

Challenges in Benchmarking Metagenomic Profilers

Yang-Yu Liu (✉ yyl@channing.harvard.edu)

Harvard Medical School <https://orcid.org/0000-0003-2728-4907>

Zheng Sun

Brigham and Women's Hospital and Harvard Medical School

Shi Huang

University of California, San Diego

Meng Zhang

Inner Mongolia Agricultural University

Qi-Yun Zhu

University of California, San Diego

Niina Haiminen

IBM T.J. Watson Research Center <https://orcid.org/0000-0002-8663-1019>

Anna Paola Carrieri

IBM Research

Yoshiki Vázquez-Baeza

University of California, San Diego

Laxmi Parida

IBM Research - Thomas J. Watson Research Center <https://orcid.org/0000-0002-7872-5074>

Ho-Cheol Kim

IBM Almaden Research Center

Robin Knight

University of California, San Diego

Analysis

Keywords: metagenomics analysis, genomics, metagenomic profiling

Posted Date: November 30th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-109702/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Methods on May 13th, 2021. See the published version at <https://doi.org/10.1038/s41592-021-01141-3>.

Challenges in Benchmarking Metagenomic Profilers

Zheng Sun^{1,*}, Shi Huang^{2,3,*}, Meng Zhang⁴, Qi-Yun Zhu^{2,3}, Niina Haiminen⁵, Anna-Paola Carrieri⁶,
Yoshiki Vázquez-Baeza^{2,3}, Laxmi Parida⁵, Ho-Cheol Kim⁷, Rob Knight^{2,3,8,9,#}, Yang-Yu Liu^{1,#}

¹ Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

² Department of Pediatrics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

³ Center for Microbiome Innovation, Jacobs School of Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

⁴ Key Laboratory of Dairy Biotechnology and Engineering, Ministry of Education, Inner Mongolia Agricultural University, Hohhot, 010018, China

⁵ IBM T. J. Watson Research Center, Yorktown Heights, New York, USA

⁶ IBM Research UK, The Hartree Centre, Warrington, United Kingdom

⁷ AI and Cognitive Software, IBM Research-Almaden, San Jose, California, USA

⁸ Department of Computer Science & Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

⁹ Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

* These authors contributed equally

Correspondence: yyl@channing.harvard.edu and robknight@eng.ucsd.edu

Accurate microbial identification and abundance estimation are crucial for metagenomics analysis. Various methods for classifying metagenomic data and estimating taxonomic profiles, broadly referred to as metagenomic profilers, have been developed. Yet, benchmarking metagenomic profilers remains challenging because some tools are designed to report relative sequence abundance while others report relative taxonomic abundance. Here, we show how misleading conclusions can be drawn by neglecting this distinction between relative abundance types when benchmarking metagenomic profilers. Moreover, we show compelling evidence that interchanging sequence abundance and taxonomic abundance will influence both per-sample summary statistics and cross-sample comparisons. We suggest that the microbiome research community should pay attention to potentially misleading biological conclusions arising from this issue when benchmarking metagenomic profilers, by carefully considering the type of abundance data that was analyzed and interpreted, and clearly stating the strategy used for metagenomic profiling.

36 Identifying microbial species present in complex biological and environmental samples is one
37 of the major challenges in microbiology^{1,2}. By directly interrogating the community
38 composition in an unbiased and culture-independent manner, metagenomic sequencing is
39 transforming microbiology by enabling more rapid species detection and discovery². This has
40 a wide range of applications from surveying the bacteria in an environmental soil sample to
41 determining the etiology of an infection from a patient's blood or stool sample. Such
42 applications drive the development of various computational methods to analyze genomic data
43 generated by metagenomic sequencing to identify all of the species contained in the samples
44 and estimate their relative abundances^{2,3}. Those computational methods are broadly referred to
45 as metagenomic profilers.

46 Following a previous benchmarking study³, metagenomic profilers can be categorized
47 based on their reference database type (**Fig.1a**): (1) DNA-to-DNA methods (e.g., Kraken^{4,5},
48 Bracken⁶ and PathSeq⁷), which compare sequence reads with comprehensive metagenome
49 databases; (2) DNA-to-Protein methods (e.g., Kaiju⁸ and Diamond⁹), which compare sequence
50 reads with genomic databases of protein-coding sequences; or (3) DNA-to-Marker methods
51 (e.g., MetaPhlAn^{10,11} and mOTU^{12,13}), which only include specific gene families in their
52 reference databases. Note that those metagenomic profilers all rely on reference databases.
53 They should not be confused with *de novo* assembly-based methods that do not use any
54 reference databases^{14,15}. Those reference-free binning methods cannot taxonomically classify
55 sequences^{14,15} and are not directly comparable with the metagenomic profilers evaluated here.

56 Many studies have benchmarked metagenomic profilers^{3,16-19}, finding that the
57 performance of different profilers varies considerably even on the same benchmark datasets.
58 For example, in a recent benchmarking study³, the performance of 20 metagenomic profilers
59 were evaluated based on two key metrics: the area under the precision-recall curve (AUPRC)
60 for organism presence/absence, and the L2 distance between the observed and true relative
61 abundance profiles. It was found that DNA-to-DNA methods were among the best-scoring
62 methods, with typical average L2 distance < 0.1, while DNA-to-Marker methods had much
63 higher L2 distance, indicating less favorable performance.

64 Here we show that this apparently high performance variation largely arises because
65 the methods report one of two fundamentally different types of relative abundances: *sequence*
66 *abundance* or *taxonomic abundance*. For example, the raw output of DNA-to-DNA methods
67 is the relative abundance of a given taxon calculated as the proportion of sequences assigned
68 to it out of the total number of sequences, i.e., the sequence abundance. By contrast, DNA-to-

69 Marker methods directly output the relative abundance of each microbial taxon calculated as
70 the number of genomes of that taxon relative to the total number of genomes detected, i.e., the
71 taxonomic abundance. For DNA-to-Protein methods, the output type is the relative sequence
72 abundance of protein-coding sequences^{8,9}.

73 Unfortunately, the distinction between the two types of relative abundances has rarely
74 been carefully considered in previous benchmarking studies. In this paper, we show that the
75 two types of relative abundances are not related by any simple algebraic relation. Moreover,
76 interchanging them leads to very misleading performance assessments of metagenomic
77 profilers. These results imply that many benchmarking results presented in the literature require
78 re-examination. Beyond examining the previous benchmarking results, we further point out the
79 serious issues in microbiome data analysis based on sequence abundances, which are typically
80 produced by DNA-to-DNA methods and have been applied in thousands of published
81 microbiome studies (e.g., Kraken: 1,283 citations; Kraken2: 95 citations; Bracken: 139
82 citations by November 2020, according to their official websites). We find that microbiome
83 data analysis based on sequence abundance will underestimate (or overestimate) the relative
84 abundances of microbes with smaller (or larger) genome sizes. This will fundamentally affect
85 differential abundance analyses and other analytical methods that rely on accurate counts in
86 their input contingency matrix. Without careful consideration, these issues could impede cross-
87 study comparisons of differentially abundant taxa identified from different methods. We think
88 this point needs more attention from the entire microbiome research community.

89

90 **Results**

91 **Illustration of the caveat in benchmarking metagenomic profilers.** To illustrate the caveat
92 of confusing sequence abundance and taxonomic abundance in benchmarking metagenomic
93 profilers, we simulated a simple microbial community with only two genomes, where genome
94 A (*Bacillus pseudofirmus*, GCF_000005825.2, size: 4.2MB) is twice the size of genome B
95 (*Lactobacillus salivarius*, GCF_000008925.1, size: 2.1MB), corresponding to **Fig.1b**. In this
96 simulated community, the sequence abundance ratio of genome A: genome B = 1:1, while the
97 taxonomic abundance ratio of genome A: genome B = 1:2. DNA-to-DNA profilers Bracken,
98 Kraken2 and PathSeq reported that this sample contains 49.9% (or 50.1% in Kraken2 and 50.6%
99 in PathSeq) *Bacillus pseudofirmus* and 50.1% (or 49.9% in Kraken2 and 49.4% in PathSeq)
100 *Lactobacillus salivarius*, respectively (**Fig.1c**). DNA-to-Markers profilers MetaPhlan2 and
101 mOTUs2 reported the relative abundance of *Bacillus pseudofirmus* as 33.8% (or 33.6%) and

102 *Lactobacillus salivarius* as 66.2% (or 66.4%, **Fig.1c**), respectively. This simple example
103 clearly illustrates how the sequence abundance profile produced by DNA-to-DNA profilers
104 does not represent the true taxonomic abundance of a microbiome sample.

