCG-based approach, drawing only from its community composition data and from publicly available genomic databases. Using mainly genome scale metabolic models and population dynamics modeling, we obtained ecological insights on metabolic niche structure and population dynamics comparable to those gained after canonical experimentation. Thus, leveraging phylogenetic signal to help unravel microbiome function and assembly rules offers a potential avenue to gain further insight on Earth's microbial ecosystems.

1. Introduction

Microbes represent a large fraction of the Earth's biomass and most of its biodiversity. They also drive global biogeochemical cycles and significantly impact the fitness of most multicellular organisms with whom they develop symbiotic relationships. Microorganisms in nature normally appear as communities, or groups of potentially interacting populations that co-exist in space and time [1]. The rules that govern the assembly of these populations, which are each formed of genetically homogeneous individuals, are still poorly understood [2]. Greater knowledge of microbial community assembly will not only increase our understanding of the role that microbiomes play in sustaining life on Earth but will also foster the development of microbiome-based technological solutions. In this regard, the bottom-up design of functional synthetic consortia would benefit from an appropriate understanding of microbial community assembly, as would top-down strategies if eco-evolutionary forces are to be controlled to produce functional consortia.

Despite their evident complexity, microbial communities often present shared characteristics: they are highly diverse (comprising various bacterial phyla), species rich (large number of species), feature coexisting populations that should theoretically exclude one another given their genomic characteristics, can show remarkable functional stability despite large species turnover in the community, and trait-based selection can significantly impact their assembly [3]. Moreover, in most microbial communities, bacteria tend to co-occur with phylogenetically related populations more often than expected by chance, a phenomenon termed phylogenetic clustering (i.e. microbial communities often bear a phylogenetic signal) [4–6]. These common patterns of microbial communities lead to the idea that a set of common principles governs microbial community assembly [7]. The prevailing view in the field is that microbiomes assemble on the basis of function, without a significant role for phylogenetic assembly [8]. This idea is supported by the predominant observation that different species compositions can translate into functionally equivalent microbial ecosystems [8]. However, the lack of a significant role for phylogenetic assembly is undermined by the extensive phylogenetic signal observed in microbial communities.

In an effort to better understand microbial community assembly, we recently investigated this community characteristic [3] by taking into consideration that traits and ecological function are, to some extent, conserved from an evolutionary standpoint [9,10]. Considering these points and phylogeny as a proxy for evolutionary history, we proposed a conceptual framework for the phylogenetically constrained assembly of microbial communities [3]. The framework is linked to Vellend's synthesis of community ecology [1] and its four basic assembly principles (drift, dispersal, selection, and diversification), and extends it to account

* Correspondence to: Departamento de Biología, Universidad Autónoma de Madrid, Madrid 28049 Spain.
E-mail address: daniel.aguirre@uam.es (D. Aguirre de Cárcer).

https://doi.org/10.1016/j.csbj.2023.10.039
Received 5 June 2023; Received in revised form 17 October 2023; Accepted 17 October 2023
Available online 18 October 2023

for ecosystem patchiness, sampling bias, and phylogeny-related selection. The framework is centered around two facts: first, phylogenetic clustering in a microbial ecosystem can be studied in terms of phylogenetic core groups (PCGs), discrete portions (i.e. specific nodes) of the bacterial phylogeny that are present in all instances of a given ecosystem type; and second, the 16S rRNA gene-based phylogeny of bacteria presents significant functional coherence [10]. However, the strength of this coherence varies with phylogenetic depth along the phylogenetic tree [10]; thus, deep branching nodes may not maintain functional coherence. So far, PCGs have been clearly detected in the rice rhizosphere [3] and human gut [11] environments. The framework contends that the most plausible explanation for the existence of a PCG in a given environment is that populations belonging to that PCG present a phylogenetically conserved set of traits that improve the fitness of those populations under the particular biotic and abiotic factors in that environment. The framework also proposes that populations belonging to the same PCG would be ecologically cohesive (to the extent that they are affected by the same selective forces), and hence, its intra-group structure would be governed mainly by immigration and drift (neutral processes). Significantly, as PCGs can be easily detected in replicated ecosystem samples by sequencing the 16S rRNA gene, it was proposed that the analysis of PCG phylogeny and genomic databases could elucidate the shared niche characteristics of PCGs, potentially offering a rapid approach that can be used to characterize microbial ecosystem functioning and identify the role that resident populations play in it.

However, it is still unclear whether PCGs are a predominant feature of microbial ecosystems or a rare phenomenon. Also, predicted intra-PCG characteristics could prove invalid, hence limiting the utility of the framework. To gauge the practical usefulness of the framework, we evaluate the existence of PCGs in a wide array of diverse microbial ecosystems, including various human and plant-associated environments, as well as some animal-associated and environmental microbial communities. Then, we assess the predicted PCG characteristics relating to local community assembly. Taking into account, with noteworthy exceptions (e.g. see [12,13]), that microbial ecosystem samples comprise different microenvironments and patches [14], we then re-analyze the community assembly data of the simple artificial communities published by Goldford et al. [7]. We show that leveraging phylogenetic signal has great potential to illuminate the selective pressures experienced by microbes in natural environments, providing a systematic computational strategy to identify functional groups without requiring exhaustive experimentation [15].

2. Methods

2.1. PCGs in the environment

PCGs have previously been detected as 16S rRNA gene sequence clusters (Operational Taxonomic Units; OTUs) of varying depth (i.e. clustering threshold) present in all instances of a given microbial ecosystem, which represents a reasonable proxy [3,11]. However, sequence clustering lacks true transitivity, which, with differential initial seeding between clustering runs, may translate into slightly different clusters for the same input dataset generated by different runs or clustering algorithms. Thus, in addition to 16S rRNA gene sequence clusters, we analyzed PCGs on the basis of nodes of a phylogenetic tree detected in all instances of the ecosystem type, an approach that provides increased phylogenetic resolution. We analyzed the PCGs in nine datasets from the literature that present a comparatively high number of ecosystem replicates and sequencing depth (Suppl. Table 1). The human microbiome was represented by the following datasets: *FlemishGut* (fecal) [16], *TwinsUK* (fecal) [17], *Illeum* (mucosa) [18], *Rectum* (mucosa) [18], and *Vagina* (mucosa) [19]. Plant-associated environments were represented by *Rice* (root samples) [20] and *Leaf* [21]; animal microbiomes, by *Sponge* (*Carteriospongia foliascens*) [22] and *Mice* [23]; and environmental communities, by *Wastewater* [24]. *Rice* was further subdivided by root environment (rhizosphere, rhizoplane, and endosphere); *Mice*, by origin (wild or lab); and *Vagina*, by previously reported community types [19]. For each dataset, samples presenting very low sequence depths were removed; chimeric sequences, identified using QIIME (v1.9.1) [25] (usearch61), were also removed. Then, all remaining samples were subsampled to a (minimum) common depth (Suppl. Table 1) using QIIME (v1.9.1). Finally, the normalized datasets were analyzed with *BacterialCore.py* (https://git.io/Je5V3). The script employs a clustering-based core detection approach as previously described [11], and a new approach based on a 16S rRNA gene phylogeny. The former approach clusters sequences at all 0.01 distance steps between 0.75 and 0.97, and defines core OTUs (i.e. PCGs) at each step as those present in all samples after the removal of all sequence data belonging to the core OTUs detected at higher similarity clustering thresholds [11]. For the latter approach, the algorithm traverses the 16S rRNA phylogenetic tree from the leaves to the root; if a leaf/node is present in a selected percentage of samples (here, 100%), it is flagged as "core" and its abundance values are removed from all parental nodes before continuing; thus, reported core groups are non-overlapping (Suppl. Fig. 1). Additionally, *BacterialCore.py* provides per core-group information, statistics, and consensus taxonomies.

