frontiers in High Performance Computing

TYPE Original Research
PUBLISHED 05 August 2026
DOI 10.3389/fhpcp.2026.1860848

Check for updates

OPEN ACCESS

EDITED BY
Umberto Ferraro Petrillo,
Sapienza University of Rome, Italy

REVIEWED BY
Ximei Luo,
University of Electronic Science and Technology of China, China
Johan Nystrom Persson,
JNP Solutions, Japan

*CORRESPONDENCE
Matteo Comin
matteo.comin@unipd.it

RECEIVED 20 April 2026
REVISED 28 May 2026
ACCEPTED 16 July 2026
PUBLISHED 05 August 2026

CITATION
Lazzaro L, Rossignolo E and Comin M (2026) When bigger is not better: the impact of RefSeq growth on Kraken2 classification accuracy.
Front. High Perform. Comput. 4:1860848.
doi: 10.3389/fhpcp.2026.1860848

COPYRIGHT
© 2026 Lazzaro, Rossignolo and Comin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

When bigger is not better: the impact of RefSeq growth on Kraken2 classification accuracy

Leonardo Lazzaro, Enrico Rossignolo and Matteo Comin*

Department of Information Engineering, University of Padova, Padova, Italy

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 this study, we delve into an evaluation of how variations in the reference database impact classification performance. This aspect assumes paramount significance, particularly when considering the continuous evolution and expansion of the NCBI Reference Sequence database (RefSeq) over time. The aim of this study is to analyze the influence of RefSeq database size and composition on taxonomic identification performance using Kraken 2, a widely used taxonomic classification and profiling method. We found that resource-efficient Kraken 2 Lite (capped) databases, which can be run on standard laptops, generally exhibit lower classification accuracy compared to full reference databases that require high-performance computing infrastructure. The size of the RefSeq database and its growth affect the performance of the k-mer-based algorithms in addition to the computing resources need. We reported that changes in RefSeq reference database over time influenced the accuracy of metagenomic taxonomic classification, and training the classifier with more data, for example, new species, worsen the results especially on the most complex microbial communities.

KEYWORDS
k-mers, Kraken2, metagenomics, reference database, taxonomic classification

1 Introduction

Metagenomics is the study of the heterogeneous microbes samples (e.g., soil, water, and human microbiome) directly extracted from the natural environment with the primary goal of determining the taxonomical identity of the microorganisms residing in the samples. It represents an evolutionary shift in focus, moving from the study of individual microorganisms to the analysis of complex microbial communities. The classical genomic-based approaches require the prior clone and culturing for further investigation (Mande et al., 2012). However, not all Bacteria can be cultured. The advent of metagenomics succeeded to bypass this difficulty. Microbial communities can be analyzed and compared through the detection and quantification of the species they contain.

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) sequence-composition-based methods, which are based on the nucleotide composition (e.g., *k*-mers usage).

Traditionally, the first strategy was to use BLAST (Altschul et al., 1990) to align each read with all sequences in NCBI RefSeq. Later, faster methods have been deployed for this task, popular examples are MegaBlast (Zhang et al., 2004) and Megan (Huson et al., 2007). However, as the reference databases and the size of sequencing data sets have grown, alignment has become computationally infeasible, leading to the development of metagenomics classifiers that provide much faster results.

The fastest and most promising approaches belong to the composition-based one (Lindgreen et al., 2016). The basic principles can be summarized as follows: each genome of reference organisms is represented by its k-mers, and the associated taxonomic label of the organisms, then the reads are searched and classified throughout this k-mers database. Several k-mer based methods have been proposed, such as Clark (Ounit et al., 2015), Centrifuge (Kim et al., 2016), or Kraken2 (Wood et al., 2019). The precision of these methods is as good as MegaBlast (Zhang et al., 2004), nevertheless, the processing speed is much faster (Lindgreen et al., 2016). The comparison between the several classification tools has been already deeply investigated in different papers (Meyer et al., 2022; McIntyre et al., 2017; Ye et al., 2019). 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 and usage of custom reference databases (Ye et al., 2019). We thus selected Kraken2 as representative tool from the genre of k-mer-based classification methods. The focus of this study was not to examine a specific software tool, but rather to evaluate the performance of a representative k-mer-based method given database variability over time.

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 Gray et al. (2022) and Nasko et al. (2018), 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 170,000 (Goldfarb et al., 2024).

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, that is, Kraken 2. We tested different RefSeq databases, in terms of contained domains (e.g., Virus, Bacteriae, etc.) and in terms of progressive releases, and we observed that the database growth is not always beneficial for taxonomic classification. We also found that, although the Kraken2 Lite (capped) reference databases can be run on laptops with limited computational resources, their classification accuracy is generally lower than that achieved using the full reference databases, which typically require high-performance computing systems.

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 study, we assess the performance of Kraken 2 under varying reference RefSeq database configurations, considering databases constructed in different years and containing different numbers of organisms[1]. For completeness, we first provide a brief overview of Kraken 2. We then describe our methodology for evaluating the impact of different RefSeq releases on classification performance. To examine how Kraken 2 behaves as the available reference knowledge expands, we downloaded multiple historical RefSeq releases, each representing a "snapshot in time" of the reference genome database. These reference datasets were used to classify several metagenomic datasets, and classification accuracy was evaluated at the level of individual reads and systematically compared across different RefSeq releases.

2.1 Kraken 2: background

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 (Meyer et al., 2022; McIntyre et al., 2017; Ye et al., 2019), and for this reason is widely adopted.

