Impact of simulation and reference catalogues on the evaluation of taxonomic profiling pipelines (Supplementary Information)

Vadim Puller[1], Florian Plaza Oñate[1], Edi Prifti[2,3*] and Raynald de Lahondès[1*]

[1]GMT science, 75 route de Lyons-La-Foret, Rouen, F-76000, France.
[2]IRD, Sorbonne Université, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, UMMISCO, 32 avenue Henri Varagnat, Bondy, F-93143, France.
[3]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; raynald.delahondes@gmt.bio;
Contributing authors: vadim.puller@gmt.bio; fplazaonate@gmt.bio;

Supplementary figures

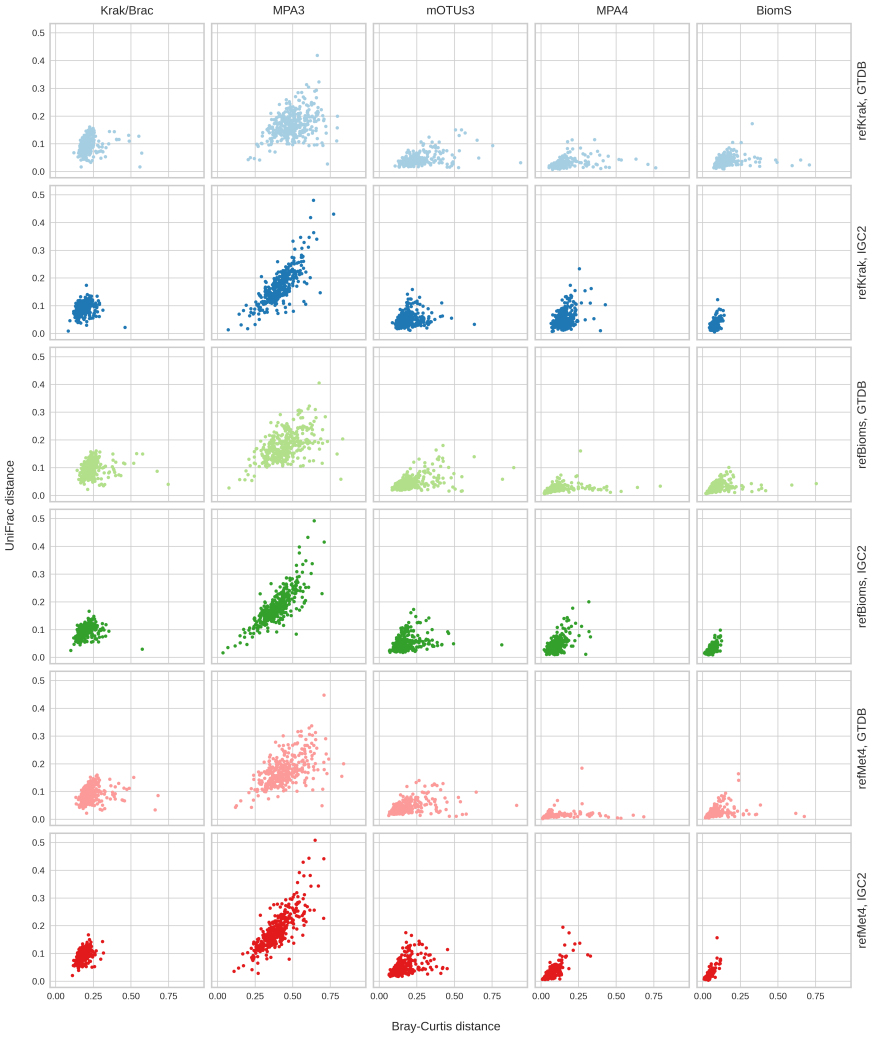

Fig. S1 Scatterplot of weighted UniFrac distance from estimation to reference vs. the Bray-Curtis distance. Each point corresponds to a sample.

