

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

Haris Zafeiropoulos¹, Ermis Ioannis Michail Delopoulos¹, Andi Erega², Annelies Geirnaert², John Morris³, 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

Background: Co-occurrence networks are widely used for deducing microbial associations from metagenomic data, yet spurious associations and tool dependencies impede their accuracy. “Reverse ecology” utilizes genomic and metabolic modeling to predict ecological traits for less-understood microorganisms and their interactions within microbial communities. **microbetag**, a microbial co-occurrence network annotator, employs reverse ecology to enhance/reduce confidence in network associations.

Results: microbetagDB consists of 34,608 high-quality Genome Taxonomy DataBase genomes, along with their corresponding KEGG-annotations and genome-scale metabolic reconstructions. Seed sets (externally required compounds) were retrieved from the latter, supporting the prediction of over 1 billion potential pairwise metabolic interactions. Relying on KEGG MODULE definitions, **342,000** unique pathway complements were exported based on pairwise

overlaps between the annotated genomes. Predictions for 33 phenotypic traits (e.g., pH optima, oxygen tolerance etc.) are also incorporated in microbetagDB based on annotated genomes. **microbetag** maps node taxonomies to their closest genome in microbetagDB and annotates them with phenotypic traits, coupled with a literature-based assessment of functional potential. Edges are then annotated by screening the pairwise associations over the pathway complements and the seed complementarities.

microbetag successfully identified potential metabolic interactions between *Variovorax* and neighboring species, enabling the latter to produce thiamine in a network derived from a mining waste microbial consortium, as the authors independently verified. Accessible as a CytoscapeApp, **microbetag** offers a user-friendly platform called microbetag GUI (MGG), that facilitates the prompt transition from an abundance table to a fully annotated co-occurrence network, providing a quick and intuitive mean to yielding explore hypotheses about the microbial interactions.

Conclusions: microbetagDB provides insight on the range of potential complementarities and their distribution among prokaryotes and benefits eco-evolutionary studies.

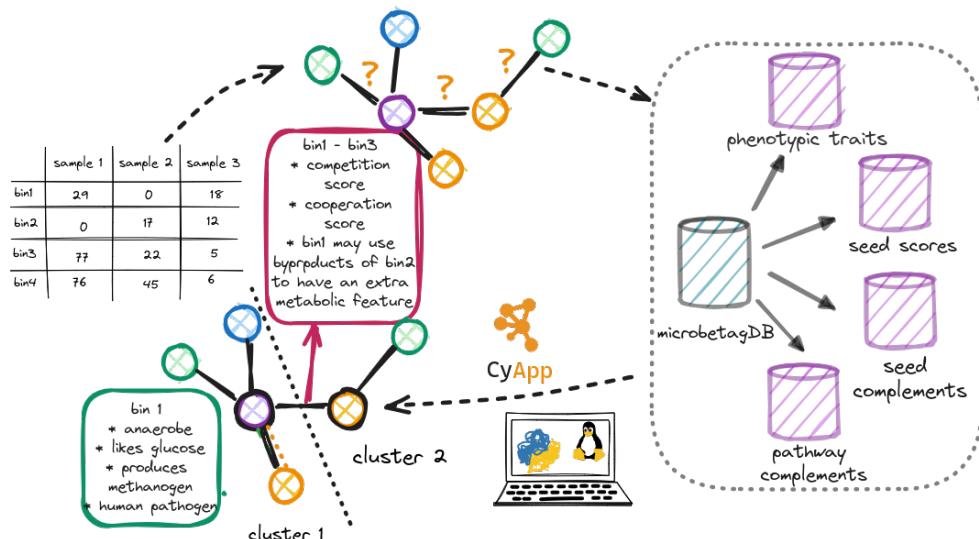


Figure abstract.

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

Background

Microbial interactions play 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.

Depending on the net fitness effects that result for the taxa involved, interactions range from cooperation, competition, parasitism, commensalism and amensalism [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 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 [6].

High-throughput sequencing (HTS) has provided insight into the diversity and composition of microbial communities [7]. Uncultivated species can now be detected, and their features can be inferred through their genomic information [8]. 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 extract such patterns is the creation of co-occurrence networks based on metagenomic read data (amplicon and/or shotgun) [9]. Several approaches are available for co-occurrence network inference based on a range of statistical concepts such as: correlation (e.g., CoNet [10], SparCC [11]), linear regression (e.g., SpecEasi [12]) and indirect causal inference (FlashWeave [13]). Nevertheless, microbial co-occurrence networks continue to encounter various challenges [14]. Each inference approach comes with its own assumptions and parameterization, leading to variations in network structure. The choice of correlation metrics, the data preprocessing techniques applied and the handling of sparsity and zero inflation, and considerations for indirect causal relationships all influence the resulting network [15]. As a consequence, the result of network construction is tool-dependent [16, 17]. Moreover, their inference inherits the challenges of metagenomic data and analysis (e.g., sampling scale, compositionality, parameters tuning) and the returned network is a "hairball" of densely interconnected taxa. Thus, additional analysis is necessary to generate testable hypotheses [14]. Addressing the question of *What can we learn from the hairballs* posed by Röttjers et al. [4] could provide insight into the mechanisms of the interactions.

The assessment of interaction predictions derived from microbial co-occurrence networks has underscored their low accuracy for this task [18]. Data integration and clustering have been suggested to address this challenge [14]. Clusters identified in microbial association networks have demonstrated their ability to mirror key drivers of community composition [19] and several algorithms and implementations are available [20]. 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 or reduce the confidence of the associations suggested by the network and generate hypotheses 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 [21]. Reverse ecology leverages genomics to explore community ecology with no *a priori* assumptions about the taxa involved. 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 [22].

