More Data is Not Always Better: The Influence of RefSeq Database Grows on Metagenomic Taxonomic Classification

Leonardo Lazzaro[1], *Enrico* Rossignolo[1], and *Matteo* Comin[1,*]

[1]Department of Information Engineering, University of Padua, Italy

Abstract. Current technologies allow for the sequencing of microbial communities directly from the environment without prior culturing. One of the major problems when analyzing a microbial sample is to taxonomically annotate its reads to identify the species it contains. In order to determine the role of the reference database in taxonomic sequence classification, we examine the influence of the database over time on the performance of Kraken 2 a widely used taxonomic classification and profiling method. We reported that changes in reference database over time influenced the accuracy of metagenomic taxonomic classification, and training the classifier with more data, e.g. new species, worsen the results.

1 Introduction

Metagenomics is the scientific discipline dedicated to investigating diverse microbial samples directly sourced from their natural environments, encompassing locations such as soil, oceans, and the human body. These samples undergo scrutiny through the identification and quantification of the constituent species, as well as an examination of the gene functions present within them. This wealth of information empowers us to fathom the roles fulfilled by different microbes within a community and, ultimately, to pinpoint any pathogens that may disrupt their equilibrium.

A typical metagenomic sample comprises a collection of small DNA fragments, referred to as "reads", initially lacking taxonomic identification. The process of attributing each read to its respective taxonomic category is recognized as metagenomic classification or taxonomic binning. Over the years, various classifiers have been developed, falling into two primary categories: (1) alignment-based methods and (2) k-mers-based methods.

Tools such as BLAST[1], MegaBlast[2], and Megan[3], belonging to the first category, are nowadays difficult to use due to the inefficiency of the alignment when large reference databases are utilized. K-mers based methods, such as Clark[4], Centrifuge[5] or Kraken2[6], are way faster with respect to the alignment ones, and they reach similar performance in terms of precision.

The comparison between the several classification tools has been already deeply investigated in different papers [7–9]. Among the array of taxonomic binning tools, Kraken 2 stands out as the top performer, for this reason it is widely used. Furthermore it allows the creation

*e-mail: matteo.comin@unipd.it

and usage of custom reference databases [9]. Therefore, in our research, we have selected Kraken 2 as the representative tool of *k*-mer-based classification methods.

While the choice of a taxonomic binning tool plays a pivotal role in classification quality, the reference database is equally crucial. In this study, we delve into an evaluation of how variations in the reference database impact classification performance. This aspect, as previously highlighted in [10, 11], assumes paramount significance, particularly when considering the continuous evolution and expansion of the Reference Sequence database (RefSeq) over time. Notably, since its inaugural release in 2003, the number of organisms cataloged in RefSeq has surged from 2,005 to more than 162,000 [12]. The aim of this study is to analyze the influence of RefSeq database on the performance of taxonomic identification using a k-mer-based tool, i.e. Kraken 2 . We tested different RefSeq databases, in terms of contained domains and in terms of progressive releases, and we observed that the database growth is not always beneficial for taxonomic classification.

2 Methods

Two questions are essential to metagenomic analysis: How to accurately identify the microbes in samples and how to efficiently update the taxonomic classifier as new microbe genomes are sequenced and added to the reference database. In this work we analyzed the Kraken 2 performance while varying the database, by considering databases built in different years or containing a different amount of organisms. To investigate how classifiers change as they work on more knowledge, we downloaded different reference databases, composed of genomes that existed in past years, that served as "snapshots in time" of the RefSeq reference genome database.

2.1 Kraken 2 and RefSeq Databases

As previously mentioned, we opted for Kraken 2 to serve as a representative among *k*-mer-based classifiers. Kraken 2 has proved to be the best tool on several benchmarking studies [7–9], and for this reason is widely adopted. The fundamental operation of Kraken 2 can be succinctly outlined as follows: it constructs a reference database by associating each reference genome with its corresponding taxonomic label and a collection of representative *k*-mers/minimizers. In cases where the same *k*-mer/minimizer is shared among multiple species, it is attributed to their lowest common ancestor (LCA) within the taxonomic tree. During the classification process, representative *k*-mers/minimizers are extracted from each read and compared to those stored in the Kraken2 database. This comparison yields a set of root-to-leaf (RTL) paths within the taxonomic tree, all of which share common *k*-mers/minimizers with the target read [6]. Among these RTL paths, the one containing the greatest number of shared *k*-mers/minimizers with the read being classified is the one that dictates its taxonomic assignment. A schematic view of the Kraken 2 pipeline is summarized in Figure 1. The Kraken 2 DB is constructed from the Reference RefSeq DB. The reads are used as a query on the Kraken2 DB in order to find the taxonomic label.

