

001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046

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.errega@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 like Chris Quince is our editor.

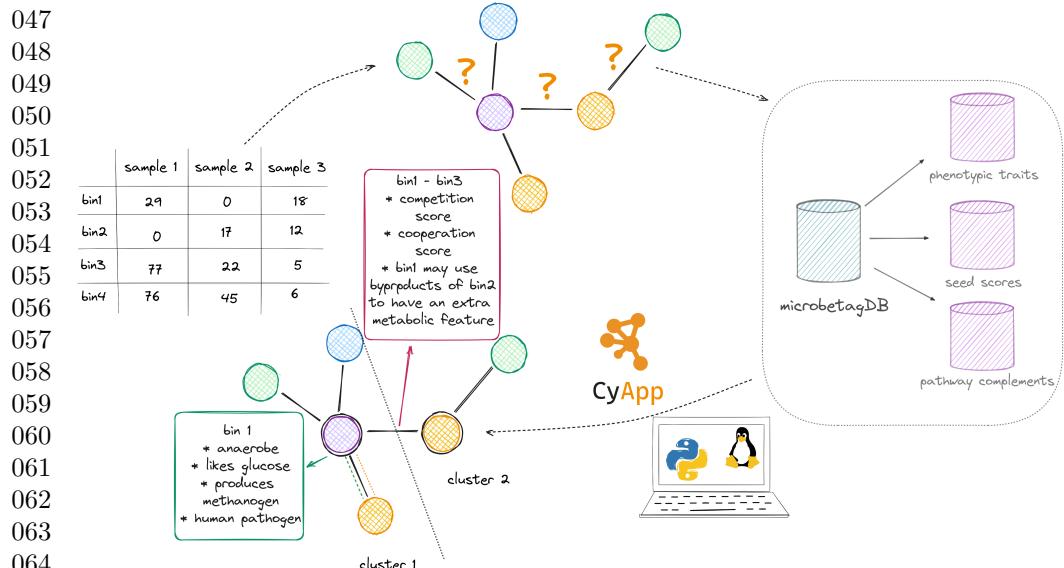


Figure abstract.

065
066
067
068 **Keywords:** microbial associations, enrichment analysis, pathway
069 complementarity, seed set

072 Background ¹ ² 073

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

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

089
090
091 ¹We are to submit in the Microbiome journal as a "Software" manuscript, thus we follow [these rules](#)
092 ²The 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].	093
	094
	095
	096
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 features 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 extract 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., SpecEasi [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., compositionality, 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 <i>What can we learn from the hairballs</i> posed by Röttjers et al. [4] could provide essential insight on the mechanisms of the interactions.	097
	098
	099
	100
	101
	102
	103
	104
	105
	106
	107
	108
	109
	110
	111
	112
	113
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 several 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 hypotheses for further investigation both at the taxon pair and the community level.	114
	115
	116
	117
	118
	119
	120
	121
	122
	123
	124
	125
	126
	127
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 <i>reverse ecology</i> [24]. Reverse ecology leverages genomics to explore community ecology with no <i>a priori</i> 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].	128
	129
	130
	131
	132
	133
	134
	135
	136
	137
	138

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

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

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

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

182

183

184

Implementation	³	185
		186
		187
		188
		189
		190
		191
		192
		193
		194
		195
		196
		197
		198
		199
		200
		201
		202
		203
		204
		205
		206
		207
		208
		209
		210
		211
		212
		213
		214
		215
		216
		217
		218
		219
		220
		221
		222
		223
		224
		225
		226
		227
		228
		229
		230

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 between 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

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

231 the [microbetag preprocess Docker image](#). Likewise, when the abundance table consists
232 of more than 1,000 taxa, providing a network as an input is mandatory. Again, to help
233 the user, [microbetag](#) preprocess Docker image supports the inference of a network
234 using FlashWeave.

