microbetag: simplifying microbial network interpretation through annotation, enrichment and metabolic complementarity analysis

Haris Zafeiropoulos[1], Ermis Ioannis Michail Delopoulos[1], Andi Erega[2], Annelies Geirnaert[2], John Morris[3], Karoline Faust[1*]

[1*] Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research , KU Leuven, Herestraat, Leuven, 3000, , Belgium .

[2] Institute of Food, Nutrition and Health, ETH Zurich, Street, Zurich, 8092, , Switzerland .

[3] Department of Pharmaceutical Chemistry, University of California San Francisco, Street, San Francisco, 94143, California, USA .

*Corresponding author(s). E-mail(s): karoline.faust@kuleuven.be;
Contributing authors: haris.zafeiropoulos@kuleuven.be;
ermisioannis.michaildelopoulos@student.kuleuven.be;
andi.erega@hest.ethz.ch; annelies.geirnaert@hest.ethz.ch;
scooter@cgl.ucsf.edu;

Abstract

*

Up to 350 words.
The abstract must include the following separate sections:
Background: the context and purpose of the study
Results: the main findings
Conclusions: a brief summary and potential implications

*Looks liks Chris Quince is our editor.

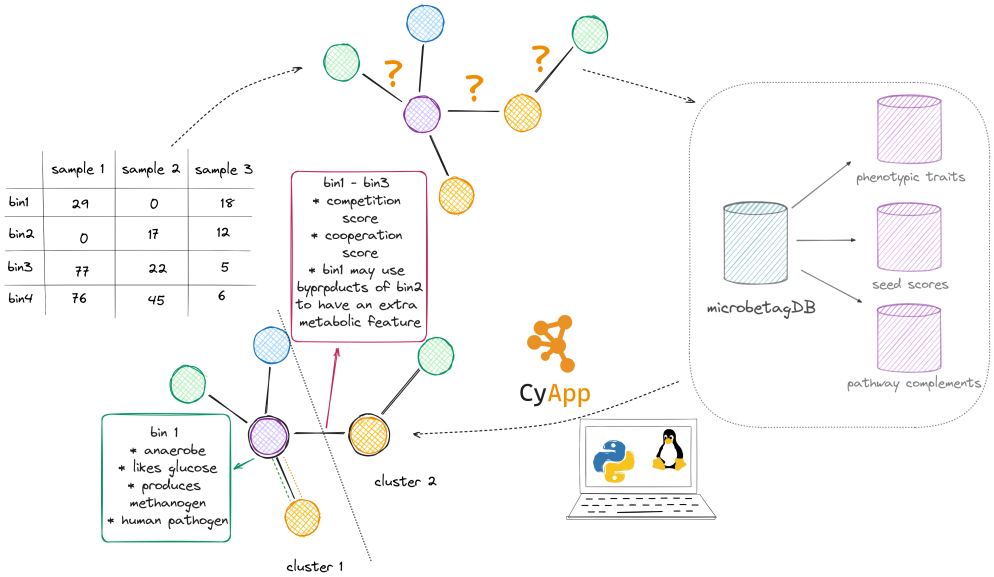

Figure abstract.

Keywords: microbial associations, enrichemnt analysis, data integration, pathway complementarity, seed set

Background [1] [2]

Microbial ecology plays a fundamental role in the stability and resilience of ecosystems and their processes; from soils, aquatic environments and biogeochemical cycles [1] to host-associated environments and the human health [2, 3]. Most microbial species live only in communities [4] and most natural microbial communities consist of hundreds or even thousands of species [5]. Each species exhibits a unique repertoire of reactions and adapts to various niches, each with specific nutrient and environmental requirements. Understanding the dynamics governing interactions among microbial species and their relationships with the surrounding environment would shed light in several aspects microbial ecology [6].

Based on the net fitness effects that result for the taxa involved, the notion of an interaction varies including cooperation, competition, parasitism, commensalism and ammensalism [3]. Metabolic interactions can be established through a range of contact-independent- and contact-dependent mechanisms leading to both positive and negative interactions. These interactions can involve either one-way (unidirectional) or two-way (bidirectional) exchanges of metabolites. Depending on the biosynthetic

[1] We are to submit in the Microbiome journal as a "Software" manuscript, thus we follow these rules
[2] The introduction should not include subheadings. The Background section should explain the relevant context and the specific issue that the software described is intended to address.

cost borne by the interacting partners, two types of metabolite exchanges occur: by-product cross-feeding, where metabolites result from a selfish act of the producer, and cooperative cross-feeding, where one partner actively invests resources to produce metabolites benefiting the interaction partner [7].

High-throughput sequencing (HTS) has provided great insight into the diversity and composition of microbial communities [8]. Uncultivated species can now be detected and their featues can be inferred through their genomic information [9]. Moreover, the composition of thousands of microbiome samples is now accessible allowing for the inference of patterns among sets of samples. A widely used approach to extraxt such patterns, is the creation of co-occurrence networks based on metagenomic read data (amplicon and/or shotgun) [10]. A great number of approaches is available for co-occurrence network inference based on a range of statistical concepts such as: correlation (e.g., CoNet [11], SparCC [12]), linear regression (e.g., SpiecEasi [13]) and causal inference (FlashWeave [14]). Nevertheless, microbial co-occurrence networks continue to encounter various challenges [15]. Their inference inherits the challenges of metagenomic data analysis (e.g., compsitionality, parameters inference) [16]. As a result, network construction remains a tool-dependent analysis [17, 18]. Moreover, more often than not, the returned network looks like a "hairball" of densely interconnected taxa. Thus, additional analysis is necessary to generate testable hypotheses [15]. Addressing the question of *What can we learn from the hairballs* posed by Röttjers et *al.* [4] could provide essential insight on the mechanisms of the interactions.

The assessment of interaction predictions derived from microbial co-occurrence networks has underscored their limitations in accuracy for this task [19]. Theoretical principles derived from network studies might provide indications of emergent biological characteristics [4, 20]. For example, modules (highly interconnected nodes) within microbial co-occurrence networks could serve as indicators of ecological processes that govern community structure, including niche filtering and habitat preference [21]. Data integration and clustering have been suggested to address this challenge [15]. Clusters identified in microbial association networks have demonstrated their ability to mirror key drivers of community composition [22] and sevaral algorithms and implementations are available [23]. However, data integration approaches in microbial co-occurrence networks are so-far limited. Here, we present `microbetag` , a microbial co-occurrence network annotator that exploits several channels of information to enhance/diminish the confidence of the associations suggested by the network and generate hypotheseses for further investigation both at the taxon pair and the community level.

