Critical Assessment of Metagenome Interpretation

the second round of challenges

Meyer, Fernando; Fritz, Adrian; Deng, Zhi-Luo; Koslicki, David; Lesker, Till Robin; Gurevich, Alexey; Robertson, Gary; Alser, Mohammed; Antipov, Dmitry; Beghini, Francesco

Total number of authors:
103

Published in:
Nature Methods

Link to article, DOI:
10.1038/s41592-022-01431-4

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

nature methods

ANALYSIS

https://doi.org/10.1038/s41592-022-01431-4

Check for updates

OPEN

Critical Assessment of Metagenome Interpretation: the second round of challenges

Fernando Meyer[1,2,76], Adrian Fritz[1,2,3,76], Zhi-Luo Deng[1,2,4], David Koslicki[5], Till Robin Lesker[3,6], Alexey Gurevich[7], Gary Robertson[1,2], Mohammed Alser[8], Dmitry Antipov[9], Francesco Beghini[10], Denis Bertrand[11], Jaqueline J. Brito[12], C. Titus Brown[13], Jan Buchmann[14], Aydin Buluç[15,16], Bo Chen[15,16], Rayan Chikhi[17], Philip T. L. C. Clausen[18], Alexandru Cristian[19,20], Piotr Wojciech Dabrowski[21,22], Aaron E. Darling[23], Rob Egan[24,25], Eleazar Eskin[26], Evangelos Georganas[27], Eugene Goltsman[24,25], Melissa A. Gray[19,28], Lars Hestbjerg Hansen[29], Steven Hofmeyr[15,16], Pingqin Huang[30], Luiz Irber[13], Huijue Jia[31,32], Tue Sparholt Jørgensen[33,34], Silas D. Kieser[35,36], Terje Klemetsen[37], Axel Kola[38], Mikhail Kolmogorov[39], Anton Korobeynikov[9,40], Jason Kwan[41], Nathan LaPierre[26], Claire Lemaitre[42], Chenhao Li[11], Antoine Limasset[43], Fabio Malcher-Miranda[44], Serghei Mangul[12], Vanessa R. Marcelino[45,46], Camille Marchet[43], Pierre Marijon[47], Dmitry Meleshko[9], Daniel R. Mende[48], Alessio Milanese[49,50], Niranjan Nagarajan[51,52], Jakob Nissen[53], Sergey Nurk[54], Leonid Oliker[15,16], Lucas Paoli[49], Pierre Peterlongo[42], Vitor C. Piro[44], Jacob S. Porter[55], Simon Rasmussen[56], Evan R. Rees[41], Knut Reinert[57], Bernhard Renard[44,58], Espen Mikal Robertsen[37], Gail L. Rosen[19,28,59], Hans-Joachim Ruscheweyh[49], Varuni Sarwal[26], Nicola Segata[10], Enrico Seiler[57], Lizhen Shi[60], Fengzhu Sun[61], Shinichi Sunagawa[49], Søren Johannes Sørensen[62], Ashleigh Thomas[24,63], Chengxuan Tong[11], Mirko Trajkovski[35,64], Julien Tremblay[65], Gherman Uritskiy[66], Riccardo Vicedomini[17], Zhengyang Wang[30], Ziye Wang[67], Zhong Wang[68,69,70], Andrew Warren[55], Nils Peder Willassen[37], Katherine Yelick[15,16], Ronghui You[30], Georg Zeller[50], Zhengqiao Zhao[19], Shanfeng Zhu[71,72], Jie Zhu[31,32], Ruben Garrido-Oter[73], Petra Gastmeier[38], Stephane Hacquard[73], Susanne Häußler[6], Ariane Khaledi[6], Friederike Maechler[38], Fantin Mesny[73], Simona Radutoiu[74], Paul Schulze-Lefert[73], Nathiana Smit[6], Till Strowig[6], Andreas Bremges[1,3], Alexander Sczyrba[75] and Alice Carolyn McHardy[1,2,3,4] ✉

Evaluating metagenomic software is key for optimizing metagenome interpretation and focus of the Initiative for the Critical Assessment of Metagenome Interpretation (CAMI). The CAMI II challenge engaged the community to assess methods on realistic and complex datasets with long- and short-read sequences, created computationally from around 1,700 new and known genomes, as well as 600 new plasmids and viruses. Here we analyze 5,002 results by 76 program versions. Substantial improvements were seen in assembly, some due to long-read data. Related strains still were challenging for assembly and genome recovery through binning, as was assembly quality for the latter. Profilers markedly matured, with taxon profilers and binners excelling at higher bacterial ranks, but underperforming for viruses and Archaea. Clinical pathogen detection results revealed a need to improve reproducibility. Runtime and memory usage analyses identified efficient programs, including top performers with other metrics. The results identify challenges and guide researchers in selecting methods for analyses.

Over the last two decades, advances in metagenomics have vastly increased our knowledge of the microbial world and intensified development of data analysis techniques[1–3]. This created a need for unbiased and comprehensive assessment of these methods, to identify best practices and open challenges in the field[4–7]. CAMI, the Initiative for the Critical Assessment of Metagenome Interpretation, is a community-driven effort addressing this need, by offering comprehensive benchmarking challenges on datasets representing common experimental settings, data generation techniques and environments in microbiome research. In addition to its open and collaborative nature, data FAIRness and reproducibility are defining principles[8].

A full list of affiliations appears at the end of the paper.

The first CAMI challenge[4] delivered insights into the performances of metagenome assembly, genome and taxonomic binning and profiling programs across multiple complex benchmark datasets, including unpublished genomes with different evolutionary divergences and poorly categorized taxonomic groups, such as viruses. The robustness and high accuracy observed for binning programs in the absence of strain diversity supported their application to large-scale data from various environments, recovering thousands of metagenome-assembled genomes[9,10] and intensified efforts in advancing strain-resolved assembly and binning. We here describe the results of the second round of CAMI challenges[11], in which we assessed program performances and progress on even larger and more complex datasets, including long-read data and further performance metrics such as runtime and memory use.

Results

We created metagenome benchmark datasets representing a marine, a high strain diversity environment ('strain-madness') and a plant-associated environment including fungal genomes and host plant material. Datasets included long and short reads sampled from 1,680 microbial genomes and 599 circular elements (Methods and Supplementary Table 1). Of these, 772 genomes and all circular elements were newly sequenced and distinct from public genome sequence collections (new genomes), and the remainder were high-quality public genomes. Genomes with an average nucleotide identity (ANI) of less than 95% to any other genome were classified as 'unique', and as 'common' otherwise, as in the first challenge[4]. Overall, 901 genomes were unique (474 marine, 414 plant-associated, 13 strain-madness), and 779 were common (303 marine, 81 plant-associated, 395 strain-madness). On these data, challenges were offered for assembly, genome binning, taxonomic binning and profiling methods, which opened in 2019 and 2020 and allowed submissions for several months (Methods). In addition, a pathogen detection challenge was offered, on a clinical metagenome sample from a critically ill patient with an unknown infection. Challenge participants were encouraged to submit reproducible results by providing executable software with parameter settings and reference databases used. Overall, 5,002 results for 76 programs were received from 30 teams (Supplementary Table 2).