105 Note that for this simple synthetic community, DNA-to-Protein profilers Kaiju and
106 Diamond reported the relative abundance of *Bacillus pseudofirmus* as 22.8% (or 7.0%) and
107 *Lactobacillus salivarius* as 19.9% (or 8.0%), respectively (**Fig.1c**). Besides the false positives
108 (57.3% in Kaiju and 85.0% in Diamond), the ratio between the relative abundances of the two
109 species is roughly 1:1, indicating the methods are indeed reporting sequence abundance.
110 However, these classifiers reported a large number of false positive species identified due to
111 the conservation of protein sequence²⁰. Going forward, we will focus on benchmarking the
112 DNA-to-DNA and DNA-to-Markers methods.

113

114 **No simple algebraic relation between the two types of relative abundances.** We emphasize
115 that mathematically there is no simple algebraic relation between the two types of relative
116 abundances, even in the ideal case (when all genomes/taxa are known). Denote R_i as the
117 number of metagenomic reads assigned to the genome of a microbial taxon i with genome size
118 L_i and ploidy P_i (i.e., the number of copies of the genome in one cell, however most methods
119 did not consider the ploidy into the abundance estimation as the information is still lacking for
120 many genomes). The number of microbial cells classified as taxon i is then given by $C_i =$
121 $R_i/(L_i P_i)$. Let n be the number of identified taxa in the sample. Then the sequence abundance
122 of taxon i is given by

$$124 \quad S_i = \frac{R_i}{\sum_{i=1}^n R_i}, \quad [1]$$

123 and its taxonomic abundance is given by

$$125 \quad T_i = \frac{C_i}{\sum_{i=1}^n C_i} = \frac{R_i/(L_i P_i)}{\sum_{i=1}^n R_i/(L_i P_i)}. \quad [2]$$

126 Eqs.[1-2] imply that as long as L_i and P_i vary across different taxa in a community, S_i and T_i
127 are not connected by any simple algebraic relation.

128 The variation of genome size L_i of different taxa can be very large. Indeed, in the
129 recently updated microbial genome database (NCBI RefSeq, 2020 Nov 6th), the sizes of fully
130 sequenced and assembled microbial genomes vary considerably (**Fig.2a**). For example, just
131 within the bacteria kingdom, the genome size variation can be more than 100-fold, e.g.,
132 *Candidatus Nasuia deltocephalinicola* (GCF_000442605.1) with 112,091 bp vs. *Sorangium*

133 *cellulosum* (GCF_000418325.1) with 14,782,125 bp. Therefore, microbial genome sizes could
134 vary radically within a single microbiome sample, including when viruses (which tend to have
135 shorter genomes, Fig. 2a) are analyzed together with bacteria in shotgun metagenomics.

136 Regrading ploidy P_i , although prokaryotes are usually thought to contain one copy of a
137 circular chromosome, previous studies have demonstrated that many species of archaea and
138 bacteria are polyploid and can contain more than ten copies of their chromosome²¹. In fact,
139 extreme polyploidy has been observed in a large bacterium *Epulopiscium*, which contains tens
140 of thousands of copies of its genome²².

141 The variations in L_i and P_i drive the theoretical distinction between sequence
142 abundance and taxonomic abundance. This point can be seen clearly from simulated microbial
143 communities based on the NCBI RefSeq database. As shown in **Fig.2b**, where we investigate
144 a complex microbial community consisting of all different kingdoms of microbes (fungi,
145 bacteria and virus), S_i tends to overestimate the abundances of species with larger genome sizes
146 (e.g., fungi) and underestimate the abundances of species with smaller genome sizes (e.g.,
147 viruses). This is true even if we investigate a community consisting of microbes from the same
148 kingdom (**Fig.2c**). Note that here, for the sake of simplicity, in our simulations we did not
149 consider the variation of ploidy, but only focused on the variation of genome sizes. Hence, the
150 demonstrated difference between sequence abundance and taxonomic abundance is
151 conservative.

152 In reality, unknown genomes/taxa will further complicate the relation between S_i and
153 T_i , and affect metagenomic profiler benchmarking on real data (because different profilers
154 handle unknown genomes/taxa differently). Moreover, instead of converting S_i to T_i through
155 L_i and P_i correction, DNA-to-Marker methods directly calculate T_i as the ratio of sequence
156 coverage of single-copy marker genes of each taxon to that of all taxa. This also affects the
157 metagenomic profiler benchmarking.

158
159 **Benchmarking results depend on the abundance type.** To further illustrate the problem of
160 mixing sequence abundance and taxonomic abundance in benchmarking metagenomic
161 profilers, we simulated metagenomic sequencing reads for 25 communities from distinct
162 habitats (e.g., gut, oral, skin, vagina and building, five communities for each habitat, see
163 **Methods**). To avoid database biases of different metagenomic profilers, the selection of
164 genomes for simulated data was based on the intersection between MetaPhlAn2, mOTUs2
165 reference database, and Kraken2 reference database (which was also used by Bracken). Then

166 we calculated the dissimilarity or distance between the ground truth abundance profiles and the
167 estimated ones from different profilers, based on the following five measures: Bray-Curtis
168 dissimilarity (BC), L1 distance, L2 distance, root Jensen-Shannon divergence (rJSD), and
169 robust Aitchison distance (rAD)²³ (**Fig.3a,b**). Note that the Aitchison distance (based on
170 centered log-ratio transform) is a compositionally aware distance measure²³. However, it
171 suffers from the inflated zero counts in microbiome data because log-transform of zero counts
172 is undefined unless arbitrary pseudocounts are added to each taxon. Here the calculation of
173 rAD does not involve any pseudocounts, and it naturally avoids the issue of dealing with sparse
174 zero counts using the classical Aitchison distance²³.

175 We found that for BC, L1, L2 and rJSD, if the sequence abundance is used as the ground
176 truth, Bracken and Kraken2 outperform MetaPhlan2 and mOTUs2; while if the taxonomic
177 abundance is used as the ground truth, MetaPhlan2 and mOTUs2 outperform Bracken and
178 Kraken2. Interestingly, with rAD as the evaluation metric, regardless of if sequence or
179 taxonomic abundance profiles were taken as the ground truth, mOTUs2 and MetaPhlan2
180 always outperform Bracken and Kraken. This could be due to the compositionally aware
181 distance measure rAD weighing low-abundance taxa more than the other measures. To test this
182 idea, we sought to rule out the bias introduced by false positives and calculated rAD based on
183 taxonomic profilers where false positives have been removed (**Methods**). This is denoted as
184 modified rAD in **Fig.3**. We found that, with the modified rAD as the evaluation metric, the
185 benchmarking result is the same as that of using BC, L1, L2 and rJSD, or their modified
186 versions (**Fig.S1**). We always found the same pattern: if the sequence abundance is used as the
187 ground truth, Bracken and Kraken2 outperform MetaPhlan2 and mOTUs2; while if the
188 taxonomic abundance is used as the ground truth, MetaPhlan2 and mOTUs2 outperform
189 Bracken and Kraken2. This result strongly indicates that the benchmarking result of
190 metagenomic profilers depends on the selected abundance type.

191 We emphasize that the above contradicting performance evaluations due to different
192 abundance types cannot be detected by using the AUPRC metric, because the calculation of
193 the Precision-Recall Curve only concerns the difference of presence/absence patterns in the
194 ground truth and predicted abundance profiles. By definition, the ground truth sequence
195 abundance and taxonomic abundance profiles of our simulated microbiome samples share
196 exactly the same presence/absence pattern.

197 Moreover, we emphasize that even though the five distance/dissimilarity measures (BC,
198 L1, L2, rJSD, and rAD) all showed the similar results in the performance evaluation (after the

199 removal of false positives), L2 was not designed for compositional data analysis. To investigate
200 whether the discriminating power of these distance measures for the two sequence types
201 persists with varied microbial diversity, we simulated a set of abundance tables (for both
202 taxonomic abundance and sequence abundance) with different species counts ranging from 10
203 to 500 (see **Methods**). We then calculated the distance or dissimilarity between the sequence
204 abundance and taxonomic abundance profiles (**Fig.4**). We found that with an increasing
205 number of species, L2 keeps decreasing while L1, BC, rJSD and rAD can still distinguish the
206 two abundance types. This result suggests that L2 distance cannot discriminate the two types
207 of relative abundances in microbiome samples of high species richness. This might be due to
208 the fact that L2 distance is not appropriate for compositional data analysis.