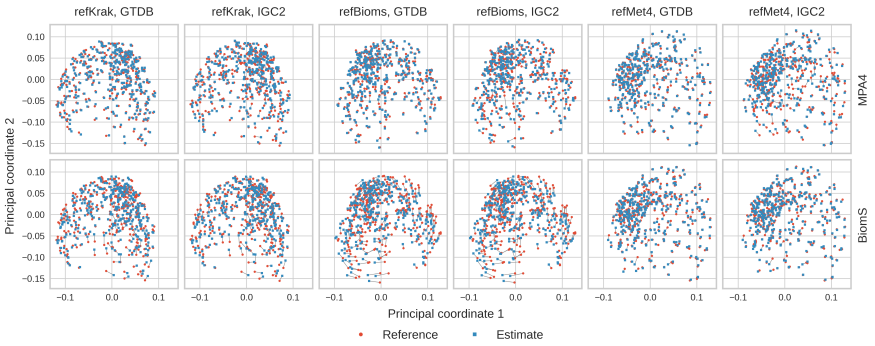

Fig. S2 Principal Coordinates decomposition of pairwise distance matrices (Bray-Curtis distance.)

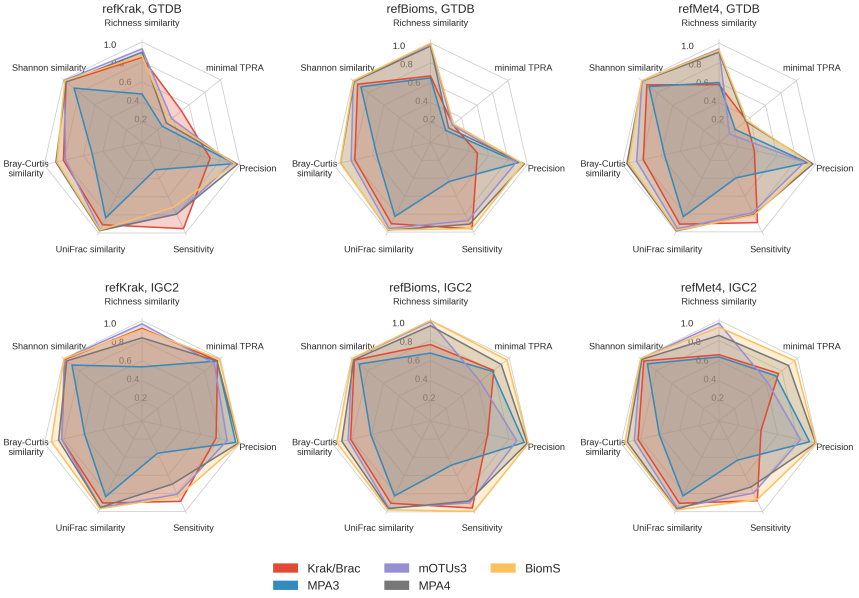

Fig. S3 Comparison of tool performances across different metrics, simulations, and projection spaces.

Overview of the most popular metagenomic tools

The table below summarizes some of the most popular metagenomic tools, as reviewed in [1–14]. We briefly describe the columns of these tables:

Stand alone

Some of these tools require pre-processing, i.e., they function as a part of a pipeline with other tools: e.g., Bracken is a profiling tool used with Kraken output, whereas metaBAT, CONCOCT and MaxBin operate with reads already assembled in contigs. Moreover several tools are not available for download, but function as web-servers, where the users may download their data for processing.

Function

Here we refer to the tools that only determine presence of species as *classification* tools, whereas those that determine the abundance profile/prevalence as *profiling tools*.

Method

Marker genes and *whole genome* refers to the tools mapping reads to marker genes or whole genomes, with the implication that marker genes usually constitute a tool-specific catalogue (even if borrowed from a public database), whereas whole genome tools are generally capable of working with a user-specified genome database. For binning tools this entry says *contigs*, whereas for Bracken and MEGAN, which function in pipeline with Kraken2 and DIAMOND respectively, this field is not specified.

The remaining columns designate the year when the first version of tool was published and cite the relevant publications.