235

236 **Literature based nodes annotation**

237 Using a set of Tara Ocean samples [46] FAPROTAX [47] estimates the functional
238 potential of the bacterial and archaeal communities, by classifying each taxonomic unit
239 into functional group(s) based on current literature, descriptions of cultured represen-
240 tatives and/or manuals of systematic microbiology. In this manually curated approach,
241 a taxon is associated with a function if and only if all the cultured species within the
242 taxon have been shown to exhibit that function. In its current version, FAPROTAX
243 includes more than 80 functions based on 7600 functional annotations and covering
244 more than 4600 taxa. Contrary to gene content based approaches, e.g. PICRUSt2 [48],
245 FAPROTAX estimates metabolic phenotypes based on experimental evidence.
246

247 [microbetag](#) invokes the accompanying script of FAPROTAX and converts the
248 taxonomic microbial community profile of the samples included in the user's abun-
249 dance table or of the taxa present in the provided network, into putative functional
250 profiles. Then, it parses FAPROTAX's subtables to annotate each taxonomic unit
251 present in the user's data with all the functions for which they had a hit. FAPROTAX
252 annotations are not part of the microbetagDB but are computed on the fly.
253

254 **Genomic based nodes annotation**

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

264 In the framework of microbetagDB, classifiers were re-trained using the genomes
265 provided by phenDB for each trait to sync with the latest version of eggNOG [50]
266 and the [phenotrex](#) [49] software tool. Genomes were downloaded from NCBI using
267 the [Batch Entrez](#) program. Then, *genotype* files were produced for all the high quality
268 GTDB representative genomes. Each model was then used against all the GTDB
269 *genotype* files to annotate each with the presence or the absence of the trait. A list of all
270 the phenotypic traits available for the genomes present in microbetagDB is available
271 on [microbetag](#)'s [documentation site](#). The updated models are also available
272

273 **Pathway complementarity**

274

275 To infer potential pathway complementarities we consider the modules described in
276 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 potential donor species, and one vice-versa.

Seed scores 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 an upper bound assessment of the potential for syntrophic interactions between two

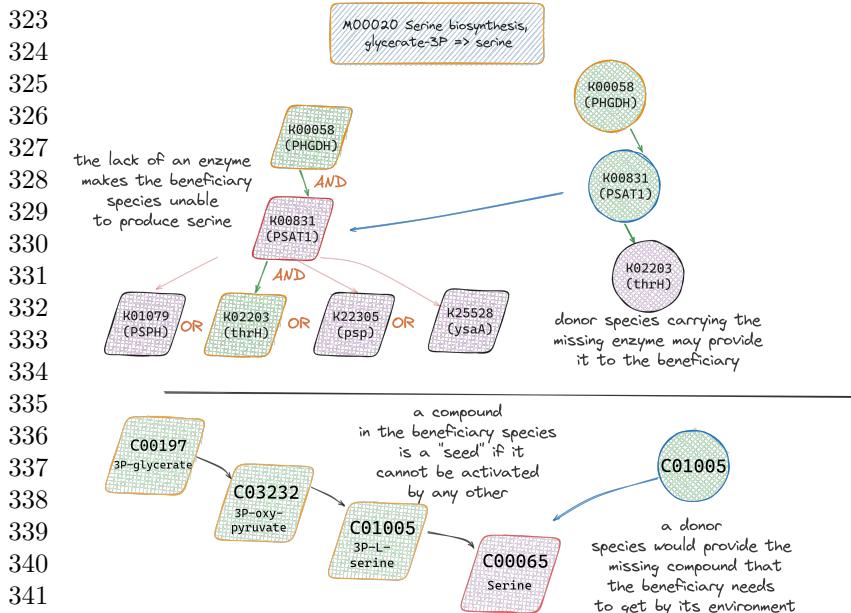


Fig. 1: Pathway complementarity approach. The high quality GTDB genomes were annotated with KEGG ORTHOLOGY (KO) terms. The various ways of getting a KEGG module complete were enumerated and all the possible ways a donor species could "fill" a beneficiary's non-complete module were calculated. In this 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.

