ganon2: up-to-date and scalable metagenomics analysis

Vitor C. Piro[1] and Knut Reinert[1]

[1]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, 14195, Germany.

Abstract

The fast growth of public genomic sequence repositories greatly contributes to the success of metagenomics. However, they are growing at a faster pace than the computational resources to use them. This challenges current methods, which struggle to take full advantage of the massive and fast data generation. We propose a generational leap in performance and usability with ganon2, a sequence classification method that performs taxonomic binning and profiling for metagenomic analysis. It indexes large datasets with a small memory footprint, maintaining fast, sensitive, and precise classification results. Based on the full NCBI RefSeq and its subsets, ganon2 indices are on average 50% smaller than state-of-the-art methods. Using 16 simulated samples from various studies, including the CAMI 1+2 challenge, ganon2 achieved up to 0.15 higher median F1-Score in taxonomic binning. In profiling, improvements in the F1-Score median are up to 0.35, keeping a balanced L1-norm error in the abundance estimation. ganon2 is one of the fastest tools evaluated and enables the use of larger, more diverse, and up-to-date reference sets in daily microbiome analysis, improving the resolution of results. The code is open-source and available with documentation at https://github.com/pirovc/ganon.

1 Introduction

The study of the collective complete genomic content from environments bypassing the cultivation of clonal cultures is known as metagenomics. Through DNA extraction and sequencing, the collection of microorganisms in an environment can be profiled in terms of taxonomic content, abundance, and, consequently, their function [1]. Metagenomics enables the analysis of virtually any community, ranging from sewage, subway systems, soil, oceans, and the human gut, to cite a few. Applications are limitless: clinical,

industrial, and agricultural; effects of medication, diet, and lifestyle; antimicrobial resistance; pathogen detection; outbreak investigation; bioforensics; etc.

Computational metagenomics enables those applications via automated methods and analysis of data [2]. However, analyzing communities is not trivial due to their complexity, high diversity, and the limitations of available technologies in translating genomic content into understandable and error-free data. Therefore, computational metagenomics faces several methodological challenges. One of the most critical aspects of current analysis is data complexity and size.

First, obtaining sequences from unknown environments through high-throughput sequencing equipment requires deep sequencing to detect low-abundant organisms, thus producing massive data sets. Large-scale studies can generate hundreds or even thousands of samples in high throughput (millions to billions of base pairs sequenced per sample), amplifying the issue. Second, the size of assembled genomic reference sequence repositories is growing exponentially [3]. Yet, they are invariably necessary in the analysis process of most metagenomics studies. This growth is closely related to the decreased cost of sequencing that has outpaced Moore's law [4] since 2008. This all means that data is being generated at a much faster pace than the availability of computational resources to analyze it. Specialized tools and algorithms are therefore essential to scale with massive datasets, usually requiring high-performance computational resources. The goal is to enable metagenomic analyzes to scale to larger sets of data, using a limited amount of memory and in manageable time (minutes to hours per sample) while keeping sensitivity and precision high.

Due to the decrease in sequencing costs and improvements in related technologies and algorithms, large amounts of assembled genomic data are being generated and constantly deposited in public data repositories [5]. With large-scale studies, some very well-studied communities are now mostly covered by reference sequences of well-characterized species (e.g. the human gut microbiome). However, reference data is not equally distributed in the tree of life [6]. Well-described biomes represent a very small fraction of the still unknown microbial world living and evolving around us [7]. The extreme data growth is mainly biased by human-related bacteria and model organisms (e.g. *Escherichia coli*).

To extract the most information out of novel clinical or environmental data, it is essential to be up-to-date with this exponential growth of publicly available assembled genome sequences [8]. The NCBI RefSeq [6], a common resource for metagenomics reference data, had in the last 15 years (November 2009 to November 2024) an annual growth rate of around 20% in the number of species. However, there is an imbalanced representation in genomic repositories. For example, in the NCBI RefSeq, the 187 most represented species in the Archaeal, Bacterial, Fungal, and Viral Complete Genomes subsection have as many base pairs as the remaining 27,662 species (as of April 2024). In the full RefSeq, the ratio is 82 to 87,979. Moreover, the lack of metadata and clear documentation makes data selection a difficult and error-prone step. This leads most studies to use the "default", outdated, or pre-built databases, leaving behind a large amount of information. For example, a subset of RefSeq Complete Genomes has approximately only 37% of the diversity in terms of number of species compared to

the full RefSeq (Table 1). Another example: a 2-year-old outdated full RefSeq release (November 2022) has 34,208 fewer species than the actual release (November 2024).

This information is difficult to track, and it is up to the researcher to mine those repositories. Extensive studies on the subject confirmed the strong improvement of metagenomic classification with larger reference sequence databases [8, 9]. Other studies confirm the importance of reference choice to research outcomes [10–12]. The use of an outdated version of references or the inability to use all references currently available leads most methods, in the best case, to underperform and most commonly to generate inconsistent, wrong, or incomplete results.

