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 repositories of sequences greatly contributes to the success of metagenomics applications. However, they are growing at a much faster pace than the resources to use them properly. 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 metagenomics analysis. It indexes large datasets with a small memory footprint, maintaining fast, sensitive, and precise classification results. This is possible with the Hierarchical Interleaved Bloom Filter data structure paired with minimizers and several other improvements and optimizations. Based on the full NCBI RefSeq and its sub-sets, ganon2 indices are on average 50% smaller than state-of-the-art methods, providing a great compression rate for large and diverse genomic reference sets. Using 16 simulated samples from various studies, including the CAMI 1+2 challenge, ganon2 achieved up to 0.17 higher median F1-Score in taxonomic binning. In profiling, improvements in the F1-Score median are up to 0.32 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, ocean, 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, bio forensics, 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 them. Specialized tools and algorithms are therefore essential to scale with massive datasets, usually requiring high-performance computational resources. The goal is to enable metagenomics analyzes to scale to larger sets of data, using a limited amount of memory and in manageable time (minutes-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 13 years (January 2010 to January 2023) 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 76 most represented species in the Bacterial and Archaeal Complete Genomes subsection have as many base pairs as the remaining 5,201 species (as of April 2020). In the full Bacterial and Archaeal RefSeq the ratio is 15 to 28,254. 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, the RefSeq Microbial Complete Genomes subset has approximately only one third of the diversity in terms of number of species compared to the full RefSeq. Another example: a 2-year-old outdated full RefSeq release has 14,292 fewer species than the actual release (from December 2021-2023).

Unfortunately, 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 metagenomics classification with larger reference sequences databases [8, 9]. Other studies confirm the importance of reference choice to research outcome [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 under-perform 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 re-assignment 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 accessed. 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 (v2.0.0) with the state-of-the-art methods: kmcp (v0.9.2) [18], kraken2 (v2.1.2) [19], bracken (v2.8) [20], and metacache (v2.3.1) [21]. Methods were selected based on the following criteria: open-source code, actively maintained and/or highly used, 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 3 different configurations: (1) full, (2) complete genomes (CG), and (3) representative genomes (RG) (Table 1). RefSeq is commonly used 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 sub-set, represented only by genomes marked as complete, providing a theoretically higher quality. The full RefSeq has the highest diversity (52.781 species), and includes both CG and RG, being almost 10 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 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 was chosen here since it 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**	**Gbp**
RefSeq	29.808.096	295.219	52.781	1.212,4
RefSeq CG	86.649	44.121	19.715	136,4
RefSeq RG	1.029.894	18.073	18.073	93,8

Table 1 Reference sets used in this article. Data obtained in 2023-03-14 for archaea, bacteria, fungi and viral groups.

We selected 16 short-read simulated samples with available ground truth 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, well-represented organisms (in relation to our 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)	14.784,8
medium complexity metagenome (225 genomes)	74.568.473 (2 x 100bp)	14.913,7
high complexity metagenome (450 genomes)	74.557.111 (2 x 100bp)	14.911,4
CAMI 1 challenge [22]		
low complexity metagenome	49.898.179 (2 x 150bp)	14.969,5
medium complexity metagenome	49.918.839 (2 x 150bp)	14.975,7
high complexity metagenome	49.905.935 (2 x 150bp)	14.971,8
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
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
in silico mock community, high diversity, **ani 97%**, multi-strain	6.999.988 (2 x 150bp)	2.100
in silico mock community, high diversity, **ani 100%**, multi-strain	7.000.009 (2 x 150bp)	2.100

Table 2 16 simulated short-read samples used in this article, with the 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 metagenomics 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 positive 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 every method, removing low-abundant taxa. This improves the performance of 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 7702P 64-Core Processor (128 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. The tool was also used to optimize the default parameter configuration for ganon2.

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 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 variation. Therefore, this is the most sensible way to evaluate them. 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 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 truly fair 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.

3 Results

3.1 Build

We first evaluated the performance of tools to build custom databases, based on 3 RefSeq subsets (Table 1). ganon2 was the fastest method to build the RefSeq RG (19 minutes), while kmcp was the fastest for the RefSeq CG (34 minutes) and full RefSeq (4 hours 34 minutes). kraken2+bracken took the longest to build all databases, between 8 hours and more than 3 days. Metacache has an intermediate time performance, between 1 and a half hours and 10 and a half hours. In terms of memory necessary to build, kmcp has the lowest requirements on average for all databases evaluated, with ganon2 and kraken2 (without bracken) coming in second with similar requirements. 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. 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 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 performed well in terms of build time, with some bigger differences in memory usage, disk usage, and database size.