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

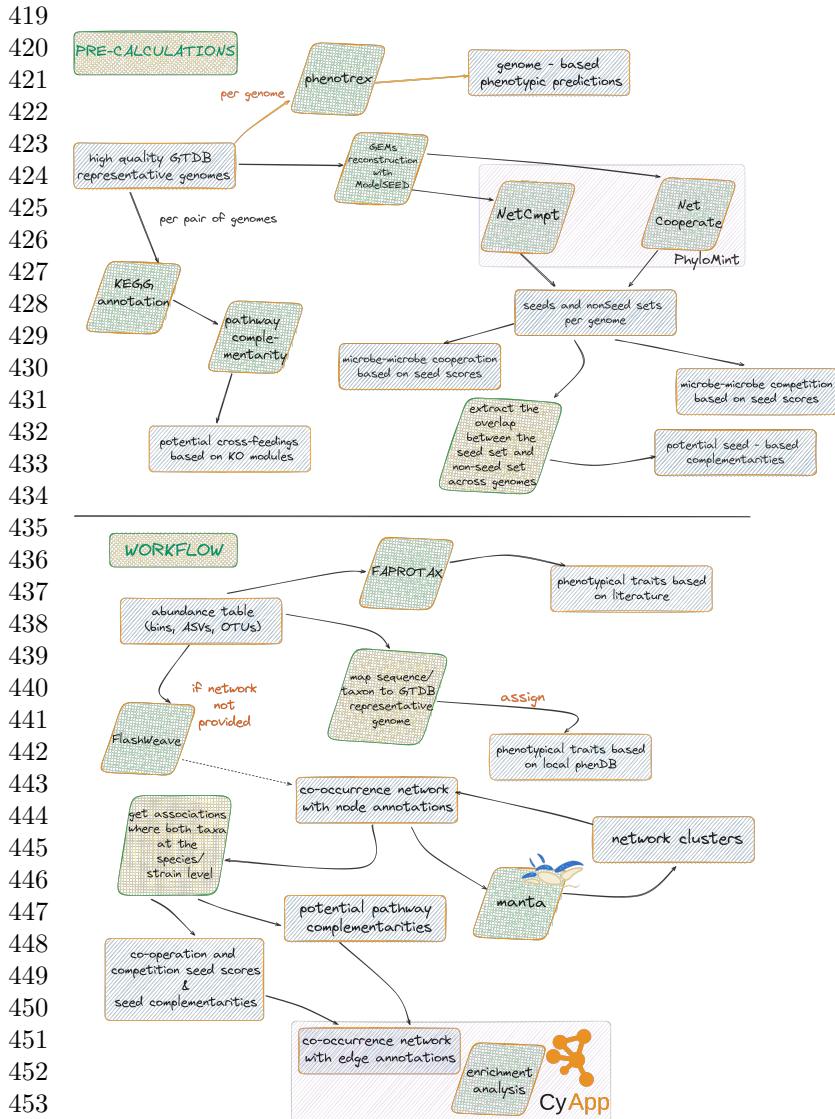
overlap among all the pairwise combinations of those genomes. More specifically, the overlap of <i>seed set</i> _{species_A} with the <i>non seed set</i> _{species_B} was retrieved. microbetag then annotates again the edges of the co-occurrence network where both taxa have been mapped to at least one GTDB genome, mentioning all the KEGG maps for which there is at least one seed compound of the potentially beneficiary species	369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
Clustering network	375
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 microbetag invokes manta [23] to infer clusters from the microbial network. A taxonomically-informed layout is	376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
strong cluster assignments. ++ taxonomy layout	382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
Groups of annotations	383
Biologically meaningful groups were built using the micrO ontology [60].	384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
Building the CytoscapeApp	386
The microbetag CytoscapeApp was build based on the source code of the scVizNet [61]. Java @Ermis to add	387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
Enrichment analysis is supported. Hypergeometric distribution FDR +++	390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
Dependencies, Web server and API	392
The microbetag web service is container - based and consists of three Docker [62] (v24.0.2) images: a. the MySQL database b. an nginx [63] 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 .	393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
python 3.11 slim docker image julia 1.7.1 for flashweave mysql.connector 8.0.27	402 403 404 405 406 407 408 409 410 411 412 413 414
python library pandas 2.1.1. numpy 1.26.0 multiprocessing	403 404 405 406 407 408 409 410 411 412 413 414
text processing using awk	404 405 406 407 408 409 410 411 412 413 414
KEGG API	405 406 407 408 409 410 411 412 413 414

