

Manuscript Number:	GIGA-D-21-00316
Full Title:	Inferring microbiota functions from taxonomic genes: a review
Article Type:	Review
Funding Information:	
Abstract:	Deciphering microbiota functions is crucial to predict ecosystem sustainability in response to global change. High-throughput sequencing at the individual or community level has revolutionized our understanding of microbial ecology, leading to the big data era and improving our ability to link microbial diversity with microbial functions. Recent advances in bioinformatics have been key for developing functional prediction tools based on DNA metabarcoding data and using taxonomic gene information. This cheaper approach in every aspect serves as an alternative to shotgun sequencing. Although these tools are increasingly used by ecologists, an objective evaluation of their modularity, portability and robustness is lacking. Here, we reviewed one hundred scientific papers on functional inference and ecological trait assignment to rank the advantages, specificities and drawbacks of these tools, using a scientific benchmarking. To date, inference tools have been mainly devoted to bacterial functions, and ecological trait assignment tools to fungal functions. A major limitation is the lack of reference genomes – compared with the human microbiota –, especially for complex ecosystems like soils. In fine, we explore applied research prospects. These tools are very promising and already provide relevant information on ecosystem functioning, but standardized indicators and corresponding repositories are still lacking for them to be used for operational diagnosis.
Corresponding Author:	Lionel Ranjard, Ph.D. INRA UMR 1347: Agroecologie Dijon, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	INRA UMR 1347: Agroecologie
Corresponding Author's Secondary Institution:	
First Author:	Christophe Djemiel, Ph.D.
First Author Secondary Information:	
Order of Authors:	Christophe Djemiel, Ph.D. Pierre-Alain Maron, Ph.D. Sébastien Terrat, Ph.D. Samuel Dequiedt Aurélien Cottin, Ph.D. Lionel Ranjard, Ph.D.
Order of Authors Secondary Information:	
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes

<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

1 **Inferring microbiota functions from taxonomic genes: a review**

2

3 Christophe Djemiel¹, Pierre-Alain Maron¹, Sébastien Terrat¹, Samuel Dequiedt¹,
4 Aurélien Cottin¹, Lionel Ranjard¹

5

6 Authors' affiliations

7 ¹ Agroécologie, AgroSup Dijon, INRAE, Univ. Bourgogne, Univ. Bourgogne Franche-
8 Comté, F-21000 Dijon, France.

9

10 **Corresponding author**

11 Correspondence to Lionel Ranjard: lionel.ranjard@inrae.fr

12

13 **Abstract**

14 Deciphering microbiota functions is crucial to predict ecosystem sustainability in re-
15 sponse to global change. High-throughput sequencing at the individual or community
16 level has revolutionized our understanding of microbial ecology, leading to the big data
17 era and improving our ability to link microbial diversity with microbial functions. Recent
18 advances in bioinformatics have been key for developing functional prediction tools
19 based on DNA metabarcoding data and using taxonomic gene information. This
20 cheaper approach in every aspect serves as an alternative to shotgun sequencing.
21 Although these tools are increasingly used by ecologists, an objective evaluation of
22 their modularity, portability and robustness is lacking. Here, we reviewed one hundred
23 scientific papers on functional inference and ecological trait assignment to rank the
24 advantages, specificities and drawbacks of these tools, using a scientific benchmark-

25 ing. To date, inference tools have been mainly devoted to bacterial functions, and eco-
26 logical trait assignment tools to fungal functions. A major limitation is the lack of refer-
27 ence genomes – compared with the human microbiota –, especially for complex eco-
28 systems like soils. *In fine*, we explore applied research prospects. These tools are very
29 promising and already provide relevant information on ecosystem functioning, but
30 standardized indicators and corresponding repositories are still lacking for them to be
31 used for operational diagnosis.

32

33 **Keywords**

34 Microbiota; Metabarcoding; Taxonomy; Functional inference; Ecological traits; Soil

35

36 **1. Background**

37 Microorganisms are present in all habitats on Earth and are essential for
38 animals, plants, and therefore for the sustainability of human activities [1]. The
39 extraordinary diversity of microbial communities plays an essential role in the various
40 biogeochemical cycles, allows aquatic and terrestrial ecosystems to function properly
41 and ensures their ability to provide ecological services (*e.g.*, soil structuring, organic
42 matter renewal, nutrient recycling, pollution control, regulation of / barrier to pathogens,
43 or even plant productivity) [2–4]. Their fabulous capacity to adapt to different
44 environmental stresses over time is now well known, and the regulation process of
45 their diversity is better and better deciphered. Despite these tremendous
46 improvements in the approaches targeting indigenous microbiotas, our understanding
47 of the link between microbes and their associated functions remains limited [5]. A
48 workshop hosted by the British Ecological Society’s Microbial Ecology Special Interest
49 Group (June 2016) recently identified fifty important research questions in microbial

50 ecology. One of the main ones was “What methods can we use to marry microbial
51 diversity with function; how do we link transcriptomics, proteomics and metabolomics?”
52 [6]. This sums up the future challenges facing the scientific community when it comes
53 to improving our understanding of the regulation of the microbiome diversity and
54 functions [7].

55 Microbial functions can be characterized from genomic, proteomic or metabolic data
56 (Fig. 1) [8–10]. Considering genomics, quantitative PCR (qPCR) and microarrays were
57 the first technologies used to describe functional genes or taxa from complex
58 environmental samples [11]. Initially designed to determine the absolute copy number
59 of a single given gene, the latest technical advances can analyze thousands of
60 combinations of samples and targets in parallel [12]. Standardized methods even make
61 it possible to quantify genes of interest (*e.g.*, involved in biogeochemical cycles,
62 pesticide degradation, etc.) to estimate soil quality [13]. DNA microarrays were the first
63 high-throughput technologies giving access to gene expression profiles at the
64 individual or community levels [11,14]. There exist different kinds of microarrays (*e.g.*,
65 PhyloChip, GeoChip; PathoChip; StressChip; CAZyChip). They provide a snapshot of
66 microbial diversity (bacteria, fungi, viruses) and / or of the functional genes present in
67 a given sample (*e.g.*, genes coding for enzymes involved in polysaccharide
68 degradation) [15–18]. Some of these microarrays have become diagnostic tools in
69 many fields, in particular for targeting viruses, bacterial or fungal pathogens or harmful
70 organisms [19]. More recent and cheaper, various high-throughput sequencing (HTS)
71 alternatives have been developed to explore microbial communities (Fig. 1) [20].
72 Genome and metagenome sequencing have changed the microbial ecology field:
73 thanks to genome sequencing and meta-omics approaches, gene catalogs can be
74 assessed, and new microorganisms can be discovered [21,22].

75 For example, by implementing a metabarcoding approach, microbial ecologists
76 were first very enthusiastic about such huge taxonomic information, but quickly pointed
77 out the lack of associated functional information [22]. Taxonomic profiles can indeed
78 change to varying degrees among samples, and predicting to what extent these
79 changes impact the overall functional capacity of the community has remained a
80 technical and scientific challenge to date [6,23,24]. Metabarcoding may well be used
81 to directly target functional genes and classify them by taxonomic group, but
82 applications remain limited to a few families [25–29]. In the face of these limitations,
83 two solutions have emerged to indirectly obtain functional information from taxonomic
84 profiles, *i.e.* (i) functional inference, and (ii) ecological trait assignment, using
85 (meta)genome and microbiome big data (Fig. 1). Functional inference predicts the
86 putative functions (*e.g.*, gene catalogs, metabolic pathways) of microbial communities,
87 while ecological trait assignment directly retrieves a trait common to all taxa by linking
88 taxonomic names with a dedicated database. The major difference between these two
89 solutions for obtaining functional information is that functional inference retrieves
90 functions even for OTUs without a taxonomic name thanks to phylogenetic placement
91 of sequences (taxonomic markers) in a reference tree and different evolutionary
92 models.

93 Many bioinformatic tools have been developed since the first publication about
94 a functional prediction tool using metabarcoding data. To date, only one review has
95 addressed functional inference tools; it is focused on aquaculture and on a limited
96 subset of all the tools available to predict functions from 16S rDNA metabarcoding
97 datasets [30]. Therefore, in the present context where new solutions are proposed
98 regularly to predict putative function profiles, the state of the art needs to be scrutinized
99 more exhaustively to build a scientific and technical benchmark. More precisely, we

100 provide a detailed description of each tool and evaluate their advantages, specificities
101 and drawbacks by paying special attention to their methods, modularity, portability, and
102 robustness. One of the main objectives of this review is to provide a rationale on the
103 use of the different tools currently available for prokaryote and fungal communities and
104 draw perspectives, with a few suggestions to enhance their usefulness in microbial
105 ecology. Finally, we illustrate the application of these methods with studies focusing
106 on the soil environment. The choice of this particular system is justified by the fact that
107 it is the most diverse and complex one in terms of microbial diversity, ecology and
108 functional reservoir [4,31]; therefore, it represents the most challenging environmental
109 matrix for linking diversity and functions. We believe that this work will help scientists
110 working on microbial communities make choices to best take advantage of their high
111 amount of microbial data. This work also shows that although those approaches are
112 promising, they still need improvements to make them operational tools for microbial
113 soil quality diagnosis. A repository using standardized and robust metrics is still lacking
114 when it comes to interpreting the results.

115

116 **2. Historical and recent increase of microbial datasets**

117 The emergence of HTS in the mid 2000's generated a huge volume of data, leading
118 to a revolution in our way of describing biodiversity. This rise of microbial data can be
119 directly linked to the improvement of high-throughput sequencing technologies,
120 concomitantly with a tremendous drop of sequencing costs (Fig. 2). This was reflected,
121 with a small time lag, by an increase in the number of sequence read archives (SRAs)
122 linked to metabarcoding data deposited on the NCBI website (Fig. 2).
123 Thanks to the contribution of ecologists, microbiologists, taxonomists and computer
124 scientists, the databases are continuously enriched and are key to enhance our

125 knowledge about the description and determinism of environmental and human
126 microbiotas [32,33]. For example, the 16S rDNA sequences data available to analyze
127 bacterial/archaeal diversity was multiplied by 4 and 10 in the RDP and SILVA
128 databases, respectively, between 2007 and 2019 (Fig. 3A). The trend is the same for
129 fungal diversity, with a doubling of ITS sequences in the UNITE/INSD database within
130 the last five years (Fig. 3B). 16S rDNA sequences are much more numerous than ITS
131 sequences. However, there were 30 times more fungal species referenced than
132 bacterial ones in 2017 (Fig. 3A, 3B). The numbers of microbial genomes available, in
133 particular in the JGI platform, have increased continuously, and they outpaced Moore's
134 Law mostly from 2013 for bacteria and archaea (Fig. 3C, 3D).

135 The number of known microbial genes, enzymes or metabolic pathways available in
136 specialized databases has also considerably increased in the last few years [34–36].
137 Thousands of functional information files are currently accessible in the KEGG, CAZy
138 or MetaCyc databases (Table 1). A recent survey predicted the total global estimated
139 bacterial and fungal functions based on KEGG Orthology to reach 35.5 and 3.2 million,
140 respectively [37]. The authors also indicated that only a tiny fraction of these functions
141 is known today, representing 0.02% and 0.14% for bacteria and fungi, respectively.

142 Although the characterization of gene catalogs using metagenomic approaches was
143 recently criticized [38], the number of non-redundant genes provides an overview of
144 the potential functional reservoir available across various ecosystems [39]. The soil by
145 far appears to harbor the largest pool of functions, followed by the ocean, and then
146 animal microbiomes (Fig. 4).

147 The rapid growth of available genomes is a unique opportunity to predict the putative
148 microbial functions from metabarcoding data by linking taxonomic markers (*i.e.*, rDNA
149 amplicons) and their reference genomes or ecological traits. Therefore, the next

150 section is devoted to the different tools and databases dedicated to functional inference
151 and ecological trait assignment for bacterial and fungal communities.

152

153 3. **Overview of the available tools for predicting the potential functions of the** 154 **microbiotas**

155 HTS and the presently increasing collection of functional or ecological traits on a
156 more regular and rigorous basis are promising cues for linking biodiversity and
157 associated functions in the near future [24,40]. In the literature, the term "function" is
158 used in different ways depending on the study model, the time scale, or even the
159 habitat [41–44]. The notion of function may refer to genes, enzymes, or metabolic
160 pathways, but may also represent ecological traits that bring together phenotypic and
161 biochemical notions [45–47].

162 Based on the analysis of twenty papers since 2013, we classified the databases and
163 tools according to the granularity of the results (Fig. 5A), from general information such
164 as ecological traits to more detailed information such as genes or metabolic pathways
165 (Fig. 5). The tools used to obtain fine results, *i.e.*, at the metabolic pathway or gene
166 levels for any taxonomic resolution, are known as functional inference tools (Fig. 5B).
167 On the other hand, we grouped existing tools or databases under the term “ecological
168 trait assignment” when functional information referred to phenotypic or ecological traits
169 and was accessible only for a specific taxonomic rank (Fig. 5C). Indeed, there is a
170 wealth of information often linked to ecological traits in published scientific articles, or
171 of partially formatted metadata (*i.e.*, partial taxonomy or data not linked to the ID of a
172 taxonomic database) [48].

173 Tools or methods exist, known under the term “text mining”, to automatically collect
174 data from various sources (*e.g.*, a website, a document in pdf format) through

175 automatic language processing (e.g., natural language processing (NLP)) [49]. For
176 example, @Minter [50] retrieves information related to microbial interactions from
177 abstracts of papers thanks to a supervised machine learning model. Other tools are
178 based on ontologies, *i.e.*, they use a structured set of terms and concepts from a
179 particular domain by specifying the relationships between these terms and their
180 properties, and thus have a common reference for the use of a common vocabulary.
181 For example, OntoBiotope [51] ontology in the food field retrieves the phenotypes and
182 habitats of microbes from the literature based on the NCBI taxonomy. Another ontology
183 exists, called Ontology of Microbial Phenotype [52]; it brings together a structured set
184 of terms and concepts around microbial phenotypes, and specifies the relationships
185 between these terms and their properties. Tools also based on machine learning such
186 as ProTraits [53] can automatically annotate prokaryotic species based on phenotypic
187 or genomic data from scientific articles or online resources (<http://protraits.irb.hr>).
188 To date, we have recorded about twenty tools or databases that retrieve functional or
189 ecological data from microbial taxonomic markers, with two to four developments *per*
190 year (Fig. 6 and Table 2). The timeline shows that most of these tools (18/23 in total)
191 are only dedicated to bacteria/archaea, two are dedicated to bacteria/archaea + fungi,
192 and only three are specifically dedicated to fungal organisms. It is important to also
193 underline that most of these tools are devoted to functional inference (13/23). The most
194 cited tool is PICRUSt v1 [54], which remains on top of all others with more than 4,000
195 citations in 2020. While FUNGuild [55], Tax4Fun v1 [56] or FAPROTAX [57] are
196 reasonably cited with a few hundred citations, the others are very less so with only a
197 dozen citations (Fig. 7A). Interestingly, the articles citing functional inference and
198 ecological trait assignment tools fall within the same scope as the scopes for which
199 they were initially developed (Fig 7B.): PICRUSt, FUNGuild and PAPERICA are mainly

200 cited in papers about human health, the soil and the marine environments,
201 respectively.

202

203 **3.1. Functional Inference**

204 3.1.1. Definition

205 Functional inference consists in predicting the functional potential of a microbial
206 community from metabarcoding data. The functional potential of a taxon or of a
207 microbial community represents the metabolic capacities based on the presence /
208 absence of genes involved in these pathways. Functional inference methods are based
209 on the assumption that phylogenetic information from marker gene sequences
210 correlates well enough with the genomic content to produce accurate predictions when
211 associated reference genomes are available. In other words, it assumes a significant
212 relationship between (i) the phylogenetic distance between taxonomic markers and (ii)
213 the conservation of the genetic content, referring to vertical gene descent during the
214 evolution of microbial genomes. This is made possible through the relationship
215 between the phylogenetic relatedness of organisms and their gene content [58,59]
216 (Fig. 5B).

217 It should be emphasized that the presence of one or more genes involved in a function
218 remains “potential” and may not be expressed under environmental conditions. From
219 this point of view, functional inference results may be similar to shotgun metagenomics
220 data, which is often observed in the literature, especially when focusing on a family of
221 genes or a specific biogeochemical cycle [60].

222

223 3.1.2. Available tools

224 *Picrust*

225 Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
226 (PICRUSt) v1 [54] is the first tool to have been developed to predict potential functional
227 genes from 16S rRNA metabarcoding and has been the most popular one since it was
228 launched in 2013 (Fig. 5B). PICRUSt v1 needs three things: (i) a reference OTU, (ii) a
229 reference genome, and (iii) a reference phylogenetic tree. As regards the reference
230 OTU, the file (in BIOM or tabulated format) is expected to contain a standard OTU
231 abundance table with sequences picked only against the Greengenes taxonomic
232 reference (18 May 2012 or v13.5/v13.8). This tool based on a modified method of
233 ancestral state reconstruction (ASR) deduces functional information for taxa without a
234 match in the reference genomes. The reference genomes are functional proxies that
235 provide a weighting of the functional profiles for the phylogenetically close taxa within
236 a reference phylogenetic tree. The PICRUSt method is divided into three main steps
237 that are necessary to obtain relevant information on functional profiles: (i) genome
238 prediction, (ii) metagenome prediction, and (iii) analysis of predictions.

239 The genome prediction step consists in preparing the trees and checking the quality of
240 the input datasets; then comes the reconstruction of ancestral states in the reference
241 tree (ASR, 4 methodologies are available). Using the output files, the software program
242 predicts traits for leaves of the phylogenetic tree lacking sequenced genomes.

