Reference genome-independent taxonomic profiling of microbiomes with mOTUs3

Hans-Joachim Ruscheweyh[1,*] hansr@ethz.ch

Alessio Milanese[1,2,*] alessio.milanese@biol.ethz.ch

Lucas Paoli[1] lucas.paoli@biol.ethz.ch

Nicolai Karcher[2] nicolai.karcher@embl.de

Quentin Clayssen[1] quentin.clayssen@gmail.com

Marisa Isabell Metzger[2] marisa.metzger@embl.de

Jakob Wirbel[2] jakob.wirbel@embl.de

Peer Bork[2,3,4] bork@embl.de

Daniel R. Mende[5] danielrmende@gmail.com

Georg Zeller[2#] zeller@embl.de

Shinichi Sunagawa[1,#] ssunagawa@ethz.ch

**Contributed equally to this work*

#Corresponding authors

[1]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland

[2]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[3]Max Delbrück Centre for Molecular Medicine, Robert-Rössle-Str. 10, 13092, Berlin, Germany

[4]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074, Würzburg, Germany

[5]Department of Medical Microbiology, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

Abstract

Background: Taxonomic profiling is a fundamental task in microbiome research that aims to detect and quantify the relative abundance of microorganisms in biological samples. Available methods using shotgun metagenomic data generally depend on the availability of sequenced and taxonomically annotated reference genomes. However, the majority of microorganisms have not been cultured yet and lack such reference genomes. Thus, a substantial fraction of microbial community members remains unaccounted for during taxonomic profiling of metagenomes, particularly in samples from underexplored environments. To address this issue, we have developed the mOTU profiler, a tool that enables reference genome-independent species-level profiling of metagenomes. As such, it supports the identification and quantification of both "known" and "unknown" species based on a set of select marker genes.

Results: Here, we present mOTUs3, a command line tool that enables the profiling of metagenomes for >33,000 species-level operational taxonomic units. To achieve this, we leveraged the reconstruction and analysis of >600,000 draft genomes, most of which are metagenome assembled genomes (MAGs), from diverse microbiomes, including soil, freshwater systems, and the gastrointestinal tract of ruminants and other animals, which we found to be greatly underrepresented by reference genomes. Overall, two-thirds of all species-level taxa lacked a reference genome. The cumulative relative abundance of these newly included taxa was low in well-studied microbiomes, such as the human body sites (6-11%). By contrast, they accounted for substantial proportions (ocean, freshwater, soil: 43-63%) or even the vast majority (pig, fish, cattle: 60-80%) of the relative abundance across diverse non-human-associated microbiomes. Using community-developed benchmarks and datasets, we found mOTUs3 to be more accurate than other methods and to be more congruent with 16S rRNA gene-based methods for taxonomic profiling. Furthermore, we demonstrate that mOTUs3 greatly increases the resolution of well-known microbial groups into species-level taxa and helps identify new differentially abundant taxa in comparative metagenomic studies.

Conclusions: We developed mOTUs3 to enable accurate species-level profiling of metagenomes. Compared to other methods, it provides a more comprehensive view of prokaryotic community diversity, in particular for currently underexplored microbiomes. To facilitate comparative analyses by the research community, it is released with >11,000 precomputed profiles for publicly available metagenomes and is freely available at: https://github.com/motu-tool/mOTUs.

Keywords: Metagenomics, microbial community, benchmarking, taxonomic profiling, marker gene, metagenome-assembled genome, single-cell genome, reference genome

63 **Background**

64 Identifying and quantifying the abundance of taxa (i.e., taxonomic profiling) is a critical step in
65 linking the composition of microbial communities to environmental functions and host health-related
66 phenotypes [1,2]. Metagenomic sequencing of DNA directly extracted from an environmental or host-
67 derived sample has enabled researchers to taxonomically profile microbial communities in an
68 unbiased and cultivation-independent manner. The development of tools to generate accurate
69 taxonomic profiles from metagenomic data has therefore become important to our understanding of
70 microbial communities [3]. However, existing tools rely on the availability of informative sequences
71 (such as k-mers or marker genes [4,5]), which are predominantly extracted from taxonomically
72 annotated reference genomes (RefGs).

73 In recent years, high-throughput culturing of microorganisms coupled with RefG sequencing (known
74 as culturomics) [6] has substantially expanded the proportion of microbial taxa with whole genome
75 sequences in data repositories (e.g., NCBI RefSeq) benefitting taxonomic profiling tools. However,
76 there is a strong bias toward microorganisms from well-studied habitats (e.g., human body sites)
77 and/or those that can be readily cultivated using standard laboratory methods. Thus, most microbes on
78 Earth remain uncultivated and lack a representative RefG [7,8], although they can be both globally
79 prevalent [9] and numerically dominant in many environments [10, 11, 12, 13]. As a result, the
80 incorporation of RefGs from newly isolated microbes into taxonomic profiling tools can be slow and
81 disproportional across environments. This poses an additional challenge for accurate taxonomic
82 profiling, given that microorganisms that remain undetected bias the abundance estimates of those
83 that are detected [14,15].

84 To close the gap between the detectable and actual diversity present in microbial community samples,
85 we developed mOTUs [14,16], a software tool that uses universal, protein-coding, single-copy
86 phylogenetic marker gene (MG) sequences to quantify the taxonomic composition of microbial
87 communities from metagenomic sequence data (for further applications, see also Ruscheweyh et al.
88 2021 [17]). As these MGs are present in all organisms, they can be identified not only in RefGs, but
89 also in metagenomic assemblies. Conceptually, mOTUs is based on clustering sets of MGs

representing individual organisms by sequence similarity into species-level units. In the absence of a generalizable species concept for prokaryotes [18,19], we refer to these units as MG-based operational taxonomic units (abbreviated as 'mOTUs').

As an alternative to RefG sequencing, draft genomes are increasingly reconstructed by computational binning of metagenomic assemblies into metagenome-assembled genomes (MAGs [20]) or by sequencing amplified DNA from individual cells, resulting in single cell genomes (SAGs [21]). These cultivation-independent methods have provided genomic access to microbial diversity in previously underexplored environments. Here, in addition to MGs found in RefG and metagenomic data, we now incorporate those found in MAGs and SAGs to more than double the number of taxa represented, adding >20,000 new mOTUs compared to the previous major release [14]. Our evaluations show that mOTUs3 outperforms other methods as assessed using metrics for taxonomic tool benchmarking developed independently from our study [3,22]. Furthermore, we found mOTUs3 to provide an unprecedented view of the species-level diversity within the most dominant heterotrophic bacterial clade in the ocean and to greatly extend the number of detected and differentially abundant species in cross-sectional studies, as exemplified in a comparison between rumen microbiomes of high- and low-level methane-emitting sheep.

Results

Taxonomic profiling of diverse environments with mOTUs3

We developed mOTUs3 to facilitate the metagenomic profiling of 33,570 mOTUs, which is a 4.3-fold increase compared to mOTUs2 (Figure 1a). Among all mOTUs, 35% were represented by a RefG (n=11,915; ref-mOTUs), while an additional 21,655 were derived using MGs from either metagenomic contigs (n=2,297; meta-mOTUs) or extended sources, such as MAGs (*de novo*-assembled or imported) and a smaller number of SAGs and isolate genomes (n=19,358; ext-mOTUs), to substantially extend the database coverage for reference genome-independent taxonomic profiling of diverse environments. MGs not assigned to any mOTU were additionally added to the database and merged into a single 'unassigned' group to improve the quantification accuracy of taxonomic profiles, as previously demonstrated [14].

118

119 The newly established database allowed us to determine and systematically compare the fraction of
120 taxa currently not represented by RefGs in various environments. These environments include
121 extensively studied human-associated ones, for which metagenomic studies are complemented by
122 several culturomics efforts (e.g., Lagier et al. [23]). Furthermore, we included data from >20
123 environmental and animal-associated microbiomes (Supplementary Tables 1 and 2) that have been
124 primarily studied by metagenomic approaches. Overall, we found that more than half (11,882) of all
125 meta/ext-mOTUs (i.e., mOTUs not represented by any RefG) could not be assigned to any known
126 family (Supplementary Table 3; Methods), illustrating the taxonomic novelty covered by mOTUs3.
127 The distribution of the newly included data into ref/meta/ext-mOTUs was highly variable across the
128 different environments (Supplementary Figure 1). As expected, 97% of the ~400,000 MAGs from
129 human microbiome samples (Supplementary Table 1) had already been represented by 2,360 pre-
130 existing (i.e., ref/meta-)mOTUs (Supplementary Table 4). Notably, the remaining 3% represented
131 2,750 new ext-mOTUs, showing that novel species can still be uncovered by studying
132 underrepresented populations, dietary habits and/or disease states [24,25]. By contrast, we found that
133 only ~25% of the 6,479 MAGs from mouse gut metagenomes (Supplementary Table 1) corresponded
134 to pre-existing mOTUs (n=72), despite ongoing cultivation efforts [6]; the remaining 75% were
135 grouped into 587 ext-mOTUs (Supplementary Table 4). However, the vast majority of ext-mOTUs
136 (n=16,021) resulted from the inclusion of other animal-associated (e.g., ruminants, fish, chicken, pig,
137 bee, dog, cat) and environmental (e.g., soil, freshwater, wastewater, ocean, air) microbiomes
138 (Supplementary Table 1) for which the generation of representative RefGs is lagging.