415 **Results**⁴

416 **microbetag and microbetagDB**

418

419



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

458

459

460 ⁴Significant 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.

Table 1: Summary of Data⁵

Description	Entries	
GTDB representative genomes	34,608	461
Phen-model-oriented metabolic functions	32	462
FAPROTAX functions	92	463
Unique complement*	341,568	464
GENREs leading to ~ 1 billion competition and complementarity scores	30,755	465
		466
		467
		468
		469
annotated network returned in .cyjs format		470
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 [64]. A co-occurrence network is also built using FlashWeave [14], as microbetag also does.		471
		472
		473
		474
		475
		476
		477
		478
		479
		480
		481
		482
		483
		HP ⁶
		487
		488
		489
		490
		491
		492
		493
		494
		495
		496
		497
		498
		499
		500
		501
		502
		503
		504
		505
		506

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.

A. GTDB-tk: 480 bins			B. GTDB 16S: 3000 ASVs		
Step	Time(sec)	Notes	Step	Time(sec)	Notes
Taxonomy mapping	Cell 1,2	on the fly	Taxonomy assignment		Docker image on HP ⁶
Network inference	Cell 2,2	on the fly	Taxonomy mapping	Cell 1,2	Cell 1,3 484
microbetag annotations	Cell 3,2	on the fly	Network inference	Cell 2,2	Cell 2,3 485
manta clustering	Cell 4,2	on the fly	microbetag annotations	Cell 3,2	Cell 3,3 486
			manta clustering	Cell 4,2	Cell 4,3 487
					488
					489
					490
					491
					492
					493
					494

Table 2: Computing times per step using an abundance table of 400 taxa with taxonomy: A. taxonomy scheme B. C. D. ⁶ specs of the laptop used.

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

Validation of microbetag potential

vitamin dataset [39]

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

Study
those
2 to
under-
stand
our
find-
ings

507 that control community organization in mine tailings-derived laboratory microbial
508 consortia grown under dozens of conditions.

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

511 Thiamine alternative pathway [65, 66]

512

513 Discussion ⁷

514

515 Interpreting a real-world network with **microbetag**

516 Annelies' dataset.

517

518 **microbetag** as a resource

519 Limitations

520

521 As shown in [67] (see Figure 6b), the original version of CheckM [68] that is still used on
522 GTDB returns lower completeness scores to genomes that correspond to phyla known
523 for having shorter genomes in general, e.g. Patescibacteria representative genomes on
524 GTDB have an average completeness 65%. **microbetag** inherits this in the filtering
525 process for getting only high quality genomes and thus, only few representatives from
526 these taxonomic groups are present on microbetagDB.

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

531

532 Future work

533

534 Further indices using the seed concept have been also presented such as the metabolic
535 interaction potential (*MIP*) and the metabolic resource overlap (*MRO*). *MIP* is
536 defined as the difference between the minimal number of components required for the
537 growth of all members in a noninteracting community and an interacting community,
538 i.e. the maximum number of essential nutritional components that a community can
539 provide for itself through interspecies metabolic exchanges [30]. Similarly, *MRO* is
540 defined as the maximum possible overlap between the minimal nutritional require-
541 ments of all member species [30]. Regression and association rule mining [69] can be
542 applied to address this challenge.