`microbetag` serves as a comprehensive platform that provides information on taxa along with their potential metabolic interactions from multiple channels (see Implementation 3). The key concept here is the reverse ecology approach *reverse ecology* [24]. Reverse ecology leverages genomics to explore community ecology with no *a priori* assumptions about the taxa involved. Making the most of advancements in systems biology and genomic metabolic modeling, as well as system-level analysis of intricate biological networks, the reverse ecology framework enables the prediction of ecological traits for less-understood microorganisms, their interactions with others, and the overall ecology of microbial communities [25].

A metabolic network's "seed set" is the set of compounds that, based on the network topology, need to be acquired exogenously [26] (see Figure 1). Such nodes might be independent, i.e. they cannot be activated by any other node in the network, or they can be interdependent forming groups of seed nodes. Seeds are a useful proxy for the habitat of the organism and an essential tool in the frameowrk of reverse ecology [26, 27]. Based on the seed concept, several graph theory-based metrics (indices) have been described to predict species interactions directly from their networks' topologies [28–31]. Over the last years, the seed apprach has been implemented at the Genome-scale metabolic network reconstructions (GENREs) level. GENREs encapsulate mathematical representations capturing the biochemical reactions that could take place within an organism [32–34].

Metabolic complementarity among species, serving as a reflection of potential cooperation within communities, assesses the capacity for collaboration; cross-feeding or syntrophy interactions are typical examples of such a collaboration. In contrast, metabolic competition refers to the metabolic overlap between two species leading to exploitative competition, e.g. for nutrient resources. Seed and non-seed sets can be used to compute such indices. Thorough examination of such complements can reveal metabolic interactions leading to patterns observed on the co-occurrence network.

However, Bacteria may complement each other not only for getting what is absolutely necessary for them to survive (seeds). For example, microbial species are recognized to exchange metabolites in order to provide support for other advantageous services, such as detoxifying harmful metabolites or offering protection against predators [35, 36]. They can additionally contribute to the production of metabolites essential for the entire community, even if the species itself does not require them [37]. To explore the potenial of a species metabolism given they benefit from a partner of theirs, genome annotations combined with collections of functional units to highlight can provide a valid proxy. We present here a naive approach exporting all possible complements between a pair of species based on their KEGG ORTHOLOGY (KOs) annotations and the KEGG MODULES database [38].

`microbetag` annotates a user's co-occurrence network by integrating phenotypic traits on the taxa present on the network (nodes) and potential metabolic interactions to their suggested associations (edges). A Graphical User Interface (GUI) is supported as a CytoscapeApp providing a user-friendly environment to investigate annotations in a straightforward way. All annotations present in microbetagDB are also available through an Application Programming Interface (API). `microbetag` 's source code is distributed under a GNU GPL v3 license and available on GitHub. Documentation and further support on how to use `microbetag` is available at documentation web-site. To the best of our knowledge there is not a software with which `microbetag` could be compared with directly. To validate our annotations we used a recently published network with partially known interactions between some pairs of species found associated [39] (see Results section, paragraph 3). To demonstrate `microbetag` 's potential, we present the main features of its interface and we discuss a real-world use-case (see Discussion section, paragraph 3).

Implementation [3]

Genomes included

Using the Genome Taxonomy Database (GTDB) v207 metadata files, we retrieved the NCBI genome accessions of the high quality representative genomes, i.e. completeness $\geq 95\%$ and contamination $\leq 5\%$. A set of $26,778$ genomes was obtained, representing $22,009$ unique NCBI Taxonomy Ids. Using these accession numbers, we were able to download their corresponding `.faa` files when available leading to a set of $16,900$ amino acid sequence files. The latter were annotated and used to obtain potential pathway complementarities beween pairs of genomes (see paragraph 3). Last, when available, their corresponding annotations on PATRIC database [40] were retrieved to reconstruct GENREs (see paragraph 3).

Taxonomy schemes

`microbetag` maps the taxonomy of each entry in the abundance table to their corresponding NCBI Taxonomy Id and, if available, their closest GTDB representative genome(s), since several GTDB representative genomes may map to the same NCBI Taxonomy Id. Two well established taxonomy schemes are supported: the GTDB [41] that is being broadly used for bins and/or MAGs taxonomical classification and the Silva database [42] that is widely used in amplicon studies. Both taxonomy schemes link their taxonomies to NCBI Taxonomy Ids [43]. In case none of those two taxonomies was used and the abundance table contains less than 1,000 taxa, `microbetag` maps the user provided taxonomies to NCBI Taxonomy. To this end, `microbetag` makes use of the `fuzzywuzzy` library that implements the Levenshtein Distance Metric to get the closest NCBI taxon name and thus its corresponding NCBI Taxonomy Id; a relatively high similarity score is used (90) to avoid false positives. Also, using the nodes dump file of NCBI Taxonomy, `microbetag` may retrieve the child taxa of a taxon in user's data, along with their corresponding NCBI Taxonomy Ids, if requested by the user. If the user provides their abundance table with taxonomies already mapped to the GTDB taxonomy, `microbetag` will report the best possible annotations in a time efficient manner.

Network inference

When a co-occurrence network is not provided by the user, `microbetag` exploits FlashWeave [14] to build one on the fly. Yet, `microbetag` supports the annotation of networks built from any algorithm/software, in any format Cytoscape can load.

`microbetag` pre-processing

In order to aid the user to map their sequences to the GTDB taxonomy, DADA2-formatted 16S rRNA gene sequences for both bacteria and archaea [44] were used to trained the TAXID classifier of the DECIPHER package [45] and are available through

[3] This should include a description of the overall architecture of the software implementation, along with details of any critical issues and how they were addressed.

the microbetag preprocess Docker image. Likewise, when the abundance table consists of more than 1,000 taxa, providing a network as an input is mandatory. Again, to help the user, `microbetag` preprocess Docker image supports the inference of a network using FlashWeave.

Literature based nodes annotation

Using a set of Tara Ocean samples [46] FAPROTAX [47] estimates the functional potential of the bacterial and archaeal communities, by classifying each taxonomic unit into functional group(s) based on current literature, descriptions of cultured representatives and/or manuals of systematic microbiology. In this manually curated approach, a taxon is associated with a function if and only if all the cultured species within the taxon have been shown to exhibit that function. In its current version, FAPROTAX includes more than 80 functions based on 7600 functional annotations and covering more than 4600 taxa. Contrary to gene content based approaches, e.g. PICRUSt2 [48], FAPROTAX estimates metabolic phenotypes based on experimental evidence.

