

# Impact of simulation and reference catalogues on the evaluation of taxonomic profiling pipelines

Vadim Puller

GMT Science

Florian Plaza Oñate

GMT Science

Edi Prifti

IRD

Raynald de Lahondès (✉ [raynald.delahondes@gmt.bio](mailto:raynald.delahondes@gmt.bio))

GMT Science <https://orcid.org/0009-0000-2862-9589>

---

## Article

**Keywords:** taxonomic profiling

**Posted Date:** October 27th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3433959/v2>

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

[Read Full License](#)

**Additional Declarations:** Yes there is potential Competing Interest. V.P., F.P.O. and R. de L. are employees of GMT Science, E.P. is a scientific advisor to GMT Science.

---

# Impact of simulation and reference catalogues on the evaluation of taxonomic profiling pipelines

Vadim Puller<sup>1</sup>, Florian Plaza Oñate<sup>1</sup>, Edi Prifti<sup>2,3\*</sup>  
and Raynald de Lahondès<sup>1\*</sup>

<sup>1</sup>GMT science, 75 route de Lyons-La-Foret, Rouen, F-76000, France.

<sup>2</sup>IRD, Sorbonne Université, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, UMMISCO, 32 avenue Henri Varagnat, Bondy, F-93143, France.

<sup>3</sup>Sorbonne Université, INSERM, Nutrition et Obesities ; systemic approaches, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière, 91 boulevard de l'Hôpital, Paris, F-75013, France.

\*Corresponding author(s). E-mail(s): [edi.prifti@ird.fr](mailto:edi.prifti@ird.fr);

[raynald.delahondes@gmt.bio](mailto:raynald.delahondes@gmt.bio);

Contributing authors: [vadim.puller@gmt.bio](mailto:vadim.puller@gmt.bio);

[fplazaonate@gmt.bio](mailto:fplazaonate@gmt.bio);

## Abstract

Microbiome profiling tools rely on reference catalogues, which significantly affect their performance. Comparing them is however challenging, mainly due to differences in their native catalogues. In this study, we present a novel standardized benchmarking framework that makes such comparisons more accurate. Specifically, we conducted three realistic simulations of gut microbiome samples, each based on a specific taxonomic profiler, and used two taxonomic references to project their results: the Genome Taxonomy Database and the Integrated Gene Catalogue. To demonstrate the importance of using such framework, we evaluated four established profilers along with a novel one, BiomScope. We evaluated the impact of these simulations and that of the common taxonomic references on the perceived performance of these profilers and provide guidelines.

## Introduction

Over the past decade, the field of microbiome research has grown substantially. A series of landmark publications has firmly established connections between the microbial ecosystem and numerous human diseases. Notably, the gut microbiome has gained prominence as a critical organ and as a sentinel for human diseases [1–9]. Furthermore, research has explored the predictive potential of the gut microbiome, not only for disease diagnosis but also for assessing disease severity [10–12].

Characterizing microbiome samples primarily involves identifying the various species and strains present, along with quantifying their abundance as well as their functional capabilities. Numerous methodologies have been devised to address this challenge. They notably diverge in their sequencing strategies: some are designed to target highly conserved genomic regions, such as the 16S rRNA genes, while others employ a broader sequencing approach known as whole metagenomic sequencing (WMS). Each of these sequencing approaches imposes distinct analytical requirements. This study focuses on the analytical methods tailored for WMS data.

The manipulation of WMS data presents several complex challenges owing to their inherent properties such as sparsity, compositionality, and variable interdependence. Consequently, a multitude of WMS taxonomic profilers are available to address these complexities, each offering distinct advantages and limitations.

Irrespective of the specific method employed, taxonomic profilers designed for WMS data have a common reliance on reference catalogues. These catalogues have evolved in terms of methodology, size, and quality over the past decade. To illustrate, the initial gene reference catalogue for the human gut microbiome was introduced in 2010 [13] and has since expanded exponentially from a mere couple of million genes to tens of millions [14], and eventually to hundreds of millions of genes [15]. The diversity in catalogues across the literature highlights the dynamism of the field, yet also presents challenges in comparing results across different studies.

Some pipelines assign sequenced reads to representative genome catalogues such as Kraken [16–18], Centrifuge [19], Kaiju [20] or DIAMOND [18]. These pipelines are typically compatible with different genomic databases, notably RefSeq [21, 22], Genome Taxonomy Database (GTDB) [23–26] and MGnify [27] (See Supplementary Tables S1 and S2 for more complete summaries of the popular metagenomic tools and recent benchmarking studies).

Another class of taxonomic profilers relies on tailor-made catalogues of marker genes such as MetaPhlAn or mOTUs [28–34]. These tools benefit from their selection of marker genes and associated taxonomy. However, they may underperform in specific scenarios, such as simulations involving species poorly or not represented in their catalogues [35–37].

It is worth emphasizing that most commonly used metrics for assessing species presence (e.g., those derived from confusion matrices [36]) and metrics for comparing abundance profiles (e.g., various distance measures [36]),

assume that the compared tools operate within the same "feature" space. In this context, we use the Machine Learning term "feature", to denote the various taxonomic units (such as species, OTUs, clades, MSPs, etc.) employed by different profilers to describe sample composition.

Novel microbiome profiling pipelines and methods are systematically compared with existing ones using either mock communities or simulated samples by tools such as CAMISIM [38]. Simulations remain inherently reliant on the input genomes and specified abundances. The realism of such simulations further hinges on several factors, including the number of distinct species (richness of the ecosystem), the specific composition in terms of biome specialization, genetic distance between species, and the distribution of their relative abundance.

When comparing pipelines, the diversity of reference catalogues also poses a specific challenge: three distinct approaches emerge. The first involves changing default pipeline catalogues to a common one. An alternative approach conducts comparisons at different taxonomic levels (e.g., species, genus or phylum), where disparities between catalogues tend to be less pronounced [39–42]. In the third approach, pipelines use their native catalogues but project their results onto a shared feature space.

In this paper, we introduce a standardized benchmarking approach using shared feature space projections. Our objective was to take advantage of the capabilities of simulation, while maintaining a close semblance to actual gut microbial samples obtained from colorectal cancer patients and control subjects (n=343) [11, 12]. To achieve this, we applied three quantification pipelines to real samples, generating realistic community descriptions for subsequent simulations. These three pipelines are each representative of a specific class of profilers: Kraken for the genome catalogue-based pipelines, MetaPhlAn for the marker gene catalogue-based pipelines and a novel one, BiomScope, developed by our research group, relying on a non-redundant gene catalogue [43, 44]. Non-redundant gene catalogue pipelines, such as Meteor [45, 46], are similar to marker gene catalogue pipelines, but may take advantage of non-marker genes to provide a more comprehensive description. These real samples served as reference abundance profiles for three distinct simulation *scenarii*. Simulated samples were then analysed using a panel of five different profilers, including the three used for the simulation, to estimate the distribution of taxonomic profiles and the impact of feature spaces. These results were then compared with the reference abundance profiles employed during the simulations. Our evaluation of these tools encompassed a range of standard performance metrics, including assessments of alpha and beta diversity disparities, as well as precision and sensitivity analyses.

## Results

### Features lost in projection

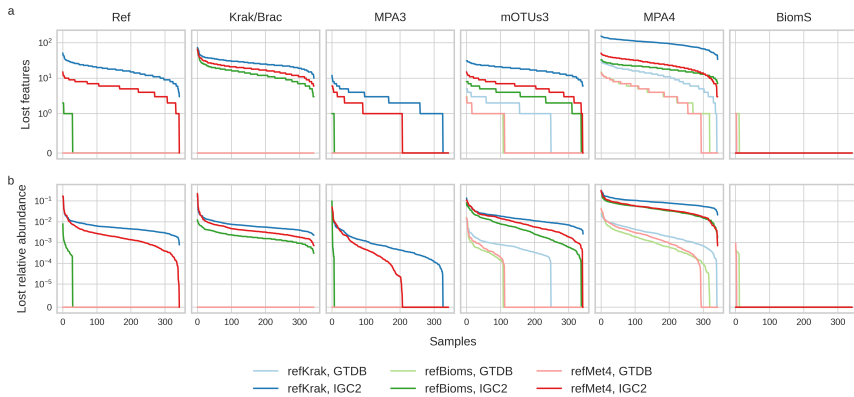
We encountered several notable challenges in our examination of microbiome profiling tools, each operating with distinct reference catalogues and native feature spaces. The primary objective was to assess the comparative performance of four well-established profiling tools (mOTUs3, Kraken/Bracken, MetaPhlAn3, MetaPhlAn4) alongside a novel tool, BiomScope, using simulated metagenomics data. These selected tools represent a subset of a larger pool of available options (see supplementary table S1). We observed that no single common feature space could impartially represent all profiling tools, thereby inevitably introducing biases that could favor some tools over others. Additionally, we noted that the choice of the tool used to generate the simulation could further influence the performance assessment. To ensure a fair and comprehensive evaluation of these tools, we meticulously considered these influential factors. Consequently, we devised an experimental framework encompassing three distinct simulations, each driven by reference abundance data generated by a different tool (refKrak for Kraken, refMet4 for MetaPhlAn4, refBioms for BiomScope), in conjunction with two common feature spaces, GTDB and IGC2, resulting in a total of six distinct projections (see Methods).