Kraken 2 is an improved version of the classifier Kraken (Wood and Salzberg, 2014) regarding memory usage and classification time. These improvements were obtained thanks to the use of the minimizers and a probabilistic compact data structure, instead of the k-mers and a sorted list used in Kraken. Instead of utilizing the complete genome as reference, Kraken 2 considers only its minimizers, that are a special type of *k*-mers, thus a genome sequence is alternatively represented by a set of minimizers, which plays a role of efficiently indexing a large volume of target-genomes database, for example, all the genomes in RefSeq. This idea is borrowed from alignment-free methods (Vinga and Almeida, 2003).

Kraken2 constructs a data structure that is an augmented taxonomic tree in which a list of significant minimizers are associated to each node, leaves and internal nodes. Given a node on this taxonomic tree, its list of minimizers is considered representative for the taxonomic label and it will be used for the classification of metagenomic reads. In the classification step, each read is decomposed into its minimizers which are then searched for in the tree.

At first, Kraken 2 needs to build a database starting from a set of reference genomes. To build the database Kraken 2 downloads from the NCBI the taxonomy and reference sequences libraries (RefSeq) required. With the taxonomy data, Kraken 2 builds a tree where each node is associated to a taxonomy ID. In each tree node, a list of minimizers is stored that is useful for the classification. Precisely, for each minimizer if it is unique to a reference sequence then, it is associated with the node with the sequence's taxonomy ID. Instead, if the minimizer belongs to more

1 A preliminary version of this work appeared in the proceedings of ICBBB 2026 (Lazzaro et al., 2026).

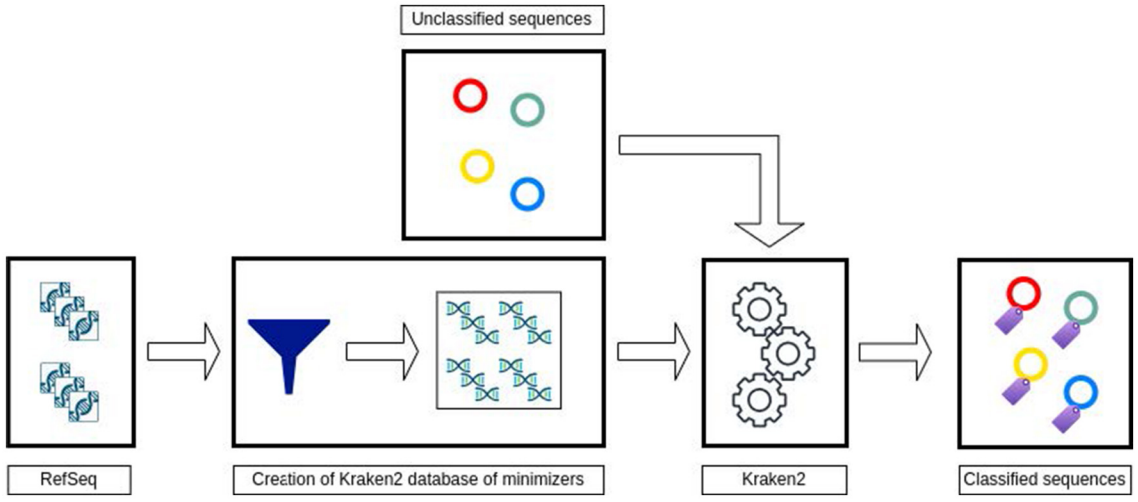

FIGURE 1
The construction of the Kraken 2 DB from the RefSeq database of genomic sequences **(left)** and the taxonomic classification of reads **(right)**.

than one reference sequence then, the minimizer is moved to the node with taxonomy ID equals to the Lowest Common Ancestor (LCA) of the two sequences it belongs. All the minimizer-taxonomy ID pairs are saved in a probabilistic Compact Hash Table (CHT) which allows a reduction of memory with respect to Kraken. During the classification process, minimizers are extracted from each read and compared to those stored in the Kraken2 database. When the minimizer is queried in the compact table, the node taxonomy ID of the augmented tree is returned, and the counter in the node is increased. Once all the minimizers in the read have been analyzed, the read is classified by choosing the deepest node from the root with the highest-weight path in the taxonomy tree.

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 and each read is classified and assigned to a taxonomic label.

2.2 Kraken 2 and RefSeq databases

The NCBI RefSeq database serves as a comprehensive repository of curated reference genomes spanning all domains of life. In recent years, RefSeq has experienced substantial growth due to the accelerated discovery of new species. This expansion has been largely driven by the decreasing cost of high-throughput sequencing technologies and the availability of efficient genome assembly pipelines. As a result, the increasing taxonomic diversity and genomic content are directly reflected in successive RefSeq releases.

To systematically evaluate Kraken2 across different RefSeq versions, it is necessary to reconstruct the database as it existed at specific time points and generate the corresponding k-mer indices. Since 2020, the authors of Kraken2 have maintained pre-built indices corresponding to multiple RefSeq releases, thereby enabling reproducible benchmarking across database versions. For this study, we selected a subset of these publicly available Kraken2 indices (see Table 1).