Since 2020, the creators of Kraken 2 have been actively maintaining a series of indexes for various versions of the RefSeq database. For the purposes of this research, we acquired a subset of these databases (refer to Table 1 for details). Initially, our selection encompassed a range of databases from the latest release (12/24), each featuring a distinct array of species, spanning from Viruses to Plants. The databases contain NCBI taxonomic information as well as the complete genomes in the RefSeq archive. In particular, the Standard database consists of sequences in the Archea, Bacterial, and Viral domains as well as the Human Genome and a

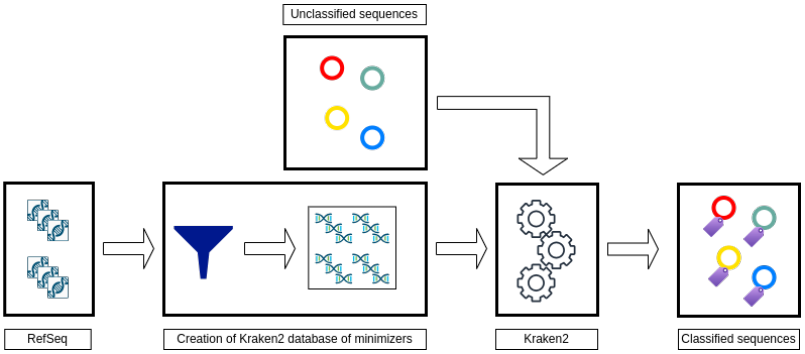

Figure 1. The construction of the Kraken 2 DB from the RefSeq database of genomic sequences (left) and the taxonomic classification of reads (right).

collection of known vectors. The PlusPF database contains Protozoa and Fungi in addition to the Standard's sequences while PlusPFP comprises the PlusPF databases and Plants genomes. The MinusB database comprises the sequences found in the Standard database, excluding bacterial sequences. Conversely, the Viral database contains only viral sequences.

Additionally, we integrated capped versions of certain databases (8GB and 16GB) into our analysis to assess Kraken 2's performance under conditions of limited available RAM. Furthermore, we expanded our collection to include different releases of the Standard database, spanning from 9/20 to 12/24. Table 1 provides an overview of all the databases employed in this study, detailing their total number of species and respective sizes in gigabytes.

Collection	Contains	No. of species	Size (GB)
Viral 12/24	Viral	13,951	0.6
MinusB 12/24	Archaea, Viral, Plasmid, Human, UniVec_Core	19,402	10.5
Standard 12/24	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	27,634	84.1
PlusPF 12/24	Standard plus Protozoa, Fungi	27,817	90.5
PlusPFP 12/24	Standard plus Protozoa, Fungi, Plant	27,953	195.2
Standard 12/24-8Gb	Standard with DB capped at 8 GB	27,476	7.5
Standard 12/24-16Gb	Standard with DB capped at 16 GB	27,543	14.9
PlusPFP 12/24-8Gb	PlusPFP with DB capped at 8 GB	27,682	7.5
PlusPFP 12/24-16Gb	PlusPFP with DB capped at 16 GB	27,770	14.9
Std 9/20	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	16,674	47
Std 12/20	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	16,823	46.8
Std 5/21	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	17,532	50.1
Std 6/22	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	19,481	58
Std 9/22	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	19,963	60
Std 3/23	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	21,439	64
Std 6/23	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	24,196	67
Std 10/23	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	24,841	70
Std 1/24	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	25,150	72
Std 6/24	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	26,512	78
Std 9/24	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	26,938	80
Std 12/24	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	27,634	84,1

Table 1. Reference Databases for different releases and number of species. Downloaded from https://benlangmead.github.io/aws-indexes/k2

2.2 Reads Datasets

We employed three distinct datasets for our testing: one simulated dataset and two real datasets. The simulated dataset was generated using the same procedure employed by the authors of Kraken 2 [6] in their strain exclusion experiment. It encompasses 40 species sourced from Bacteria and Archaea, along with an additional 10 species from Viruses.