We initiated our assessment by examining the impact of projections from each native feature space onto the two common feature spaces. Our initial observation revealed that even before evaluating the profiling tools, certain features were lost from the reference profiles used for simulating the sequence data (Figure 1a,b facet *Ref*). Notably, the IGC2 common feature space experienced the highest loss of features from the native space. This was primarily due to the specialized nature of IGC2, which comprises a smaller number of features (*i.e.*,  $n=1990$  species) compared to GTDB (*i.e.*,  $n=317,542$  species). Furthermore, we noted that refBioms exhibited the least impact from feature loss, followed by refMet4 and refKrak. This same pattern extended to the loss of relative abundance from the native feature space to the common spaces.

Likewise, we compared for each tool the number of lost features and their associated lost abundance after projecting them onto common feature spaces (1a, b facets 2:6). The results were similar to the trend seen above. The IGC2 projections displayed the most pronounced effects, particularly impacting mOTUs and MetaPhlAn4. For MetaPhlAn3 and Kraken, some lost features were also observed in the GTDB feature space, albeit to a lesser degree. Notably, BiomScope experienced the smallest number of lost features across all three distinct reference simulations.

### Richness and Shannon diversity estimation

Richness and Shannon diversity are two widely used metrics in microbiome studies to estimate the complexity of the studied ecosystems. Microbiome



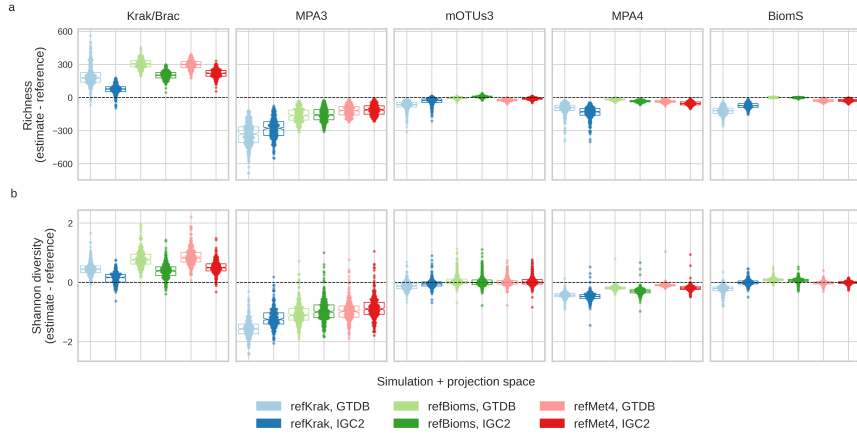
**Figure 1** The number of features (a) and the relative abundance (b) lost when projecting the abundance tables from the native feature space to the two common feature spaces for the simulation source (i.e. Ref) and the five tools. Samples are ordered independently according to the number of lost species and abundance for each of the simulations and common feature space.

diversity is shown to play an important role not only in the health of the ecosystem but also associates with a healthier phenotype of the host [10, 47]. The simulation-based approach used here allowed us not only to precisely compare the performance of the five studied profilers in terms of accuracy but also to evaluate the impact of the common features spaces. Our observations revealed notable discrepancies among the tools. Specifically, the Kraken pipeline consistently exhibited a significant overestimation of both richness and Shannon diversity, while MetaPhlAn3 exhibited underestimations. Comparatively, the remaining tools (mOTUs, MetaPhlAn4, and BiomScope) demonstrated smaller deviations from the ground truth. They performed most effectively in the refBiomS simulation and displayed their poorest performance in the refKrak simulation, potentially due to the heightened complexity of the latter (see Figure 2). BiomScope and mOTUs exhibited the highest accuracy, albeit with mOTUs displaying somewhat greater variability.

## Abundance estimation performance

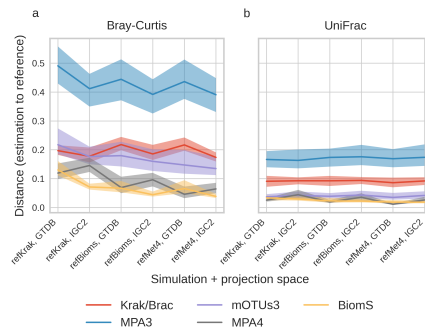
As microbiome data inherently exhibit a compositional nature, the precise estimation of species proportions remains a pivotal requirement for microbiome profiling tools. We examined the impact of simulations and common feature spaces on the tools' ability to provide accurate abundance estimation. This assessment was based on two widely used distance metrics, namely, Bray-Curtis and the weighted UniFrac [48, 49], which quantify the dissimilarity between the estimated abundance profiles and the ground truth (see Figure 3).

MetaPhlAn3 displayed the highest level of discrepancy in accurately estimating species proportions, followed by Kraken and mOTUs, while MetaPhlAn4 and BiomScope performed better. In terms of UniFrac distances,

6 *Impact of simulation on taxonomic profiling pipeline evaluation*

**Figure 2** (a) Difference of the estimated species richness with the reference richness depicted by boxplots and jittered points. (b) Difference of the estimated Shannon diversity with the reference Shannon diversity depicted by boxplots and jittered points.

Kraken consistently lagged behind mOTUs, which approached the performance levels of MetaPhlan4 and BiomScope. Notably, UniFrac distances minimized differences compared to Bray-Curtis distances across simulations and common feature spaces, as seen in Supplementary Figure S1. This suggests that incorrect species identifications often involved phylogenetically closely related species, especially for mOTUs.

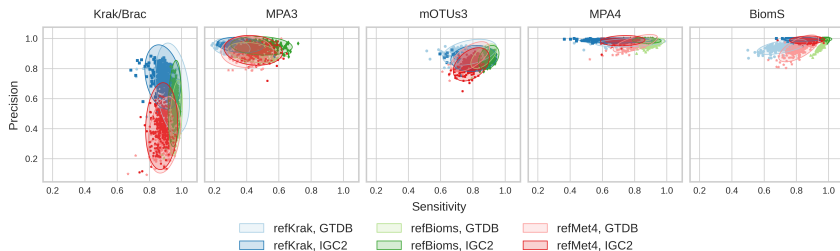


**Figure 3** Bray-Curtis (a) and UniFrac (b) pairwise distances between the estimated and the reference abundance profiles, displayed as mean  $\pm$  standard deviation. Colors depict the five different tools.

## Impact of simulation in species discovery and abundance estimation

After the overall estimation of relative abundance, we evaluated the impact of the simulation and the common feature spaces on the ability of the studied

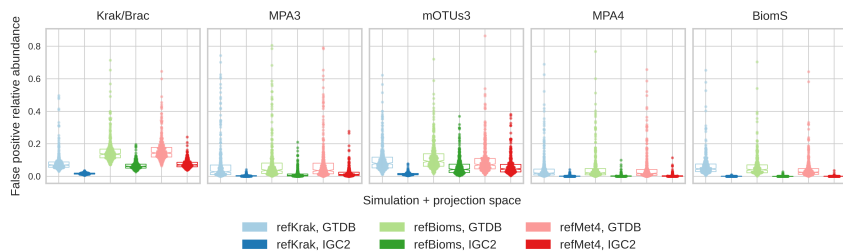
tools to correctly discover the abundance and presence of the microbial species. Sensitivity and precision, depicted in Figure 4, respectively measure the tools' ability to identify truly present species and avoid false positives, *i.e.* species erroneously detected by a pipeline. Kraken emerged as one of the most sensitive tools, indicating a lower rate of false negatives, although its specificity (i.e. the number of false positives), was notably impacted by the simulation, particularly in the case of refMet4 and refBioms. However, the influence of the common feature space was less pronounced. In contrast, MetaPhlAn3 exhibited lower sensitivity, although its precision remained high. mOTUs demonstrated lower precision than MetaPhlAn4 and BiomScope but struck a balance between false positives and false negatives. MetaPhlAn4 and BiomScope, on the other hand, leaned towards precision over sensitivity with overall very good performance. BiomScope, however, displayed a greater susceptibility to the impact of the simulation, particularly benefiting from its native reference simulation or when projecting on its native space.



