

	001
	002
	003
	004
	005
Genomic, metabolic and literature oriented annotation of microbial co-occurrence networks enhances associations confidence level and hypothesis generation	006
	007
	008
	009
	010
	011
	012
	013
	014
Haris Zafeiropoulos ¹ , Ermis Ioannis Michail Delopoulos ¹ , Andi Erega ² , Annelies Geirnaert ² , John Morris ³ , Karoline Faust ^{1*}	015
^{1*} Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research , KU Leuven, Herestraat, Leuven, 3000, , Belgium .	016
² Institute of Food, Nutrition and Health, ETH Zurich, Street, Zurich, 8092, , Switzerland .	017
³ Department of Pharmaceutical Chemistry, University of California San Francisco, Street, San Francisco, 94143, California, USA .	018
	019
	020
	021
	022
	023
	024
	025
	026
	027
*Corresponding author(s). E-mail(s): karoline.faust@kuleuven.be ;	028
Contributing authors: haris.zafeiropoulos@kuleuven.be ;	029
ermisioannis.michaildelopoulos@student.kuleuven.be ;	030
andi.erega@hest.ethz.ch ; annelies.geirnaert@hest.ethz.ch ;	031
scooter@cgl.ucsf.edu ;	032
	033
	034
Abstract	035
Up to 350 words.	036
The abstract must include the following separate sections:	037
Background: the context and purpose of the study	038
Results: the main findings	039
Conclusions: a brief summary and potential implications	040
	041
	042
	043
	044
	045
	046

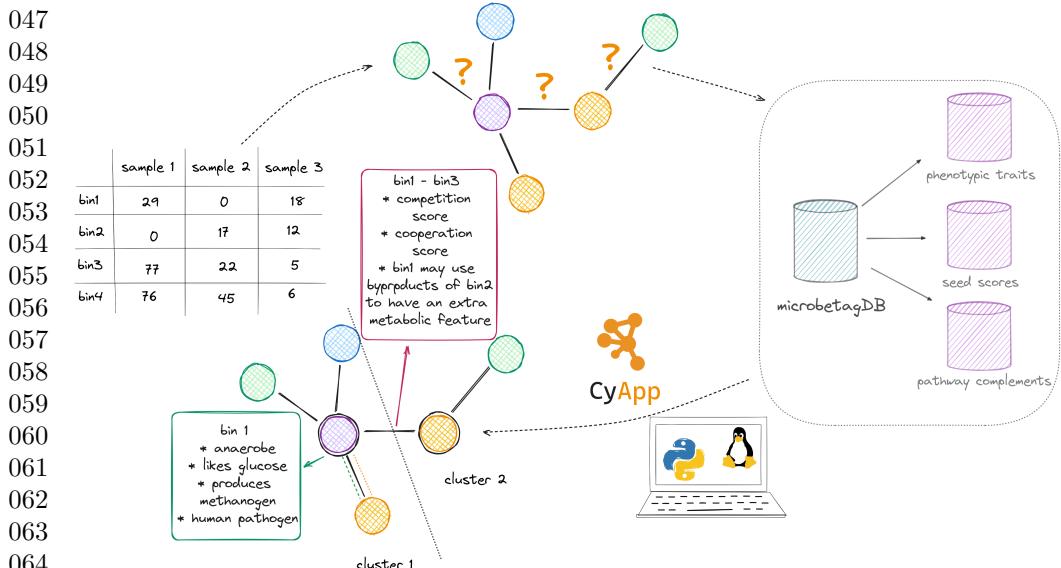


Figure abstract.

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

072 1 Introduction

- 073 1 2
074
075
076
-
- 077 • Role of microbial associations in the study of microbial communities
 - 078 • HTP sequencing
 - 079 • cooccurrence networks to analyse HTP data
 - 080 • challenges of cooccurrence networks / refs: [1]: microbial network inference as a tool
081 for interaction prediction has highlighted this tool's low accuracy and the biological
082 implications of network properties are unclear / [2] : modules in microbial co-
083 occurrence networks may be indicative of ecological processes governing community
084 structure, such as niche filtering and habitat preference
 - 085 • current approaches to deal with those challenges
 - 086 • our contribution

087
088 It is well known that most microbial species live not in isolation but in communi-
089 ties [3]. Such communities play a crucial role in ecosystem functioning in almost
090

091
092 ¹We are to submit in the Microbiome journal as a "Software" manuscript, thus we follow [these rules](#)
092 ²The introduction should not include subheadings.