The first group of indices was derived from release 12/24 and includes databases with varying taxonomic compositions, ranging from viral-only datasets to comprehensive multi-domain collections. Each database integrates complete RefSeq genomes alongside the associated NCBI taxonomic annotations. Specifically, the Standard database includes sequences from the Archaea, Bacteria, and Viruses domains, as well as the human genome and a curated set of known vectors. The PlusPF database extends the Standard collection by incorporating Protozoa and Fungi, whereas PlusPFP further expands the database to include Plant genomes. The MinusB database corresponds to the Standard collection with Bacterial sequences removed, and the Viral database contains exclusively viral genomes. Although all these indices originate from the same RefSeq release (12/24), they differ in taxonomic scope. The spectrum ranges from the Viral index, containing only viral sequences, to the comprehensive PlusPFP database encompassing all domains of life. Across these builds, the number of represented species ranges from 13,951 to 27,953. Database sizes vary accordingly, from less than 1 GB for the Viral index to approximately 195 GB for the most comprehensive index. This collection enables a systematic assessment of how increasing taxonomic breadth affects Kraken2 classification performance. In particular, we evaluate the trade-off between expanded domain coverage and computational efficiency.

Most full-scale indices require substantial RAM to load the k-mer database into memory. To address hardware limitations, the Kraken2 developers provide “capped” versions of selected databases, reduced to predefined target sizes (8 GB and 16 GB). These compact indexes decrease memory requirements at the expense of reduced sensitivity and classification accuracy. We incorporated capped versions (8 GB and 16 GB) of the Standard and PlusPFP indexes into our analysis to evaluate Kraken2 performance under constrained memory conditions. This approach allows us to quantify the impact of index compression on classification outcomes.

Finally, we expanded our dataset to include multiple historical releases of the Standard index, spanning from 09/2020 to 12/2024.

TABLE 1 Reference databases for different releases and number of species.

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	46.8
Std 12/20	Archaea, Bacteria, Viral, Plasmid, Human, UniVec_Core	16,823	47
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

Downloaded from https://benlangmead.github.io/aws-indexes/k2.

This longitudinal series enables the assessment of the most important question of this study: how does the temporal expansion of RefSeq influence Kraken2 classification performance? Notably, even within the Standard index—whose taxonomic domains remain constant (Archaea, Bacteria, and Viruses)—the number of species increased from 16,674 in 2020 to 27,634 in 2024. Correspondingly, the database size nearly doubled, growing from 47 GB to 84 GB over four years. Table 1 summarizes all databases analyzed in this study, including their respective species counts and total sizes in gigabytes.

2.3 Reads datasets

To assess the classification performance of Kraken2, we employed three different reads datasets: a simple simulated dataset, a complex simulated dataset, and a real metagenomic reads dataset.

The first reads dataset was constructed following the same protocol used by the Kraken2 authors in their strain exclusion experiment (Wood et al., 2019). It comprises 40 species from the Bacteria and Archaea domains, supplemented with 10 viral species. This dataset provides a controlled and relatively low-complexity benchmark for evaluating classification accuracy under well-defined conditions.

The second reads dataset consists of a complex marine metagenomic sample from the CAMI2 benchmark (Sczyrba et al., 2017). This realistic simulated dataset has been widely adopted for method evaluation in metagenomic studies. It contains long single-end reads and represents a highly complex microbial community with more than 800 distinct species, thereby offering a challenging scenario for taxonomic classification.

For the real-data evaluation, we selected a human gut metagenomic sample composed of short paired-end reads. The dataset was obtained from the European Nucleotide Archive (accession SRR1804065) and originates from the Human Microbiome Project. This sample has been used in several previous benchmarking studies (Sobih et al., 2016).

As no "ground truth" information is available for real reads metagenome, we employed a similar evaluation procedure to previous works, such as Wood et al. (2019) and others. Specifically, reads were aligned against the complete set of reference genomes using BLAST. Reads mapping to multiple genomes were discarded to avoid ambiguous assignments. From the remaining alignments, we retained only those reads that could be uniquely assigned to a single reference genome with at least 95% sequence identity. These high-confidence, uniquely mapped reads were considered as the ground truth for performance assessment. Moreover, as the reference genome collection changes across RefSeq releases, the ground truth set was reconstructed separately for each database version to ensure consistency with the corresponding reference set. A summary of the three reads datasets is reported in Table 2.

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

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

2.4 Evaluation metrics

At first we run Kraken 2 to classify each read dataset (Table 2) using each reference database (Table 1). For each run we evaluate the classification and the performance are compared using the same evaluation metrics as in Wood et al. (2019); precisely we use Sensitivity, Precision, F-measure, and Pearson Correlation Coefficient (PCC) of the abundance profile. To compute these measures, we use the same methodology used in Wood et al. (2019) as following explained. We firstly compare the classifier's results with the ground truth file of the dataset classified, at a specific taxonomy level, to obtain the number of reads belonging to the following categories: true positive (TP), vague positive (VP), false negative (FN), and false positive (FP). A read belongs to the TP category if its taxonomy classification rank is the same or is a descendant of the truth rank. If the read taxonomy classification rank is an ancestor of the truth rank of interest, it is considered as Vague Positive (VP). A read belongs to the FN category if the classifier fails to classify the sequence. Lastly, a read belongs to FP if its classification is incorrect.

In line with previous studies (Wood et al., 2019), we define sensitivity as the number of reads that are correctly classified divided by the total number of reads. Precision is defined as the number of correctly classified reads divided by the number of predictions made, note that reads that have no match (FN) or that are not classified at a given taxonomic rank or below (VP) are not counted. The evaluation metrics are computed as follows (Equations 1–4):

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

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

$$F\text{–}Score = \frac{2 \cdot Sensitivity \cdot Precision}{Sensitivity + Precision} \tag{3}$$

It is worth noting that the effect of vague positives could be evaluated using more sophisticated metrics, such as the Read Index (Nyström-Persson et al., 2025; Wright et al., 2023). However, the computation and analysis of these measures are beyond the scope of this work.