Metabolic networks are comprehensive representations of biochemical reactions occurring within an organism's cells or a biological system, while genome-scale metabolic network reconstructions (GENREs) are metabolic networks capable of encapsulating the biochemical reactions that could take place within an organism in a mathematical representation [23–25]. Over the last decade, (semi-) automated approaches support such reconstructions in a time-efficient way, such as in [26, 27] and more. A metabolic network's "seed set" is the set of compounds that, based on the network topology, cannot be produced by the organism and need to be acquired exogenously [28] (see Figure 2). Such seeds might be independent, i.e. they cannot be produced by any other biochemical reaction in the metabolic network, or they can be interdependent forming groups of seed compounds. Seeds are a useful proxy for the essential nutrients of an organism [28, 29]. Based on the seed concept, several graph theory-based metrics (indices) have been described to predict species interactions directly from their metabolic networks' topologies [30–33].

Metabolic complementarity among species reflects the potential for cooperation through cross-feeding. 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 compound sets can be used to compute complementarity and overlap indices. The examination of such indices can reveal metabolic interactions leading to patterns observed in a co-occurrence network.

However, such indices consider a species genomic repertoire as a bag of genes and thus, they cannot resolve how exactly a species could benefit another. To explore whether a species may benefit from a partner, it is helpful to move from the overall network to the pathway level and check whether their pathways complement each other. 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 [34].

microbetag annotates a user's co-occurrence network by integrating phenotypic traits of the taxa present in the network (nodes) and by mapping potential metabolic interactions onto their associations (edges). **microbetag** includes a Graphical User Interface (GUI) implemented 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 directly. To validate our annotations, we used a recently published network with partially known interactions between some pairs of species found to be associated [35] (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

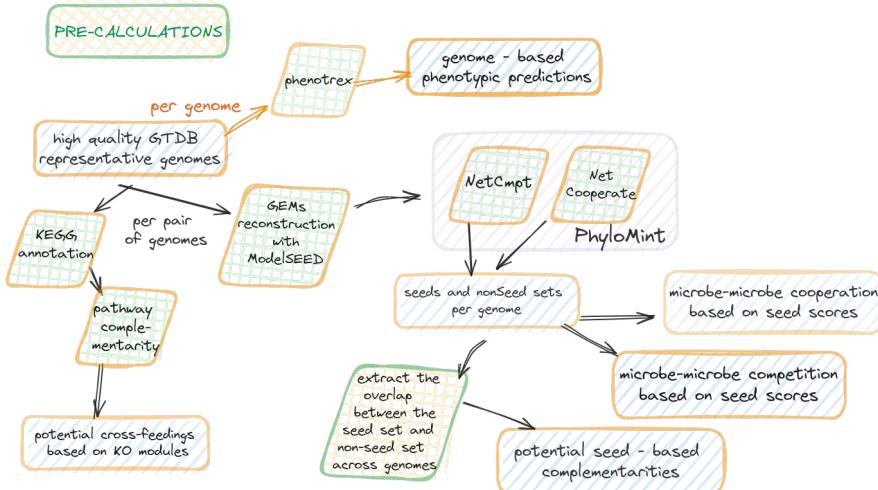


Fig. 1: Diagram of the `microbetag` pre-calculations. GTDB v207 representative genomes were filtered and for those of high-quality, 33 phenotypic traits were predicted using `phenotrex`. To this end, models were re-trained to sync with version 5.0 of eggNOG.

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 between pairs of genomes (see paragraph 3). Last, when available, their corresponding annotations in the PATRIC database [36] 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 [37] that is being broadly used for bins and/or MAGs taxonomic classification and the Silva database [38] that is widely used in amplicon studies. Both taxonomy schemes link their taxonomies to NCBI TaxonomyIds [39]. In case neither of those two taxonomies was used, and the abundance table contains less than 1,000 taxa, `microbetag` maps the user provided taxonomies to the 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 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` relies on FlashWeave [13] 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

To aid the user to map their sequences to the GTDB taxonomy, DADA2-formatted 16S rRNA gene sequences for both bacteria and archaea [40] were used to train the IDTAXA classifier of the DECIPHER package [41] and are available through 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, the `microbetag` preprocess Docker image supports the inference of a network using FlashWeave.

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 [41] of the DECIPHER R package [42]. A co-occurrence network is also built using FlashWeave [13], as `microbetag` also does.

Literature based nodes annotation

Using a set of Tara Ocean samples [43] FAPROTAX [44] estimates the functional potential of the bacterial and archaeal communities, by classifying each taxonomic unit into functional group(s) based on the current literature, descriptions of cultured representatives and/or manuals of systematic microbiology. In this manually curated approach, a taxon is associated with a function 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 covers more than 4600 taxa. Contrary to gene content based approaches, e.g., PICRUSt2 [45], 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 sub-tables 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 [46] 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 a 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.

For microbetagDB, classifiers were re-trained using the genomes provided by phenDB for each trait to sync with the latest version of eggNOG [47] and the `phenotrex` [46] 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 [34]. 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 [48]. Such a unit consists of several *steps*, each of which may have more than one molecular way to occur (Figure 2). 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 2). 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, only if all the KOs of at least one alternative are present on it. In Appendix A we show an example of a module along with its alternatives.

Within this framework, `kofamscan` [49] was used to annotate with KEGG ORTHOLOGY terms (KOs) the 16,900 high quality GTDB representative genomes for which a `.faa` was available [50]. The KOs of each genome were then mapped to their corresponding KEGG modules; a KO may map to more than one module ($1 : n$).