- 543
- 544 • pathway and seed complementarities for higher-order interactions
 - 545 • spatial dimension
 - 546 • transcriptomics data integration: compare potential complementarities with what
 - 547 is going on
 - 548 •

549

550 ⁷The user interface should be described and a discussion of the intended uses of the software, and the
551 benefits that are envisioned, should be included, together with data on how its performance and functionality
552 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.

Conclusions	553
8	554
Data integration	555
Supplementary information.	556
9	557
	558
Declarations	559
• Availability of data and materials	560
– Raw sequences for the use case:	561
– Raw data for the validations case:	562
• Funding	563
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.	564
• Conflict of interest/Competing interests	565
The authors declare that they have no other competing interests.	566
• Authors' contributions ¹⁰	567
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.	568
• Acknowledgements	569
We would like to thank Dr Christina Pavloudi and ++ for the insight on how to organise the trait groups.	570
• Ethics approval	571
Not applicable	572
• Consent to participate	573
Not applicable.	574
• Code availability:	575
– microbetagDB related scripts: https://github.com/hariszaf/microbetag	576
– microbetagApp and webserver: https://github.com/msysbio/microbetagApp .	577
– CytoscapeApp: https://github.com/ermismd/MGG/	578
– Validation and use case: {think of having that under the 3D'omics organization}	579
– Documentation web-site: https://hariszaf.github.io/microbetag/	580

⁸This should state clearly the main conclusions and provide an explanation of the importance and relevance of the case, data, opinion, database or software reported.

⁹If your article has accompanying supplementary file(s) please state so here. E.g. supplementary figures and tables captions.

¹⁰Based on the [CRediT system](#). Current list is indicative.

599 **Appendix A Mappings**

600
601 $n : 1 n : n$ etc

602
603 **Appendix B Background on seed scores and**
604 **complementarities**
605

606 **B.1 Background on seed scores**

607
608 In that case, once a seed is assured, it activates all the rest of that group. Therefore,
609 a confidence level (C) ranging from 0 to 1, has been previously described to quantify
610 the relevance of each seed:

611
612
$$C_i = 1 / \text{seed}'s \text{ group with } i \text{ size} \quad (\text{B1})$$

613 $C = 0$ corresponds to a non-seed node, while $C = 1$ represents an independent
614 node.

615
616
$$MI_{\text{Complementarity}} = \frac{|\text{SeedSet}_A \cap \neg \text{SeedSet}_B|}{|\text{SeedSet}_A \cap (\text{SeedSet}_B \cup \neg \text{SeedSet}_B)|} \quad (\text{B2})$$

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

620
621
$$MI_{\text{Competition}} = \frac{\sum C(\text{SeedSet}_A \cap \text{SeedSet}_B)}{\sum C(\text{SeedSet}_A)} \quad (\text{B3})$$

622 **B.2 Background on pathway complementarity**

623 For example, the definition of the D-Galacturonate degradation in Bacteria ([M00631](#))
624 is:

625 K01812 K00041 (K01685,K16849+K16850) K00874 (K01625,K17463)
626 that once breaking down, it leads to 4 alternative sets of KOs (pathways):
627
628 K01812 K00041 K01685 K00874 K01625
629 K01812 K00041 K16849+K16850 K00874 K01625
630 K01812 K00041 K01685 K00874 K17463
631 K01812 K00041 K16849+K16850 K00874 K17463
632
633
634
635
636
637
638
639

640 **B.3 Complementarities**

641 KEGG compound ModelSEED compounds ModelSEED compounds mapped to
642 KEGG compounds and kept only those related to KEGG modules.
643