Assembly challenge. Sequence assemblies are key for metagenome analysis and used to recover genome and taxon bins. Assembly quality degrades for genomes with low evolutionary divergences, resulting in consensus or fragmented assemblies[12,13]. Due to their relevance for understanding microbial communities[14,15], we assessed methods' abilities to assemble strain-resolved genomes, using long- and short-read data (Methods).

Overall trends. We evaluated 155 submissions for 20 assembler versions, including some with multiple settings and data preprocessing options (Supplementary Table 2). In addition, we created gold standard co- and single-sample assemblies as in refs. [4,16]. The gold standards of short, long and hybrid marine data comprise 2.59, 2.60 and 2.79 gigabases (Gb) of assembled sequences, respectively, while the strain-madness gold standards consist of 1.45 Gb each.

Assemblies were evaluated with MetaQUAST v.5.1.0rc (ref. [17]), adapted for assessing strain-resolved assembly (Supplementary Text). We determined strain recall and precision, similar to ref. [18] (Methods and Supplementary Table 3). To facilitate comparisons, we ranked assemblies produced with different versions and parameter settings for a method based on key metrics (Methods) and chose the highest-ranking as the representative (Fig. 1, Supplementary Fig. 1 and Supplementary Tables 3–7).

Short-read assemblers achieved genome fractions of up to 10.4% on strain-madness and 41.1% on marine data, both by MEGAHIT[19]. The gold standard reported 90.8 and 76.9%, respectively (Fig. 1a and Supplementary Table 3). HipMer[20] ranked best across metrics and datasets, and on marine data, as it produced few mismatches with a comparably high genome fraction and NGA50 (Table 1). On strain-madness data, GATB[21,22] ranked best, with HipMer in second place. On the plant-associated dataset, HipMer again ranked best, followed by Flye v.2.8 (ref. [23]), which outperformed other short-read assemblers in most metrics (Supplementary Fig. 2).

The best hybrid assembler, A-STAR, excelled in genome fraction (44.1% on marine, 30.9% on strain-madness), but created more misassemblies and mismatches (773 mismatches per 100 kb on marine) than others. HipMer had the fewest mismatches (67) per 100 kb on the marine and GATB on the strain-madness data (98, Fig. 1b). GATB introduced the fewest mismatches (173) among hybrid assemblers on the marine dataset. ABySS[24] created the fewest misassemblies for the marine and GATB for the strain-madness data (Fig. 1c). The hybrid assembler OPERA-MS[25] created the most contiguous assemblies for the marine data (Fig. 1d), with an average NGA50 of 28,244 across genomes, compared to 682,777 for the gold standard. The SPAdes[26] hybrid submission had a higher NGA50 of 43,014, but was not the best ranking SPAdes submission. A-STAR had the highest contiguity for the strain-madness data (13,008 versus 155,979 for gold standard). For short-read assembly, MEGAHIT had the highest contiguity on the marine (NGA50 26,599) and strain-madness data (NGA50 4,793). Notably, Flye performed well on plant-associated long-read data but worse than others across most metrics on the marine data (Supplementary Fig. 2), likely due to different versions or parameter settings (Supplementary Table 2).

For several assemblers, preprocessing using read quality trimming or error correction software, such as trimmomatic[27] or DUK[28], improved assembly quality (Supplementary Tables 2 and 3). Genome coverage was also a key factor (Fig. 1g). While gold standards for short and hybrid assemblies included genome assemblies with more than 90% genome fraction and 3.3× coverage, SPAdes best assembled low coverage marine genomes, starting at 9.2×. MEGAHIT, A-STAR, HipMer and Ray Meta[29] required 10×, 13.2×, 13.9× and 19.5× coverage, respectively. Several assemblers reconstructed high-copy circular elements well, with HipMer, MEGAHIT, SPAdes and A-STAR reconstructing all (Fig. 1g). Compared to software assessed in the first CAMI challenge, A-STAR had a 20% higher genome fraction on strain-madness data, almost threefold that of MEGAHIT. HipMer introduced the fewest mismatches (67 mismatches per 100 kb) on the marine data. This was 30% less than Ray Meta, the best performing method also participating in CAMI 1. OPERA-MS improved on MEGAHIT in NGA50 by 1,645 (6%), although using twice as much (long- and short-read) data. SPAdes, which was not assessed in the first challenge, was among the top submissions for most metrics.

Closely related genomes. The first CAMI challenge revealed substantial differences in assembly quality between unique and common strain genomes[4]. Across metrics, datasets and software

Fig. 1 | Metagenome assembler performances on the marine and strain-madness datasets. a, Radar plots of genome fraction. **b**, Mismatches per 100 kilobases (kb). **c**, Misassemblies. **d**, NGA50. **e**, Strain recall. **f**, Strain precision. For methods with multiple evaluated versions, the best ranked version on the marine data is shown (Supplementary Fig. 1 and Supplementary Table 3). Absolute values for metrics are log scaled. Lines indicate different subsets of genomes analyzed, and the value of the GSAs indicates the upper bound for a metric. The metrics are shown for both unique and common strain genomes. **g**, Genome recovery fraction versus genome sequencing depth (coverage) on the marine dataset. Blue indicates unique genomes (<95% ANI), green common genomes (ANI ≥ 95%) and orange high-copy circular elements. Gray lines indicate the coverage at which the first genome is recovered with ≥90% genome fraction.

results, unique genome assemblies again were superior, for marine genomes by 9.7% in strain recall, 19.3% genome fraction, sevenfold NGA50 and 6.5% strain precision, resulting in more complete and less fragmented assemblies (Fig. 1 and Supplementary Tables 4–7). This was even more pronounced on the strain-madness dataset, with a 79.1% difference in strain recall, 75.9% genome fraction, 20.6% strain precision and 50-fold NGA50. Although there were more misassemblies for unique than for common genomes (+1.5 in