**Figure 4** Sensitivity and Precision evaluation of the five tools and the impact of the simulations and common feature spaces. Each point corresponds to a sample, in the simulation defined by the color scheme in the legend. Error ellipses indicate standard deviation of the sample distribution.

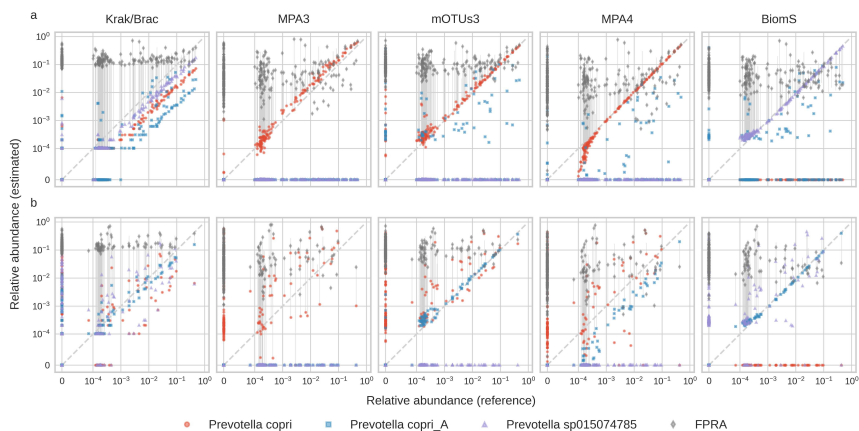
Moreover, we evaluated the impact of the simulation on the compared tools, on their ability to estimate species abundance. For this we computed the false positive relative abundance (FPRA). We observed a significant number of outlier samples with FPRA as high as 80%, but only when using the GTDB feature space (see Figure 5).

This high error rate is primarily due to a small number of closely related species that were inverted, which is consistent with our earlier observations regarding the UniFrac vs. Bray-Curtis distances. An example of such confusion is illustrated in Figure 6, where the estimated abundance of *Prevotella* sp.015074785, *Prevotella copri* and *Prevotella copri* A are compared with the abundance of these species in the refBioms simulation. *P. copri* was not present in this simulation, as it did not exist in the UHGG collection (see Methods). With the exception of MetaPhlAn3, all the pipelines identified correctly *P. copri* A, which was present in the catalogues of all tools except MetaPhlAn3. On the other hand, *P. sp.015074785* was misidentified by both

8 *Impact of simulation on taxonomic profiling pipeline evaluation*

**Figure 5** False positive relative abundance, computed between the estimated abundance and the ground truth. colors indicate the different simulations and common feature spaces across the five profiling tools.

MetaPhlAns and mOTUs as *P. copri*. It was also correctly identified by BiomScope, although its catalogue did not contain a species corresponding to *P. copri* of GTDB, only those mapped to *P. sp.015074785* and *P. copri* A. Kraken found all three species in correlated proportions. However, none of them was equal to their reference abundances.



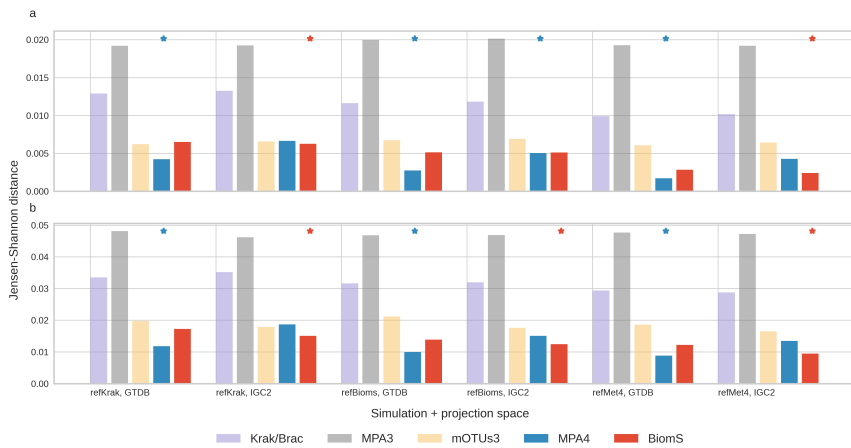
**Figure 6** Estimated abundance of three different *Prevotella* species and the total false positive relative relative abundance vs. the relative abundance in reference of (a) *Prevotella* sp.015074785 and (b) *Prevotella copri* A (data for the refBioms simulation).

## Community structure

A frequently used method for representing community structure involves conducting a principal coordinates analysis (PCoA) of the pairwise distance matrix. Supplementary Fig S2 illustrates the first two components of this decomposition for MetaPhlAn4 and BiomScope. However, since the initial principal components only capture a portion of the information embedded in

the pairwise distance matrix, we opted for a direct comparison of these matrices. This comparison involved calculating pairwise distance matrices for the reference simulation profiles and the estimations provided by each tool.

By employing the Jensen-Shannon distance (JSD) to measure dissimilarity between the ground truth and the estimates, we noted that the top-performing tools were MetaPhlAn4 and BiomScope, with mOTUs, Kraken, and MetaPhlAn3 following in that order. While MetaPhlAn4 demonstrated slightly superior performance when using the Bray-Curtis distance, it was on par with BiomScope when UniFrac was employed. Notably, BiomScope consistently outperformed when benefiting from its native feature space, as illustrated in Figure 7.



**Figure 7** Similarity between pairwise distance matrices for the simulated data and the estimations. (a) Bray-Curtis pairwise distances, (b) weighted UniFrac pairwise distances. Best tools are annotated with a '\*'. \*

## Summary tables and tool ranking

Table 1 summarizes the median values (median across the samples) of the abundance and features lost in projecting onto the common feature space, as well as the abundance and Shannon diversity of the projected data. Table 2 summarizes the median values of the results presented above. On the basis of different metrics the tools were ranked as follows:

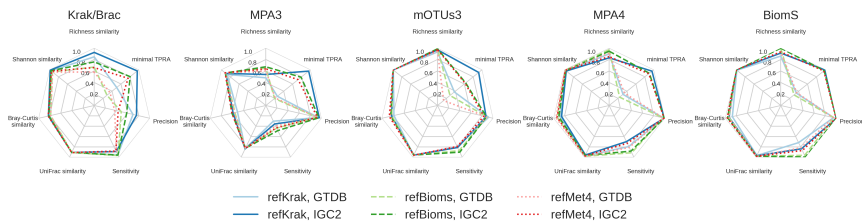
1. *Richness estimate.* Kraken consistently overestimated species richness, whereas MetaPhlAn3 systematically underestimated it. BiomScope and MetaPhlAn4 tended to underestimate richness, with the former being closer to the ground truth. mOTUs was the best in this category.
2. *Shannon diversity.* Kraken overestimated the Shannon diversity, while MetaPhlAn3 underestimated it. MetaPhlAn4 also underestimated it, while mOTUs and BiomScope performed on the same level, with BiomScope

showing fewer errors across samples, while mOTUs was less sensitive to the choice of the simulation and the projection space.

3. *Bray-Curtis distance*. MetaPhlan3 was furthest from the ground truth, followed by Kraken and mOTUs, with BiomScope and MetaPhlan4 being closest to the ground truth.
4. *UniFrac distance*. MetaPhlan3 was furthest from the ground truth, followed by Kraken. mOTUs, BiomScope, and MetaPhlan4 performed equally.
5. *Sensitivity*. Kraken had the best sensitivity, surpassed by BiomScope only when the latter was given the advantage of the native simulation feature space. For other simulations BiomScope performed similarly to MetaPhlan4 and mOTUs. MetaPhlan3 had the lowest sensitivity (it generated more false negatives.)
6. *Precision*. BiomScope and MetaPhlan4 were the most precise tools, followed by MetaPhlan3 (due to its small catalogue), mOTUs, and then by Kraken (producing more false positives.)
7. *False positive relative abundance (FPRA)*. Kraken had the greatest abundance of false positives, followed by mOTUs and MetaPhlan3. BiomScope and both MetaPhlans performed similarly. FPRA was significantly lower in the IGC2 feature space.