`microbetag` invokes the accompanying script of FAPROTAX and converts the taxonomic microbial community profile of the samples included in the user's abundance table or of the taxa present in the provided network, into putative functional profiles. Then, it parses FAPROTAX's subtables to annotate each taxonomic unit present in the user's data with all the functions for which they had a hit. FAPROTAX annotations are not part of the microbetagDB but are computed on the fly.

Genomic based nodes annotation

phenDB [49] is a publicly available resource that supports the analysis of bacterial (meta)genomes to identify 47 distinct functional traits, e.g. whether a species is producing butanol or has an halophilic lifestyle. It relies on support vector machines (SVM) trained with manually curated datasets based on gene presence/absence patterns for trait prediction. More specifically, the model for a particular trait is trained using a collection of EggNOG annotated genomes where the knowledge of whether that trait is present or absent among its members is available. These models (classifiers) are used to predict presence/absence of their corresponding traits in non-studied species.

In the framework of microbetagDB, classifiers were re-trained using the genomes provided by phenDB for each trait to sync with the latest version of eggNOG [50] and the `phenotrex` [49] software tool. Genomes were downloaded from NCBI using the Batch Entrez program. Then, *genotype* files were produced for all the high quality GTDB representative genomes. Each model was then used against all the GTDB *genotype* files to annotate each with the presence or the absence of the trait. A list of all the phenotypic traits available for the genomes present in microbetagDB is available on `microbetag` 's documentation site. The updated models are also available

Pathway complementarity

To infer potential pathway complementarities we consider the modules described in KEGG MODULES database [38]. A KEGG module is defined as a functional unit

within the KEGG framework that represents a set of enzymes and reactions involved in a specific biological process or pathway [51]. Such a unit consists of several *steps*, each of which may have more than one molecular ways to occur (Figure 1). A module's definition is a logical expression and consists of KOs that may be coupled with one another as: a. connected steps of the pathway b. parts of a molecular complex, c. alternatives of the same step, and d. optional entities of a complex. Both (a) and (b) cases should be considered as the `AND` logical operator, while (c) would be the `OR` (Figure 1). Given a module's definition, we will consider as an *alternative* any subset of the KO terms mentioned in the definition, that has exactly one way to perform each step, provided that all the steps of the module are covered. We define a genome as having a *complete* module, if and only if all of the KOs of at least one alternative are present on the genome.

Within this framework, `kofamscan` [52] was used to annotate with KEGG ORTHOLOGY terms (KOs) the 16,900 high quality GTDB representative genomes for which a `.faa` was available [53]. The KOs of each genome were then mapped to their corresponding KEGG modules; a KO may map to more than one modules ($1 : n$).

All module definitions were retrieved using the KEGG API and parsed to enumerate their alternatives. Each pair of the KEGG annotated genomes was then investigated for potential pathway complementarities, i.e. whether a genome lacking a number of KOs ($genome_A$) to have a complete module ($module_x$) could benefit from another's species genome(s) ($genome_B$). In that case, $genome_B$ does not necessarily have a complete alternative of $module_x$; as long as it has the missing KOs that $genome_A$ needs to complete an alternative of it, $genome_B$ potentially complements $genome_A$ with respect to $module_x$. In total, 341,568 unique complementarities were exported.

Thanks to the graphical user interface (GUI) of the KEGG pathway map viewer [54, 55], each complementarity can be visualised as part of the closest KEGG metabolic map; where the KOs contributed by the donor are shown in blue-green whereas those coming from the beneficiary genome are coloured in rose.

`microbetag` annotates the edges of a co-occurrence network by identifying pairs where both taxa map to an annotated genome present on microbetagDB. Since co-occurrence networks are undirected, both nodes of a suggested association are considered as potential donors and beneficiary species. When more than one GTDB representative genome map to the same NCBI Taxonomy Id all the possible genome combinations are considered. Finally, two edges are added in such pairs of taxa in the annotated network: one considering $species_A$ as the potential beneficiary and $species_B$ as the potenial donor species, and one vice-versa.

Seed scores and complements using genome scale metabolic reconstructions

The Metabolic Complementarity Index ($MI_{Complementarity}$) measures the degree to which two microbial species can mutually assist each other by complementing each other's biosynthetic capabilities. As described in [56], it is defined as the proportion of seed compounds of a species that can be synthesized by the metabolic network of another, but are not included in the seed set of the latter. $MI_{Complementarity}$ offers

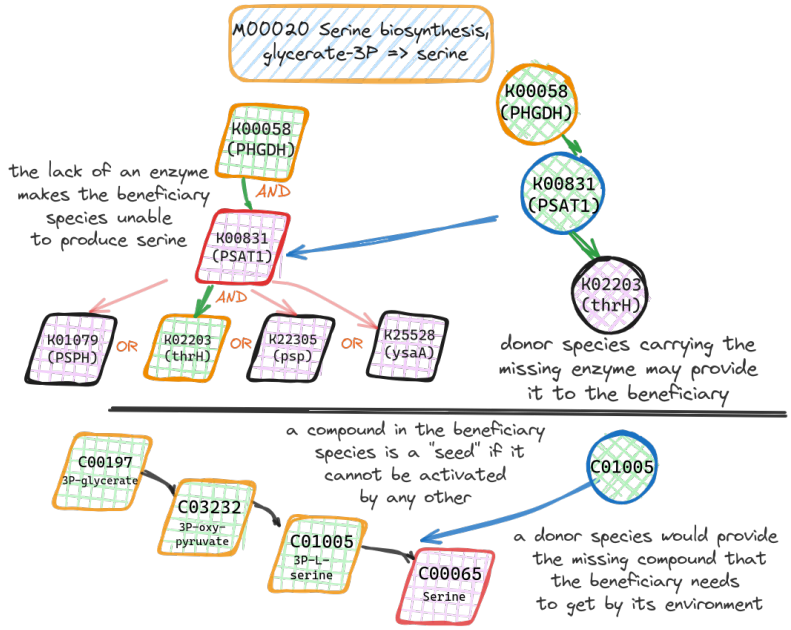

Fig. 1: Pathway complementarity approach. The high quality GTDB genomes were annotated with KEGG ORTHOLOGY (KO) terms. The various ways of getting a KEGG module complete were enumerated and all the possible ways a donor species could "fill" a beneficiary's non-complete module were calculated. In this case, there are 4 unique ways for having the serine biosynthesis module complete; in all of them K00831 is required. However, it is missing from the beneficiary species that supports the 2 out of the 3 steps of the module's definition. A donor species having and potentially sharing the corresponding enzyme of K00831 may enable the beneficiary species to produce serine.

an upper bound assessment of the potential for syntrophic interactions between two species. Further, the Metabolic Competition Index ($MI_{Competition}$) represents the similarity in two species' nutritional profiles. This index establishes an upper limit on the level of competition that one species may face from another. Those indices have been thoroughly described and implemented in the NetCooperate [28] and NetCompt [29] tools correspondingly. We will be referring to those two indices as "seed scores". Recently, the `PhyloMint`tool [56] was released supporting the calculation of the seed scores of GENREs in SBML format.