any ecosystem type [4, 5]. High-throughput sequencing (HTP) has provided great insight into the diversity and composition of microbial communities [6]. Uncultivated species can now be detected and their features can be inferred through their genomic information [7].	093
NGS facilitates culture-independent sampling of the microorganisms in an area with the potential for both taxonomic and functional annotation;	094
NGS not only provides information about the taxonomic composition of the microbial community but also enables the annotation of their functional capabilities. This combination of taxonomic and functional annotation provides a more detailed and holistic view of the role and activities of microorganisms.	095
Understanding the structure, composition, and dynamics of microbial communities is crucial for unraveling the intricate web of interactions that shape ecosystems. To come up with a comprehensive understanding of these interactions, we need to comprehend the relationships between microorganisms and thus their own interactions.	096
The notion of an interaction varies including cooperation, competition, parasitism, commensalism and amensalism [5].	097
Microbial interactions play a crucial role in shaping ecosystems and influencing various biological processes.	098
These interactions contribute to the overall stability and functioning of ecosystems by influencing nutrient cycling, disease dynamics, and the overall diversity and composition of microbial communities. For example, some microorganisms can produce and release certain compounds that benefit neighboring microbes or inhibit the growth of competing species. Other interactions involve symbiotic relationships, where different microbes rely on each other for survival and perform complementary functions. Understanding and studying microbial interactions is vital for unraveling the complexity of microbial communities and their impact on human health, agriculture, and environmental processes.	099
Physical Interactions: Microorganisms can interact physically by forming biofilms, which are communities of microorganisms attached to surfaces and enclosed in a matrix of extracellular substances. Biofilms provide protection and promote cooperation among the microorganisms within them.	100
Quorum Sensing: Many bacteria use quorum sensing to communicate with each other and coordinate their behavior. They release and detect signaling molecules called autoinducers, which allow them to sense the local population density. This mechanism enables bacteria to coordinate processes like biofilm formation, virulence factor expression, and nutrient acquisition.	101
Antibiosis: Antibiosis refers to the production of antimicrobial substances by microorganisms that inhibit the growth or survival of other microorganisms. This can be a competitive strategy to gain an advantage in a particular environment.	102
metabolic interactions specifically, microorganisms often engage in metabolic exchanges during their interactions. For instance, one microbe may produce metabolites that serve as nutrients or signaling molecules for another microbe. These metabolic interactions can involve the transfer of essential nutrients, the breakdown of complex compounds, or the production of secondary metabolites with antimicrobial properties. Overall, understanding the different types of microbial interactions,	103
	104
	105
	106
	107
	108
	109
	110
	111
	112
	113
	114
	115
	116
	117
	118
	119
	120
	121
	122
	123
	124
	125
	126
	127
	128
	129
	130
	131
	132
	133
	134
	135
	136
	137
	138

139 including metabolic interactions, provides insights into the complexity and dynamics
140 of microbial communities and their impact on various processes in nature.

141 A widely used approach is the creation of co-occurrence networks based on commu-
142 nity data. To build such networks, there is a great number of approaches: Spearman
143 and Pearson correlations, CoNet [8] SparCC [9] SpieEasi [10], MAGMA [11] and
144 FlashWeave [12] are just a few of them. However, challenges [13] the outcome is usually
145 tool-dependent [3, 14, 15].

146 Metagenomic or metabarcoding data are often used to predict microbial interac-
147 tions in complex communities, but these predictions are rarely explored experimentally
148 and/or validated.

149 FlashWeave

150 microbeAnnotator [16]

151 Related literature: Karaoz U and Brodie EL (2022) microTrai [17], a computational
152 pipeline that infers and distills ecologically relevant traits from microbial genome
153 sequences. It does not apply networks

154

155 **2 Implementation**

156

157 ³

158

159 **Genomes included**

160

161 Using the GTDB v202 [metadata files](#), we retrieved the NCBI genome accessions of the
162 representative genomes of high quality, i.e. completeness $\geq 95\%$ and contamination
163 $\leq 5\%$. That resulted a set of 26,778 covering 22,009 unique NCBI Taxonomy Ids.
164 Using these accession numbers, we were able to download their corresponding .faa
165 files when available ([get_gtdb_faa.py](#)) leading to a set of 16,900 amino acid sequence
166 files.