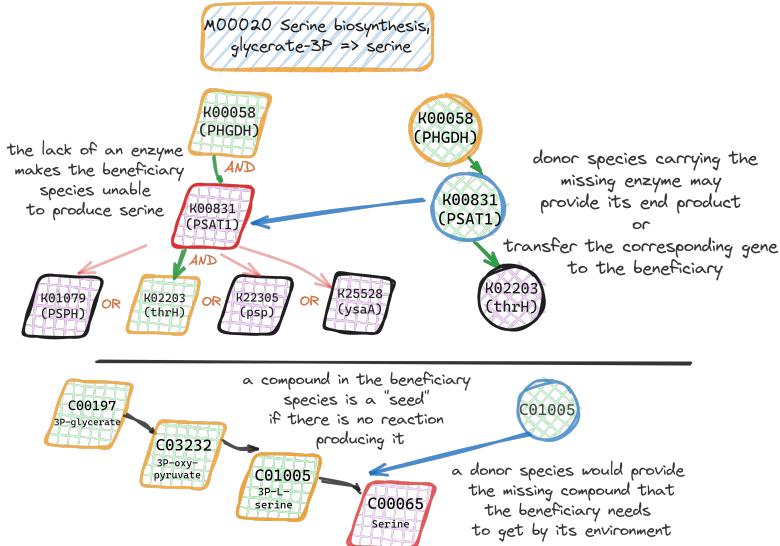


Fig. 2: Pathway complementarity approach. The high-quality GTDB genomes were annotated with KEGG ORTHOLOGY (KO) terms. The various ways of completing a KEGG module were enumerated and all the possible ways a donor species could "gap-fill" a beneficiary's non-complete module were calculated. In this case, there are 4 unique ways to complete the serine biosynthesis module; in all of them K00831 is required. However, it is missing from the beneficiary species that supports two out of the three steps of the module's definition. A donor species including the missing KO (K00831) could potentially transfer the respective gene or provide the beneficiary species with the relevant intermediary product, enabling the beneficiary species to produce serine.

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](#) [51, 52], 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 red.

`microbetag` annotates the edges of a co-occurrence network by identifying pairs where both taxa map to an annotated genome present in `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 potential 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 [53], 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 an upper bound assessment of the potential for cross-feeding 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 described and implemented in the NetCooperate [30] and NetCompt [31] tools respectively. We will be referring to those two indices as "seed scores". Recently, the PhyloMint tool [53] was released supporting the calculation of the seed scores of GENREs in SBML format.

In the `microbetag` framework, 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 [54] through its Python interface `ModelSEEDpy`. The latter requires RAST annotated genomes [55]; if available through the PATRIC database [36], annotations were retrieved. For the rest of the genomes, RAST annotation was performed through RASTtk [56].

Moreover, the computed seed and the non-seed (i.e., 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 further. Focusing on the KEGG MODULE - related KO terms as terms of interest, the overlap of $seed\ set_{species_A}$ with the $non\ seed\ set_{species_B}$ was retrieved. Such $seed\ complementarities$ were calculated for all pairwise GENREs and are now available through `microbetagDB`. Edges of the co-occurrence network where both taxa have been mapped to at least one GTDB genome can be further annotated 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` [20] to cluster the microbial network. In case `manta` is performed, the annotated network inherits the layout that `manta` returns.

The `microbetag` workflow

As shown in Figure 3, the `microbetag` workflow expects an abundance table representing either amplicon or shotgun data. If a co-occurrence network is already available, the user may provide it too as input. The `microbetag` workflow will first map the taxa present on the abundance table to their corresponding GTDB representative genomes if that is possible, i.e., in case the taxonomy provided does reach the species or the strain level (see paragraph 3). If a network is not provided, `microbetag` will then build one using FlashWeave [13]. Then the abundance table will be used for a literature-based annotation using FAPROTAX [44]. This is the only annotation step that is `microbetag`DB independent within the web-service workflow. The nodes of the network will be further annotated with phenotypic traits based on the model predictions [46]. Edges linking taxa assigned to the species or strain level will be annotated with pathway and seed complementarities and seed scores. Last, a network clustering will be performed assigning each node to a cluster. The annotated network is then returned in a .cx format. The user may skip any of these annotation steps if not needed for their analysis.

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: a. the lifestyle of a species, for example being halophilic or thermophilic etc., b. the biogeochemical processes a species metabolic potential has been found related to, for example Nitrite-oxidizing bacteria (NOB) bacteria and c. important metabolites a species is suggested to produce, e.g. butanol. The aim of these groups are to facilitate filtering of the taxa present. Enrichment analysis for members of such groups (e.g., based on the clusters identified by an algorithm like `manta`) can be performed through the CytoscapeApp.

Software architecture

`microbetag` is a Docker-based application. We deployed the `microbetag` application using Docker containers [57] (v24.0.2) managed by Docker Compose (see Supplementary Figure B1). Docker Compose is a tool for defining and running multi-container Docker applications using a YAML file to configure the services required for the application. Containers of three Docker images are being used simultaneously: a. a MySQL database including the `microbetag`DB b. a nginx [58] web server and c. the application itself, including the API and the `microbetag` workflow. The latter uses Gunicorn



Fig. 3: Diagram of `microbetag`'s on-the-fly workflow. `microbetag` expects either an abundance table only as input and infers a co-occurrence network using Flash-Weave or an abundance table along with an already inferred co-occurrence network and after mapping taxa present to GTDB reference genomes, for those possible, phenotypic attributes are assigned to the nodes. Literature-based annotations of the nodes are also added using FAPROTAX. On the edge level, `microbetag` assigns the pre-calculated potential complements based on the pathway and the seed complementarities approaches. `microbetag` supports optional network clustering with manta. The annotated network can then be parsed into Cytoscape using the MGG app.

(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. The API has a route for performing the `microbetag` workflow, either through any Python console or the Cytoscape MGG app, but also several other routes that enable quick and easy access to the `microbetag`DB content, i.e. the genomes present, their phenotypic traits predicted annotations, pathway and seed complementarities among specific

genomes or NCBI Taxonomy Ids and their corresponding seed scores if available. A thorough description of the `microbetag` API is available at the [ReadTheDocs web site](#). The source code of the `microbetag` web service is available on [GitHub](#).