139 We used mOTUs3 to profile 10,541 available shotgun metagenomic data sets across the 23
140 environments covered by its database (Supplementary Table 1). For comparative analyses, we subset
141 the data to 5,756 high-quality samples (Methods; Supplementary Table 5) from 16 environments and
142 found the overall number of detected mOTUs to range from 247 (honey bee) to >6,000 (ocean,
143 wastewater and cattle microbiomes). To illustrate the proportion of quantifying taxa currently not
144 represented by RefGs (Figure 1b), we summarized the cumulative relative abundances of unassigned
145 taxa and the different types of mOTUs (ref-mOTUs, meta-mOTUs, ext-mOTUs). The fraction of

146 unassigned taxa was highest for soil samples (33%; s.d. 8%), which reflects the high microbial
147 diversity in soil as well as challenges in reconstructing genomes from this environment [26]. By
148 contrast, more than 87% (s.d. 0.7%) of the relative abundance was represented by ref-mOTUs in
149 human skin samples mainly due to the dominance of few taxa with cultivated representatives [27].
150 Similarly, the fraction of relative abundance assigned to ext-mOTUs varied considerably between
151 environments: on average, only ~6% of the bacterial abundance in human-associated samples was
152 assigned to newly added taxa, while this fraction was as high as ~80% in cattle rumen microbiomes.
153
154 *Comparison with other taxonomic profilers*
155 As in other fields of bioinformatics, there is broad consensus that the performance of analysis tools
156 needs to be carefully evaluated. However, best practices (e.g., balancing precision and recall,
157 selecting criteria for 'best' performance) are often debated [28,29], and in microbiome research, an
158 agreement on some fundamental concepts (e.g., sequence vs. taxonomic abundance, representation of
159 unknown taxa in ground truth data) is still lacking [30,31]. In an attempt to address some of these
160 issues in a community-driven effort, modeled after successful examples in other fields [32,33], the
161 Critical Assessment of Metagenome Interpretation (CAMI) has provided curated ground truth datasets
162 along with a tool (OPAL) to reproducibly evaluate metagenomic analysis tools [3,22].

163 Using the latest CAMI datasets with disclosed results [34], we compared mOTUs3 to its prior major
164 release version (mOTUs2) [14] and other selected metagenomic profiling tools (MetaPhlAn3 [5] and
165 Bracken [4,35], Methods) representing conceptually different, well-performing approaches to
166 taxonomic profiling [30]. Using the OPAL tool for scoring and evaluation, we first evaluated
167 presence/absence (F_1-score) and relative abundance predictions (L1 norm error) at the species level.
168 For the different datasets, which represented samples from five human body sites and the mouse gut
169 microbiome, mOTUs3 and MetaPhlAn3 performed generally better than Bracken and mOTUs2
170 (Figure 2a/b). At higher taxonomic ranks, mOTUs3 had similar or higher scores than the other tools.
171 For some datasets, taxonomic ranks and tools, there was little to no room for improvements of the F_1-
172 score or L1 norm error. This may be due to the simulated datasets being mainly based on taxa for
173 which RefGs are available and/or result from incongruencies of taxonomic annotations used by the

174 different profilers compared to the ground truth. In addition to the L1 norm error, OPAL computes
175 additional metrics for profiling quality (completeness, purity, weighted UniFrac error) and
176 summarizes them across taxonomic ranks into a composite score. Based on this evaluation criterion,
177 mOTUs3 outperformed the other tools (Figure 2c), as well as additional tools assessed in the CAMI
178 challenge (Methods; Supplementary Figure 2).

179 In the absence of independent ground truth data sets to benchmark taxonomic profiling tools for less
180 well-studied environments, we correlated taxonomic profiles obtained by mOTUs3 and other tools to
181 those obtained by analyzing 16S rRNA gene (16S) fragments. This approach leverages both the
182 availability of comprehensive 16S databases for taxonomic classification [36] and the possibility of
183 estimating taxonomic abundances based on 16S-based data from metagenomes [37]. Briefly, we
184 extracted 16S fragments from the same datasets we used for metagenomic profiling and generated
185 relative abundance profiles for them (Methods). To ensure comparability between 16S and
186 metagenomic profiles, the analysis was performed at the genus and higher taxonomic ranks (for
187 discussion, see Salazar et al. [37]). We found that mOTUs3 had consistently higher correlations with
188 16S profiles than the other tools across all environments, except for the human gut for which
189 MetaPhlAn3 showed correlation coefficients similar to those of mOTUs3 (Figure 3).
190
191 *Resolving the diversity of Pelagibacterales with mOTUs3*
192 In addition to the broader taxonomic coverage by mOTUs3 across environments, we sought to
193 investigate the capability of mOTUs3 to resolve microbial clades into more fine-grained taxonomic
194 units. To this end, we focused on Pelagibacterales (also referred to as the SAR11 clade), which is the
195 most abundant heterotrophic bacterial group in the global oceans [38]. Members of the
196 Pelagibacterales have previously been shown to display high genomic variability while maintaining
197 highly conserved 16S sequences [39]. This prompted us to evaluate the species-level resolution of
198 mOTUs3 and to compare the diversity represented by mOTUs to the diversity represented by
199 operational taxonomic units (OTUs) defined by 16S sequence similarity.

200 For this analysis, we selected from all mOTUs annotated as Pelagibacterales (n=1,029; 2,063
201 genomes) those that were represented by genomes with complete 16S sequences (n=602; 1,105
202 genomes). The number of mOTUs was comparable to the number resulting from a 95% average
203 nucleotide identity (ANI)-based clustering of the 1,105 genome sequences into species-level groups
204 (n=700; Figure 4a), which is common practice in the field of microbial phylogenomics [7,40].
205 Moreover, we found sequence identities of mOTUs-representing MGs to linearly correlate with those
206 of whole genomes across the whole range of observed values (r^2=0.71; Figure 4b). By contrast, 16S
207 sequence-based OTUs using a 97% or 99% sequence similarity cutoff resulted in a 31.7-fold (n=19)
208 or 5.8-fold (n=104) lower number of taxonomic units, respectively, compared to mOTUs (Figure 4a).
209 This discrepancy is also reflected by a weaker correlation (r^2=0.45; Figure 4b) of identities between
210 16S sequences and corresponding whole genome sequences. The minimum 16S identities were ca.
211 87% and started saturating at approximately 97% at which point genome identities were still as low as
212 ~70-80% (Figure 4b). Similar findings were reported previously albeit on smaller datasets [39].
213 Finally, comparing the grouping of genomes by mOTUs and ANI into species-level clusters, we
214 found almost perfect congruence (Figure 4c, Methods).
215
216 *Differential abundance of novel archaea in low/high methane-emitting sheep rumen metagenomes*
217 High-resolution taxonomic profiling of metagenomes from underexplored environments can be
218 achieved by custom-made marker gene or genome databases selected for the microbial community
219 under study [12,41]. However, this approach is often labor- and resource-intensive and requires
220 specialized expertise, and its results cannot easily be compared across studies and communities. To
221 demonstrate the utility of mOTUs3 to address these challenges, we reanalyzed rumen metagenomes
222 from high- and low-methane emitting (HME and LME) sheep [41]. Importantly, these data were not
223 used for the database construction of mOTUs3.
224 Based on mOTUs3 taxonomic profiles, we identified 131 microbial species that differed significantly
225 in abundance between HME and LME samples and showed an at least tenfold increase or decrease in
226 relative abundance (corresponding to a generalized fold change of >= 1 [42]). Among these
227 differentially abundant species, 92% were represented by ext-mOTUs. These were therefore not

228 expected to be detectable by reference-based profilers. To test this, we applied the same workflow
229 using MetaPhlAn3 and Bracken (see Methods), which yielded only 10 and 30 differentially abundant
230 species for the respective tools (Figure 5a).

231 Given the metabolic importance of methanogenic archaea in ruminants as well as previous evidence
232 of uncharted archaeal diversity in the sheep rumen [12], we further investigated the species-level
233 diversity of known and unknown archaeal species. To this end, we reconstructed a phylogenetic tree
234 of the archaeal mOTUs detected in the sheep rumen metagenomes (n=15) and contextualized them
235 with reference genomes from members of the genera *Methanobrevibacter* and *Methanosphaera*
236 (Figure 5b). This analysis revealed that all six differentially abundant archaea in the sheep rumen
237 corresponded to ext-mOTUs. Two of them, which were significantly more abundant in high-methane
238 emitters, were most closely related to *Methanobrevibacter gottschalkii*, which itself was not detected.
239 Notably, the MG sequence similarity between these ext-mOTUs and *M. gottschalkii* was <85%
240 (Figure 5b), which is well below the species-level cutoff of 96.5% used by mOTUs [16] and therefore
241 suggests that these ext-mOTUs represent novel *Methanobrevibacter* spp.

242

243 **Discussion**

244 With mOTUs3, we have developed a taxonomic profiler that combines state-of-the-art accuracy, as
245 demonstrated in competitive benchmarks based on simulated datasets, with an innovative database
246 construction approach to detect and quantify underrepresented microbes from diverse environments at
247 high (i.e., species-level) taxonomic resolution. The ability to incorporate MG sequences from any
248 MAG and SAG to generate mOTUs *de novo* and independently from the availability of RefGs and/or
249 prior existence of taxonomic annotations (such as NCBI or GTDB species names) will allow users to
250 continuously extend the core database of mOTUs to represent microbial diversity from newly
251 explored microbiomes. Such future extensions could also target eukaryotic microorganisms, as these
252 are an integral part of many microbial communities, but are not well represented in databases of
253 existing taxonomic profiling tools.