167

168 **Taxonomy schemes**

169

170 microbetag maps the taxonomy of each entry in the abundance table to its correspond-
171 ing NCBI Taxonomy id and if available its closest GTDB representative genome(s).
172 Two well established taxonomy schemes are supported. The Genome Taxonomy
173 DataBase (GTDB) [18] that is being broadly used in bins and/or MAGs taxonomical
174 classification and the Silva database [19] that has NCBI Taxonomy [20]. The primer
175 links the representative genomes included to their corresponding NCBI Taxonomy ids
176 too.

177

178 There is a great number of taxonomies that are being used in such studies, e.g.
179 Silva [19], Ribosomal Database Project (RDP) [21], manually curated ones and more,
180 As a consequence, there is not a standardised format of the taxonomies assigned, from
181 bioinformatics pipelines used for the analysis of such data. microbetag makes use of
182 the [fuzzywuzzy](#) library that implements the Levenshtein Distance Metric to get the

183

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.

184

closest NCBI taxon name and thus its corresponding NCBI Taxonomy id. ++ ncbi
nodes dump A relatively high similarity score is used (90) to avoid false positives. 185
186

DADA2 formatted 16S rRNA gene sequences for both bacteria and archaea [22]
were used to trained the TAXID classifier [23] of the DECIPHER package. 187
188
189
190
191
192
193
194
195
196
197
198
199
200
201

Network inference

FlashWeave [12]

a computational approach based on a flexible Probabilistic Graphical Model
framework that integrates metadata and predicts direct microbial interactions from
heterogeneous microbial abundance data sets with hundreds of thousands of samples.

A flexible Probabilistic Graphical Model framework is used in a computational
approach that incorporates metadata and predicts direct microbial interactions. This
is done using heterogeneous microbial abundance datasets consisting of hundreds of
thousands of samples.

Literature oriented node annotation

Using a set of Tara Oceans samples [24] FAPROTAX [25] estimates the functional
potential of the bacterial and archaeal communities, by classifying each taxonomic
unit into functional group(s) based on current literature, announcements of cultured
representatives and/or manuals of systematic microbiology. In this manually curated
approach, a taxon is associated with a function if and only if all the cultured species
within the taxon have been shown to exhibit that function. In its current version,
FAPROTAX includes more than 80 functions based on 7600 functional annotations
and covering more than 4600 taxa. Contrary to gene content based approaches,
e.g. PICRUSt2, FAPROTAX estimates metabolic phenotypes based on experimental
evidence.

microbetag invokes the accompanying script of FAPROTAX and converts the taxonomic
microbial community profile of the samples included in the user's abundance
table or of the taxa present in the provided network, into putative functional profiles.
Then, it parses FAPROTAX's subtables to annotate each taxonomic unit present
on 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 oriented node annotation

phenDB [26] is a publicly available resource that supports the analysis of bacterial
(meta)genomes to identify 47 distinct functional traits. 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. The
`compute-genotype` program of phenotrex supports the creation of such tabular *genotype*
files. A *genotype* file can be used along with a *phenotype* one, i.e., a file containing
true phenotypic trait values for each input genome on which to train the model, and
the `train` program of phenotrex can then be performed. Last, the models can now be

231 used to predict their corresponding traits; based on the completeness/contamination
232 of the genomes, the accuracy varies.

233 In the frameowrk of microbetagDB, phenotrex classifiers were re-trained using the
234 genomes provided by phenDB for each trait to sync with the latest version of eggNOG.
235 Genomes were downloaded from NCBI using the [Batch Entrez](#) program. Then, *geno-*
236 *type* files were produced for all the high quality GTDB representative genomes. Each
237 model was then used against all the GTDB *genotype* files to annotate each with the
238 presence or the absence of the trait.

239

240 Pathway complementarity

241

242 For the subset of the 16,900 high quality GTDB representative genomes that a *.faa*
243 was available, *kofamscan* [27] was performed to annotate them with KEGG ORTHOL-
244 OGY terms (KOs) [28]. Their KOs were then mapped to their corresponding KEGG
245 modules. A KEGG module is defined as a functional unit within the KEGG frame-
246 work, that represents a set of enzymes and reactions involved in a specific biological
247 process or pathway [29]. A module's definition is a logical expression and consists of
248 KOs and the following symbols: a. the space, representing a connection in the pathway
249 b. plus sign, representing a molecular complex, c. comma, representing alternatives
250 and d. minus sign, designates an optional item in the complex. Both (a) and (b) cases
251 should be considered as "AND" logical operators, while (c) would be the "OR".