The MGG CytoscapeApp

We developed a Cytoscape app to enable a straightforward, user-friendly way to perform the `microbetag` workflow and visualise `microbetag`-annotated networks. The `microbetag` CytoscapeApp (called MGG) was built based on the [source code](#) of the scVizNet [59]. A visual style was developed to facilitate distinguishing annotated nodes and edges. Nodes are colored based on the level of the taxonomic assignment with those being annotated highlighted in green. Similarly, edges are light green when they carry a positive weight and red when negative. Black edges denote pathway and/or seed complementarities. The latter were not accounted for in the edge weights since edges represent an undirected relationship while complementarity/overlap scores assume a direction, i.e. the complementarity score of species A versus B is not necessarily the same as that of B versus A. 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 4 shows an example of the CyPanels. In the node panel (4.A), the node name, the taxonomy as well as the NCBI Taxonomy Id and the GTDB genome to which the sequence was mapped can be viewed. Depending on the user's settings and the available annotations for a node, genomic based predictions may be present and/or literature-based ones. Further, the annotation groups mentioned in paragraph 3 are on top of this panel allowing for the selection of the nodes carrying either one among several attributes (OR logical relationship) or all of them (AND). Accordingly, in the edges panel (4.B), the beneficiary taxon is specified along with their corresponding GTDB representative sequence identifier. Pathway and seed complementarities are shown each in a table. Potential metabolic interactions are shown in a sub-table entitled with the genome pair under consideration, as several GTDB genomes may have been assigned to a node. In case of pathway complementarities, these tables consist of six columns: a. the KEGG MODULE id of the module to be completed, b. its description, c. a more general metabolic category to which the module belongs, d. the complement itself as a list of KEGG terms, e. the alternative that now represents a complete module in the beneficiary and f. a URL that points to a coloured KEGG map highlighting the complement. If clicked, the user's default browser pops up showing a coloured KEGG map as shown in an example in Figure 4.C. Last, MGG allows to test for enrichment or depletion of the phenotypic traits assigned to the nodes in each of the network clusters. Clusters may have been returned from `manta` [20] while performing the `microbetag` workflow or users may assign them on their own or using any other network clustering algorithm. For thorough instructions on how to use MGG and `microbetag` the reader may visit the [ReadTheDocs web site](#).

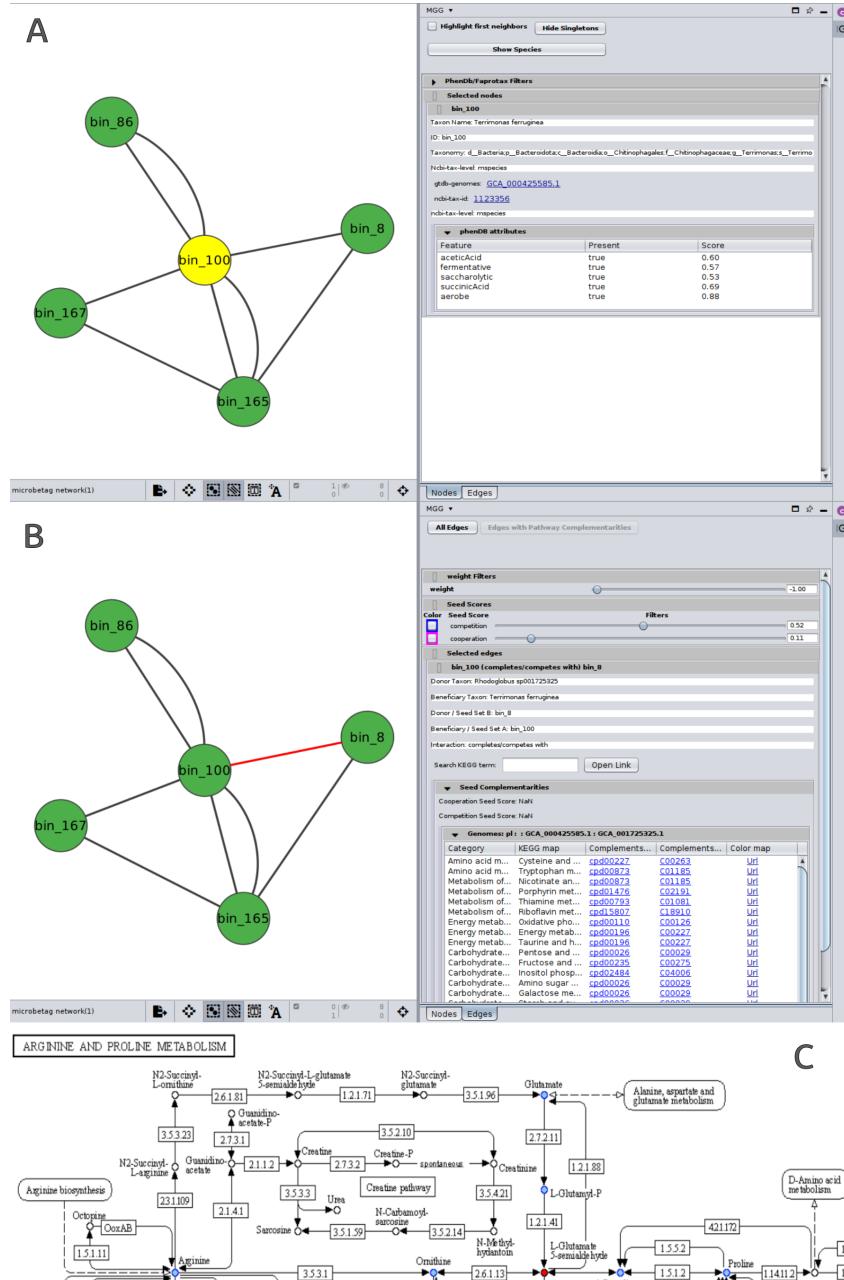


Fig. 4: CyPanels of the MGG CytoscapeApp. A. *Nodes* panel display the annotations of each taxon (node) mapped to one or more GTDB genomes. In this example, genomic predicted phenotypic attributes are shown along with their prediction score. B. *Edges* panel displays the list of potential metabolic complementarities between two nodes, specifying which is the potential donor and the potential beneficiary taxon; thus giving a direction to the corresponding edge in the graph. There are two cases of complementarities in *microbeta*. The first are seed complementarities, which are computed based on the species' reconstructed metabolic networks (case shown above). *Seed complementarities* are first exported based on ModelSEED complements (column three) and mapped in KEGG COMPOUNDS (column four). The provided URL points to a colored KEGG map that highlights seeds complementing a beneficiary's metabolic network. The same applies for the case of the *Pathway complementarities* only there is no ModelSEED ids as they are computed directly from the KEGG annotated genomes and not from the Genome-Scale Metabolic Reconstructions; that is the case for the *Seed complementarities*. The second are pathway complementarities. These are directly computed from the KEGG annotated genomes and do not involve ModelSEED ids since the metabolic network construction step is not necessary in this case. C. Part of a colored KEGG map returned based on the seed complementarities. Compounds that belong to the beneficiary taxon are colored in cyan while the potential complementing seeds from the donor are highlighted in red.