To deal with the aforementioned challenges, we propose ganon2, a sequence classification and metagenomics profiling tool that is able to classify large amounts of sequences against large amounts of references in a short time, using low amounts of memory with high sensitivity and precision. ganon2 builds on its first version, and it provides a generational leap in terms of computational performance and results. ganon2 is open-source and available at: https://github.com/pirovc/ganon

2 Methods

2.1 ganon2

ganon2 improves and builds on its first version [13], providing a complete analytical software for metagenomics analysis. It ships an efficient algorithm for sequence classification with the Hierarchical Interleaved Bloom Filter [14, 15] from the SeqAn3 [16] library with several pre- and post-processing procedures to facilitate and improve its use in metagenomics analysis. ganon2 includes an integrated step for better database selection, download, and construction using standard (RefSeq or GenBank) or custom reference sequences, pairing them with NCBI or GTDB taxonomy, with support for assembly, strain, or sequence-level analysis. It further processes classification results to provide either sequence or taxonomic profiles, with abundance estimation including correction for genome sizes, multi-matching read reassignment with the Expectation-Maximization (EM) and/or the Lowest Common Ancestor (LCA) algorithm with multiple reporting filters and outputs for better integration into downstream analyzes. ganon2 is a user-friendly software, that automates most of its processes and provides a data-based optimized set of parameters and clear trade-offs. Thus, it helps researchers to focus on their results and analysis rather than on the data handling and tool execution.

2.2 Evaluations

ganon2 was evaluated in terms of sequence classification (binning) and abundance estimation (profiling), both using taxonomy for evaluation. Although similar, those two evaluations should be independently assessed. In binning, the ability to assign each read or sequence to its correct origin is evaluated. In profiling, the ability to detect members of a sample and their respective abundance is evaluated [17]. Binning and profiling are also evaluated with different metrics and supported by different tools.

We compare ganon2 (v2.1.0) with the state-of-the-art methods: kmcp (v0.9.4) [18], kraken2 (v2.1.3) [19], bracken (v2.9) [20], and metacache (v2.4.2) [21]. Metacache was evaluated only for binning since it does not perform taxonomic profiling. Methods were selected based on the following criteria: open-source code, actively maintained and/or highly used and cited, scalable to handle very large data in terms of database construction and sequence classification, execution in doable time (hours/few days for building, minutes/hours per sample), possibility to create custom databases with nucleotide sequences, ability to perform taxonomic binning and/or profiling.

To evaluate ganon2 and the aforementioned methods, we use the RefSeq database in 4 different configurations: (1) full, (2) complete genomes (CG), (3) representative genomes (RG), and (4) complete and representative genomes (CG+RG) (Table 1). RefSeq is a commonly used source to perform metagenomics analysis and provides a good standard of reference sequences of high quality and broad diversity. RefSeq RG contains just one selected assembly for each species but with a higher number of sequences in comparison to the RefSeq CG, some of them of lower quality with fragmented contigs. RefSeq CG is the most commonly used subset, represented only by genomes marked as complete, providing a theoretically higher quality. The full RefSeq has the highest diversity (64.616 species), and includes both CG and RG, being almost 9 times larger in raw size than the RefSeq CG. We further evaluated similar subsets with the GenBank database, and although it contains higher diversity, it achieved the overall worst results in our evaluations. ganon2 automates the download, build, and update of databases based on RefSeq and GenBank and any of their possible subsets (representative or complete genomes, by organism group or taxa, top assemblies). RefSeq provides a very generalized set of possible genomes to analyze metagenomics data. However, in studies from well-characterized environments (e.g. human gut), a customized database is recommended. ganon2 has an advanced custom build procedure with examples for commonly used data sources in the documentation: https://pirovc.github.io/ganon/custom_databases/

References

	# sequences	**# assemblies**	**# species**	**Mbp**
RefSeq	36.725.285	366.941	64.616	18.767,31
RefSeq CG	110.470	55.114	24.238	2.173,06
RefSeq RG	1.083.138	19.890	19.888	1.306,45
RefSeq CG+RG	1.183.909	69.600	37.864	3.195,85

Table 1 Reference sets used in this article. Data obtained in 2024-04-20 for archaea, bacteria, fungi, and viral groups. RefSeq CG+RG is the union of the complete and reference genome subsets.

16 short-read simulated samples with available ground truth were selected to evaluate the performance of ganon2 and other tools (Table 2). The 16 samples are very diverse and range from low- to high-complexity environments from different studies, with well-represented organisms (in relation to our chosen reference sets) to poorly represented, with some easier sets (Mende and Microba) to more challenging sets (CAMI1 and CAMI2). For the CAMI sets, we used the first replicate of each study when more than one was available. The goal here is to evaluate all methods using a spectrum

of variable diversity and difficulty, instead of performing a one-sample analysis. That way, methods can be better accessed in different scenarios, and misinterpretation of results can be partially avoided.

CAMI 1 toy [22]	**# seqs. (length)**	**Mbp**
low complexity metagenome (30 genomes)	73.924.013 (2 x 100bp)	14.784,80
medium complexity metagenome (225 genomes)	74.568.473 (2 x 100bp)	14.913,70
high complexity metagenome (450 genomes)	74.557.111 (2 x 100bp)	14.911,40
CAMI 1 challenge [22]		
low complexity metagenome	49.898.179 (2 x 150bp)	14.969,50
medium complexity metagenome	49.918.839 (2 x 150bp)	14.975,70
high complexity metagenome	49.905.935 (2 x 150bp)	14.971,80
CAMI 2 [23]		
metagenome from a **marine** environment	16.647.395 (2 x 150bp)	4.994,22
metagenome from the guts of **mice**	16.585.160 (2 x 150bp)	4.975,55
metagenome from a **plant** rhizosphere environment	16.627.054 (2 x 150bp)	4.988,12
metagenome with large amount of **strain**-level variation	6.655.023 (2 x 150bp)	1.996,51
Mende [24]		
low complexity metagenome (**10 species**)	26.666.674 (2 x 75bp)	4.000,00
medium complexity metagenome (**100 species**)	26.667.004 (2 x 75bp)	4.000,05
high complexity metagenome (**400 species**)	26.665.698 (2 x 75bp)	3.999,85
Microba [25]		
in silico mock community, medium diversity, **ani 95%**, multi-strain	6.999.997 (2 x 150bp)	2.100,00
in silico mock community, high diversity, **ani 97%**, multi-strain	6.999.988 (2 x 150bp)	2.100,00
in silico mock community, high diversity, **ani 100%**, multi-strain	7.000.009 (2 x 150bp)	2.100,00

Table 2 16 simulated short-read samples used in this article, with a total of 124.781 Mbp.