252 We define a genome as having a "complete" module if and only if all of the
253 KOs present in any of the module's alternatives are also found among the anno-
254 tated KOs of the genome. All modules definitions were retrieved using the KEGG
255 API and parsed ([parse_module_definitions.py](#)). A dictionary was built with all
256 the alternatives, i.e. alternative sets of KOs, for a module to be complete ([mod-](#)
257 [ule_definition_map.json](#)). Each pair of the KEGG annotated genomes was then
258 investigated for potential pathway complementarities, i.e. whether a genome lacking
259 a number of KOs (*genome_A*) to have a complete module (*module_x*) could benefit
260 from another's species genome(s) (*genome_B*). In that case, *genome_B* does not nec-
261 essarily have a complete alternative of *module_x*; as long as it has the missing KOs
262 that *genome_A* needs to complete an alternative of it, *genome_B* potentially comple-
263 ments *genome_A* with respect to *module_x*. In total, 341,568 unique complementarities
264 were exported ([pathway_complementarity.py](#)). Thanks to the graphical user inter-
265 face (GUI) of the [KEGG pathway map viewer](#) [30, 31], each complementarity can
266 be visualised as part of the closest KEGG metabolic map; where the KOs coming
267 from the donor are shown with a blue-green colour, while those from the beneficiary's
268 genome itself with rose.

269 As several GTDB representative genomes might map to the same NCBI Taxonomy
270 Id, all the possible genomes' combinations are annotated in the edge of a pair of species
271 level taxonomically annotated OTUs/ASVs/bins. On top of that, as co-occurrence net-
272 works are undirected, both nodes of a suggested association are considered as potential
273 donors and beneficiary species.

274

275

276

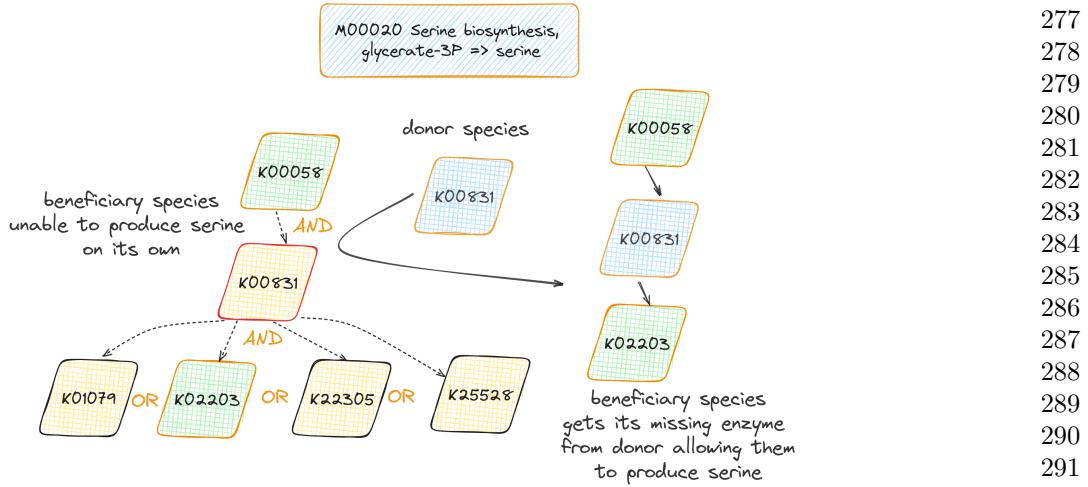


Fig. 1: Pathway complementarity approach. The high quality GTDB genomes were annotated with KEGG ORTHOLOGY (KO) terms. The various ways of getting a KEGG module complete were enumerated and all the possible ways a donor species could "fill" a beneficiary's non-complete module were calculated. In this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

Seed scores using genome scale metabolic reconstructions

A metabolic network's "seed set" is the set of compounds that, based on the network topology, need to be acquired exogenously [32]. Such nodes might be independent, i.e. they cannot be activated by any other node in the network, or they can be interdependent forming groups of seed nodes.

Based on the seed concept, several graph theory-based metrics have been described to predict species interactions directly from their networks' topologies. 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 [33], 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 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 [34] and NetCompt [35] tools correspondingly. We will be referring to those two indices as "seed scores". Most recently, the PhyloMint Python package [33] was

323 released supporting the calculation of the seed scores of genome scale metabolic
324 network reconstructions (GENREs) in SBML format.