Results and discussion

Annotating microbial co-occurrence networks with `microbetag`

The `microbetag` software ecosystem consists of five main modules: a. `microbetagDB` including `microbetag` precalculations, b. the `microbetag` workflow to annotate the co-occurrence network, c. a web server hosting both the `microbetagDB` and the `microbetag` application, d. a CytoscapeApp called `MGG` that enables users to easily invoke the workflow and investigation of the annotated network, and e. a stand-alone pre-processing workflow provided as Docker image for data sets with more than 1,000 sequence identifiers (OTUs/ASVs/bins etc.).

Currently, `microbetagDB` includes more than 34,000 genomes (Table 1) along with their corresponding annotations. The vast majority of these genomes represent bacterial taxa (364 taxa are Archaea). The presence/absence of more than 30 phenotypic traits have been predicted for those genomes. About 1.4 billion potential metabolic interactions leading to pathway or seed complementarities have been precomputed as well. There are one order of magnitude more seed complements than those corresponding to pathway complementarities as for all GENREs present in `microbetagDB` all pairwise complements were calculated ($33,755^2$) and stored, even if empty. In the case of pathway complementarities, a genome pair is present in the database and thus counted only if a potential complementarity was found. Yet, in the first case, the number of genomes with no potential seed complements ranges from zero to a few dozen. All annotations can be accessed directly from `microbetagDB` through the API. Using GENREs for the seed complements rather than the genomes themselves facilitates a more realistic simulation of the essential requirements of corresponding taxa for growth (seeds) and of potential secretions (non-seeds) assuming species (model) growth. However, this comes with its own challenges (see paragraph 3).

Running the `microbetag` workflow is straightforward and can be done using an abundance table with taxonomic assignments as input. When the taxonomy scheme being used is not one among the GTDB, Silva or the GTDB-oriented taxonomy for 16S rRNA amplicon data (see pre-process paragraph 3), the most time-consuming step of the workflow is the one mapping user's taxonomy to a NCBI Taxonomy Id and from that to GTDB representative genomes. Network inference can be a computationally intensive step too, particularly as the number of sequences in the abundance table increases. To enable annotation of large data sets, a stand-alone pre-process workflow is provided with `microbetag`. The user can either assign their amplicon data to the GTDB-oriented taxonomy and/or reconstruct a network locally. Once a network is available and the taxonomy used is among the standard ones for `microbetag`, the computational time required for annotation ranges from several seconds to a few minutes based on user's settings. An annotated network in `cx2` format is returned, which can be viewed in Cytoscape. A tutorial, frequently asked questions and hints to address the idiosyncrasies of various data sets are available on the [ReadTheDocs](#) web site while a Gitter community allows users to exchange experience and ask for more specific help. In the following two sections, we present a validation and use case, highlighting our approach's potential.

Table 1: Summary of the data in `microbetag`DB

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	33,755
Seed complements	1,139,400,025
Seed scores	1,105,250,048

A validation case

To validate `microbetag` we used the correlation network of Hessler *et al.* [35] describing mine tailing-derived laboratory microbial consortia. In this study, *Variovorax*, a thiamine producer, and its co-occurrence with a series of thiamine auxotrophs are discussed. The study was selected as a validation case as the authors tested the network's predictions by performing co-culture experiments measuring the thiamine production. Sequence bins corresponding to network nodes and the original network were obtained from the authors. Using GTDB-tk [60], bins were annotated to GTDB taxonomies; those retrieved were added in the original network, which was then annotated with `microbetag`. Supplementary Figure B2 highlights bin_55 that corresponds to *Variovorax* and its first neighbors. The annotated network is available on `microbetag`'s [GitHub repository](#). GTDB-tk returned GCA_001899795.1 as the one closer to bin_55 assigning it to *Variovorax* sp001899795. `microbetag` then suggested that this specific genome corresponds to an aerobe [61], that can grow autotrophically if needed [62] and utilizes D-glucose, while producing ethanol and lactic acid [63]. Last, the Type VI secretion system was suggested to be available on its genome [64].

Hessler *et al.* argue that *Variovorax* is an important thiamine source and can supply neighboring species that cannot produce it (auxotrophs). Indeed, `microbetag` was able to suggest several thiamine-related potential seed complements among the potential metabolic interactions between *Variovorax* as well as its neighbors (Table 2.A). Potential interactions were also found in some cases between the neighbors themselves (Table 2.B). The authors also argue that isolates of that *Variovorax* strain required the addition of pantothenic acid to grow. However, based on the KEGG annotation of the genome that bin_55 was mapped to, it contains both KEGG MODULES related to pantothenate biosynthesis, M00119 (valine/L-aspartate \Rightarrow pantothenate) and M00913 (2-oxoisovalerate/spermine \Rightarrow pantothenate). Other genomes though are not capable of either one or any of those (Supplementary Table) B1. This example highlights a challenge of the `microbetag` approach (see Section 3). The complementarities between the nodes would have been different if bin_55 was classified as another *Variovorax* genome, with incomplete modules. For example, if GCF_001577265.1 was picked and its complementarities with the neighboring taxa were retrieved, it would have revealed that all its neighboring species can actually provide it with pantothenate (C00864) as suggested by their seed complementarities (see [coloured map](#)).