209

210 **Impact of abundance type on the alpha diversity calculation.** Interchanging sequence
211 abundance and taxonomic abundance strongly influences per-sample summary statistics. To
212 demonstrate this issue, we simulated 500 abundance profiles representing microbiota from
213 distinct habitats (gut, oral, skin, vagina, and building, 100 profiles for each, see **Methods**) with
214 known sequence abundance and taxonomic abundance profiles. We found that the Shannon
215 and Simpson indices calculated from taxonomic abundances are significantly higher than those
216 calculated from sequence abundances ($p < 0.001$, Wilcoxon rank-sum test) regardless of the
217 habitat (**Fig.5**). Moreover, when ranking the samples according to their alpha diversity
218 measures calculated from sequence abundance and from taxonomic abundance, the orderings
219 are not fully concordant with each other (Spearman correlation of the rank vectors is $0.929 \pm$
220 0.020 for Shannon index and 0.835 ± 0.042 for Simpson index). These results suggest that alpha
221 diversity calculations and comparisons can be strongly affected by the type of relative
222 abundance used.

223

224 **Impact of abundance types on the beta diversity and ordination analyses.** To check if
225 mixing sequence abundance and taxonomic abundance will also influence between-sample
226 attributes such as beta diversity and ordination analyses, we reanalyzed the 500 samples
227 generated for Fig.5. In order to quantify the influence on beta diversity introduced by
228 abundance type, Mantel test^{24, 25} was employed to compare the beta-diversity (in terms of BC,
229 rJSD, L1, L2 and rAD) calculated from the taxonomic abundance and sequence abundance
230 profiles of those samples (see **Methods**). Interestingly, regardless of the species richness in the
231 habitats, the abundance type has some influence on the cross-sample comparisons based on the

232 BC, rJSD and L1 measures (Spearman coefficient $r=0.944\pm 0.006$, 0.947 ± 0.009 , 0.944 ± 0.006 ,
233 respectively; p -value = $1e-4$ for all), but affects the L2 and rAD measures more strongly (r
234 $=0.844\pm 0.026$, 0.519 ± 0.137 , respectively; p -value= $1e-4$ for both). Moreover, we found that
235 species richness of samples associates with the correlation coefficient in the rAD calculation.
236 These results demonstrate the inconsistent relative relationships between samples that are
237 introduced by different abundance types in beta diversity calculation.

238 We then performed ordination analyses using four different methods: Non-metric
239 Multidimensional Scaling (NMDS)²⁶, Principal Coordinates Analysis (PCoA)²⁷, t-distributed
240 stochastic neighbor embedding (t-SNE)²⁸, and Uniform Manifold Approximation and
241 Projection (UMAP)²⁹. We found that, regardless of the distance/dissimilarity measures used
242 (e.g. rJSD, BC and rAD), taxonomic abundance and sequence abundance profiles are
243 drastically different in all the four ordination results (**Fig.6, Figs.S2-S3**). Procrustes analysis
244 was then employed to analyze the congruence of two-dimensional shapes produced from
245 superimposition of ordination analyses from two datasets^{30, 31}. Indeed, Procrustes analysis
246 revealed very low similarity between the ordination results calculated from sequence and
247 taxonomic abundance (Fig.6, Figs.S2-S3, Monte Carlo p -value <0.05). These results indicate
248 that both beta diversity (especially for L2 and rAD) and ordination analyses can be heavily
249 affected by the relative abundance type used.

250

251 **Discussion**

252 Taken together, these analyses emphasize the importance of differentiating between relative
253 sequence abundance and relative taxonomic abundance in metagenomic profiling. Ignoring this
254 distinction can potentially underestimate the relative abundance of organisms with small
255 genome sizes. Sequence abundances are typically produced by DNA-to-DNA or DNA-to-
256 Protein methods, which rely on microbial genomes or genes as the reference database, report
257 relative sequence abundance, i.e. the fraction of sequence reads assigned to each entity in the
258 database. By contrast, DNA-to-Marker methods output relative taxonomic abundance
259 representing the fraction of each detected taxon.

260 Our results demonstrate that misleading performance assessment of metagenomic
261 profilers and spurious alpha and beta diversity patterns can arise from interchanging sequence
262 abundance with taxonomic abundance. For alpha diversity, Shannon index and Simpson index
263 are not simply higher based on taxonomic abundance than that based on sequence abundance,
264 the relative ranking of alpha diversity is not consisting in the two abundance types either.

265 Dramatic changes in the relative position between samples are also shown in the ordination
266 analysis. Therefore, interchanging abundance types could have a deleterious effect on the
267 interpretation of alpha and beta diversity analyses and meta-analyses.

268 The distinction between the two types of relative abundances was known to the field of
269 microbiome research (at least to the developers of various metagenomic profilers), and has
270 been conceptually considered in some benchmark studies (e.g., CAMI¹⁹). However, the
271 consequences of ignoring this distinction for benchmarking metagenomic classifiers and per-
272 sample summary statistics have not been quantitatively studied or clearly illustrated so far. In
273 particular, the vast majority of users of those metagenomic profilers should be clearly aware of
274 the distinction between sequence abundance and taxonomic abundance, and of the
275 consequences of ignoring this distinction in selecting metagenomics tools, data interpretation,
276 and cross-study comparison of differentially abundant taxa identified by different tools.

277 In summary, we suggest that the microbiome research community should pay more
278 attention to potentially misleading biological conclusions arising from this issue by carefully
279 considering which type of abundance data was analyzed and interpreted, and, going forward,
280 the strategy used for taxonomy assignment should be clearly represented. We also suggest that,
281 in future development or evaluation of metagenomic profilers, both types of relative abundance
282 should be strictly distinguished, and both should be reported. This would substantially improve
283 the comparison of abundance estimations of metagenomic profilers and enhance the
284 reproducibility and biological interpretation of microbiome studies.

285

286 **Methods**

287 **Simulation of microbiome profiles.** In the simulation of microbiome profiles based on
288 different species counts (from 10 to 500), the abundance was created randomly from a log-
289 normal distribution using “rlnorm” function in R language with parameters: meanlog = 0 and
290 sdlog = 1, and 10 repeats were simulated for each species count. In the simulation of
291 microbiome profiles for alpha diversity calculation, 100 profiles were simulated for each
292 habitat, and species counts in different habitats were set up as: 10-50 (vaginal), 50-100 (skin),
293 100-150 (gut), 150-200 (oral), 200-300 (building). The representative species in each specific
294 habitat were selected based on the set of microbial species identified in the HMP³² and by Hsu
295 et al.³³.

296
297 **Simulation of sequencing reads.** Firstly, the 25 microbiome profiles (five for each habitat)
298 were simulated using the above method. Then the simulation of sequencing data is illustrated
299 as the process in Fig.1a: Given a specified species composition (taxonomic abundance), their
300 sequence abundance can be inferred accordingly (taxonomic abundance equals to sequence
301 abundance divide by their genome length) and “Wgsim” (<https://github.com/lh3/wgsim>) was
302 then used (with default parameters) to simulate the sequences. The selection of genomes for
303 simulated data was based on the intersection between MetaPhlAn2 and mOTUs2 reference
304 database and Bracken’s database to avoid database biases.

305 Currently, there are many more DNA-to-DNA profilers (e.g., Bracken and Kraken2)
306 than DNA-to-Marker profilers (e.g., MetaPhlAn2 and mOTU2). In this paper we focused on
307 two DNA-to-DNA profilers for the following reasons. First, as representative DNA-to-DNA
308 methods, Bracken and Kraken/Kraken2 demonstrated the best performance in previous
309 benchmarking studies^{4,6,34}, and have been cited in more than one thousand microbiome studies.
310 Second, mOTU2 and MetaPhlAn2 do not support custom reference databases, and the
311 reference database is a critical factor affecting profiler performance. As such we decided to use
312 the intersection of organisms in mOTU2, MetaPhlAn2, and Kraken2 reference databases as
313 the source for our simulation data. Introducing more DNA-to-DNA profilers could further
314 reduce the reference database size of the simulated data and affect the diversity of genome sizes
315 **(Fig.S4)**.