For our real datasets, the first one is derived from a genuine human metagenome, consisting of short paired-end reads. This dataset was retrieved from the European Nucleotide Archive and represents a metagenomic sample from human feces, originally part of the Human Microbiome Project (SRR1804065). It has been utilized in previous studies, such as [13].

Dataset	No. of reads	Reads length	No. of species
Simulated	3,125,000	100	40 Bacteria and Archaea, 10 Virus
SRR1804065	5,500,983	100	591
Marine	1,150,123	[1,000-3,000]	817

Table 2. A summary of the reads datasets used for testing.

The second real dataset originates from a marine sample available from the benchmark CAMI2 [14]. It comprises long single-end reads and is a highly complex dataset, featuring more than 800 distinct species. As no 'ground truth' information is available for real metagenomes, we employed a similar evaluation procedure to previous works, such as [6, 15], and others. This procedure entails utilizing BLAST to map the reads against all reference genomes and subsequently filtering out the reads that map to two or more genomes. Following this filtering process, we retain only the reads that can be unequivocally assigned to a single genome, supported by a high-confidence match of at least 95% identity. These reads serve as our 'ground truth' for testing purposes. A summary of the three datasets is reported in Table 2.

2.3 Evaluation Metrics

We run Kraken 2 for each dataset and each reference database. The performance of the classification is evaluated and compared, using the same evaluation metrics as in [6, 16], on the basis of the three standard metrics of: Sensitivity, Precision, and F1-Score. These metrics are computed as follows:

$$Sensitivity = \frac{TP}{TP + VP + FN + FP} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$\begin{aligned} F1\text{–}Score &= \frac{2 \cdot Sensitivity \cdot Precision}{Sensitivity + Precision} \\ &= \frac{2 \cdot TP}{2 \cdot TP + VP + FN + 2 \cdot FP} \end{aligned} \tag{3}$$

where TP, VP, FN, FP indicate, respectively: True Positive (*TP*) correctly classified read; False Negative (*FN*) non-classified read; Vague Positive (*VP*) the read taxonomy classification rank is an ancestor of the rank of interest; False Positive (*FP*) incorrectly classified read.

3 Results and Discussion

We evaluate the performance of Kraken 2 for the three samples while varying the set of reference genomic sequences.

3.1 Results Varying the Reference Domains

In the first experiment, we tested how the performance of Kraken 2 varies, by adding more domains to the reference database. The results are reported in Figure 2.

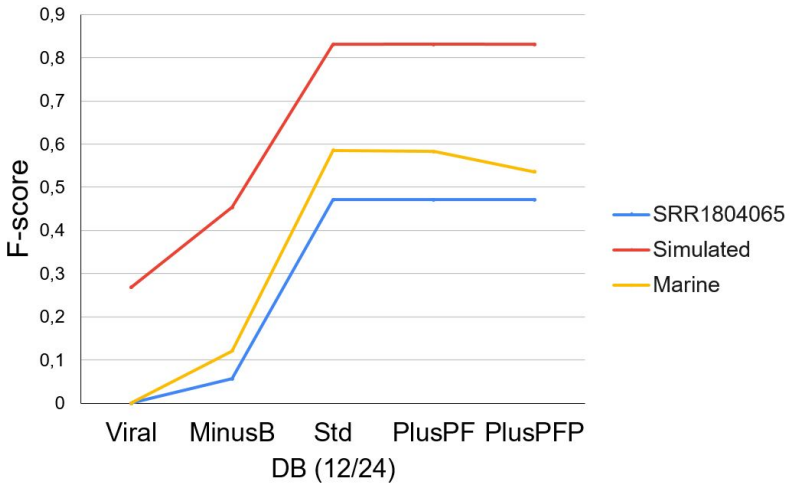

Figure 2. The species classification F1-score varying the number of species in different reference databases.