In the framework of `microbetag` , seed scores were computed using GENREs derived from the high quality GTDB representative genomes and the `PhyloMint` tool. GENREs were reconstructed using the Model SEED pipeline [57] through its Python interface ModelSEEDpy. The latter requires RAST annotated genomes [58]; if available through the PATRIC database [40], annotations were retrieved. For the rest of the genomes, RAST annotation was performed through RASTtk [59].

Moreover, the computed seed and the non-seed (i.e., set of metabolic compounds a genome can build on its own) sets of each genome were used to compute their overlap among all the pairwise combinations of those genomes. More specifically, seed and non-seed compounds of each genome were mapped to their corresponding KO terms and those related to any KEGG MODULE were considered. The latter were then used for the calculation of the overlap of $seed\ set_{species_A}$ with the $non\ seed\ set_{species_B}$ was retrieved. `microbetag` then annotates again the edges of the co-occurrence network where both taxa have been mapped to a at least one GTDB genome, mentioning all the KEGG maps for which there is at least one seed compound of the potentially beneficiary species

Clustering network

`manta` is a heuristic network clustering algorithm that clusters nodes within weighted networks effectively, leveraging the presence of negative edges and discerning between weak and strong cluster assignments. `microbetag` invokes manta [23] to infer clusters from the microbial network. In case `manta` is performed, the annotated network inherits the layout that `manta` returns.

Groups of annotations

Biologically meaningful groups were described to group phenotypic traits returned from FAPROTAX and phenDB-like annotation steps. The main groups supported are related to: 1. the lifestyle of a species, for example being halophilic or thermophyllic etc., 2. the biogeochemical processes a species metabolic potenial has been found related to, for example Nitrite-oxidizing bacteria (NOB) bacteria and 3. important metabolites a species is suggested to produce, e.g. butanol. Aim of these groups are to facilitate filtering of the taxa present. Enrichemnt analysis for members of such groups (e.g., based on the finfings of a clustering algorithm like `manta`) can be performed through the CytoscapeApp.

The `MGG` CytoscapeApp

`microbetag` is accessible as a CytoscapeApp. The `microbetag` CytoscapeApp (called `MGG`) was built based on the source code of the scVizNet [60]. A visual style was developed to facilitate the distinguish of nodes and edges annotated. `MGG` allows the user to import their data, retrieve an annotated network and investigate the annotations through a series of CyPanels both for node and edge annotations. Figure 3 shows an example of the edges CyPanel.

Dependencies, Web server and API

The `microbetag` web service is container - based and consists of three Docker [61] (v24.0.2) images: a. the MySQL database b. an nginx [62] web server and c. the app itself. The latter uses Gunicorn (20.1.0) to build an application server which communicates with the web server using the Web Server Gateway Interface (WSGI) protocol

and handles incoming HTTP requests. `microbetag` is implemented as a Flask application (v2.3.2); Flask is a micro web framework for developing Python web applications and RESTful APIs. A thorough description of `microbetag` 's API is available at the ReadTheDocs web site. The source code of the `microbetag` web service is available on GitHub.

Results [4]

`microbetag` and microbetagDB

`microbetag` is

Table 1: Summary of the data in microbetagDB

Description	Entries
GTDB representative genomes	34,608
Phen-model-oriented metabolic functions	32
FAPROTAX functions	92
Unique pathway complements	341,568
Pairwise pathway complementarities	184,184,548
GENREs leading	33,755
Seed complements	1,139,400,025
Seed scores	1,105,250,048

annotated network returned in `.cyjs` format

For a computationally efficient way to annotate large networks, a Docker image is provided so the user runs a taxonomy assignment using the IDTAXA algorithm [45] of the DECIPHER R package [63]. A co-occurrence network is also built using FlashWeave [14], as `microbetag` also does.

`microbetag` CytoscapeApp

Overall comment, the CytoscapeApp returns averages and s.d. for example in seed scores. If you want the exact values, go through the API.

The app was based on the StringApp and supported by the NRNB group.

Validation of `microbetag` potential

vitamin dataset [39] supplementary file **??**

Metagenomic or metabarcoding data are often used to predict microbial interactions in complex communities, but these predictions are rarely explored experimentally. Here, we use an organism abundance correlation network to investigate factors that control community organization in mine tailings-derived laboratory microbial consortia grown under dozens of conditions.

[4] Significant advance over previously published software (usually demonstrated by direct comparison with available related software) This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures. This section may be combined with the Discussion section for Software articles.