316
317 **Alpha and beta diversity calculation.** Alpha diversity calculation e.g. Shannon and Simpson
318 indices were performed in R language by the “Vegan 2.5-6” package. As for the beta diversity,

319 we employed "Vegan 2.5-6" for distance/dissimilarity calculation e.g. L1 ("Manhattan" in
320 vegdist function), L2 ("Euclidean") and BC ("Bray"), while rJSD and rAD were calculated by
321 self-programmed script (see **code availability**). In the ordination analyses, R packages "ade4
322 1.7-15", "Rtsne 0.15", "ape 5.4-1" and "umap 0.2.6.0" were used to conduct the NMDS, t-SNE,
323 PCoA and UMAP analyses separately. Since the iterative algorithm of NMDS, t-SNE and
324 UMAP find different solutions depending on the starting point of the calculation (which is
325 a randomly chosen configuration) we performed 101 repeats of NMDS, t-SNE, UMAP and
326 their Procrustes test, the median result (sorting by the Mote-Caro test) was selected for
327 presentation of similarity and p-value in **Fig.6**, **Fig.S2** and **Fig.S3**. The ordination analyses
328 based on the ground truth of the sequence abundance and taxonomic abundance for the 500
329 profiles (from five habitats) were conducted separately before Procrustes analysis.

330

331 **Robust Aitchison distance calculation.** We applied DEICODE
332 (<https://github.com/biocore/DEICODE>) to calculate the robust Aitchison distance (rAD) to
333 benchmark the performance of metagenomics profilers. DEICODE represents a form of
334 Aitchison Distance that is robust to high levels of sparsity. It preprocesses the compositional
335 data using the centered log-ratio (CLR) transform only on the non-zero values of the data
336 (hence no pseudo counts are used). Then it performs dimensionality reduction through robust
337 PCA based on the non-zero values of the data. The Euclidean distance of the robust CLR-
338 transformed abundance profiles (i.e., rAD) was finally employed to evaluate the performance
339 of metagenomic profilers. To avoid the impact of false positives on the benchmarking results,
340 we further filtered out false positives in all output taxonomic profiles (Kraken2:
341 29.26%±12.13%; Bracken: 36.91%±12.11%; mOTUs2: 11.47%±4.62%; MPA2:
342 11.29%±4.19%) and compared the performance of different profilers using rAD calculated
343 from the true positives only. This is termed as the modified rAD in **Fig.3**. For other evaluation
344 measures, the same procedure was performed and presented in **Fig.S1**.

345

346 **Mantel Test.** Mantel test was used as a correlation test to determine the correlation between
347 two beta diversity (BC, rJSD, L1, L2 and rAD) matrices based on sequence abundance and
348 taxonomic abundance. In order to calculate the correlation, the matrix values of both matrices
349 are 'unfolded' into long column vectors, which are then used to determine correlation.
350 Permutations (n=9999) of one matrix are used to determine significance. Whether distances
351 between samples in one matrix are correlated with the distances between samples in the other

352 matrix is revealed by the p-value.

353

354 **Procrustes analysis.**

355 Procrustes analysis (by R package “ade4 1.7-15”) typically takes as input two coordinate
356 matrices with matched sample points, and transforms the second coordinate set by rotating,
357 scaling, and translating it to maximize the similarity between corresponding sample points in
358 the two shapes. It allows us to determine whether we would come to same conclusions on the
359 beta diversity, regardless of which distance/dissimilarity measure was used to compare the
360 samples. To assess the significance level of observed similarity between two matrices,
361 empirical p-values are calculated using a Monte Carlo simulation. Basically, sample labels are
362 shuffled in one of the coordinate matrices, and then the similarity between them is re-computed
363 for 9999 times. Here, similarity is calculated as the sum of the squared residual deviations
364 between sample points for each measurement. The proportion of similarity values that are equal
365 to or lower than the observed similarity value is then the Monte Carlo or empirical p-value.

366

367 **Data availability**

368 All the simulated datasets can be downloaded here:

369 https://figshare.com/projects/Challenges_in_Benchmarking_Metagenomic_Profilers/79916.

370

371 **Code availability**

372 R scripts used in this paper is available at <https://github.com/shihuang047/re-benchmarking>.

373

374 **Acknowledgements**

375 Research reported in this publication was supported by grants R01AI141529, R01HD093761,
376 UH3OD023268, U19AI095219, and U01HL089856 from National Institutes of Health. This
377 work was also supported by IBM Research through the AI Horizons Network, UC San Diego
378 AI for Healthy Living program in partnership with the UC San Diego Center for Microbiome
379 Innovation.

380

381 **Author contributions**

382 Y.-Y.L. and R.K. conceived and designed the analysis. Z.S. and H.S. led the analysis. M.Z.,
383 Q.Z., N.H., A.-P.C., Y.V., L.P., and H.-C.K. contributed evaluation strategies. All authors
384 analyzed the results. Z.S., H.S., Y.-Y.L., and R.K. wrote the paper. All authors edited the paper.

385

386 **Competing interests**

387 The authors declare no competing interests.

388

389

390 **Reference:**

- 391 1. Garud, N.R. & Pollard, K.S. Population Genetics in the Human Microbiome. *Trends*
392 *Genet* **36**, 53-67 (2020).
- 393 2. Knight, R. et al. Best practices for analysing microbiomes. *Nature Reviews*
394 *Microbiology* **16**, 410-422 (2018).
- 395 3. Ye, S.H., Siddle, K.J., Park, D.J. & Sabeti, P.C. Benchmarking Metagenomics Tools
396 for Taxonomic Classification. *Cell* **178**, 779-794 (2019).
- 397 4. Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2.
398 *Genome Biology* **20** (2019).
- 399 5. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification
400 using exact alignments. *Genome Biol* **15**, R46 (2014).
- 401 6. Lu, J., Breitwieser, F.P., Thielen, P. & Salzberg, S.L. Bracken: estimating species
402 abundance in metagenomics data. *Peerj Computer Science* (2017).
- 403 7. Kostic, A.D. et al. PathSeq: software to identify or discover microbes by deep
404 sequencing of human tissue. *Nat Biotechnol* **29**, 393-396 (2011).
- 405 8. Menzel, P., Ng, K.L. & Krogh, A. Fast and sensitive taxonomic classification for
406 metagenomics with Kaiju. *Nature Communications* **7** (2016).
- 407 9. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using
408 DIAMOND. *Nature Methods* **12**, 59-60 (2015).
- 409 10. Truong, D.T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling.
410 *Nature Methods* **12**, 902-903 (2015).
- 411 11. Segata, N. et al. Metagenomic microbial community profiling using unique clade-
412 specific marker genes. *Nature Methods* **9**, 811-+ (2012).
- 413 12. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with
414 mOTUs2. *Nature Communications* **10** (2019).
- 415 13. Sunagawa, S. et al. Metagenomic species profiling using universal phylogenetic marker
416 genes. *Nature Methods* **10**, 1196-1199 (2013).
- 417 14. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. metaSPAdes: a new
418 versatile metagenomic assembler. *Genome Res* **27**, 824-834 (2017).
- 419 15. Li, D. et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by
420 advanced methodologies and community practices. *Methods* **102**, 3-11 (2016).
- 421 16. Mavromatis, K. et al. Use of simulated data sets to evaluate the fidelity of metagenomic
422 processing methods. *Nature Methods* **4**, 495-500 (2007).
- 423 17. McIntyre, A.B.R. et al. Comprehensive benchmarking and ensemble approaches for
424 metagenomic classifiers. *Genome Biology* **18** (2017).
- 425 18. Meyer, F. et al. Assessing taxonomic metagenome profilers with OPAL. *Genome*
426 *Biology* **20** (2019).
- 427 19. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation-a benchmark of
428 metagenomics software. *Nature Methods* **14**, 1063-+ (2017).
- 429 20. Chen, F., Mackey, A.J., Vermunt, J.K. & Roos, D.S. Assessing performance of
430 orthology detection strategies applied to eukaryotic genomes. *PloS one* **2**, e383-e383
431 (2007).
- 432 21. Soppa, J. Polyploidy in archaea and bacteria: about desiccation resistance, giant cell
433 size, long-term survival, enforcement by a eukaryotic host and additional aspects. *J Mol*
434 *Microbiol Biotechnol* **24**, 409-419 (2014).
- 435 22. Mendell, J.E., Clements, K.D., Choat, J.H. & Angert, E.R. Extreme polyploidy in a
436 large bacterium. *Proc Natl Acad Sci U S A* **105**, 6730-6734 (2008).
- 437 23. Martino, C. et al. A Novel Sparse Compositional Technique Reveals Microbial
438 Perturbations. **4**, e00016-00019 (2019).