If the reference database is composed just of Viruses, no reads can be classified for the two real samples, indicating that there are no Viral species inside. Only for the simulated dataset, which contains 10 Viral species, a portion of the reads is correctly classified. As the number of species in the reference database increases, the F-score for all samples increases too, reaching its maximum value in the case of the Standard database. However, adding to the reference database species that are not present in the sample, can also be counter-productive. This is the case for PlusPFP when utilized on the marine dataset: the F-score decreased by around 5% if compared to the one obtained utilizing the Standard database.

The variation in species classifications across different database versions indicates that, even when using the same tool to analyze the same sample, the conclusions derived from the analysis can vary substantially depending on which database you are searching against. In principle, one should use the smallest reference database that contains all the species that are present in the sample. However, for a real dataset, it is impossible to know in advance all the species present in the sample, and so the selection of the reference database is crucial.

3.2 Results Using the Memory Friendly Reference DB

The use of Kraken 2 requires a large amount of RAM, for some databases more than 195 GB. However, there are cases, like real-time in-the-field analyses [17], where the users have limited computing power. For these reasons, Kraken 2 offers the possibility to cap the database

size through the downsampling of minimizers. The Standard database capped at 8 or 16 GB contains exactly the same species as the original one but allows Kraken2 to be run on an average laptop. We tested this scenario and the results are shown in Figure 3.

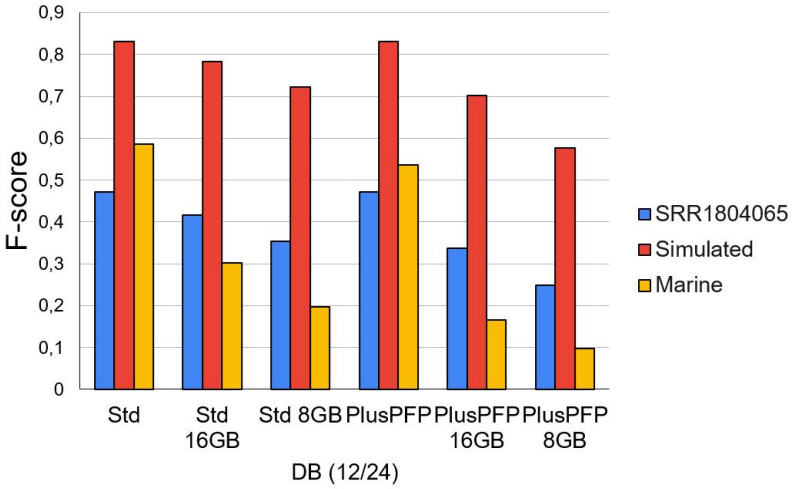

Figure 3. The species classification F1-score using different reference databases: a comparison of the original DB with the capped variants.

For all input samples, the best results are obtained with the complete reference database. If the database is capped, the classification results for all datasets drop and the lowest F-score is obtained for the smallest database (8GB). It is interesting to note that the largest drop is reported for the marine sample, which is the most complex dataset, with more than 800 species. Although these capped reference databases are extremely helpful in real-time in-the-field applications, the performances are in general not accurate and the user should be advised to re-run the classification with the non-capped database version.

3.3 Results Varying the Reference DB Over Time

In the third experiment, we evaluate the classification performance over time. We chose the Standard reference database because it is the one with the highest F-score for all samples. We tested the performance for different releases of the Standard database over time, from 2020 to 2024 (Figure 4).

For the simulated sample, that contains only 50 species, the F-score remains constant or it decrease slightly. We can observe that, for the two real samples, the F-score decreases over time, and it reaches the lowest value for the most recent release of 12/24. In particular the F-score of SRR real dataset decreases from 0.66, obtained on the 9/20 release, to 0.47 for most recent release of 12/24. A similar drop in performance is also reported for the other marine real dataset. In general, the decrease in F-score is caused by a reduction of precision and sensitivity and a higher number of false positives (see Appendix). A similar reduction of performance over time is observe also for the most comprehensive DB the PlusPFP DB (see Figure 6 in the Appendix).