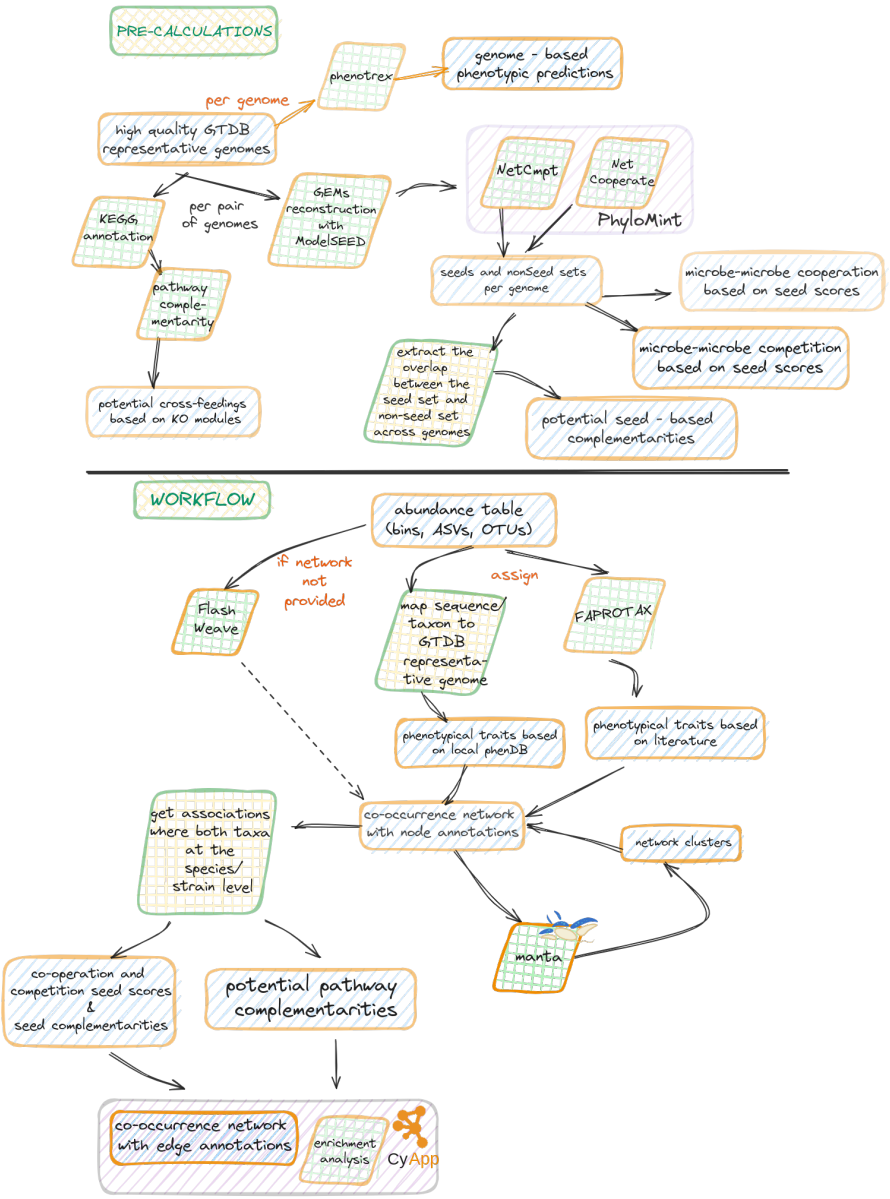

Fig. 2: Diagram of the `microbetag` pre - calculations (top panel) and the on the fly workflow (bottom panel). GTDB v207 representative genomes were filtered and for those of high-quality 33 phenotypic traits were predicted using `phenotrex`. To this end, models were re-trained to sync with recent version of eggNOG.

The network is overlaid with metagenomic information about functional capacities to generate testable hypotheses.

A. GTDB-tk: 480 bins

Step	Time (sec)	Notes
Taxonomy	Cell 1,2	on the fly
Network	Cell 2,2	on the fly
microbetag	Cell 3,2	on the fly
manta clustering	Cell 4,2	

B. GTDB 16S: 3000 ASVs

Step	Time (sec)	Notes
Taxonomy assignment*		
Taxonomy mapping		
Network inference		
microbetag annotations		
manta clustering		

C. Silva:

Step	Time (sec)	Notes
Taxonomy mapping		
Network inference		
microbetag annotations		
manta clustering		

D. fuzzywuzzy:

Step	Time (sec)	Notes
Taxonomy mapping		
Network inference		
microbetag annotations		
manta clustering		

Table 2: Computing times per step using an abundance table of 400 taxa with taxonomy: A. taxonomy scheme B. C. D. *specs of the laptop used. Docker image on HP*

Thiamine alternative pathway [64, 65] [5]

A. *Variovorax* thiamine-related benefits to its neighbors

Neighboring taxon	node id	KEGG compounds	url
Kapabacteria thiocyanatum	bin_59	C15809	url
Terrimonas ferruginea	bin_100	C15809;C01081	url
Tahibacter sp001725155	bin_167	C15809	url
Microbacterium sp900156455	bin_28	C15809; C20246	url
Sphingobium sp001899715	bin_155	Iminoglycine C15809;	url
Nitrosospira sp001899235	bin_176	None	None
62-47 sp001899255*	bin_233	None	None
Bosea sp001898115	bin_273	C04327;C01279	url
54-19 sp001898225**	bin_41	C15809	url
Rhodoglobus sp001725325	bin_8	C15809	url

B. Neighbors thiamine-related benefits among them

Beneficiary	Donor	potential complement
T. ferruginea	Tahibacter	C01081
T. ferruginea	Rhodoglobus	C01081
Nitrosospira	Bosea sp001898115	C04327;C01279
Chloroflexi	Bosea sp001898115	C15809
Chloroflexi	Xanthobacteraceae	C15809
Chloroflexi	Nitrosospira	C15809

Table 3: Computing times per step using an abundance table of 400 taxa with taxonomy: A. taxonomy scheme B.

Regarding pantothenate, the Variovorax genome mapped from GTDB-tk $GCA\_001899795$. brings two complete KEGG modules for that....

a["md:M00119"]
{'0': 'K01918', '1': 'K00826', '2': 'K00077', '3': 'K00606'}
a["md:M00913"]
{'0': 'K01918', '1': 'K00077', '2': 'K00128', '3': 'K00606'}

[5] to understand our findings

thus either some Variovorax species can actually produce that or this is a limitation of our method In fact, *Variovorax* seems to be able to benefit some partners of their.

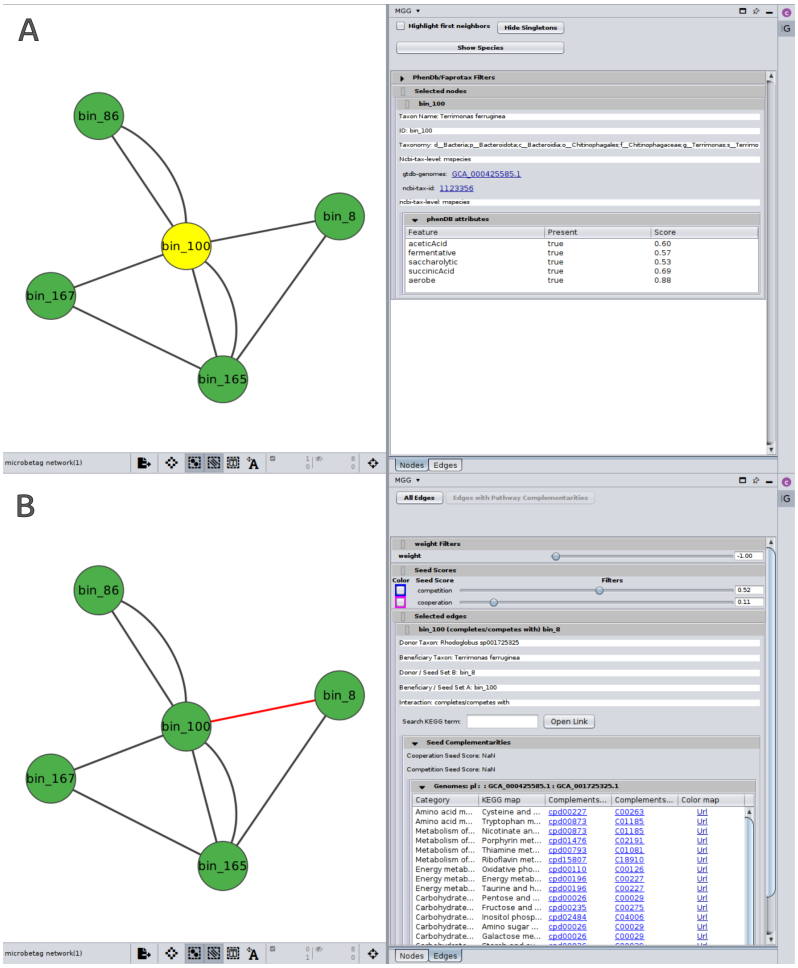

Fig. 3: CyPanels of the MGG CytoscapeApp.