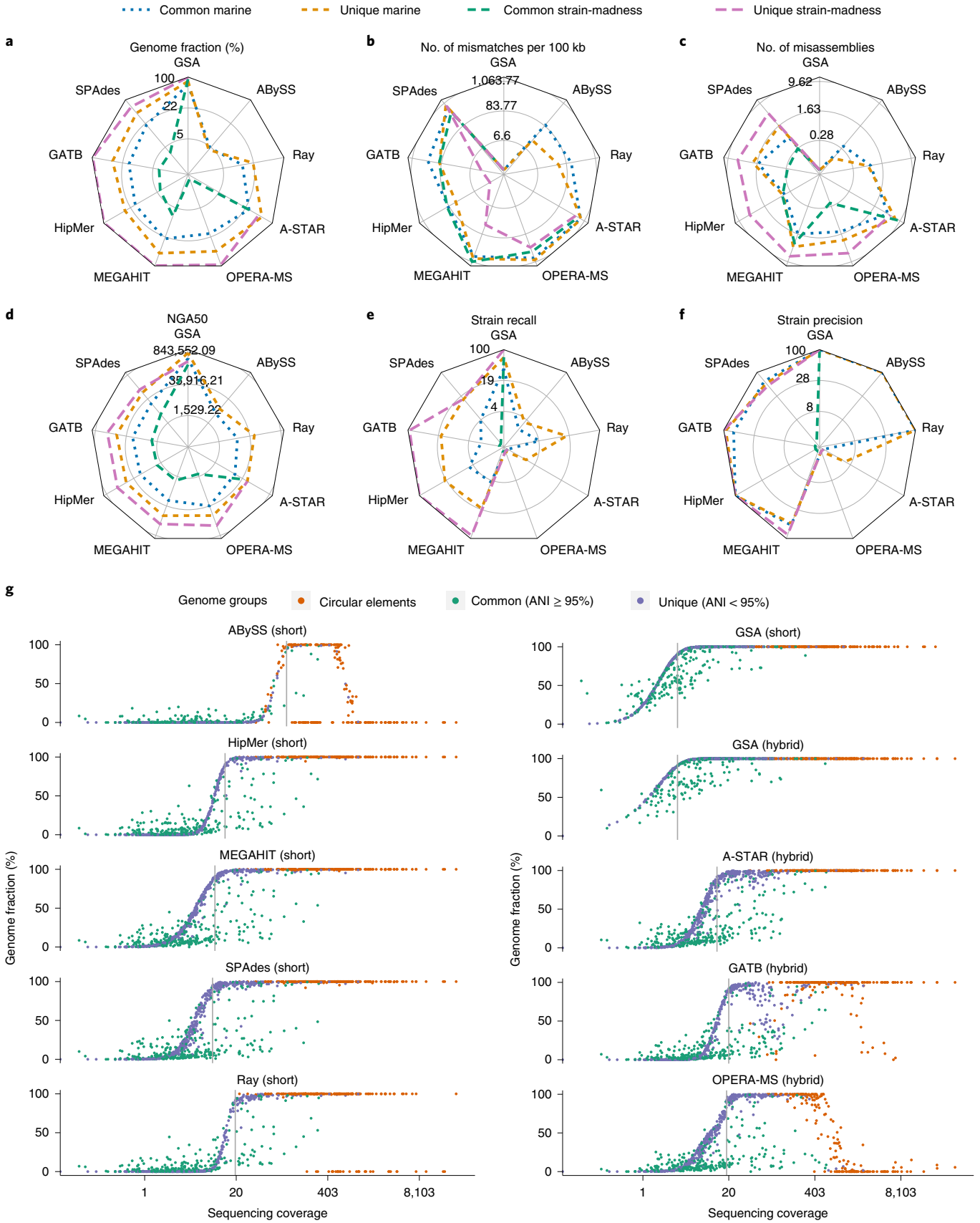

Table 1 | Best ranked software for four categories across datasets, in presence or absence of strain diversity and by computational requirements

	Assembly	Genome binning	Taxon binning	Taxon profiling
Metrics	Strain recall and precision, mismatches per 100 kb, duplication ratio, misassemblies, genome fraction, NGA50	Average completeness and purity, ARI, percentage of binned base pairs	Average completeness and purity, F1-score, accuracy	Completeness, purity, F1-score, L1- norm error, Bray–Curtis, Shannon equitability, weighted UniFrac error
Best methods across metrics				
Marine	HipMer 1.0, metaSPAdes 3.13.1, ABySS 2.1.5 (all on SR)	MetaBinner 1.0, UltraBinner 1.0, MetaWRAP 1.2.3	Kraken 2.0.8 beta (GSA), Ganon 0.1.4 (SR), MEGAN 6.15.2 (GSA)	mOTUs 2.5.1, MetaPhlAn 2.9.22 and v.cami1
Strain-madness	GATB 1.0 (hybrid), HipMer 1.0 (SR), OPERA-MS 0.8.3 (hybrid)	CONCOCT 0.4.1, MetaBinner 1.0, UltraBinner 1.0	PhyloPythiaS+ 1.4 (GSA), Kraken 2.0.8 beta (GSA), LSHVec (GSA)	mOTUs v.cami1, MetaPhlAn 2.9.22, DUDes v.cami1
Plant-associated	MetaHipMer 2.0.1.2 (SR), metaFlye 2.8.1 (hybrid), metaSPAdes 3.13.1 (SR)	CONCOCT 0.4.1 and 1.1.0, MaxBin 2.2.7	MEGAN 6.15.2 (GSA), Ganon 0.3.1 (SR), DIAMOND 0.9.28 (GSA)	mOTUs 2.5.1, MetaPhlAn 2.9.21, Bracken 2.6
Strain diversity	GATB 1.0 (hybrid), HipMer 1.0	CONCOCT 0.4.1	PhyloPythiaS+ 1.4 (on strain-madness data)	NA
No strain diversity	HipMer 1.0	UltraBinner 1.0	NA	NA
Fastest	MetaHipMer 2.0.1.2, MEGAHIT 1.2.7	MetaBAT 2.13, Vamb fa045c0	Kraken 2.0.8 beta (GSA, SR), DIAMOND (GSA)	FOCUS 1.5, Bracken 2.2
Most memory efficient	MEGAHIT 1.2.7, GATB 1.0	MetaBAT 2.13, MaxBin 2.0.2	Kraken 2.0.8 beta (GSA, SR), DIAMOND (GSA)	FOCUS 1.5, mOTUs 1.1.1

GSA denotes run on contigs of the GSAs and SR run on short reads. Submission deadlines for the different method categories and datasets are provided in the Methods. The numbers given are the software version numbers.

marine, +5.4 in strain-madness), this was due to the larger assembly size of the former, evident by a similar fraction of misassembled contigs (2.6% for unique genomes, 3.1% for common). While the duplication ratio was similar for unique and common genomes (+0.01 marine, −0.08 strain-madness), unique marine genome assemblies had 12% more mismatches than common ones (548 versus 486 mismatches per 100 kb). In contrast, there were 62% fewer mismatches for unique than common strain-madness genome assemblies (199 mismatches per 100 kb versus 511 mismatches per 100 kb), likely due to the elevated strain diversity.