The reduction in the number of correct species classifications is due to more closely related genomes appearing in RefSeq over time, making it difficult for the classifier to distinguish between them. In particular, as genomes in the reference databases increase, clade-

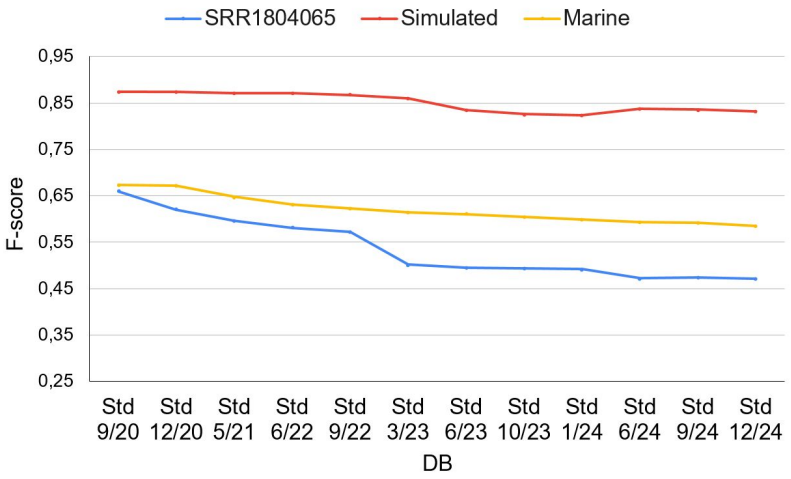

Figure 4. The species classification F-score using the Standard reference database for different releases over time.

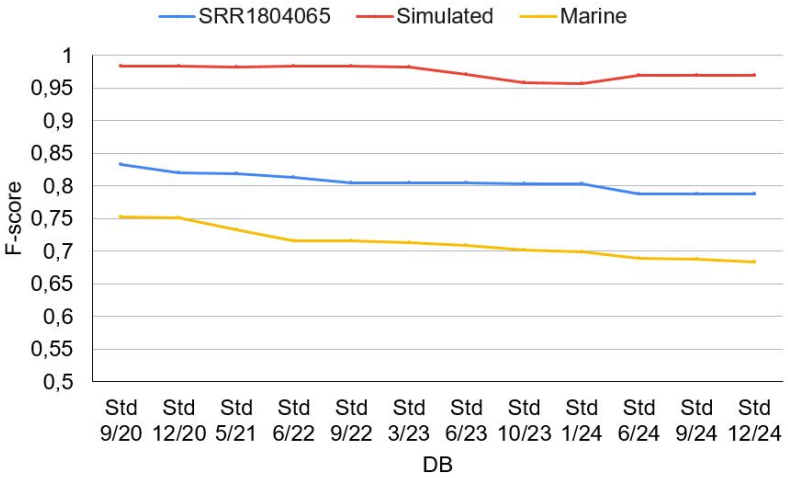

Figure 5. The genera classification F-score using the Standard reference database for different releases over time.

identifying k-mers/minimizers, that may have been previously discriminating between taxa before, are forced to move up to the genus level, as that is the lowest common ancestor (LCA). Therefore, the size of the database and its growth affect the performance of k-mer-based classification tools, where the most recent release, which contains the largest number of species, is also the worst performing.

In the last experiment, we shifted from species to genera classification and re-evaluated the classification performance of the Standard database over time (Figure 5). The behavior observed for genera classification closely mirrors that of species classification. As the database grows over time, the F-score exhibits a decreasing trend, with the most recent re-

lease showing the lowest performance. This decline is again attributed to the increasing number of closely related genomes in the reference set, which shifts taxonomic resolution upward and reduces the classifier's ability to distinguish between genera. Similar to species classification, the accumulation of additional genomes in RefSeq results in k-mers that were previously informative at the genus level becoming shared among multiple genera, thus decreasing specificity. These findings reinforce the notion that while database growth enhances overall genomic representation, it can also introduce ambiguity in taxonomic classification, particularly for closely related taxa.

4 Conclusion

The advancement of sequencing technologies has led to a constant growth of reference databases, such as RefSeq. The aim of this study is to elucidate the influence of this growth on the performance of taxonomic binning, using Kraken 2 a k-mer-based classification tools. Our analysis revealed that the database growth is not always beneficial for the taxonomic classification. In particular, filling the reference database with many species that aren't present in the analyzed sample damages the results of the classification, especially at the deeper levels of the taxonomic tree. This emerged both varying the database in terms of contained domains and in terms of progressive releases of the same database over time.