Table S1 Most popular metagenomic tools (as reviewed in [1–9])

Tool	Stand alone	Function	Method	Year	Reference
Kraken	yes	classification	whole genome	2014	[15, 16]
Bracken	use with Kraken	profiling		2017	[17, 18]
mOTUs	yes	profiling	marker genes	2013	[14, 19]
MetaPhlAn	yes	profiling	marker genes	2012	[20–22]
MetaPhlAn4	yes	profiling	marker genes	2022	[23, 24]
BiomScope	proprietary software	profiling	marker genes	2023	this article
MetaPhyler	yes	profiling	marker genes	2011	[25, 26]
TIPP	yes	profiling	marker genes	2014	[27, 28]
FOCUS	yes	profiling	whole genome	2014	[29]
DIAMOND	yes	classification	whole genome	2015	[30]
MEGAN6	use with DIAMOND or BLAST	profiling		2007	[31–37]
MetaPalette	yes	profiling	whole genome	2016	[38]
GOTTCHA	yes	profiling	marker genes ?	2015	[39]
MICROBA	proprietary software	profiling	whole genome	2021	[7]
METEOR		profiling		2010	
QIIME	yes	profiling	marker genes, whole genomes (via plugins)	2010	[40, 41]
MOCAT2	yes	profiling	whole genome	2012	[42, 43]
NGLess	yes	profiling	marker genes, whole genome (via plugins)	2019	[44]
Centrifuge	yes	classification	whole genome	2016	[45]
CLARK	yes	classification	whole genome	2015	[46, 47]
Kaiju	yes	classification	whole genome	2016	[48]
NBC++	Web-server	classification	whole genome	2010	[49, 50]
PhymmBL	yes	classification	whole genome	2009	[51, 52]
Ganon	yes	classification	whole genome	2020	[53]
DUDes	yes	classification	whole genome	2016	[54]
Metacache	yes	classification	whole genome	2017	[55]
LMAT	yes	classification	whole genome	2013	[56–58]
LCA*	classifies asembled contigs	classification	whole genome	2016	[59]
MetaWatt	bin assembled contigs	classification	contigs	2012	[60]
taxator-tk	use with BLAST or assembled contigs	classification	whole genome	2015	[61]
CCMetagen	yes	classification	whole genome	2020	[62, 63]
specI	Web-server	classification	marker genes	2013	[64]
metaBAT	bins assembled contigs	binning	contigs	2015	[65, 66]
CONCOCT	bin assembled contigs	binning	contigs	2014	[67]
MaxBin	bins assembled contigs	binning	contigs	2014	[68, 69]

State-of-the-art benchmarking studies

In Table S2 we provide a short summary of the benchmarking studies involving metagenomic profilers [1–10, 12–14, 70].

Diversity of profilers and versions

Table S1 presents only part of the tools analyzed in the cited benchmarking studies. The number of tools compared in a given study varies from two (particularly where it mainly concerns introducing a new tool [14]) to several dozens [2, 71]. Various versions of Kraken (and Bracken), mOTUs and MetaPhlAn appear in most studies and can be viewed as the most popular representatives of the most recognized algorithms. However, the published studies, although recent, consider older versions of these tools: mOTUs2 and MetaPhlAn3 or even lower. This is less of a problem for Kraken, which remains unchanged since the emergence of Kraken2 and Bracken.

Correspondence between feature spaces

Establishing correspondence between the feature space in which simulation is performed and the native feature spaces of the studied profilers is crucial for using most of the traditional performance metrics. Indeed, even simple metrics such as the species richness, is highly dependent on the catalogue used (*i.e. the same number of genomes can be grouped into different number of species or be grouped differently; see Fig. 9b in the main text). To our knowledge the influence of the feature space used for comparison had not been previously studied. Previous state-of-the-art studies focused exclusively on the tools using the NCBI taxonomy, or performed taxonomic comparison at higher taxonomic levels - family or genus (where the discrepancy is less visible, and the performance is consequently better than when focusing at the species level, see, e.g., [14]). This noticeably put at disadvantage mOTUs, using its own original taxonomy [2, 4], and, to a lesser extent MetaPhlAn, mostly based on the NCBI taxonomy. (This shortcoming is alleviated only partially by using UniFrac distance, and almost never alongside the related, but phylogenetically insensitive, Bray-Curtis distance.)*

The impact of the reference databases

A Recent study [7] used the GTDB database as a standardized feature catalogue for all tools that were compared [7]. However, although this permitted the comparison between the classification/profiling algorithms, it deprived some of them from any advantage that they might have had using their own unique catalogues. The influence of the database choice has been recently the focus of interest in [11, 13]. However, these studies were limited to varying the size of the Kraken2 database, and always focused on the same feature space (GTDB or NCBI).