## Best profilers on our radar

Fig. 8 and Fig. S3 compare the performance of the profiling tools studied across different metrics, simulations, and projection spaces. To facilitate the representation we calculated Bray-Curtis similarity for richness and Shannon diversity estimates (i.e., we present  $1 - |u - v| / (u + v)$ , where  $u, v$  are the true and the estimated richness/Shannon diversity). Bray-Curtis and UniFrac distances to the ground truth were also transformed to similarities as  $1 - d_{BC}$ ,  $1 - d_{WUF}$ . Likewise, instead of FPRA we took the minimal abundance of false positives across the samples,  $TPRA = 1 - FPRA$  (For all other metrics we present the median value.) Finally, no transformation was necessary for Sensitivity and Precision, which naturally change between 0 and 1, with higher value corresponding to the best results.



**Figure 8** Comparison of tool performances across different metrics, simulations, and projection spaces.

**Table 1** Median values (across the samples) of lost abundance, lost features, richness, and Shannon diversity

Tool	Simulation	Common space	Lost abundance	Lost clades	Richness	Shannon
Ref	refMet4	GTDB	0.0000	0	214	5.5574
Ref	refMet4	IGC2	0.0017	5	211	5.5539
Ref	refBioms	GTDB	0.0000	0	308	5.8201
Ref	refBioms	IGC2	0.0000	0	312	5.8735
Ref	refKrak	GTDB	0.0000	0	489	6.6404
Ref	refKrak	IGC2	0.0050	17	447	6.4082
Krak/Brac	refMet4	GTDB	0.0000	0	521	6.4017
Krak/Brac	refMet4	IGC2	0.0035	18	433	6.0816
Krak/Brac	refBioms	GTDB	0.0000	0	624	6.5778
Krak/Brac	refBioms	IGC2	0.0018	13	524	6.2752
Krak/Brac	refKrak	GTDB	0.0000	0	692	7.0674
Krak/Brac	refKrak	IGC2	0.0059	26	530	6.5234
MPA3	refMet4	GTDB	0.0000	0	91	4.5695
MPA3	refMet4	IGC2	0.0001	1	99	4.6506
MPA3	refBioms	GTDB	0.0000	0	147	4.7341
MPA3	refBioms	IGC2	0.0000	0	159	4.8901
MPA3	refKrak	GTDB	0.0000	0	159	5.0266
MPA3	refKrak	IGC2	0.0006	2	167	5.1590
mOTUs	refMet4	GTDB	0.0000	0	191	5.5497
mOTUs	refMet4	IGC2	0.0075	6	204	5.5394
mOTUs	refBioms	GTDB	0.0000	0	303	5.8622
mOTUs	refBioms	IGC2	0.0039	3	325	5.8636
mOTUs	refKrak	GTDB	0.0004	1	420	6.4917
mOTUs	refKrak	IGC2	0.0130	18	416	6.3575
MPA4	refMet4	GTDB	0.0014	4	176	5.4500
MPA4	refMet4	IGC2	0.0424	25	154	5.3320
MPA4	refBioms	GTDB	0.0008	4	289	5.6378
MPA4	refBioms	IGC2	0.0389	19	278	5.5678
MPA4	refKrak	GTDB	0.0025	12	392	6.2038
MPA4	refKrak	IGC2	0.0859	100	309	5.9547
BiomS	refMet4	GTDB	0.0000	0	184	5.5273
BiomS	refMet4	IGC2	0.0000	0	185	5.5273
BiomS	refBioms	GTDB	0.0000	0	309	5.9148
BiomS	refBioms	IGC2	0.0000	0	311	5.9173
BiomS	refKrak	GTDB	0.0000	0	371	6.3988
BiomS	refKrak	IGC2	0.0000	0	375	6.4054

## Discussion

The benchmarking approach we designed and explored in this paper revolves around the concept of common feature spaces and the impact of simulation on various microbiome profiling tools. We implemented different standardized metrics to compare their ability in estimating the abundance and prevalence of microbiome features. The general outline of the results is consistent with the expectations based on previous benchmarking studies [35–37, 39–42, 50–55] and the tools’ design. Kraken for instance is known for its high sensitivity while generating an excessive number of false positives, whereas MetaPhlan3 is known for having a relatively poor native feature space resulting in an excess number of false negatives.

The loss of a certain number of features and their corresponding abundance in the projections remains relatively low across all tools (see Figure 1). This

**Table 2** Median values (across the samples) of major metrics (error in estimated richness and Shannon diversity, Bray-Curtis and UniFrac distances between the estimations and the reference, sensitivity, precision, false positive relative abundance (FPRA)).

Tool	Simulation	Space	Richness diff.	Shannon diff.	Bray-Curtis	UniFrac	Sensitivity	Precision	FPRA
Krak/Brac	refMet4	GTDB	299	0.8331	0.2171	0.0863	0.8966	0.3659	0.1436
Krak/Brac	refMet4	IGC2	220	0.4961	0.1738	0.0924	0.8793	0.4289	0.0708
Krak/Brac	refBioms	GTDB	306	0.7643	0.2182	0.0929	0.9655	0.4848	0.1363
Krak/Brac	refBioms	IGC2	202	0.3875	0.1858	0.0943	0.9577	0.5831	0.0616
Krak/Brac	refKrak	GTDB	179	0.4414	0.1972	0.0912	0.9517	0.6966	0.0698
Krak/Brac	refKrak	IGC2	77	0.1645	0.1778	0.0931	0.8887	0.7588	0.0167
MPA3	refMet4	GTDB	-118	-0.9837	0.4361	0.1694	0.3962	0.9172	0.0357
MPA3	refMet4	IGC2	-111	-0.8999	0.3908	0.1741	0.4306	0.9259	0.0112
MPA3	refBioms	GTDB	-163	-1.1060	0.4442	0.1738	0.4362	0.9098	0.0372
MPA3	refBioms	IGC2	-157	-0.9961	0.3918	0.1760	0.4843	0.9542	0.0054
MPA3	refKrak	GTDB	-331	-1.5697	0.4907	0.1668	0.3021	0.9455	0.0281
MPA3	refKrak	IGC2	-279	-1.2490	0.4117	0.1635	0.3561	0.9595	0.0027
mOTUs	refMet4	GTDB	-23	-0.0070	0.1476	0.0353	0.7853	0.8868	0.0710
mOTUs	refMet4	IGC2	-10	0.0045	0.1352	0.0422	0.7949	0.8333	0.0453
mOTUs	refBioms	GTDB	-5	0.0250	0.1799	0.0392	0.8714	0.8880	0.0925
mOTUs	refBioms	IGC2	8	-0.0065	0.1603	0.0438	0.9008	0.8787	0.0425
mOTUs	refKrak	GTDB	-64	-0.1239	0.2177	0.0375	0.7930	0.9179	0.0793
mOTUs	refKrak	IGC2	-27	-0.0522	0.1769	0.0427	0.8106	0.8718	0.0127
MPA4	refMet4	GTDB	-35	-0.0994	0.0463	0.0117	0.8083	0.9732	0.0176
MPA4	refMet4	IGC2	-55	-0.1971	0.0647	0.0265	0.7290	0.9878	0.0006
MPA4	refBioms	GTDB	-18	-0.1914	0.0694	0.0201	0.9116	0.9691	0.0209
MPA4	refBioms	IGC2	-33	-0.2989	0.0965	0.0357	0.8806	0.9869	0.0006
MPA4	refKrak	GTDB	-91	-0.4335	0.1196	0.0260	0.7923	0.9756	0.0194
MPA4	refKrak	IGC2	-129	-0.4763	0.1458	0.0448	0.6977	0.9879	0.0008
BiomS	refMet4	GTDB	-28	-0.0179	0.0661	0.0189	0.8160	0.9433	0.0249
BiomS	refMet4	IGC2	-26	-0.0068	0.0389	0.0190	0.8694	0.9923	0.0005
BiomS	refBioms	GTDB	1	0.0851	0.0674	0.0233	0.9619	0.9574	0.0406
BiomS	refBioms	IGC2	-1	0.0674	0.0447	0.0227	0.9942	0.9969	0.0002
BiomS	refKrak	GTDB	-119	-0.2077	0.1304	0.0338	0.7167	0.9483	0.0454
BiomS	refKrak	IGC2	-73	-0.0094	0.0715	0.0299	0.8351	0.9962	0.0005

observation holds true when comparing the median number of lost features to the median total number of features (richness), notably in the case of GTDB feature space. MetaPhlAn4 deviates from this trend, exhibiting a higher rate of lost features, particularly in the case of the IGC2 feature space.

Two factors contribute to the observed differences in lost features by MetaPhlAn4 when projecting to the GTDB feature space. Firstly, MetaPhlAn4 employs its own projection tool, which occasionally yields results that are not assigned at the species level. Secondly, the greater loss in both abundance and feature number when converting to the IGC2 feature space can be attributed to the specificity of the IGC2 catalogue. This catalogue exclusively comprises species from the gut microbiome, whereas all the tools under scrutiny include a proportion of non-gut species in their catalogues, except for BiomScope. The UHGG collection for simulation exclusively contains gut microbiome features. Consequently, non-gut features identified by other tools were mostly false positives. As a result, a relatively high feature loss, as observed in the case of MetaPhlAn4, partially skewed the comparison in a favorable direction.