This phenomenon can be explained if we consider the LCA approach of the k-mer-based methods. As more and more species are added to the reference database, the number of k-mers mapped to higher levels of the taxonomic tree progressively grows, making classification at deeper levels more difficult.

This raises the question of whether these methods will continue to be the best choice as databases progressively grow, or whether new approaches to the taxonomic binning problem will need to be explored. In the meantime, researchers should pay attention to the choice of the reference database to use based on the expected species in the sample. In particular, an exploratory analysis of the sample could be useful, when possible, to tailor the reference database to the sample.

Acknowledgments

M.C. is supported by the Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU and by the EUAqua Project funded by the European Union under Grant Agreement 101181589.

References

[1] S.F. Altschul, W. Gish, W. Miller et al., Basic local alignment search tool, Journal of Molecular Biology **215**, 403 (1990). https://doi.org/10.1016/S0022-2836(05)80360-2

[2] Z. Zhang, S. Schwartz, L. Wagner et al., A greedy algorithm for aligning dna sequences, Journal of Computational Biology **7**, 203 (2004). https://doi.org/10.1089/10665270050081478

[3] D.H. Huson, A.F. Auch, J. Qi et al., Megan analysis of metagenomic data, Genome Research **17**, 377 (2007). https://doi.org/10.1101/gr.5969107

[4] R. Ounit, S. Wanamaker, T.J. Close et al., Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers, BMC Genomics **16** (2015). https://doi.org/10.1186/s12864-015-1419-2

D. Kim, L. Song, F. Breitwieser et al., Centrifuge: Rapid and sensitive classification of metagenomic sequences, Genome Research **26**, gr.210641.116 (2016). 10.1101/gr.210641.116

D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with kraken 2, Genome biology **20**, 257 (2019). https://doi.org/10.1186/s13059-019-1891-0

F. Meyer, A. Fritz, Z.L. Deng et al., Critical assessment of metagenome interpretation: the second round of challenges, Nature Methods **19**, 429 (2022). 10.1038/s41592-022-01431-4

A.B.R. McIntyre, R. Ounit, E. Afshinnekoo et al., Comprehensive benchmarking and ensemble approaches for metagenomic classifiers, Genome Biology **18**, 182 (2017). 10.1186/s13059-017-1299-7

S.H. Ye, K.J. Siddle, D.J. Park, P.C. Sabeti, Benchmarking metagenomics tools for taxonomic classification, Cell **178**, 779 (2019). https://doi.org/10.1016/j.cell.2019.07.010

M. Gray, Z. Zhao, G.L. Rosen, How scalable are clade-specific marker k-mer based hash methods for metagenomic taxonomic classification?, Frontiers in Signal Processing **2** (2022). 10.3389/frsip.2022.842513

D.J. Nasko, S. Koren, A.M. Phillippy, T.J. Treangen, Refseq database growth influences the accuracy of k-mer-based lowest common ancestor species identification, Genome Biology **19**, 165 (2018). 10.1186/s13059-018-1554-6

Refseq growth statistics, https://www.ncbi.nlm.nih.gov/refseq/statistics/, accessed: 2025-03-24

M. Cavattoni, M. Comin, Classgraph: Improving metagenomic read classification with overlap graphs, Journal of Computational Biology **30**, 633 (2023). 10.1089/cmb.2022.0208

A. Sczyrba, P. Hofmann, A.C. McHardy, Critical assessment of metagenome interpretation—a benchmark of metagenomics software, Nature Methods **14**, 1063–1071 (2017). https://doi.org/10.1038/nmeth.4458

D. Storato, M. Comin, K2mem: Discovering discriminative k-mers from sequencing data for metagenomic reads classification, IEEE/ACM Transactions on Computational Biology and Bioinformatics **19**, 220 (2022). 10.1109/TCBB.2021.3117406

J. Qian, D. Marchiori, M. Comin, Fast and Sensitive Classification of Short Metagenomic Reads with SKraken, in *Biomedical Engineering Systems and Technologies*, edited by N. Peixoto, M. Silveira, H.H. Ali, C. Maciel, E.L. van den Broek (Springer International Publishing, Cham, 2018), pp. 212–226, ISBN 978-3-319-94806-5