2.2. In vitro experimental dataset processing

Goldford et al. [7] cultivated microbial communities derived from soil and leaf environments in M9 minimal medium supplemented with either glucose, leucine, or citrate as the sole energy and carbon source. Twelve initial communities were subjected to 12 cycles of dilution and growth in microtiter plates, and each initial community was grown in eight replicates. DNA was extracted from the initial community samples and each experimental one, and the V3-V4 region of the 16S rRNA gene was sequenced as 2 × 250 reads in an Illumina MiSeq sequencer. We obtained the resulting sequences from the repository and the metadata from the authors. Filtering, trimming, sample inference, merging of paired reads, and chimera removal were performed with the *R* package DADA2 as described by Goldford et al. [7]. For PCG detection, *BacterialCore.py* was used with all endpoint samples from the same culture medium subsampled to the (minimum) depth of 15904, 6382, and 20037 sequences for glucose, leucine, and citrate, respectively. Intragenomic 16S rRNA gene diversity is known to confound diversity estimates and the quantification of individual populations [27]. However, the sequence identity of most intragenomic 16S rRNA gene sequences differs by $< 1\%$ [28]; thus, for the rest of the analyses using community composition data, we used the community table obtained after clustering the experimental sequences against the Greengenes v13.5 99% reference dataset and reference trees [26].

2.3. Mapping 16S rRNA gene sequences to genomes and core pangenome exploration

The 16S rRNA gene sequences of intra-PCG experimental populations were mapped to the GTDB [29] 16S rRNA gene bacterial database (bac120_ssu_reps_r95) using *Nucmer* 3.1 [30], with alignment cutoffs of $> 97\%$ identity and $> 90\%$ coverage. Then, complete amino acid sequence-coding genomes were obtained from the same resource (gtdb_proteins_aa_reps_r95) for best matches passing the above cutoffs.

Genomes obtained for each PCG were annotated using *eggNOG-mapper* v2.0.1 with default parameters [31]. Pangenomes were defined as annotations present in at least 90% of the available genomes for each PCG, and thus considered as the "core genome" of each PCG. Functions not related to metabolism or transport were excluded, and the rest were explored manually using the online tool *KEGG Mapper Reconstruct* [32].

2.4. Metabolic modeling

We produced genome-scale metabolic models for each genome retrieved using the automated tool *CarveMe* v1.4.1 [33] without gap-filling and with Gram-specific universal models. Some of the models produced were not able to grow on their respective media and were removed from further analyses (Suppl. Table 2).

A species metabolic interaction analysis was carried out using *SMETANA* [34]. Using the genome scale metabolic models previously obtained, several metabolic metrics were calculated for each model. Then, for each experimental carbon source, pairwise interaction metrics were obtained for inter-PCG models of species co-existing in at least one endpoint experimental sample.

The biomass reaction flux of each metabolic model was obtained by a Flux Balance Analysis (FBA) and a Constrained Allocation FBA (CAFBA) [35], as implemented in *ReFramed* (github.com/cdanielmachado/reframed) using default settings. The results of the two methods were highly similar (i.e. variations below the second decimal place), thus, we reported only those of the CAFBA here. Initially, growth was assessed in M9 minimal medium supplemented with the respective carbon source. Later, we conducted inter-PCG simulations in which one metabolic model grew on the other's spent media, and vice versa. Spent media compositions were designed based on the exchange reactions predicted to have a positive flux for the "donor" metabolic model grown on the original medium as simulated by CAFBA. For chosen inter-PCG pairs, results were parsed into reactions present in only one of the models or in both models but with a different reaction sign and then graphically depicted to show the possible nature of their metabolic interaction.

2.5. Drift modeling

We implemented a computational model that simulates drift in communities undergoing dilution-growth cycles. At each cycle, the community is first randomly subsampled to mimic experimental dilution. Then, the community grows by adding one new individual at each step until the pre-dilution community size is reached. At each growth step, the probability for each population to grow equals its actual relative abundance. This process models neutral growth along a passage experiment in which all individuals have equal fitness (Suppl. Fig. 2). We also introduced a second scenario in which the initial community populations are initially split into different groups, which then undergo the above dilution-growth cycles independently with different pre-set per group community sizes. The output contains each final community or all trajectory communities, depending on the settings. The same number of transfers and dilution rates employed by Goldford et al. (12 and 0.008, respectively) were also used here. The initial simulated communities reflected the starting community compositions of their experiment, and the community size matched its respective variable sequencing depths (49470 and 7271 for simulations starting at transfers 0 and 1, respectively). The second restricted scenario drove growth using the observed average relative abundance of the PCGs in the glucose experiment: 71.5% for Node 35562 (Enterobacteriaceae), 21% for Node 27828 (Pseudomonadaceae), and 7.5% for a group containing all other OTUs. For the restricted scenario, we chose single per group community size values (the average in the endpoint experimental communities), as well as per-group percentages randomly drawn from observed experimental distributions. Finally, we followed the same approach starting at timepoint 1. Due to missing values in the Goldford et al. dataset, we modeled growth for only those original community trajectories presenting all 8 timepoint replicates (10 and 2 communities for simulations starting at transfer 0 and 1, respectively). For each starting community, we ran 100 simulations.

For experimental and drift-simulated endpoint communities arising from the same initial community composition, OTU tables were subsampled to a common depth (the minimum number of sequences in the dataset) before obtaining the following indices: richness (number of OTUs), Shannon (diversity), Pielou (evenness), and Faith (phylogenetic diversity). Next, we derived empirical cumulative distribution functions from each community type (experimental, simulated neutral, and simulated phylogenetically constrained), which were then compared using the DTS test as implemented in the package *twosample* in the *R* statistical environment.