325 In the framework of microbetag, seed scores were computed using PhyloMint and
326 draft GENREs for all pair-wised combinations of GTDB representative genomes that
327 have been RAST annotated in the framework of the PATRIC database [36]. GENREs
328 were reconstructed using the Model SEED pipeline [37] through its Python interface
329 [ModelSEEDpy](#).

330

331 **Clustering network**

332 manta is a heuristic network clustering algorithm that clusters nodes within weighted
333 networks effectively, leveraging the presence of negative edges and discerning between
334 weak and microbetag invokes manta [38] to infer clusters from the microbial network.
335 A taxonomically-informed layout is
336 strong cluster assignments. ++ taxonomy layout

337

338 **Groups of annotations**

340 Biologically meaningful groups were built using the micrO ontology [39].

341

342 **Building the CytoscapeApp**

344 The microbetag CytoscapeApp was build based on the [source code](#) of the scVizNet [40].

345 Java @Ermis to add
346 Enrichment analysis is supported. Hypergeometric distribution FDR +++

347

348 **Dependencies, Web server and API**

349 The microbetag web service is container - based and consists of three Docker [41]
350 (v24.0.2) images: a. the [MySQL](#) database b. an nginx [42] web server and c. the app
351 itself. The latter uses [Gunicorn](#) (20.1.0) to build an application server which commu-
352 nicates with the web server using the Web Server Gateway Interface (WSGI) protocol
353 and handles incoming HTTP requests. microbetag is implemented as a [Flask](#) applica-
354 tion (v2.3.2); Flask is a micro web framework for developing Python web applications
355 and RESTful APIs. A thorough description of microbetag's API is available at the
356 [ReadTheDocs web site](#). The source code of the microbetag web service is available
357 on [GitHub](#).
358 python 3.11 slim docker image julia 1.7.1 for flashweave mysql.connector 8.0.27
359 python library pandas 2.1.1. numpy 1.26.0 multiprocessing
360 text processing using awk
361 KEGG API

363

364

365

366

367

368

2.1 Running large datasets	369
3 Results	370
4	371
<hr/>	
⁴ Significant advance over previously published software (usually demonstrated by direct comparison with available related software) This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures. This section may be combined with the Discussion section for Software articles.	372
373	373
374	374
375	375
376	376
377	377
378	378
379	379
380	380
381	381
382	382
383	383
384	384
385	385
386	386
387	387
388	388
389	389
390	390
391	391
392	392
393	393
394	394
395	395
396	396
397	397
398	398
399	399
400	400
401	401
402	402
403	403
404	404
405	405
406	406
407	407
408	408
409	409
410	410
411	411
412	412
413	413
414	414

415 microbetag and microbetagDB

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

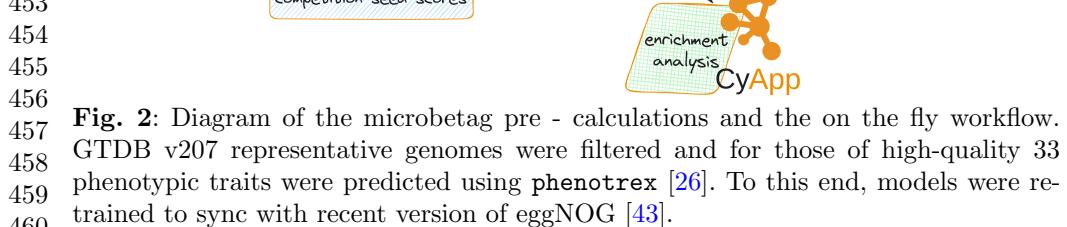


Fig. 2: Diagram of the microbetag pre-calculations and the on-the-fly workflow. GTDB v207 representative genomes were filtered and for those of high-quality 33 phenotypic traits were predicted using *phenotrex* [26]. To this end, models were re-trained to sync with recent version of eggNOG [43].