For common marine genomes, HipMer ranked best across metrics and GATB for common strain-madness genomes. On unique genomes, HipMer ranked first for the marine and strain-madness datasets. HipMer had the highest strain recall and precision for common and unique marine genomes (4.5 and 20.4% recall, 100% precision each). For the strain-madness dataset, A-STAR had the highest strain recall (1.5%) on common strain-madness genomes, but lower precision (23.1%). GATB, HipMer, MEGAHIT and OPERA-MS assembled unique genomes with 100% recall and precision. A-STAR excelled in genome fraction, ranking first across all four data partitions and HipMer had the fewest mismatches. HipMer also had the fewest misassemblies on the common and unique marine genomes, while GATB had the fewest misassemblies on common strain-madness genomes and SPAdes on unique ones. The highest NGA50 on common marine genomes was achieved by OPERA-MS, on common strain-madness genomes by A-STAR and on unique genomes in both datasets by SPAdes.

Difficult to assemble regions. We assessed assembly performances for difficult to assemble regions, such as repeats or conserved elements (for example, 16S ribosomal RNA genes) on high-quality public genomes included in the marine data. These regions are important for genome recovery, but often missed[30]. We selected 50 unique, public genomes with annotated 16S sequences and present as a single contig in the gold standard assembly (GSA). We mapped assembly submissions to these 16S sequences using Minimap2 (ref. [31]) and measured their completeness (% genome fraction) and divergence[31] (Supplementary Fig. 3a,b,e). A-STAR partially recovered 102 (78%) of 131 16S sequences. The hybrid assemblers GATB (mean completeness 60.1%) and OPERA-MS (mean 47.1%) recovered the most complete 16S sequences. Mean completeness for short-read assemblies ranged from 29.6% (HipMer) to 36.9% (MEGAHIT). Assemblies were very accurate for ABySS and HipMer (<1% divergence). The hybrid assemblers GATB and OPERA-MS produced the longest contigs aligning to 16S rRNA genes, with a median length of 8,513 and 4,430 base pairs (bp), respectively, while for other assemblers median contig length was less than the average 16S rRNA gene length (1,503 bp). For all assemblers and 16S sequences, there were 17 cross-genome chimeras, reported by MetaQUAST as interspecies translocations: ten for MEGAHIT, five for A-STAR and one each for HipMer and SPAdes, while GATB, ABySS and OPERA-MS did not produce chimeric sequences. We performed the same evaluation for CRISPR cassettes found in 30 of the 50 genomes using different methods[32–34]. CRISPR cassette regions were easier to assemble, as evident by a higher (5–50%) completeness and longer assembled CRISPR-carrying contigs (up to 22× median length) than for 16S rRNA genes (Supplementary Fig. 3c,d,f). Across assemblies and methods, average assembly quality was better for public than for new genomes in key metrics, such as genome fraction and NGA50 (Supplementary Fig. 4).

Single versus coassembly. For multi-sample metagenome datasets, common assembly strategies are pooling samples (coassembly) and single-sample assembly[10,20,35]. We evaluated the assembly quality for both strategies using genomes spiked into the plant-associated data with specific coverages (Supplementary Table 8) across results for five assemblers (Supplementary Fig. 5). Only HipMer recovered a unique genome split across 16 samples from pooled samples, while a unique, single-sample genome was reconstructed well by all assemblers with both strategies. For genomes unique to a single sample,

but common in pooled samples (LjRoot109, LjRoot170), HipMer performed better on single samples, while OPERA-MS was better on pooled samples (Supplementary Fig. 5), and other assemblers traded a higher genome fraction against more mismatches. Thus, coassembly could generally improve assembly for OPERA-MS and for short-read assemblers on low coverage genomes without expected strain diversity across samples. For HipMer, single-sample assembly might be preferable if coverage is sufficient and closely related strains are expected.

Genome binning challenge. Genome binners group contigs or reads to recover genomes from metagenomes. We evaluated 95 results for 18 binner versions on short-read assemblies: 22 for the strain-madness GSAs, 17 for the strain-madness MEGAHIT assembly (MA), 19 for marine MA, 15 for marine GSA, 12 for plant-associated GSA and ten for the plant-associated MA (Supplementary Tables 9–15). In addition, seven results on the plant-associated hybrid assemblies were evaluated. Methods included well performing ones from the first CAMI challenge and popular software (Supplementary Table 2). While for GSA contigs the ground truth genome assignment is known, for the MA, we considered this to be the best matching genomes for a contig identified using MetaQUAST v.5.0.2. We assessed the average bin purity and genome completeness (and their summary using the F1-score), the number of high-quality genomes recovered, as well as the adjusted Rand index (ARI), using AMBER v.2.0.3 (ref. [36]) (Methods). The ARI, together with the fraction of binned data, quantifies binning performance for the overall dataset.

The performance of genome binners varied across metrics, software versions, datasets and assembly type (Fig. 2), while parameters affected performance mostly by less than 3%. For the marine GSA, average bin purity was 81.3 ± 2.3% and genome completeness was 36.9 ± 4.0% (Fig. 2a,b and Supplementary Table 9). For the marine MA, average bin purity (78.3 ± 2.6%) was similar, while average completeness was only 21.2 ± 1.6% (Fig. 2a,c and Supplementary Table 10), due to many short contigs with 1.5–2 kb, which most binners did not bin (Supplementary Fig. 6). For the strain-madness GSA, average purity and completeness decreased, by 20.1 to 61.2 ± 2.3% and by 18.7 to 18.2 ± 2.2%, respectively, relative to the marine GSA (Fig. 2a,d and Supplementary Table 11). While the average purity on the strain-madness MA (65.3 ± 4.0%) and GSA were similar, the average completeness dropped further to 5.2 ± 0.6%, again due to a larger fraction of unbinned short contigs (Fig. 2a,e and Supplementary Table 12). For the plant-associated GSA, purity was almost as high as for marine (78.2 ± 4.5%; Fig. 2a,f and Supplementary Table 13), but bin completeness decreased relative to other GSAs (13.9 ± 1.4%), due to poor recovery of low abundant, large, fungal genomes. Notably, the *Arabidopsis thaliana* host genome (5.6x coverage) as well as fungi with more than eight times coverage were binned with much higher completeness and purity (Supplementary Fig. 7). Binning of the hybrid assembly further increased average purity to 85.1 ± 6.3%, while completeness remained similar (11.9 ± 2.1%, Supplementary Table 14). For the plant-associated MA, average purity (83 ± 3.3%) and completeness (12.4 ± 1.5%, Fig. 2a,g and Supplementary Table 15) were similar to the GSA.