2.6. Interactions between populations

Co-abundance correlations in the experimental passage data were evaluated following the same strategy recently employed by Goyal et al. [36] for a similar dilution-growth passage experiment. The approach, which assumes that fluctuations in OTU abundance are independent of one another (i.e. no interactions) and follow a gamma distribution, assesses if the abundance trajectory of two populations is more coupled than expected by chance. For this analysis, first, the temporal abundance trajectories of all OTUs were selected. Then, the Pearson correlation coefficient was calculated for each pair of OTUs based on their trajectories. The statistical significance of the correlations were calculated against the expected correlation distribution produced following a null model. For the null model, a gamma distribution was constructed for each OTU based on the experimental data. Then, random communities were generated from the gamma distributions, and the resulting community compositions were renormalized by dividing each individual abundance by the communities' total sum. The simulated communities were then arranged in passage trajectories and the correlations between each pair of OTUs obtained as per the experimental data. In our case, we assessed the existence of interactions within the last five passages, when communities remained compositionally stable (i.e. the PCGs had emerged, and their relative abundances were comparatively stable). The analysis was conducted independently for experimental communities arising from each initial community, and only OTUs appearing in at least one endpoint experimental community were evaluated. To avoid spurious correlations, experimental and simulated correlations were recorded only if both members of the pair had at least 10 reads in a minimum of three of the five trajectory points. Experimental correlations were obtained only from trajectories presenting no missing values. As a result, we used a single trajectory for six of the initial communities and the average of eight trajectories for two communities. Finally, interactions were deemed as present if their correlation value fell within the top 5% of the extreme values of the null correlation distribution. The results were depicted as a network using the *R* package *iGraph.*

All scripts and additional data are available at https://github.com/silvtal/phyloassembly. Repository information for the 16S rRNA gene datasets used in this study is available in their corresponding original publications.

3. Results

3.1. Pervasive existence of PCGs in microbial ecosystems

Most of the microbial ecosystems analyzed presented a considerable number of PCGs detected at different phylogenetic depths along the bacterial phylogeny (Fig. 1, Suppl. Table 1, Suppl. Material 1). The exceptions to this pattern were the mucosal environments (*Illeum*, *Rectum*, and *Vagina*) and the *Leaf* ecosystem, which all presented very few PCGs. The low number of PCGs detected in the mucosal ecosystems could be related to their comparatively low sequencing depth (Suppl. Table 1) and/or host immune response. Overall, the detected PCGs represented a preeminent fraction of the total community (Suppl. Table 1), with the lowest pooled abundance values being 18.5% (*Leaf*) and 34.9% (*Illeum*), and the largest, 77.6% (*Sponge*) and 93.4% (*Vagina*). In general, the results of the clustering and tree-based approaches were largely congruent (Fig. 1, Suppl. Material 1), particularly in terms of PCG number and phylogenetic depth.

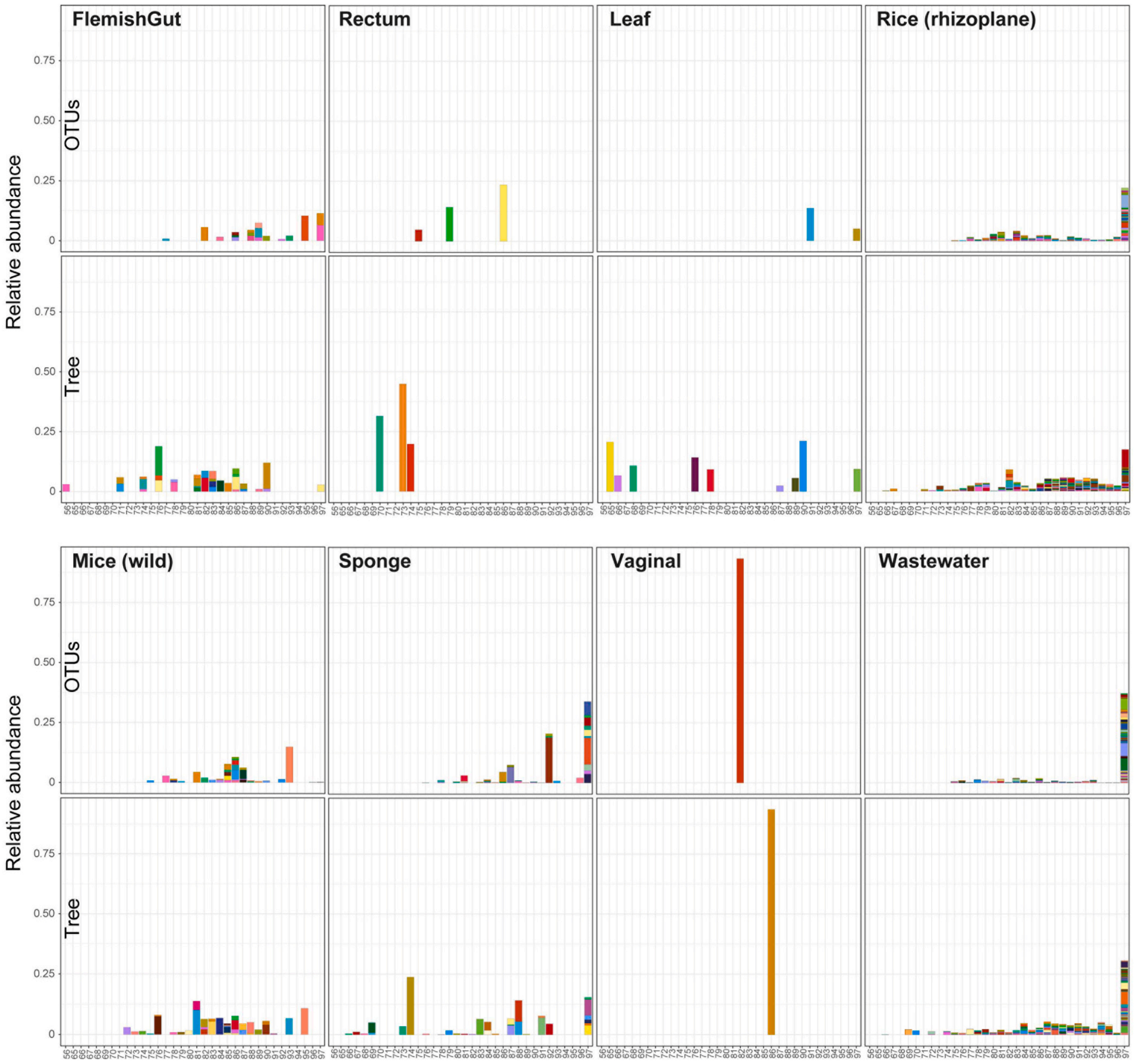


Fig. 1. Detection of PCGs in selected datasets. Results for selected datasets based on the dynamic clustering of 16S rRNA gene sequences [OTUs] and the phylogenetic tree-based approach [Tree]. The x-axis values represent, in the first case, clustering distance thresholds and, in the second, maximum intra-node distance among all reference sequences belonging to that node. The two value types are not directly equivalent and are aligned in the figure for comparative purposes. For each threshold, OTUs/nodes present in all samples (i.e. core) appear vertically stacked with individual heights representing the average relative abundance of each core OTU/node in the dataset.