Methods were evaluated in a range of metrics, with the F1-score being central to summarize the results, since it has the power to evaluate the balance between precision and sensitivity. This balance is very important in metagenomic evaluations, since it is easy to achieve high precision or high sensitivity by classifying just a few or many sequences, respectively. In the binning evaluation, each read is classified to one taxon and set as true or false positive given the ground truth. The number of true and false positives per-read classification is then used to calculate the F1-score at a certain taxonomic level (e.g. species). In the profiling evaluation, the F1-Score is based on the final taxonomic profile generated by each method, and the presence of a taxon in the ground truth defines a true positive. Taxa reported that are not present in the truth set are false positives. Additionally, in profiling, we evaluate the abundance estimation for each taxon, and the difference from the ground truth is accounted for in the L1-norm error. For the profiling evaluation, a threshold is applied for all results, removing low-abundant taxa. This improves the performance for all methods, reducing the long tail of false positives that are usually reported, highly increasing precision with a small decrease in sensitivity. Finally, all methods were evaluated in terms of time and memory consumption, including the customized build of databases. Computational resources used to run all evaluations: AMD EPYC 9454 48-Core Processor (192 threads), 1TB RAM.

During the development of ganon2, we further created a pipeline to run tools, evaluate metrics, and visualize results called MetaBench

(https://github.com/pirovc/metabench). This pipeline was used to generate all the results presented in this article. MetaBench can evaluate taxonomic profiling and binning tools in a combination of databases, parameters, thresholds, and metrics. It outputs standardized file formats (JSON and BioBoxes [26]) and generates a dashboard to visualize them. It was also used to optimize the default parameter configuration for ganon2.

3 Results

3.1 Build

We first evaluated the performance of tools to build custom databases, based on 4 RefSeq subsets (Table 1). ganon2 was the fastest method to build the RefSeq RG (20 minutes), while kmcp was the fastest for the RefSeq CG (14 minutes) and full RefSeq (6 hours 48 minutes). kraken2+bracken took the longest to build all databases, between 8 hours and more than 4 days. Metacache has an intermediate time performance, between 1 and a half hours and 12 hours. In terms of memory necessary to build, kmcp has the lowest requirements on average for all databases evaluated, with ganon2 coming in second with intermediate memory usage. kmcp required very large amounts of disk space (several TBs for the full RefSeq) to build the database. bracken and metacache had the highest memory consumption overall, with significantly higher memory usage for the full RefSeq. Finally, when considering database size, ganon2 has, on average, the smallest sizes, with some mixed results for other tools: kmcp has the largest full RefSeq database, while kraken2 has the largest for RefSeq RG. Bracken (profiling) is not a standalone tool and depends on results from kraken2 (binning) to run; therefore, time and database sizes for bracken have to be added to kraken2 for realistic numbers. In summary, all tools besides kraken2 can efficiently build indices for large reference sets in reasonable times, with some bigger differences in memory and disk usage and database size.

We further compared the ability and easiness of tools to generate custom databases. ganon2 provides a highly automated build system. It downloads, builds, and updates any subset of RefSeq and GenBank with only one command, being the most user-friendly tool of the ones tested. Additionally, custom database creation with non-standard files, full integration with NCBI or GTDB taxonomy are also included. kraken2 has the ability to download reference data automatically, but is restricted to fewer options. However, kraken2 does not support compressed input files to build databases, requiring high storage capacities for larger databases, keeping the compressed, uncompressed, and database files in the disk at the same time. Bracken builds upon kraken2 databases, it is straightforward but adds a second step not directly integrated to kraken2 to perform taxonomic profiling. Bracken also requires users to know the size of the reads to be analyzed beforehand, which may require the construction of several databases. kmcp does not automatically download files, but there is an online documentation with guidance on how to perform customized builds. However, the commands are quite extensive and may be a challenge for less experienced users. metacache also has an online documentation on building custom databases, which is

Time

	ganon2	kmcp	kraken2	bracken	metacache
RefSeq	09:47:27	06:48:31	4 days, 01:29:26	02:06:18	12:22:11
RefSeq CG	00:52:44	00:14:15	11:43:31	00:13:51	01:37:27
RefSeq RG	00:20:01	00:21:09	08:46:14	00:09:48	01:03:34
RefSeq CG+RG	01:19:46	00:32:22	18:50:03	00:21:18	02:06:37

Memory

(GB)	ganon2	kmcp	kraken2	bracken	metacache
RefSeq	331	48	278	505	472
RefSeq CG	78	29	69	83	146
RefSeq RG	146	108	144	157	137
RefSeq CG+RG	168	79	173	188	224

Database size

(GB)	ganon2	kmcp	kraken2	bracken	metacache
RefSeq	215	494	274	0.019	406
RefSeq CG	42	57	69	0.003	119
RefSeq RG	77	34	143	0.004	100
RefSeq CG+RG	100	84	172	0.006	179

Table 3 Build time and memory consumption. Time is demonstrated in wall time (lower is better). Memory usage and database size are shown in GB (lower is better). The bracken database is not standalone, and it needs a matching kraken2 database. 64 threads were used for all tools to build the database.

well explained but adds a level of complexity that may confuse less technically inclined users.