The Pearson Correlation Coefficient is defined as follows. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, PCC is defined as:

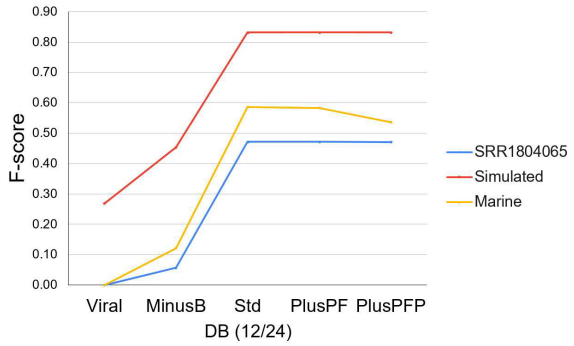

FIGURE 2
The species classification F-score varying the number of domains in the reference databases.

$$PCC = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{4}$$

where: n is the amount of species, x_i, y_i are, respectively, the number of reads correctly classified for the i-th species and the total number of reads belonging to the i-th species present in the ground truth, and $\bar{x}, \bar{y}$ are, respectively, the sample mean of x and y. The Pearson Correlation Coefficient reflects the ability of a classifier to detect the correct abundance ratio of species.

We used Kraken 2, version 2.1.6, with default parameters k = 35, m = 31, s = 7, and confidence threshold 0.0. We didn't experiment with different parameters settings like in Wright et al. (2023). All our tests are executed on a server equipped with Intel Xeon Platinum 8260 CPU @ 2.40GHz and 1TB of RAM. Kraken 2 was run on multithreading using 16 threads.

3 Results and discussion

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

3.1 Results varying the reference domains

In the first experiment, we evaluated how Kraken 2 classification performance changes as additional taxonomic domains are incorporated into the reference database, while keeping the RefSeq release fixed (12/24). The corresponding results are shown in Figure 2.

When the reference database included only viral genomes, no reads were classified for either of the more complex datasets, indicating the absence of viral species in those samples. In contrast, for the simple simulated dataset—which contains 10 viral species—a fraction of the reads was correctly assigned.

As the taxonomic breadth of the reference database increased, the F-score improved across all datasets, reaching its highest value with the Standard database. This suggests that expanding the

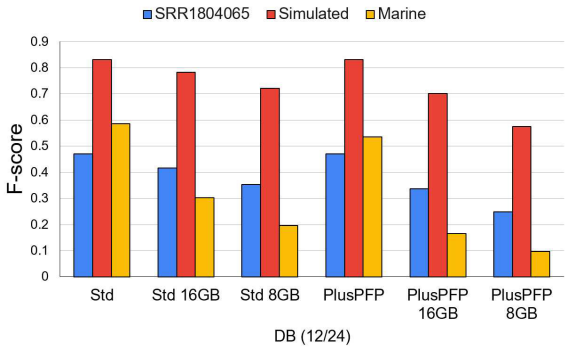

FIGURE 3
The species classification F-score using different reference databases: a comparison of the original DB with the capped variants.

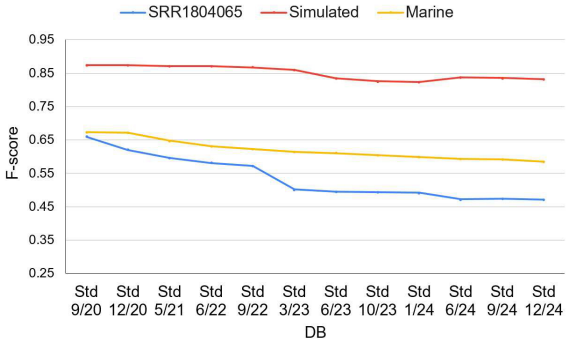

FIGURE 4
The species classification F-score using the Standard reference database for different releases over time.

reference to include the relevant domains enhances classification accuracy when those domains are represented in the sample.

However, incorporating genomes from domains not present in the sample may negatively affect performance. This effect was observed for the PlusPFP database when applied to the marine dataset, where the F-score decreased by approximately 5% compared to the Standard database. As detailed in the Supplementary material, this decline was primarily driven by an increase in false positives, leading to an overall reduction in the F-score. One possible explanation is that the inclusion of plant genomes in the PlusPFP database might introduce spurious matches in the marine sample.

Overall, the differences observed across RefSeq databases with varying domain compositions demonstrate that classification outcomes can differ substantially depending on the chosen reference database, even when analyzing the same sample with the same tool. Ideally, the reference database should be restricted to the smallest set of domains that fully represents the taxa present in the sample. In practice, however, the true taxonomic composition of real metagenomic samples is unknown a priori, making the selection of an appropriate reference database a critical and nontrivial step in the analysis pipeline.

3.2 Effect of reference database capping

Kraken2 requires substantial memory resources because the entire k-mer index must be loaded into RAM during execution. For the largest reference builds, memory requirements can exceed 195 GB. Such demands may be prohibitive in scenarios involving limited computational infrastructure, such as real-time, in-field sequencing applications (Latorre-Pérez et al., 2020), where only portable or low-memory devices are available. To address this limitation, Kraken2 provides the option to cap the size of the k-mer index by downsampling k-mers during database construction. For example, the Standard database capped at 8 GB or 16 GB retains approximately the same set of species as the full version but reduces the total number of indexed k-mers, allowing the software to run on a typical laptop. However, because capped databases contain fewer k-mers, a decrease in classification performance is expected. To evaluate this trade-off, we tested the 8 GB and 16 GB capped versions of both the Standard database and the largest PlusPFP database. The results are summarized in Figure 3.