3.2. Phylogenetic signal exploration reveals two specific PCGs in the in vitro experimental datasets

We assessed the phylogenetic signal in microbial community assembly using a previously published in vitro experimental dataset from a study featuring the *ex situ* cultivation of complex natural microbial communities on a single carbon and energy source [7]. This dataset was deemed ideal owing to its high replication from different starting communities, which allows a phylogenetic signal to be appropriately delimited, and its relative simplicity as a microbial ecosystem, which we consider as a reasonable proxy for local community assembly, thus avoiding high-scale sampling bias [3]. Moreover, it had previously been shown to harbor two predominant bacterial families, thus presenting a clear phylogenetic signal. Indeed, the phylogenetic signal detected in the endpoint experimental samples could be delimited in terms of PCGs (Suppl. Table 3). We discarded from further analyses the PCGs presenting an extremely high phylogenetic depth (Suppl. Table 3) as they unlikely present eco-functional cohesion [10]. All three experimental media presented two PCGs that, in each case, accounted for a very large fraction of the total community abundance (Suppl. Table 4). For each of the three media, one PCG was taxonomically affiliated with the Pseudomonadaceae family, and the other, with the Enterobacteriaceae family or the phylum Proteobacteria. For the glucose and citrate media, the same Enterobacteriaceae-affiliated PCG (E) was detected, however the Pseudomonadaceace-affiliated PCGs (P) differed in that citrate's P was phylogenetically shallower (Suppl. Table 5). In comparison with Goldford et al., who described the phylogenetic signal of their experimental dataset in terms of taxonomic family dominance, our results indicate that E (in citrate and glucose) and P (in citrate) are phylogenetically more restricted than its corresponding taxonomic family (roughly half of the phylogenetic distance and a third of the reference sequences).

3.3. *Core pangenome analysis shows the potential energetic advantage of E over P on the experimental media*

To explore the mechanism of persistence and reproducibility of these PCGs along the course of our drift experimental setup, we obtained matching database genomes of the experimental populations of each PCG and constructed 90% consensus pangenome annotations. We then looked for metabolic and transport genes exclusive to each core pangenome. The results indicated that glucose could be transported into the cytoplasm differently for each PCG. E's pangenome presents a PEP phosphotransferase whereas that of P presents an ABC transporter, as was also noted by Goldford and colleagues for the respective families. Both transport systems spend energy, but the former saves it by transporting phosphorylated glucose into the cytoplasm. Furthermore, glucose degradation can follow the Entner-Doudoroff or the Embden-Meyerhoff pathway towards pyruvate, with the latter pathway providing one extra ATP. Coded in P's but not E's pangenome are key enzymes for the less efficient Entner-Doudoroff pathway, such as phosphogluconate dehydratase and 2-dehydro-3-deoxyphosphogluconate aldolase (Fig. 2).

Also, P's pangenome includes glucose dehydrogenase, a membrane bound enzyme able to transform extracellular glucose to gluconate, which in turn can eventually convert to other compounds that, like gluconate itself, are less utilizable to other microbes [37], thus possibly increasing its fitness. However, the experimental glucose medium lacked the necessary pyrroloquinoline quinone cofactor; therefore, this theoretical advantage could have only played a role in the present experimental system if this cofactor was stably produced by co-existing populations. Without additional information, it is not possible to discern the degree to which P populations may use the two pathways. Thus far, the results suggest that E populations could have an energetic advantage over P ones when grown on glucose, which is in accordance with the higher abundance of E in these samples; however, they do not provide an explanation for the persistence of P populations in the harsh drift experiment. The pangenome exploration approach did not yield cues for the existence of PCGs in the citrate environment. As for leucine, P's pangenome includes 3-metilcrotonil-CoA carboxylase, an enzyme able to assimilate inorganic CO_2 during leucine metabolism, which is advantageous in starvation scenarios [38]. Indeed, such a scenario may have occurred at the end of each growth cycle in Goldford et al.'s [7] experiments.

3.4. *Metabolic modeling of inter-PCG interactions provides a potential rationale for community niche structure*

Our modeling results using universal metabolic models as implemented by the automated modeling tool *CarveMe* are based on only the glucose and citrate datasets as the generated models were unable to grow on leucine. Our initial metabolic interaction analysis using *SMETANA* suggested the lack of obligate syntropy between inter-PCG pairs in the glucose medium. However, for citrate, two dependent OTUs were found in the Enterobacteriaceae-related PCG. According to our results, OTU 4454257 could only grow in the presence of either the Pseudomonadaceae-affiliated OTUs 4370747 and 4419276 or the Enterobacteriaceae-affiliated OTUs 4475144, 691423, 9994, and 3944484. The second dependent OTU (3944484) could only grow in the presence of the Pseudomonadaceae-affiliated OTU 4419276. Metabolic resource overlap values between inter-PCG models indicated high metabolic similarity (Suppl. Tables 6 and 7), and also the potential exchange of compounds by inter-PCG models grown on the same medium (Suppl. Tables 8 and 9).

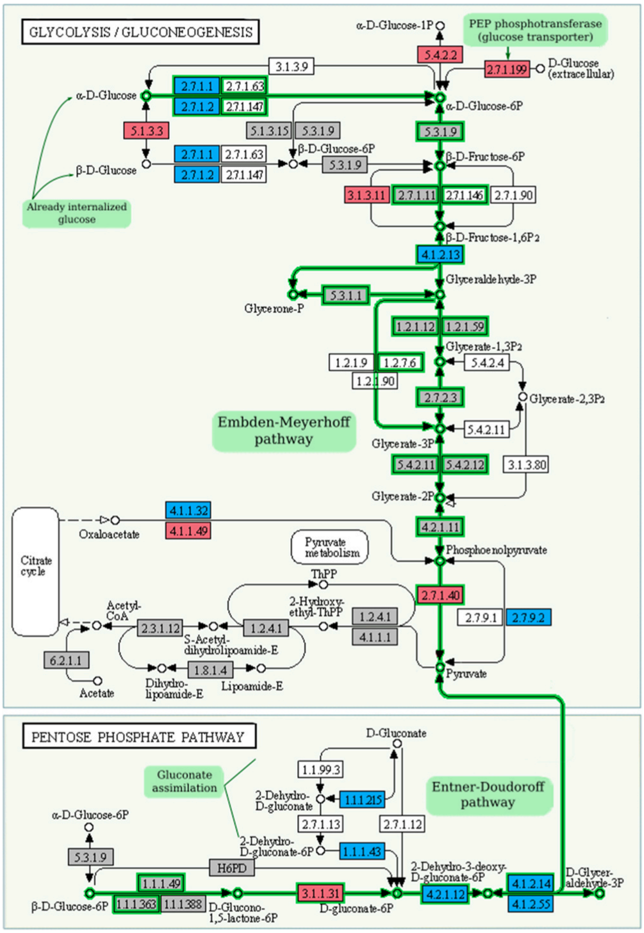


Fig. 2. Differential presence of glucose metabolism pathways in the core pangenomes of E and P. The diagram depicts only the parts of the glycolysis and pentose phosphate pathways relevant to the present study. The Embden-Meyernoff and Entner-Doudoroff pathways are highlighted in green. Reactions represented in P's pangenome are colored in blue and E's reactions, in red. Reactions represented in both pangenomes are colored in gray. In the case of E, extracellular D-glucose is directly converted into D-glucose-6-phosphate for glycolysis by the phosphotransferase transport system (EC 2.7.1.199), whereas in P, D-glucose-6-phosphate is converted from already internalized D-glucose, which needs an extra ATP-expending step. In the Entner-Doudoroff pathway, the essential enzymes phosphogluconate dehydratase (EC 4.2.1.12) and 2-dehydro-3-deoxyphosphogluconate aldolase (EC 4.1.2.14) are only present in P's core pangenome. [Diagram generated using the KEGG Mapper Reconstruct tool].