3.2 Binning

Binning results in terms of F1-Score at the species level are presented in Figure 1. ganon2 shows a significant improvement over other methods, especially when using more diverse databases (full RefSeq, RefSeq CG, and RefSeq CG+RG). For the full RefSeq, ganon2 achieved a median F1-score of 0.77, with 75% of the samples with an F1-score of 0.53 or higher, with a well-balanced performance between precision and sensitivity as shown in Figure 2. kraken2 comes in second with a median of 0.61, followed by metacache with 0.45. kmcp could not be evaluated with the full RefSeq due to very long run times. The same trend with slightly lower values can be observed for RefSeq CG+RG, followed by RefSeq CG, both providing very good results with significantly smaller databases (Table 1). With the smallest set of references (RefSeq RG), tools have similar results, with a median of around 0.55. Metacache performed badly in this sample, having most of the assignments to lower ranks (e.g. genus, family). Those results show how tools can perform drastically differently with mixed databases and the benefits of using a highly diverse and bigger set of references. ganon2 has the smallest memory footprint classifying reads against all databases, besides kmcp which just finished the run for the smallest reference set (Table 4). Using RefSeq CG, ganon2 also achieved very good results in terms of F1-score and requires 5.1 times less memory (42 GB) compared to the full RefSeq (217 GB). metacache was the fastest tool to perform binning; however, it had a higher memory usage and poor results in this scenario. kraken2 and ganon2 have similar performance in terms of speed, with kraken2 being significantly faster with the full RefSeq. kmcp has the smallest memory

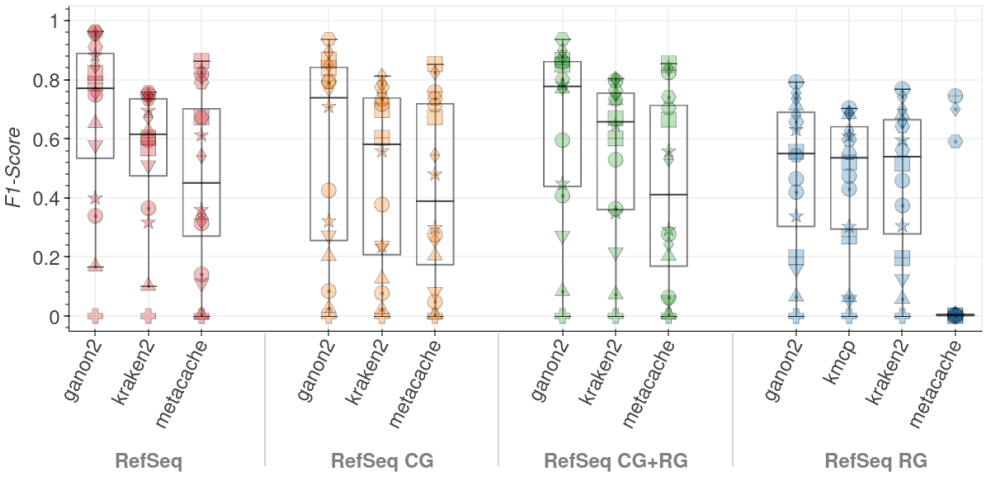

Fig. 1 F1-Score results for taxonomic binning at the species level (higher is better). Each tool and database combination (x-axis) has 16 points, one for each sample analyzed (Table 2), identified by a distinct marker, with a boxplot showing their overall distribution. The CAMI 2 mice dataset (cross) scored 0 for all methods because the ground truth is provided at the genus level.

footprint using the RefSeq RG, but it was the slowest among the tested methods, failing to provide results in a doable time (several days) for any of the largest databases.

3.3 Profiling

In profiling, ganon2 shows a highly improved F1-score at the species level with the full RefSeq, with 50% of the samples with an F1-score above 0.67 (Figure 3 TOP). In the same scenario, kraken2 + bracken achieved a median of 0.31. kmcp shows comparably good results to ganon2 with the RefSeq RG databases; however, it takes on average at least 6 times longer to run (Table 4). To exemplify the difference in runtime using the RefSeq RG in profiling: ganon2 took 43 minutes to run all 16 samples (Table 2) in this benchmark, kraken2+bracken took 59 minutes, while kmcp needed 4 hours and 47 minutes. Although fast, kraken2 + bracken performed poorly in our profiling evaluation in terms of F1-score. Here again, the results clearly show the benefits of using a larger database (full RefSeq, RefSeq CG, or RefSeq CG+RG) to profile diverse samples, yielding a substantial improvement in the results. With diverse databases, ganon2 runs fast, with the smallest memory footprint, and achieves the best performance in terms of F1-Score among 16 samples.