microbetag in numbers: 34, 608 GTDB representative genomes 32 phen-model-oriented metabolic functions 92 FAPROTAX functions 341, 568 unique complements involved in > 184 million beneficiary - donor pairs' complementarities 30, 755 GENREs leading to 1 billion competition and complementarity scores	461		
annotated network returned in .cyjs format	462		
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 [23] of the DECIPHER R package [44]. A co-occurrence network is also built using FlashWeave [12], as microbetag also does.	463		
	464		
	465		
	466		
	467		
	468		
	469		
	470		
	471		
	472		
microbetag CytoscapeApp	473		
Overall comment, the CytoscapeApp returns averages and s.d. for example in seed scores. If you want the exact values, go through the API.	474		
	475		
A. GTDB-tk: 480 bins	476		
<hr/>			
Step	Time(sec)	Notes	
Taxonomy mapping	Cell 1,2	on the fly	
Network inference	Cell 2,2	on the fly	
microbetag annotations	Cell 3,2	on the fly	
manta clustering	Cell 4,2	on the fly	
B. GTDB 16S: 3000 ASVs	477		
<hr/>			
Step	Time(sec)	Notes	
Taxonomy assignment		Docker image on ^{HP} ⁵	478
Taxonomy mapping	Cell 1,2	Cell 1,3	479
Network inference	Cell 2,2	Cell 2,3	480
microbetag annotations	Cell 3,2	Cell 3,3	481
manta clustering	Cell 4,2	Cell 4,3	482
			483
C. Silva:	484		
<hr/>			
Step	Time(sec)	Notes	
Taxonomy mapping	Cell 1,2	Cell 1,3	
Network inference	Cell 2,2	Cell 2,3	
microbetag annotations	Cell 3,2	Cell 3,3	
manta clustering	Cell 4,2	Cell 4,3	
D. fuzzywuzzy:	485		
<hr/>			
Step	Time(sec)	Notes	
Taxonomy mapping	Cell 1,2	Cell 1,3	486
Network inference	Cell 2,2	Cell 2,3	487
microbetag annotations	Cell 3,2	Cell 3,3	488
manta clustering	Cell 4,2	Cell 4,3	489
The app was based on the StringApp and supported by the NRNB group.	490		
	491		
	492		
	493		
	494		
	495		
Validation of microbetag potential	496		
vitamin dataset [45]	497		
Metagenomic or metabarcoding data are often used to predict microbial interactions in complex communities, but these predictions are rarely explored experimentally. Here, we use an organism abundance correlation network to investigate factors that control community organization in mine tailings-derived laboratory microbial consortia grown under dozens of conditions.	498		
The network is overlaid with metagenomic information about functional capacities to generate testable hypotheses.	499		
	500		
	501		
	502		
	503		
	504		
	505		
	506		

507 **Interpetating a real-world network with microbetag**

508 Annelies' dataset.
509

510 **4 Discussion**

512 [6](#)
513

514 **5 Conclusions**

516 [7](#)
517 Data integration
518

519 **Supplementary information.** [8](#)

520

521 **Declarations**

522

523 • **Availability of data and materials**

524 – Raw sequences for the use case:
525 – Raw data for the validations case:

527 • **Funding**

528 This work was initiated thanks to an EMBO Scientific Exchange Grant to HZ. It
529 was then supported by the 3D'omics Horizon project (101000309). We would also
530 like to thank the National Resource for Network Biology (NRNB) and the Google
531 Summer of Code 2023 for the support of E.I.M.D.

532 • **Conflict of interest/Competing interests**

533 The authors declare that they have no other competing interests.

534 • **Authors' contributions** [9](#)

535 Conceptualization: K.F. Methodology: K.F. and H.Z. Software: H.Z., E.I.M.D. and
536 J.M Validation: H.Z. and K.F. Formal analysis: H.Z. and K.F. Investigation: H.Z.
537 Resources: K.F., A.E. and A.G. Data Curation: H.Z. Writing - Original Draft: H.Z.
538 and K.F. Writing - Review & Editing: all Visualization: H.Z. Supervision: K.F.,
539 H.Z. and S.M. Project administration: K.F. Funding acquisition: K.F., A.E.

540 • **Acknowledgements**

541 We would like to thank Dr Christina Pavloudi and ++ for the insight on how to
542 organise the trait groups.

543 • **Ethics approval**

544 Not applicable

545

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

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

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

552 ⁹Based on the [CRediT system](#). Current list is indicative.