Simulated growth of the individual models with FBA and CAFBA indicated that they could potentially grow alone, and that their simulated growth ratios on the same carbon source were only slightly higher for E (Suppl. Tables 10 and 11). Goldford et al. [7] hypothesized that highly similar growth rates sustain the presence of both groups in the system. However, we propose an alternative, potentially more plausible, explanation, particularly considering that the experiment was conducted for ca. 84 generations with 12 severe dilution-related drift bottlenecks, and a stable composition still resulted over the course of eight replicates from 12 initial communities.

Given that inter-PCG pairs can potentially exchange metabolites when grown together, we simulated growth of inter-PCG pairs reciprocally on fresh media and the partner's spent media. For P grown on the spent media of E, which was grown on glucose or citrate, the observed growth flux ratios of P to E were highly similar to the corresponding experimental ratios (Suppl. Table 12) (glucose experiment 0.29, average simulated 0.29 ± 0.03; citrate experiment 0.63, average simulated 0.49

± 0.10). These results strongly support the idea that the interaction between PCGs could have driven the experimental compositions. Next, we parsed the modeling results for the inter-PCG pairs (in which E grew on fresh media and P on the spent media of E) and found that, for both glucose and citrate, acetate represented the main exchanged metabolite (Fig. 3, Suppl. Fig. 3). In both cases, the acetate released by E is first transported into the cytoplasm by P, which then follows two alternative routes: i) production of acetyl-phosphate, then transformation to acetyl-CoA, or ii) direct production of acetyl-CoA via a succinyl-CoA:acetyl-CoA transferase. Both succinate and acetyl-CoA can then be fed into the Krebs cycle to produce reductive power and thus the means to generate ATP and sustain anabolism. For both glucose and citrate, half of the inter-PCG pairs assayed presented both active routes, though a slightly higher flux was observed for route (ii); the other half of the pairs presented only route (ii).

3.5. *Intra-PCG neutral dynamics partly explain in vitro experimental diversity patterns*

Having acquired an understanding of the metabolic determinants of niche structure based on phylogenetic composition, genomic databases, and metabolic modeling, we next explored the population dynamics of the PCGs. First, we modeled neutral growth and drift along the passage experiment for all starting communities, also including a second scenario in which the community size of each of the three groups (E, P, and Others) was fixed according to their average relative abundance in all end-point experimental communities. The three resulting datasets (experimental, neutral, and phylogenetically constrained neutral) were then compared in terms of richness, diversity, evenness, and phylogenetic diversity. Our initial expectation was that intra-PCG ecological cohesiveness leads to intra-PCG neutral dynamics, and hence, the phylogenetically constrained scenario would most resemble the experimental results. Though we observed this result for most of the starting communities and metrics (Suppl. Fig. 4), the values of the phylogenetically constrained scenario were generally higher than those for the experimental one. To assess if the difference in values was associated with the use of single, fixed per group community sizes, we reanalyzed the data using per-group sizes at each step that were randomly sampled from the experimental distribution. In general, we obtained the same results (Suppl. Fig. 4). We also repeated the analysis starting at transfer 1 and, in this case, the diversity values of our simulation more closely matched those of the experimental observations (Fig. 4).

Thus, intra-group neutrality with fixed niche sizes could recapitulate the experimental observations except at the initial step when the starting community faced the first experimental selection-drift cycle. To better understand this phenomenon, we compared the abundance of individual populations in the experimental vs. simulated datasets (Suppl. Fig. 5). In most cases, one or two of the initial intra-PCG group populations dominated the experimental endpoint communities. Although a few of these initial populations presented a higher abundance than expected, in most cases, the abundance was lower than expected. These results indicate unequal fitness among the initial populations of the same PCG, contradicting our initial hypothesis, followed by strong selection established during the first transfer that resulted in the removal of intra-group populations with low fitness. After this point, the behavior of the experimental ecosystem was more in line with intra-group neutral dynamics and group relative abundances, which were likely imposed by the metabolic considerations mentioned above.

The observed differences in intra-group fitness could be related to variations in growth rate in the experimental media and to the existence of strong biotic interactions. To test the latter, we modeled co-abundance correlations in the late experimental transfers, which

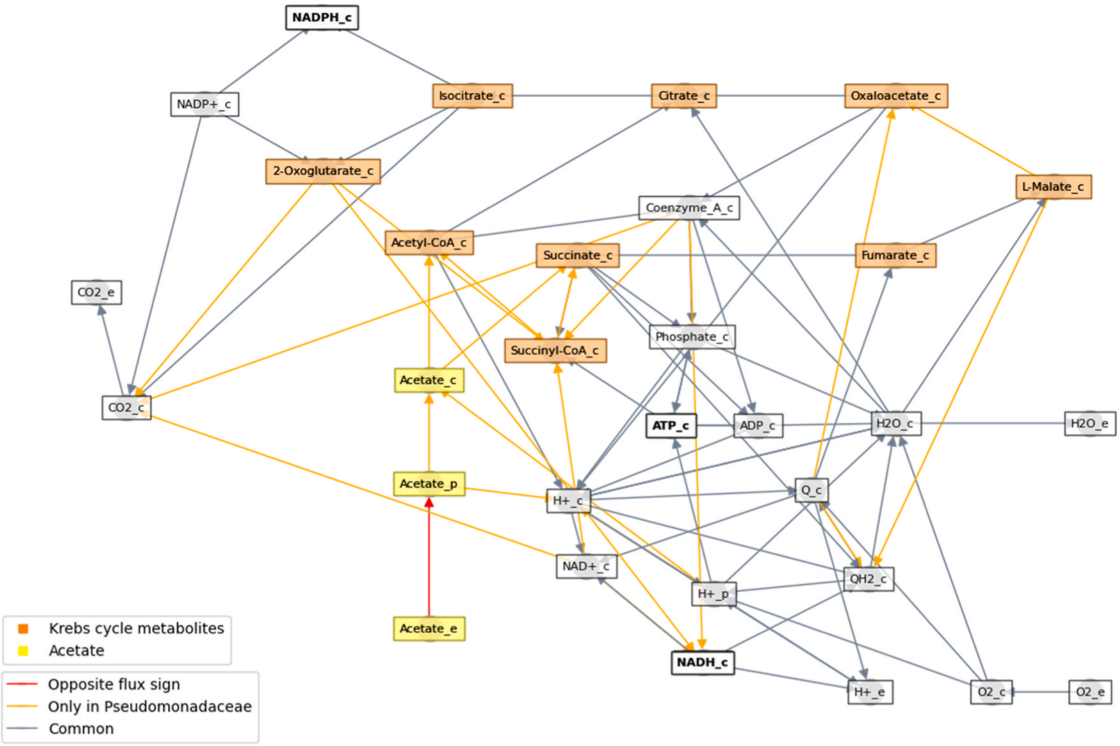


Fig. 3. Simulated inter-PCG metabolism for E-P pair grown on glucose. The diagram depicts the simulation scenario for E grown on M9 + glucose media and P, on the spent media of E. Edges represent reactions from P with absolute flux values > 3.5. Red reactions present opposite flux signs in both P and E models. Orange reactions are active exclusively in P, and gray reactions are active and with same flux sign in both P and E models. Metabolic compartments are designated with suffixes "_e", "_p" and "_c" for extracellular, periplasm, and cytoplasm, respectively. The diagram shows that P takes extracellular acetate, while E exports it (red arrow). Intracellular acetate in P can then be transformed into acetyl-CoA by spending ATP or through a succinyl-CoA:acetyl-CoA transferase. Both acetyl-CoA and succinate can be fed into the Krebs cycle (orange compounds), generating reductive power and ATP (bold compounds).