254 However, the flexibility in defining operational taxonomic units *de novo* comes with a need for
255 taxonomic annotation, as is also the case for 16S rRNA-based *de novo* clustered OTUs. Despite the
256 calibration of MG sequence identity cutoffs to maximize congruence with the NCBI taxonomy [16],
257 this procedure can lead to conflicts with existing taxonomies. Irrespective of the ongoing debate on
258 whether prokaryotic species should be consistent with genomic similarity-based criteria, delineating
259 species by sequence identity puts mOTUs at a disadvantage in benchmarks, such as CAMI, which
260 rely on rigid matching of taxonomic labels. The high performance of mOTUs [34] despite this
261 disadvantage is likely due to the higher number of quantified taxa and the resulting reduction in
262 compositionality-related biases.
263
264 **Conclusions**
265 The present work introduces mOTUs3 as a reference-genome independent tool that allows for
266 charting the taxonomic landscape of many environments at species-level resolution. Its independence
267 from taxonomically annotated reference genomes, makes it generally applicable also beyond well-
268 studied environments to quantify and reveal yet uncharacterized microbial species of potential
269 biological relevance. To support the research community, mOTUs3 is documented and available as
270 open source software at https://github.com/motu-tool/mOTUs.
271
272 **Methods**
273 *Collection and processing of data to compile the mOTUs3 database*
274 To extend the taxonomic coverage of the mOTUs3 database, 4,531 publicly available metagenomic
275 datasets from 23 environments (Supplementary Table 1) were processed to generate 150,880 MAGs
276 as previously described [43]. Briefly, BBMap (v.38.71) was used to quality control sequencing reads
277 from all samples by removing adapters from the reads, removing reads that mapped to quality control
278 sequences (PhiX genome) and discarding low-quality reads (*trimq=14, maq=20, maxns=1* and
279 *minlength=45*). For metagenomic data of human origin, human genome-derived reads were removed
280 using the masked human reference genome provided by BBMap. Quality-controlled reads were
281 merged using bbmerge.sh with a minimum overlap of 16 bases, resulting in merged, unmerged paired

282 and single reads. The reads were assembled into scaffolded contigs (hereafter scaffolds) using the
283 SPAdes assembler (v3.14 or v3.12) [44] in metagenomic mode. Genes were predicted on length-
284 filtered (≥ 500 bp) scaffolded contigs (hereafter scaffolds) using Prodigal (v2.6.3) [45]. Universal
285 single-copy phylogenetic marker genes (MGs) were extracted using fetchMGs (v1.2; *-m extraction*)
286 [16].

287 Scaffolds were length-filtered (≥ 1000 bp) and within each study, quality-controlled reads from each
288 sample were mapped against the scaffolds of each sample. Mapping was performed using BWA
289 (v0.7.17-r1188; *-a*) [46]. Alignments were filtered to be at least 45 bp in length, with an identity of ≥
290 97% and a coverage of ≥ 80% of the read sequence. The resulting BAM files were processed using
291 the *jgi_summarize_bam_contig_depths* script of MetaBAT2 (v2.12.1) [20] to compute within- and
292 between-sample coverages for each scaffold. The scaffolds were binned by running MetaBAT2 on all
293 samples individually (*--minContig 2000* and *--maxEdges 500* for increased sensitivity). These
294 metagenomic bins were complemented with 454,773 external draft genomes (~96% MAGs; ~4%
295 isolate and single-cell genomes) from previous work (Supplementary Table 1). Complete genes in
296 external draft genomes and metagenomic bins were predicted using Prodigal (v2.6.3; *-c -m -g 11 -p*
297 *single*) and MGs were extracted using fetchMGs (v1.2) *(-m extraction -v -i*).

298 Metagenomic bins and draft genomes were annotated with Anvio (v5.5.0) [47], quality controlled
299 using the CheckM (v1.0.13) [48] lineage workflow (completeness ≥ 50% and contamination < 10%)
300 and filtered for genomes containing at least six out of the 10 MGs used by mOTUs [16] to produce
301 the dataset of MGs from a total of 499,512 *de novo*-generated MAGs (i.e., quality-controlled
302 metagenomic bins) and external draft genomes used for the construction of the mOTUs3 database.
303
304 *Construction of the mOTUs3 database*
305 MGs from 499,512 genomes were mapped against the latest mOTUs database (v2.5.1), which was an
306 update of version 2.0 to account for a more recent release of the progenomes2 database [49] (Figure
307 1a) using vsearch [50] (v2.14.1; *--usearch_global --strand both --id 0.8 --maxaccepts 10000 --*
308 *maxrejects 10000*). MGs from a total of 283,250 and 136,429 genomes were assigned to existing ref-

309 mOTUs and meta-mOTUs, respectively. These genomes were removed since they were already
310 represented. The remaining 79,833 genomes resulted in an extension of the mOTUs database by
311 19,358 new mOTUs (ext-mOTUs). For consistency with the taxonomic annotation of ref-mOTUs,
312 ext-mOTUs were annotated using the STAG classifier (https://github.com/zellerlab/stag, version 0.7;
313 default parameters) trained on genomes in the proGenomes2 database [49] (NCBI taxonomy, version:
314 8 January 2019). MGs identified on scaffolds that were not binned into MAGs were used to update
315 the 'unassigned' mOTU, which contain unbinned MGs that are used to estimate the quantity of
316 unknown species, by aligning these MGs against the extended database using vsearch (v2.14.1;
317 *usearch_global --maxaccepts 1000 --maxrejects 1000 --strand both*). MGs that did not align within
318 MG-specific cutoffs [51] were clustered using vsearch (v2.14.1; *--cluster_fast*) using MG-specific
319 cutoffs and the representative sequence was added to the unassigned mOTU.
320
321 *Computation of mOTUs3 profiles for comparative analyses*
322 A total of 11,164 metagenomic and metatranscriptomic samples (Supplementary Table 1,
323 Supplementary Table 2) were quality controlled and merged as described above and profiled with
324 mOTUs3 using default parameters and the *-c* option to build a community resource of taxonomic
325 profiles. For comparative analyses across environments, 5,756 of these samples were used after
326 removing all (n=623) metatranscriptomic samples, metagenomic samples from environments with too
327 few samples (termite, panda, aerosols and bioreactor) or from studies comprising samples from
328 different environments and samples with less than 5,000 mapped inserts. To calculate the total
329 number of detected mOTUs for a given environment, we counted the number of mOTUs with a
330 prevalence greater than 0.1% (Supplementary Table 5). To compare the median number of detected
331 mOTUs across different environments, we downsampled the insert counts to 5,000 using the *rrarefy*
332 function of the vegan package [52].
333
334 *Comparison of taxonomic profilers using the CAMI framework*
335 The performance of mOTUs3 was evaluated and compared to mOTUs2 and other taxonomic profilers
336 by analyzing 113 publicly available samples (49 human-associated, 63 mouse gut metagenomes)

337 provided by the second CAMI challenge (https://cami-challenge.org/participate). The samples were
338 profiled with mOTUs3 (v3.0.1; *-C precision*), mOTUs2 (v2.1.1; *-C precision*), MetaPhlAn3 (v3.0.7; -
339 *-CAMI_format_output --index mpa_v30_CHOCOPhlAn_201901*) [5] and Kraken/Bracken (v2.1.2; --
340 *db=k2_standard_20201202 --paired* / v2.6.1; *--db=k2_standard_20201202 -r 100 -l S|G|F|O|C|P|D*)
341 [4,35]. Kraken/Bracken reports were further translated into the CAMI format ed files using the
342 *tocami.py* script provided at https://github.com/hzi-bifo/cami2_pipelines. For comparative analyses,
343 the OPAL framework (v1.0.9) [22] was used with default parameters providing the gold standard with
344 the parameter *--gold_standard_file,* the names of the tools with *--labels*, the description with *-d*, the
345 output with *--output_dir* and the taxonomic profiles files as positional arguments.
346
347 *Comparison of metagenomic profiles with 16S rRNA gene-based profiles*
348 The 16S rRNA-based taxonomic profiler mTAGs [37] (v1.0.1; *-ma 1000 -mr 1000*) was used to
349 generate relative abundance profiles for metagenomic samples (Supplementary Table 1). The output
350 of mTAGs was mapped to the NCBI taxonomy to facilitate comparative analysis. The same samples
351 were profiled with MetaPhlAn3 (v3.0.7; *--index mpa_v30_CHOCOPhlAn_201901*) and
352 Kraken/Bracken (v2.1.2; *--db=k2_standard_20201202 --paired* / v2.6.1; --
353 *db=k2_standard_20201202 -r 100 -l S*). Samples with small read/insert coverages (mTAGs<10,000,
354 mOTUs<1,000, Kraken/Bracken<10,000, no filtering was done on MetaPhlAn3 as profiles contain
355 relative abundances) were removed, leaving 6,119 samples for comparative analysis. Spearman
356 correlations were calculated for each taxonomic rank based on concatenated relative abundances
357 between mTAGs and the metagenomic profiling tools.
358
359 *Comparison of Pelagibacterales genome clusters with marker gene and 16S rRNA gene sequences*
360 Out of 2,063 genomes belonging to 1,029 mOTUs annotated as Pelagibacterales, 1,105 genomes
361 (from 602 mOTUs) that contained a complete copy of the 16S rRNA gene were selected. These
362 genomes were also clustered based on average nucleotide identity using dRep [53] (v2.5.4; *-comp 0 -*
363 *con 1000 -sa 0.95 -nc 0.2*) using a 95% cutoff as part of the OMD [43]. In addition, these genomes
364 were clustered based on their 16S rRNA gene identity (99% and 97%) using vsearch [50] (v2.14.1; --

365 *cluster_smallmem --id 0.97 / 0.99*). The consistency between the different clustering approaches was
366 evaluated using the V-measure, which combines both the homogeneity and completeness metrics
367 [54].