Simulated samples and mock communities

CAMI and CAMI2 [8, 9, 71, 72] are the two benchmarking datasets that were explicitly designed for comparing metagenomic pipelines and used in numerous studies [2, 5, 8, 11, 13, 14]. Other studies have used their own simulations [1, 7, 10] or sometimes combination of own simulations and CAMI datasets [5, 11, 14]. Finally, there are studies that used the data obtained from synthetic communities [3, 6, 12]. The number of samples in these studies could vary from a few samples, e.g., [4, 10] to several hundreds [2, 7, 8]. The number of species in simulated samples was typically a few hundred, whereas in synthetic samples only a few dozen or less.

In terms of the effort made to make simulated samples to represent the real-world data, those based on CAMISIM simulator are probably the most advanced. However, no specific study until now had used CAMISIM for generating its own samples (with the obvious exceptions of the studies carried within the context of CAMI challenges [2, 8]). Finally, the studies vary greatly in terms of the type of the microbiome addressed: ocean, plant, human gut, mouse gut or simply generic microbiome based on the reference genome database.

Table S2 Previous benchmarking studies

Study	Method of comparison	Test data	Number of samples	Type of microbiome	Reference taxonomy	Number of tools	Tools studied	Reference
Lindgreen (2016)	different taxonomic ranks	own simulations	6	unspecified	NCBI	14	Kraken, MetaPhlAn, mOTU, other	[1]
Sczyrba (2017)	different taxonomic ranks	CAMI challenge	700	unspecified	NCBI	10	Kraken (binning only), MetaPhlAn2, mOTUs, other	[2]
Mcintyre (2017)	different taxonomic ranks	synthetic and simulations	6	human mouth, human gut, other	NCBI	11	Kraken, MetaPhlAn2, other	[3]
Ye (2019)	standartized database, different taxonomic ranks	own simulations	10	unspecified	NCBI	18	Kraken2+Bracken, MetaPhlAn2, mOTUs2, other	[4]
Milanese (2019)	different taxonomic ranks	own simulations, CAMI challenge	10	human gut	NCBI	11	Kraken+Bracken, MetaPhlAn2, mOTUs2, other	[14]
Seppey (2020)	imposing NCBI taxonomy	own simulations, CAMI1	8	unspecified	NCBI	8	Kraken2+Bracken, MetaPhlAn2, other	[5]
Amos (2020)	limited number of known species	synthetic communities	2	human gut	NCBI	5	Kraken+Bracken, MetaPhlAn2, other	[6]
Mosesc (2020)	different taxonomic ranks	own simulations	8	gut, ocean	NCBI	8	Kraken, MetaPhlAn2, other	[10]
Parks (2021)	standartized database	own simulations	140	unspecified, gut	GTDB, NCBI	10	Kraken2+Bracken, MetaPhlAn2, mOTUs2	[7]
Meyer (2022)	different taxonomic ranks	CAMI II	113	human microbiome, mouse gut	NCBI	20	Kraken2+Bracken, MetaPhlAn2, mOTUs2, other	[8, 9]
Portik (2022)	limited number of known species, different taxonomic levelss	synthetic communities	10	unspecified	NCBI	10	Kraken2, MetaPhlAn3, mOTUs2, other	[12]
Wright (2023)	same taxonomy	CAMI, other	164	mouse gut, marine, other	NCBI, GTDB	2	Kraken2+Bracken, MetaPhlAn3	[11]
Xu (2023)	different taxonomic ranks	CAMI, biological samples	10	mouse gut	NCBI	9	Kraken2+Bracken, MetaPhlAn3, other	[13]

Ranking tools

We ranked the performance of the tools according to different metrics by performing pairwise one-sided Wilcoxon sign tests. The results are presented in the tables S3-S9.

Table S3 Ranking of the tools according to the error in estimating species richness.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	mOTUs3	MPA4	BiomS	Krak/Brac	MPA3
refKrak	IGC2	mOTUs3	BiomS	Krak/Brac	MPA4	MPA3
refBioms	GTDB	BiomS	mOTUs3	MPA4	MPA3	Krak/Brac
refBioms	IGC2	BiomS	mOTUs3	MPA4	MPA3	Krak/Brac
refMet4	GTDB	mOTUs3	BiomS	MPA4	MPA3	Krak/Brac
refMet4	IGC2	mOTUs3	BiomS	MPA4	MPA3	Krak/Brac