243 During the metagenome prediction step, normalization of the abundance of each OTU
244 is carried out based on rRNA gene copy numbers to predict the functional category
245 abundances of the metagenome. The user obtains an abundance table for each
246 functional category *per* sample. The correcting step of the rRNA gene copy numbers
247 (GCNs) allows normalizing to correct the biases towards microorganisms with greater
248 GCNs and improve the estimation of microbial diversity [61]. This step is recommended
249 when the OTUs are phylogenetically closely linked to the genomes [62]. To assess the

250 robustness of the predictions, *i.e.*, to obtain the representativeness of the database
251 towards a community of interest, a nearest sequenced taxon index (NSTI) is generated
252 for each sample. It is calculated using the average of the branches that separate the
253 sequences of interest (OTUs, ASVs) in a sample from the reference microbial genome,
254 with a weighting by their relative abundance in the sample. This confidence score is
255 one of the major strengths of this tool. Regarding functional categories, information can
256 be obtained at different levels (genes or metabolic pathways) with more or less detailed
257 descriptions (EC numbers, KEGG pathway [35], COG). Information about all functional
258 categories can also be obtained for each OTU. The last step consists in analyzing the
259 predicted data. This step is essential for interpreting the large number of results
260 generated from a robust statistical analysis.

261 The major strength of PICRUS_t v1 lies in its evolutionary models that infer functions
262 for the complete bacterial community. The portability of this tool with the support of a
263 broad stakeholder community including a forum (google group), blogs, are advantages
264 that make it a central tool for functional predictions (Table 2). Despite all its benefits,
265 PICRUS_t v1 has drawbacks such as focusing only on the 16S rDNA marker and using
266 only Greengenes taxonomy (Table 2). Several specialized tools have emerged to
267 integrate PICRUS_t as a sub-layer in order to carry out diagnoses in the medical field
268 [63] or directly in a pipeline [64].

269

270 *PAPRICA*

271 Pathway Prediction by Phylogenetic Placement (PAPRICA) [65] infers the metabolic
272 potential of prokaryotic and eukaryotic communities from metabarcoding data based
273 on rRNA gene amplicons. It was the first tool that allowed for the functional prediction
274 of 16S and 18S rRNA amplicons. It comes in the form of a pipeline taking the OTU

275 reads as inputs to place them in an rRNA reference tree built from complete genomes.
276 To build this tree, a consensus genome is found for each node in the tree, which then
277 makes it possible to predict metabolic pathways for the sequences of interest without
278 a match in the complete reference genomes. The abundance of metabolic pathways
279 is weighted by rRNA gene copy numbers from known genomes. A strength of this tool
280 is that it also provides an indicator of genomic stability depicting the robustness of the
281 results. However, PAPRICA, like all the tools using a reference phylogenetic tree and
282 sequence placement methods, is dependent on the quality of rRNA resolution, and this
283 represents a drawback when some clades may be affected (Table 2).

284

285 *Tax4Fun*

286 Tax4Fun [56] is an R [66] package published in 2015 for predicting functional profiles
287 from targeted metagenomic 16S rRNA data. However, the algorithm and statistical
288 efficiency based on a metabolic mixture model in terms of a mixture of pathways (MoP)
289 was developed in 2013. This R-based architecture is inherently a cross-platform tool,
290 and it may be more accessible for a large number of users with low experience in
291 bioinformatics. This tool uses pre-calculated functional profiles like PICRUST v1 and
292 taxonomic data formatted from the SILVA database. One of the differences with
293 PICRUST the rRNA sequence placement in the reference genomes, which is achieved
294 by a BLAST search (instead of a tree placement approach for PICRUST). It is a very
295 convenient tool because it provides a confidence score (FTU and FSU) to determine
296 the fraction of OTUs that was not mapped to KEGG organisms or the number of
297 sequences without KEGG Orthology (KO) hits (Table 2). Like PICRUST v1, it cannot be
298 used for fungal diversity predictions.

299

300 *Piphillin*

301 Piphillin [67] differs from the PICRUSt or PAPRICA approaches because it does not
302 use a phylogenetic tree or database (16S) but directly maps the OTU sequences on
303 the rRNA of the reference genomes using a nearest-neighbor algorithm. This
304 specificity could avoid faulty sequence placements in the reference phylogenetic tree.
305 It is used online only, which represents both a strength and a weakness: it benefits
306 from computing power (a strength), whose strength depends on the hosting server
307 (e.g. quota management, cluster configuration) (a weakness). A Piphillin sub-layer
308 also exists to complete the analysis of the results [68].

309

310 The quality of prediction represents a prerequisite for the application of the above-
311 presented tools to study indigenous microbial communities. It may depend on the tool,
312 but also on the type of targeted ecosystem. To test the quality of functional prediction
313 according to the tool and the studied ecosystem, we compiled the NSTI scores for
314 PICRUSt v1 and the FTUs for Tax4Fun from a subsampling of articles that covered a
315 range of ecosystems – human, marine, plant, and soil (Fig. 8). Whatever the tool, the
316 best predictions were obtained for the human microbiotas, and the most approximate
317 ones for the soil samples. The variability of quality scores across the different soil
318 studies seemed to be lower with PICRUSt than with Tax4Fun. Nevertheless, some soil
319 studies using Tax4fun indicate a good-quality survey with only about 30% of OTUs
320 unmapped to a reference. This likely reflects the discrepancy between human
321 reference genome availability and soil microbiota genome availability. In addition,
322 microbial diversity is much more complex in soils than in the human microbiotas. In
323 this case, it is essential that the quality scores from functional inference tools should

324 be taken into account because it is a key to a robust interpretation of the results.
325 Unfortunately, we found few studies indicating these quality scores.

326

327 **3.2. Ecological trait assignment**

328

329 3.2.1. Definition

330 Ecological trait assignment differs from functional inference since it consists in
331 obtaining information on the life strategy, phenotypic and quantitative genomic traits
332 (e.g., trophic modes, growth strategy) of a taxon from its nomenclature, whatever its
333 taxonomic rank. If the taxon is not present in the database, it will not be possible to
334 know its traits (Fig. 5C). This approach is faster than functional inference for retrieving
335 an item of functional information, but tools dedicated to metabarcoding outputs are
336 lacking, and only a few ecological traits are available (Table 2). The main interest is to
337 get functional information with a possibly not so fine granularity as functional inference
338 does, but obviously more accurate. Ecological traits are indeed often based on results
339 with biochemical experimentations from curated databases or scientific publications.
340 Practically speaking, only the guild will be recovered and for example the fungal
341 sequences identified as belonging to the *Serpula* genus will be assigned to a wood
342 saprotroph when an ecological trait tool is used; with an inference tool, the abundance
343 of various genes related to polysaccharide degradation will be attributed to all fungal
344 sequences.

345

346 3.2.2. Tools

347 *FUNGuild*

348 FUNGuild [55] is the pioneer and one of the few tools that assigns ecological traits to
349 fungi based on their taxonomy (Table 2). These assignments rely on metabarcoding
350 data. They require providing a contingency table (OTUs or sequence counts *per*
351 sample) and the link between each OTU and its taxonomy. To carry out the
352 assignment, FUNGuild uses its own curated database, and searches it for the taxon.
353 This database contains several taxonomic levels (e.g., phylum, genus, species).
354 However, the taxonomic name at the genus or species level is necessary to assign
355 traits to the taxa of interest. Trait information is available in 66% of the cases at the
356 genus level, and only in 34% of the cases at the species level [55]. The user obtains a
357 summary table of the different possible ecological traits for each taxon with a
358 robustness indicator and a confidence range (“possible”, “probable”, and “highly
359 probable”).

360 The strength of this database is that the provided data are based on the literature
361 (primary research), or on reference websites or their own collective research
362 experience if the datum is missing. The authors recommend the use of the UNITE
363 database for taxonomic assignment and therefore the use of the internal transcribed
364 spacer (ITS) marker, but it can be easily transposed to data based on the 18S rRNA
365 marker. It just requires creating a wrapper to make a link between the taxonomy of the
366 data and FUNGuild to retrieve the traits of interest.

367 A new database called Fun^{Fun} [69] is now available. It encompasses 80 fungal
368 ecological traits. In reality, this database is a FUNGuild database overlay with
369 information on genetic, enzymatic, morphological, stoichiometric, life history, and
370 physiological aspects. In addition, the authors mention that Fun^{Fun} will be updated in
371 terms of taxonomy and associated guilds, which is not necessarily the case with
372 FUNGuild. However, although this database is promising, a lot of information is missing

373 because it integrates literature data for the first time ever, and its improvement relies
374 on the progress of research as well as the contribution of scientists. This caused an
375 impulse leading to a community of scientists proposing a new database: FungalTraits
376 [70] links information from FUNGuild and Fun^{Fun}. It is very complete, and offers
377 different levels of life styles. Please note that this database includes species from the
378 fungal kingdom but also fung-like stramenopiles (e.g., the Oomycota phylum). This
379 may be especially useful because various species are identified as major plant
380 pathogens within Oomycota. For example, the genus *Phytophthora* gathers several
381 crop pathogens causing important losses and can represent a risk to global food
382 security [71].

383 To conclude, the minor drawbacks of FUNGuild, with rare updates or a tool oriented to
384 ITS sequences, have been offset by the new Fun^{Fun} and FungalTraits databases.

385 To complete the tools concerning fungal communities, DEEMY [72] is an information
386 system only available online and specialized in ectomycorrhizas
387 (<http://www.deemy.de>). This website references 554 species associated with their
388 respective symbiotic organisms, including 104 genera. To characterize each species,
389 a summary sheet provides taxonomic nomenclature, bibliographical references and
390 photographs, as well as information on morphology, anatomy, potential chemical
391 reactions, or even ecology traits.

392

393 *FAPROTAX*

394 Functional Annotation of Prokaryotic Taxa (FAPROTAX) [57] is used to assign
395 metabolic functions, ecological traits or large functional groups relevant to prokaryotes
396 (Table 2). This database was built manually from the scientific literature of the
397 *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) and

398 Bergey's *Manual of Systematic Bacteriology*. It contains about 4,700 unique
399 prokaryotic taxonomies (mostly at the species level) and 90 functional groups.
400 FAPROTAX is based on the implicit assignment of a trait / function to a taxon (whether
401 cultivated or not) if all the cultivated members display this trait / function. Its main
402 limitation is that it is focused on marine prokaryotic organisms, so that communities
403 from other biomes can be missing. Another point to be considered is that if the taxa of
404 interest do not have a species name, the tool cannot draw inferences at the upper
405 levels (e.g., genus) to assign an ecological trait.

406

407 *IJSEM phenotypic database*

408 IJSEM [73] compiles phenotypic and environmental tolerance data about more than
409 5,000 bacterial strains. It is an official and unique reference for publishing and
410 validating new strains. These strains cover about 23 phyla from various habitats
411 (mainly soils). The database appears as a TSV file
412 ([https://figshare.com/articles/International_Journal_of_Systematic_and_Evolutionary_](https://figshare.com/articles/International_Journal_of_Systematic_and_Evolutionary_Microbiology_IJSEM_phenotypic_database/4272392)
413 [Microbiology_IJSEM_phenotypic_database/4272392](https://figshare.com/articles/International_Journal_of_Systematic_and_Evolutionary_Microbiology_IJSEM_phenotypic_database/4272392)), and available information can
414 be grouped into five categories: ancillary data (e.g., article's digital object identifier;
415 taxonomic nomenclature), morphology/phenotype (e.g., Gram stain status; motility),
416 metabolism (e.g., BIOLOG information), environmental preferences (e.g., habitat of
417 isolation; oxygen requirement), and sequence data (e.g., 16S rRNA accession no.).

418

419 *BacDive*

420 BacDive [74] is one of the largest metadatabases (<https://bacdive.dsmz.de>)
421 referencing information on bacterial and archaeal diversity (Table 2). No tool links
422 taxonomy and phenotypic information directly because the database can only be

423 browsed on a website or data can be downloaded from it. However, it provides a
424 complete application programming interface (API) to achieve scripts and retrieve the
425 desired information. In the first months of 2020, it offered data on 81,827 bacterial and
426 archaeal strains, including 14,091 type strains, and thereby covered approximately
427 90% of the described species according to their website. This database is very
428 interesting because it provides different levels of robust information on taxonomy,
429 morphology, physiology (API®-tests), molecular data, and cultivation conditions. As for
430 physiological data, it provides – for example – the main substrates used for culturing a
431 species and the enzymes present (a link with the EC classification number is available).
432 These data have been more broadly incorporated into a tool (bacteria-archaea-traits)
433 that encompasses numerous traits of bacteria and archaea from 26 sources [46].

434

435 To complete this list, a few specialized databases target only one or a few traits. For
436 example, Engqvist [75] recently grouped the growth temperatures of 21,498 non-
437 redundant organisms across the whole tree of life. This study showed a strong
438 correlation between the growth temperature of organisms and enzymatic optima, with
439 temperature-dependent increases or decreases of enzymatic functions. This
440 information can be very interesting and complementary to the interpretation of
441 functional inference results, and can be linked – for example – to environmental
442 conditions.

443

444 4. Application of these new approaches to the functions of the soil microbial
445 ecosystem

446 4.1. Functional Inference

447 In recent years, meta-omics approaches have been increasingly included in soil
448 monitoring, whether in fundamental research programs or in more operational projects
449 [76]. Most studies (about 60% based on keywords in the titles or abstracts of the
450 publications, see Fig. 7B) have focused on PICRUSt to generate functional predictions
451 from taxonomic data of the soil microbiota. We summarized the most valuable
452 outcomes about soils by grouping them into categories: anthropogenic gradient,
453 agricultural practices, and biogeochemical cycle or soil properties (Fig. 9). For
454 example, a study showed that plant-bacteria interactions in the rhizosphere were
455 mainly related to beneficial cooperation [77] involving the release of root exudates by
456 the plants on the one hand, and hormone production or the ability to break down toxic
457 chemicals by bacteria on the other hand. Another study investigated the stoichiometric
458 regulation of soil carbon cycling by comparing functional predictions by metabarcoding
459 (via PICRUSt) and shotgun sequencing on a wide C:N:P soil gradient in a rice field
460 [60]. A strong correlation was evidenced between the functional predictions from
461 metabarcoding and metagenomics as regards the abundance of some metabolic
462 families involved in the C, N and P cycles. Still using PICRUSt, another study examined
463 the effects of intercropping by predicting the soil microbial functional profiles. It
464 evidenced that an intercropping system increased the functional potential in terms of
465 carbon fixation pathways and the citrate cycle [78]. Finally, a study focused on the
466 impact of long-term land-use practices (forest, grassland, crops) on soil bacterial
467 communities [79] showed that forest soils harbored the largest reservoir of genes,
468 followed by no-till soils and then grasslands. The plowed soils presented the lowest
469 functional richness.

470 Based on Tax4Fun predictions, a study investigated the impact of different irrigation
471 practices with various water qualities (freshwater, treated or untreated wastewater)

472 along with the different land use systems in drylands [80]. The authors compared the
473 potential functional and taxonomic profiles of bacteria. Irrigation with wastewater had
474 an effect on bacterial responses by shaping communities and functional profiles. By
475 bringing more nitrogen, wastewater favored the response of certain genera, in
476 particular *Nitrosospira*, and increased the relative abundance of the genes involved in
477 nitrification and denitrification.

478 Among all the functional inference tools available today, two of them stand out, *i.e.*,
479 PICRUSt and Tax4Fun. A benchmark study of these tools found no major differences
480 in terms of performance, especially for soil samples [81]. Another benchmark study
481 indicated that these two tools provided similar functional profiles but could be
482 complementary for certain gene families found only in one or the other [82]. Moreover,
483 the characterization of the fungal functional potential by PICRUSt2 is too recent for us
484 to have any insights into its robustness concerning soil communities. Compared to trait
485 assignment, the links between diversity and functions still remain tenuous concerning
486 certain biogeochemical cycles or the impact of climate change and plant diversity (Fig.
487 9).

488

489 4.2. Ecological trait assignment

490 The complexity of microbial traits is variable, with simple traits like organic
491 phosphate utilization, and more complex ones like methanogenesis [24,83]. The
492 conservation of prokaryotic traits or core genes varies according to phylogenetic depth
493 [58]. For example, the complex methanogenesis trait appears to be very conserved at
494 the order and family levels, while contrastingly with the resistance to specific
495 bacteriophages appears to vary at the species level due to particular point mutations

496 [24]. Below are a few examples of the possible benefits of ecological traits to the
497 analysis of the diversity of soil microbial communities (Fig. 9).