Across all datasets, the highest F-scores were consistently achieved using the full, non-capped reference databases. When capped databases were used, performance declined for every sample, with the lowest F-scores observed for the smallest (8 GB) versions. These findings indicate that reducing the number of indexed k-mers substantially impacts classification accuracy, underscoring the importance of k-mer redundancy in Kraken2's decision process. The performance degradation was particularly pronounced for the marine dataset, which represents the most complex community, comprising more than 800 species. This suggests that complex microbial communities are especially sensitive to reductions in reference database resolution.

Although capped databases are highly valuable for real-time and resource-constrained applications, their reduced accuracy highlights the need for caution. When feasible, results obtained with capped databases should be validated by re-running the analysis using the full, non-capped reference build.

3.3 Results varying the reference database over time

In the third experiment, we investigated how the temporal evolution of the RefSeq database influences classification performance. We focused on the Standard reference database, as it consistently achieved the highest F-score across all datasets in previous analyses. Performance was evaluated across multiple Standard database releases spanning from September 2020 (9/20) to December 2024 (12/24) (Figure 4).

For the simple simulated dataset, which includes only 50 species, the F-score remained largely stable over time, with only a slight decrease observed in later releases. In contrast, both complex datasets exhibited a progressive decline in performance as the database expanded. The lowest F-scores were consistently obtained with the most recent release (12/24).

In particular, for the real SRR dataset, the F-score decreased from 0.66 with the 9/20 release to 0.47 with the 12/24 release.

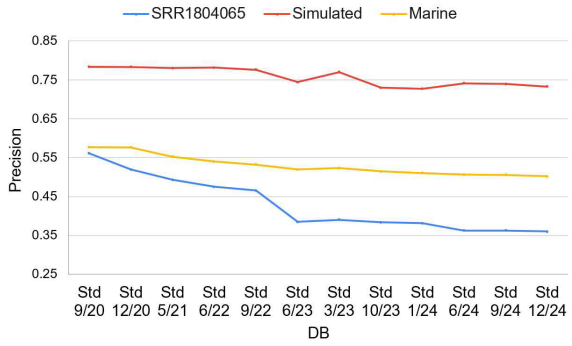

FIGURE 5
The species classification Precision using the Standard reference database for different releases over time.

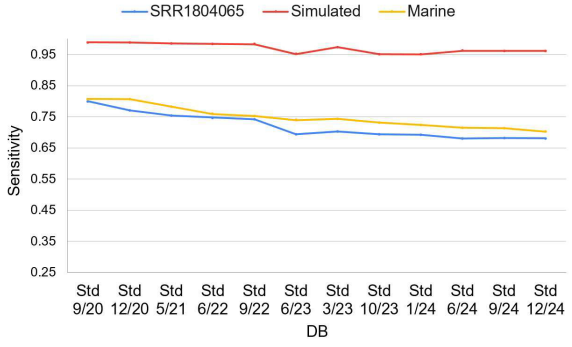

FIGURE 6
The species classification Sensitivity using the Standard reference database for different releases over time.

A comparable reduction was observed for the marine dataset. A similar downward trend was also observed when using the most comprehensive database (PlusPFP), which includes all domains (see Supplementary Figure S1 in the Supplementary material).

In Figure 4, we discussed the F-score as a measure of overall classification performance. To gain a more detailed understanding of the classifier's behavior, we further analyzed the precision and sensitivity obtained in the same experiment. Figures 5, 6 report the precision and sensitivity of species-level classification for the Standard reference database across different RefSeq releases over time.

For the simple simulated dataset, sensitivity remains relatively stable across database versions, whereas precision decreases slightly, from 0.78 for the 9/20 release to 0.73 for the 12/24 release. In contrast, both precision and sensitivity progressively decline over time for the two complex datasets. For the real SRR dataset, precision decreases from 0.56 to 0.36, while sensitivity declines from 0.79 to 0.68. A similar trend is observed for the marine dataset, where precision decreases from 0.58 to 0.50 and sensitivity from 0.80 to 0.70.

The degradation of these metrics is associated with two concurrent effects observed in the more recent database releases: fewer reads are correctly assigned to the true species (TP), and a larger number of reads are incorrectly classified (FP). The overall reduction in both precision and sensitivity is driven by the combined decrease in TP and increase in FP (see Supplementary material). A possible strategy to improve these results would be to carefully tune the confidence threshold in order to reduce the number of false positives, although this may come at the cost of a reduction in true positives (Wright et al., 2023).

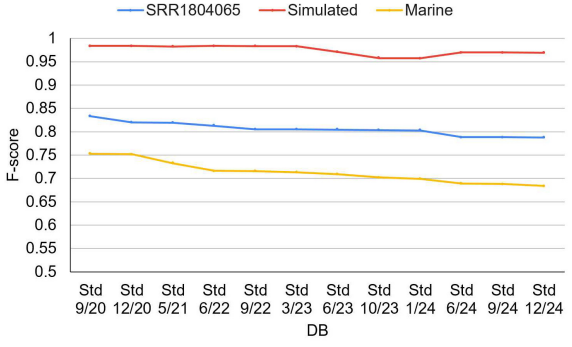

FIGURE 7
The genera classification F-score using the Standard reference database for different releases over time.

The observed decrease in correct species-level classifications appears to be associated with the growing number of closely related genomes incorporated into RefSeq over time. As additional related genomes are added, k-mers that were previously unique to specific taxa may become shared across multiple lineages. Consequently, these k-mers are reassigned to higher taxonomic levels according to the lowest common ancestor (LCA) principle. This shift reduces the discriminative power of species-level classification. Thus, while database expansion increases genomic representation, it may simultaneously reduce the specificity of k-mer-based classification methods. Notably, the most recent release—containing the largest number of species—also exhibited the lowest classification performance.