The L1-norm error analysis shows how the tools perform in terms of abundance estimation (Figure 3 BOTTOM). Here tools behave quite differently, with ganon2 showing the smallest interquartile range in all scenarios, followed by kraken2 + bracken and kmcp. kmcp always reports a full 100% abundance in the taxonomic profiles, not considering unmatched reads or unknown taxa. This is beneficial when samples are well represented and mostly covered by the database (Microba and Mende sets) but can generate inflated abundances when that is not the case (CAMI sets), as shown

Binning

Time (Mbp/m)	**ganon2**	**kmcp**	**kraken2**	**metacache**
RefSeq	1114 (±527)	-	2237 (±591)	1296 (±683)
RefSeq CG	2523 (±765)	-	2908 (±611)	4640 (±1192)
RefSeq RG	2345 (±547)	483 (±85)	2127 (±555)	5041 (±930)
RefSeq CG+RG	2196 (±657)	-	2285 (±688)	3603 (±1070)
Mem. (GB)	**ganon2**	**kmcp**	**kraken2**	**metacache**
RefSeq	217	-	279	435
RefSeq CG	42	-	72	124
RefSeq RG	78	34	145	108
RefSeq CG+RG	100	-	177	187

Profiling

Time (Mbp/m)	**ganon2**	**kmcp**	**kraken2+bracken**
RefSeq	1664 (±937)	-	2218 (±591)
RefSeq CG	3007 (±726)	-	2892 (±605)
RefSeq RG	2868 (±746)	434 (±103)	2118 (±556)
RefSeq CG+RG	2734 (±755)	-	2272 (±685)
Mem. (GB)	**ganon2**	**kmcp**	**kraken2+bracken**
RefSeq	216	-	279
RefSeq CG	42	-	72
RefSeq RG	78	34	145
RefSeq CG+RG	100	-	177

Table 4 Binning and profiling time and memory consumption, averaged from 16 short-read samples for every tool and database combination. Time is demonstrated in Mbp/m (higher is better) with standard deviation in brackets. Peak memory is shown in GB. Every combination of tool, database, and sample was executed 3 consecutive times, and only the best time was considered. Time represents a full run, including pre- and post-processing steps (load the database in memory, classification, read reassignment, etc.). kmcp did not finish the benchmark after several days with the full RefSeq, RefSeq CG, and RefSeq CG+RG databases; therefore it is not included. 64 threads were used.

in Figure 3. kraken2+bracken also has a higher number of reads classified, explaining the higher variation in L1-norm values. In all scenarios, ganon2 did not exceed L1-norm values of approximately 100 for all samples, keeping a low error in abundance estimation. kmcp has very low L1-norm values for some samples with RefSeq CG, with the downside of also having some of the highest L1-norm for others.

Further metrics, thresholds, and plots for the results presented in this article can be interactively explored in the Supplementary Material HTML report.

4 Discussion and Conclusion

ganon2 is a complete package for metagenomics analysis, integrating database download and construction with binning and profiling classification methods. It further brings an extensive set of exclusive functions: multiple and hierarchical database support, customizable database construction for local or non-standard sequence files, NCBI taxonomy and GTDB integration, optional LCA or EM-algorithm to solve multi-matching reads, classification of short and long reads, optimized taxonomic binning and classification configurations, advanced report generation and filtering, and easy integrated analysis at any taxonomic rank, strain, assembly, file, sequence, or

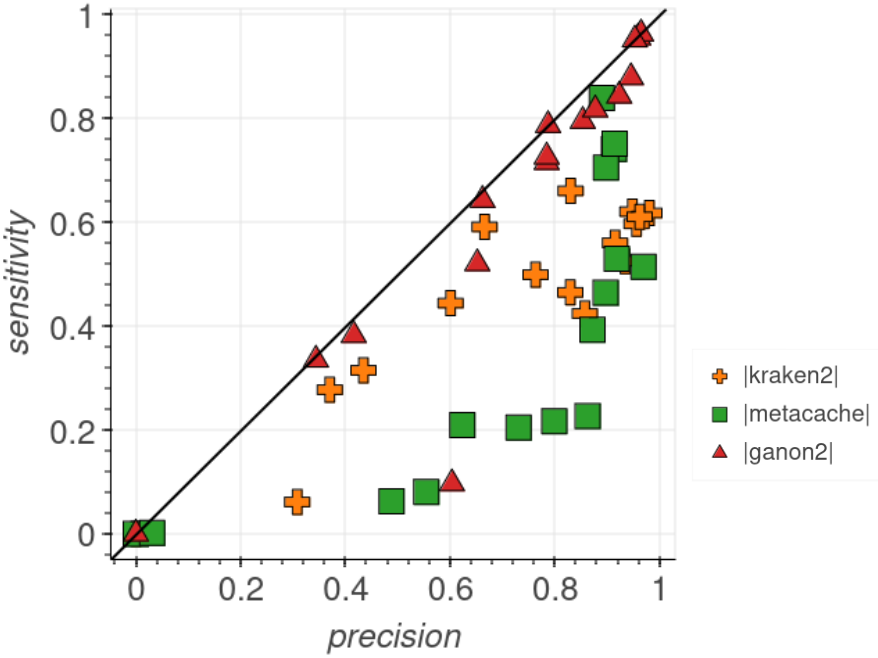

Fig. 2 Sensitivity and precision plot for taxonomic binning at the species level against the full RefSeq. Each tool has 16 points, one for each sample analyzed (Table 2). ganon2 achieved balanced results between the metrics.

custom specialization level. Coupled with a set of simple parameters and extensive documentation, ganon2 is a very strong and robust method for exploratory metagenomic classification and data analysis.