The absence of a one-to-one correspondence between different feature spaces is prone to generate discrepancies, such as false positives and false negatives, with the extent of these discrepancies contingent on the proximity of the native and projection feature spaces. To mitigate the impact of such disparities on the evaluation of profiling tools, we devised an experimental framework comprising different simulations and common feature spaces.

The performance of the tools was markedly influenced by the simulations. However, it is evident that Kraken+Bracken and MetaPhlAn3 consistently lagged behind in performance, irrespective of the simulation or projection space chosen. Conversely, the three remaining tools (mOTUs3, MetaPhlAn4, and BiomScope) demonstrated significantly superior performance, with variations across simulations.

We conducted a further assessment to quantify the extent of these disparities by computing distances between the estimated relative abundance and the ground truth, utilizing both Bray-Curtis and phylogenetically sensitive weighted UniFrac distances (Figure 3 and Supplementary Figure S1). Notably, the observation that UniFrac distances exhibited greater resilience to the impact of various simulations compared to Bray-Curtis distances suggests that mismatches primarily arose from phylogenetically closely related features. These discrepancies can be attributed to differences in the categorization of individual genomes or strains into species, as illustrated in Figure 9b. Such mismatches assume reduced significance when analyses are conducted at higher phylogenetic levels, such as genus or family, as is commonly practiced in benchmarking studies.

Specifically, we illustrated that the elevated levels of false positive relative abundance detected in certain samples could be wholly attributed to the confusion between phylogenetically closely related features. In instances where one feature exhibited high abundance in the simulation, this phenomenon became particularly evident (see Figure 6).

The Kraken+Bracken pipeline consistently exhibited a tendency to identify an excessive number of false positives. This resulted in an overestimation of diversity metrics (richness and Shannon), elevated distances (Bray-Curtis and UniFrac) from the reference abundance profile, and a lower precision score. Nevertheless, this pipeline compensated for these drawbacks with its high sensitivity. Kraken proved to be exceptionally valuable in facilitating cross-referencing between different feature spaces (see Methods), highlighting its versatility in utilizing various catalogues, including both GTDB and MGnify, as an advantageous trait.

Although still widely used, MetaPhlAn3 [53–55] (along with its previous versions [35–37, 39–42, 50–52], see Supplementary Table S2), did not withstand the competition with more recent tools. Its primary handicap stemmed from its highly constricted native feature space, resulting in a substantial number of false negatives, thereby impairing all performance metrics.

Among the three most modern tools studied here (mOTUs3, MetaPhlAn4, and BiomScope) mOTUs3 was the only tool that was not given the advantage of a simulation using its own measurements. This was primarily due to resource constraints, and secondarily because its simulation results would have been quite similar to MetaPhlAn4's (both are not specialized marker-gene-based pipelines, and MetaPhlAn4 is more recent). Nevertheless, mOTUs3 displayed little variation across the simulations and the projection spaces (see Figure

3). Importantly, mOTUs3 struck a balance between the numbers of false negatives and false positives, reflecting a good trade-off between sensitivity and precision. In terms of Bray-Curtis distance from the ground truth, mOTUs3 was similar to Kraken but clearly outperformed this pipeline when using the phylogenetically sensitive UniFrac distance. Additionally, the error in estimating species richness or Shannon diversity remained relatively insensitive to the simulation and the projection.

MetaPhlAn4 exhibited strong overall performance, excelling in specificity with minimal false positives. However, this specificity came at the cost of sensitivity, as it demonstrated relatively higher numbers of false negatives with notable variability across the simulations. Unlike mOTUs3 and BiomScope, MetaPhlAn4 did not appear to significantly benefit from being used for the simulation. Notably, when the simulation was based on Kraken, MetaPhlAn4 was disadvantaged in terms of richness and Shannon diversity estimation. Furthermore, projecting MetaPhlAn4 onto the IGC2 feature space resulted in a disadvantage, both in diversity estimates and in proximity to the ground truth.

BiomScope, like MetaPhlAn4, demonstrated excellent performance while exhibiting a trade-off between sensitivity and precision/specificity. Notably, BiomScope significantly benefited from both the initial simulation and the subsequent projection using its native IGC2 feature space. In these conditions, BiomScope consistently outperformed the other four tools across various metrics. However, it faced challenges with the species-rich refKrak simulation. Interestingly, BiomScope even outperformed MetaPhlAn4 when the refMet4 simulation was used (where MetaPhlAn4 provided the reference abundance). When given the advantage of its native feature space, BiomScope accurately predicted species richness but tended to slightly overestimate Shannon diversity.

The performance of the profiling tools exhibited sensitivity to the simulation details. This sensitivity can be attributed, in part, to the specific features chosen for representation in the simulations, such as cases where certain tools could not distinguish between two species of *Prevotella* or had differing annotations in the simulation and tool catalogues. Additionally, feature size in the simulations varied, with refKrak containing more than twice as many features as refMet4. While this variability did not impact our conclusions about Kraken and MetaPhlAn3, it did influence the performance of the remaining tools. Overall, mOTUs3 displayed the most consistent performance across the simulations, while MetaPhlAn4 and BiomScope were notably affected by the feature-rich refKrak simulation, which may include numerous species not present in their catalogues. It is worth noting that the high number of false positives generated by Kraken potentially makes the refKrak simulated dataset the least realistic of the three. BiomScope, and to a lesser extent MetaPhlAn4, also exhibited improved performance when the simulation was based on their own measurements (see Figure 3).

When comparing the performance of the tools, the choice of a common feature space had a notably greater impact than the initial tool selected for the

simulation. BiomScope, in particular, exhibited a significant advantage when projected onto the IGC2 feature space compared to the GTDB projection (see Figure 4). This relative advantage was more pronounced than that observed for any other pipeline, with Kraken, for instance, experiencing only marginal improvements from using its native space. Our interpretation is that the choice of a common feature space fundamentally alters the nature of the comparison, whereas the initial choice of the profiling tool is more of a technical detail. The GTDB projection is better suited for assessing tools' performance in a general bacterial detection context, while the IGC2 projection better represents their performance in the specific context of human gut microbiome detection. Therefore, BiomScope, being specialized for this environment, benefits more significantly from this projection advantage than Kraken.

In this study, we aimed to explore a fundamental conceptual question within the field, necessitating substantial computational resources and reliance on extensive public databases. Given the inherent complexity of the question, we intentionally limited our focus to the human gut microbiome context. However, we acknowledge this limitation as a constraint of our study and advocate for applying this approach to evaluate additional databases and tools across diverse microbiome settings.

## Methods

### State of the art taxonomic profilers

The taxonomic profilers and classifiers evaluated in this study are: Kraken [16, 17] (used together with Bracken [56, 57]), MetaPhlAn3 [30–32], mOTUs [28, 29], MetaPhlAn4 [33, 34], and BiomScope (which is showcased here for the first time). These tools were selected on different criteria, including the type of method, their popularity in the scientific community measured by an yearly averaged number of citations, and also how they are maintained.

#### *Kraken and Bracken*

Kraken[16, 17] uses K-mers to match metagenomic reads to whole genomes. We used it with the Genome Taxonomy Database (GTDB) release R207 [23–26] downloaded from the webpage of Struo2 tool <https://github.com/leylabmpi/Struo2> [58, 59]. Kraken was also used with UHGG database[27, 60] downloaded from the official ftp repository [http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify\\_genomes/human-gut/](http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/). As Kraken alone does not determine the species abundance, it is often used together with its sister tool called Bracken (Bayesian Reestimation of Abundance with KrakEN) [56, 57]. Subsequently, we refer to the pipeline composed of Kraken and Bracken simply as *Kraken*, only when Kraken was used to project native feature spaces to GTDB or to UHGG that it was used without Bracken.

### ***MetaPhlAn and mOTUs***

MetaPhlAn and mOTUs are two metagenomic profilers relying on marker genes. Both of these tools come with their own custom marker gene database. Importantly, MetaPhlAn uses species-specific marker genes (*i.e.*, genes shared by all strains of a given species and not found in strains of other species), whereas mOTUs uses a collection of universal single copy marker-genes corresponding to orthologous gene families present in all prokaryotic species. The version of mOTUs used here is mOTUs 3.0.3 [28, 29].