A. Latorre-Pérez, J. Pascual, M. Porcar, C. Vilanova, A lab in the field: applications of real-time, in situ metagenomic sequencing, Biology Methods and Protocols **5** (2020), bpaa016, `https://academic.oup.com/biomethods/article-pdf/5/1/bpaa016/34010873/bpaa016.pdf`. 10.1093/biomethods/bpaa016

Appendix

	Simulated			SRR1804065			CAMI2 Marine		
Database	Sens.	Prec.	F1	Sens.	Prec.	F1	Sens.	Prec.	F1
Viral 12/24	0.159	0.8817	0.269	0.000	0.001	0.000	0.001	0.000	0.000
MinusB 12/24	0.330	0.725	0.453	0.035	0.158	0.057	0.075	0.320	0.121
Std 12/24	0.733	0.961	0.831	0.360	0.681	**0.471**	0.502	0.702	**0.585**
PlusPF 12/24	0.733	0.961	**0.832**	0.360	0.680	**0.471**	0.500	0.698	**0.583**
PlusPFP 12/24	0.733	0.960	0.831	0.364	0.680	**0.471**	0.468	0.626	0.536

Table 3. Sensitivity, Precision and F-score varying reference DB with different number of species.

	Simulated			SRR1804065			CAMI2 Marine		
Database	Sens.	Prec.	F1	Sens.	Prec.	F1	Sens.	Prec.	F1
Std 12/24	0.733	0.961	**0.831**	0.360	0.681	**0.471**	0.502	0.702	**0.585**
Std 12/24-16GB	0.661	0.959	0.782	0.301	0.676	0.412	0.201	0.567	0.303
Std 12/24-8GB	0.582	0.949	0.722	0.245	0.637	0.354	0.123	0.495	0.197
PlusPFP 12/24	0.733	0.960	**0.831**	0.364	0.680	**0.471**	0.468	0.626	**0.536**
PlusPFP 12/24-16Gb	0.558	0.944	0.701	0.230	0.626	0.337	0.105	0.392	0.165
PlusPFP 12/24-8Gb	0.417	0.932	0.576	0.159	0.578	0.249	0.058	0.320	0.098

Table 4. Sensitivity, Precision and F-score varying reference DB with capped size.

	Simulated			SRR1804065			CAMI2 Marine		
Database	Sens.	Prec.	F1	Sens.	Prec.	F1	Sens.	Prec.	F1
Std 9/20	0.784	0.989	**0.875**	0.562	0.800	**0.660**	0.577	0.807	**0.673**
Std 12/20	0.783	0.989	0.874	0.519	0.770	0.620	0.577	0.806	0.672
Std 5/21	0.781	0.986	0.872	0.493	0.754	0.597	0.553	0.783	0.648
Std 6/22	0.782	0.983	0.871	0.475	0.747	0.581	0.540	0.758	0.631
Std 9/22	0.776	0.983	0.867	0.465	0.742	0.572	0.532	0.752	0.623
Std 3/23	0.771	0.973	0.860	0.390	0.703	0.502	0.523	0.744	0.614
Std 6/23	0.745	0.951	0.835	0.385	0.694	0.496	0.520	0.739	0.611
Std 10/23	0.731	0.951	0.826	0.384	0.693	0.494	0.515	0.731	0.604
Std 1/24	0.728	0.950	0.824	0.381	0.692	0.492	0.511	0.724	0.599
Std 6/24	0.741	0.962	0.837	0.362	0.681	0.473	0.506	0.715	0.593
Std 9/24	0.739	0.961	0.836	0.362	0.682	0.473	0.505	0.714	0.592
Std 12/24	0.733	0.961	0.831	0.360	0.681	0.471	0.502	0.702	0.585

Table 5. Sensitivity, Precision and F-score varying reference DB for different releases.