Using 16 diverse, realistic, and complementary samples against 4 different genomic reference sets based on RefSeq, ganon2 achieved consistently better results than state-of-the-art methods, keeping a top performance in terms of run-time and memory consumption. Although the tools evaluated perform similarly well using standard size databases, ganon2 excels when using more data, scaling well. This gives more options for researchers, especially when describing under-represented environments, where a combination of many resources is necessary to minimally profile samples. ganon2 is able to classify more reads correctly, increasing the resolution of results.

The improvements in ganon2 were achieved with a combination of factors. First, the use of the Hierarchical Interleaved Bloom Filter paired with minimizers over the Interleaved Bloom Filter improved performance and at the same time reduced drastically database sizes and memory requirements. Second, an extensive optimization of parameters and methodologies based on real data for the post-processing procedure (e.g. read reassignment), refined the results, enabling ganon2 to achieve a strong overall improvement in evaluation metrics compared to the current methodologies.

We are aware of and acknowledge the difficulties and biases in evaluating self-developed methods in an article against others fairly [27]. In our evaluations, we ran all

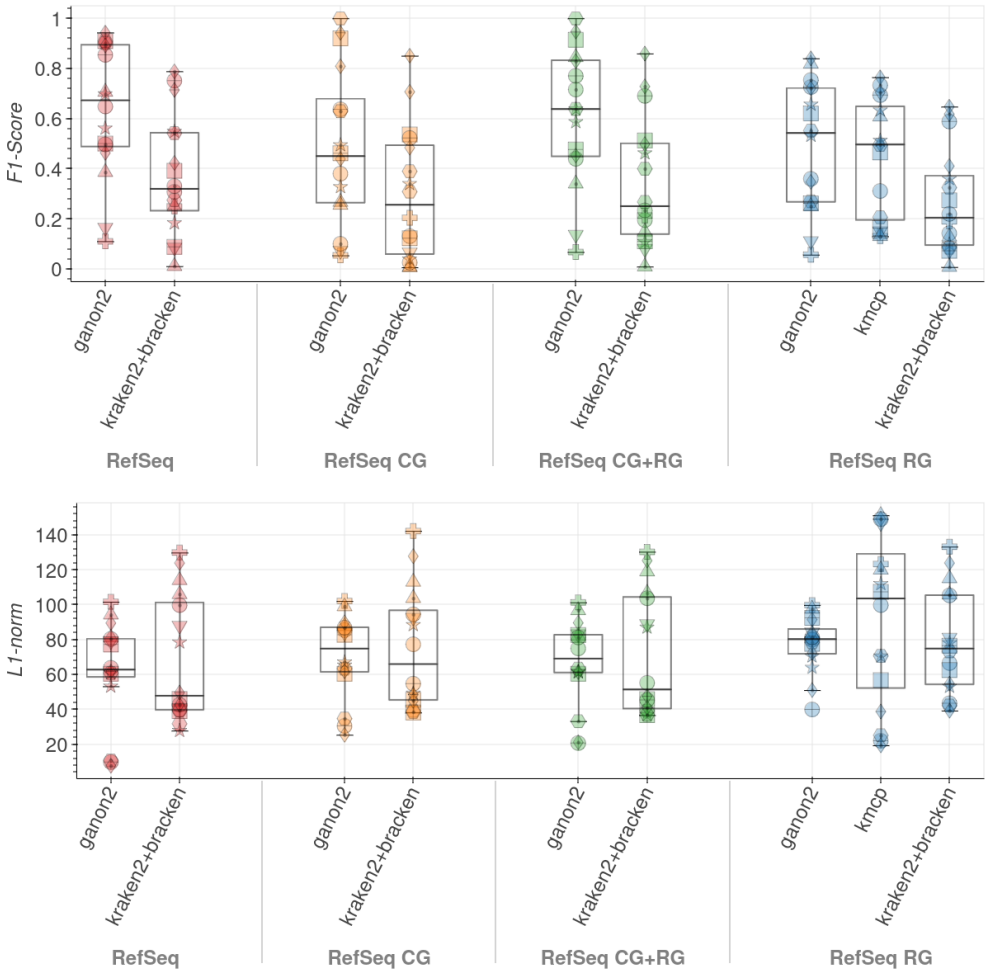

Fig. 3 **TOP**) F1-Score results for taxonomic profiling at the species level (higher is better). **BOTTOM**) L1-norm error results at the species level (lower is better). Each tool and database combination (x-axis) has 16 points, one for each sample analyzed (Table 2), identified by a distinct marker, with a boxplot showing their overall distribution. Results considering only species with abundance above 0.005% for all tools.

methods in "default" mode using the same underlying set of reference sequences. We believe that this is the fairest way to evaluate methods in a "user-level" approach. As noted by other studies [9, 23], changing parameters on a sample basis can drastically change the results of methods. However, real users in a real case application mostly do not have the time and resources to fine-tune parameters for every analysis and do not have the ground truth to evaluate the effects of the changes. Observing the usage of such methods in the literature, in most cases they are executed with the provided

default parameters with little to no variation. Therefore, this is the most sensible way to evaluate them in our opinion.

On a more technical level, many methods share characteristics in terms of data structure and parametrization (e.g. k-mer size), but the equivalence of them can be complex and not directly comparable. For example, kmcp uses a variation of bloom filters with a false-positive rate of 30% and 1 hash function by default, while ganon2 also uses variations of bloom filters but with a way lower false-positive rate and more hash functions. The way both tools use the outcome from the bloom filters is different, and matching those values would not make the comparison fairer. Readers should be aware of possible bias in software articles and interpret the results with discretion [27].