• Consent to participate	553
Not applicable.	554
• Code availability:	555
– microbetagDB related scripts: https://github.com/hariszaf/microbetag	556
– microbetagApp and webserver: https://github.com/msysbio/microbetagApp .	557
– CytoscapeApp: https://github.com/ermismd/MGG/	558
– Validation and use case: {think of having that under the 3D'omics organization};	559
– Documentation web-site: https://hariszaf.github.io/microbetag/	560
	561
	562
	563
	564
	565

Appendix A Background on seed scores and complementarities

A.1 Background on seed scores

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

$$C_i = 1/\text{seed's group with } i \text{ size} \quad (\text{A1})$$

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

$$MI_{Complementarity} = \frac{|SeedSet_A \cap \neg SeedSet_B|}{|SeedSet_A \cap (SeedSet_B \cup \neg SeedSet_B)|} \quad (\text{A2})$$

As also described in [33], it is calculated as the proportion of compounds in a species' seed set that coincide with those in an other's, while also factoring in the confidence scores associated with seed compounds.

$$MI_{Competition} = \frac{\sum C(SeedSet_A \cap SeedSet_B)}{\sum C(SeedSet_A)} \quad (\text{A3})$$

A.2 Background on pathway complementarity

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

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

K01812 K00041 K01685 K00874 K01625

K01812 K00041 K16849+K16850 K00874 K01625

K01812 K00041 K01685 K00874 K17463

K01812 K00041 K16849+K16850 K00874 K17463

599 **A.3 Complementarities**

600 KEGG compound ModelSEED compounds ModelSEED compounds mapped to
601 KEGG compounds and kept only those related to KEGG modules.
602

603 **References**

- 605 [1] Berry, D., Widder, S.: Deciphering microbial interactions and detecting keystone
606 species with co-occurrence networks. *Frontiers in microbiology* **5**, 219 (2014)
- 608 [2] Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J.A., Faust, K., Knight, R.,
609 Jansson, J.K., Cardona, C., *et al.*: Earth microbial co-occurrence network reveals
610 interconnection pattern across microbiomes. *Microbiome* **8**, 1–12 (2020)
- 612 [3] Röttjers, L., Faust, K.: From hairballs to hypotheses—biological insights from
613 microbial networks. *FEMS microbiology reviews* **42**(6), 761–780 (2018)
- 615 [4] Raes, J., Bork, P.: Molecular eco-systems biology: towards an understanding of
616 community function. *Nature Reviews Microbiology* **6**(9), 693–699 (2008)
- 618 [5] Faust, K., Raes, J.: Microbial interactions: from networks to models. *Nature
619 Reviews Microbiology* **10**(8), 538–550 (2012)
- 621 [6] Finn, R., Balech, B., Burgin, J., Chua, P., Corre, E., Cox, C., Donati, C., Santos,
622 V., Fosso, B., Hancock, J., Heil, K., Ishaque, N., Kale, V., Kunath, B., Médigue,
623 C., Pafilis, E., Pesole, G., Richardson, L., Santamaria, M., Van Den Bossche, T.,
624 Vizcaíno, J., Zafeiropoulos, H., Willassen, N., Pelletier, E., Batut, B.: Establishing
625 the elixir microbiome community [version 1; peer review: awaiting peer review].
626 *F1000Research* **13**(50) (2024) <https://doi.org/10.12688/f1000research.144515.1>
- 628 [7] Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle,
629 C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., *et al.*: A new view
630 of the tree of life. *Nature microbiology* **1**(5), 1–6 (2016)
- 631 [8] Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J.,
632 Huttenhower, C.: Microbial co-occurrence relationships in the human microbiome.
633 *PLoS computational biology* **8**(7), 1002606 (2012)
- 635 [9] Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data.
636 *PLoS computational biology* **8**(9), 1002687 (2012)
- 638 [10] Kurtz, Z.D., Müller, C.L., Miralldi, E.R., Littman, D.R., Blaser, M.J., Bon-
639 neau, R.A.: Sparse and compositionally robust inference of microbial ecological
640 networks. *PLoS computational biology* **11**(5), 1004226 (2015)
- 642 [11] Cougoul, A., Bailly, X., Wit, E.C.: Magma: inference of sparse microbial
643 association networks. *BioRxiv*, 538579 (2019)

- [12] 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) 645
646
647
648
649
650
651
652
653
654
655
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
- [13] Faust, K.: Open challenges for microbial network construction and analysis. *The ISME Journal* **15**(11), 3111–3118 (2021)
- [14] 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)
- [15] 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)
- [16] Ruiz-Perez, C.A., Conrad, R.E., Konstantinidis, K.T.: Microbeannotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC bioinformatics* **22**, 1–16 (2021)
- [17] Karaoz, U., Brodie, E.L.: microtrait: a toolset for a trait-based representation of microbial genomes. *Frontiers in Bioinformatics* **2**, 918853 (2022)
- [18] 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)
- [19] 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)
- [20] 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)
- [21] Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M.: Ribosomal database project: data and tools for high throughput rrna analysis. *Nucleic acids research* **42**(D1), 633–642 (2014)
- [22] 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>
- [23] Murali, A., Bhargava, A., Wright, E.S.: Idtaxa: a novel approach for accurate