498 Regarding the assignment of fungal traits, FUNGuild is currently and by far the most
499 implemented tool, if not the only tool implemented by ecologists wishing to supplement
500 their diversity analyses with data on the ecological traits of fungal communities, and
501 mainly in studies on soil fungal communities [84–87]. A study on fungal communities
502 in subtropical forest soils highlighted a negative relationship between the abundance
503 of pathogenic fungi and the phylogenetic diversity of plant communities [88]. Another
504 study showed a positive correlation between soil fungal community dissimilarities
505 (plant pathogens, saprotrophs and ectomycorrhizas) and plant phylogenetic distances
506 in forest soils [89]. Tropical land uses also impact the functional guild. A massive shift
507 of fungal trophic modes has been showed – notably a decrease in mycorrhizal fungi
508 and an increase in saprophytic and pathogenic fungi –along with increased
509 anthropization levels [90]. Interestingly, several large-scale (national or global) studies
510 have characterized the distribution of trophic types while identifying the environmental
511 parameters that influence them [84,91–93]. The distribution of these trophic modes
512 seems to vary greatly depending on temperature and precipitation [93]. This supports
513 a recent global study focused on the distribution of pathogens and indicating higher
514 abundance in warm regions [92]. A recent study compared the trophic modes
515 (synonym: life strategies) assigned to the ITS and 18S rDNA molecular markers by
516 FUNGuild [84]. This study indicated that the saprotroph and pathotroph richness levels
517 were directly and negatively correlated with the organic matter content and elevation,
518 and positively correlated with the pH and bulk density. For symbiotroph richness, the
519 relationship differed depending on the molecular marker used: it was positively
520 correlated with the C:N ratio when ITS sequences were used, but negatively correlated

521 when 18S rDNA sequences were used. Similarly, the pH was positively correlated
522 based on 18S rDNA data, but negatively correlated based on ITS data [84]. These
523 differences may come from the fact that the two molecular markers do not cover the
524 same taxonomic range. Therefore, the choice of molecular markers and primers is
525 essential because it impacts the global picture obtained by possibly enhancing or
526 decreasing the representation of particular functional groups in the community. For
527 example, arbuscular mycorrhizal fungi are better represented, in particular the
528 *Glomeromycota* group, when the 18S rDNA marker is used [94,95]. A study at a
529 smaller scale also showed that saprotroph richness was directly driven by the soil
530 physico-chemical parameters and confirmed the results mentioned above. The authors
531 showed a positive correlation with the pH but a negative one with the C:N ratio [96]. All
532 these studies used the FUNGuild tool dedicated to characterizing fungal community
533 traits.

534 Regarding the assignment of bacterial traits, various databases exist but few tools have
535 been developed to assign ecological traits from metabarcoding datasets. Only
536 FAPROTAX stands out as a powerful tool for analyzing the functional potential of soil
537 communities [97], although it is dedicated to marine organisms.

538

539 5. Technical and conceptual limitations and biases

540 The metabarcoding approaches have significant advantages for characterizing
541 indigenous prokaryotic and eukaryotic microbial communities. Standard protocols now
542 exist, from sample preparation to bioinformatic and statistical analyses, and scientists
543 have acquired an important feedback on biases, costs, and efficiency [98–100].

544 A fundamental limitation of functional inference tools, represented by gene gain and
545 loss, is due to horizontal gene transfer [101], which is addressed in the literature and

546 taken into account to some extent in these tools. However, horizontal gene transfer
547 remains difficult to consider accurately for functional prediction, and its influence on
548 microbial communities is hard to estimate. Moreover, the horizontal gene transfer rate
549 varies substantially within the tree of life and according to gene families / pathways
550 [24,83,101]. This process is mainly described in prokaryotes, but is also found to a
551 lesser extent in eukaryotes, in particular fungi [102]. Microorganisms can gain a
552 function through plasmid transfer, but no information was found in the literature about
553 functional prediction [54]. However, plasmids are extrachromosomal DNA molecules
554 that play a role in the rapid adaptation of microbial communities to environmental
555 changes across all microbiomes [103,104]. In particular, they are transferred between
556 phylogenetically distant populations for them to acquire genes and beneficial traits for
557 their adaptation (*e.g.*, resistance to antibiotics, biocides, pollutants). This is key for all
558 environments, especially soils where biotic and abiotic fluctuations are tremendous
559 [105].

560 From a technical point of view, most of the studies on microbial diversity using
561 metabarcoding approaches are based on the sequencing of one or more hypervariable
562 regions and remain limited by the size of the amplicon to be sequenced. The most
563 commonly used Illumina sequencing platforms (MiSeq, HiSeq and NovaSeq) can
564 provide maximum readings of 600 bp (~550 bp after adapter/tag/primer trimming).
565 Several studies have questioned the most suitable regions for obtaining the best
566 taxonomic resolution [106,107]; the use of full-length rRNA (~1,800 bp) seems to be
567 the most appropriate solution [108]. It would significantly enhance phylogenetic
568 resolution for prokaryotic and eukaryotic microorganisms [109] (Fig. 10, second box).
569 Short reads do not allow good enough resolution in taxonomic assignment either (*i.e.*,
570 not down to the species level) although this point is crucial for placing sequences/taxa

571 in the phylogenetic tree to achieve functional inference. With third-generation HTS
572 platforms (*e.g.*, PacBio, Oxford Nanopore), full-length molecular markers can be
573 sequenced, *e.g.*, 16S/18S rRNA genes or the full ITS1 and ITS2 sequences [110,111].
574 This will considerably improve taxonomic assignment, and make it possible to assign
575 sequences at the species or even the strain level in certain cases [111]. This way,
576 functional inference and ecological trait assignment will be improved. However, if the
577 objective is to obtain the best taxonomic resolution possible, the study of ecological
578 traits at high taxonomic ranks (*e.g.*, the phylum) remains very promising, especially for
579 highly conserved traits [112]. For example, the carbon mineralization rate was
580 positively (*e.g.*, Bacteroidetes) or negatively (*e.g.*, Acidobacteria) correlated with their
581 relative abundance [113].

582 A good practice complementary to the use of full-length amplicon sequencing would
583 be the use of amplicon sequence variants (ASVs, also called ZOTUs) to increase the
584 rate of inference with a better sequence placement on the reference tree [114,115].
585 Indeed, for those using an OTU clustering approach with a similarity threshold, one
586 solution would be to use all the sequences within the OTUs instead of one
587 representative sequence for each OTU seed, which could be less accurate. However,
588 this would also increase the analysis time.

589

590 6. Importance of taxonomy and genome references: from accuracy to resolution

591 Many tools use taxonomic data to obtain information about microbial functions
592 through a metabarcoding approach. Therefore, it is very important to check the
593 bioinformatic strategy used to analyze the amplicon sequences, from the filtering steps
594 to OTU clustering or not (see ASV), including taxonomic assignment.

595 The use of tools on ecological traits is highly dependent on taxonomic resolution. For
596 example, when using FUNGuild, special attention must also be paid to the fact that a
597 sequence assigned at the genus level may be associated with several trophic types,
598 and that plant-pathogenic fungi are highly host-specific and may be non-pathogenic in
599 the context of the study. For the sequences (or OTUs) without any taxonomic
600 assignment, functions cannot be obtained using tools on ecological traits (Fig. 10,
601 second box). In order to improve this point, especially for fungal communities,
602 inferences may be drawn based on phylogeny, as done for bacteria, archaea or
603 macroorganisms [116–120]. One of the avenues to be explored is the use of ASR tools
604 such as PICANTE [121] or CASTOR [122], which infer traits for taxa devoid of
605 ecological data from a phylogenetic tree.

606 Functional inference tools depend on the reference genomes to establish predictions,
607 so that the accuracy of the results can vary among samples. Samples with well
608 described host-associated communities such as the human microbiome have many
609 reference genomes available, and allow good predictive accuracy (Fig. 8, Fig. 10 third
610 box). Contrastingly, in more complex and highly biodiverse environments like soils
611 [123], the genomes representing the total taxonomic diversity are much more difficult
612 to obtain. The proportion of cultivable terrestrial strains remains very low
613 (approximately 25%) compared to the human microbiotas (80%) [124]. Thus, the
614 results estimated for the communities from complex biomes are approximate and
615 debatable.

616 In order to improve functional prediction results, it is advisable to provide genomes
617 specific to the habitat of interest [125]. Considerable efforts have to be made to
618 increase the number of habitat-specific reference genomes (animal / human, water,
619 plant, soil), with special attention to the most complex and unknown environments

620 [126]. Tools to routinely update the databases will also need to be developed [127].
621 This is an ongoing dynamic at the international scale. For example, the annotation of
622 reference genomes in databases is not yet representative of soil microbial diversity
623 [128]. To fill this gap, an effort has been made by creating the Refsoil database (which
624 does not seem to be maintained (https://github.com/germs-lab/ref_soil)) [128] or a
625 Refsoil + plasmid database [104].

626

627 7. Discussion and future prospects

628 The possible retrieval of a putative functional potential or ecological traits directly
629 from taxonomic markers and metabarcoding approaches opens new perspectives for
630 our understanding of microbial communities, both from a fundamental and/or
631 operational point of view (*e.g.*, functional redundancies, diagnostic tool) [63,129]. This
632 information can be used to (i) understand the main functions potentially expressed in
633 a given environment and identify the possible drivers, (ii) examine the distribution of
634 functions among taxonomic group, or (iii) supplement the classical diversity metrics
635 used to evaluate the ecological state of environmental matrices (Fig. 10, first box).
636 Beyond providing an overview of the putative functions of an ecosystem, prediction
637 tools could also provide more detailed information than taxonomic markers do for users
638 to significantly distinguish sample groups from each other in certain habitats [108] (Fig.
639 10A, first box).

640 A new generation of tools solves the main limitations of the previous generation tools
641 by including improvements in terms of taxonomic marker targeting, methodology and
642 flexibility.

643

644 Future prospects with second-generation tools

645 Second-generation tools are currently emerging, *e.g.* PICRUSt2 [115],
646 Tax4Fun2 [125] or iVikodak [130] (Fig. 6). Indeed, Langille's team of developers
647 bridged the gap for the scientific community working on fungal ecology. PICRUSt2 now
648 includes 18S rDNA and ITS amplicons from the fungal kingdom. Another great
649 improvement is flexibility: the sequence can be used directly, instead of taxonomy
650 based on Greengenes nomenclature. Users are no longer dependent on taxonomy to
651 infer functions; this is a great comfort, and provides better robustness of the analyses.
652 However, users should be wary of the results because the number of sequenced fungal
653 genomes currently integrated in the tool is much lower than the number of bacterial
654 genomes. It is recommended to check the quality score (*e.g.*, NSTI) for the robustness
655 of the results and interpretation. However, this limitation can be lifted. For example, the
656 1000 Fungal Genomes Project [131] is aimed at high-quality sequencing and
657 annotation of fungal genomes so as to build a reference dataset to be used for meta-
658 omics data analysis.

659 Another downside of these tools is the absence of data support for micro-eukaryotic
660 communities, which are essential to the soil ecosystem. Protists are abundant and
661 diverse, with a large range of functional diversity, and are highly involved in soil food
662 webs and functioning [132,133]. It would be particularly useful to develop tools
663 dedicated to protists from data on ecological traits available in the literature [134].

664

665 **Challenges: from fundamental research to diagnosis**

666 Switching from fundamental research to practical applications would be really
667 interesting because although operational microbial diversity bioindicators are
668 increasingly emerging, there is a huge gap in the functional information of microbial
669 communities. Even if the number of species can be an indicator of the impact of biotic

670 and abiotic factors [135,136], the need to characterize the associated functions at the
671 ecosystem level has become obvious to obtain a complete diagnosis with functional
672 information on the soil microbial quality [137,138].

673 As regards human health, identifying taxonomic and functional changes to estimate
674 the contributions of taxa associated with a disease is an emerging topic [139], but
675 examples for the soil microbial quality are still scarce. The huge complexity and
676 diversity of the soil microbial community probably still limits such applications to the
677 soil ecosystem, along with a lack of genome references. However, initiatives at the
678 global level are in progress to access the soil biodiversity using taxonomic, functional
679 and environmental data [140]. We can also note that a real dynamic seems to be
680 developing at the international scale to collect, standardize and disseminate traits
681 through the tree of life *via* an open science tool called the Open Traits Network (OTN)
682 [83].

683 To our knowledge, providing robust and operational indicators based on putative
684 functions derived from metabarcoding data is impossible today. The main challenges
685 are to (i) aggregate and summarize the mass of data currently generated, (ii) test the
686 predictions on datasets and compare them with “real” functional measurements, (iii)
687 validate these indicators on datasets under diverse experimental conditions (*e.g.*, land
688 use gradient, agricultural practices) at the local and global scales, and (iv) develop
689 representative repositories to ensure the validity of the diagnosis made from these new
690 tools.

691 Regarding aggregation and data reduction [(i)], a track would be to use a constrained
692 non-negative matrix factorization approach [141], an alternative to the concept of
693 community-aggregated traits (CATs) [142]. This method has already been used to
694 aggregate functional traits from meta-genomes [141]. The authors demonstrated that

695 significant data reduction made it possible to propose simple models to describe a set
696 of complex functions at the scale of an ecosystem (here the potential for fiber
697 degradation in the human intestinal microbiota) while preserving biological data quality
698 [141]. Concerning [(ii)], it will be interesting, for example, to confront functional
699 predictions with volatile organic compound emissions or microbial respiration rates
700 from soil measurements. Moreover, to suggest these tools as robust indicators of the
701 soil quality [(iii)], it will be essential to use large datasets in order to determine the best
702 metrics (e.g., functional richness, relative gene abundance, aggregation of traits) and
703 the most sensitive genes or groups of genes depending on the various scientific issues.
704 Once these limitations have been lifted, these tools will provide results of great interest
705 to the scientific community at relatively affordable human, technological and financial
706 costs. However, maintaining the associated scientific expertise will be essential to
707 support their transfer for operational applications and avoid erroneous interpretations
708 that could potentially have disastrous consequences for soil users and soil policy
709 makers [(iv)]. For example, interpreting trophic types requires strong expertise, with
710 particular attention to the exploitation of potential pathogenicity information – a highly
711 sensible task. The responses of the traits vary according to the disturbances applied
712 to the ecosystem [143], and the results must be contextualized to ensure correct
713 interpretation.

714

715 **Conclusion**

716 The exploration of the microbial functional diversity based on taxonomic marker genes
717 in order to improve our knowledge of microbial diversity and functions is just starting.
718 As highlighted in this review, various solutions have emerged over a number of years
719 and are being improved quickly thanks to technological advances. Functional inference

720 results are already robust and representative for some ecosystems with low diversity
721 (specific richness) and with well characterized genomes such as the human
722 microbiotas. Progress now needs to be made for more complex environments. The
723 upcoming challenge, notably for environmental samples, will be to establish the link
724 between functional predictions on reference datasets and environmental
725 measurements. The new network SoilBON dedicated to monitoring soil biodiversity
726 and functional ecosystems at a global scale, with particular attention to microbial
727 diversity, is a step in this direction [3]. This ambitious framework aims to collect and
728 analyze soil diversity based on soil ecological indicators (*i.e.*, essential biodiversity
729 variables [144]). One purpose of this framework is to inform policy makers and
730 stakeholders for them to adapt measures and preserve this biodiversity.

731

732 **Abbreviations**

733 DNA: Deoxyribonucleic acid; qPCR: quantitative polymerase chain reaction; HTS:
734 high-throughput sequencing; rDNA ribosomal DNA; rRNA: ribosomal ribonucleic acid;
735 SRA: sequence read archive; NCBI: National Center for Biotechnology Information;
736 RDP: Ribosomal Database Project; ITS: internal transcribed spacer; INSD:
737 International Nucleotide Sequence Database; JGI: Joint Genome Institute; KEGG:
738 Kyoto Encyclopedia of Genes and Genomes; CAZy: carbohydrate-active enzymes;
739 NLP: natural language processing; OTU: operational taxonomic unit; GCN: gene copy
740 number; NSTI: nearest sequenced taxon index; FTU: fraction of OTUs; EC number:
741 enzyme commission number; COG: cluster of orthologous groups; KO: KEGG
742 orthology; IJSEM: *International Journal of Systematic and Evolutionary Microbiology*;
743 API: application programming interface; C, N and P cycles: carbon, nitrogen and

744 phosphorus cycles; bp: base pairs; ASV: amplicon sequence variant; ZOTU: zero-
745 radius OTU; OTN: open traits network; CAT: community-aggregated trait.

746

747 **Competing interests**

748 The authors declare that they have no competing interests.

749

750 **Funding**

751 This work was funded by the ADEME (French Environment and Energy Management
752 Agency).

753

754 **Author contributions**

755 C.D and L.R conceptualized the manuscript. C.D drafted the manuscript with
756 contributions from S.T, S.D, A.C, P-A.M and L.R. All authors read and approved the
757 final manuscript.