From the hub species:

Table 4: *Variovorax* genomes present on microbetagDB and their corresponding complete/incomplete presence of the pantothenate-related KEGG modules.

Genome	md:M00119	md:M00913
GCA_004210915.1	incomplete	complete
GCA_902506565.1	incomplete	incomplete
GCF_000184745.1	complete	complete
GCF_000282635.1	complete	complete
GCF_000463015.1	complete	complete
GCF_000834655.1	complete	complete
GCF_001424835.1	complete	complete
GCF_001425205.1	complete	complete
GCF_001426505.1	complete	complete
GCF_001577265.1	incomplete	incomplete
GCF_002157355.1	complete	complete
GCF_002754375.1	complete	complete
GCF_003019815.1	incomplete	complete
GCF_003852515.1	complete	complete
GCF_003951285.1	complete	complete
GCF_003952165.1	complete	complete
GCF_003952185.1	complete	complete
GCF_003984625.1	complete	complete
GCF_003984645.1	complete	complete
GCF_006438845.1	complete	complete
GCF_007828835.1	complete	complete
GCF_009498455.1	complete	complete
GCF_009755665.1	complete	complete
GCF_010499245.1	complete	complete
GCF_013376045.1	complete	complete
GCF_014170375.1	complete	complete
GCF_014302995.1	complete	complete
GCF_014303735.1	incomplete	incomplete
GCF_901827175.1	complete	complete
GCF_901827205.1	complete	complete

Discussion [6]

Interpreting a real-world network with `microbetag`

Annelies' dataset.

`microbetag` as a resource

Eco-evolutionary studies Distribution of the complements among taxa // su

[6] The user interface should be described and a discussion of the intended uses of the software, and the benefits that are envisioned, should be included, together with data on how its performance and functionality compare with, and improve, on functionally similar existing software. A case study of the use of the software may be presented. The planned future development of new features, if any, should be mentioned.

Limitations

As shown in [66] (see Figure 6b), the original version of CheckM [67] that is still used on GTDB returns lower completeness scores to genomes that correspond to phyla known for having shorter genomes in general, e.g. Patescibacteria representative genomes on GTDB have an average completeness $\tilde{6}5\%$. `microbetag` inherits this in the filtering process for getting only high quality genomes and thus, only few representatives from these taxonomic groups are present on microbetagDB.

It is well known that higher-order interactions, i.e. interactions involving more than two species [30] Pairwise relationships do not capture more complex forms of ecological interactions, in which one species depends on (or is influenced by) multiple other species. [3]

Future work

Further indices using the seed concept have been also presented such as the metabolic interaction potential (MIP) and the metabolic resource overlap (MRO). MIP is defined as the difference between the minimal number of components required for the growth of all members in a noninteracting community and an interacting community, i.e. the maximum number of essential nutritional components that a community can provide for itself through interspecies metabolic exchanges [30]. Similarly, MRO is defined as the maximum possible overlap between the minimal nutritional requirements of all member species [30]. Regression and association rule mining [68] can be applied to address this challenge.

- pathway and seed complementarities for higher-order interactions
- spatial dimension
- transcriptomics data integration: compare potential complementarities with what is going on
-

Conclusions

[7]

Data integration

Supplementary information. [8]

Declarations

- **Availability of data and materials**
  - Raw sequences for the use case:
  - Raw data for the validations case:

[7] This should state clearly the main conclusions and provide an explanation of the importance and relevance of the case, data, opinion, database or software reported.

[8] If your article has accompanying supplementary file(s) please state so here. E.g. supplementary figures and tables captions.

- **Funding**
  This work was initiated thanks to an EMBO Scientific Exchange Grant to HZ. It was then supported by the 3D'omics Horizon project (101000309). We would also like to thank the National Resource for Network Biology (NRNB) and the Google Summer of Code 2023 for the support of E.I.M.D.
- **Conflict of interest/Competing interests**
  The authors declare that they have no other competing interests.
- **Authors' contributions** [9]
  Conceptualization: K.F. Methodology: K.F. and H.Z. Software: H.Z., E.I.M.D. and J.M Validation: H.Z. and K.F. Formal analysis: H.Z. and K.F. Investigation: H.Z. Resources: K.F., A.E. and A.G. Data Curation: H.Z. Writing - Original Draft: H.Z. and K.F. Writing - Review & Editing: all Visualization: H.Z. Supervision: K.F., H.Z. and S.M. Project administration: K.F. Funding acquisition: K.F., H.Z.
- **Acknowledgements**
  We would like to thank Dr Christina Pavloudi and ++ for the insight on how to organise the trait groups.
- **Ethics approval**
  Not applicable
- **Consent to participate**
  Not applicable.
- **Code availability:**
  - microbetagDB related scripts: https://github.com/hariszaf/microbetag
  - microbetagApp and webserver: https://github.com/msysbio/microbetagApp.
  - CytoscapeApp: https://github.com/ermismd/MGG/
  - Validation and use case: ¡think of having that under the 3D'omics organization¿
  - Documentation web-site: https://hariszaf.github.io/microbetag/

Appendix A Mappings

$n:1$ $n:n$ etc

Appendix B Background on seed scores and complementarities

B.1 Background on seed scores

In that case, once a seed is assured, it activates all the rest of that group. Therefore, a confidence level (C) ranging from 0 to 1, has been previously described to quantify the relevance of each seed:

$$C_i = 1/seed's\ group\ with\ i\ size \tag{B1}$$

$C = 0$ corresponds to a non-seed node, while $C = 1$ represents an independent node.