- 439 24. Legendre, P., Borcard, D. & Peres-Neto, P.R. ANALYZING BETA DIVERSITY:
440 PARTITIONING THE SPATIAL VARIATION OF COMMUNITY COMPOSITION
441 DATA. **75**, 435-450 (2005).
- 442 25. Mantel, N. The detection of disease clustering and a generalized regression approach.
443 *Cancer research* **27**, 209-220 (1967).
- 444 26. Faith, D.P., Minchin, P.R. & Belbin, L. Compositional dissimilarity as a robust measure
445 of ecological distance. *Vegetatio* **69**, 57-68 (1987).
- 446 27. Legendre, P. & Gallagher, E.D. Ecologically meaningful transformations for ordination
447 of species data. *Oecologia* **129**, 271-280 (2001).
- 448 28. Hinton, G.E.J.J.o.M.L.R. Visualizing High-Dimensional Data Using t-SNE. **9**, 2579-
449 2605 (2008).
- 450 29. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold
451 Approximation and Projection. *Journal of Open Source Software* **3** (2018).
- 452 30. Dray, S., Chessel, D. & Thioulouse, J. Procrustean co-inertia analysis for the linking of
453 multivariate datasets. *Écoscience* **10**, 110-119 (2003).
- 454 31. Digby, P. & Kempton, R. *Multivariate Analysis of Ecological Communities*. (1987).
- 455 32. Human Microbiome Project, C. Structure, function and diversity of the healthy human
456 microbiome. *Nature* **486**, 207-214 (2012).
- 457 33. Hsu, T. et al. Urban Transit System Microbial Communities Differ by Surface Type
458 and Interaction with Humans and the Environment. *mSystems* **1**, e00018-00016 (2016).
- 459 34. McIntyre, A.B.R. et al. Comprehensive benchmarking and ensemble approaches for
460 metagenomic classifiers. *Genome Biol* **18**, 182 (2017).
- 461

462 **Figures**

463 **Figure 1. Comparison of profiling results.** **a**, Illustration of the reference databases and the default
464 output abundance type for DNA-to-DNA, DNA-to-Protein and DNA-to-Marker profilers on a mixture
465 of two species A (1 cell) and B (2 cells). **b**, A simulated microbial community with only two genomes:
466 *Bacillus pseudofirmus* (genome size 4.2MB) and *Lactobacillus salivarius* (genome size 2.1MB). We
467 merged one copy of *Bacillus pseudofirmus* genome (genome A) with two copies of *Lactobacillus*
468 *salivarius* genome (genome B) sequences into one metagenome file. Then we sheared the merged
469 metagenomic sequences into 150bp to simulate a typical metagenomic dataset. **c**, Profiling results
470 (default output) of different profilers for the simulated microbial community shown in **a**. The bar plots
471 show the estimated relative abundance of the two microbial members A and B using different
472 metagenomics profilers.

473

474 **Figure 2. Correlation between sequence abundance and taxonomic abundance in synthetic**
475 **profiles based on different kingdoms.** **a**, Genome size distribution of microorganisms calculated from
476 the microbial genome database (NCBI RefSeq 2020 Nov 6th) that includes 171,927 bacteria, 293 fungi,
477 945 archaea, and 9,362 viruses. **b**, The scatter plot shows the correlation between taxonomic abundance
478 (x axis) and sequence abundance (y axis) of 600 randomly selected species in a simulated profile which
479 includes bacteria (species number=200), fungi (species number=200) and virus (species number=200).
480 **c**, Each scatter plot shows the correlation between taxonomic abundance (x axis) and sequence
481 abundance (y axis) of 200 randomly selected species in three simulated profiles which represent
482 different kingdoms e.g. bacteria, fungi, and virus.

483

484 **Figure 3. Differential benchmarking results of four representative metagenomics profilers using**
485 **two types of relative abundance as ground truth:** **a**, sequence abundance and **b**, taxonomic
486 abundance. The boxplots indicate the dissimilarities based on L1, L2, root Jensen-Shannon divergence
487 (rJSD), Bray-Curtis (BC), and robust Aitchison distance (rAD) between the ground-truth profiles and
488 the profiles predicted by different metagenomics profilers (Bracken, Kraken2, mOTUs2, and
489 MetaPhlan2) at the species level. For each metagenomic profiler, we performed the dissimilarity
490 calculations based on 25 simulated microbial communities from five representative environmental
491 habitats (gut, oral, skin, vagina and building) separately. Note that for each profiler based on any
492 evaluation metric, its performance variation across different synthetic communities is due to
493 microbiome complexity difference (e.g. species composition and richness). The asterisks in the boxplots
494 refer to the statistical significance: “*” refers to p-value <0.05, “**” refers to <0.01, “***” refers to <
495 0.001.

496

497 **Figure 4. Dissimilarity between sequence abundance and taxonomic abundance with varied**

498 **species number measured by different distance measures.** For each species number, we simulated
499 10 repeats of profiles. The distance/dissimilarity was then measured by different measures: rAD (red),
500 L1 (blue), L2 (purple), Bray-Curtis (yellow) and rJSD (green). rAD between these types of abundance
501 profiles positively correlated with the species richness when < 200 microbial species presented in a
502 community, yet saturated after the number of species reaching 200. L1, BC and rJSD can also reveal
503 the difference between the two abundance types yet they were not affected by the species-level richness.
504 L2 distance between the two abundance types dramatically dropped with the increase in the species-
505 level richness. In the complex community with the number of species over 200, L2 distance metric
506 almost lost the discriminatory power of these two abundance profiles.

507

508 **Figure 5. Alpha diversity based on sequence abundance and taxonomic abundance.** Alpha
509 diversity (Shannon index and Simpson index) based on ground truth of simulated data from different
510 habitats revealed the influence of abundance types. The index within sample between two abundance
511 type were connected to illustrate the change trend of the indices, the asterisks representing significantly
512 differences are based on paired Wilcoxon test, “***” refers to $P < 0.001$.

513

514 **Figure 6. Ordination analyses of simulated profiles based on rJSD.** Scatter plots of NMDS, PCoA,
515 t-SNE and UMAP illustrate the dissimilarities between the sequence abundance (red dots) and
516 taxonomic abundance (blue dots), which are the ground truth of the simulated 100 gut profiles. Root
517 Jensen-Shannon divergence (rJSD) was used to for the ordination analyses. The plots of the ordination
518 analyses based on sequence abundance and taxonomic abundance were adjusted to overlap with each
519 other first, then the similarity was calculated by the Monte-Carlo test. The two abundance types from
520 the same profile were connected using grey lines to show the change of its position.