To further examine this phenomenon, we repeated the analysis at the genus level (Figure 7). The overall trend closely mirrored that observed for species-level classification. As the database size increased over time, the F-score progressively declined, with the lowest performance again corresponding to the most recent release. This reduction can be attributed to the same mechanism: the accumulation of closely related genomes in the reference set shifts informative k-mers upward in the taxonomic hierarchy, diminishing the classifier's ability to discriminate between genera. K-mers that were previously genus-specific may become shared among multiple genera, thereby decreasing taxonomic resolution and specificity.

3.4 Species detection and abundance estimation

Beyond per-read classification accuracy, metagenomic analysis also requires accurate estimation of species presence and relative abundance. To evaluate abundance estimation, we computed the Pearson Correlation Coefficient (PCC) between true and predicted

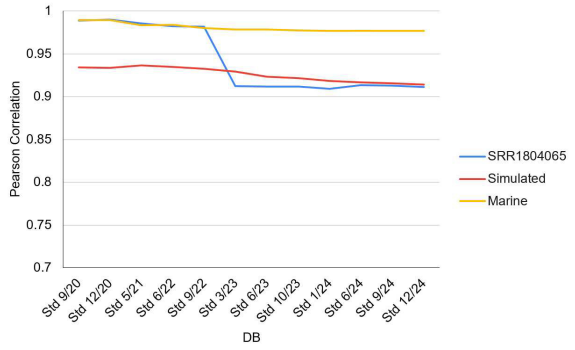

FIGURE 8
Pearson Correlation using the Standard reference database for different releases over time.

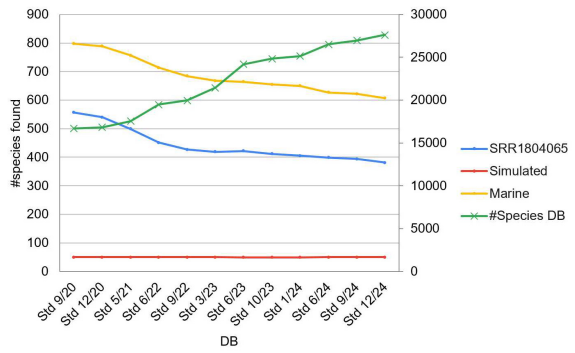

FIGURE 9
Number of Species detected using the Standard reference database for different releases over time.

species abundances for each Standard database release (Figure 8). Note that these results are based only on the standard classification output of Kraken 2. It is worth noting that this classification can be further improved with Braken (Lu et al., 2016) or other similar tools. Consistent with the F-score results, PCC values decreased in more recent releases across all datasets. A drop is observed for the SRR dataset between the Std 9/22 and Std 3/23 database releases. This reduction in PCC is primarily driven by a decrease in true positives (TP) accompanied by a simultaneous increase in false positives (FP) under this specific configuration. Overall PCC values remain relatively high, consistently above 0.9, although a slight decreasing trend can be observed over time. This indicates that database growth not only affects read-level accuracy but also impairs the ability to estimate correct species abundances.

To better understand these trends, we quantified the number of correctly detected species for each dataset across database releases (Figure 9). We consider a species to be found if at least one read is correctly classified to that species. Note that a more stringent definition can be applied, with a higher threshold. However, in this step we are not analyzing the overall number of species detected by Kraken 2, but rather whether the true species present in the sample were successfully identified. For the simple simulated dataset, all 50 species were consistently detected across releases. However, for the more complex datasets, the number of correctly identified species decreased over time. In the marine dataset, the number of detected species declined from 797 using the 9/20 release to 607 with the 12/24 release. A similar pattern was observed for the real SRR dataset. For reference, the total number of species contained in the Standard index increased from approximately 16,000 in the 9/20 release to about 27,000 in the 12/24 release. These findings suggest that the expansion of the reference database negatively affects classification performance, not only at the level of individual read assignment, as reflected by decreasing F-scores, but also in terms of species detection and abundance estimation.

Overall, the variation in classification results across database versions demonstrates that analyses performed with the same tool on the same sample can yield substantially different conclusions depending on the specific reference release used. This effect is particularly pronounced for taxa that are inherently difficult to resolve. While the growth of RefSeq enhances genomic coverage and representation, it also introduces increasing taxonomic ambiguity, especially among closely related organisms, thereby challenging the resolution of k-mer-based classifiers.

3.5 Impact of reference database growth on minimizer distribution

In the previous experiments, we evaluated the performance of Kraken2 primarily in terms of read classification accuracy. In this section, we instead analyze the internal composition of the Kraken2 databases by examining the number of minimizers associated with different taxa. Kraken2 provides the built-in command kraken2-inspect, which allows inspection of the taxonomic composition of a database by reporting the number of minimizers assigned to each taxon.

Here, we investigate the number of minimizers associated with five bacterial species: *Escherichia coli*, *Salmonella enterica*, *Pseudomonas fluorescens*, *Serratia marcescens*, and *Proteus mirabilis*. We selected *E. coli*, *S. enterica*, and *P. fluorescens* because they are extensively studied organisms with a large number of available genome assemblies and strain variants. In contrast, *S. marcescens* and *P. mirabilis* are represented by comparatively fewer genomic data. This selection does not comprehensively represent the entire taxonomic diversity contained in RefSeq. However, it provides useful insight into how the number of minimizers evolves over time and how this evolution may affect the classification observed in the previous experiments.