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

Neighboring taxon	node id	KEGG compounds	url
<i>Kapabacteria thiocyanatum</i>	bin..59	C15809	url
<i>Terrimonas ferruginea</i>	bin..100	C15809;C01081	url
<i>Tahibacter</i> sp001725155	bin..167	C15809	url
<i>Microbacterium</i> sp900156455	bin..28	C15809; C20246	url
<i>Sphingobium</i> sp001899715	bin..155	Iminoglycine C15809;	url
<i>Nitrosospira</i> sp001899235	bin..176	None	None
62-47 sp001899235*	bin..233	None	None
<i>Bosea</i> sp001898115	bin..273	C04327;C01279	url
54-19 sp001898225**	bin..41	C15809	url
<i>Rhodoglobus</i> sp001725325	bin..8	C15809	url

B. Potential thiamine-related complements among *Variovorax* neighbors

Beneficiary	Donor	potential complement
<i>T. ferruginea</i>	<i>Tahibacter</i> sp001725155	C01081
<i>T. ferruginea</i>	<i>Rhodoglobus</i> sp001725325	C01081
<i>Nitrosospira</i> sp001899235	<i>Bosea</i> sp001898115	C04327;C01279
Chloroflexi	<i>Bosea</i> sp001898115	C15809
Chloroflexi	Xanthobacteraceae	C15809
Chloroflexi	<i>Nitrosospira</i> sp001899235	C15809

Table 2: Thiamine biosynthesis related seed complements between *Variovorax* and its first neighbors in the network of Hessler *et al.*[35] (A), and between pairs of the neighbors (B). Bin sequence files were mapped against GTDB using GTDB-tk. Chloroflexi refers to the GTDB taxonomy of: 54-19 sp001898225

Interpreting a real-world network with microbetag

Annelies' dataset.

One last visual component from the use case would be nice to have.

Potential and limitations

The previous paragraph shows the potential of **microbetag** in the interpretation of co-occurrence networks and how it can be used to generate new hypotheses derived from those. However, **microbetag** benefits the microbiome community in several other ways. The microbetagDB provides a vast number of annotations; 31 predicted traits for more than 30,000 genomes, their GENREs along with their corresponding seed sets, potential metabolic complementarities and cooperation/competition scores. Such a resource may support a range of studies; from a more theoretical perspective regarding the distribution of the complements among taxonomic groups or how often a complement potentially appears, to applications such as eco-evolutionary studies and the investigation of interactions.

Yet, there are several challenges involved in our approach. First, **microbetag** inherits all the biases and drawbacks of both the data and the software it is based on. Functional annotation comes with its own limitations. Some domains boast richer annotations and more comprehensive descriptions compared to others, thus exhibiting a wealth of detail and employing more precise terminology, particularly for widely recognized processes (e.g., glycolysis and secondary metabolites biosynthesis correspondingly).

In the validation case, the bin representing the *Variovorax* strain was mapped to a genome that is supposed to contain the pantothenate KEGG modules. Thus, the fact that it requires to receive pantothenate from its environment to grow, as the authors mention, would not have been predicted in the **microbetag** framework. Beyond the sequencing and the annotation challenges, we also need to consider the fact that a pathway may not be fully represented in a KEGG module. Of course, various factors can prevent the actual production of an enzyme even if its genetic information is included in a species' genome.

Pathway complementarity can only be as accurate as the KEGG MODULE database and as precise as the software annotating genomes with KO terms. It is well known that automated Genome-Scale Metabolic Reconstruction comes with a great number of challenges and different software for this task come with their intrinsic limitations [65]. Using ModelSEED with a complete medium may limit potential metabolic interactions but, on the other hand, the retrieved ones will be of higher confidence.

It is also well known that higher-order interactions, i.e. interactions involving more than two species [32], should also be considered. Pairwise relationships do not capture the more complex forms of ecological interactions, in which species depend on (or are influenced by) multiple other species [3]. However, since **microbetag** is decoupled from network inference, it could annotate a network with hyperedges (i.e. edges connecting more than two taxa) produced by a future tool capable of inferring higher-order interactions.

Last, the limited number of Archaea in microbetagDB is also the result of a software limitation. As shown in [66] (Figure 6b), the original version of CheckM [67] that is still being used by GTDB returns lower completeness scores for genomes that correspond to phyla known for having smaller genomes in general, e.g. Patescibacteria representative genomes on GTDB have an average completeness of ~ 65%. Thus, only few representatives from these taxonomic groups passed our filters leading to an important under-representation of Archaea.

Future work

In the near future, we plan to develop two main features: a. the integration of transcriptomics data provided by the user, which would enhance or lower the probability for a potential metabolic interaction to occur based on whether the KO terms involved are present or not, and b. the integration of spatial data; it is well-known that the distance between cells determines whether an interaction occurs [68]. For this, we intend to support data with spatial dimensions. Thus, potential metabolic interactions between taxa that are closer to one another would be more probable to occur.

Conclusions

Co-occurrence networks are widely used in microbiome studies to explore associations [4]. However, their inference and their interpretation come with a range of challenges [14]. Metabolic exchanges among microbial taxa are considered ubiquitous [69] in a large number of environments. In our study, we exploit reverse-ecology approaches and publicly available genomic data and software to predict phenotypic traits and construct metabolic networks to annotate co-occurrence networks derived from amplicon or shotgun data. Our annotation was in-line with the study of Hessler et al. [35] predicting thiamine-related metabolic interactions among *Variovorax* and its closest neighbors, suggesting several ways to achieve them. Using Enrichment analysis using them combined with network clustering algorithms can further benefit their interpretation. Both microbetagDB and **microbetag** will benefit microbiome studies, as a resource and as a hypothesis generation tool.

Use
case

Supplementary information. List of supplementary figures and tables.

Supplementary Figure 1: `microbetag` software ecosystem architecture.

Supplementary Figure 2: *Variovorax* (node_55) and its closest neighbors. *Variovorax* annotations are shown in the node CyPanel.

Supplementary Table 1: *Variovorax* genomes present on `microbetagDB` and their corresponding complete/incomplete presence of the pantothenate - related KEGG modules

Supplementary Table 2: Computing times per step of the `microbetag` workflow using four different data sets.

Declarations

- **Availability of data and materials**

- Raw sequences for the use case:
- Raw data for the validations case:

- **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**

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 for the insight on how to organise the trait groups. We would also like to thank Dr. Hessler and Prof. Jillian F. Banfield for sharing both the bins and the network of their study [35].

- **Ethics approval**

Not applicable

- **Consent to participate**

Not applicable.