Time

	ganon2	**kmcp**	**kraken2**	**bracken**	**metacache**
RefSeq	10:19:33	04:34:35	3 days, 5:48:04	05:12:47	10:27:35
RefSeq CG	01:00:01	00:34:24	10:07:55	00:16:51	01:33:49
RefSeq RG	00:18:50	00:33:17	08:02:03	00:19:25	01:33:29

Memory

(GB)	**ganon2**	**kmcp**	**kraken2**	**bracken**	**metacache**
RefSeq	229	50	226	404	415
RefSeq CG	61	26	57	70	125
RefSeq RG	136	95	131	144	126

Database size

(GB)	**ganon2**	**kmcp**	**kraken2**	**bracken**	**metacache**
RefSeq	159	399	223	0.270	352
RefSeq CG	34	45	57	0.003	100
RefSeq RG	69	32	130	0.015	92

Table 3 Build Time and Memory consumption. Time is demonstrated in wall time (lower is better). Memory usage and database size is shown in GB (lower is better). bracken database is not standalone and it needs a matching kraken2 database. 64 threads were used.

We also compared the ability and easiness of tools to generate custom databases. ganon2 provides a highly automated build system and can download and build and update 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 well explained but adds a level of complexity that may confuse less technically inclined users.

3.2 Binning

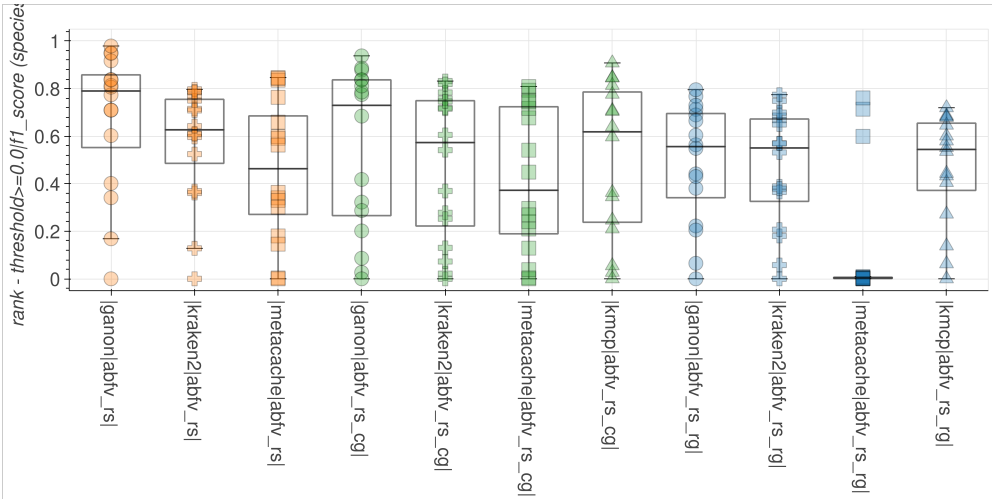

Fig. 1 F1-Score results for taxonomic binning at species level. Each tool and database combination (x-axis) has 16 points, one for each sample analyzed (Table 2), with a box-plot showing their overall distribution. Databases: abfv_rs = full RefSeq (yellow), abfv_rs_cg = RefSeq CG (green), abfv_rs_rg = RefSeq RG (blue).

Binning results in terms of F1-Score at species level are presented in Figure 1. ganon2 shows a significant improvement over other methods, especially when using more diverse databases (full RefSeq and RefSeq CG). For the full RefSeq, ganon2 achieved a median F1-Score of 0.79, with 75% of the samples with a F1-Score of 0.6 or higher. kraken2 comes in second with a median of 0.62 followed by metacache with 0.46. 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, providing very good binning results with a significantly smaller database. Here, kmcp was evaluated and performed well, with the second best results after ganon2. With 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 is the tool with the smallest memory footprint classifying reads against the full RefSeq (Table 4), requiring around 160GB of RAM. Using RefSeq CG, ganon2 achieved also very good results in terms of F1-Score and requires 4.7 times less memory (34GB) compared to the full RefSeq. metacache and kraken2 were the fastest tools, however with a higher memory usage. kmcp was significantly slower among the tested methods, failing to provide results in doable time for the full RefSeq. However, kmcp has the smallest memory footprint using the RefSeq RG database.

Binning

Time (Mbp/m)	**ganon2**	**kmcp**	**kraken2**	**metacache**
RefSeq	1021 (±395)	-	1860 (±546)	1773 (±695)
RefSeq CG	1956 (±589)	124 (±129)	2770 (±452)	3556 (±832)
RefSeq RG	1902 (±437)	289 (±35)	2229 (±435)	3675 (±760)
Mem. (GB)	**ganon2**	**kmcp**	**kraken2**	**metacache**
RefSeq	160	-	227	377
RefSeq CG	34	49	60	106
RefSeq RG	69	32	135	94

Profiling

Time (Mbp/m)	**ganon2**	**kmcp**	**kraken2+bracken**
RefSeq	1377 (±597)	-	1643 (±546)
RefSeq CG	2395 (±658)	124 (±129)	2750 (±447)
RefSeq RG	2426 (±559)	289 (±35)	2196 (±437)
Mem. (GB)	**ganon2**	**kmcp**	**kraken2+bracken**
RefSeq	159	-	227
RefSeq CG	34	49	60
RefSeq RG	69	32	135

Table 4 Binning and Profiling Time and Memory consumption, average from 16 short-read samples. 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 were executed 3 consecutive times and only the best time was considered. Time represent full run, including pre- and post-processing steps (load database in memory, classification, read re-assignment, etc.). kmcp did not finish running after several days with the full RefSeq database, therefore is not included. 64 threads were used.