758

759 **Acknowledgements**

760 Thanks to Annie Buchwalter for correction and improvement of English language in the
761 manuscript.

762

763 **References**

764 1. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, et al.
765 Scientists' warning to humanity: microorganisms and climate change. Nat Rev
766 Microbiol [Internet]. 2019;17:569–86. Available from:
767 [http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L628](http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L628318774%0Ahttp://dx.doi.org/10.1038/s41579-019-0222-5)
768 [318774%0Ahttp://dx.doi.org/10.1038/s41579-019-0222-5](http://dx.doi.org/10.1038/s41579-019-0222-5)

769 2. Maron PA, Mougél C, Ranjard L. Soil microbial diversity: Methodological strategy,
770 spatial overview and functional interest. *Comptes Rendus - Biol* [Internet]. Académie
771 des sciences; 2011;334:403–11. Available from:
772 <http://dx.doi.org/10.1016/j.crv.2010.12.003>

773 3. Guerra CA, Bardgett RD, Caon L, Crowther TW, Delgado-Baquerizo M,
774 Montanarella L, et al. Tracking, targeting, and conserving soil biodiversity: A
775 monitoring and indicator system can inform policy. *Science* (80-) [Internet].
776 2021;371:239–41. Available from:
777 <https://www.sciencemag.org/lookup/doi/10.1126/science.abd7926>

778 4. Bardgett RD, Van Der Putten WH. Belowground biodiversity and ecosystem
779 functioning. *Nature* [Internet]. 2014;515:505–11. Available from:
780 <http://www.nature.com/articles/nature13855>

781 5. Rivett DW, Bell T. Abundance determines the functional role of bacterial
782 phylotypes in complex communities. *Nat Microbiol* [Internet]. 2018;3:767–72.
783 Available from: <http://www.nature.com/articles/s41564-018-0180-0>

784 6. Antwis RE, Griffiths SM, Harrison XA, Aranega-Bou P, Arce A, Bettridge AS, et al.
785 Fifty important research questions in microbial ecology. *FEMS Microbiol Ecol*
786 [Internet]. 2017;93. Available from: [https://academic.oup.com/femsec/article-](https://academic.oup.com/femsec/article-lookup/doi/10.1093/femsec/fix044)
787 [lookup/doi/10.1093/femsec/fix044](https://academic.oup.com/femsec/article-lookup/doi/10.1093/femsec/fix044)

788 7. Sergaki C, Lagunas B, Lidbury I, Gifford ML, Schäfer P. Challenges and
789 Approaches in Microbiome Research: From Fundamental to Applied. *Front Plant Sci*
790 [Internet]. 2018;9:1–12. Available from:
791 <https://www.frontiersin.org/article/10.3389/fpls.2018.01205/full>

792 8. Starr AE, Deeke SA, Li L, Zhang X, Daoud R, Ryan J, et al. Proteomic and
793 Metaproteomic Approaches to Understand Host-Microbe Interactions. *Anal Chem*

794 [Internet]. 2018;90:86–109. Available from:
795 <https://pubs.acs.org/doi/10.1021/acs.analchem.7b04340>

796 9. Aldridge BB, Rhee KY. Microbial metabolomics: Innovation, application, insight.
797 *Curr Opin Microbiol* [Internet]. 2014;19:90–6. Available from:
798 <https://linkinghub.elsevier.com/retrieve/pii/S1369527414000794>

799 10. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best
800 practices for analysing microbiomes. *Nat Rev Microbiol* [Internet]. 2018;16:410–22.
801 Available from: <http://www.nature.com/articles/s41579-018-0029-9>

802 11. Porter TM, Hajibabaei M. Scaling up: A guide to high-throughput genomic
803 approaches for biodiversity analysis. *Mol Ecol*. 2018;27:313–38.

804 12. Mehle N, Dreo T. Quantitative Analysis with Droplet Digital PCR. *Notes Greek*
805 *Text Genes* [Internet]. 2019. p. 171–86. Available from:
806 http://link.springer.com/10.1007/978-1-4939-8837-2_14

807 13. Thiele-Bruhn S, Schloter M, Wilke BM, Beaudette LA, Martin-Laurent F, Cheviron
808 N, et al. Identification of new microbial functional standards for soil quality
809 assessment. *Soil* [Internet]. 2020;6:17–34. Available from:
810 <https://soil.copernicus.org/articles/6/17/2020/>

811 14. Sessitsch A, Hackl E, Wenzl P, Kilian A, Kostic T, Stralis-Pavese N, et al.
812 Diagnostic microbial microarrays in soil ecology. *New Phytol*. 2006;171:719–36.

813 15. He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, et al. GeoChip: A
814 comprehensive microarray for investigating biogeochemical, ecological and
815 environmental processes. *ISME J*. 2007;1:67–77.

816 16. Lee YJ, Van Nostrand JD, Tu Q, Lu Z, Cheng L, Yuan T, et al. The PathoChip, a
817 functional gene array for assessing pathogenic properties of diverse microbial
818 communities. *ISME J* [Internet]. Nature Publishing Group; 2013;7:1974–84. Available

819 from: <http://dx.doi.org/10.1038/ismej.2013.88>

820 17. Zhou A, He Z, Qin Y, Lu Z, Deng Y, Tu Q, et al. StressChip as a high-throughput
821 tool for assessing microbial community responses to environmental stresses. *Environ*
822 *Sci Technol*. 2013;47:9841–9.

823 18. Abot A, Arnal G, Auer L, Lazuka A, Labourdette D, Lamarre S, et al. CAZyChip:
824 dynamic assessment of exploration of glycoside hydrolases in microbial ecosystems.
825 *BMC Genomics* [Internet]. *BMC Genomics*; 2016;17:671. Available from:
826 <http://www.ncbi.nlm.nih.gov/pubmed/27552843>
827 [http://www.pubmedcentral.nih.g](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4994258)
828 19. Tedersoo L, Drenkhan R, Anslan S, Morales-Rodriguez C, Cleary M. High-
829 throughput identification and diagnostics of pathogens and pests: Overview and
830 practical recommendations. *Mol Ecol Resour* [Internet]. 2019;19:47–76. Available
831 from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12959>

832 20. Franzosa E a., Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al.
833 Sequencing and beyond: integrating molecular “omics” for microbial community
834 profiling. *Nat Rev Microbiol* [Internet]. Nature Publishing Group; 2015;13:360–72.
835 Available from:
836 <http://www.nature.com/doi/abs/10.1038/nrmicro3451>
837 [http://www.ncbi.nlm.nih.g](http://www.ncbi.nlm.nih.gov/pubmed/25915636)

838 21. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C.
839 Computational meta’omics for microbial community studies. *Mol Syst Biol* [Internet].
840 Nature Publishing Group; 2013;9:1–15. Available from:
841 <http://dx.doi.org/10.1038/msb.2013.22>

842 22. Frioux C, Singh D, Korcsmaros T, Hildebrand F. From bag-of-genes to bag-of-
843 genomes: metabolic modelling of communities in the era of metagenome-assembled

844 genomes. *Comput Struct Biotechnol J* [Internet]. 2020;18:1722–34. Available from:
845 <https://linkinghub.elsevier.com/retrieve/pii/S2001037020303172>

846 23. Fierer N. Embracing the unknown: Disentangling the complexities of the soil
847 microbiome. *Nat Rev Microbiol* [Internet]. Nature Publishing Group; 2017;15:579–90.
848 Available from: <http://dx.doi.org/10.1038/nrmicro.2017.87>

849 24. Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A
850 phylogenetic perspective. *Science (80-)* [Internet]. 2015;350:aac9323–aac9323.
851 Available from: <https://www.sciencemag.org/lookup/doi/10.1126/science.aac9323>

852 25. Penton CR, Johnson TA, Quensen JF, Iwai S, Cole JR, Tiedje JM. Functional
853 genes to assess nitrogen cycling and aromatic hydrocarbon degradation: primers and
854 processing matter. *Front Microbiol* [Internet]. 2013;4. Available from:
855 <http://journal.frontiersin.org/article/10.3389/fmicb.2013.00279/abstract>

856 26. Hannula SE, van Veen JA. Primer sets developed for functional genes reveal
857 shifts in functionality of fungal community in soils. *Front Microbiol*. 2016;7.

858 27. Barbi F, Bragalini C, Vallon L, Prudent E, Dubost A, Fraissinet-Tachet L, et al.
859 PCR primers to study the diversity of expressed fungal genes encoding
860 lignocellulolytic enzymes in soils using high-throughput sequencing. *PLoS One*
861 [Internet]. 2014 [cited 2015 Jan 5];9:e116264. Available from:
862 <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116264#pone->
863 [0116264-g004](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116264#pone-0116264-g004)

864 28. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, et al. FunGene: the
865 functional gene pipeline and repository. *Front Microbiol* [Internet]. 2013;4. Available
866 from: <http://journal.frontiersin.org/article/10.3389/fmicb.2013.00291/abstract>

867 29. Angel R, Nepel M, Panhölzl C, Schmidt H, Herbold CW, Eichorst SA, et al.
868 Evaluation of primers targeting the diazotroph functional gene and development of

869 NifMAP - A bioinformatics pipeline for analyzing nifH amplicon data. *Front Microbiol*
870 [Internet]. 2018;9. Available from:
871 <http://journal.frontiersin.org/article/10.3389/fmicb.2018.00703/full>

872 30. Ortiz-Estrada AM, Gollas-Galván T, Martínez-Córdova LR, Martínez-Porchas M.
873 Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to
874 understanding the microbial ecology of aquaculture systems. *Rev Aquac* [Internet].
875 2019;11:234–45. Available from: <http://doi.wiley.com/10.1111/raq.12237>

876 31. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA,
877 Bodegom PM, et al. Structure and function of the global topsoil microbiome. *Nature*
878 [Internet]. 2018;560:233–7. Available from: [http://www.nature.com/articles/s41586-](http://www.nature.com/articles/s41586-018-0386-6)
879 018-0386-6

880 32. Hahn AS, Konwar KM, Louca S, Hanson NW, Hallam SJ. The information
881 science of microbial ecology. *Curr Opin Microbiol* [Internet]. 2016;31:209–16.
882 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1369527416300492>

883 33. Farley SS, Dawson A, Goring SJ, Williams JW. Situating ecology as a big-data
884 science: Current advances, challenges, and solutions. *Bioscience*. 2018;68:563–76.

885 34. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al.
886 The MetaCyc database of metabolic pathways and enzymes - a 2019 update.
887 *Nucleic Acids Res* [Internet]. 2020;48:D445–53. Available from:
888 <https://academic.oup.com/nar/article/48/D1/D445/5581728>

889 35. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a
890 reference resource for gene and protein annotation. *Nucleic Acids Res*.
891 2016;44:D457–62.

892 36. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The
893 carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* [Internet].

894 2014 [cited 2014 Jul 10];42:D490-5. Available from:
895 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965031&tool=pmcentrez>
896 &rendertype=abstract

897 37. Starke R, Capek P, Morais D, Callister SJ, Jehmlich N. The total microbiome
898 functions in bacteria and fungi. *J Proteomics* [Internet]. 2020;213:103623. Available
899 from: <https://linkinghub.elsevier.com/retrieve/pii/S1874391919303951>

900 38. Commichaux S, Shah N, Ghurye J, Stoppel A, Goodheart JA, Luque GG, et al. A
901 critical assessment of gene catalogs for metagenomic analysis. Birol I, editor.
902 *Bioinformatics* [Internet]. 2021; Available from:
903 [https://academic.oup.com/bioinformatics/advance-](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btab216/6207961)
904 [article/doi/10.1093/bioinformatics/btab216/6207961](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btab216/6207961)

905 39. Zhou Y, Coventry DR, Gupta VVSR, Fuentes D, Merchant A, Kaiser BN, et al.
906 The preceding root system drives the composition and function of the rhizosphere
907 microbiome. *Genome Biol* [Internet]. 2020;21:89. Available from:
908 <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-01999-0>

909 40. Baldrian P. The known and the unknown in soil microbial ecology. *FEMS*
910 *Microbiol Ecol* [Internet]. 2019;95. Available from:
911 <https://academic.oup.com/femsec/article/doi/10.1093/femsec/fiz005/5281230>

912 41. Blondel J. Guilds or functional groups: Does it matter? *Oikos* [Internet].
913 2003;100:223–31. Available from: [http://doi.wiley.com/10.1034/j.1600-](http://doi.wiley.com/10.1034/j.1600-0706.2003.12152.x)
914 [0706.2003.12152.x](http://doi.wiley.com/10.1034/j.1600-0706.2003.12152.x)

915 42. Mlambo MC. Not all traits are “functional”: Insights from taxonomy and
916 biodiversity-ecosystem functioning research. *Biodivers Conserv* [Internet].
917 2014;23:781–90. Available from: <http://link.springer.com/10.1007/s10531-014-0618-5>

918 43. Voltaire F, Gleason SM, Delzon S. What do you mean “functional” in ecology?

919 Patterns versus processes. *Ecol Evol* [Internet]. 2020;10:11875–85. Available from:
920 <https://onlinelibrary.wiley.com/doi/10.1002/ece3.6781>

921 44. Escalas A, Hale L, Voordeckers JW, Yang Y, Firestone MK, Alvarez-Cohen L, et
922 al. Microbial functional diversity: From concepts to applications. *Ecol Evol* [Internet].
923 2019;9:12000–16. Available from:
924 <https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5670>

925 45. Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, Allison SD. Defining
926 trait-based microbial strategies with consequences for soil carbon cycling under
927 climate change. *ISME J* [Internet]. 2020;14:1–9. Available from:
928 <http://www.nature.com/articles/s41396-019-0510-0>

929 46. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, et al. A
930 synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* [Internet].
931 2020;7:170. Available from: <http://www.nature.com/articles/s41597-020-0497-4>

932 47. Lajoie G, Kembel SW. Making the Most of Trait-Based Approaches for Microbial
933 Ecology. *Trends Microbiol* [Internet]. 2019;27:814–23. Available from:
934 <https://linkinghub.elsevier.com/retrieve/pii/S0966842X19301581>

935 48. Reimer LC, Söhngen C, Vetschinova A, Overmann J. Mobilization and integration
936 of bacterial phenotypic data—Enabling next generation biodiversity analysis through
937 the Bac Dive metadatabase. *J Biotechnol* [Internet]. 2017;261:187–93. Available
938 from: <https://linkinghub.elsevier.com/retrieve/pii/S0168165617302067>

939 49. Endara L, Cui H, Burleigh JG. Extraction of phenotypic traits from taxonomic
940 descriptions for the tree of life using natural language processing. *Appl Plant Sci*
941 [Internet]. 2018;6:e1035. Available from: <http://doi.wiley.com/10.1002/aps3.1035>

942 50. Lim KMK, Li C, Chng KR, Nagarajan N. @MInter: Automated text-mining of
943 microbial interactions. *Bioinformatics* [Internet]. 2016;32:2981–7. Available from:

944 <https://academic.oup.com/bioinformatics/article->
945 [lookup/doi/10.1093/bioinformatics/btw357](https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw357)
946 51. Chaix E, Deléger L, Bossy R, Nédellec C. Text mining tools for extracting
947 information about microbial biodiversity in food. *Food Microbiol* [Internet].
948 2019;81:63–75. Available from:
949 <https://linkinghub.elsevier.com/retrieve/pii/S0740002017310638>
950 52. Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, et al. An
951 ontology for microbial phenotypes. *BMC Microbiol* [Internet]. 2014;14:294. Available
952 from: <http://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-014-0294-3>
953 53. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. Phenotype Inference
954 from Text and Genomic Data. *Lect Notes Comput Sci (including Subser Lect Notes*
955 *Artif Intell Lect Notes Bioinformatics)* [Internet]. 2017. p. 373–7. Available from:
956 http://link.springer.com/10.1007/978-3-319-71273-4_34
957 54. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al.
958 Predictive functional profiling of microbial communities using 16S rRNA marker gene
959 sequences. *Nat Biotechnol* [Internet]. 2013;31:814–21. Available from:
960 <http://www.ncbi.nlm.nih.gov/pubmed/23975157>
961 55. Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, et al. FUNGuild:
962 An open annotation tool for parsing fungal community datasets by ecological guild.
963 *Fungal Ecol* [Internet]. Elsevier Ltd; 2016;20:241–8. Available from:
964 <http://dx.doi.org/10.1016/j.funeco.2015.06.006>
965 56. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: Predicting functional
966 profiles from metagenomic 16S rRNA data. *Bioinformatics* [Internet]. 2015 [cited 2015
967 May 13];31:2882–4. Available from:
968 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4547618&tool=pmcentrez>

969 &rendertype=abstract

970 57. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global
971 ocean microbiome. *Science* (80-) [Internet]. 2016;353:1272–7. Available from:
972 <https://www.sciencemag.org/lookup/doi/10.1126/science.aaf4507>

973 58. Segata N, Huttenhower C. Toward an Efficient Method of Identifying Core Genes
974 for Evolutionary and Functional Microbial Phylogenies. Gibas C, editor. *PLoS One*
975 [Internet]. 2011;6:e24704. Available from:
976 <https://dx.plos.org/10.1371/journal.pone.0024704>

977 59. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. *Nat*
978 *Genet* [Internet]. 1999;21:108–10. Available from:
979 http://www.nature.com/articles/ng0199_108

980 60. Hartman WH, Ye R, Horwath WR, Tringe SG. A genomic perspective on
981 stoichiometric regulation of soil carbon cycling. *ISME J* [Internet]. 2017;11:2652–65.
982 Available from: <http://www.nature.com/articles/ismej2017115>

983 61. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S Gene Copy Number
984 Information Improves Estimates of Microbial Diversity and Abundance. von Mering C,
985 editor. *PLoS Comput Biol* [Internet]. 2012;8:e1002743. Available from:
986 <https://dx.plos.org/10.1371/journal.pcbi.1002743>

987 62. Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in
988 microbiome surveys remains an unsolved problem. *Microbiome* [Internet]. 2018;6:41.
989 Available from:
990 <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L620>

991 [952556%0Ahttp://dx.doi.org/10.1186/s40168-018-0420-9](http://dx.doi.org/10.1186/s40168-018-0420-9)

992 63. Woloszynek S, Mell JC, Zhao Z, Simpson G, O'Connor MP, Rosen GL. Exploring
993 thematic structure and predicted functionality of 16S rRNA amplicon data. Loor JJ,

994 editor. PLoS One [Internet]. 2019;14:e0219235. Available from:
995 <https://dx.plos.org/10.1371/journal.pone.0219235>

996 64. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a
997 web-based tool for comprehensive statistical, visual and meta-analysis of microbiome
998 data. Nucleic Acids Res [Internet]. 2017;45:W180–8. Available from:
999 <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx295>

1000 65. Bowman JS, Ducklow HW. Microbial communities can be described by metabolic
1001 structure: A general framework and application to a seasonally variable, depth-
1002 stratified microbial community from the coastal West Antarctic Peninsula. PLoS One.
1003 2015;10:1–18.