- **Code availability:**

- `microbetagDB` related scripts: <https://github.com/hariszaf/microbetag>
- `microbetag` application: <https://github.com/msysbio/microbetagApp>.
- MGG CytoscapeApp: <https://github.com/ermismd/MGG/>
- Validation and use case:
- Documentation web-site: <https://hariszaf.github.io/microbetag/>

Appendix A Background on pathway and seed complementarity

For a genome to have a KEGG module *complete* means it provides at least one complete *alternative*. Alternatives are considered as the unique combinations of KOs that connect an input compound to an output compound of the module. For example, the definition of the D-Galacturonate degradation in Bacteria ([M00631](#)) is:

K01812 K00041 (K01685,K16849+K16850) K00874 (K01625,K17463)

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

In alternatives two and four, the K16849+K16850 is a *complex*, meaning both KO terms are required for the step to be available.

In case of seed complementarity, in `microbetag` we focus on the effect that a metabolic exchange between two taxa might have if the seed of the beneficiary taxon is linked to a KEGG MODULE. Therefore, the KOs that were found linked to modules were mapped to ModelSEED ids. The initial seed and non-seed sets that were exported as sets of ModelSEED ids were then mapped to KOs too. When the non-seed set of a genome (donor) provided a seed related to a KEGG module to another genome (beneficiary), this is considered a potential metabolic interaction.

Appendix B Validation

Table B1: *Variovorax* genomes present in microtagDB and their corresponding complete/incomplete presence of the pantothenate-related KEGG modules. The genome that bin_55 was mapped to is shown in bold.

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
GCA_001899795.1	complete	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

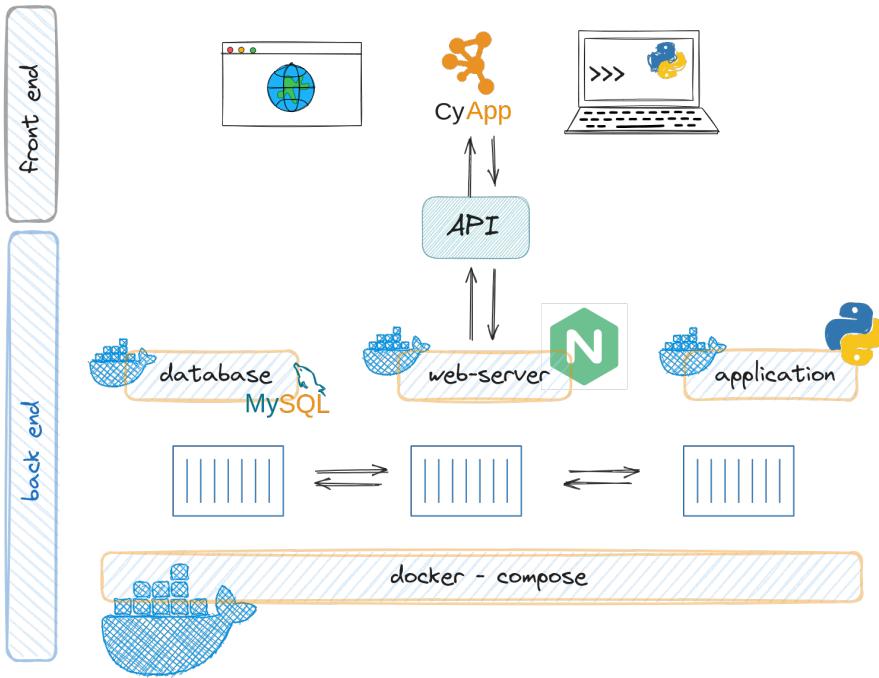


Fig. B1: `microbetag` software ecosystem. Three Docker containers are combined: a nginx web server connected to a MySQL database and the `microbetag` workflow. An API, as part of the last container, enables the communication between the client (front) and the server side (back end). The content of the `microbetagDB` and the `microbetag` workflow are accessible through a web-browser, a terminal and a CytoscapeApp (MGG).

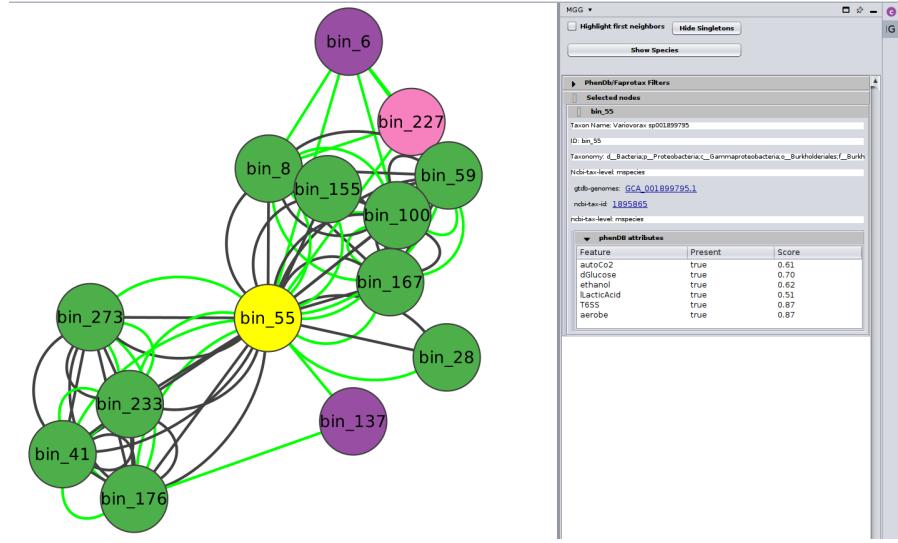


Fig. B2: *Variovorax* node (bin_55) and its neighbors annotated by **microbetag**. All but three of them were not mapped to a GTDB representative genome. Green edges represent the positive association weights. The black edges represent pairwise seed complementarities and scores.