3.3 Profiling

In profiling, ganon2 shows improved F1-Score at species level with the full RefSeq, with 50% of the samples with an F1-Score above 0.7 (Figure 2 TOP). kmcp shows comparable good results to ganon2 with the RefSeq CG and RG databases, however, it takes on average at least 10 times longer to run (Table 4). To exemplify the difference in runtime using the RefSeq CG in profiling: kraken2+bracken took 45 minutes to run all 16 samples (Table 2) in this benchmark, ganon2 took 52 minutes while kmcp needed 16 hours and 46 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) to profile diverse samples, yielding a substantial improvement in the results. With this database ganon2 runs fast, with the smallest memory footprint and achieved with 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 2 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). 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 over approximately 100

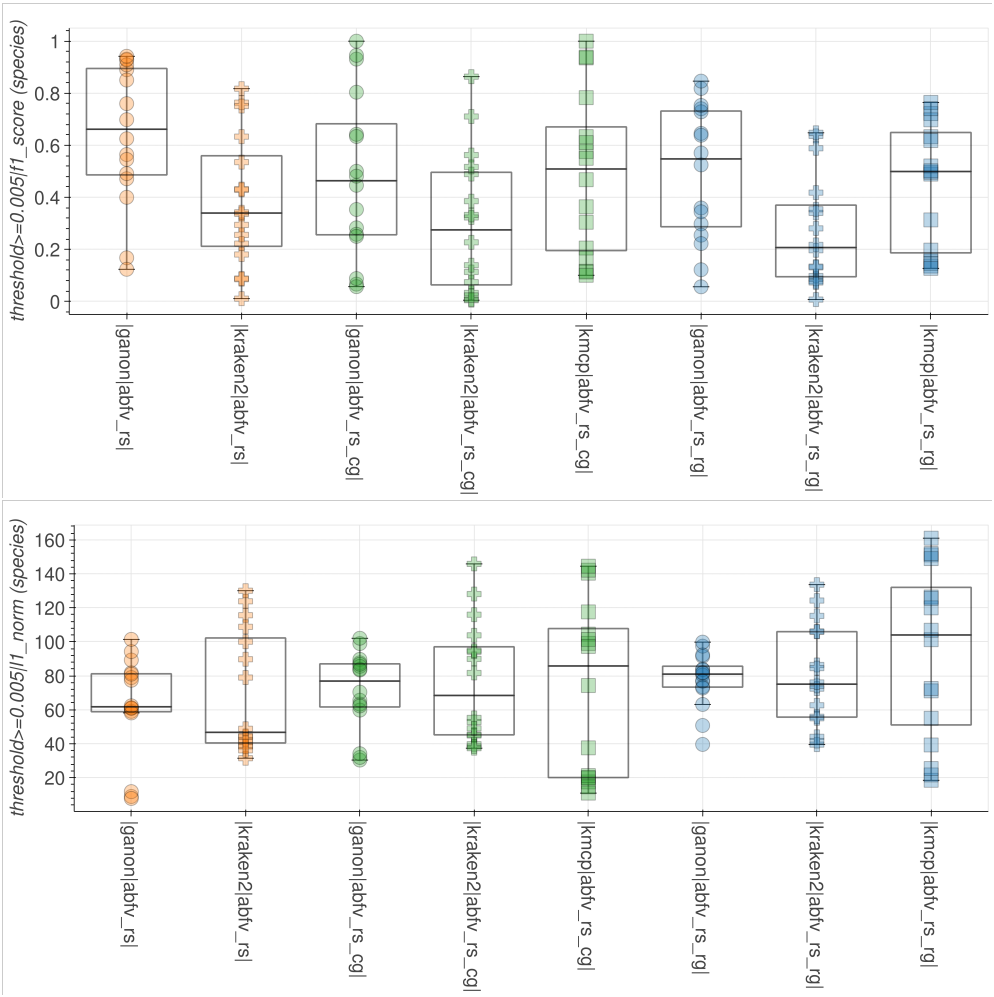

Fig. 2 **TOP**) F1-Score results for taxonomic profiling at species level. **BOTTOM**) L1-norm error results at species level (lower is better). Each tool and database combination (x-axis) has 16 points, one for each sample analyzed (Table 2), with a box-plot showing their overall distribution. Databases: abfv_rs = full RefSeq (yellow), abfv_rs_cg = RefSeq CG (green), abfv_rs_rg = RefSeq RG (blue). kraken2 here stands for kraken2+bracken. Results considering only species with abundance above 0.005% for all tools.

for all samples, keeping a low error in abundance estimation. kmcp has very low L1-norm values for some samples, with the downside of having also 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 provides an efficient and scalable method to analyze large amounts of sequences against numerous references 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 the easy-of-use in mind, enabling the use of diverse, large and up-to-date reference sequences into metagenomics analysis. ganon2 is open-source and comes with an extensive documentation online with a detailed explanation of its functions, databases, parameters, and trade-offs and can be found at https://github.com/pirovc/ganon

Acknowledgments. This work was financially supported by the 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