Table S4 Ranking of the tools according to the error in estimating Shannon diversity.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	mOTUs3	BiomS	Krak/Brac	MPA4	MPA3
refKrak	IGC2	BiomS	mOTUs3	Krak/Brac	MPA4	MPA3
refBioms	GTDB	mOTUs3	BiomS	MPA4	Krak/Brac	MPA3
refBioms	IGC2	BiomS	mOTUs3	MPA4	Krak/Brac	MPA3
refMet4	GTDB	BiomS	mOTUs3	MPA4	Krak/Brac	MPA3
refMet4	IGC2	BiomS	mOTUs3	MPA4	Krak/Brac	MPA3

Table S5 Ranking of the tools according to the Bray-Curtis distance from the ground truth.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	MPA4	BiomS	Krak/Brac	mOTUs3	MPA3
refKrak	IGC2	BiomS	MPA4	Krak/Brac	mOTUs3	MPA3
refBioms	GTDB	BiomS	MPA4	mOTUs3	Krak/Brac	MPA3
refBioms	IGC2	BiomS	MPA4	mOTUs3	Krak/Brac	MPA3
refMet4	GTDB	MPA4	BiomS	mOTUs3	Krak/Brac	MPA3
refMet4	IGC2	BiomS	MPA4	mOTUs3	Krak/Brac	MPA3

Table S6 Ranking of the tools according to the weighted UniFrac distance from the ground truth.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	MPA4	BiomS	mOTUs3	Krak/Brac	MPA3
refKrak	IGC2	BiomS	mOTUs3	MPA4	Krak/Brac	MPA3
refBioms	GTDB	MPA4	BiomS	mOTUs3	Krak/Brac	MPA3
refBioms	IGC2	BiomS	MPA4	mOTUs3	Krak/Brac	MPA3
refMet4	GTDB	MPA4	BiomS	mOTUs3	Krak/Brac	MPA3
refMet4	IGC2	BiomS	MPA4	mOTUs3	Krak/Brac	MPA3

Table S7 Ranking of the tools according to their sensitivity.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	Krak/Brac	mOTUs3	MPA4	BiomS	MPA3
refKrak	IGC2	Krak/Brac	BiomS	mOTUs3	MPA4	MPA3
refBioms	GTDB	Krak/Brac	BiomS	MPA4	mOTUs3	MPA3
refBioms	IGC2	BiomS	Krak/Brac	mOTUs3	MPA4	MPA3
refMet4	GTDB	Krak/Brac	BiomS	MPA4	mOTUs3	MPA3
refMet4	IGC2	Krak/Brac	BiomS	mOTUs3	MPA4	MPA3

Table S8 Ranking of the tools according to their precision.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	MPA4	BiomS	MPA3	mOTUs3	Krak/Brac
refKrak	IGC2	BiomS	MPA4	MPA3	mOTUs3	Krak/Brac
refBioms	GTDB	MPA4	BiomS	MPA3	mOTUs3	Krak/Brac
refBioms	IGC2	BiomS	MPA4	MPA3	mOTUs3	Krak/Brac
refMet4	GTDB	MPA4	BiomS	MPA3	mOTUs3	Krak/Brac
refMet4	IGC2	BiomS	MPA4	MPA3	mOTUs3	Krak/Brac

Table S9 Ranking of the tools according to the false positive relative abundance.

Simulation	Common space	Best	2nd best	3rd best	4th best	Last
refKrak	GTDB	MPA4	MPA3	BiomS	Krak/Brac	mOTUs3
refKrak	IGC2	BiomS	MPA4	MPA3	mOTUs3	Krak/Brac
refBioms	GTDB	MPA4	MPA3	BiomS	mOTUs3	Krak/Brac
refBioms	IGC2	BiomS	MPA4	MPA3	mOTUs3	Krak/Brac
refMet4	GTDB	MPA4	BiomS	MPA3	mOTUs3	Krak/Brac
refMet4	IGC2	BiomS	MPA4	MPA3	mOTUs3	Krak/Brac

References

[1] 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)

[2] 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)

[3] McIntyre, A.B., Ounit, R., Afshinnekoo, E., Prill, R.J., Hénaff, E., Alexander, N., Minot, S.S., Danko, D., Foox, J., Ahsanuddin, S., *et al.*: Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome biology **18**(1), 1–19 (2017)