References	645
[1] Yuan, M.M., Guo, X., Wu, L., Zhang, Y., Xiao, N., Ning, D., Shi, Z., Zhou, X., Wu, L., Yang, Y., <i>et al.</i> : Climate warming enhances microbial network complexity and stability. <i>Nature Climate Change</i> 11 (4), 343–348 (2021)	646
[2] Raes, J., Bork, P.: Molecular eco-systems biology: towards an understanding of community function. <i>Nature Reviews Microbiology</i> 6 (9), 693–699 (2008)	647
[3] Faust, K., Raes, J.: Microbial interactions: from networks to models. <i>Nature Reviews Microbiology</i> 10 (8), 538–550 (2012)	648
[4] Röttjers, L., Faust, K.: From hairballs to hypotheses—biological insights from microbial networks. <i>FEMS microbiology reviews</i> 42 (6), 761–780 (2018)	649
[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., <i>et al.</i> : Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. <i>FEMS microbiology reviews</i> 40 (5), 686–700 (2016)	650
[6] Robinson, C.J., Bohannan, B.J., Young, V.B.: From structure to function: the ecology of host-associated microbial communities. <i>Microbiology and Molecular Biology Reviews</i> 74 (3), 453–476 (2010)	651
[7] D’Souza, G., Shitut, S., Preussger, D., Yousif, G., Waschina, S., Kost, C.: Ecology and evolution of metabolic cross-feeding interactions in bacteria. <i>Natural Product Reports</i> 35 (5), 455–488 (2018)	652
[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., Paflis, 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]. <i>F1000Research</i> 13 (50) (2024) https://doi.org/10.12688/f1000research.144515.1	653
[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., <i>et al.</i> : A new view of the tree of life. <i>Nature microbiology</i> 1 (5), 1–6 (2016)	654
[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. <i>Computational and structural biotechnology journal</i> 19 , 2687–2698 (2021)	655
[11] Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.: Microbial co-occurrence relationships in the human microbiome. <i>PLoS computational biology</i> 8 (7), 1002606 (2012)	656
	657
	658
	659
	660
	661
	662
	663
	664
	665
	666
	667
	668
	669
	670
	671
	672
	673
	674
	675
	676
	677
	678
	679
	680
	681
	682
	683
	684
	685
	686
	687
	688
	689
	690

- 691 [12] Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data.
692 PLoS computational biology **8**(9), 1002687 (2012)
- 693
- 694 [13] Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bon-
695 neau, R.A.: Sparse and compositionally robust inference of microbial ecological
696 networks. PLoS computational biology **11**(5), 1004226 (2015)
- 697
- 698 [14] Tackmann, J., Rodrigues, J.F.M., Mering, C.: Rapid inference of direct interac-
699 tions in large-scale ecological networks from heterogeneous microbial sequencing
700 data. Cell systems **9**(3), 286–296 (2019)
- 701
- 702 [15] Faust, K.: Open challenges for microbial network construction and analysis. The
703 ISME Journal **15**(11), 3111–3118 (2021)
- 704
- 705 [16] Cao, H.-T., Gibson, T.E., Bashan, A., Liu, Y.-Y.: Inferring human microbial
706 dynamics from temporal metagenomics data: Pitfalls and lessons. BioEssays
707 **39**(2), 1600188 (2017)
- 708
- 709 [17] Kishore, D., Birzu, G., Hu, Z., DeLisi, C., Korolev, K.S., Segrè, D.: Inferring
710 microbial co-occurrence networks from amplicon data: a systematic evaluation.
711 Msystems, 00961–22 (2023)
- 712
- 713 [18] Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y.,
714 Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., *et al.*: Correlation detection strategies
715 in microbial data sets vary widely in sensitivity and precision. The ISME journal
716 **10**(7), 1669–1681 (2016)
- 717
- 718 [19] Berry, D., Widder, S.: Deciphering microbial interactions and detecting keystone
719 species with co-occurrence networks. Frontiers in microbiology **5**, 219 (2014)
- 720
- 721 [20] Guo, B., Zhang, L., Sun, H., Gao, M., Yu, N., Zhang, Q., Mou, A., Liu, Y.: Micro-
722 bial co-occurrence network topological properties link with reactor parameters
723 and reveal importance of low-abundance genera. npj Biofilms and Microbiomes
724 **8**(1), 3 (2022)
- 725
- 726 [21] Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J.A., Faust, K., Knight, R.,
727 Jansson, J.K., Cardona, C., *et al.*: Earth microbial co-occurrence network reveals
728 interconnection pattern across microbiomes. Microbiome **8**, 1–12 (2020)
- 729
- 730 [22] Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi,
731 Y., Audic, S., Berline, L., Brum, J.R., *et al.*: Plankton networks driving carbon
732 export in the oligotrophic ocean. Nature **532**(7600), 465–470 (2016)
- 733
- 734 [23] Röttjers, L., Faust, K.: Manta: A clustering algorithm for weighted ecological
735 networks. Msystems **5**(1), 10–1128 (2020)
- 736