Figures

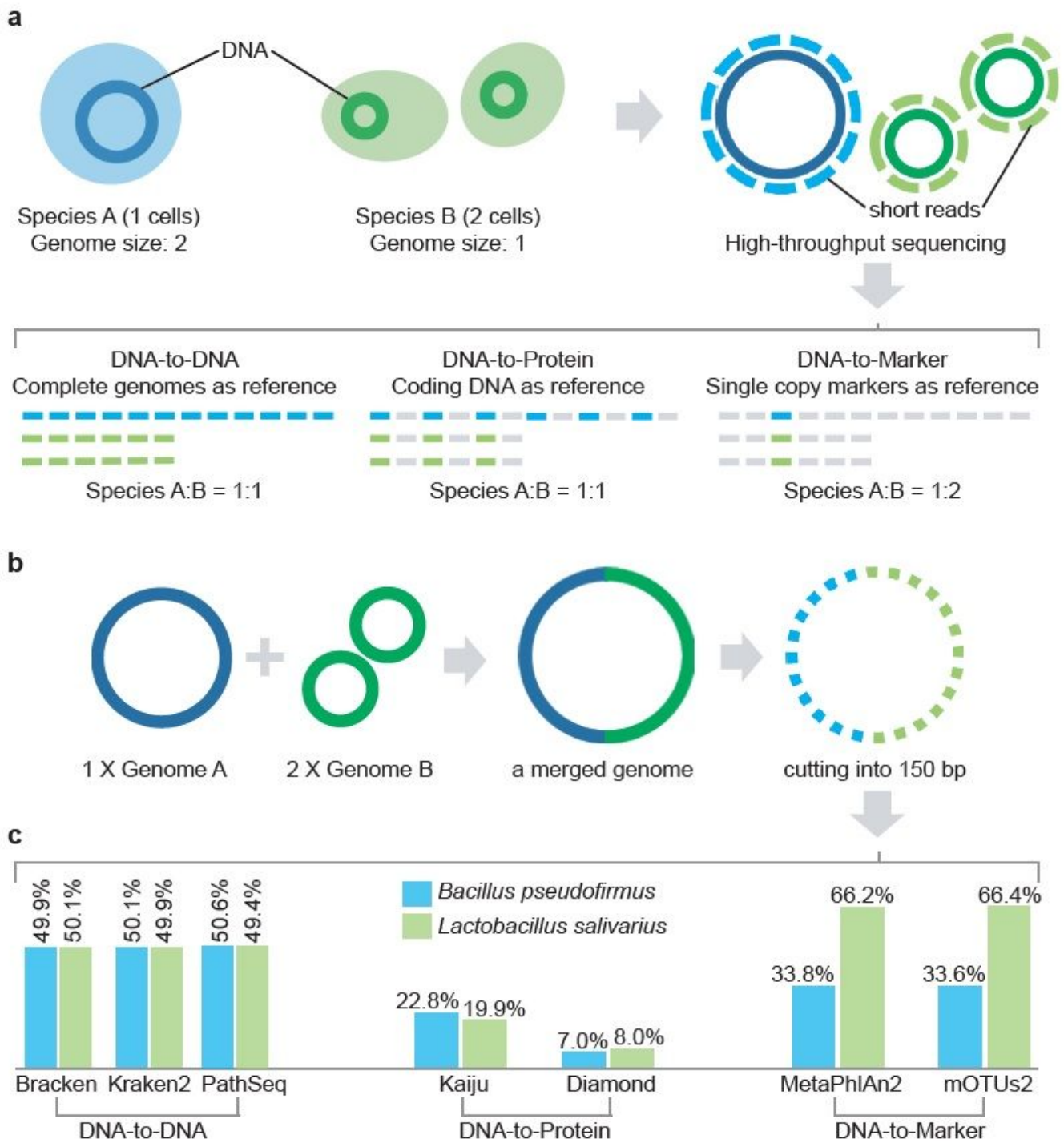


Figure 1

Comparison of profiling results. a, Illustration of the reference databases and the default output abundance type for DNA-to-DNA, DNA-to-Protein and DNA-to-Marker profilers on a mixture of two species A (1 cell) and B (2 cells). b, A simulated microbial community with only two genomes: *Bacillus pseudofirmus* (genome size 4.2MB) and *Lactobacillus salivarius* (genome size 2.1MB). We merged one

copy of *Bacillus pseudofirmus* genome (genome A) with two copies of *Lactobacillus salivarius* genome (genome B) sequences into one metagenome file. Then we sheared the merged metagenomic sequences into 150bp to simulate a typical metagenomic dataset. c, Profiling results (default output) of different profilers for the simulated microbial community shown in a. The bar plots show the estimated relative abundance of the two microbial members A and B using different metagenomics profilers.

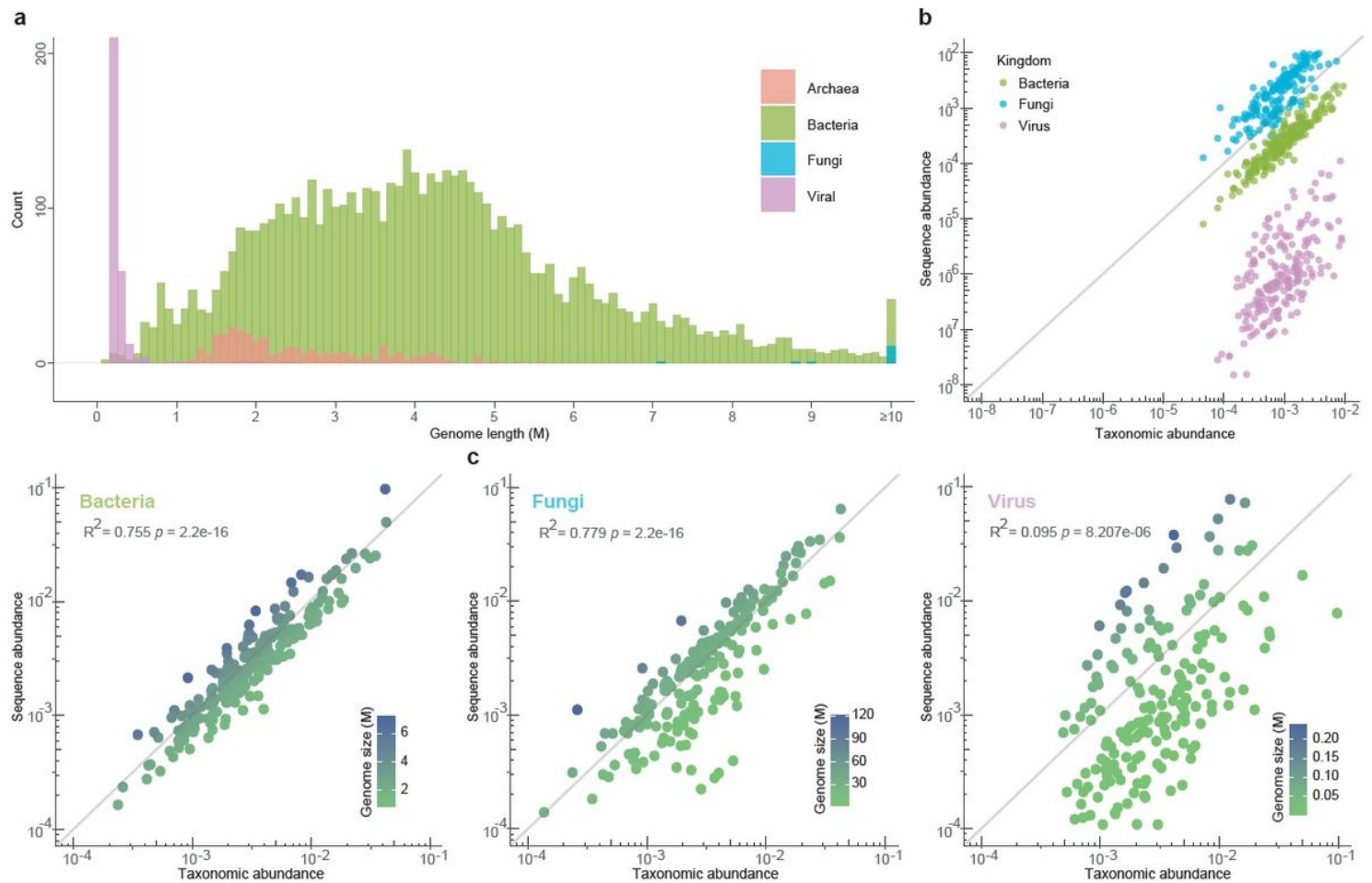


Figure 2

Correlation between sequence size abundance and taxonomic abundance in synthetic profiles based on different kingdoms. a, Genome size distribution of microorganisms calculated from the microbial genome database (NCBI RefSeq 2020 Nov 6th) that includes 171,927 bacteria, 293 fungi, 945 archaea, and 9,362 viruses. b, The scatter plot shows the correlation between taxonomic abundance (x axis) and sequence abundance (y axis) of 600 randomly selected species in a simulated profile which includes bacteria (species number=200), fungi (species number=200) and virus (species number=200). c, Each scatter plot shows the correlation between taxonomic abundance (x axis) and sequence abundance (y axis) of 200 randomly selected species in three simulated profiles which represent different kingdoms e.g. bacteria, fungi, and virus.