1004 66. R Core Team. R: A Language and Environment for Statistical Computing
1005 [Internet]. Vienna, Austria; 2020. Available from: <https://www.r-project.org/>

1006 67. Iwai S, Weinmaier T, Schmidt BL, Albertson DG, Poloso NJ, Dabbagh K, et al.
1007 Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from
1008 Human Microbiomes. He Z, editor. PLoS One [Internet]. 2016;11:e0166104.
1009 Available from: <https://dx.plos.org/10.1371/journal.pone.0166104>

1010 68. Mitchell K, Ronas J, Dao C, Freise AC, Mangul S, Shapiro C, et al. PUMAA: A
1011 Platform for Accessible Microbiome Analysis in the Undergraduate Classroom. Front
1012 Microbiol [Internet]. 2020;11. Available from:
1013 <https://www.frontiersin.org/article/10.3389/fmicb.2020.584699/full>

1014 69. Zanne AE, Abarenkov K, Afkhami ME, Aguilar-Trigueros CA, Bates S, Bhatnagar
1015 JM, et al. Fungal functional ecology: bringing a trait-based approach to plant-
1016 associated fungi. Biol Rev [Internet]. 2020;95:409–33. Available from:
1017 <https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12570>

1018 70. Pölme S, Abarenkov K, Henrik Nilsson R, Lindahl BD, Clemmensen KE,

1019 Kauserud H, et al. FungalTraits: a user-friendly traits database of fungi and fungus-
1020 like stramenopiles. *Fungal Divers* [Internet]. 2020;105:1–16. Available from:
1021 <http://link.springer.com/10.1007/s13225-020-00466-2>

1022 71. Fones HN, Beber DP, Chaloner TM, Kay WT, Steinberg G, Gurr SJ. Threats to
1023 global food security from emerging fungal and oomycete crop pathogens. *Nat Food*
1024 [Internet]. 2020;1:332–42. Available from: [http://www.nature.com/articles/s43016-](http://www.nature.com/articles/s43016-020-0075-0)
1025 [020-0075-0](http://www.nature.com/articles/s43016-020-0075-0)

1026 72. Agerer R, Rambold G. DEEMY—an information system for characterization and
1027 determination of ectomycorrhizae. München, Ger [Internet]. 2004; Available from:
1028 <http://www.deemy.de/>

1029 73. Barberán A, Caceres Velazquez H, Jones S, Fierer N. Hiding in Plain Sight:
1030 Mining Bacterial Species Records for Phenotypic Trait Information. Hallam SJ, editor.
1031 *mSphere* [Internet]. 2017;2. Available from:
1032 <https://msphere.asm.org/content/2/4/e00237-17>

1033 74. Reimer LC, Vetschinova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, et al.
1034 Bac Dive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis.
1035 *Nucleic Acids Res* [Internet]. 2019;47:D631–6. Available from:
1036 <https://academic.oup.com/nar/article/47/D1/D631/5106998>

1037 75. Engqvist MKM. Correlating enzyme annotations with a large set of microbial
1038 growth temperatures reveals metabolic adaptations to growth at diverse
1039 temperatures. *BMC Microbiol* [Internet]. 2018;18:177. Available from:
1040 <https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-018-1320-7>

1041 76. Nkongolo KK, Narendrula-Kotha R. Advances in monitoring soil microbial
1042 community dynamic and function. *J Appl Genet* [Internet]. 2020;61:249–63. Available
1043 from: <http://link.springer.com/10.1007/s13353-020-00549-5>

1044 77. Jin T, Wang Y, Huang Y, Xu J, Zhang P, Wang N, et al. Taxonomic structure and
1045 functional association of foxtail millet root microbiome. *Gigascience*. 2017;6:1–12.

1046 78. Lian T, Mu Y, Jin J, Ma Q, Cheng Y, Cai Z, et al. Impact of intercropping on the
1047 coupling between soil microbial community structure, activity, and nutrient-use
1048 efficiencies. *PeerJ* [Internet]. 2019;7:e6412. Available from:
1049 <https://peerj.com/articles/6412>

1050 79. Sengupta A, Hariharan J, Grewal PS, Dick WA. Bacterial community dissimilarity
1051 in soils is driven by long-term land-use practices. *Agrosystems, Geosci Environ*.
1052 2020;3.

1053 80. Lüneberg K, Schneider D, Siebe C, Daniel R. Drylands soil bacterial community
1054 is affected by land use change and different irrigation practices in the Mezquital
1055 Valley, Mexico. *Sci Rep* [Internet]. 2018;8:1413. Available from:
1056 <http://www.nature.com/articles/s41598-018-19743-x>

1057 81. Sun S, Jones RB, Fodor AA. Inference-based accuracy of metagenome
1058 prediction tools varies across sample types and functional categories. *Microbiome*
1059 [Internet]. 2020;8:46. Available from:
1060 <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00815-y>

1061 82. Koo H, Hakim JA, Morrow CD, Eipers PG, Davila A, Andersen DT, et al.
1062 Comparison of two bioinformatics tools used to characterize the microbial diversity
1063 and predictive functional attributes of microbial mats from Lake Obersee, Antarctica.
1064 *J Microbiol Methods* [Internet]. 2017;140:15–22. Available from:
1065 <https://linkinghub.elsevier.com/retrieve/pii/S0167701217301781>

1066 83. Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits
1067 in microorganisms. *ISME J* [Internet]. 2013;7:830–8. Available from:
1068 <http://www.nature.com/articles/ismej2012160>

- 1069 84. George PBL, Creer S, Griffiths RI, Emmett BA, Robinson DA, Jones DL. Primer
1070 and Database Choice Affect Fungal Functional but Not Biological Diversity Findings
1071 in a National Soil Survey. *Front Environ Sci* [Internet]. 2019;7. Available from:
1072 <https://www.frontiersin.org/article/10.3389/fenvs.2019.00173/full>
- 1073 85. Yang T, Adams JM, Shi Y, He JS, Jing X, Chen L, et al. Soil fungal diversity in
1074 natural grasslands of the Tibetan Plateau: associations with plant diversity and
1075 productivity. *New Phytol* [Internet]. 2017;215:756–65. Available from:
1076 <https://onlinelibrary.wiley.com/doi/10.1111/nph.14606>
- 1077 86. Makiola A, Dickie IA, Holdaway RJ, Wood JR, Orwin KH, Glare TR. Land use is a
1078 determinant of plant pathogen alpha- but not beta-diversity. *Mol Ecol* [Internet].
1079 2019;28:3786–98. Available from:
1080 <https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15177>
- 1081 87. Buscardo E, Souza RC, Meir P, Geml J, Schmidt SK, da Costa ACL, et al. Effects
1082 of natural and experimental drought on soil fungi and biogeochemistry in an Amazon
1083 rain forest. *Commun Earth Environ* [Internet]. 2021;2:55. Available from:
1084 <http://www.nature.com/articles/s43247-021-00124-8>
- 1085 88. Liang M, Liu X, Parker IM, Johnson D, Zheng Y, Luo S, et al. Soil microbes drive
1086 phylogenetic diversity-productivity relationships in a subtropical forest. *Sci Adv*
1087 [Internet]. 2019;5:eaax5088. Available from:
1088 <https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aax5088>
- 1089 89. Yang T, Tedersoo L, Soltis PS, Soltis DE, Gilbert JA, Sun M, et al. Phylogenetic
1090 imprint of woody plants on the soil mycobiome in natural mountain forests of eastern
1091 China. *ISME J* [Internet]. 2019;13:686–97. Available from:
1092 <http://www.nature.com/articles/s41396-018-0303-x>
- 1093 90. Brinkmann N, Schneider D, Sahner J, Ballauff J, Edy N, Barus H, et al. Intensive

1094 tropical land use massively shifts soil fungal communities. *Sci Rep* [Internet].
1095 2019;9:3403. Available from: <http://www.nature.com/articles/s41598-019-39829-4>
1096 91. Egidi E, Delgado-Baquerizo M, Plett JM, Wang J, Eldridge DJ, Bardgett RD, et al.
1097 A few Ascomycota taxa dominate soil fungal communities worldwide. *Nat Commun*
1098 [Internet]. 2019;10:2369. Available from: <http://www.nature.com/articles/s41467-019->
1099 10373-z

1100 92. Delgado-Baquerizo M, Guerra CA, Cano-Díaz C, Egidi E, Wang J-T, Eisenhauer
1101 N, et al. The proportion of soil-borne pathogens increases with warming at the global
1102 scale. *Nat Clim Chang* [Internet]. 2020;10:550–4. Available from:
1103 <http://www.nature.com/articles/s41558-020-0759-3>

1104 93. Větrovský T, Kohout P, Kopecký M, Machac A, Man M, Bahnmann BD, et al. A
1105 meta-analysis of global fungal distribution reveals climate-driven patterns. *Nat*
1106 *Commun* [Internet]. 2019;10:5142. Available from:
1107 <http://www.nature.com/articles/s41467-019-13164-8>

1108 94. Öpik M, Davison J, Moora M, Zobel M. DNA-based detection and identification of
1109 Glomeromycota: the virtual taxonomy of environmental sequences. *Botany* [Internet].
1110 2014;92:135–47. Available from: <http://www.nrcresearchpress.com/doi/10.1139/cjb->
1111 2013-0110

1112 95. Berruti A, Desirò A, Visentin S, Zecca O, Bonfante P. ITS fungal barcoding
1113 primers versus 18S AMF-specific primers reveal similar AMF-based diversity patterns
1114 in roots and soils of three mountain vineyards. *Environ Microbiol Rep* [Internet].
1115 2017;9:658–67. Available from: <http://doi.wiley.com/10.1111/1758-2229.12574>

1116 96. Anthony MA, Frey SD, Stinson KA. Fungal community homogenization, shift in
1117 dominant trophic guild, and appearance of novel taxa with biotic invasion. *Ecosphere*
1118 [Internet]. 2017;8:e01951. Available from: <http://doi.wiley.com/10.1002/ecs2.1951>

1119 97. Sansupa C, Wahdan SFM, Hossen S, Disayathanoowat T, Wubet T, Purahong
1120 W. Can we use functional annotation of prokaryotic taxa (Faprotax) to assign the
1121 ecological functions of soil bacteria? *Appl Sci* [Internet]. 2021;11:1–17. Available
1122 from: <https://www.mdpi.com/2076-3417/11/2/688>

1123 98. Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L.
1124 Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat Rev*
1125 *Microbiol* [Internet]. 2019;17:95–109. Available from:
1126 <http://www.nature.com/articles/s41579-018-0116-y>

1127 99. Comeau AM, Douglas GM, Langille MGI. Microbiome Helper: a Custom and
1128 Streamlined Workflow for Microbiome Research. Eisen J, editor. *mSystems* [Internet].
1129 2017;2. Available from: <https://msystems.asm.org/content/2/1/e00127-16>

1130 100. Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, et al.
1131 Prospects and challenges of implementing DNA metabarcoding for high-throughput
1132 insect surveillance. *Gigascience* [Internet]. 2019;8:1–22. Available from:
1133 <https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giz092/55416>
1134 30

1135 101. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and
1136 evolution. *Nat Rev Microbiol* [Internet]. 2005;3:679–87. Available from:
1137 <http://www.nature.com/articles/nrmicro1204>

1138 102. Treseder KK, Lennon JT. Fungal Traits That Drive Ecosystem Dynamics on
1139 Land. *Microbiol Mol Biol Rev* [Internet]. 2015;79:243–62. Available from:
1140 <http://mmbbr.asm.org/lookup/doi/10.1128/MMBR.00001-15>

1141 103. Smalla K, Jechalke S, Top EM. Plasmid Detection, Characterization, and
1142 Ecology. *Microbiol Spectr* [Internet]. 2015;3. Available from:
1143 <http://www.asmscience.org/content/journal/microbiolspec/10.1128/microbiolspec.PLA>

1144 S-0038-2014

1145 104. Dunivin TK, Choi J, Howe A, Shade A. RefSoil+: a Reference Database for
1146 Genes and Traits of Soil Plasmids. Sharpton TJ, editor. mSystems [Internet]. 2019;4.
1147 Available from: <https://msystems.asm.org/content/4/1/e00349-18>

1148 105. Aminov RI. Horizontal Gene Exchange in Environmental Microbiota. Front
1149 Microbiol [Internet]. 2011;2. Available from:
1150 <http://journal.frontiersin.org/article/10.3389/fmicb.2011.00158/abstract>

1151 106. Banos S, Lentendu G, Kopf A, Wubet T, Glöckner FO, Reich M. A
1152 comprehensive fungi-specific 18S rRNA gene sequence primer toolkit suited for
1153 diverse research issues and sequencing platforms. BMC Microbiol [Internet].
1154 2018;18:190. Available from:
1155 <https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-018-1331-4>

1156 107. Bukin YS, Galachyants YP, Morozov I V., Bukin S V., Zakharenko AS,
1157 Zemskaya TI. The effect of 16S rRNA region choice on bacterial community
1158 metabarcoding results. Sci Data [Internet]. 2019;6:190007. Available from:
1159 <http://www.nature.com/articles/sdata20197>

1160 108. Xu Z, Malmer D, Langille MGI, Way SF, Knight R. Which is more important for
1161 classifying microbial communities: who's there or what they can do? ISME J
1162 [Internet]. International Society for Microbial Ecology; 2014 [cited 2015 Aug
1163 31];8:2357–9. Available from: <http://dx.doi.org/10.1038/ismej.2014.157>

1164 109. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M.
1165 Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene
1166 sequences without primer bias. Nat Biotechnol [Internet]. Nature Publishing Group;
1167 2018;36:190–5. Available from: <http://dx.doi.org/10.1038/nbt.4045>

1168 110. Tedersoo L, Anslan S. Towards PacBio-based pan-eukaryote metabarcoding

1169 using full-length ITS sequences. *Environ Microbiol Rep* [Internet]. 2019;11:659–68.
1170 Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-2229.12776>
1171 111. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et
1172 al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome
1173 analysis. *Nat Commun* [Internet]. 2019;10:5029. Available from:
1174 <http://www.nature.com/articles/s41467-019-13036-1>
1175 112. Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB, et
1176 al. The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol*
1177 [Internet]. 2010;8:523–9. Available from: <http://www.nature.com/articles/nrmicro2367>
1178 113. Fierer N, Bradford MA, Jackson RB. TOWARD AN ECOLOGICAL
1179 CLASSIFICATION OF SOIL BACTERIA. *Ecology* [Internet]. 2007;88:1354–64.
1180 Available from: <http://doi.wiley.com/10.1890/05-1839>
1181 114. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, et al. High-
1182 throughput amplicon sequencing of the full-length 16S rRNA gene with single-
1183 nucleotide resolution. *Nucleic Acids Res* [Internet]. 2019;47:e103–e103. Available
1184 from: <https://academic.oup.com/nar/article/47/18/e103/5527971>
1185 115. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al.
1186 PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* [Internet]. 2020;
1187 Available from: <http://www.nature.com/articles/s41587-020-0548-6>
1188 116. Feldbauer R, Schulz F, Horn M, Rattei T. Prediction of microbial phenotypes
1189 based on comparative genomics. *BMC Bioinformatics* [Internet]. 2015;16:S1.
1190 Available from: [http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S14-S1)
1191 [2105-16-S14-S1](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S14-S1)
1192 117. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC. From
1193 Genomes to Phenotypes: TraitAr, the Microbial Trait Analyzer. Segata N, editor.

1194 mSystems [Internet]. 2016;1:043315. Available from:
1195 <https://msystems.asm.org/content/1/6/e00101-16>
1196 118. Goberna M, Verdú M. Predicting microbial traits with phylogenies. ISME J
1197 [Internet]. 2016;10:959–67. Available from:
1198 <http://www.nature.com/articles/ismej2015171>
1199 119. Levatić J, Brbić M, Perdih TS, Kocev D, Vidulin V, Šmuc T, et al. Phenotype
1200 Prediction with Semi-supervised Classification Trees. Lect Notes Comput Sci
1201 (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. 2018. p.
1202 138–50. Available from: http://link.springer.com/10.1007/978-3-319-78680-3_10
1203 120. Zanne AE, Powell JR, Flores-Moreno H, Kiers ET, van 't Padjé A, Cornwell WK.
1204 Finding fungal ecological strategies: Is recycling an option? Fungal Ecol [Internet].
1205 2020;46:100902. Available from:
1206 <https://linkinghub.elsevier.com/retrieve/pii/S1754504819302181>
1207 121. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et
1208 al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics [Internet].
1209 2010;26:1463–4. Available from:
1210 <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link>
1211 [&LinkName=pubmed_pubmed&LinkReadableName=Related](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link&LinkName=pubmed_pubmed&LinkReadableName=Related)
1212 [Articles&IdsFromResult=20395285&ordinalpos=3&itool=EntrezSystem2.PEntrez.Pub](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link&LinkName=pubmed_pubmed&LinkReadableName=Related&IdsFromResult=20395285&ordinalpos=3&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum)
1213 [med.Pubmed_ResultsPanel.Pubmed_RVDocSum](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link&LinkName=pubmed_pubmed&LinkReadableName=Related&IdsFromResult=20395285&ordinalpos=3&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum) [http://www.ncbi.n](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link&LinkName=pubmed_pubmed&LinkReadableName=Related&IdsFromResult=20395285&ordinalpos=3&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum)
1214 122. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees.
1215 Valencia A, editor. Bioinformatics [Internet]. 2018;34:1053–5. Available from:
1216 <https://academic.oup.com/bioinformatics/article/34/6/1053/4582279>
1217 123. Walters KE, Martiny JBH. Alpha-, beta-, and gamma-diversity of bacteria varies
1218 across habitats. Nabout JC, editor. PLoS One [Internet]. 2020;15:e0233872.

1219 Available from: <https://dx.plos.org/10.1371/journal.pone.0233872>

1220 124. Martiny AC. High proportions of bacteria are culturable across major biomes.

1221 ISME J [Internet]. Springer US; 2019;3–6. Available from:

1222 <http://dx.doi.org/10.1038/s41396-019-0410-3>

1223 125. Wemheuer F, Taylor JA, Daniel R, Johnston E, Meinicke P, Thomas T, et al.