[4] Simon, H.Y., Siddle, K.J., Park, D.J., Sabeti, P.C.: Benchmarking metagenomics tools for taxonomic classification. Cell **178**(4), 779–794 (2019)

[5] Seppey, M., Manni, M., Zdobnov, E.M.: Lemmi: a continuous benchmarking platform for metagenomics classifiers. Genome research **30**(8), 1208–1216 (2020)

[6] 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)

[7] 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)

[8] 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)

[9] McHardy, A.C., Meyer, F.: CAMI II: identifying best practices and issues for metagenomics software. NATURE PORTFOLIO HEIDELBERGER PLATZ 3, BERLIN, 14197, GERMANY (2022)

[10] 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)

[11] 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)

[12] 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)

[13] 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)

[14] 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)

[15] Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology **15**(3), 1–12 (2014)

[16] Wood, D.E., Lu, J., Langmead, B.: Improved metagenomic analysis with kraken 2. Genome biology **20**(1), 1–13 (2019)

[17] Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L.: Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science **3**, 104 (2017)

[18] 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)

[19] 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)

[20] 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)

[21] 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)

[22] 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)

[23] 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)

[24] 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)

[25] Liu, B., Gibbons, T., Ghodsi, M., Pop, M.: Metaphyler: Taxonomic profiling for metagenomic sequences. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 95–100 (2010). IEEE

[26] Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., Pop, M.: Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. Genome biology **12**, 1–27 (2011)

[27] Nguyen, N.-p., Mirarab, S., Liu, B., Pop, M., Warnow, T.: Tipp: taxonomic identification and phylogenetic profiling. Bioinformatics **30**(24), 3548–3555 (2014)

[28] Shah, N., Molloy, E.K., Pop, M., Warnow, T.: Tipp2: metagenomic taxonomic profiling using phylogenetic markers. Bioinformatics **37**(13), 1839–1845 (2021)

[29] Silva, G.G.Z., Cuevas, D.A., Dutilh, B.E., Edwards, R.A.: Focus: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ **2**, 425 (2014)

[30] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using diamond. Nature methods **12**(1), 59–60 (2015)

[31] Gautam, A., Felderhoff, H., Bağci, C., Huson, D.H.: Using annotree to get more assignments, faster, in diamond+ megan microbiome analysis. Msystems **7**(1), 01408–21 (2022)

[32] Gautam, A., Zeng, W., Huson, D.H.: Meganserver: facilitating interactive access to metagenomic data on a server. Bioinformatics **39**(3), 105 (2023)

[33] Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C.: Megan analysis of metagenomic data. Genome research **17**(3), 377–386 (2007)

[34] Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., Schuster, S.C.: Integrative analysis of environmental sequences using megan4. Genome research **21**(9), 1552–1560 (2011)

[35] Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., Tappu, R.: Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. PLoS computational biology **12**(6), 1004957 (2016)

[36] Huson, D.H., Albrecht, B., Bağcı, C., Bessarab, I., Gorska, A., Jolic, D., Williams, R.B.: Megan-lr: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biology direct **13**(1), 1–17 (2018)

[37] Arumugam, K., Bağcı, C., Bessarab, I., Beier, S., Buchfink, B., Górska, A., Qiu, G., Huson, D.H., Williams, R.B.: Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. Microbiome **7**(1), 1–13 (2019)

[38] Koslicki, D., Falush, D.: Metapalette: ak-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. MSystems **1**(3), 00020–16 (2016)

[39] Freitas, T.A.K., Li, P.-E., Scholz, M.B., Chain, P.S.: Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic acids research **43**(10), 69–69 (2015)

[40] Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., *et al.*: Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. Nature biotechnology **37**(8), 852–857 (2019)

[41] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., *et al.*: Qiime allows analysis of high-throughput community sequencing data. Nature methods **7**(5), 335–336 (2010)

[42] 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)

[43] 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)

[44] 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)

[45] 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)

[46] 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)

[47] Ounit, R., Lonardi, S.: Higher classification sensitivity of short metagenomic reads with clark-s. Bioinformatics **32**(24), 3823–3825 (2016)

[48] Menzel, P., Ng, K.L., Krogh, A.: Fast and sensitive taxonomic classification for metagenomics with kaiju. Nature communications **7**(1), 11257 (2016)