To quantitatively assess binners across gold standard and real assemblies for the datasets, we ranked submissions (Supplementary Tables 16–19 and Supplementary Fig. 8) across metrics (Methods). For marine and strain-madness, CONCOCT[37] and MetaBinner had the best trade-off performances for MAs, UltraBinner for GSAs and MetaBinner overall. CONCOCT also performed best on plant-associated assemblies (Table 1). UltraBinner had the best completeness on the marine GSA, CONCOCT on the strain-madness GSA and plant-associated MA, MetaWRAP on marine and strain-madness MAs and MaxBin[38] on the plant-associated GSA. Vamb always had the best purity, while UltraBinner had the best ARI for the marine GSA, MetaWRAP for the strain-madness GSA and MetaBAT[39,40] for MAs and plant-associated assemblies. MetaWRAP and MetaBinner assigned the most for the marine and plant-associated assemblies, respectively. Many methods assigned all strain-madness contigs, although with low ARI (Fig. 2b–g). UltraBinner recovered the most high-quality genomes from the marine GSA, MetaWRAP from the marine MA, CONCOCT from strain-madness assemblies and plant-associated GSA, and MetaBinner from the plant-associated GSA and hybrid assemblies (Fig. 2 and Supplementary Table 20). For plasmids and other high-copy circular elements, Vamb performed best, with an F1-score of 70.8%, 54.8% completeness and 100% purity, while the next best method, MetaWRAP, had an F1-score of 12.7% (Supplementary Table 21).

Effect of strain diversity. For marine and strain-madness GSAs, unique strain binning was substantially better than for common strains (Supplementary Fig. 9 and Supplementary Tables 9 and 11). Differences were more pronounced on strain-madness, for which unique strain bin purity was particularly high (97.9 ± 0.4%). UltraBinner ranked best across metrics and four data partitions for unique genomes and overall, and CONCOCT for common strains (Supplementary Table 22). UltraBinner had the highest completeness on unique strains, while CONCOCT ranked best for common strains and across all partitions. Vamb always ranked first by purity, UltraBinner by ARI and MetaBinner by most assigned. Due to the dominance of unique strains in the marine and common strains in the strain-madness dataset, the best binners in the respective data and entire datasets were the same (Supplementary Tables 9 and 11) and performances similar for most metrics.

Taxonomic binning challenge. Taxonomic binners group sequences into bins labeled with a taxonomic identifier. We evaluated 547 results for nine methods and versions: 75 for the marine, 405 for strain-madness and 67 for plant-associated data, on either reads or GSAs (Supplementary Tables 2). We assessed the average purity and completeness of bins and the accuracy per sample at different taxonomic ranks, using the National Center for Biotechnology Information (NCBI) taxonomy version provided to participants (Methods).

On the marine data, average taxon bin completeness across ranks was 63%, average purity 40.3% and accuracy per sample bp 74.9% (Fig. 3a and Supplementary Table 23). On the strain-madness data, accuracy was similar (76.9%, Fig. 3b and Supplementary Table 24), while completeness was around 10% higher and purity lower by that much. On the plant-associated data, purity was between those of the first two datasets (35%), but completeness and accuracy were lower (44.2 and 50.8%, respectively; Fig. 3c and Supplementary Table 25). For all datasets, performances declined at lower taxonomic ranks, most notably from genus to species rank by 22.2% in completeness, 9.7% in purity and 18.5% in accuracy, on average.

Across datasets, MEGAN on contigs ranked first across metrics and ranks (Supplementary Table 26), closely followed by Kraken v.2.0.8 beta on contigs and Ganon on short reads. Kraken on contigs was best for genus and species, and on marine data across metrics and in completeness and accuracy (89.4 and 96.9%, Supplementary Tables 23 and 27 and Supplementary Fig. 10). Due to the presence of public genomes, Kraken's completeness on marine data was much higher than in the first CAMI challenge, particularly at species and genus rank (average of 84.6 and 91.5%, respectively, versus 50 and 5%), while purity remained similar. MEGAN on contigs ranked highest for taxon bin purity on the marine and plant-associated data (90.7 and 87.1%, Supplementary Tables 23, 25, 27 and 28). PhyloPythiaS+ ranked best for the strain-madness data across metrics, as well as in completeness (90.5%) and purity (75.8%) across ranks