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] 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)
- [7] 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>
- [8] 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)
- [9] 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)
- [10] 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)
- [11] Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data. *PLoS computational biology* **8**(9), 1002687 (2012)

- [12] 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)
- [13] 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)
- [14] Faust, K.: Open challenges for microbial network construction and analysis. *The ISME Journal* **15**(11), 3111–3118 (2021)
- [15] 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)
- [16] 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)
- [17] 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)
- [18] Berry, D., Widder, S.: Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology* **5**, 219 (2014)
- [19] 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)
- [20] Röttjers, L., Faust, K.: Manta: A clustering algorithm for weighted ecological networks. *Msystems* **5**(1), 10–1128 (2020)
- [21] Levy, R., Borenstein, E.: Reverse ecology: from systems to environments and back. In: *Evolutionary Systems Biology*, pp. 329–345. Springer, ??? (2012)
- [22] 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)
- [23] Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**(1), 93–121 (2010)
- [24] Durot, M., Bourguignon, P.-Y., Schachter, V.: Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS microbiology reviews* **33**(1), 164–190 (2008)

- [25] 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
- [26] Machado, D., Andrejev, S., Tramontano, M., Patil, K.R.: Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research* **46**(15), 7542–7553 (2018)
- [27] Seaver, S.M., Liu, F., Zhang, Q., Jeffryes, J., Faria, J.P., Edrizinghe, J.N., Mundy, M., Chia, N., Noor, E., Beber, M.E., *et al.*: The modelseed biochemistry database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic acids research* **49**(D1), 575–588 (2021)
- [28] 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)
- [29] Parter, M., Kashtan, N., Alon, U.: Environmental variability and modularity of bacterial metabolic networks. *BMC evolutionary biology* **7**, 1–8 (2007)
- [30] 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)
- [31] 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)
- [32] 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)
- [33] Belcour, A., Frioux, C., Aite, M., Bretaudau, A., Hildebrand, F., Siegel, A.: Metage2metabo, microbiota-scale metabolic complementarity for the identification of key species. *Elife* **9**, 61968 (2020)
- [34] 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

- [35] 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)
- [36] 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)
- [37] 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)
- [38] 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)
- [39] 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)
- [40] 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>
- [41] Murali, A., Bhargava, A., Wright, E.S.: Idtaxa: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* **6**(1), 1–14 (2018)
- [42] Wright, E.S.: Using decipher v2. 0 to analyze big biological sequence data in r. *R Journal* **8**(1) (2016)
- [43] 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)
- [44] Louca, S., Parfrey, L.W., Doebeli, M.: Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**(6305), 1272–1277 (2016)
- [45] 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)
- [46] Feldbauer, R., Schulz, F., Horn, M., Rattei, T.: Prediction of microbial phenotypes based on comparative genomics. *BMC bioinformatics* **16**(14), 1–8 (2015)

- [47] 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)
- [48] 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)
- [49] 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)
- [50] 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)
- [51] Kanehisa, M., Sato, Y.: Kegg mapper for inferring cellular functions from protein sequences. *Protein Science* **29**(1), 28–35 (2020)
- [52] Kanehisa, M., Sato, Y., Kawashima, M.: Kegg mapping tools for uncovering hidden features in biological data. *Protein Science* **31**(1), 47–53 (2022)
- [53] 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)
- [54] Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Lindsay, B., Stevens, R.L.: High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28**(9), 977–982 (2010)
- [55] 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)
- [56] 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)
- [57] Merkel, D., *et al.*: Docker: lightweight linux containers for consistent development and deployment. *Linux j* **239**(2), 2 (2014)
- [58] Reese, W.: Nginx: The high-performance web server and reverse proxy. *Linux J.* **2008**(173) (2008)

- [59] 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)
- [60] Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H.: GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press (2020)
- [61] Carbajal-Rodríguez, I., Stöveken, N., Satola, B., Wübbeler, J.H., Steinbüchel, A.: Aerobic degradation of mercaptosuccinate by the gram-negative bacterium *variovorax paradoxus* strain b4. *Journal of Bacteriology* **193**(2), 527–539 (2011) <https://doi.org/10.1128/jb.00793-10> <https://journals.asm.org/doi/pdf/10.1128/jb.00793-10>
- [62] Han, J.-I., Choi, H.-K., Lee, S.-W., Orwin, P.M., Kim, J., LaRoe, S.L., Kim, T.-g., O’Neil, J., Leadbetter, J.R., Lee, S.Y., Hur, C.-G., Spain, J.C., Ovchinnikova, G., Goodwin, L., Han, C.: Complete genome sequence of the metabolically versatile plant growth-promoting endophyte *variovorax paradoxus* s110. *Journal of Bacteriology* **193**(5), 1183–1190 (2011) <https://doi.org/10.1128/jb.00925-10> <https://journals.asm.org/doi/pdf/10.1128/jb.00925-10>
- [63] Sun, J., Matsumoto, K., Ndoko, J.M., Ooi, T., Taguchi, S.: Enzymatic characterization of a depolymerase from the isolated bacterium *variovorax* sp. c34 that degrades poly(enriched lactate-co-3-hydroxybutyrate). *Polymer Degradation and Stability* **110**, 44–49 (2014) <https://doi.org/10.1016/j.polymdegradstab.2014.08.013>
- [64] Astafyeva, Y., Gurschke, M., Qi, M., Bergmann, L., Indenbirken, D., Grahl, I., Katzowitsch, E., Reumann, S., Hanelt, D., Alawi, M., Streit, W.R., Krohn, I.: Microalgae and bacteria interaction—evidence for division of diligence in the alga microbiota. *Microbiology Spectrum* **10**(4), 00633–22 (2022) <https://doi.org/10.1128/spectrum.00633-22> <https://journals.asm.org/doi/pdf/10.1128/spectrum.00633-22>
- [65] Mendoza, S.N., Olivier, B.G., Molenaar, D., Teusink, B.: A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome biology* **20**(1), 1–20 (2019)
- [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] Dal Co, A., Vliet, S., Kiviet, D.J., Schlegel, S., Ackermann, M.: Short-range

interactions govern the dynamics and functions of microbial communities. *Nature ecology & evolution* 4(3), 366–375 (2020)

- [69] Kost, C., Patil, K.R., Friedman, J., Garcia, S.L., Ralser, M.: Metabolic exchanges are ubiquitous in natural microbial communities. *Nature Microbiology*, 1–9 (2023)