368 To correlate distances of the 1,105 genomes between the different clustering techniques we performed
369 exhaustive distance calculations at the whole-genome level, the 10 MGs used by mOTUs and the 16S
370 rRNA gene. Whole genome distances were computed using MASH [55] as implemented in dRep
371 (v2.5.4). MG- and 16S rRNA gene-based distances were computed using vsearch (v2.14.1; --
372 *allpairs_global --id 0.0*) and MG distances were averaged across the 10 genes prior to computing
373 correlations.

374

375 *Differential abundance of mOTUs between low/high methane-emitting sheep*

376 Samples from sheep rumen metagenomes (n=16) [41] were profiled with mOTUs3 (v3.0.1; *-c*),
377 MetaPhlAn3 (v3.0.7; *--index mpa_v30_CHOCOPhlAn_201901*) and Kraken/Bracken (v2.1.2; --
378 *db=k2_standard_20201202 --paired* / v2.6.1; *--db=k2_standard_20201202 -r 100 -l S*). To test for
379 differentially abundant species between low methane emitters (LMEs) and high methane emitters
380 (HMEs), the respective profiles were analyzed using SIAMCAT default workflows [42]. This
381 workflow includes filtering of species/mOTUs with a relative abundance of >0.1% in at least one
382 sample [42]. Wilcoxon test results were corrected for multiple testing using the Benjamini–Hochberg
383 method [56] at 5% FDR. The reported effect size measure is the generalized fold change (gFC),
384 calculated as the log10 of the geometric mean of quantile differences between groups as defined in
385 SIAMCAT [42].

386 A phylogeny was constructed for all archaeal mOTUs belonging to the *Methanobrevibacter* and
387 *Methanosphaera* genera or the *Thermoplasmata* class that passed the relative abundance filtering (14
388 ext-mOTUs, 1 ref-mOTU) together with ref-mOTUs from *Methanobrevibacter* and *Methanosphaera*
389 (n=15) and a randomly selected Thermoplasmata ref-mOTU as an outgroup. Representative genomes
390 from these 31 mOTUs were selected either by picking the centroid genome (for ext-mOTUs) or the
391 reference genome (for ref-mOTUs). Marker genes were individually aligned (*mafft* [57], v7.458), the

alignments were concatenated and a maximum-likelihood phylogeny was calculated using RAxML [58] (v8.2.12; *raxmlHPC -p 12345 -m PROTGAMMAAUTO*). The distance between the 14 ext-mOTUs and their closest ref-mOTU was calculated based on averaged marker gene distances across the 10 genes (v2.14.1; *vsearch --allpairs_global --id 0.0*).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The mOTUs3 software is documented and publicly available as open source software (GPL 3) at https://github.com/motu-tool/mOTUs. The updated mOTUs3 database can be found at Zenodo (https://doi.org/10.5281/zenodo.5140350) and contains all MGs used in this study and the public profiles generated with mOTUs3. A complete list with all sequencing samples used for building the database and/or for profiling can be found in Supplemental Tables 1 and 2.

Competing interests

none declared

Funding

We acknowledge funding from the Swiss National Foundation through project grants 205321_184955, the NCCR Microbiomes (51NF40_180575) and the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (#2018-521) to SS, and the European Molecular Biology Laboratory and the German Federal Ministry of Education and Research (BMBF, the de.NBI network, grant no. 031A537B and grant no. 031L0181A) to GZ.

Authors' contributions

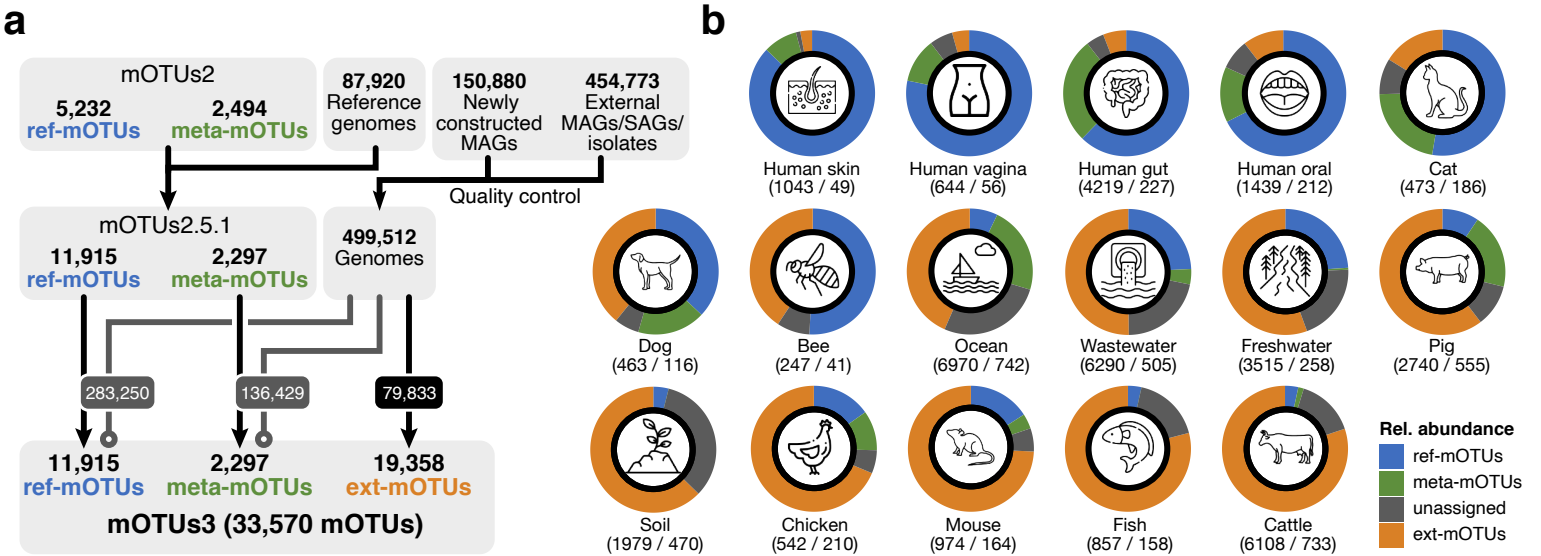
GZ and SS conceived and supervised the work. HJR and AM developed code, generated the database with support from DRM, and performed the benchmark analysis. LP and NK performed the taxonomic diversity analysis of the SAR11 clade and the comparative metagenomic analysis, respectively. QC supported the collection and processing of data. MIM and JW contributed to the taxonomic annotation of mOTUs. HJR, AM, LP, NK, PB, DRM, GZ and SS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

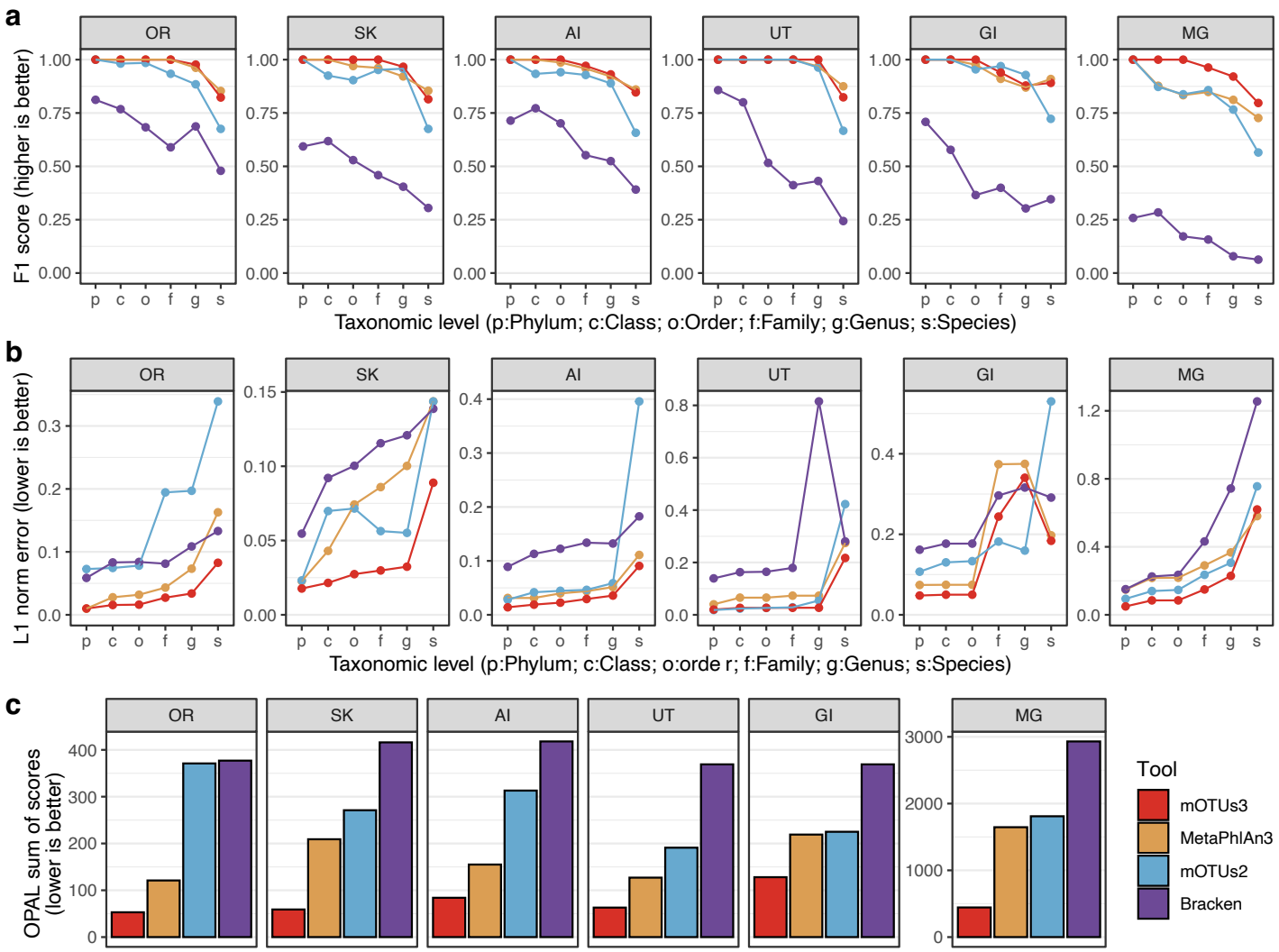
We would like to thank the ETH IT Services and HPC facilities for granting access to the EULER high performance cluster. We also thank Thea Van Rossum for her input on the taxonomic annotation of mOTUs and the users of mOTUs for their feedback and continuous support.