[49] Rosen, G.L., Reichenberger, E.R., Rosenfeld, A.M.: Nbc: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics **27**(1), 127–129 (2011)

[50] Rosen, G.L., Sokhansanj, B.A., Polikar, R., Bruns, M.A., Russell, J., Garbarine, E., Essinger, S., Yok, N.: Signal processing for metagenomics: extracting information from the soup. Current genomics **10**(7), 493–510 (2009)

[51] Brady, A., Salzberg, S.L.: Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. Nature methods **6**(9), 673–676 (2009)

[52] Brady, A., Salzberg, S.: Phymmbl expanded: confidence scores, custom databases, parallelization and more. Nature methods **8**(5), 367–367 (2011)

[53] Piro, V.C., Dadi, T.H., Seiler, E., Reinert, K., Renard, B.Y.: ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. Bioinformatics **36**(Supplement_1), 12–20 (2020)

[54] Piro, V.C., Lindner, M.S., Renard, B.Y.: Dudes: a top-down taxonomic profiler for metagenomics. Bioinformatics **32**(15), 2272–2280 (2016)

[55] Müller, A., Hundt, C., Hildebrandt, A., Hankeln, T., Schmidt, B.: Metacache: context-aware classification of metagenomic reads using minhashing. Bioinformatics **33**(23), 3740–3748 (2017)

[56] Ames, S.K., Hysom, D.A., Gardner, S.N., Lloyd, G.S., Gokhale, M.B., Allen, J.E.: Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics **29**(18), 2253–2260 (2013)

[57] Ames, S., Allen, J.E., Hysom, D.A., Lloyd, G.S., Gokhale, M.B.: Design

and optimization of a metagenomics analysis workflow for nvram. In: 2014 IEEE International Parallel & Distributed Processing Symposium Workshops, pp. 556–565 (2014). IEEE

[58] Van Essen, B., Hsieh, H., Ames, S., Gokhale, M.: Di-mmap: A high performance memory-map runtime for data-intensive applications. In: 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, pp. 731–735 (2012). IEEE

[59] Hanson, N.W., Konwar, K.M., Hallam, S.J.: Lca*: an entropy-based measure for taxonomic assignment within assembled metagenomes. Bioinformatics **32**(23), 3535–3542 (2016)

[60] Strous, M., Kraft, B., Bisdorf, R., Tegetmeyer, H.E.: The binning of metagenomic contigs for microbial physiology of mixed cultures. Frontiers in microbiology **3**, 410 (2012)

[61] Dröge, J., Gregor, I., McHardy, A.C.: Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. Bioinformatics **31**(6), 817–824 (2015)

[62] Marcelino, V.R., Clausen, P.T., Buchmann, J.P., Wille, M., Iredell, J.R., Meyer, W., Lund, O., Sorrell, T.C., Holmes, E.C.: Ccmetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome biology **21**(1), 1–15 (2020)

[63] Clausen, P.T., Aarestrup, F.M., Lund, O.: Rapid and precise alignment of raw reads against redundant databases with kma. BMC bioinformatics **19**, 1–8 (2018)

[64] Mende, D.R., Sunagawa, S., Zeller, G., Bork, P.: Accurate and universal delineation of prokaryotic species. Nature methods **10**(9), 881–884 (2013)

[65] Kang, D.D., Froula, J., Egan, R., Wang, Z.: Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3**, 1165 (2015)

[66] Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z.: Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ **7**, 7359 (2019)

[67] Alneberg, J., Bjarnason, B.S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C.: Binning metagenomic contigs by coverage and composition. Nature methods **11**(11), 1144–1146 (2014)

[68] Wu, Y.-W., Tang, Y.-H., Tringe, S.G., Simmons, B.A., Singer, S.W.:

Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**, 1–18 (2014)

[69] Wu, Y.-W., Simmons, B.A., Singer, S.W.: Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics **32**(4), 605–607 (2016)

[70] 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)

[71] Meyer, F., Lesker, T.-R., Koslicki, D., Fritz, A., Gurevich, A., Darling, A.E., Sczyrba, A., Bremges, A., McHardy, A.C.: Tutorial: assessing metagenomics software with the cami benchmarking toolkit. Nature protocols **16**(4), 1785–1801 (2021)

[72] 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)