MetaPhlAn3 [30–32] is a popular version of the tool, which has been superseded by the newest MetaPhlAn4 version [33, 34]. MetaPhlAn4 was released when the current study was in preparation, drastically increasing the reference catalogue size by including metagenome-assembled genomes (MAGs) in addition to the cultured species already included in MetaPhlAn3. Therefore we chose to keep both MetaPhlAn versions in our comparison.

### ***BiomScope***

BiomScope is a new pipeline for shotgun metagenomic data that generates gene and species abundance tables. It maps reads on a reference gene catalogue (here, the updated Integrated Gene Catalogue of the human gut microbiome, abbreviated IGC2) [61]) with a nucleotide identity threshold of 95% using bowtie2 [62], and quantifies genes using a two-step procedure. First, uniquely mapped reads (reads mapped to a single gene in the catalogue) are attributed to their corresponding genes. Second, shared reads (reads that map with the same alignment score to multiple genes) are assigned to their corresponding genes in proportion to counts obtained with uniquely mapped reads. This approach is similar to the one developed in different softwares such as Meteor [45, 46], Mocat [63, 64] and NGLess [65]. These gene counts are divided by gene length, to account for the fact that more reads map on longer genes. Then, abundance of metagenomic species (*i.e.* *MSP*) [43, 44] are inferred from these normalized gene counts. The 100 best core genes for each species (ordered by decreasing pairwise correlation) are then used as marker genes. A species is considered present in the sample, if at least 5 of these 100 marker genes (5%) are detected. The species abundance is then estimated as the robust mean value (trimming first and last quintiles) of the non-zero normalized counts. The final relative species abundance is obtained by normalizing the total abundance of all the species present in the sample to unity.

### **Task orchestration with scitq**

The computational burden behind generation and processing of high sequencing depth simulated samples (3Gbp) required an orchestration solution, enabling the distribution of computational tasks over different servers, with the ability of quickly integrating different scientific programs (*i.e.* the above-mentioned state-of-the-art taxonomic profilers and CAMISIM). Different orchestration and workflow solutions were tested (including Nextflow[66] and Celery[67]) and while these solutions have strong qualities, the amount of

reworking required for each pipeline and some performance issues made them not suited for our purpose. A *de-novo* orchestration solution was developed upon the simple idea of a distributed task queue of containerized tasks, which is now proposed as an open-source solution: scitq, provided in the context of the present work (see code and data availability section).

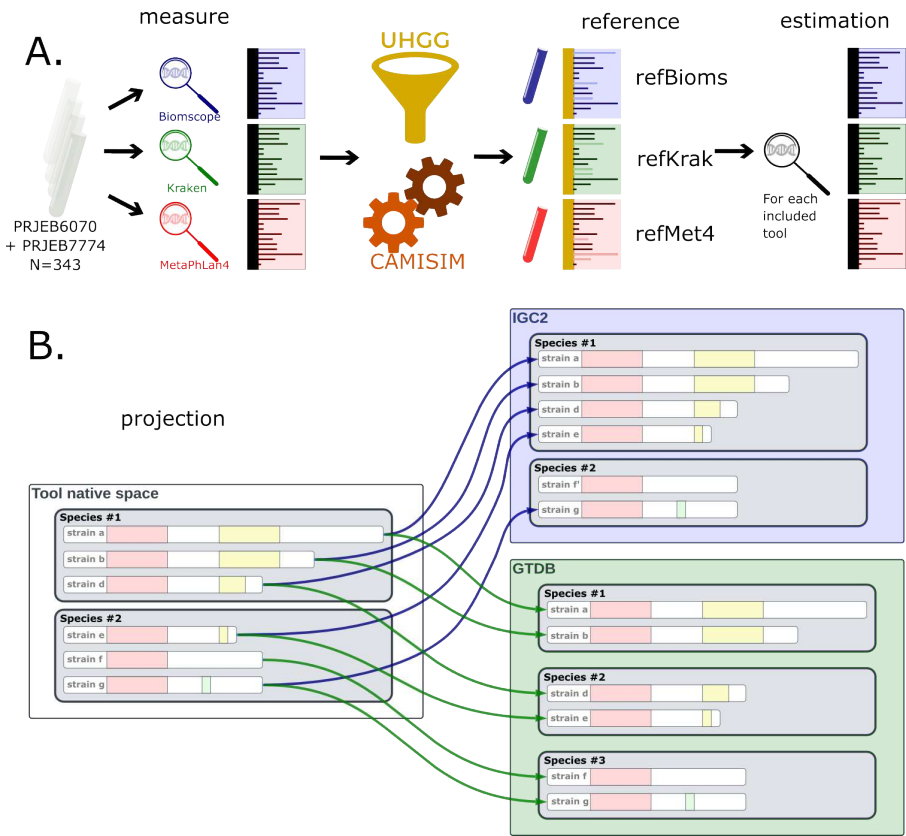
## Experimental workflow

The experimental workflow of this benchmark study is illustrated in Figure 9. Each one of the pipelines analyzed here assigns reads and estimates species abundance according to its own database and corresponding taxonomic annotation *native feature space*. This can be different from *the reference feature space* (*i.e.* the genomes used in the simulation) because some of the tools' features may not be represented or may be wrongly assigned. Here, we have explicitly chosen the reference feature space to be independent from those of the studied tools, in order to avoid biases in favor of any particular tool. For this, we have drawn reads from the genomes chosen from the Unified Human Gastrointestinal Genome (UHGG) collection available on MGnify [27, 60] and have projected the results of all the pipelines onto a *common feature space*. In particular, we have used two feature spaces: one based on the GTDB taxonomy R207 (*the GTDB feature space*) and the other based on the taxonomy provided with the metagenomic species catalogue [43, 44] (*IGC2 feature space*).

Figure 9 (a) illustrates our strategy. In short, it starts by computing reference abundance profiles with three different profilers based on real gut metagenomic data from two studies. These abundances along with the set of genomes from MGnify were used to simulate three sets of different metagenomes using CAMISIM, which were next used as input to all compared pipelines for profiling. Figure 9 (b) illustrates the projection procedure and the potential difficulties involved, notably that a strain/genome can be associated with different species, depending on the feature space used or none at all. In other words, this means that there is no unique one-to-one correspondence between features of any two spaces, and a feature in one space may potentially correspond to multiple features in another one. Thus, the projection would inevitably introduce errors in estimating species diversity and abundance profiles. We now consider how different steps of this workflow are implemented in our study.

### *Simulating metagenomic data*

The source material of the simulation is a typical meta-cohort of real metagenomic data from human fecal samples (n=343), comprising two public studies (*i.e.* PRJEB6070 [12] and PRJEB7774 [11]) including patients with colorectal cancer (CRC) (n=180) and controls (n=163). Three different simulations were run using CAMISIM [38] based on the initial species abundance profiles obtained with three different tools: MetaPhlan4, Kraken+Bracken/UHGG (that is, Kraken using the UHGG collection, version 2.0.1) and BiomScope.



**Figure 9** Schematic representation of the simulation and the analysis carried out in this study. Panel (a): samples from the PRJEB6070 and PRJEB7774 bioprojects were grouped and analyzed with three different pipelines (BiomScope, Kraken and MetaPhlan4) to obtain initial species abundance profiles for the simulation. The results were projected on UHGG species representative genomes, and used as input for the CAMISIM metagenome simulator, resulting in three reference datasets (resp. refBioms, refKrak and refMet4). Each dataset was then analyzed with the five compared pipelines (Kraken, MetaPhlan3 mOTUs, MetaPhlan4 and BiomScope). Panel (b): each compared pipeline output was then projected to a reference taxonomy, either IGC2 or GTDB.

The term feature was retained to designate how species/clades/OTU/MSP are described by the different tools.

The three simulated datasets were respectively named refMet4, refKrak and refBioms. For refMet4 and refBioms, profiles were based upon internal catalogue features not directly translatable to the UHGG collection, so a representative species in the UHGG was found using Kraken/UHGG on marker genes for each feature. For refKrak, the number of identified species was much higher compared with the other tools, therefore some filtering steps were added in Bracken post-treatment as suggested by the authors [56]: all species with less than 2000 reads ( $-t$  2000) were discarded as this was the minimal filtering to ensure no sample had more than a thousand species, which is an upper

bound for species richness in refMet4 and refBioms. Kraken tendency of over-estimating species richness was already reported in previous studies [68] and was also confirmed in the present work (see Results). For each dataset, 343 simulated samples made up of 10 million 2x150bp paired-end reads were generated with CAMISIM v1.3 using HiSeq profile. In total 1029 (343\*3) different samples were simulated.