Table 3 reports the number of minimizers for both the Standard and PlusPFP databases for the 9/20 and 9/24 releases. The absolute number of minimizers varies considerably among the selected species. However, when comparing the Standard and PlusPFP databases within the same release (column Difference), a reduction in species-associated minimizers can be observed. These minimizers are reassigned to higher taxonomic levels as additional genomes are incorporated into the database.

Interestingly, the number of minimizers associated with a species does not evolve uniformly across releases. For species such as *Escherichia coli*, *Serratia marcescens*, and *Proteus mirabilis*, the total number of minimizers increases in the more recent release. Whereas, for *Salmonella enterica* and *Pseudomonas fluorescens*, the

TABLE 3 Number of minimizers associated to a Species, for various DB version.

	Standard		PlusPFP		Difference	
Species	9/2020	9/2024	9/2020	9/2024	9/2020	9/2024
Escherichia coli	15,351,495	23,196,177	15,343,722	23,172,019	7,773	24,158
Salmonella enterica	14,291,134	12,026,911	14,285,953	12,016,671	5,181	10,240
Pseudominas fluorescens	22,351,073	16,341,701	22,348,653	16,336,653	2,420	5,048
Serratia marcescens	7,401,286	7,997,582	7,399,532	7,992,573	1,754	5,009
Proteus mirabilis	3,457,807	3,764,977	3,453,363	3,749,991	4,444	14,986

TABLE 4 Percentage of minimizers associated to a Species, for various DB version.

	Standard		PlusPFP	
Species	9/2020	9/2024	9/2020	9/2024
Escherichia coli	0.17	0.15	0.09	0.07
Salmonella enterica	0.16	0.08	0.08	0.03
Pseudominas fluorescens	0.25	0.11	0.13	0.05
Serratia marcescens	0.08	0.05	0.04	0.02
Proteus mirabilis	0.04	0.03	0.02	0.01

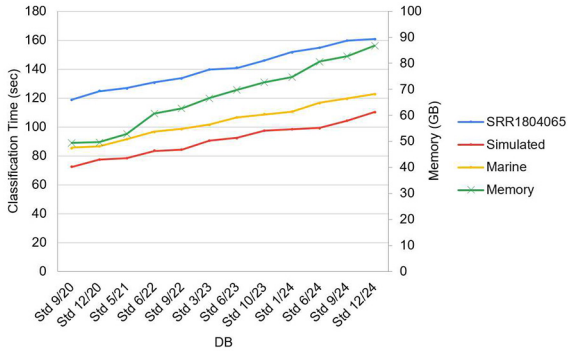

FIGURE 10
Classification time and memory usage for all inputs using the Standard reference database for different releases over time.

number decreases over time. Nevertheless, in all cases, the number of minimizers reassigned upward in the taxonomic tree is greater in the 9/24 release than in the older 9/20 release. This trend is observed regardless of whether the total number of minimizers for the species increases or decreases. As discussed previously, minimizers reassigned to higher taxonomic ranks lose their discriminative value for species-level classification.

Another potential source of classification ambiguity in Kraken2 arises from the minimizer matching strategy itself. With the default parameter setting ($s = 7$), Kraken2 masks s positions of each minimizer during the search phase. Consequently, two minimizers may be considered equivalent even if they differ at the masked positions. We suspect that, as the total number of minimizers stored in the database increases, the probability of unrelated minimizers being treated as identical may also increase.

To investigate this effect, we computed, for each selected species, the proportion of minimizers associated with that species relative to the total number of minimizers contained in the database. These percentages are reported in Table 4. Although, the absolute number of minimizers associated with a species may increase in more recent releases (see Table 3), the relative proportion of species-specific minimizers with respect to the entire database consistently decreases over time. This trend is observed for both the Standard and PlusPFP databases. Given that the total number of minimizers in the database continuously increases while the relative fraction associated with individual species decreases, this phenomenon may contribute to the reduction in classification accuracy observed for recent database releases. This can be the case, especially when using the default masking parameter ($s = 7$), where a larger number of minimizers may be considered matches. However, additional experiments will be required to further quantify the impact of this effect on Kraken2 classification performance. Furthermore, these observations should be interpreted cautiously, as a broader range of species would be necessary to establish more general conclusions.

3.6 Effect of reference database growth on classification runtime

To evaluate the computational impact of the increasing size of the RefSeq database, we measured the classification runtime of Kraken2 across different releases of the Standard reference database, spanning from September 2020 (9/20) to December 2024 (12/24). The results are summarized in Figure 10.

Across all datasets, the classification runtime increases steadily as the reference database grows. For the SRR1804065 dataset, runtime rises from 118.78 seconds using the 9/20 release to 160.78 seconds with the 12/24 release, corresponding to an increase of approximately 35%. A similar trend is observed for the simulated dataset, where runtime increases from 72.39 s to 110.39 s over the same period. The marine dataset exhibits comparable behavior, with runtime increasing from 85.63 s to 122.63 s.

This increase in computational cost is directly associated with the growth of the reference database. Over the same time span, the memory required to load the Standard index increases from 49.5 GB to 86.8 GB, reflecting the substantial expansion of the

RefSeq collection. As the database grows, Kraken2 must process a larger k-mer index during classification, which leads to longer processing times.

Despite Kraken2 being designed for fast metagenomic classification, these results indicate that the continuous expansion of reference genome collections has a measurable impact on computational performance. The increase in runtime is consistent across datasets of different complexity, suggesting that database size is a primary factor influencing classification speed.

When considered together with the previous experiments showing a decline in classification accuracy for larger database releases, these findings highlight an important trade-off. While the expansion of reference databases improves genomic coverage, it also increases both computational requirements and classification ambiguity.