Authors' information

Hans-Joachim Ruscheweyh and Alessio Milanese contributed equally to this work.


a
mOTUs2
5,232
ref-mOTUs
2,494
meta-mOTUs
87,920
Reference genomes
150,880
Newly constructed MAGs
454,773
External MAGs/SAGs/ isolates
Quality control
mOTUs2.5.1
11,915
ref-mOTUs
2,297
meta-mOTUs
499,512
Genomes
283,250
136,429
79,833
11,915
ref-mOTUs
2,297
meta-mOTUs
19,358
ext-mOTUs
mOTUs3 (33,570 mOTUs)
b
Human skin
(1043 / 49)
Human vagina
(644 / 56)
Human gut
(4219 / 227)
Human oral
(1439 / 212)
Cat
(473 / 186)
Dog
(463 / 116)
Bee
(247 / 41)
Ocean
(6970 / 742)
Wastewater
(6290 / 505)
Freshwater
(3515 / 258)
Pig
(2740 / 555)
Soil
(1979 / 470)
Chicken
(542 / 210)
Mouse
(974 / 164)
Fish
(857 / 158)
Cattle
(6108 / 733)
Rel. abundance
ref-mOTUs
meta-mOTUs
unassigned
ext-mOTUs

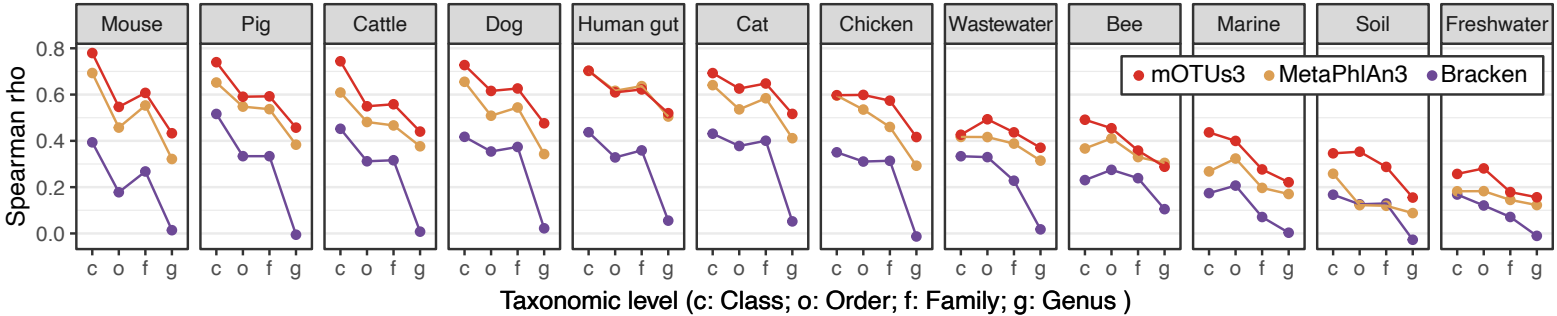
436 **Figure 1. The mOTUs3 database enables species-level profiling across diverse environments.**
437 **(a)** The database of the previous major release of mOTUs (version 2)[14] was updated to version 2.5
438 to account for the current release of the progenomes2 database[49]. Based on version 2.5, the
439 mOTUs3 database was constructed by adding universal, single-copy phylogenetic marker genes
440 (MGs) from 605,653 genomes (metagenome-assembled genomes (MAGs) and a smaller number of
441 isolate and single amplified genomes (SAGs)). This addition resulted in the extension of the database
442 by 19,358 new species-level, MG-based operational taxonomic units (ext-mOTUs). Genomes already
443 represented by ref- and meta-mOTUs in version 2.5 were not added (gray lines). **(b)** Breakdown by
444 the three types of mOTUs shows that mOTUs3 enables the reference genome-independent profiling
445 of a substantial fraction of microbial diversity across different environments. The numbers below the
446 ring charts represent the total number of mOTUs that were detected per environment (left)
447 considering only species with a prevalence of 0.1% and the median number of mOTUs per sample
448 that were detected after downsampling to 5,000 inserts (right).
449


a
OR
SK
AI
UT
GI
MG
F1 score (higher is better)
1.00
0.75
0.50
0.25
0.00
p c o f g s
Taxonomic level (p:Phylum; c:Class; o:Order; f:Family; g:Genus; s:Species)
b
L1 norm error (lower is better)
Taxonomic level (p:Phylum; c:Class; o:orde r; f:Family; g:Genus; s:Species)
c
OPAL sum of scores (lower is better)
Tool
mOTUs3
MetaPhlAn3
mOTUs2
Bracken

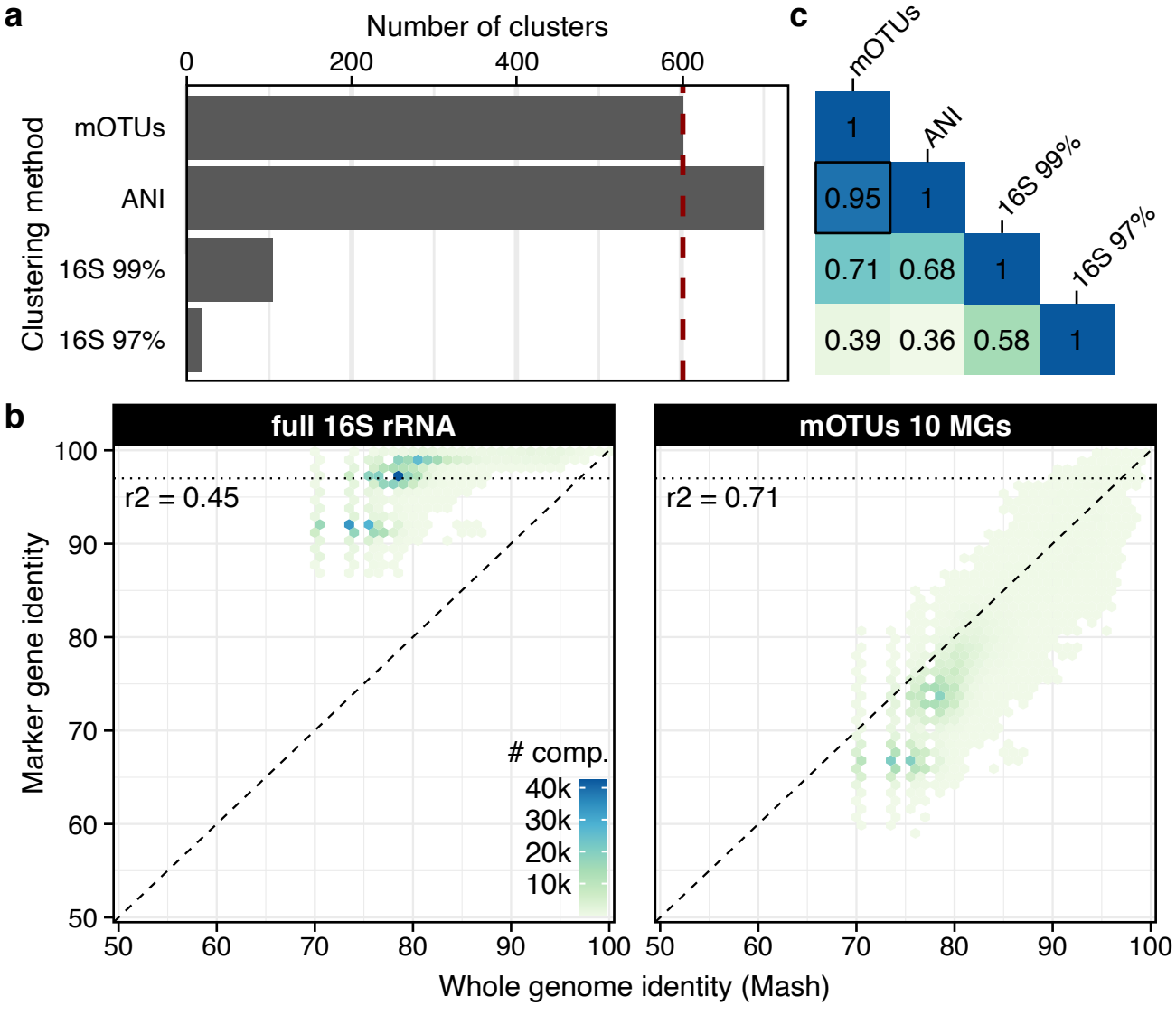
450 **Figure 2. Comparison of mOTUs to other taxonomic profilers.**

451 The performance of mOTUs3 was compared to other taxonomic profiling tools based on the dataset
452 from the second Critical Assessment of Metagenome Interpretation (CAMI) challenge (see Methods).
453 The F1 score **(a)** and L1 norm error **(b)** are shown as reported by the OPAL tool[22] for each
454 taxonomic rank (x-axis). High L1 norm error values at the family and genus levels of GI samples
455 mostly derive from an updated taxonomy of the highly abundant Oscillospiraceae (previously
456 Ruminococcaceae)[59]. **(c)** Each method was ranked across all samples and for each taxonomic rank
457 using four measures (completeness, purity, L1 norm error and weighted UniFrac error), and the
458 OPAL sum of scores was calculated as a sum of these ranks (lower rank indicates better
459 performance). OR: oral cavity, SK: skin, AI: airways, UT: urogenital tract, GI: gastrointestinal tract,
460 MG: mouse gut.