1224 Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy

1225 based on 16S rRNA gene sequences. Environ Microbiome [Internet]. 2020;15:11.

1226 Available from:

1227 [https://environmentalmicrobiome.biomedcentral.com/articles/10.1186/s40793-020-](https://environmentalmicrobiome.biomedcentral.com/articles/10.1186/s40793-020-00358-7)

1228 [00358-7](https://environmentalmicrobiome.biomedcentral.com/articles/10.1186/s40793-020-00358-7)

1229 126. Cheifet B. Where is genomics going next? Genome Biol [Internet]. 2019;20:17.

1230 Available from: [https://genomebiology.biomedcentral.com/articles/10.1186/s13059-](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1626-2)

1231 [019-1626-2](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1626-2)

1232 127. Piro VC, Dadi TH, Seiler E, Reinert K, Renard BY. ganon: precise

1233 metagenomics classification against large and up-to-date sets of reference

1234 sequences. Bioinformatics [Internet]. 2020;36:i12–20. Available from:

1235 https://academic.oup.com/bioinformatics/article/36/Supplement_1/i12/5870470

1236 128. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, et al.

1237 Strategies to improve reference databases for soil microbiomes. ISME J [Internet].

1238 2017;11:829–34. Available from: <http://www.nature.com/articles/ismej2016168>

1239 129. Louca S, Jacques SMS, Pires APF, Leal JS, Srivastava DS, Parfrey LW, et al.

1240 High taxonomic variability despite stable functional structure across microbial

1241 communities. Nat Ecol Evol [Internet]. 2017;1:0015. Available from:

1242 <http://www.nature.com/articles/s41559-016-0015>

1243 130. Nagpal S, Haque MM, Singh R, Mande SS. iVikodak—A Platform and Standard

1244 Workflow for Inferring, Analyzing, Comparing, and Visualizing the Functional
1245 Potential of Microbial Communities. *Front Microbiol* [Internet]. 2019;9. Available from:
1246 <https://www.frontiersin.org/article/10.3389/fmicb.2018.03336/full>

1247 131. Grigoriev I V., Nikitin R, Haridas S, Kuo A, Ohm R, Otilar R, et al. MycoCosm
1248 portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 2014;42:699–704.

1249 132. Bonkowski M, Dumack K, Fiore-Donno AM. The Protists in Soil—A Token of
1250 Untold Eukaryotic Diversity. *Mod Soil Microbiol* [Internet]. Third edition. | Boca Raton :
1251 Taylor & Francis, 2019.: CRC Press; 2019. p. 125–40. Available from:
1252 <https://www.taylorfrancis.com/books/9780429607929/chapters/10.1201/9780429059>
1253 186-8

1254 133. Xiong W, Jousset A, Guo S, Karlsson I, Zhao Q, Wu H, et al. Soil protist
1255 communities form a dynamic hub in the soil microbiome. *ISME J* [Internet].
1256 2018;12:634–8. Available from: <http://www.nature.com/articles/ismej2017171>

1257 134. Fiore-Donno AM, Richter-Heitmann T, Degruene F, Dumack K, Regan KM,
1258 Marhan S, et al. Functional Traits and Spatio-Temporal Structure of a Major Group of
1259 Soil Protists (Rhizaria: Cercozoa) in a Temperate Grassland. *Front Microbiol*
1260 [Internet]. 2019;10. Available from:
1261 <https://www.frontiersin.org/article/10.3389/fmicb.2019.01332/full>

1262 135. Delgado-Baquerizo M, Trivedi P, Trivedi C, Eldridge DJ, Reich PB, Jeffries TC,
1263 et al. Microbial richness and composition independently drive soil multifunctionality.
1264 Bennett A, editor. *Funct Ecol* [Internet]. 2017;31:2330–43. Available from:
1265 <https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.12924>

1266 136. Terrat S, Horrigue W, Dequietd S, Saby NPA, Lelièvre M, Nowak V, et al.
1267 Mapping and predictive variations of soil bacterial richness across France. *PLoS*
1268 *One.* 2017;12:5–8.

1269 137. Schloter M, Nannipieri P, Sørensen SJ, van Elsas JD. Microbial indicators for
1270 soil quality. *Biol Fertil Soils. Biology and Fertility of Soils*; 2018;54:1–10.

1271 138. Hariharan J, Sengupta A, Grewal P, Dick WA. Functional Predictions of
1272 Microbial Communities in Soil as Affected by Long-term Tillage Practices. *Agric*
1273 *Environ Lett [Internet]*. 2017;2:170031. Available from:
1274 <https://onlinelibrary.wiley.com/doi/10.2134/ael2017.09.0031>

1275 139. Manor O, Borenstein E. Systematic Characterization and Analysis of the
1276 Taxonomic Drivers of Functional Shifts in the Human Microbiome. *Cell Host Microbe*
1277 *[Internet]*. 2017;21:254–67. Available from:
1278 <https://linkinghub.elsevier.com/retrieve/pii/S1931312816305261>

1279 140. Ramirez KS, Döring M, Eisenhauer N, Gardi C, Ladau J, Leff JW, et al. Toward
1280 a global platform for linking soil biodiversity data. *Front Ecol Evol [Internet]*. 2015;3.
1281 Available from: <http://journal.frontiersin.org/Article/10.3389/fevo.2015.00091/abstract>

1282 141. Raguideau S, Plancade S, Pons N, Leclerc M, Laroche B. Inferring Aggregated
1283 Functional Traits from Metagenomic Data Using Constrained Non-negative Matrix
1284 Factorization: Application to Fiber Degradation in the Human Gut Microbiota.
1285 Maranas CD, editor. *PLOS Comput Biol [Internet]*. 2016;12:e1005252. Available
1286 from: <https://dx.plos.org/10.1371/journal.pcbi.1005252>

1287 142. Fierer N, Barberán A, Laughlin DC. Seeing the forest for the genes: Using
1288 metagenomics to infer the aggregated traits of microbial communities. *Front Microbiol*
1289 *[Internet]*. 2014;5. Available from:
1290 <http://journal.frontiersin.org/article/10.3389/fmicb.2014.00614/abstract>

1291 143. Allison SD, Martiny JBH. Resistance, resilience, and redundancy in microbial
1292 communities. *Proc Natl Acad Sci [Internet]*. 2008;105:11512–9. Available from:
1293 <http://www.pnas.org/cgi/doi/10.1073/pnas.0801925105>

1294 144. Navarro LM, Fernández N, Guerra C, Guralnick R, Kissling WD, Londoño MC,
1295 et al. Monitoring biodiversity change through effective global coordination. *Curr Opin*
1296 *Environ Sustain* [Internet]. 2017;29:158–69. Available from:
1297 <https://linkinghub.elsevier.com/retrieve/pii/S1877343517301665>

1298

1299 **Tables**

1300 **Table 1: Numbers of organisms, genes, enzymes and metabolic pathways**
1301 **available in the CAZy, KEGG and MetaCyc databases.** When possible, we detailed
1302 the number of organisms for the three domains of the tree of life. CAZy includes
1303 glycoside hydrolases (GH), glycosyl transferases (GT), carbohydrate esterases (CE),
1304 polysaccharide lyases (PL), and auxiliary activities (AA).

1305

Databases	Organisms	Metabolic Pathways	Enzymes/Genes
CAZy (Carbohydrate-Active Enzymes)	Eukaryotes (344), Bacteria (20,421), Archaea (413)	NA	GH (171), GT (114), PL (41), CE (19), AA (16)
KEGG (Kyoto Encyclopedia of Genes and Genomes)	Eukaryotes (557), Bacteria (6,317), Archaea (344)	547	KEGG Orthology (KO) groups 24,402
MetaCyc (metabolic pathways and enzymes)	Total (3,295)	2,937	13,356

1306

1307 Table 2: List of the functional inference tools, ecological trait assignment tools and databases.

1308

Tools	Implementation	Targeted genes	Functional Prediction	Approaches	Methods	Inputs used	Strengths and Specificities	Limitations
PanFP	Perl (recently Python)	16S rRNA	KEGG Orthology; Gene Ontology; Pfam; TIGRFAM	Functional inference	builds a pangenome	NCBI taxonomy	- uses functional profile of the pangenome so could be less sensitive to horizontal gene transfer	- evolutionary models are not taken into account - no confidence score generated - not yet available for microbial eukaryotes
PAPRICA	Python	16S/18S rRNA	MetaCyc ontology	Functional inference	phylogenetic placement	based on rDNA amplicon sequences	- 18S rRNA amplicons are taken into account - examples on the developer's blog	- errors may occur with sequence placement due to poor resolution of rRNA amplicons in some clades
PICRUST	Python	16S rRNA	KEGG Orthology; KEGG Pathway; COG; CAZy*	Functional inference	ASR (Wagner Parsimony, ACE ML, ACE REML, ACE PIC)	Greengenes taxonomy (18may2012 or v13.5/v13.8)	- evolutionary models are taken into account - confidence score generated (NSTI) - correction of OTU copy numbers	- based on specific taxonomy (GreenGenes identifiers) - KEGG database not updated since 2011 - no pre-calculated table of fungal genomes available

PICRUSt2	Python / R	16S/18S rRNA/ITS	MetaCyc; KEGG Orthology; EC number, COGS, Pfam, TIGRFAM	Functional inference	HSP (maximum parsimony, empirical probabilities, subtree averaging, SCP)	based on rDNA amplicon sequences	<ul style="list-style-type: none"> - evolutionary models are taken into account - confidence score generated (NSTI) - twice as many KO scores - multiple HSP methods can be implemented (takes branch length weighting into account) - 18S rRNA and ITS amplicons are taken into account - extensive documentation and active community 	- errors may occur with sequence placement due to poor resolution of rRNA amplicons in some clades
Piphillin	Web-based	16S rRNA	BioCyc; KEGG	Functional inference	Nearest-neighbor matching of 16S rRNA gene amplicons with genomes from reference databases	based on rDNA amplicon sequences	<ul style="list-style-type: none"> - regular updates of functional databases - rRNA copy number adjustment 	<ul style="list-style-type: none"> - available online only - available for 16S rRNA only
SINAPS	USEARCH	16S rRNA	Trait annotation (e.g., energy metabolism, Gram-positive staining, presence of a flagellum)	Functional inference	word counting	Greengenes; SILVA	<ul style="list-style-type: none"> - confidence is estimated by bootstrapping - integrated to USEARCH tool 	<ul style="list-style-type: none"> - no peer-reviewed publication (biorxiv preprint) - detailed explanation is missing (e.g., how was protrait input created?)

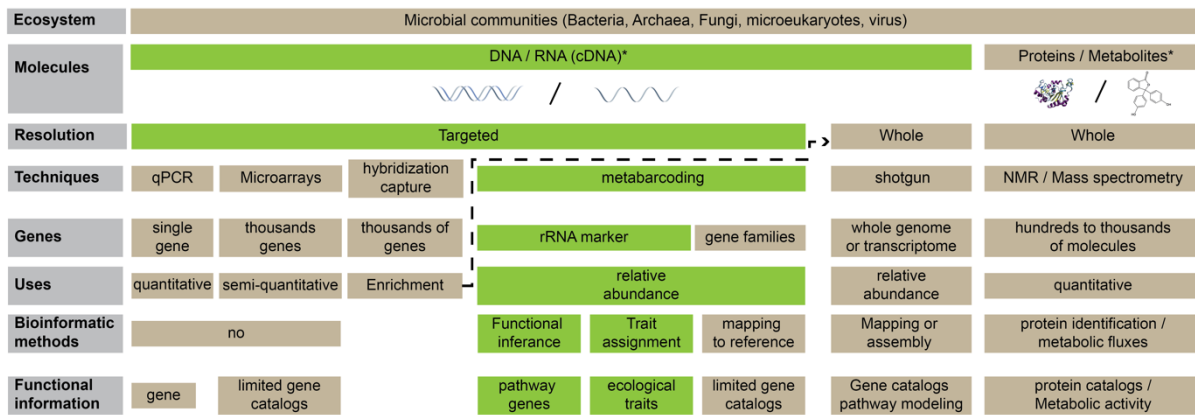
Tax4Fun	R package	16S rRNA	KEGG Orthology	Functional inference	nearest-neighbour search based on a minimum 16S rRNA sequence similarity	SILVA taxonomy	<ul style="list-style-type: none"> - uses R (multiplatform) with pre-calculated files - confidence score generated (FTU and FSU) - the algorithm could better predict poorly characterized taxa compared to approaches based on ASR with possible large distances in the tree, thanks to a minimum of similarity between sequences 	<ul style="list-style-type: none"> - based on specific taxonomy (SILVA identifiers) - KEGG database not updated since 2011
Tax4Fun2	R package	16S rRNA	KEGG Orthology	Functional inference	BLAST	based on rDNA amplicon sequences	<ul style="list-style-type: none"> - algorithm with a minimal sequence similarity - uses R (multiplatform) with pre-calculated, highly memory-efficient platform-independent files - confidence score generated (FTU and FSU) - KO update from 2018 - calculates the redundancy of specific functions directly - builds its own habitat-specific reference 	<ul style="list-style-type: none"> - not yet available for microbial eukaryotes
Vikodak	Web-based (not longer available)	16S rRNA	KEGG pathway, EC number	Functional inference	microbial co-existence patterns	RDP, SILVA	<ul style="list-style-type: none"> - pathway exclusion cut-off value is available to provide the minimum percentage of genes/enzymes belonging to a metabolic pathway required to consider the pathway as functional. - compares two datasets 	<ul style="list-style-type: none"> - not longer available - not yet available for microbial eukaryotes

iVikodak	Web-based	16S rRNA	KEGG; Pfam; COG; TIGRfam	Functional inference	microbial co-inhabitation patterns	RDP, Greengenes, SILVA	<ul style="list-style-type: none"> - user-friendly for non-expert bioinformaticians - integrated tools for statistical comparisons - graphical visualizations 	<ul style="list-style-type: none"> - available online only - not yet available for microbial eukaryotes
FUNGuild	Python / Web-based	ITS	Guild type	Trait assignment	not applicable	based on UNITE taxonomy (ITS)	<ul style="list-style-type: none"> - trait quality for taxon assignment 	<ul style="list-style-type: none"> - no regular update - 18S rRNA taxonomy with related database not included. However, the database is open-access, and a homemade wrapper can be used for 18S metabarcoding output
FAPROTAX	Python; flat file	16S rRNA	Ecological functions (e.g., nitrification, denitrification or fermentation)	Trait assignment; Database	If all type strains of a species at the genus level share the function, FAPROTAX assumes that all uncultured organisms of this genus possess the putative function	SILVA (128, 132)	<ul style="list-style-type: none"> - based on the literature of cultured taxa - availability of all literature to create the database - functions easily added to the tool 	<ul style="list-style-type: none"> - implicit assumption (see algorithm column) could be false with the increase of newly cultured organisms - does not infer upper rank when taxonomic resolution is poor
BacDive	Python and R API, R package	/	Morphology, physiology (API®-tests), molecular data, and cultivation conditions	Database	not applicable	NCBI taxonomy	<ul style="list-style-type: none"> - provides links to ENA, GenBank, SILVA, BRENDA, GBIF, ChEBI, Straininfo website data - a match with 16S rRNA sequences is available from SILVA 	<ul style="list-style-type: none"> - does not provide a tool for metabarcoding output
BugBase	R / Python	16S rRNA	KEGG	Functional inference	PICRUSt; custom trait assignment	Greengenes	<ul style="list-style-type: none"> - biologically interpretable traits (Gram staining, oxygen tolerance, biofilm formation, pathogenicity, mobile element content and oxidative stress tolerance) 	<ul style="list-style-type: none"> - no peer-reviewed publication (biorxiv preprint)

IJSEM	flat file with R script for curation	/	IJSEM	Database	not applicable	not applicable	- 16S rRNA accession numbers available	- does not provide a tool for metabarcoding output
ProTraits	Web-based; flat files	/	Wikipedia; MicrobeWiki; HAMAP proteomes; PubMed abstracts and publications; Bacmap; Genoscope; JGI, KEGG, NCBI; Karyn's Genomes	Database	not applicable	not applicable	- phenotypic inference - large resource (~545,000 phenotypes scanning 424 traits across 3,046 species) - NCBI taxonomy available	- does not provide a tool for metabarcoding output
BURRITO	Web-based	16S rRNA	KEGG Orthology	Functional inference	PICRUSt	Greengenes	- explores simultaneous and integrative studies of taxonomic and functional profiles	- based on PICRUSt v1
MACADAM	Python / web implementation	16S rRNA	MetaCyc, MicroCyc, FAPROTAX; IJSEM	Functional inference; Trait assignment	custom methods (provides functional information about upper-rank taxa when organism name is not found)	NCBI taxonomy	- pathway score and pathway frequency score are provided, allowing knowledge of number of enzymes present in the pathway	- not yet available for microbial eukaryotes
FunFun	R package; flat file	/	Ecological traits	Trait assignment	not applicable	based on UNITE taxonomy (ITS)	- uses R (multiplatform) - complementary to FUNGuild	
Fungal-Traits	flat files	/	Guild type, body type, habitat	Trait assignment	not applicable	based on UNITE taxonomy (ITS)	- expert work to propose traits at the genus level - merges the FUNGuild and FunFun tools - an excel file with vlookup function is available to assign guilds or trait data	- does not provide a tool for metabarcoding output

DEEMY	Web-based	/	Morphology, anatomy, potential for chemical reactions, or even ecology traits	Database	not applicable	not applicable	- link to tree species associated - includes images	- specialized in ectomycorrhizas only
Bacteria-archaea-traits	R package; flat file	16S rRNA	Traits, phenotypic traits, quantitative genomic traits	Database	not applicable	NCBI taxonomy, GTDB taxonomy	- groups the major bacterial and archaeal databases into one database - traits and species data condensed - R workflow available to retrieve condensed trait and species data	
OntoBiotope	Web-based	/	Habitats and phenotypes	Database	ToMap (Text to ontology mapping)	NCBI taxonomy	- term relevance is evaluated by the semantic search engine PubMed-Biotope - maintained by around 30 microbiology experts	- dedicated to the food domain
@Minter	Python	/	Microbial interactions	Machine learning	Support-vector machine (SVM)-based classifier	No specific taxonomy, just species level	- original approach to get information on microbial interactions rapidly	- species name required

Figures



*DNA: potential functional profiling, RNA/protein: expression functional profiling, Metabolite: activity profiling

Figure 1: **Schematic diagram of the various strategies available for exploring the functional diversity of the microbiota.** Green frames, metabarcoding approaches for retrieving putative functions from taxonomic genes by functional inference and ecological trait assignment.