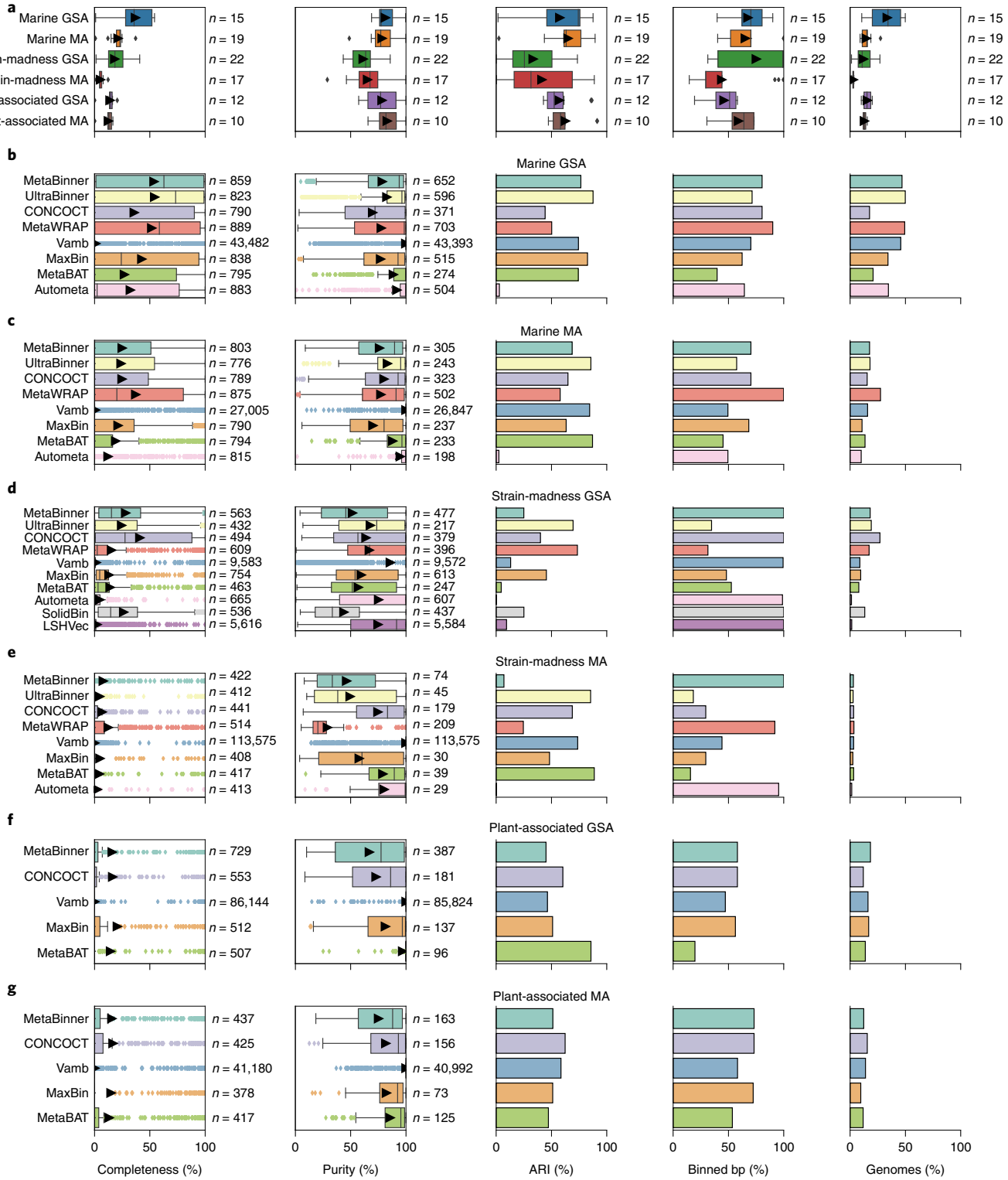

Fig. 2 | Performance of genome binners on short-read assemblies (GSA and MA, MEGAHIT) of the marine, strain-madness, and plant-associated data. a, Boxplots of average completeness, purity, ARI, percentage of binned bp and fraction of genomes recovered with moderate or higher quality (>50% completeness, <10% contamination) across methods from each dataset (Methods). Arrows indicate the average. **b**–**g**, Boxplots of completeness per genome and purity per bin, and bar charts of ARI, binned bp and moderate or higher quality genomes recovered, by method, for each dataset: marine GSA (**b**), marine MA (**c**), strain-madness GSA (**d**), strain-madness MA (**e**), plant-associated GSA (**f**) and plant-associated MA (**g**). The submission with the highest F1-score per method on a dataset is shown (Supplementary Tables 9–15). Boxes in boxplots indicate the interquartile range of *n* results, the center line the median and arrows the average. Whiskers extend to 1.5 × interquartile range or to the maximum and minimum if there is no outlier. Outliers are results represented as points outside 1.5 × interquartile range above the upper quartile and below the lower quartile.

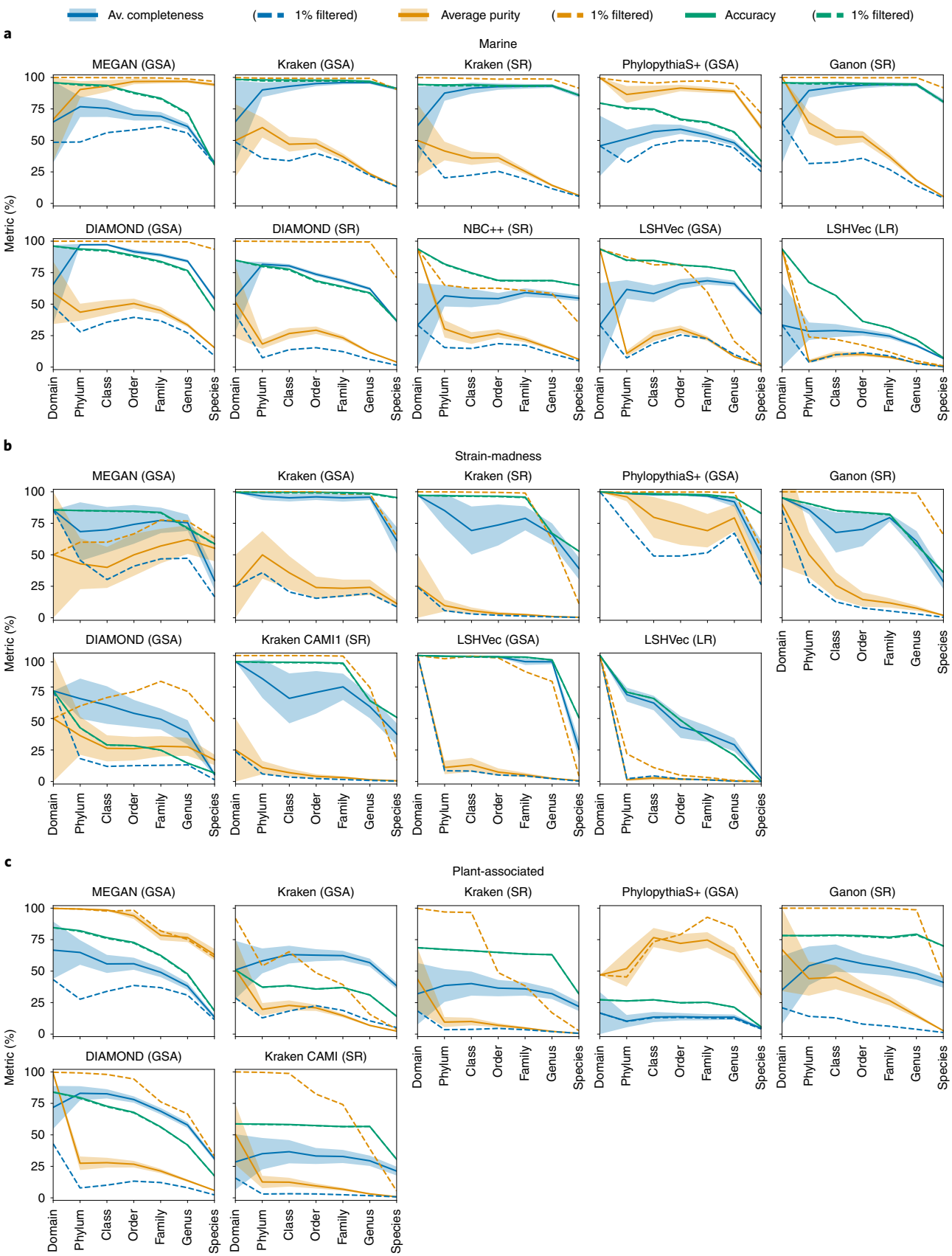

Fig. 3 | Taxonomic binning performance across ranks per dataset. a, Marine. **b**, Strain-madness. **c**, Plant-associated. Metrics were computed over unfiltered (solid lines) and 1%-filtered (that is, without the 1% smallest bins in bp, dashed lines) predicted bins of short reads (SR), long reads (LR) and contigs of the GSA. Shaded bands show the standard error across bins.