461


Mouse
Pig
Cattle
Dog
Human gut
Cat
Chicken
Wastewater
Bee
Marine
Soil
Freshwater
mOTUs3
MetaPhlAn3
Bracken
Spearman rho
0.8
0.6
0.4
0.2
0.0
c o f g
Taxonomic level (c: Class; o: Order; f: Family; g: Genus)

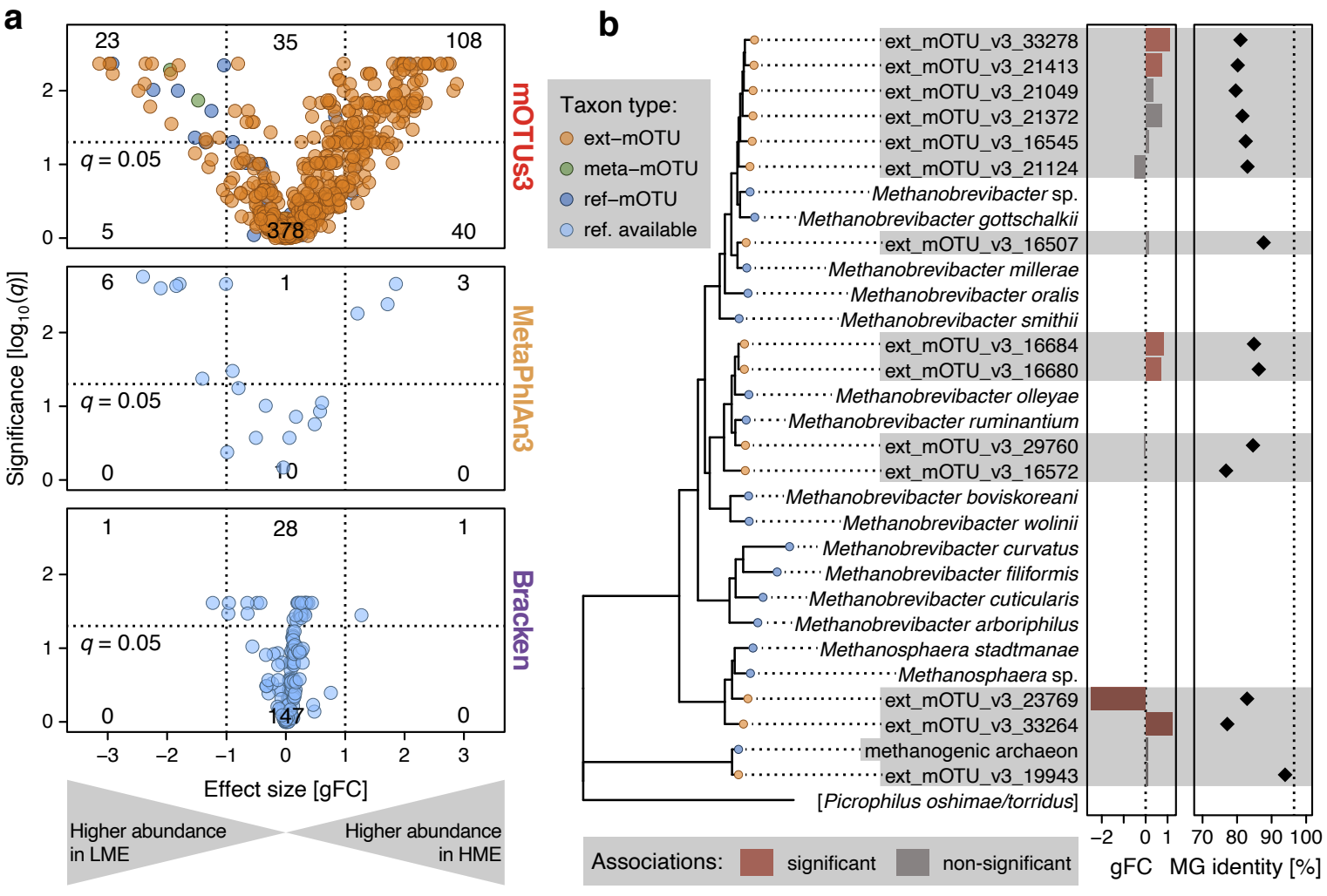
462 **Figure 3. Comparison of metagenomic profiling tools using 16S rRNA-based taxonomic profiles.**
463 Spearman correlations between relative abundances generated by different metagenomic profiling
464 tools and 16S rRNA gene-based profiles from the same samples. The correlations were calculated at
465 different taxonomic ranks (x-axis; c: class, o: order, f: family, g: genus) and showed that mOTUs3
466 generally had the highest values for the different body sites tested, except for human gut samples with
467 similar values for mOTUs3 and MetaPhlAn3.
468


a
Number of clusters
0
200
400
600
Clustering method
mOTUs
ANI
16S 99%
16S 97%
c
mOTUs
ANI
16S 99%
16S 97%
1
0.95
1
0.71
0.68
1
0.39
0.36
0.58
1
b
full 16S rRNA
mOTUs 10 MGs
r2 = 0.45
r2 = 0.71
# comp.
40k
30k
20k
10k
Marker gene identity
Whole genome identity (Mash)
50
60
70
80
90
100

469 **Figure 4. Species-level diversity of Pelagibacterales as resolved by mOTUs3.**

470 (a) The number of taxonomic units within the Pelagibacterales order varies depending on the
471 clustering method used, which was based on using marker gene (MG) sequences (used by mOTUs),
472 Average Nucleotide Identity (ANI) of whole genomes, and full length 16S rRNA gene sequences. (b)
473 mOTUs marker gene distances better capture whole genome distances compared to full length 16S,
474 explaining the patterns observed in (a). In particular, 16S rRNA gene sequence identity saturates
475 while whole genome similarity can be as low as 70-80%. (c) The different clustering approaches vary
476 in their agreement with each other as determined by the V-measure, which captures both the
477 completeness and homogeneity of the clusterings. The highest agreement was found between mOTUs
478 and with whole genome clustering by ANI.

479


a
b
23
35
108
5
378
40
q = 0.05
mOTUs3
6
1
3
0
10
0
MetaPhlAn3
1
28
1
0
147
0
Bracken
Significance [log10(q)]
Effect size [gFC]
Higher abundance in LME
Higher abundance in HME
Taxon type:
ext–mOTU
meta–mOTU
ref–mOTU
ref. available
ext_mOTU_v3_33278
ext_mOTU_v3_21413
ext_mOTU_v3_21049
ext_mOTU_v3_21372
ext_mOTU_v3_16545
ext_mOTU_v3_21124
Methanobrevibacter sp.
Methanobrevibacter gottschalkii
ext_mOTU_v3_16507
Methanobrevibacter millerae
Methanobrevibacter oralis
Methanobrevibacter smithii
ext_mOTU_v3_16684
ext_mOTU_v3_16680
Methanobrevibacter olleyae
Methanobrevibacter ruminantium
ext_mOTU_v3_29760
ext_mOTU_v3_16572
Methanobrevibacter boviskoreani
Methanobrevibacter wolinii
Methanobrevibacter curvatus
Methanobrevibacter filiformis
Methanobrevibacter cuticularis
Methanobrevibacter arboriphilus
Methanosphaera stadtmanae
Methanosphaera sp.
ext_mOTU_v3_23769
ext_mOTU_v3_33264
methanogenic archaeon
ext_mOTU_v3_19943
[Picrophilus oshimae/torridus]
Associations:
significant
non-significant
gFC
MG identity [%]

480 **Figure 5. Detection of differentially abundant taxa in low/high-level methane-emiting sheep**
481 **rumen microbiomes.**
482 **(a)** A comparison between metagenomic profilers shows that mOTUs3 detected 131 differentially
483 abundant species (*q*-value <0.05 and an absolute generalized fold change > 1; indicated by dotted
484 lines) between low- and high-level methane-emitting sheep, while MetaPhlAn3 and Bracken detected
485 nine and two species, respectively. Most of the species detected by mOTUs were represented by ext-
486 mOTUs only, demonstrating the added value of reference genome-independent profiling enabled by
487 mOTUs3. **(b)** Archaeal mOTUs present in the sheep rumen microbiome (highlighted in gray) were
488 phylogenetically contextualized with *Methanobrevibacter* spp. and *Methanosphaera* spp. represented
489 by ref-mOTUs. All differentially abundant ext-mOTUs (middle panel) correspond to distinct yet
490 undescribed *Methanobrevibacter* spp. as supported by MG sequence identities (right panel) to the
491 closest known species being below the species-level cutoff of 96.5% (dotted vertical line).
492
493

494 **References**

495 1. Fuhrman JA. Microbial community structure and its functional implications. Nature. Nature
496 Publishing Group; 2009;459:193–9.
497 2. Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev
498 Microbiol. 2016;14:508–22.
499 3. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of
500 Metagenome Interpretation—a benchmark of metagenomics software. Nat Methods. Nature
501 Publishing Group; 2017;14:1063–71.
502 4. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact
503 alignments. Genome Biol. BioMed Central; 2014;15:1–12.
504 5. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating
505 taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.
506 Elife [Internet]. 2021;10. Available from: http://dx.doi.org/10.7554/eLife.65088
507 6. Lagkouvardos I, Pukall R, Abt B, Foesel BU, Meier-Kolthoff JP, Kumar N, et al. The Mouse
508 Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and
509 functional potential of the gut microbiota. Nat Microbiol. 2016;1:16131.
510 7. Konstantinidis KT, Rosselló-Móra R. Classifying the uncultivated microbial majority: A place for
511 metagenomic data in the Candidatus proposal. Syst Appl Microbiol. 2015;38:223–30.
512 8. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and
513 dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.
514 9. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the
515 tree of life. Nat Microbiol. 2016;1:16048.
516 10. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene
517 Expression Changes and Community Turnover Differentially Shape the Global Ocean
518 Metatranscriptome. Cell. 2019;179:1068–83.e21.
519 11. Lesker TR, Durairaj AC, Gálvez EJC, Lagkouvardos I, Baines JF, Clavel T, et al. An Integrated
520 Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. Cell Rep.
521 2020;30:2909–22.e6.