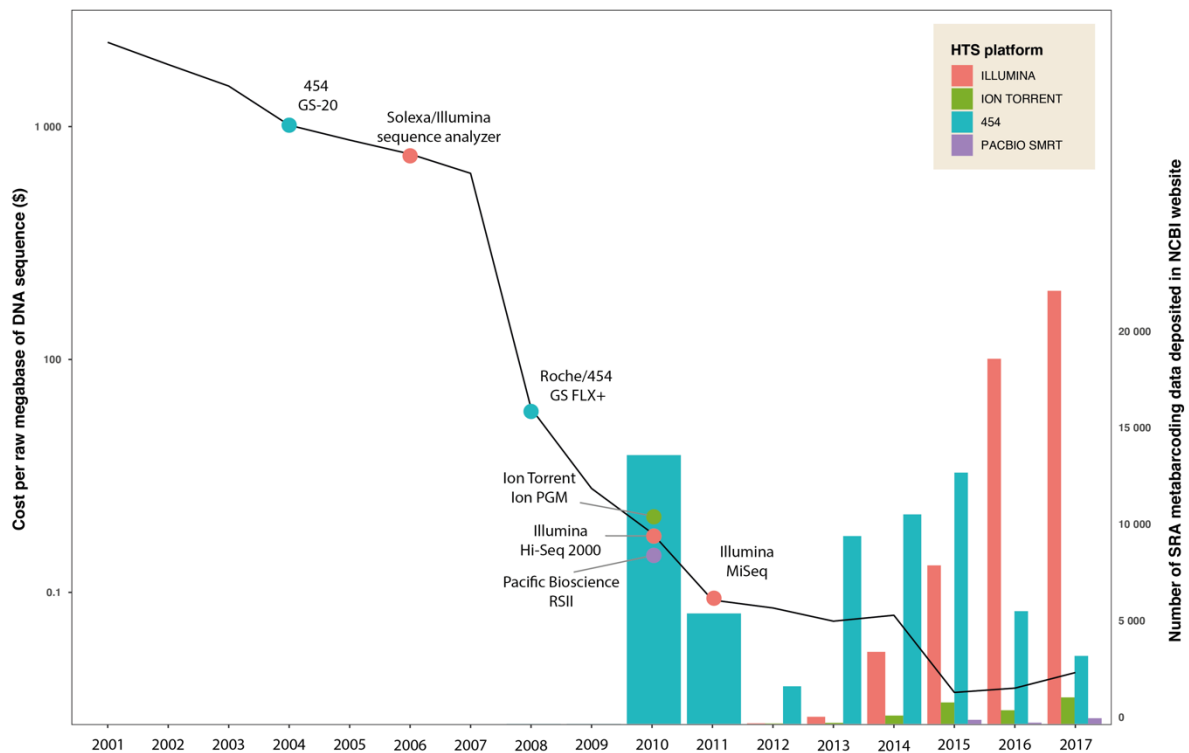


Figure 2: **Evolution of costs (dollars) per raw megabase of DNA sequence (black line with logarithmic scale), and evolution of the number of SRA metabarcoding data deposited in the NCBI website.** The data used to draw this figure is described in Additional file 1.

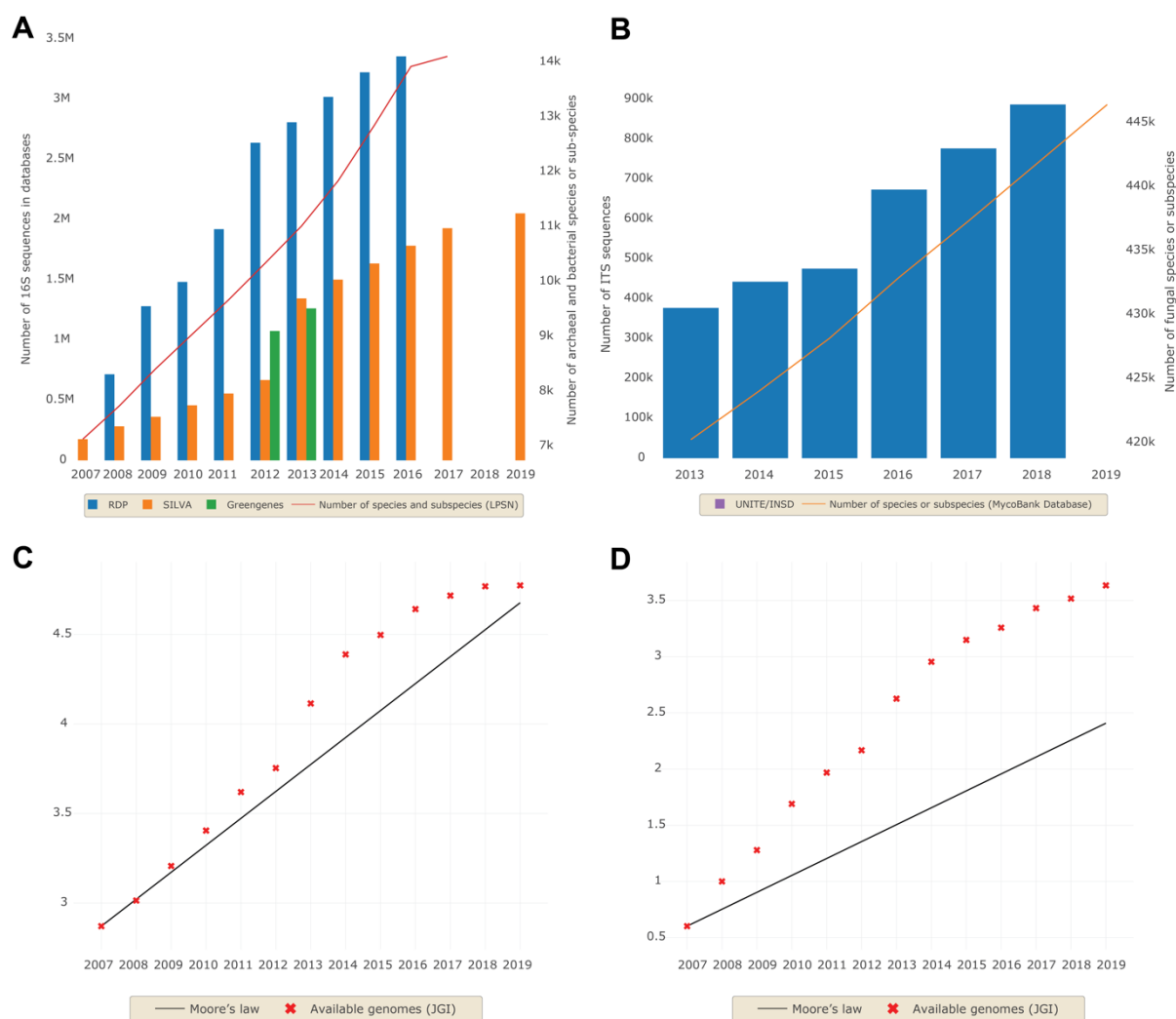


Figure 3: Annual cumulative growth of databases in terms of bacterial/archaeal (A) and fungal (B) sequences, and species/subspecies deposited *per year*. Comparison of the annual cumulative growth of bacterial/archaeal (C) and fungal (D) genomes compared to simulations of Moore's law. The plot is in logarithmic scale. Three databases were compared for 16S rRNA gene sequences: RDP (blue), SILVA (orange), Greengenes (green). Information is based on the List of Prokaryotic names with Standing in Nomenclature (LPSN [125], <http://www.bacterio.net>) website for bacterial and archaeal species, and on the MycoBank database for fungal species ([126], <http://www.mycobank.org>). Information about the bacterial, archaeal and fungal genomes is based on the Genome OnLine Database (GOLD) [127].

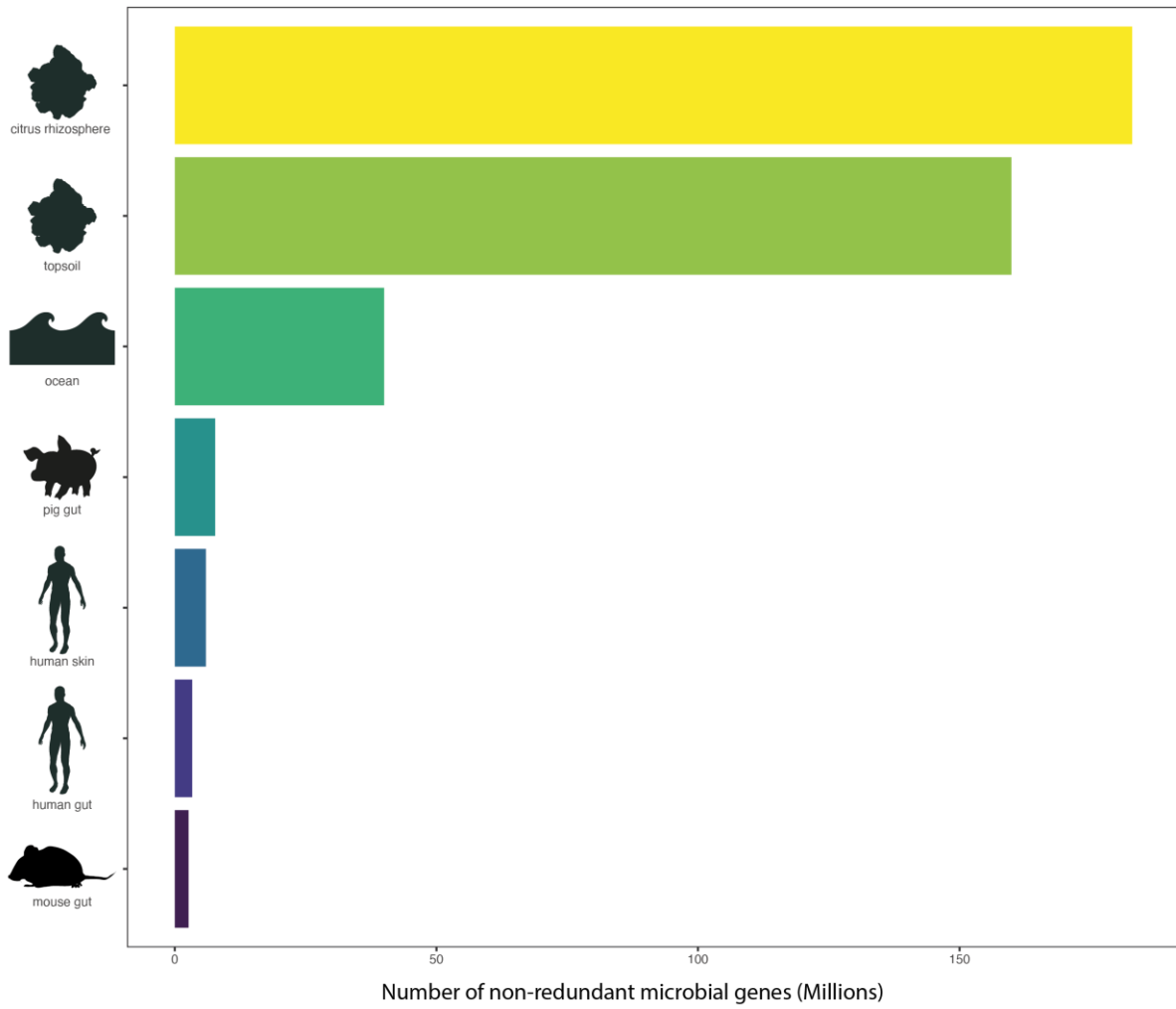


Figure 4: **Global microbial gene catalogs from various ecosystems.** The references are listed in Additional file 1.

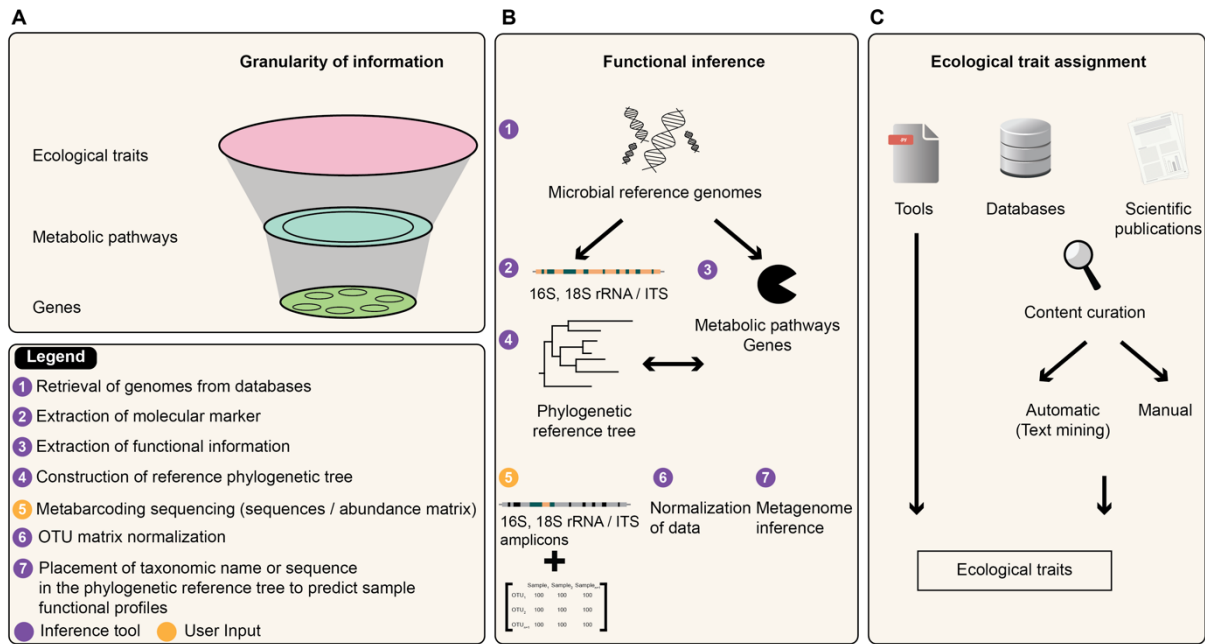


Figure 5: **Diagram of the granularity of the data (A) that can be obtained by functional inference (B) or ecological trait assignment (C).**

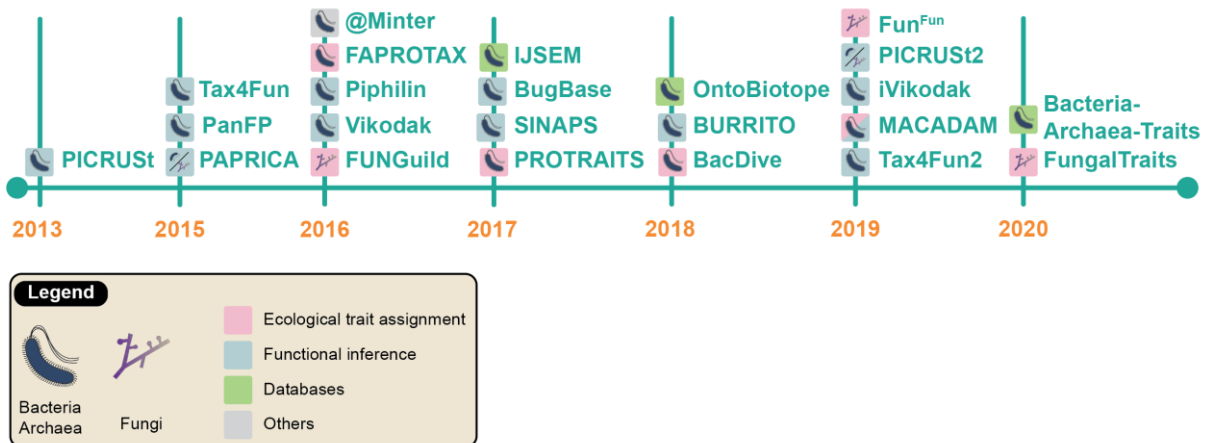


Figure 6: **Timeline depicting the historical record of the major tools developed for functional inference or ecological trait assignment.** The first version of the DEEMY database dates back to 1996; it was not included for aesthetic reasons.