(Supplementary Tables 24 and 29). DIAMOND on contigs ranked best for completeness (67.6%) and Ganon on short reads for accuracy (77.1%) on the plant-associated data.

Filtering the 1% smallest predicted bins per taxonomic level is a popular postprocessing approach. Across datasets, filtering increased average purity to above 71% and reduced completeness, to roughly 24% on marine and strain-madness and 13.4% on plant-associated data (Supplementary Tables 23–25). Accuracy was not much affected, as large bins contribute more to this metric. Kraken on contigs still ranked first in filtered accuracy and MEGAN across all filtered metrics (Supplementary Table 26). MEGAN on contigs and Ganon on short reads profited the most from filtering, ranking first in filtered completeness and purity, respectively, across all datasets and taxonomic levels.

Taxonomic binning of divergent genomes. To investigate the effect of increasing divergence between query and reference sequences for taxonomic binners, we categorized genomes by their distances to public genomes (Supplementary Fig. 11 and Supplementary Tables 30 and 31). Sequences of known marine strains were assigned particularly well at species rank by Kraken (accuracy, completeness and filtered purity above 93%) and MEGAN (91% purity, 33% completeness and accuracy). Kraken also best classified new strain sequences at species level, although with less completeness and accuracy for the marine data (68 and 80%, respectively). It also had the best accuracy and completeness across ranks, but low unfiltered purity. For the strain-madness data, PhyloPythiaS+ performed similarly well up to genus level and best assigned new species at genus level (93% accuracy and completeness, and 75% filtered purity). Only DIAMOND correctly classified viral contigs, although with low purity (50%), completeness and accuracy (both 3%).

Taxonomic profiling challenge. Taxonomic profilers quantify the presence and relative taxon abundances of microbial communities from metagenome samples. This is different from taxonomic sequence classification, which assigns taxon labels to individual sequences and results in taxon-specific sequence bins (and sequence abundance profiles)[41]. We evaluated 4,195 profiling results (292 marine, 2,603 strain-madness and 1,300 plant-associated datasets), from 22 method versions (Supplementary Table 2) with most results for short-read samples, and a few for long-read samples, assemblies or averages across samples. Performance was evaluated with OPAL v.1.0.10 (ref. [42]) (Methods). The quality of predicted taxon profiles was determined based on completeness and purity of identified taxa, relative to the underlying ground truth, for individual ranks, while taxon abundance estimates were assessed using the L1 norm for individual ranks and the weighted UniFrac error across ranks. Accuracy of alpha diversity estimates was measured using the Shannon equitability index (Methods). Overall, mOTUs v.2.5.1 and MetaPhlAn v.2.9.22 ranked best across taxonomic ranks and metrics on the marine and plant-associated datasets, and mOTUs v.cami1 and MetaPhlAn v.2.9.22 on the strain-madness dataset (Table 1, Supplementary Tables 33, 35 and 37 and Supplementary Fig. 12).

Taxon identification. Methods performed well until genus rank (marine average purity 70.4%, strain-madness 52.1%, plant-associated 62.9%; marine average completeness 63.3%, strain-madness 80.5%, plant-associated 42.1%; Fig. 4a,c, Supplementary Fig. 13 and Supplementary Tables 32, 34 and 36), with a substantial drop at species level. mOTUs v.2.5.1 (ref. [43]) had completeness and purity above 80% at genus and species ranks on marine data, and Centrifuge v.1.0.4 beta (ref. [44]) and MetaPhlAn v.2.9.22 (refs. [45,46]) at genus rank (Fig. 4a). Bracken[47] and NBC++ (ref. [48]) had completeness above 80% at either rank, and CCMetagen[49], DUDes v.0.08 (ref. [50]), LSHVec v.gsa[51], Metalign[52], MetaPalette[53] and MetaPhlAn v.cami1 more than 80% purity. Filtering the rarest (1%) predicted taxa per rank decreased completeness by roughly 22%, while increasing precision by roughly 11%.

On strain-madness data at genus rank, MetaPhlAn v.2.9.22 (89.2% completeness, 92.8% purity), MetaPhyler v.1.25 (ref. [54]) (92.3% completeness, 79.2% purity) and mOTUs v.cami1 (92.9% completeness, 69.1% purity) performed best, but no method excelled at species rank. DUDes v.0.08 and LSHVec v.gsa had high purity, while Centrifuge v.1.0.4 beta, DUDes v.cami1, TIPP v.4.3.10 (ref. [55]) and TIPP v.cami1 high completeness.

On plant-associated data at genus rank, sourmash_gather v.3.3.2_k31_sr (ref. [56]) was best overall (53.3% completeness, 89.5% purity). Sourmash_gather v.3.3.2_k31 on PacBio reads and MetaPhlAn v.3.0.7 had the highest purity for genus (98.5%, 95.5%) and species ranks (64.4%, 68.8%) and sourmash_gather v.3.3.2_k21_sr the highest completeness (genus 61.9%, species 53.8%).

Relative abundances. Abundances across ranks and submissions were on average predicted better for strain-madness than marine data, which has less complexity above strain level, with the L1 norm improving from 0.44 to 0.3, and average weighted UniFrac error from 4.65 to 3.79 (Supplementary Tables 32, 34 and 36). These weighted UniFrac values are substantially higher than for biological replicates (0.22, Methods). Abundance predictions were not as good on the plant-associated data and averaged 0.57 in L1 norm and 5.16 in weighted UniFrac. On the marine data, mOTUs v.2.5.1 had the lowest L1 norm at almost all levels with 0.12 on average, 0.13 at genus and 0.34 at species level, respectively. It was followed by MetaPhlAn v.2.9.22 (average 0.22, 0.32 genus, 0.39 species). Both methods also had the lowest weighted UniFrac error, followed by DUDEs v.0.08. On the strain-madness data, mOTUs v.cami1 performed best in L1 norm across ranks (0.05 average), and also at genus and species with 0.1 and 0.15, followed by MetaPhlAn v.2.9.22 (0.09 average, 0.12 genus, 0.23 species). The last also had the lowest weighted UniFrac error, followed by TIPP v.cami1 and mOTUs v.2.0.1. On the plant-associated data, Bracken v.2.6 had the lowest L1 norm across ranks with 0.36 on average, and at genus with 0.34. Sourmash_gather v.3.3.2_k31' on short reads had the lowest (0.55) at species. Both methods also had the lowest UniFrac error on this dataset. Several methods accurately reconstructed the alpha diversity of samples using the Shannon equitability; best (0.03 or less absolute difference to gold standards) across ranks on marine data were: mOTUs v.2.5.1, DUDes v.0.08 and v.cami1 and MetaPhlAn v.2.9.22 and v.cami1; on strain-madness data: DUDes v.cami1, mOTUs v.cami1 and MetaPhlAn v.2.9.22. On the plant-associated data, mOTU v.cami1 and Bracken v.2.6 performed best with this metric (0.08 and 0.09).