522 12. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941
523 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat
524 Biotechnol. Nature Publishing Group; 2019;37:953–61.
525 13. Wilhelm RC, Cardenas E, Leung H, Maas K, Hartmann M, Hahn A, et al. A metagenomic survey
526 of forest soil microbial communities more than a decade after timber harvesting. Sci Data.
527 2017;4:170092.
528 14. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, et al. Microbial
529 abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019;10:1014.
530 15. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are
531 Compositional: And This Is Not Optional. Front Microbiol. 2017;8:2224.
532 16. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al.
533 Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods.
534 2013;10:1196–9.
535 17. Ruscheweyh H-J, Milanese A, Paoli L, Sintsova A, Mende DR, Zeller G, et al. mOTUs: Profiling
536 Taxonomic Composition, Transcriptional Activity and Strain Populations of Microbial Communities.
537 Curr Protoc. 2021;1:e218.
538 18. Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev.
539 2001;25:39–67.
540 19. Staley JT. The bacterial species dilemma and the genomic-phylogenetic species concept. Philos
541 Trans R Soc Lond B Biol Sci. 2006;361:1899–909.
542 20. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning
543 algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ.
544 2019;7:e7359.
545 21. Woyke T, Doud DFR, Schulz F. The trajectory of microbial single-cell sequencing. Nat Methods.
546 2017;14:1045–54.
547 22. Meyer F, Bremges A, Belmann P, Janssen S, McHardy AC, Koslicki D. Assessing taxonomic
548 metagenome profilers with OPAL. Genome Biol. 2019;20:51.

549 23. Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously
550 uncultured members of the human gut microbiota by culturomics. Nat Microbiol. 2016;1:16203.
551 24. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal
552 metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med.
553 2019;25:679–89.
554 25. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored
555 Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning
556 Age, Geography, and Lifestyle. Cell. 2019;176:649–62.e20.
557 26. Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, et al. Complementary
558 Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. mSystems
559 [Internet]. 2020;5. Available from: http://dx.doi.org/10.1128/mSystems.00768-19
560 27. Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. Nat Rev Microbiol. 2018;16:143–
561 55.
562 28. Buchka S, Hapfelmeier A, Gardner PP, Wilson R, Boulesteix A-L. On the optimistic performance
563 evaluation of newly introduced bioinformatic methods. Genome Biol. 2021;22:152.
564 29. Marx V. Bench pressing with genomics benchmarkers. Nat Methods. 2020;17:255–8.
565 30. Sun Z, Huang S, Zhang M, Zhu Q, Haiminen N, Carrieri AP, et al. Challenges in benchmarking
566 metagenomic profilers. Nat Methods. 2021;18:618–26.
567 31. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic
568 Classification. Cell. 2019;178:779–94.
569 32. Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods:
570 the DREAM of high-throughput pathway inference. Ann N Y Acad Sci. 2007;1115:1–22.
571 33. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure
572 prediction methods. Proteins. 1995;23:ii – v.
573 34. Meyer F, Lesker T-R, Koslicki D, Fritz A, Gurevich A, Darling AE, et al. Tutorial: assessing
574 metagenomics software with the CAMI benchmarking toolkit. Nat Protoc. 2021;16:1785–801.
575 35. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in
576 metagenomics data. PeerJ Comput Sci. PeerJ Inc.; 2017;3:e104.

577 36. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA
578 gene database project: improved data processing and web-based tools. Nucleic Acids Res.
579 2013;41:D590–6.
580 37. Salazar G, Ruscheweyh H-J, Hildebrand F, Acinas SG, Sunagawa S. mTAGs: taxonomic profiling
581 using degenerate consensus reference sequences of ribosomal RNA genes. Bioinformatics [Internet].
582 2021; Available from: http://dx.doi.org/10.1093/bioinformatics/btab465
583 38. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann Rev Mar Sci.
584 2017;9:231–55.
585 39. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, et al. Streamlining and core
586 genome conservation among highly divergent members of the SAR11 clade. MBio [Internet]. 2012;3.
587 Available from: http://dx.doi.org/10.1128/mBio.00252-12
588 40. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes
589 with the Genome Taxonomy Database. Bioinformatics [Internet]. 2019; Available from:
590 http://dx.doi.org/10.1093/bioinformatics/btz848
591 41. Shi W, Moon CD, Leahy SC, Kang D, Froula J, Kittelmann S, et al. Methane yield phenotypes
592 linked to differential gene expression in the sheep rumen microbiome [Internet]. Genome Research.
593 2014. p. 1517–25. Available from: http://dx.doi.org/10.1101/gr.168245.113
594 42. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and
595 cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol.
596 2021;22:93.
597 43. Paoli L, Ruscheweyh H-J, Forneris C, Kautsar S, Clayssen Q, Salazar S, et al. Uncharted
598 biosynthetic potential of the ocean microbiome. submitted. 2021;
599 44. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic
600 assembler. Genome Res. 2017;27:824–34.
601 45. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene
602 recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
603 46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
604 Bioinformatics. 2009;25:1754–60.

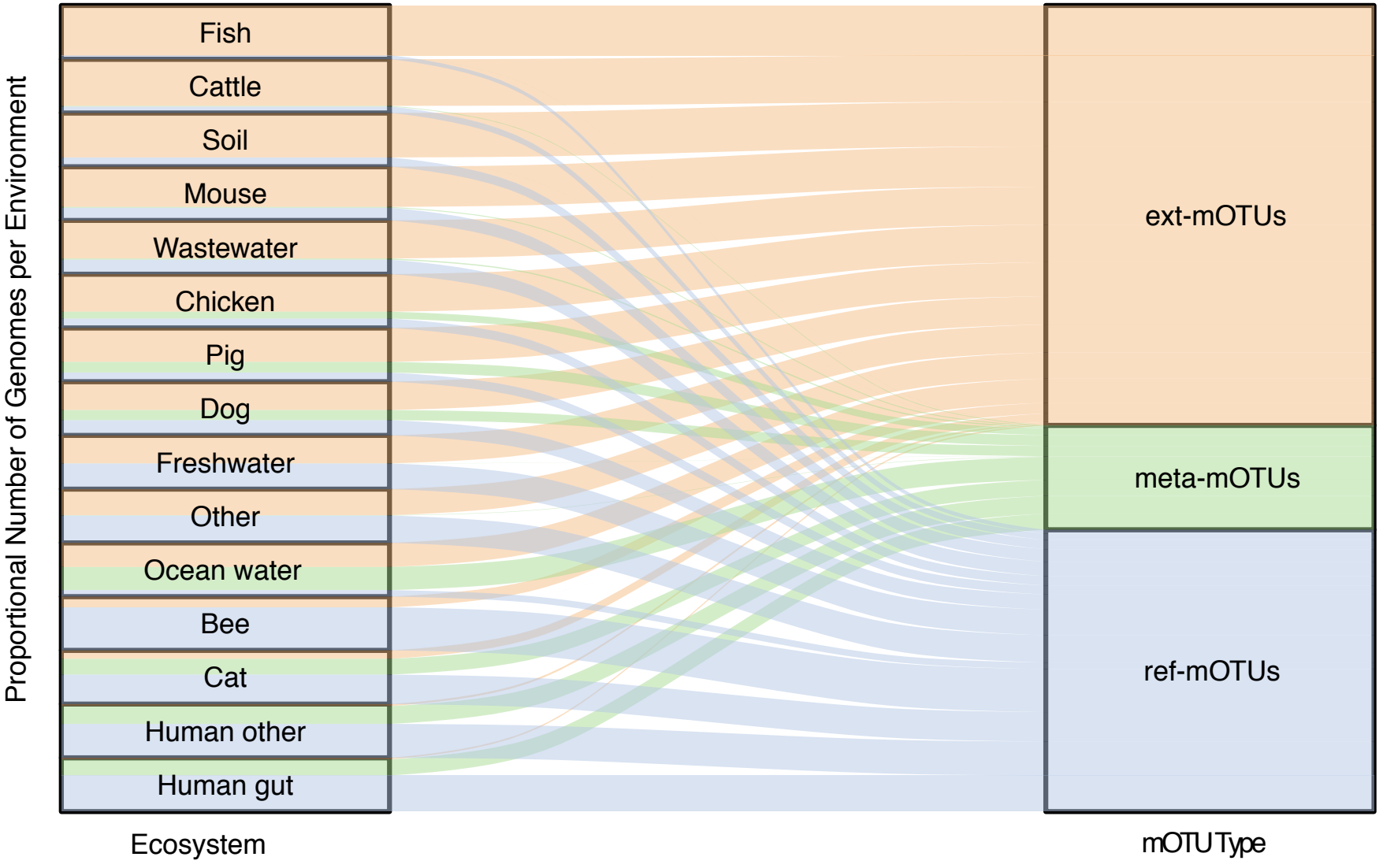
605 47. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced
606 analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
607 48. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality
608 of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res.
609 2015;25:1043–55.
610 49. Mende DR, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, et al. proGenomes2:
611 an improved database for accurate and consistent habitat, taxonomic and functional annotations of
612 prokaryotic genomes. Nucleic Acids Res. 2020;48:D621–5.
613 50. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for
614 metagenomics. PeerJ. 2016;4:e2584.
615 51. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic
616 species. Nat Methods. Nature Publishing Group; 2013;10:881–4.
617 52. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, et al. The vegan
618 package. Community ecology package. 2007;10:719.
619 53. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic
620 comparisons that enables improved genome recovery from metagenomes through de-replication.
621 ISME J. 2017;11:2864–8.
622 54. Hirschberg JB, Rosenberg A. V-Measure: A conditional entropy-based external cluster evaluation
623 [Internet]. Columbia University; 2007. Available from:
624 https://academiccommons.columbia.edu/doi/10.7916/D80V8N84
625 55. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast
626 genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
627 56. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful
628 approach to multiple testing. J R Stat Soc. Wiley; 1995;57:289–300.
629 57. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements
630 in performance and usability. Mol Biol Evol. 2013;30:772–80.
631 58. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
632 phylogenies. Bioinformatics. 2014;30:1312–3.