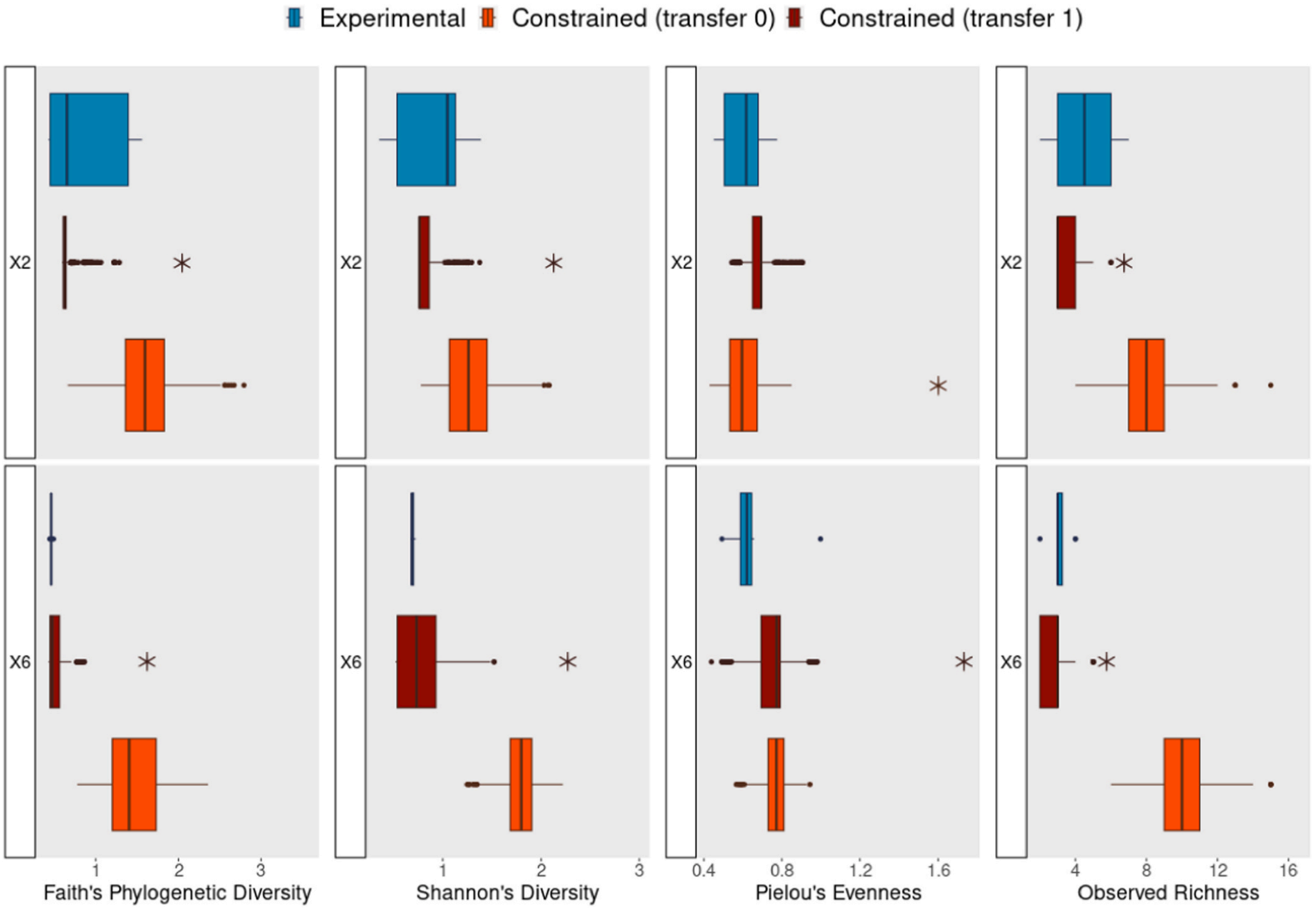


Fig. 4. Distribution of the diversity values of the experimental and simulated communities. Four diversity metrics were calculated for all the experimental and simulated communities for replicates arising from initial samples X2 and X6. Experimental observations (blue) are compared against 800 simulations for each within-group neutral drift and fixed per-group community sizes (constrained neutrality). The starting communities were either the original samples (orange) or the experimental communities after the first growth period (dark red). The asterisk indicates the simulated distribution that is most similar to the experimental one according to the DTS analysis.

showed stable compositions. Negative inter-group correlations accounted for most of the observed significant correlations, though we also observed a few instances of significant intra-group correlations, both positive and negative (Suppl. Fig. 6). The latter result could be expected a priori based on likely partial ecological redundancy, and positive intra-group interactions may be indicators of a division of labor. However, the metabolic division of labor that seems to explain the persistence of P and E in the in vitro experimental dataset, or the observation of significant positive intra-group interactions, does not explain the preeminence of the negative inter-group interactions. The proposed conceptual framework posits that the sum of intra-group populations abundance is governed by relative niche size. In this regard, both the negative inter-group and the positive intra-group interactions could be explained by fluctuations in niche size of E and P along the modeled experimental samples.

4. Discussion

We analyzed 16S rRNA gene datasets from various human and plant-associated environments, as well as from animal-associated and environmental microbial communities. PCGs were detected in terms of nodes of a phylogenetic tree present in all instances of each community type. Here, we provide evidence of the presence of PCGs in all datasets analyzed, indicating they could be a predominant feature of microbial ecosystems.

Previously, we proposed the conceptual framework of PCGs and an approach for their study [3]. In the present work, we aimed to exemplify how phylogenetic signal (i.e. PCGs) in a microbial ecosystem could be used to help understand its metabolic niches and assembly rules. Drawing from the well supported ideas that traits and ecological function show some degree of phylogenetic conservatism [10,39] and that the core pangenome of a clade can be translated to its core ecological niche [11,40], we re-analyzed Goldford et al.'s [7] experimental community composition dataset using solely phylogenetic signal and available genomic resources in order to gauge how much of this community's ecology could be inferred without further in-depth experimentation.

In Goldford et al. [7], the observed phylogenetic signal was restricted to two families, Enterobacteriaceae and Pseudomonadaceae. Our more fine-grained analysis showed the same restriction to the two families, but it, more importantly, revealed that the PCGs were significantly more restricted than the corresponding taxonomic families. Thus, we argue that phylogenetic signal should be finely delimited before attempting to map phylogeny to shared eco-functional traits based on genomic information.