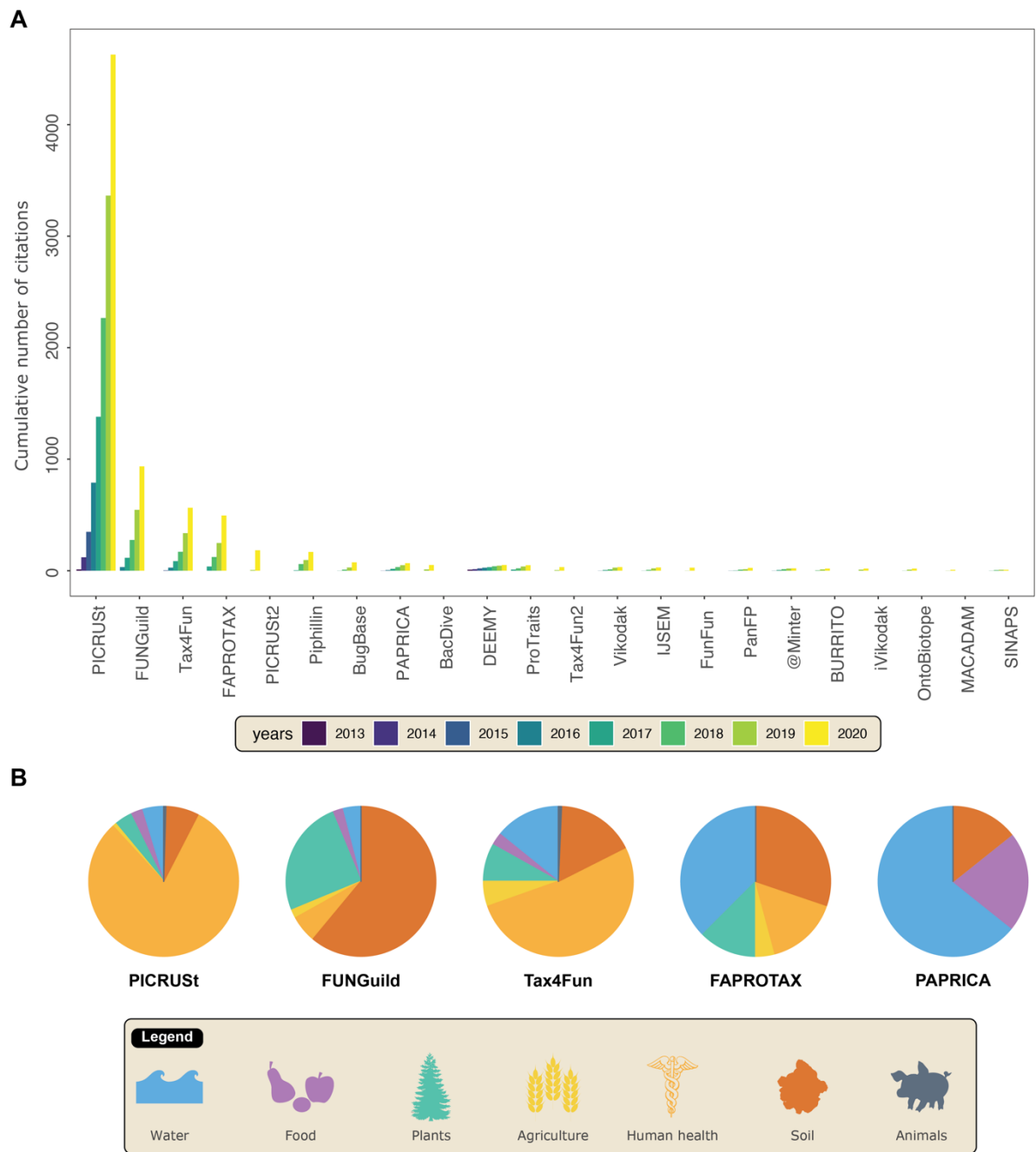


Figure 7: **Annual cumulative number of citations of the major tools (A) and their scope (B).** The keywords used for “scope” were retrieved from the titles and abstracts of the papers listed in Additional file 1.

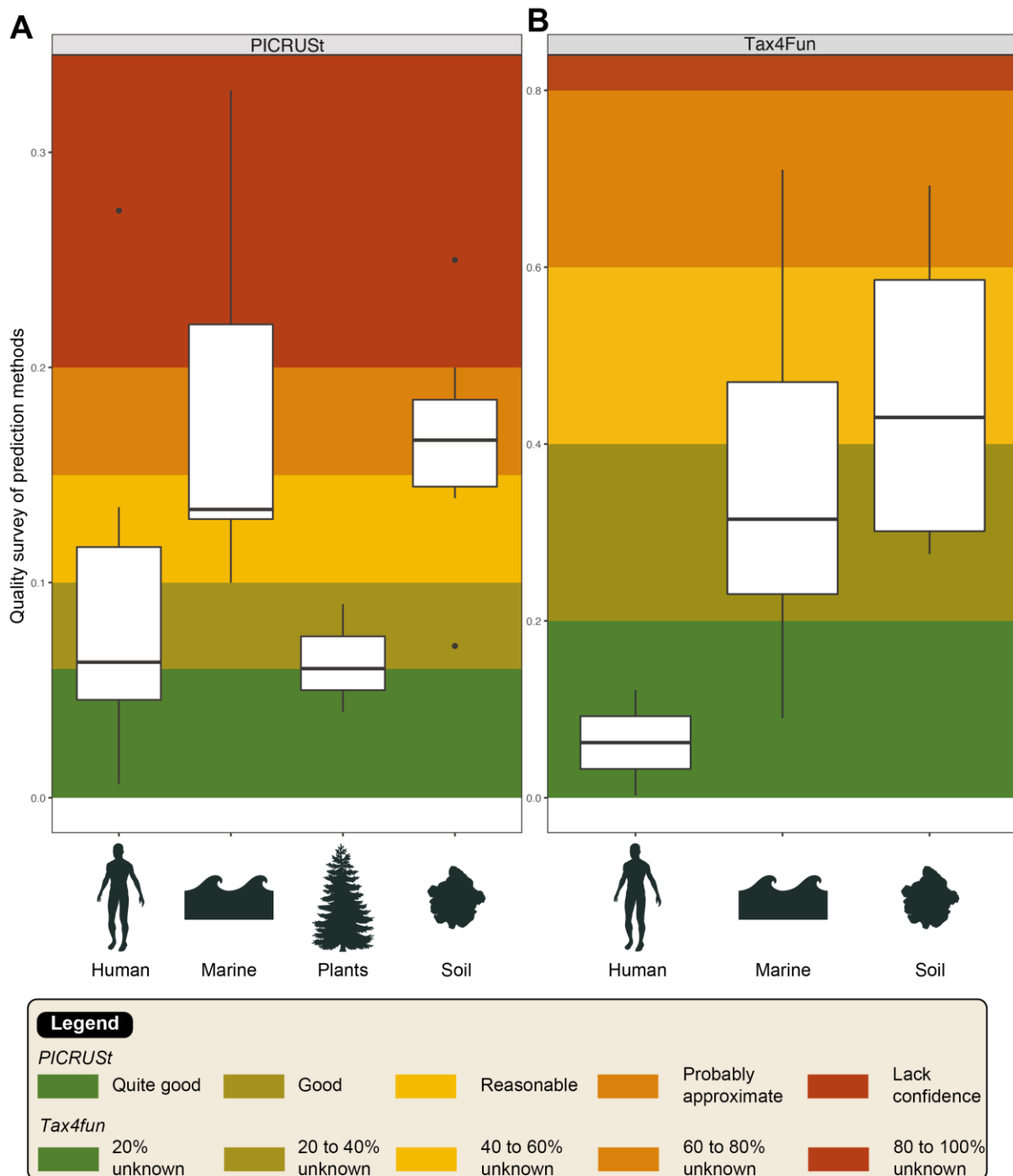


Figure 8: Overview of the quality of functional prediction based on a subsampling of articles for PICRUSt (A) and Tax4Fun (B) across various ecosystems. For PICRUSt, colors were assigned according NSTI results: < 0.06, quite good; 0.06 to 0.10, good; 0.10 to 0.15, reasonable but probably approximate; and > 0.20, probably unreliable. For Tax4Fun, we split the fraction of OTUs that could not be mapped to KEGG organisms in 5 harmonious groups. References are listed in Additional file 1.

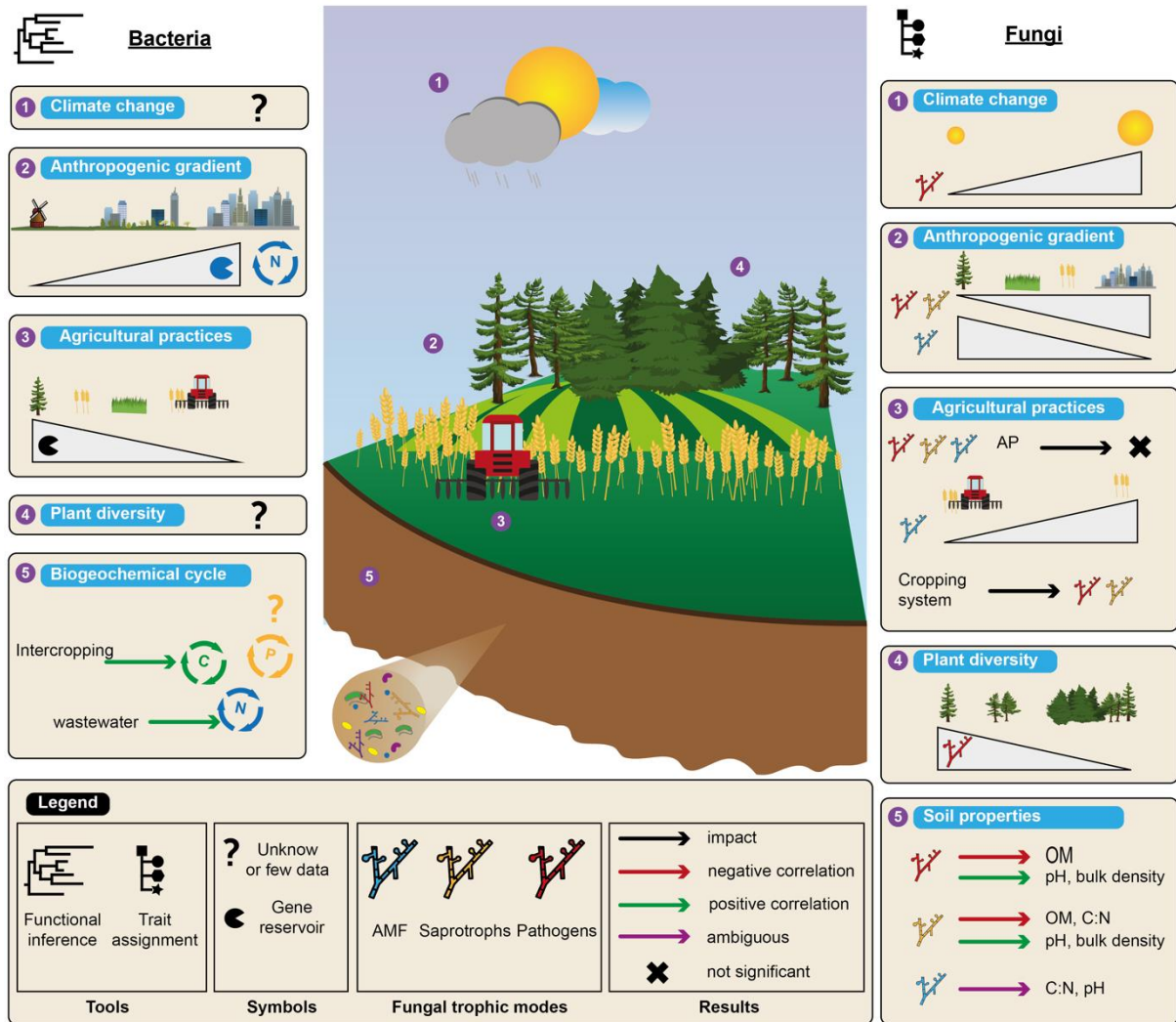


Figure 9: **Summary diagram of the most relevant microbial soil functions results based on functional inference and ecological trait assignment.**

The figure is made up of two parts: studies on bacterial communities based on functional inference on the left, and studies on fungal communities based on ecological trait assignment on the right. For all studies (climate change, anthropogenic gradient, agricultural practices, plant diversity or the biogeochemical cycle), if an impact or a correlation was found on the gene reservoir or on microbial communities with a particular ecological trait, a colored arrow indicates the effect and a cross indicates no significant effect. A triangle indicates either a decrease or an increase of the gene reservoir or microbial communities with a particular trait. References are listed in Additional file 1.

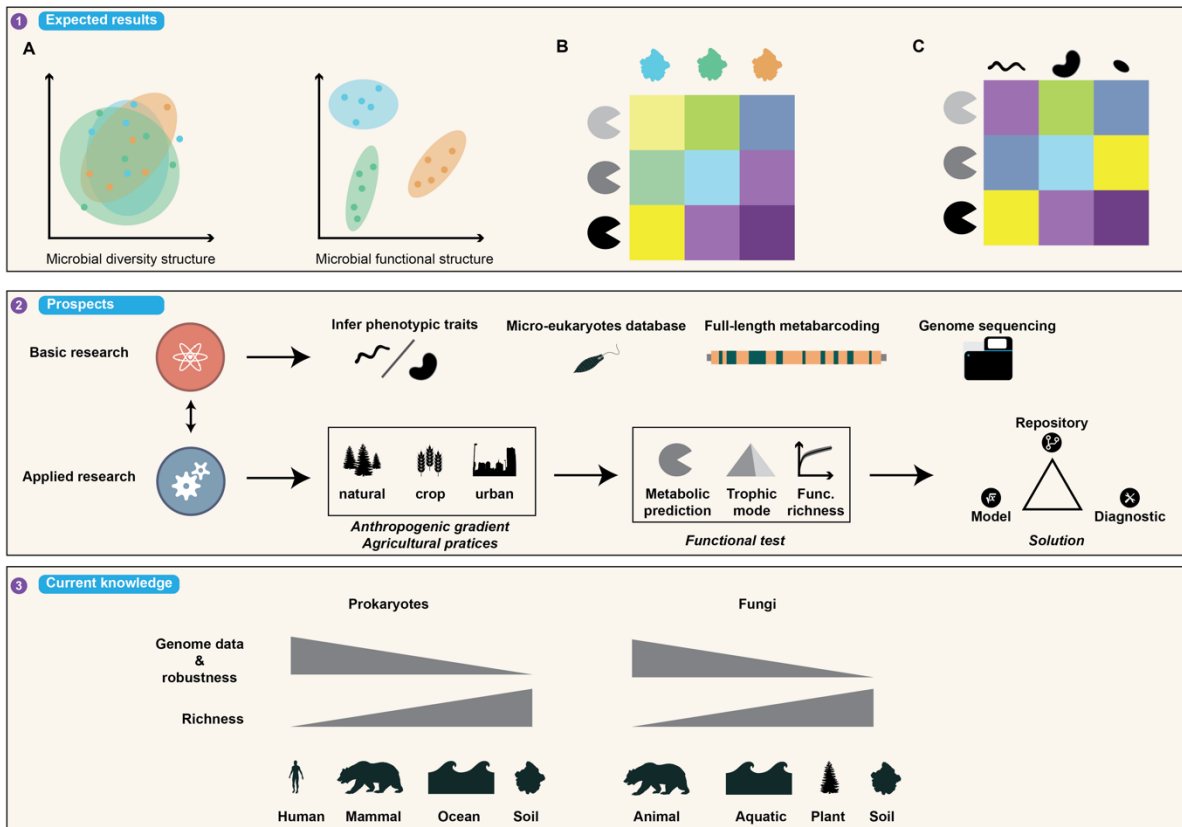
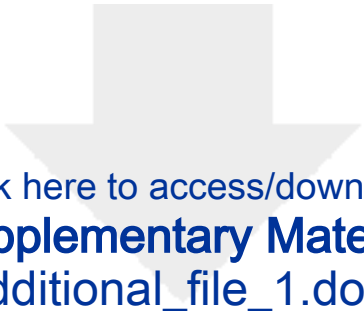



Figure 10: **Summary diagram of the expected results (first box), the functional prediction prospects (second box) and the limits of the microbial genomic data available for different habitats (third box).** The first box illustrates a comparative example of data results of community structures and functional structures through a PCA (A). This example illustrates the case when the functional community structure differentiates experimental conditions better than it differentiates the microbial community structure. Illustrative heat maps showing the relative abundance of genes *per sample* (B) or *per OTU* (C).



Click here to access/download
Supplementary Material
Additional_file_1.docx





Agroécologie

Dijon
Unité de Recherche

INRAE

UBFC



UMR 1347 - AGROECOLOGIE

RANJARD Lionel (BIOCOM Team)

Phone. : 33 (0) 03 80 69 30 88

Mail : lionel.ranjard@inrae.fr

Dijon, 04 October 2021

Dear Editor,

We would like to submit the paper entitled “Inferring microbiota functions from taxonomic genes: a review” by Christophe Djemiel, Pierre-Alain Maron, Sébastien Terrat, Samuel Dequiedt, Aurélien Cottin, and Lionel Ranjard for publication in *GigaScience*.

In this paper we review the tools and methods dedicated to functional inference and ecological trait assignment to explore the functional potential of microbial ecosystems. These approaches have been developed after the recent surge of big data in microbial ecology studies thanks to high-throughput sequencing. Some tools have become quite popular thanks to the popularization of metabarcoding, but studies allowing an overview, an evaluation and a ranking of the advantages, specificities and drawbacks of these tools are still blatantly lacking in current literature, both for bacterial and fungal communities.

Overall, our scientific and technical benchmarks show that functional inference and trait assignment are powerful methods for describing changes in the functional potential of complex microbial communities metabarcoding approaches. However, making them a robust diagnostic tool in various fields (e.g. soil studies) still remains a challenge.

We believe that this work will help scientists working on microbial communities make the appropriate choices to best take advantage of the high amounts of microbial data made available.

The work presented in this manuscript is original and has not been published or considered for publication by another journal.

We thank you for considering this manuscript for publication in *GigaScience*.

Yours sincerely,

Lionel RANJARD and Christophe DJEMIEL