633 59. Zhang X, Tu B, Dai L-R, Lawson PA, Zheng Z-Z, Liu L-Y, et al. Petroclostridium xylanilyticum
634 gen. nov., sp. nov., a xylan-degrading bacterium isolated from an oilfield, and reclassification of
635 clostridial cluster III members into four novel genera in a new Hungateiclostridiaceae fam. nov. Int J
636 Syst Evol Microbiol. 2018;68:3197–211.
637

Supplementary Information

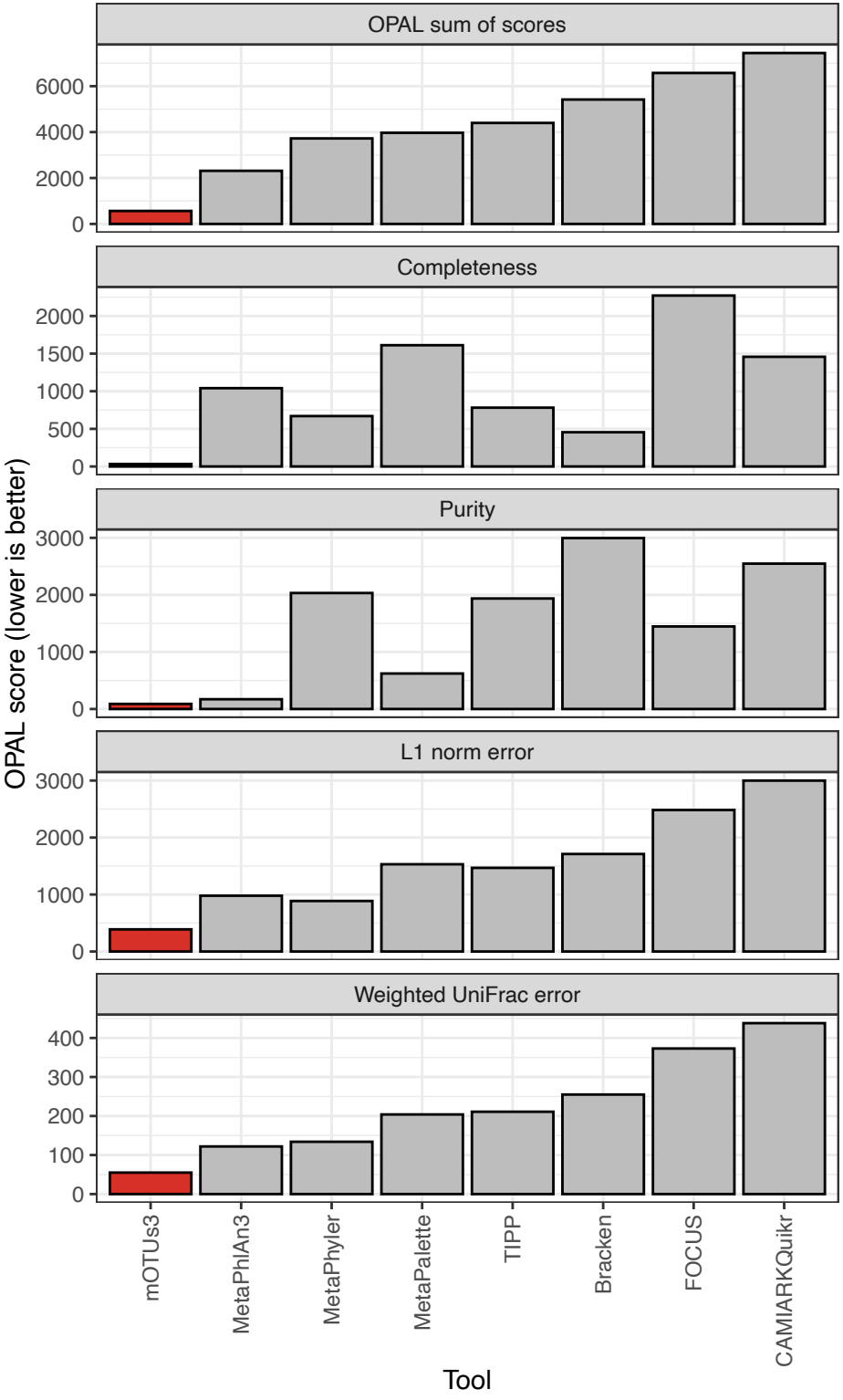
Supplementary Information for this manuscript includes:

- Legends for Supplementary Figures 1-2
- Legends for Supplementary Tables 1-5


Proportional Number of Genomes per Environment
Fish
Cattle
Soil
Mouse
Wastewater
Chicken
Pig
Dog
Freshwater
Other
Ocean water
Bee
Cat
Human other
Human gut
Ecosystem
ext-mOTUs
meta-mOTUs
ref-mOTUs
mOTU Type

Supplementary Figure 1. Environment-specific membership of genomes in ref-, meta- and ext-mOTUs.

A total of 499,512 genomes derived from 23 environments (environments with few genomes are grouped as 'Other', see Supplementary Tables 1 and 3) were used for the extension. The number of genomes was normalized by environments. The proportions of genomes per environment that are either associated with ref- and meta-mOTUs or were used to build ex-mOTUs are shown in the colors blue, green or orange, respectively. For example, the majority of genomes from the human gut match ref-mOTUs, whereas the vast majority of genomes from the fish environment are used to build ext-mOTUs.


OPAL sum of scores
Completeness
Purity
L1 norm error
Weighted UniFrac error
OPAL score (lower is better)
mOTUs3
MetaPhlAn3
MetaPhyler
MetaPalette
TIPP
Bracken
FOCUS
CAMIARKQuikr
Tool

Supplementary Figure 2. OPAL score broken down to individual metrics for the 63 mouse gut metagenomic samples.

The evaluation was performed using the OPAL tool [1] on 63 simulated mouse gut metagenomes [2], which also provided taxonomic profiles for seven different taxonomic profiling tools, and to which we have added mOTUs3 profiling results. The OPAL tool ranks the tools for each sample and for each taxonomic level. The measures considered are completeness, purity, L1 norm error and weighted UniFrac error, shown individually in the bottom 4 plots. Tools with a lower score perform better, as the OPAL score is a sum over rank. The top plot represents the OPAL sum of scores, which is the sum over the four individual measures. mOTUs3 scored best in all categories, including the OPAL sum of scores.

Supplementary Table Legends

Supplementary Table 1: Included studies and associated environments.

Data from 91 studies from 23 environments were included in the extension and/or profiling of the mOTUs database. Of these, 39 studies were selected for in-house MAG reconstruction and 11,164 sequencing samples from 67 studies were used for taxonomic profiling.

Supplementary Table 2: Sequencing samples included in the taxonomic profile.

A total of 11,164 samples were taxonomically profiled. Sample names are connected to public repositories by biosample and sequencing run ids. The project name column links the sample name to the study name used in Supplementary Table 1.

Supplementary Table 3: Breakdown of taxonomic novelty in ext-mOTUs.

Taxonomic novelty increases with higher ranks, i.e., more than 50% of ext-mOTUs were assigned to previously unknown families.

Supplementary Table 4: Contribution of genomes to ref-, meta- or ext-mOTUs.

Genomes/MAGs from different studies and environments contribute in varying proportions to the extension of the database.

Supplementary Table 5: Data for Figure 1.

For each sample that passed the filter (total 5,756), we reported the relative abundance for each mOTU type. Additionally, we added the total number of detected mOTUs and the habitat.

References

1. Meyer F, Bremges A, Belmann P, Janssen S, McHardy AC, Koslicki D. Assessing taxonomic metagenome profilers with OPAL. Genome Biol. 2019;20:51.

2. Meyer F, Lesker T-R, Koslicki D, Fritz A, Gurevich A, Darling AE, et al. Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. Nat Protoc. 2021;16:1785–801.