Having finely delimited the PCGs present in the ecosystem, we explored their metabolic niches through a core pangenome analysis. The results pinpointed a potential advantage of E when grown on glucose, derived from its more efficient transport system and the possibility that P processes some or all of the glucose through a less efficient pathway. Subsequent analyses of inter-PCG metabolic models indicated that the experimental P/E could be recapitulated if E consumed glucose or citrate and P consumed the resulting acetate by-product, thus providing an indication of the ecosystem's niche structure.

Our analysis of the community at the population level showed that, contrary to our starting hypothesis, the intra-PCG populations did not present equal fitness. Nonetheless, after the first selective transfer in our drift model with phylogenetic constraints, the resulting diversity metrics were similar to those observed experimentally. Our results are in line with those of Datta et al. [41], who showed that, during the early stages of community assembly, environmental filtering couples the dynamics of functionally equivalent populations, which, as a result, initially behave in a non-neutral fashion. The ability of the model to approximate experimental observations is noteworthy, though several a priori caveats must be taken into account. These include the inability to model populations that are initially present below the sequence depth-based detection threshold, the use of total community sizes based on normalized sequence numbers instead of experimental cell counts, and the apparent distance between a stepwise growth simulation and in vivo bacterial growth dynamics.

Another interesting finding from our general exploration of the glucose dataset was the indirect evidence of biotic interactions. For instance, a single E population (4454257; GreenGenes reference number) could either dominate or co-dominate (alongside 4399988) the endpoint experimental replicates in a non-neutral fashion, depending on the starting community context (Suppl. Fig. 5). However, we hypothesize that the results of our analytical approach to detect possible biotic interactions using co-abundance networks are more consistent with changes in niche sizes. A recent in-depth metabolism study of Goldford's experimental system indicated that each transfer cycle followed an E to P succession driven by the consumption of glucose and the concomitant accumulation of acetate (see below). Possible fluctuations in sampling times or cycle dynamics could alter PCG abundance; therefore, observed correlations would be related to succession stage instead of actual biotic interactions, as recently cautioned by Pascual García et al. [42]. Given this, sampling of multi-replicated enrichment communities [43] should be adjusted on the basis of community metabolic or abiotic parameters to better serve as model systems, though the implementation of this may prove difficult.

Recently, Pascual-García [44] provided a commentary on our proposed conceptual framework. We acknowledge that our initial description provided a tautological formulation, and agree that the detection of PCGs could benefit from using a methodological pipeline independent of the number of samples or their depth. In this regard, we detected PCGs in terms of 16S rRNA gene sequence clusters and nodes of a phylogenetic tree at different depths that are present in all samples from the same ecosystem type. While this is a useful heuristic approach, other criteria, such as a Poisson distribution [45], a competitive lottery schema [46], invariance metrics [47], or the use of neutral models, could also be employed. On the other hand, we feel that the commentary presents misinterpretations regarding our propositions and framework. Significantly, Pascual-García's commentary presents a mental exercise with different assembly scenarios, with the author exploring these scenarios in search of PCGs [44]. However, Pascual-García apparently analyzed only a single sample per scenario, despite the fact that our framework requires the analysis of a large number of samples from the same scenario [3]. This key difference likely accounts for the disparity in the results and interpretations reported in the commentary as pertaining to the framework compared to our own evaluation of the reported exercise. Nonetheless, we both seem to agree on the potential of the PCG approach, which also aligns with Goyal et al.'s recent call to clarify how phylogenetic signal maps to ecological functions [36].

As alluded to above, Estrela et al. [48] recently followed-up on Goldford et al.'s [7] results using a canonical experimental approach, which has provided an overlapping study with which we can compare our results. They first repeated the original experimental set-up with glucose, then isolated several strains that represented a large percentage of the experimental communities and measured their growth rate on the experimental glucose medium. They found that Enterobacteriaceae isolates had higher growth rates than Pseudomonadaceae isolates, contradicting their initial hypothesis that both families coexisted due to similar growth rates [7]. Rather, their results suggest that Pseudomonadaceae populations were sustained in the community owing to the higher competitive ability of the metabolic by-products secreted by the Enterobacteriaceae populations. To confirm this idea, the authors analyzed the secreted by-products of the Enterobacteriaceae strains and showed that acetate was dominant and that the Pseudomonadaceae isolates had a higher growth rate in acetate compared with the Enterobacteriaceae isolates. They proposed that the ecosystem has two phylogenetically conserved metabolic niches: fermenters (Enterobacteriaceae) and respirators (Pseudomonadaceae), which are selected according to the organic acids released by the fermenters on which they specialize. Significantly, as mentioned above, they measured the ratio of P to E at different time points during a 48-h growth cycle and, concomitantly, quantified glucose and acetate levels. In this manner, they demonstrated the differential growth advantage of the two types of isolates: Enterobacteriaceae had an advantage early during the incubation period when glucose was abundant, and Pseudomonadaceae, later, when glucose was absent and acetate was abundant. Thus, each transfer cycle represented a succession from E to P. They also modeled P and E interactions using well-curated metabolic models. They found, as we did with our automated models, that the experimental P/E could be recapitulated only when glucose is completely metabolized to acetate by Enterobacteriaceae, and Pseudomonadaceae fully respires acetate to CO_2. In summary, their valuable experimental follow-up study shows essentially the same results as those obtained by our purely bioinformatic approach, further supporting the potential usefulness of PCG analyses in evaluating microbial community assembly.

5. Conclusion

The results presented here show that the proposed PCG-based approach can provide ecological insights comparable to those obtained after canonical experimentation. While it is undisputable that appropriate experimentation provides more direct evidence than our approach, the latter can be of great value for cases in which experimentation is not possible or practical, or as a complement. As previously discussed in detail [3], the conceptual framework and proposed methodological approach has some limitations including its reliance on sufficient sequencing depth, the low resolution of the 16S rRNA phylogeny at deep and shallow nodes, and its incapacity to help explain microbial ecosystems that do not show a phylogenetic signal or those with strong succession patterns or frequent and strong perturbations [3]. Nonetheless, we show the presence and high relative abundance of PCGs in a large and diverse array of environments, suggesting PCGs are a predominant feature of microbial ecosystems that, when detectable, can be used to explore microbiome function and assembly rules. Leveraging this phylogenetic signal thus offers a relatively quick (and cost effective) way to gain further insight on Earth's microbial ecosystems.

CRediT authorship contribution statement

Daniel Aguirre de Cárcer: Conceptualization, Supervision, Project administration, Funding acquisition, Writing – original draft, Investigation. **Marcos Parras-Moltó**: Methodology, Software, Formal analysis, Data curation. **Silvia Talavera**: Methodology, Software, Formal analysis, Investigation.

Declaration of Generative AI and AI-assisted technologies in the writing process

No AI-assisted technology was used in the writing process.