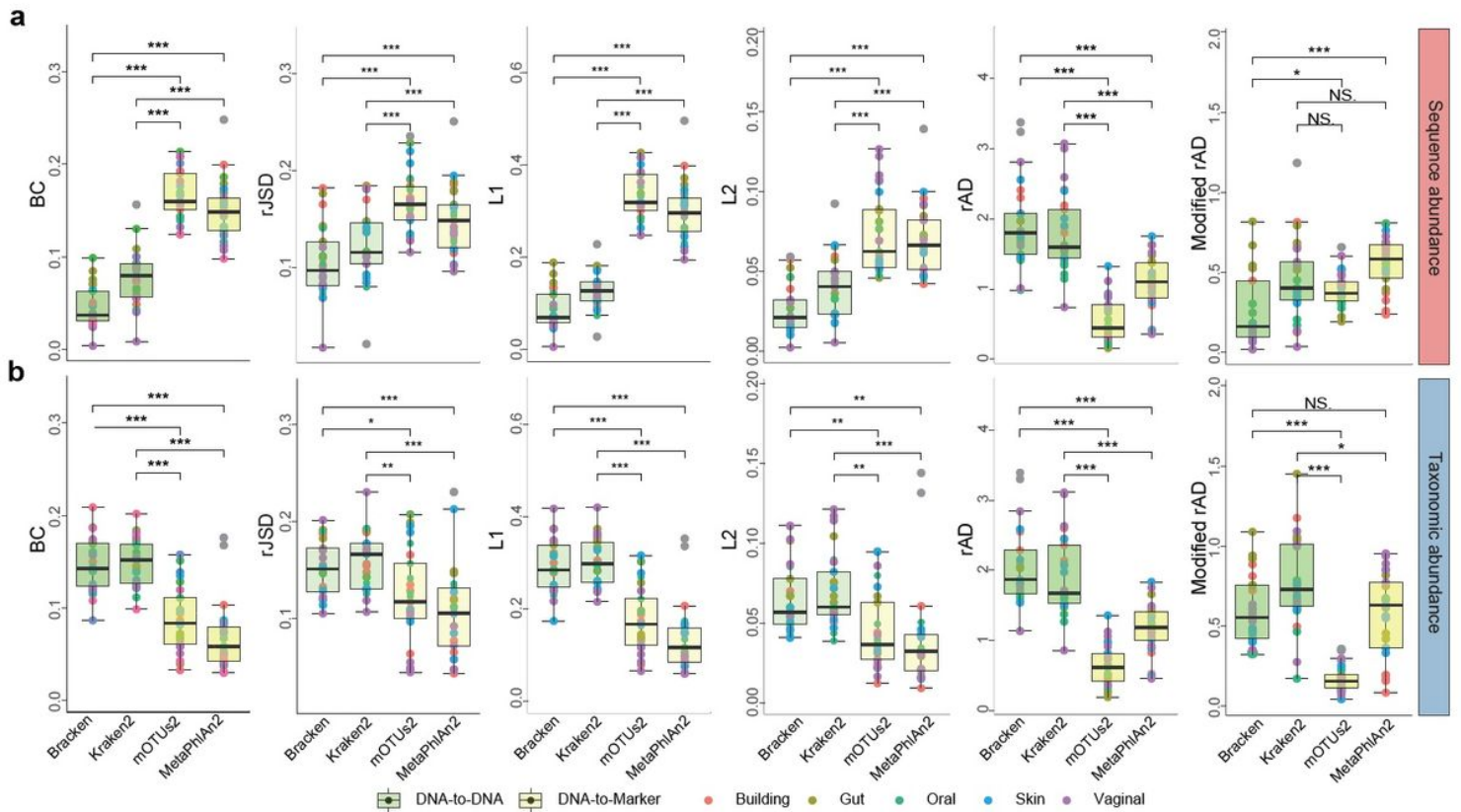


Figure 3

Differential benchmarking results of four representative metagenomics profilers using two types of relative abundance as ground truth: a, sequence abundance and b, taxonomic abundance. The boxplots indicate the dissimilarities based on L1, L2, root Jensen-Shannon divergence (rJSD), Bray-Curtis (BC), and robust Aitchison distance (rAD) between the ground-truth profiles and the profiles predicted by different metagenomics profilers (Bracken, Kraken2, mOTUs2, and MetaPhlan2) at the species level. For each metagenomic profiler, we performed the dissimilarity calculations based on 25 simulated microbial communities from five representative environmental habitats (gut, oral, skin, vagina and building) separately. Note that for each profiler based on any evaluation metric, its performance variation across different synthetic communities is due to microbiome complexity difference (e.g. species composition and richness). The asterisks in the boxplots refer to the statistical significance: “*” refers to p -value < 0.05 , “**” refers to < 0.01 , “***” refers to < 0.001 .

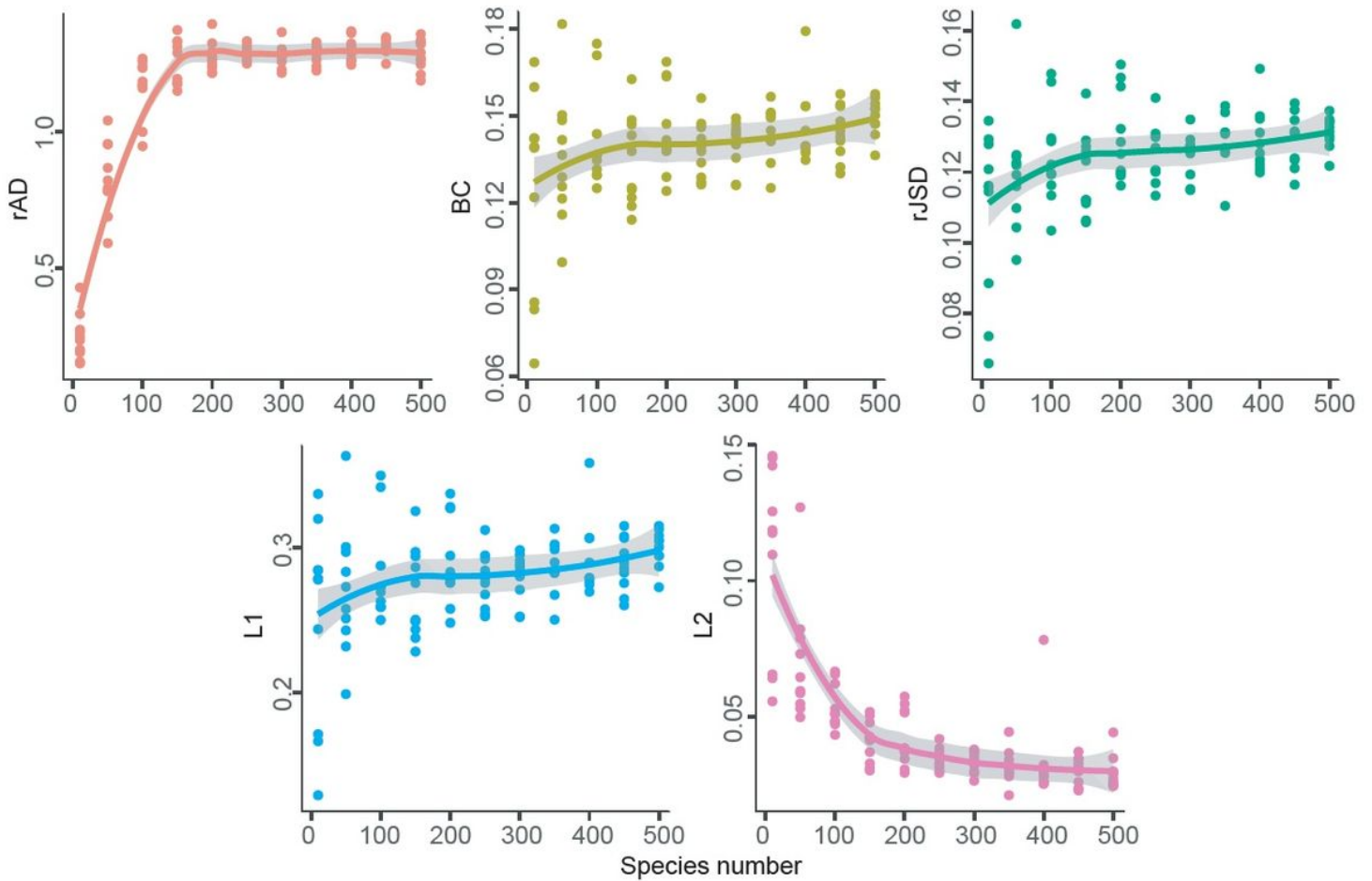


Figure 4

Dissimilarity between sequence abundance and taxonomic abundance with varied species number measured by different distance measures. For each species number, we simulated 10 repeats of profiles. The distance/dissimilarity was then measured by different measures: rAD (red), L1 (blue), L2 (purple), Bray-Curtis (yellow) and rJSD (green). rAD between these types of abundance profiles positively correlated with the species richness when < 200 microbial species presented in a community, yet saturated after the number of species reaching 200. L1, BC and rJSD can also reveal the difference between the two abundance types yet they were not affected by the species-level richness. L2 distance between the two abundance types dramatically dropped with the increase in the species-level richness. In the complex community with the number of species over 200, L2 distance metric almost lost the discriminatory power of these two abundance profiles.