- back. In: Evolutionary Systems Biology, pp. 329–345. Springer, ??? (2012) 737
 738
- [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) 739
 740
 741
 742
 743
 744
 745
- [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) 746
 747
 748
- [27] Parter, M., Kashtan, N., Alon, U.: Environmental variability and modularity of bacterial metabolic networks. BMC evolutionary biology **7**, 1–8 (2007) 749
 750
 751
 752
- [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) 753
 754
 755
 756
- [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) 757
 758
 759
 760
 761
- [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) 762
 763
 764
- [31] 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) 765
 766
 767
- [32] Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols **5**(1), 93–121 (2010) 768
 769
 770
 771
- [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) 772
 773
 774
 775
 776
 777
 778
- [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 779
 780
 781
 782
- [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) 783

- 783 [36] Zientz, E., Dandekar, T., Gross, R.: Metabolic interdependence of obligate intra-
784 cellular bacteria and their insect hosts. *Microbiology and Molecular Biology*
785 *Reviews* **68**(4), 745–770 (2004)
- 786
- 787 [37] Kallus, Y., Miller, J.H., Libby, E.: Paradoxes in leaky microbial trade. *Nature*
788 *communications* **8**(1), 1361 (2017)
- 789
- 790 [38] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., Kanehisa, M.:
791 Modular architecture of metabolic pathways revealed by conserved sequences of
792 reactions. *Journal of Chemical Information and Modeling* **53**(3), 613–622 (2013)
793 <https://doi.org/10.1021/ci3005379> <https://doi.org/10.1021/ci3005379>. PMID:
794 23384306
- 795
- 796 [39] Hessler, T., Huddy, R.J., Sachdeva, R., Lei, S., Harrison, S.T., Diamond, S.,
797 Banfield, J.F.: Vitamin interdependencies predicted by metagenomics-informed
798 network analyses and validated in microbial community microcosms. *Nature*
799 *Communications* **14**(1), 4768 (2023)
- 800
- 801 [40] Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad,
802 N., Dietrich, E.M., Disz, T., Gabbard, J.L., *et al.*: Improvements to patric, the
803 all-bacterial bioinformatics database and analysis resource center. *Nucleic acids*
804 *research* **45**(D1), 535–542 (2017)
- 805
- 806 [41] Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugen-
807 holtz, P.: Gtdb: an ongoing census of bacterial and archaeal diversity through
808 a phylogenetically consistent, rank normalized and complete genome-based
809 taxonomy. *Nucleic acids research* **50**(D1), 785–794 (2022)
- 810
- 811 [42] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies,
812 J., Glöckner, F.O.: The silva ribosomal rna gene database project: improved data
813 processing and web-based tools. *Nucleic acids research* **41**(D1), 590–596 (2012)
- 814
- 815 [43] Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya,
816 R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., *et al.*: Ncbi taxonomy:
817 a comprehensive update on curation, resources and tools. *Database* **2020**, 062
(2020)
- 818
- 819 [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>
- 820
- 821
- 822 [45] Murali, A., Bhargava, A., Wright, E.S.: Idtaxa: a novel approach for accurate
823 taxonomic classification of microbiome sequences. *Microbiome* **6**(1), 1–14 (2018)
- 824
- 825 [46] Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar,
826 G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., *et al.*: Structure and
827 function of the global ocean microbiome. *Science* **348**(6237), 1261359 (2015)
- 828