### ***Abundance quantification and projection***

Each simulated dataset was processed with the five different pipelines: Kraken+Bracken/GTDB (Kraken version: 2.1.2, Bracken version: 2.8, the catalogue used is the GTDB catalogue in this case), MetaPhlAn3 (version 3.1.0), mOTUs (version 3.0.3), MetaPhlAn4 (version 4.0.6) and BiomScope (version 2.1). The abundance estimation of features in the respective native feature spaces were obtained. They were further projected into two common reference feature spaces, GTDB (Kraken/GTDB native feature space) and IGC2 (BiomScope native feature space), respectively.

### ***Projection to GTDB feature space***

MetaPhlAn3, mOTUs and BiomScope features were projected to GTDB applying Kraken/GTDB (without Bracken) on marker genes associated with their native features. MetaPhlAn4 embarks its own *ad hoc* projection tool (`sgb_to_gtdb_profile.py`). For Kraken no projection was needed as GTDB is its native feature space.

### ***Projection to IGC2 feature space***

Similarly, no projection was needed for BiomScope in IGC2 as this is its native feature space. For other pipelines, the main projection method consisted in extracting 2000 fake 100 bp reads from reference genomes of the features of each tool, and using BiomScope to identify the IGC2 features (MSP) corresponding to the different tools' native features. This method was used with success on Kraken. For MetaPhlAn3 & MetaPhlAn4 and mOTUs the features were identified only with marker genes (whereas Kraken use complete genomes) and probably because their marker genes differ from those of BiomScope, an important proportion of native features were lost with this direct method. For those lost native features, an intermediary projection was performed using Kraken/MGNIFY to get complete genomes, which were then submitted to the first method to identify IGC2 feature. Most lost features were successfully recovered and projected using this two-step approach. Kraken proved an invaluable tool to establish both projections.

The final projected tables can be found in the Supplementary data 3.

### ***Converting abundance tables to the common feature space***

When converting the abundance tables from the tools' native feature space to the common feature space, the following procedure was followed:

- abundances of the native features corresponding to the same feature of the common feature space were summed.
- abundances of the native features corresponding to several features of the native feature space were split proportionally. The proportions were obtained during the projection (this was necessary only for conversion to IGC2, as this case does not appear from the projection onto the GTDB space).
- the native features without correspondence in the common feature space were omitted, and the abundance tables obtained after conversion were normalized to unity.
- we omitted all the features that could not be identified at the species level.

### *Computing performance metrics*

The final metrics were calculated using Python packages NumPy, SciPy, Biopython, and Scikit-bio (The code is provided as part of the Supplementary materials). All the figures were created using Matplotlib.

*Sensitivity and precision* are defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}; \text{ Precision} = \frac{TP}{TP + FP} \quad (1)$$

where  $TP$ ,  $FP$ , and  $FN$  are respectively the numbers of true positives (features correctly identified), false positives (features identified but not present in the samples), and false negatives (features not identified, although present in the samples).

*The false positive relative abundance (FPRA)* is the sum of the estimated relative abundances of false positives in a sample

*Bray-Curtis distance* is defined by

$$d_{BC}(u, v) = \frac{\sum_i |u_i - v_i|}{\sum_i |u_i + v_i|}, \quad (2)$$

where  $u_i$  and  $v_i$  are the true and the estimated abundances of species  $i$ .

*Weighted UniFrac distance* [48, 49] is defined as

$$d_{WUF}(u, v) = \frac{\sum_{i=1}^n b_i |u_i - v_i|}{\sum_{i=1}^n b_i (u_i + v_i)}, \quad (3)$$

where  $b_i$  is the length of the branch  $i$  on the phylogenetic tree relating all the features.

*Jensen-Shannon divergence* is the pairwise distance between two matrices  $p$  and  $q$ . It is defined as

$$JSD(p, q) = \sqrt{\frac{D(p||m) + D(q||m)}{2}}, \quad (4)$$

where  $m = (p + q)/2$  and the Kullback-Leibler divergence is given by

$$D(p\|m) = \sum_{i,j} p_{ij} \log \left( \frac{p_{ij}}{q_{ij}} \right). \quad (5)$$

## Data availability

For convenience, all code and data is centralized in this github repository:

<https://github.com/gmtscience/microbiome-pipeline-benchmarking>

## Acknowledgments

We are very grateful to OVHcloud, which sponsored part of the this work providing most computational time and data hosting. We would also like to thank Etienne Formstecher and Joel Dore for their thorough proofreading of the manuscript.

## Additional information

- Funding: This work was funded by grants DOS0171565/00 and DOS0171566/00 from BPI France and Région Normandie, and supported in kind by OVHcloud.
- Conflict of interest/Competing interests : V.P., F.P.O. and R. de L. are employees of GMT Science, E.P. is a scientific advisor to GMT Science.
- Ethics approval: Although some human derived samples from other public studies were used in that work, all the information was used anonymously taking no account of any specific participant characteristics (such as sex, age, or health status).
- Consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials : Simulated samples are published as PRJNA987980 on NCBI SRA archive. Codes and other data are provided with extensive documentation and details in <https://github.com/gmtscience/microbiome-pipeline-benchmarking>
- Supplementary information accompanies this paper in a separate PDF file (REPLACE BY URL AFTER PUBLICATION)
- Authors' contributions : V.P., F.P.O., E.P., and R. de L. designed the study and conceived the methodology, R. de L. carried out the simulations and processing by pipelines, V.P. carried out the initial tests of taxonomic profilers and analyzed the results, V.P., E.P. and R. de L. designed experiments and wrote the manuscript, F.P.O. contributed expertise on all stages of the project.

## References

- [1] Thomas, L.V., Ockhuizen, T.: New insights into the impact of the intestinal microbiota on health and disease: a symposium report. *British Journal of Nutrition* **107**(S1), 1–13 (2012)
- [2] Ley, R.E., Turnbaugh, P.J., Klein, S., Gordon, J.I.: Human gut microbes associated with obesity. *nature* **444**(7122), 1022–1023 (2006)
- [3] Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., Gordon, J.I.: An obesity-associated gut microbiome with increased capacity for energy harvest. *nature* **444**(7122), 1027–1031 (2006)
- [4] Turnbaugh, P.J., Hamady, M., Yatsunenkov, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., *et al.*: A core gut microbiome in obese and lean twins. *nature* **457**(7228), 480–484 (2009)
- [5] Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., Gordon, J.I.: Obesity alters gut microbial ecology. *Proceedings of the national academy of sciences* **102**(31), 11070–11075 (2005)
- [6] Lepage, P., Häslér, R., Spehlmann, M.E., Rehman, A., Zvirbliene, A., Begun, A., Ott, S., Kupcinskis, L., Doré, J., Raedler, A., *et al.*: Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* **141**(1), 227–236 (2011)
- [7] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., *et al.*: A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**(7418), 55–60 (2012)
- [8] Vijay-Kumar, M., Aitken, J.D., Carvalho, F.A., Cullender, T.C., Mwangi, S., Srinivasan, S., Sitaraman, S.V., Knight, R., Ley, R.E., Gewirtz, A.T.: Metabolic syndrome and altered gut microbiota in mice lacking toll-like receptor 5. *Science* **328**(5975), 228–231 (2010)
- [9] Yan, A.W., Fouts, D., Brandl, J., Stärkel, P., Torralba, M., Schott, E., Tsukamoto, H., Nelson, K., Brenner, D., Schnabl, B.: Enteric dysbiosis associated with a mouse model of alcoholic liver disease. *Hepatology* **53**(1), 96–105 (2011)
- [10] Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chate-lier, E., Yao, J., Wu, L., *et al.*: Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**(7516), 59–64 (2014)
- [11] Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., nauer, U.,

- Niederseer, D., Xu, X., Al-Aama, J.Y., Yang, H., Wang, J., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C., Wang, J.: Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* **6**, 6528 (2015)
- [12] Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., hm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D.R., Schneider, M.A., Schrotz-King, P., Tournigand, C., Tran Van Nhieu, J., Yamada, T., Zimmermann, J., Benes, V., Kloor, M., Ulrich, C.M., von Knebel Doeberitz, M., Sobhani, I., Bork, P.: Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**(11), 766 (2014)
- [13] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.*: A human gut microbial gene catalogue established by metagenomic sequencing. *nature* **464**(7285), 59–65 (2010)
- [14] Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., *et al.*: An integrated catalog of reference genes in the human gut microbiome. *Nature biotechnology* **32**(8), 834–841 (2014)
- [15] Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., *et al.*: A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology* **39**(1), 105–114 (2021)
- [16] Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* **15**(3), 1–12 (2014)
- [17] Wood, D.E., Lu, J., Langmead, B.: Improved metagenomic analysis with kraken 2. *Genome biology* **20**(1), 1–13 (2019)
- [18] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using diamond. *Nature methods* **12**(1), 59–60 (2015)
- [19] Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L.: Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* **26**(12), 1721–1729 (2016)
- [20] Menzel, P., Ng, K.L., Krogh, A.: Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications* **7**(1), 11257 (2016)
- [21] O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei,