In an effort towards transparency and reproducibility, we provide all results and metrics from this article in an easy-to-navigate, open-source, and user-friendly dashboard (Supplementary Material), where it is possible to verify the results of each metric for all samples, databases, and methods independently and how they were performed. A completely fair and unbiased evaluation can only be performed by researchers unrelated to the development of the evaluated methods. Independent and external evaluation efforts like LEMMI [28] and CAMI [23] are great resources; however, a continuous, updated, and actively supported benchmark using the capabilities of MetaBench, covering a range of parameter variations and databases, is still to be developed.

ganon2 provides an efficient and scalable method to analyze large amounts of sequences against numerous reference sequences. It excels in metagenomics analysis, providing substantial improvements in taxonomic binning and profiling, especially using larger and more diverse reference sets. ganon2 is one of the fastest tools available in terms of building databases and analyzing data, with a reasonable memory footprint and the best results in terms of sensitivity and precision in our simulated evaluations. It provides not only an improvement over its first version but also is the overall best-performing tool in comparison with some of the state-of-the-art methods. ganon2 was developed with ease of use in mind, enabling the use of diverse, large, and up-to-date reference sequences in metagenomic analysis. ganon2 is open-source and comes with extensive documentation online with a detailed explanation of its functions, databases, parameters, and trade-offs and can be found at https://pirovc.github.io/ganon/.

Data Availability. The source code for ganon2 can be found in GitHub (https://github.com/pirovc/ganon) and Zenodo (v2.1.0: https://doi.org/10.5281/zenodo.10648995). The source code for MetaBench can be found in GitHub (https://github.com/pirovc/metabench) and Zenodo (v1.1.0: https://doi.org/10.5281/zenodo.14330683).

Funding. This work was financially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 458163427.

References

[1] Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N.: Shotgun metagenomics, from sampling to analysis. Nature Biotechnology **35**(9), 833–844 (2017) https://doi.org/10.1038/nbt.3935

[2] Marx, V.: Microbiology: the road to strain-level identification. Nature Methods **13**(5), 401–404 (2016) https://doi.org/10.1038/nmeth.3837 . Publisher: Nature Publishing Group

[3] GenBank and WGS Statistics. https://www.ncbi.nlm.nih.gov/genbank/statistics/ Accessed 2023-09-06

[4] DNA Sequencing Costs: Data. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data Accessed 2023-09-06

[5] Arita, M., Karsch-Mizrachi, I., Cochrane, G., on behalf of the International Nucleotide Sequence Database Collaboration: The international nucleotide sequence database collaboration. Nucleic Acids Research **49**(D1), 121–124 (2021) https://doi.org/10.1093/nar/gkaa967 . Accessed 2023-09-06

[6] Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M., Durkin, A.S., Gonzales, N.R., Gwadz, M., Lanczycki, C., Song, J.S., Thanki, N., Wang, J., Yamashita, R., Yang, M., Zheng, C., Marchler-Bauer, A., Thibaud-Nissen, F.: RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Research **49**(D1), 1020–1028 (2021) https://doi.org/10.1093/nar/gkaa1105 . Accessed 2023-09-06

[7] Locey, K.J., Lennon, J.T.: Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences **113**(21), 5970–5975 (2016) https://doi.org/10.1073/pnas.1521291113 . 577 citations (Crossref) [2023-01-02] ISBN: 2167-9843

[8] Méric, G., Wick, R.R., Watts, S.C., Holt, K.E., Inouye, M.: Correcting index databases improves metagenomic studies. bioRxiv, 712166 (2019) https://doi.org/10.1101/712166 . Accessed 2019-08-28

[9] Wright, R.J., Comeau, A.M., Langille, M.G.I.: From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microbial Genomics **9**(3), 000949 (2023) https://doi.org/10.1099/mgen.0.000949 . Publisher: Microbiology Society,. Accessed 2023-03-17

[10] Zhao, Z., Cristian, A., Rosen, G.: Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life. BMC Bioinformatics **21**(1), 412 (2020) https://doi.org/10.1186/s12859-020-03744-7 . Accessed 2022-12-27

[11] Breitwieser, F.P., Lu, J., Salzberg, S.L.: A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics (2019) https://doi.org/10.1093/bib/bbx120 . Accessed 2019-03-12

[12] Nasko, D.J., Koren, S., Phillippy, A.M., Treangen, T.J.: RefSeq database growth

influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biology **19**(1) (2018) https://doi.org/10.1186/s13059-018-1554-6 . 84 citations (Crossref) [2023-01-02]. Accessed 2019-03-12

[13] 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) https://doi.org/10.1093/bioinformatics/btaa458 . Accessed 2021-11-11

[14] Seiler, E., Mehringer, S., Darvish, M., Turc, E., Reinert, K.: Raptor: A fast and space-efficient pre-filter for querying very large collections of nucleotide sequences. iScience **24**(7) (2021) https://doi.org/10.1016/j.isci.2021.102782 . Publisher: Elsevier. Accessed 2021-08-18

[15] Mehringer, S., Seiler, E., Droop, F., Darvish, M., Rahn, R., Vingron, M., Reinert, K.: Hierarchical Interleaved Bloom Filter: enabling ultrafast, approximate sequence queries. Genome Biology **24**(1), 131 (2023) https://doi.org/10.1186/s13059-023-02971-4 . Accessed 2023-06-05