- [47] Louca, S., Parfrey, L.W., Doebeli, M.: Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**(6305), 1272–1277 (2016) 829
830
831
- [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) 832
833
834
835
836
837
- [49] Feldbauer, R., Schulz, F., Horn, M., Rattei, T.: Prediction of microbial phenotypes based on comparative genomics. *BMC bioinformatics* **16**(14), 1–8 (2015) 838
839
840
841
842
843
- [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) 844
845
846
847
- [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) 848
849
850
851
- [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) 852
853
854
855
- [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) 856
857
- [54] Kanehisa, M., Sato, Y.: Kegg mapper for inferring cellular functions from protein sequences. *Protein Science* **29**(1), 28–35 (2020) 858
859
860
- [55] Kanehisa, M., Sato, Y., Kawashima, M.: Kegg mapping tools for uncovering hidden features in biological data. *Protein Science* **31**(1), 47–53 (2022) 861
862
863
864
- [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) 865
866
867
868
- [57] 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) 869
870
871
872
873
874
- [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) 869
870
871
872
873
874

- 875 [59] Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson,
876 R., Overbeek, R., Parrello, B., Pusch, G.D., *et al.*: Rasttk: a modular and extensi-
877 ble implementation of the rast algorithm for building custom annotation pipelines
878 and annotating batches of genomes. *Scientific reports* **5**(1), 1–6 (2015)
879
- 880 [60] Blank, C.E., Cui, H., Moore, L.R., Walls, R.L.: Micro: an ontology of pheno-
881 typic and metabolic characters, assays, and culture media found in prokaryotic
882 taxonomic descriptions. *Journal of biomedical semantics* **7**(1), 1–10 (2016)
883
- 884 [61] Choudhary, K., Meng, E.C., Diaz-Mejia, J.J., Bader, G.D., Pico, A.R., Morris,
885 J.H.: scnetviz: from single cells to networks using cytoscape. *F1000Research* **10**
886 (2021)
- 887 [62] Merkel, D., *et al.*: Docker: lightweight linux containers for consistent development
888 and deployment. *Linux j* **239**(2), 2 (2014)
889
- 890 [63] Reese, W.: Nginx: The high-performance web server and reverse proxy. *Linux J.*
891 **2008**(173) (2008)
892
- 893 [64] Wright, E.S.: Using decipher v2. 0 to analyze big biological sequence data in r. *R*
894 *Journal* **8**(1) (2016)
895
- 896 [65] Llavero-Pasquina, M., Geisler, K., Holzer, A., Mehrshahi, P., Mendoza-Ochoa,
897 G.I., Newsad, S.A., Davey, M.P., Smith, A.G.: Thiamine metabolism genes
898 in diatoms are not regulated by thiamine despite the presence of predicted
899 riboswitches. *New Phytologist* **235**(5), 1853–1867 (2022)
900
- 901 [66] Romine, M.F., Rodionov, D.A., Maezato, Y., Osterman, A.L., Nelson, W.C.:
902 Underlying mechanisms for syntrophic metabolism of essential enzyme cofactors
903 in microbial communities. *The ISME journal* **11**(6), 1434–1446 (2017)
904
- 905 [67] Chklovski, A., Parks, D.H., Woodcroft, B.J., Tyson, G.W.: Checkm2: a rapid,
906 scalable and accurate tool for assessing microbial genome quality using machine
907 learning. *Nature Methods* **20**(8), 1203–1212 (2023)
908
- 909 [68] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W.:
910 Checkm: assessing the quality of microbial genomes recovered from isolates, single
911 cells, and metagenomes. *Genome research* **25**(7), 1043–1055 (2015)
912
- 913 [69] Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of
914 items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International*
915 *Conference on Management of Data*. SIGMOD '93, pp. 207–216. Association for
916 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>
917
- 918
- 919
- 920