[9] Based on the CRediT system. Current list is indicative.

$$MI_{Complementarity} = \frac{|SeedSet_A \cap \neg SeedSet_B|}{|SeedSet_A \cap (SeedSet_B \cup \neg SeedSet_B)|} \quad \text{(B2)}$$

As also described in [56], it is calculated as the proportion of compounds in a species' seed set that coincide with those in an other's, while also factoring in the confidence scores associated with seed compounds.

$$MI_{Competition} = \frac{\sum C(SeedSet_A \cap SeedSet_B)}{\sum C(SeedSet_A)} \quad \text{(B3)}$$

B.2 Background on pathway complementarity

For example, the definition of the D-Galacturonate degradation in Bacteria (M00631) is:

K01812 K00041 (K01685,K16849+K16850) K00874 (K01625,K17463)
that once breaking down, it leads to 4 alternative sets of KOs (pathways):

K01812 K00041 K01685 K00874 K01625

K01812 K00041 K16849+K16850 K00874 K01625

K01812 K00041 K01685 K00874 K17463

K01812 K00041 K16849+K16850 K00874 K17463

B.3 Complementarities

KEGG compound ModelSEED compounds ModelSEED compounds mapped to KEGG compounds and kept only those related to KEGG modules.

References

[1] Yuan, M.M., Guo, X., Wu, L., Zhang, Y., Xiao, N., Ning, D., Shi, Z., Zhou, X., Wu, L., Yang, Y., *et al.*: Climate warming enhances microbial network complexity and stability. Nature Climate Change **11**(4), 343–348 (2021)

[2] Raes, J., Bork, P.: Molecular eco-systems biology: towards an understanding of community function. Nature Reviews Microbiology **6**(9), 693–699 (2008)

[3] Faust, K., Raes, J.: Microbial interactions: from networks to models. Nature Reviews Microbiology **10**(8), 538–550 (2012)

[4] Röttjers, L., Faust, K.: From hairballs to hypotheses–biological insights from microbial networks. FEMS microbiology reviews **42**(6), 761–780 (2018)

[5] Bálint, M., Bahram, M., Eren, A.M., Faust, K., Fuhrman, J.A., Lindahl, B., O'Hara, R.B., Öpik, M., Sogin, M.L., Unterseher, M., *et al.*: Millions of reads,

thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS microbiology reviews **40**(5), 686–700 (2016)

[6] Robinson, C.J., Bohannan, B.J., Young, V.B.: From structure to function: the ecology of host-associated microbial communities. Microbiology and Molecular Biology Reviews **74**(3), 453–476 (2010)

[7] D'Souza, G., Shitut, S., Preussger, D., Yousif, G., Waschina, S., Kost, C.: Ecology and evolution of metabolic cross-feeding interactions in bacteria. Natural Product Reports **35**(5), 455–488 (2018)

[8] Finn, R., Balech, B., Burgin, J., Chua, P., Corre, E., Cox, C., Donati, C., Santos, V., Fosso, B., Hancock, J., Heil, K., Ishaque, N., Kale, V., Kunath, B., Médigue, C., Pafilis, E., Pesole, G., Richardson, L., Santamaria, M., Van Den Bossche, T., Vizcaíno, J., Zafeiropoulos, H., Willassen, N., Pelletier, E., Batut, B.: Establishing the elixir microbiome community [version 1; peer review: awaiting peer review]. F1000Research **13**(50) (2024) https://doi.org/10.12688/f1000research.144515.1

[9] Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., *et al.*: A new view of the tree of life. Nature microbiology **1**(5), 1–6 (2016)

[10] Matchado, M.S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D., List, M.: Network analysis methods for studying microbial communities: A mini review. Computational and structural biotechnology journal **19**, 2687–2698 (2021)

[11] Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.: Microbial co-occurrence relationships in the human microbiome. PLoS computational biology **8**(7), 1002606 (2012)

[12] Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data. PLoS computational biology **8**(9), 1002687 (2012)

[13] Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology **11**(5), 1004226 (2015)

[14] Tackmann, J., Rodrigues, J.F.M., Mering, C.: Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. Cell systems **9**(3), 286–296 (2019)

[15] Faust, K.: Open challenges for microbial network construction and analysis. The ISME Journal **15**(11), 3111–3118 (2021)

[16] Cao, H.-T., Gibson, T.E., Bashan, A., Liu, Y.-Y.: Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. BioEssays

39(2), 1600188 (2017)

[17] Kishore, D., Birzu, G., Hu, Z., DeLisi, C., Korolev, K.S., Segrè, D.: Inferring microbial co-occurrence networks from amplicon data: a systematic evaluation. Msystems, 00961–22 (2023)

[18] Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., *et al.*: Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME journal **10**(7), 1669–1681 (2016)

[19] Berry, D., Widder, S.: Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Frontiers in microbiology **5**, 219 (2014)

[20] Guo, B., Zhang, L., Sun, H., Gao, M., Yu, N., Zhang, Q., Mou, A., Liu, Y.: Microbial co-occurrence network topological properties link with reactor parameters and reveal importance of low-abundance genera. npj Biofilms and Microbiomes **8**(1), 3 (2022)

[21] Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J.A., Faust, K., Knight, R., Jansson, J.K., Cardona, C., *et al.*: Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. Microbiome **8**, 1–12 (2020)

[22] Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J.R., *et al.*: Plankton networks driving carbon export in the oligotrophic ocean. Nature **532**(7600), 465–470 (2016)

[23] Röttjers, L., Faust, K.: Manta: A clustering algorithm for weighted ecological networks. Msystems **5**(1), 10–1128 (2020)

[24] Levy, R., Borenstein, E.: Reverse ecology: from systems to environments and back. In: Evolutionary Systems Biology, pp. 329–345. Springer, ??? (2012)

[25] Levy, R., Borenstein, E.: Metagenomic systems biology and metabolic modeling of the human microbiome: From species composition to community assembly rules. Gut Microbes **5**(2), 265–270 (2014)

[26] Borenstein, E., Kupiec, M., Feldman, M.W., Ruppin, E.: Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences **105**(38), 14482–14487 (2008)

[27] Parter, M., Kashtan, N., Alon, U.: Environmental variability and modularity of bacterial metabolic networks. BMC evolutionary biology **7**, 1–8 (2007)

[28] Levy, R., Carr, R., Kreimer, A., Freilich, S., Borenstein, E.: Netcooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. BMC bioinformatics **16**(1), 1–6 (2015)

[29] Kreimer, A., Doron-Faigenboim, A., Borenstein, E., Freilich, S.: Netcmpt: a network-based tool for calculating the metabolic competition between bacterial species. Bioinformatics **28**(16), 2195–2197 (2012)

[30] Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D.R., Bork, P., Patil, K.R.: Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proceedings of the National Academy of Sciences **112**(20), 6449–6454 (2015)

[31] Belcour, A., Frioux, C., Aite, M., Bretaudeau, A., Hildebrand, F., Siegel, A.: Metage2metabo, microbiota-scale metabolic complementarity for the identification of key species. Elife **9**, 61968 (2020)