[16] Reinert, K., Dadi, T.H., Ehrhardt, M., Hauswedell, H., Mehringer, S., Rahn, R., Kim, J., Pockrandt, C., Winkler, J., Siragusa, E., Urgese, G., Weese, D.: The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. Journal of Biotechnology **261**(July), 157–168 (2017) https://doi.org/10.1016/j.jbiotec.2017.07.017 . 67 citations (Crossref) [2023-01-02]

[17] Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A.P., Vázquez-Baeza, Y., Parida, L., Kim, H.-C., Knight, R., Liu, Y.-Y.: Challenges in benchmarking metagenomic profilers. Nature Methods, 1–9 (2021) https://doi.org/10.1038/s41592-021-01141-3 . Publisher: Nature Publishing Group. Accessed 2021-05-14

[18] Shen, W., Xiang, H., Huang, T., Tang, H., Peng, M., Cai, D., Hu, P., Ren, H.: KMCP: accurate metagenomic profiling of both prokaryotic and viral populations by pseudo-mapping. Bioinformatics **39**(1), 845 (2023) https://doi.org/10.1093/bioinformatics/btac845 . Accessed 2023-03-02

[19] Wood, D.E., Lu, J., Langmead, B.: Improved metagenomic analysis with Kraken 2. Genome Biology **20**(1), 257 (2019) https://doi.org/10.1186/s13059-019-1891-0 . Accessed 2019-12-02

[20] Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L.: Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science **3**, 104 (2017) https://doi.org/10.7717/peerj-cs.104 . ISBN: biorxiv;051813v1

[21] Müller, A., Hundt, C., Hildebrandt, A., Hankeln, T., Schmidt, B.: MetaCache: Context-aware classification of metagenomic reads using minhashing. Bioinformatics (2017) https://doi.org/10.1093/bioinformatics/btx520 . 34 citations

(Crossref) [2023-01-02]

[22] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hansen, L.H., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G.G.Z., Cuevas, D.A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.-P., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A.E., Rattei, T., McHardy, A.C.: Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature Methods **14**(11), 1063–1071 (2017) https://doi.org/10.1038/nmeth.4458

[23] Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T.R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J.J., Brown, C.T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P.T.L.C., Cristian, A., Dabrowski, P.W., Darling, A.E., Egan, R., Eskin, E., Georganas, E., Goltsman, E., Gray, M.A., Hansen, L.H., Hofmeyr, S., Huang, P., Irber, L., Jia, H., Jørgensen, T.S., Kieser, S.D., Klemetsen, T., Kola, A., Kolmogorov, M., Korobeynikov, A., Kwan, J., LaPierre, N., Lemaitre, C., Li, C., Limasset, A., Malcher-Miranda, F., Mangul, S., Marcelino, V.R., Marchet, C., Marijon, P., Meleshko, D., Mende, D.R., Milanese, A., Nagarajan, N., Nissen, J., Nurk, S., Oliker, L., Paoli, L., Peterlongo, P., Piro, V.C., Porter, J.S., Rasmussen, S., Rees, E.R., Reinert, K., Renard, B., Robertsen, E.M., Rosen, G.L., Ruscheweyh, H.-J., Sarwal, V., Segata, N., Seiler, E., Shi, L., Sun, F., Sunagawa, S., Sørensen, S.J., Thomas, A., Tong, C., Trajkovski, M., Tremblay, J., Uritskiy, G., Vicedomini, R., Wang, Z., Wang, Z., Wang, Z., Warren, A., Willassen, N.P., Yelick, K., You, R., Zeller, G., Zhao, Z., Zhu, S., Zhu, J., Garrido-Oter, R., Gastmeier, P., Hacquard, S., Häußler, S., Khaledi, A., Maechler, F., Mesny, F., Radutoiu, S., Schulze-Lefert, P., Smit, N., Strowig, T., Bremges, A., Sczyrba, A., McHardy, A.C.: Critical Assessment of Metagenome Interpretation: the second round of challenges. Nature Methods (2022) https://doi.org/10.1038/s41592-022-01431-4 . Accessed 2022-04-11

[24] Mende, D.R., Waller, A.S., Sunagawa, S., Järvelin, A.I., Chan, M.M., Arumugam, M., Raes, J., Bork, P.: Assessment of metagenomic assembly using simulated next generation sequencing data. PloS one **7**(2), 31386 (2012) https://doi.org/10.1371/journal.pone.0031386 . Accessed 2013-08-06

[25] Parks, D.H., Rigato, F., Vera-Wolf, P., Krause, L., Hugenholtz, P., Tyson, G.W., Wood, D.L.A.: Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome. Frontiers

in Microbiology **12** (2021). Accessed 2022-10-11

[26] Belmann, P., Dröge, J., Bremges, A., McHardy, A.C., Sczyrba, A., Barton, M.D.: Bioboxes: standardised containers for interchangeable bioinformatics software. GigaScience **4**(1), 47 (2015) https://doi.org/10.1186/s13742-015-0087-0 . Publisher: GigaScience

[27] Buchka, S., Hapfelmeier, A., Gardner, P.P., Wilson, R., Boulesteix, A.-L.: On the optimistic performance evaluation of newly introduced bioinformatic methods. Genome Biology **22**(1), 152 (2021) https://doi.org/10.1186/s13059-021-02365-4 . 11 citations (Crossref) [2023-01-02]. Accessed 2021-05-27

[28] Seppey, M., Manni, M., Zdobnov, E.M.: LEMMI: A continuous benchmarking platform for metagenomics classifiers. Genome Research, 260398–119 (2020) https://doi.org/10.1101/gr.260398.119 . Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Accessed 2020-08-18