Declaration of Competing Interest

The material represents original research, has not been previously published (is deposited as a preprint in Research Square 10.21203/rs.3.rs-2272005/v1) and has not been submitted for publication elsewhere while under consideration in *CSBJ*. The authors declare no conflict of interests.

Acknowledgements

This work was funded by the Spanish Ministry of Science and Innovation grant PID2019–108797RB-I00 awarded to DAC. DAC conceived the idea, MP developed *BacterialCore.py*, with input from DAC, and performed the detection and analysis of PCGs in the selected datasets. ST carried out all other analyses. DAC wrote the manuscript, with input from ST.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.039.

References

[1] Vellend M. Conceptual synthesis in community ecology. Q Rev Biol 2010;85(2):183–206.
[2] Zhou J, Ning D. Stochastic community assembly: does it matter in microbial ecology? Microbiol Mol Biol Rev 2017;81(4):00002–17.
[3] Aguirre de Cárcer D. A conceptual framework for the phylogenetically constrained assembly of microbial communities. Microbiome 2019;7(1):142.
[4] Stegen JC, Lin X, Konopka AE, Fredrickson JK. Stochastic and deterministic assembly processes in subsurface microbial communities. ISME J 2012;6:1653.
[5] Burns AR, Stephens WZ, Stagaman K, Wong S, Rawls JF, Guillemin K, et al. Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. ISME J 2016;10(3):655–64.
[6] Horner-Devine MC, Bohannan BJ. Phylogenetic clustering and overdispersion in bacterial communities. Ecology 2006;87(7 Suppl):S100–8.
[7] Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, et al. Emergent simplicity in microbial community assembly. Science 2018;361(6401):469–74.
[8] Shafquat A, Joice R, Simmons SL, Huttenhower C. Functional and phylogenetic assembly of microbial communities in the human microbiome. Trends Microbiol 2014;22(5):261–6.
[9] Martiny JB, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: a phylogenetic perspective. Science 2015;350(6261):aac9323.
[10] Parras-Moltó M, Aguirre de Cárcer D. Assessment of phylo-functional coherence along the bacterial phylogeny and taxonomy. Sci Rep 2021;11(1):8299.
[11] Aguirre de Cárcer D. The human gut pan-microbiome presents a compositional core formed by discrete phylogenetic units. Sci Rep 2018;8(1):14069.
[12] Leventhal GE, Boix C, Kuechler U, Enke TN, Sliwerska E, Holliger C, et al. Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. Nat Microbiol 2018;3(11):1295–303.
[13] Wilbanks EG, Jaekel U, Salman V, Humphrey PT, Eisen JA, Facciotti MT, et al. Microscale sulfur cycling in the phototrophic pink berry consortia of the Sippewissett Salt Marsh. Environ Microbiol 2014;16(11):3398–415.
[14] Cordero OX, Datta MS. Microbial interactions and community assembly at microscales. Curr Opin Microbiol 2016;31:227–34.
[15] Gralka M, Szabo R, Stocker R, Cordero OX. Trophic interactions and the drivers of microbial community assembly. Curr Biol 2020;30(19):R1176–88.
[16] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. Science 2016;352(6285):560–4.
[17] Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al. Genetic determinants of the gut microbiome in UK twins. Cell Host Microbe 2016;19(5):731–43.
[18] Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 2014;15(3):382–92.
[19] Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci 2011;108(Supplement 1):4680–7.
[20] Edwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S, et al. Structure, variation, and assembly of the root-associated microbiomes of rice. Proc Natl Acad Sci 2015;112(8). E911.
[21] Wagner MR, Lundberg DS, Del Rio TG, Tringe SG, Dangl JL, Mitchell-Olds T. Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. Nat Commun 2016;7(12151).
[22] Moitinho-Silva L, Nielsen S, Amir A, Gonzalez A, Ackermann GL, Cerrano C, et al. The sponge microbiome project. Gigascience 2017;6(10):1–7.
[23] Rosshart SP, Vassallo BG, Angeletti D, Hutchinson DS, Morgan AP, Takeda K, et al. Wild mouse gut microbiota promotes host fitness and improves disease resistance. Cell 2017;171(5):1015–28. e13.
[24] Saunders AM, Albertsen M, Vollertsen J, Nielsen PH. The activated sludge ecosystem contains a core community of abundant organisms. ISME J 2016;10(1):11–20.
[25] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Meth 2010;7(5):335–6.
[26] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72(7):5069–72.
[27] Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 2019;10(1):5029.
[28] Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol 2004;186(9):2629–35.
[29] Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 2022;50(D1):D785–d94.
[30] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5(2):R12.
[31] Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 2021;38(12):5825–9.
[32] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
[33] Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res 2018;46(15):7542–53.
[34] Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc Natl Acad Sci 2015;112(20):6449–54.
[35] Mori M, Hwa T, Martin OC, De Martino A, Marinari E. Constrained allocation flux balance analysis. PLoS Comput Biol 2016;12(6):e1004913.
[36] Goyal A, Bittleston LS, Leventhal GE, Lu L, Cordero OX. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. Elife 2022;11.
[37] Whiting PH, Midgley M, Dawes EA. The role of glucose limitation in the regulation of the transport of glucose, gluconate and 2-oxogluconate, and of glucose metabolism in Pseudomonas aeruginosa. J Gen Microbiol 1976;92(2):304–10.
[38] Díaz-Pérez AL, Díaz-Pérez C, Campos-García J. Bacterial l-leucine catabolism as a source of secondary metabolites. Rev Environ Sci Bio/Technol 2016;15(1):1–29.
[39] Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits in microorganisms. ISME J 2013;7(4):830–8.
[40] Lassalle F, Muller D, Nesme X. Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. Res Microbiol 2015;166(10):729–41.
[41] Datta MS, Sliwerska E, Gore J, Polz MF, Cordero OX. Microbial interactions lead to rapid micro-scale successions on model marine particles. Nat Commun 2016;7:11965.
[42] Pascual-García A, Bell T. functionInk: An efficient method to detect functional groups in multidimensional networks reveals the hidden structure of ecological communities. Methods Ecol Evol 2020;11(7):804–17.
[43] Estrela S, Sánchez Á, Rebolleda-Gómez M. Multi-replicated enrichment communities as a model system in microbial ecology. Front Microbiol 2021;12:657467.
[44] Pascual-García A. Phylogenetic Core Groups: a promising concept in search of a consistent methodological framework: Comment to ``A conceptual framework for the phylogenetically-constrained assembly of microbial communities''. Microbiome 2021;9(1):73.
[45] Gumiere T, Meyer K, Burns A, Gumiere S, Bohannan B, Andreote F. A probabilistic model to identify the core microbial community. Biorxiv 2018:491183.
[46] Verster AJ, Borenstein E. Competitive lottery-based assembly of selected clades in the human gut microbiome. Microbiome 2018;6(1):186.
[47] Bradley PH, Pollard KS. Proteobacteria explain significant functional variability in the human gut microbiome. Microbiome 2017;5(1):017–0244.
[48] Estrela S, Vila JCC, Lu N, Bajić D, Rebolleda-Gómez M, Chang C-Y, et al. Functional attractors in microbial community assembly. Cell Syst 2022;13(1):29–42.e7.