[32] Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols **5**(1), 93–121 (2010)

[33] Durot, M., Bourguignon, P.-Y., Schachter, V.: Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS microbiology reviews **33**(1), 164–190 (2008)

[34] Cerk, K., Ugalde-Salas, P., Nedjad, C.G., Lecomte, M., Muller, C., Sherman, D.J., Hildebrand, F., Labarthe, S., Frioux, C.: Community-scale models of microbiomes: Articulating metabolic modelling and metagenome sequencing. Microbial Biotechnology **n/a**(n/a), 14396 https://doi.org/10.1111/1751-7915.14396 https://ami-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/1751-7915.14396. e14396 MICROBIO-2023-392.R1

[35] Little, A.E., Robinson, C.J., Peterson, S.B., Raffa, K.F., Handelsman, J.: Rules of engagement: interspecies interactions that regulate microbial communities. Annu. Rev. Microbiol. **62**, 375–401 (2008)

[36] Zientz, E., Dandekar, T., Gross, R.: Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiology and Molecular Biology Reviews **68**(4), 745–770 (2004)

[37] Kallus, Y., Miller, J.H., Libby, E.: Paradoxes in leaky microbial trade. Nature communications **8**(1), 1361 (2017)

[38] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., Kanehisa, M.: Modular architecture of metabolic pathways revealed by conserved sequences of reactions. Journal of Chemical Information and Modeling **53**(3), 613–622 (2013) https://doi.org/10.1021/ci3005379 https://doi.org/10.1021/ci3005379. PMID: 23384306

[39] Hessler, T., Huddy, R.J., Sachdeva, R., Lei, S., Harrison, S.T., Diamond, S., Banfield, J.F.: Vitamin interdependencies predicted by metagenomics-informed network analyses and validated in microbial community microcosms. Nature

Communications **14**(1), 4768 (2023)

[40] Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E.M., Disz, T., Gabbard, J.L., *et al.*: Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. Nucleic acids research **45**(D1), 535–542 (2017)

[41] Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P.: Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic acids research **50**(D1), 785–794 (2022)

[42] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The silva ribosomal rna gene database project: improved data processing and web-based tools. Nucleic acids research **41**(D1), 590–596 (2012)

[43] Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., *et al.*: Ncbi taxonomy: a comprehensive update on curation, resources and tools. Database **2020**, 062 (2020)

[44] Alishum, A.: DADA2 Formatted 16S rRNA Gene Sequences for Both Bacteria & Archaea. https://doi.org/10.5281/zenodo.6655692 . https://doi.org/10.5281/zenodo.6655692

[45] Murali, A., Bhargava, A., Wright, E.S.: Idtaxa: a novel approach for accurate taxonomic classification of microbiome sequences. Microbiome **6**(1), 1–14 (2018)

[46] Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., *et al.*: Structure and function of the global ocean microbiome. Science **348**(6237), 1261359 (2015)

[47] Louca, S., Parfrey, L.W., Doebeli, M.: Decoupling function and taxonomy in the global ocean microbiome. Science **353**(6305), 1272–1277 (2016)

[48] Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., Langille, M.G.: Picrust2 for prediction of metagenome functions. Nature biotechnology **38**(6), 685–688 (2020)

[49] Feldbauer, R., Schulz, F., Horn, M., Rattei, T.: Prediction of microbial phenotypes based on comparative genomics. BMC bioinformatics **16**(14), 1–8 (2015)

[50] Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., *et al.*: eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research **47**(D1), 309–314 (2019)

[51] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., Kanehisa, M.: Modular architecture of metabolic pathways revealed by conserved sequences of reactions. Journal of chemical information and modeling **53**(3), 613–622 (2013)

[52] Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., Ogata, H.: Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. Bioinformatics **36**(7), 2251–2252 (2020)

[53] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: Kegg for integration and interpretation of large-scale molecular data sets. Nucleic acids research **40**(D1), 109–114 (2012)

[54] Kanehisa, M., Sato, Y.: Kegg mapper for inferring cellular functions from protein sequences. Protein Science **29**(1), 28–35 (2020)

[55] Kanehisa, M., Sato, Y., Kawashima, M.: Kegg mapping tools for uncovering hidden features in biological data. Protein Science **31**(1), 47–53 (2022)

[56] Lam, T.J., Stamboulian, M., Han, W., Ye, Y.: Model-based and phylogenetically adjusted quantification of metabolic interaction between microbial species. PLoS computational biology **16**(10), 1007951 (2020)

[57] Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., Stevens, R.L.: High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology **28**(9), 977–982 (2010)

[58] Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., *et al.*: The seed and the rapid annotation of microbial genomes using subsystems technology (rast). Nucleic acids research **42**(D1), 206–214 (2014)

[59] Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., *et al.*: Rasttk: a modular and extensible implementation of the rast algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports **5**(1), 1–6 (2015)

[60] Choudhary, K., Meng, E.C., Diaz-Mejia, J.J., Bader, G.D., Pico, A.R., Morris, J.H.: scnetviz: from single cells to networks using cytoscape. F1000Research **10** (2021)

[61] Merkel, D., *et al.*: Docker: lightweight linux containers for consistent development and deployment. Linux j **239**(2), 2 (2014)

[62] Reese, W.: Nginx: The high-performance web server and reverse proxy. Linux J. **2008**(173) (2008)

[63] Wright, E.S.: Using decipher v2. 0 to analyze big biological sequence data in r. R

Journal **8**(1) (2016)

[64] Llavero-Pasquina, M., Geisler, K., Holzer, A., Mehrshahi, P., Mendoza-Ochoa, G.I., Newsad, S.A., Davey, M.P., Smith, A.G.: Thiamine metabolism genes in diatoms are not regulated by thiamine despite the presence of predicted riboswitches. New Phytologist **235**(5), 1853–1867 (2022)

[65] Romine, M.F., Rodionov, D.A., Maezato, Y., Osterman, A.L., Nelson, W.C.: Underlying mechanisms for syntrophic metabolism of essential enzyme cofactors in microbial communities. The ISME journal **11**(6), 1434–1446 (2017)

[66] Chklovski, A., Parks, D.H., Woodcroft, B.J., Tyson, G.W.: Checkm2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nature Methods **20**(8), 1203–1212 (2023)

[67] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W.: Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research **25**(7), 1043–1055 (2015)

[68] Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. SIGMOD '93, pp. 207–216. Association for Computing Machinery, New York, NY, USA (1993). https://doi.org/10.1145/170035.170072 . https://doi-org.kuleuven.e-bronnen.be/10.1145/170035.170072