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 evaluate the influence of this growth on the performance of taxonomic binning, using Kraken 2 a k-mer-based classification tools. Our results show that classification accuracy depends strongly on the composition, size, and temporal evolution of the reference database.

We found that expanding the taxonomic scope of the RefSeq database can improve classification performance when the additional domains are present in the sample, but may introduce false positives when unrelated taxa are included. Memory-constrained database builds significantly reduce classification accuracy due to the loss of k-mer redundancy. Most notably, the continuous growth of the RefSeq database over time leads to a gradual decline in species-level classification accuracy, abundance estimation, and species detection for complex metagenomic datasets.

These findings suggest that increasing the size of reference genome collections does not necessarily improve the performance of k-mer-based classifiers. Instead, the accumulation of closely related genomes can reduce taxonomic resolution by shifting discriminative k-mers toward higher taxonomic levels. At the same time, database growth increases computational requirements, both in terms of memory consumption and runtime.

Overall, our study highlights a paradox in modern metagenomics: while reference genome collections continue to expand at an unprecedented pace, this growth does not necessarily translate into improved taxonomic classification. On the contrary, the increasing number of closely related genomes introduces greater ambiguity in k-mer-based classification, reducing taxonomic resolution and increasing computational requirements. These findings suggest that future developments in metagenomic analysis should move beyond simply enlarging reference databases and instead focus on smarter strategies for reference selection, k-mer indexing, and taxonomic discrimination. Addressing these challenges will be essential for maintaining accurate and scalable metagenomic classification as genomic repositories continue to grow.

In summary, our results highlight the importance of carefully selecting and managing reference databases in metagenomic analyses. Future work should investigate strategies to mitigate the effects of database growth, such as improved k-mer selection methods, taxonomic-aware indexing strategies, or dynamic reference database construction tailored to specific datasets (Nyström-Persson et al., 2025; Storato and Comin, 2022). Such approaches may help preserve classification accuracy while maintaining the benefits of expanding genomic resources.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

LL: Data curation, Resources, Software, Writing – review & editing. ER: Data curation, Methodology, Resources, Software, Writing – original draft. MC: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. Authors are 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.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of

artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fhpcp.2026.1860848/full#supplementary-material

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., et al. (2024). Ncbi refseq: reference sequence standards through 25ăyears of curation and annotation. *Nucleic Acids Res.* 53, D243–D257. doi: 10.1093/nar/gkae1038

Gray, M., Zhao, Z., and Rosen, G. L. (2022). How scalable are clade-specific marker k-mer based hash methods for metagenomic taxonomic classification? *Front. Signal Process.* 2:842513. doi: 10.3389/frsip.2022.842513

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). Megan analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Kim, D., Song, L., Breitwieser, F., and Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomics sequences. *Genome Res.* 26, 1721–1729. doi: 10.1101/054965

Latorre-Pérez, A., Pascual, J., Porcar, M., and Vilanova, C. (2020). A lab in the field: applications of real-time, *in situ* metagenomics sequencing. *Biol. Methods Protoc.* 5:bpaa016. doi: 10.1093/biomethods/bpaa016

Lazzaro, L., Rossignolo, E., and Comin, M. (2026). "More data is not always better: the influence of refseq database grows on metagenomic taxonomic classification," in *16th international conference on bioscience, biochemistry and bioinformatics (ICBBB 2026)*, *BIO Web Conf. Vol. 232*, 01002. doi: 10.1051/bioconf/202623201002

Lindgreen, S., Adair, K., and Gardner, P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* 6:19233. doi: 10.1038/srep19233

Lu, J., Breitwieser, F. P., Thielen, P. M., and Salzberg, S. L. (2016). Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3, e104. doi: 10.1101/051813

Mande, S. S., Mohammed, M. H., and Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.* 13, 669–681. doi: 10.1093/bib/bbs054

McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., et al. (2017). Comprehensive benchmarking and ensemble approaches for metagenomics classifiers. *Genome Biol.* 18:182. doi: 10.1186/s13059-017-1299-7

Meyer, F., Fritz, A., Deng, Z. L., Koslicki, D., Lesker, T. R., Gurevich, A., et al. (2022). Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods* 19, 429–440. doi: 10.1038/s41592-022-01431-4

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

Nyström-Persson, J., Bapatdhar, N., and Ghosh, S. (2025). Precise and scalable metagenomic profiling with sample-tailored minimizer libraries. *NAR Genom. Bioinform.* 7:lqaf076. doi: 10.1093/nargab/lqaf076

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genom.* 16:236. doi: 10.1186/s12864-015-1419-2

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

Sobih, A., Tomescu, A. I., and Mäkinen, V. (2016). "Metaflow: metagenomic profiling based on whole-genome coverage analysis with min-cost flows," in *Research in Computational Molecular Biology*, ed. M. Singh (Cham: Springer International Publishing), 111–121. doi: 10.1007/978-3-319-31957-5_8

Storato, D., and Comin, M. (2022). K2mem: discovering discriminative k-mers from sequencing data for metagenomic reads classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 220–229. doi: 10.1109/TCBB.2021.3117406

Vinga, S., and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics* 19, 513–523. doi: 10.1093/bioinformatics/btg005

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

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Wright, R. J., Comeau, A. M., and Langille, M. G. I. (2023). From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microb. Genom.* 9, 000949. doi: 10.1099/mgen.0.000949

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell* 178, 779–794. doi: 10.1016/j.cell.2019.07.010

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2004). A greedy algorithm for aligning dna sequences. *J. Comput. Biol.* 7, 203–214. doi: 10.1089/106652700500 81478