- 691 taxonomic classification of microbiome sequences. *Microbiome* **6**(1), 1–14 (2018)
- 692
- 693 [24] Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar,
694 G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., *et al.*: Structure and
695 function of the global ocean microbiome. *Science* **348**(6237), 1261359 (2015)
- 696
- 697 [25] Louca, S., Parfrey, L.W., Doebeli, M.: Decoupling function and taxonomy in the
698 global ocean microbiome. *Science* **353**(6305), 1272–1277 (2016)
- 699
- 700 [26] Feldbauer, R., Schulz, F., Horn, M., Rattei, T.: Prediction of microbial phenotypes
701 based on comparative genomics. *BMC bioinformatics* **16**(14), 1–8 (2015)
- 702
- 703 [27] Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto,
704 S., Ogata, H.: Kofamkoala: Kegg ortholog assignment based on profile hmm and
705 adaptive score threshold. *Bioinformatics* **36**(7), 2251–2252 (2020)
- 706
- 707 [28] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: Kegg for integration
708 and interpretation of large-scale molecular data sets. *Nucleic acids research*
709 **40**(D1), 109–114 (2012)
- 710
- 711 [29] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., Kanehisa, M.:
712 Modular architecture of metabolic pathways revealed by conserved sequences of
713 reactions. *Journal of chemical information and modeling* **53**(3), 613–622 (2013)
- 714
- 715 [30] Kanehisa, M., Sato, Y.: Kegg mapper for inferring cellular functions from protein
716 sequences. *Protein Science* **29**(1), 28–35 (2020)
- 717
- 718 [31] Kanehisa, M., Sato, Y., Kawashima, M.: Kegg mapping tools for uncovering
719 hidden features in biological data. *Protein Science* **31**(1), 47–53 (2022)
- 720
- 721 [32] Borenstein, E., Kupiec, M., Feldman, M.W., Ruppin, E.: Large-scale reconstruc-
722 tion and phylogenetic analysis of metabolic environments. *Proceedings of the
723 National Academy of Sciences* **105**(38), 14482–14487 (2008)
- 724
- 725 [33] Lam, T.J., Stamboulian, M., Han, W., Ye, Y.: Model-based and phylogenetically
726 adjusted quantification of metabolic interaction between microbial species. *PLoS
727 computational biology* **16**(10), 1007951 (2020)
- 728
- 729 [34] Levy, R., Carr, R., Kreimer, A., Freilich, S., Borenstein, E.: Netcooperate: a
730 network-based tool for inferring host-microbe and microbe-microbe cooperation.
731 *BMC bioinformatics* **16**(1), 1–6 (2015)
- 732
- 733 [35] Kreimer, A., Doron-Faigenboim, A., Borenstein, E., Freilich, S.: Netcmpt: a
734 network-based tool for calculating the metabolic competition between bacterial
735 species. *Bioinformatics* **28**(16), 2195–2197 (2012)
- 736 [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) 737
 738
 739
- [37] 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) 740
 741
 742
- [38] Röttgers, L., Faust, K.: Manta: A clustering algorithm for weighted ecological networks. Msystems **5**(1), 10–1128 (2020) 743
 744
 745
- [39] Blank, C.E., Cui, H., Moore, L.R., Walls, R.L.: Micro: an ontology of phenotypic and metabolic characters, assays, and culture media found in prokaryotic taxonomic descriptions. Journal of biomedical semantics **7**(1), 1–10 (2016) 746
 747
 748
 749
- [40] 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) 750
 751
 752
 753
- [41] Merkel, D., *et al.*: Docker: lightweight linux containers for consistent development and deployment. Linux j **239**(2), 2 (2014) 754
 755
 756
- [42] Reese, W.: Nginx: The high-performance web server and reverse proxy. Linux J. **2008**(173) (2008) 757
 758
 759
- [43] 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) 760
 761
 762
 763
 764
- [44] Wright, E.S.: Using decipher v2. 0 to analyze big biological sequence data in r. R Journal **8**(1) (2016) 765
 766
 767
- [45] 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) 768
 769
 770
 771
 772
 773
 774
 775
 776
 777
 778
 779
 780
 781
 782