	Simulated			SRR1804065			CAMI2 Marine		
Database	T. P.	F. P.	F. N.	T. P.	F. P.	F. N.	T. P.	F. P.	F. N.
Viral 12/24	495,550	67,191	2,476,642	34	34,878	5,464,892	12	49,330	1,100,707
MinusB 12/24	1,030,321	390,611	1,565,559	187,832	1,002,668	4,127,194	86,070	182,724	865,459
Std 12/24	2,290,085	93,367	995	1,935,300	907,982	112,783	576,534	244,434	157,602
PlusPF 12/24	2,290,454	93,560	944	1,935,592	908,859	112,425	574,378	248,461	153,344
PlusPFP 12/24	2,290,153	94,231	338	1,935,514	912,216	108,755	537,822	321,655	90,492

Table 6. True Positive, False Positive and False Negative varying reference DB with different number of species.

	Simulated			SRR1804065			CAMI2 Marine		
Database	T. P.	F. P.	F. N.	T. P.	F. P.	F. N.	T. P.	F. P.	F. N.
Std 12/24	2,290,085	93,367	995	1,935,300	907,982	112,783	576,534	244,434	157,602
Std 12/24-16GB	2,064,934	89,378	9,244	1,617,181	776,418	213,873	237,482	181,546	642,340
Std 12/24-8GB	1,819,463	97,245	103,011	1,319,082	751,217	449,506	140,934	143,838	811,554
PlusPFP 12/24	2,290,153	94,231	338	1,935,514	912,216	108,755	537,822	321,655	90,492
PlusPFP 12/24-16Gb	1,742,843	102,864	157,266	1,239,719	739,726	568,632	120,226	186,433	792,968
PlusPFP 12/24-8Gb	1,302,424	95,409	687,153	859,579	627,856	1,514,485	66,559	141,686	914,210

Table 7. True Positive, False Positive and False Negative varying reference DB with capped size.

	Simulated			SRR1804065			CAMI2 Marine		
Database	T. P.	F. P.	F. N.	T. P.	F. P.	F. N.	T. P.	F. P.	F. N.
Std 9/20	2,449,076	26,942	1,505	3,016,979	755,082	63,748	663,207	158,811	195,278
Std 12/20	2,447,603	28,154	1,446	2,789,606	833,219	169,616	662,619	159,663	195,126
Std 5/21	2,441,346	34,665	1,381	2,649,557	863,673	150,452	635,170	176,481	205,986
Std 6/22	2,443,065	41,711	1,259	2,551,445	862,361	140,747	620,050	197,425	187,237
Std 9/22	2,425,204	42,011	1,233	2,499,183	870,608	132,947	610,898	201,405	184,162
Std 6/23	2,326,771	118,764	30,889	2,069,282	910,401	129,406	597,583	211,176	176,256
Std 3/23	2,408,058	66,523	3,784	2,096,878	884,603	129,580	600,751	206,762	178,576
Std 10/23	2,282,860	118,477	29,922	2,059,836	910,448	122,669	591,254	217,671	172,699
Std 1/24	2,273,805	118,804	30,142	2,048,269	910,211	121,535	586,785	223,370	171,250
Std 6/24	2,315,107	91,914	1,038	1,944,348	911,484	115,451	581,518	231,405	165,290
Std 9/24	2,310,482	92,683	1,025	1,946,615	905,972	114,985	580,519	232,900	162,207
Std 12/24	2,290,085	93,367	995	1,935,300	907,982	112,783	576,534	244,434	157,602

Table 8. Results varying reference DB for different releases.

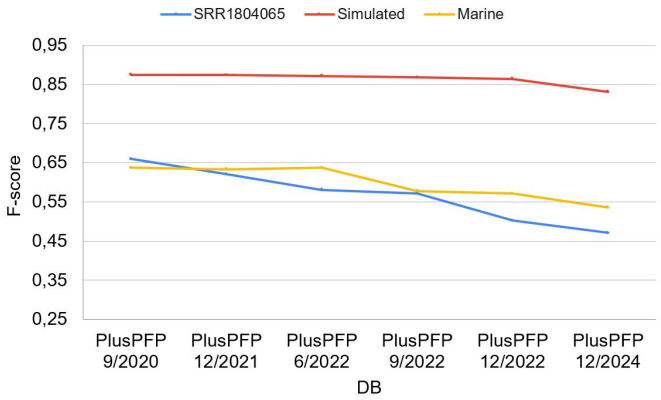

Figure 6. The species classification F-score using the PlusPFP reference database for different releases over time.