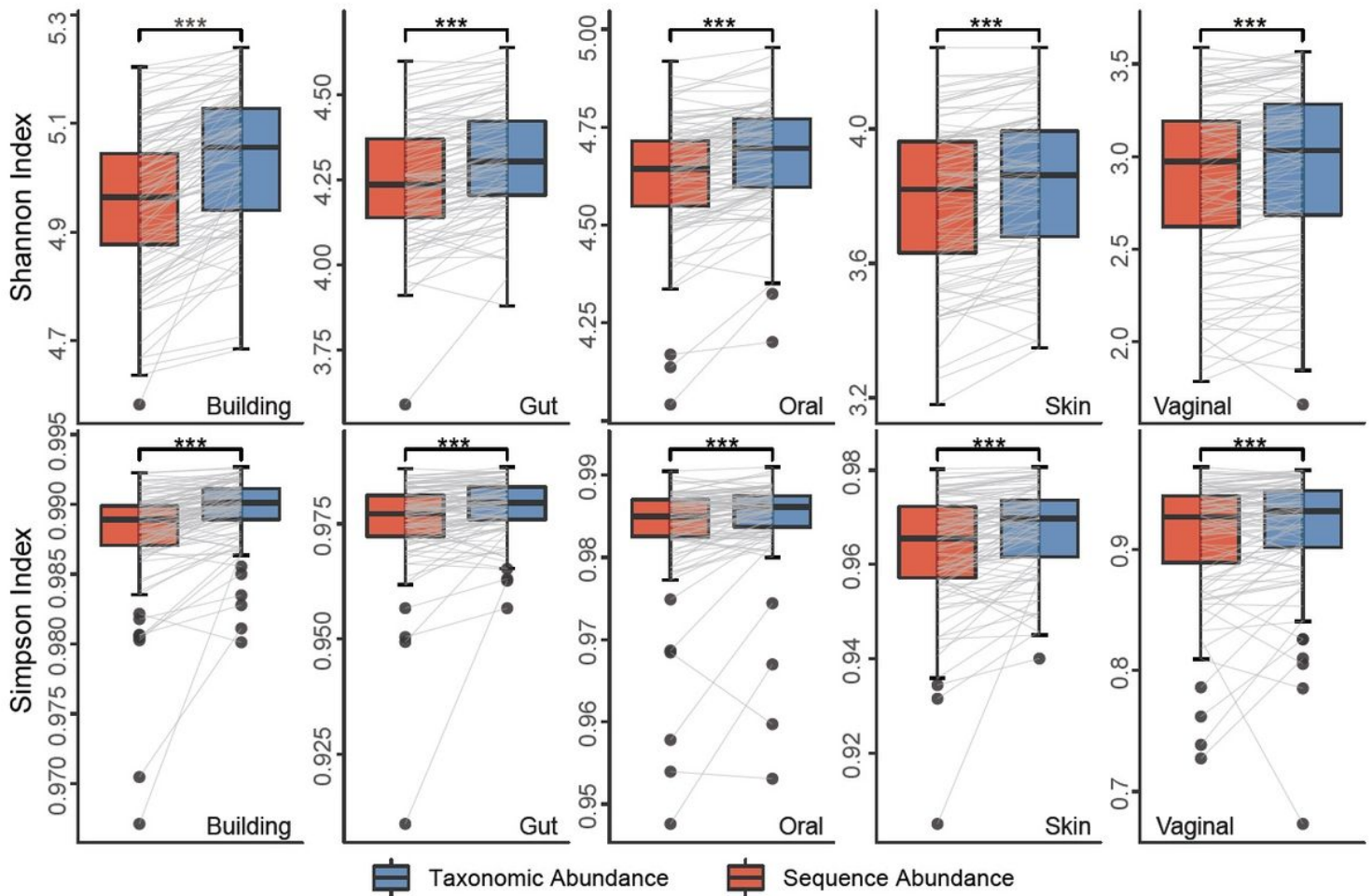


Figure 5

Alpha diversity based on sequence abundance and taxonomic abundance. Alpha diversity (Shannon index and Simpson index) based on ground truth of simulated data from different habitats revealed the influence of abundance types. The index within sample between two abundance type were connected to illustrate the change trend of the indices, the asterisks representing significantly differences are based on paired Wilcoxon test, “***” refers to $P < 0.001$.

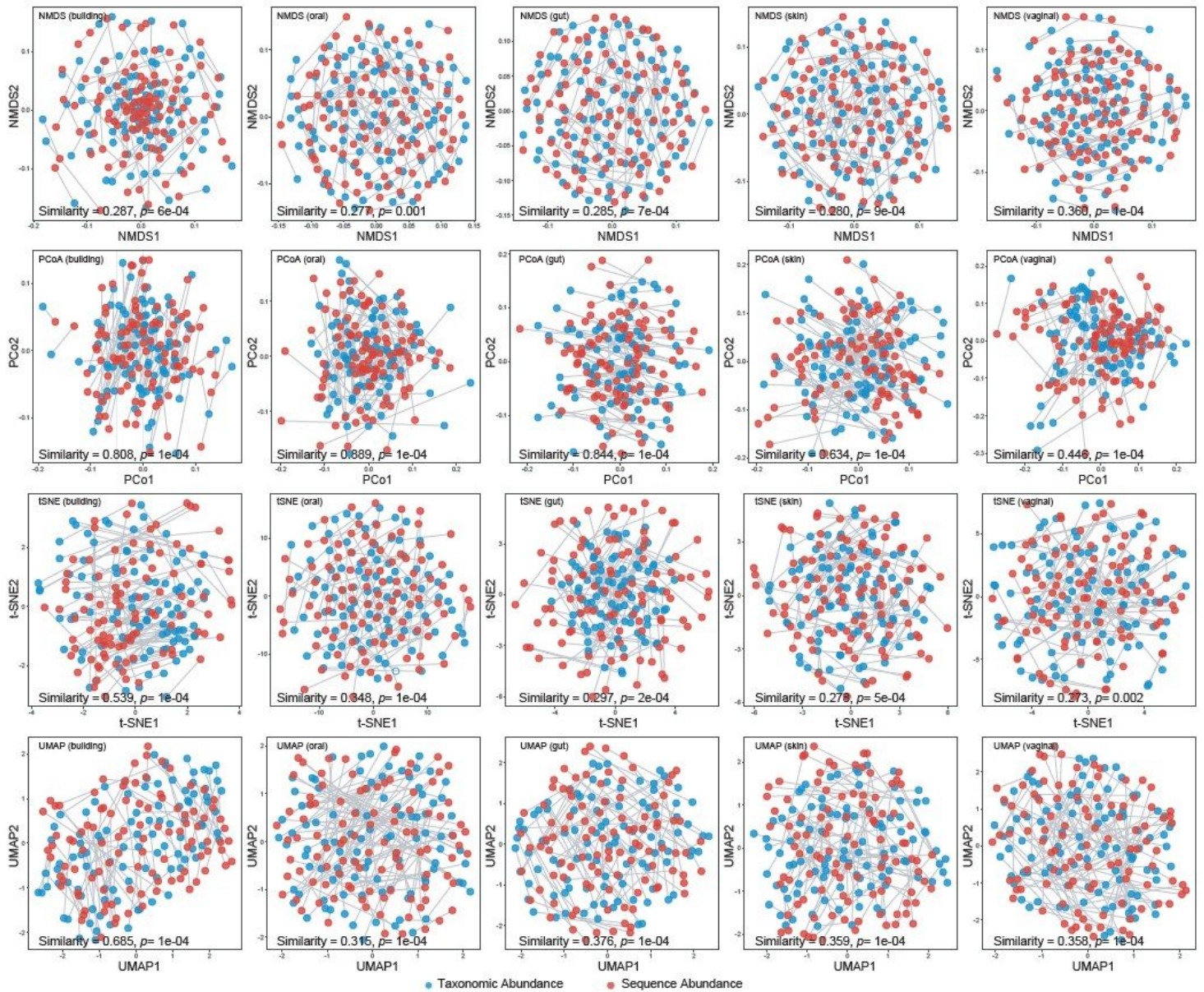


Figure 6

Ordination analyses of simulated profiles based on rJSD. Scatter plots of NMDS, PCoA, t-SNE and UMAP illustrate the dissimilarities between the sequence abundance (red dots) and taxonomic abundance (blue dots), which are the ground truth of the simulated 100 gut profiles. Root Jensen-Shannon divergence (rJSD) was used for the ordination analyses. The plots of the ordination analyses based on sequence abundance and taxonomic abundance were adjusted to overlap with each other first, then the similarity was calculated by the Monte-Carlo test. The two abundance types from the same profile were connected using grey lines to show the change of its position.

Supplementary Files

This is a list of supplementary files associated with this preprint. [Click to download.](#)

- [SupplementaryMaterials20201114.pdf](#)