- D., *et al.*: Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**(D1), 733–745 (2016)
- [22] Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., Ostell, J.: Ncbi prokaryotic genome annotation pipeline. *Nucleic acids research* **44**(14), 6614–6624 (2016)
- [23] Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P.: Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research* **50**(D1), 785–794 (2022)
- [24] Rinke, C., Chuvochina, M., Mussig, A.J., Chaumeil, P.-A., Davín, A.A., Waite, D.W., Whitman, W.B., Parks, D.H., Hugenholtz, P.: A standardized archaeal taxonomy for the genome taxonomy database. *Nature Microbiology* **6**(7), 946–959 (2021)
- [25] Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P.: A complete domain-to-species taxonomy for bacteria and archaea. *Nature biotechnology* **38**(9), 1079–1086 (2020)
- [26] Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., Hugenholtz, P.: A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology* **36**(10), 996–1004 (2018)
- [27] Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L.J., *et al.*: Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **51**(D1), 753–759 (2023)
- [28] Ruscheweyh, H.-J., Milanese, A., Paoli, L., Sintsova, A., Mende, D.R., Zeller, G., Sunagawa, S.: Motus: Profiling taxonomic composition, transcriptional activity and strain populations of microbial communities. *Current Protocols* **1**(8), 218 (2021)
- [29] Milanese, A., Mende, D.R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., Hingamp, P., Alves, R., Costea, P.I., Coelho, L.P., *et al.*: Microbial abundance, activity and population genomic profiling with motus2. *Nature communications* **10**(1), 1–11 (2019)
- [30] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C.: Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **9**(8), 811–814 (2012)

- [31] Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., Segata, N.: Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods* **12**(10), 902–903 (2015)
- [32] Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., *et al.*: Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *Elife* **10**, 65088 (2021)
- [33] Blanco-Miguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., *et al.*: Extending and improving metagenomic taxonomic profiling with uncharacterized species with metaphlan 4. *bioRxiv* (2022)
- [34] Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., Segata, N.: Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research* **27**(4), 626–638 (2017)
- [35] Parks, D.H., Rigato, F., Vera-Wolf, P., Krause, L., Hugenholtz, P., Tyson, G.W., Wood, D.L.: Evaluation of the microba community profiler for taxonomic profiling of metagenomic datasets from the human gut microbiome. *Frontiers in Microbiology* **12**, 643682 (2021)
- [36] Simon, H.Y., Siddle, K.J., Park, D.J., Sabeti, P.C.: Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**(4), 779–794 (2019)
- [37] Seppey, M., Manni, M., Zdobnov, E.M.: Lemmi: a continuous benchmarking platform for metagenomics classifiers. *Genome research* **30**(8), 1208–1216 (2020)
- [38] Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T.R., Belmann, P., DeMaere, M.Z., Darling, A.E., *et al.*: Camisim: simulating metagenomes and microbial communities. *Microbiome* **7**(1), 1–12 (2019)
- [39] Lindgreen, S., Adair, K.L., Gardner, P.P.: An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports* **6**(1), 19233 (2016)
- [40] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., *et al.*: Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods* **14**(11), 1063–1071 (2017)
- [41] Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T.R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., *et al.*: Critical

- assessment of metagenome interpretation: the second round of challenges. *Nature methods* **19**(4), 429–440 (2022)
- [42] McHardy, A.C., Meyer, F.: CAMI II: identifying best practices and issues for metagenomics software. *NATURE PORTFOLIO HEIDELBERGER PLATZ 3, BERLIN, 14197, GERMANY* (2022)
- [43] Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A.C., Gauthier, F., Magoulès, F., Ehrlich, S.D., Pichaud, M.: Mspminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* **35**(9), 1544–1552 (2019)
- [44] Plaza Onate, F., Pons, N., Gauthier, F., Almeida, M., Ehrlich, S.D., Le Chatelier, E.: Updated Metagenomic Species Pan-genomes (MSPs) of the Human Gastrointestinal Microbiota. <https://doi.org/10.15454/FLANUP>
- [45] Pons, N., Batto, J., Kennedy, S., Almeida, M., Boumezbeur, F., Moumen, B., Léonard, P., Le Chatelier, E., Ehrlich, S.D., Renault, P.: METEOR, a platform for quantitative metagenomic profiling of complex ecosystems. <http://www.jobim2010.fr/sites/default/files/presentations/27Pons.pdf>
- [46] Meteor (Metagenomic Explorer), a software for profiling metagenomic data at gene level. <https://forgemia.inra.fr/metagenopolis/meteor>
- [47] Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., *et al.*: Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**(7464), 541–546 (2013)
- [48] Lozupone, C., Knight, R.: Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71**(12), 8228–8235 (2005)
- [49] Lozupone, C.A., Hamady, M., Kelley, S.T., Knight, R.: Quantitative and qualitative  $\beta$  diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology* **73**(5), 1576–1585 (2007)
- [50] McIntyre, A.B., Ounit, R., Afshinnekoo, E., Prill, R.J., Hénaff, E., Alexander, N., Minot, S.S., Danko, D., Foon, J., Ahsanuddin, S., *et al.*: Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology* **18**(1), 1–19 (2017)
- [51] Amos, G.C., Logan, A., Anwar, S., Fritzsche, M., Mate, R., Bleazard, T., Rijpkema, S.: Developing standards for the microbiome field. *Microbiome* **8**, 1–13 (2020)

- [52] Miossec, M.J., Valenzuela, S.L., Pérez-Losada, M., Johnson, W.E., Crandall, K.A., Castro-Nallar, E.: Evaluation of computational methods for human microbiome analysis using simulated data. *PeerJ* **8**, 9688 (2020)
- [53] Portik, D.M., Brown, C.T., Pierce-Ward, N.T.: Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC bioinformatics* **23**(1), 541 (2022)
- [54] Wright, R.J., Comeau, A.M., Langille, M.G.: From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microbial Genomics* **9**(3) (2023)
- [55] Xu, R., Rajeev, S., Salvador, L.C.: The selection of software and database for metagenomics sequence analysis impacts the outcome of microbial profiling and pathogen detection. *Plos one* **18**(4), 0284031 (2023)
- [56] Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L.: Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**, 104 (2017)
- [57] Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., Steinegger, M.: Metagenome analysis using the kraken software suite. *Nature protocols*, 1–25 (2022)
- [58] de la Cuesta-Zuluaga, J., Ley, R.E., Youngblut, N.D.: Struo: a pipeline for building custom databases for common metagenome profilers. *Bioinformatics* **36**(7), 2314–2315 (2020)
- [59] Youngblut, N.D., Ley, R.E.: Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets. *PeerJ* **9**, 12198 (2021)
- [60] Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., Finn, R.D.: A new genomic blueprint of the human gut microbiota. *Nature* **568**(7753), 499–504 (2019)
- [61] Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., *et al.*: Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome biology* **18**(1), 1–13 (2017)
- [62] Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. *Nat. Methods* **9**(4), 357–359 (2012)
- [63] Kultima, J.R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D.R., Arumugam, M., Pan, Q., Liu, B., Qin, J., *et al.*: Mocat: a metagenomics

assembly and gene prediction toolkit (2012)

- [64] Kultima, J.R., Coelho, L.P., Forslund, K., Huerta-Cepas, J., Li, S.S., Driessen, M., Voigt, A.Y., Zeller, G., Sunagawa, S., Bork, P.: Mocat2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* **32**(16), 2520–2523 (2016)
- [65] Coelho, L.P., Alves, R., Monteiro, P., Huerta-Cepas, J., Freitas, A.T., Bork, P.: Ng-meta-profiler: fast processing of metagenomes using ngless, a domain-specific language. *Microbiome* **7**(1), 1–10 (2019)
- [66] Tommaso, P.D., Floden, E.W., Magis, C., Palumbo, E., Notredame, C.: *Nextflow*: un outil efficace pour l'amélioration de la stabilité numérique des calculs en analyse génomique. *Biologie Aujourd'hui* **211**(3), 233–237 (2017)
- [67] Celery: Celery - Distributed Task Queue. <https://docs.celeryq.dev/en/stable/>
- [68] Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S.: Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics* **16**(1), 1–13 (2015)

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Impactofsimulationsupplementarymaterial.pdf](#)