Difficult and easy taxa. For all methods, viruses, plasmids and Archaea were difficult to detect (Supplementary Fig. 14 and Supplementary Table 38) in the marine data. While many Archaeal taxa were detected by several methods, others, such as Candidatus Nanohaloarchaeota, were not detected at all. Only Bracken and Metalign detected viruses. In contrast, bacterial taxa in the Terrabacteria group and the phyla of Bacteroidetes and Proteobacteria were always correctly detected. Based on taxon-wise precision and recall for submissions, methods using similar information tended to cluster (Supplementary Fig. 15).

Clinical pathogen prediction: a concept challenge. Clinical pathogen diagnostics from metagenomics data is a highly relevant translational problem requiring computational processing[57]. To raise awareness, we offered a concept challenge (Methods): a short-read metagenome dataset of a blood sample from a patient with hemorrhagic fever was provided for participants to identify pathogens and to indicate those likely to cause the symptoms described in a case report. Ten manually curated, hence not reproducible results were

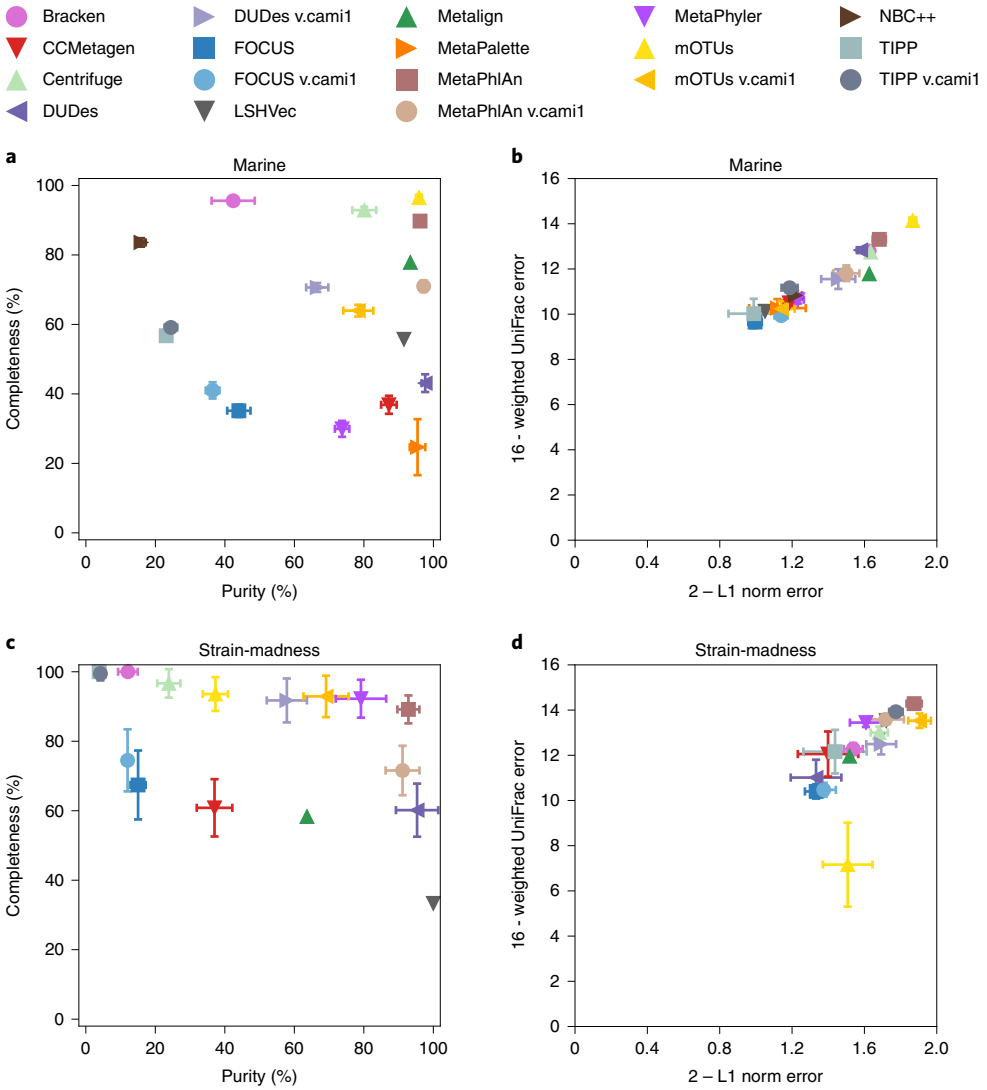

Fig. 4 | Taxonomic profiling results for the marine and strain-madness datasets at genus level. a,**b**, Marine datasets. **c**,**d**, Strain-madness datasets. Results are shown for the overall best ranked submission per software version (Supplementary Tables 33 and 35, and Supplementary Fig. 12). **a**,**c**, Purity versus completeness. **b**,**d**, Upper bound of L1 norm (2) minus actual L1 norm versus upper bound of weighted UniFrac error (16) minus actual weighted UniFrac error. Symbols indicate the mean over ten marine and 100 strain-madness samples, respectively, and error bars the standard deviation. Metrics were determined using OPAL with default settings.

received (Supplementary Table 39). The number of identified taxa per result varied considerably (Supplementary Fig. 16). Three submissions correctly identified the causal pathogen, Crimean–Congo hemorrhagic fever orthonairovirus (CCHFV), using the taxonomic profilers MetaPhlAn v.2.2, Bracken v.2.5 and CCMetagen v.1.1.3 (ref. [49]). Another submission using Bracken v.2.2 correctly identified orthonairovirus, but not as the causal pathogen.

Computational requirements. We measured the runtimes and memory usage for submitted methods across the marine and strain-madness data (Fig. 5, Supplementary Table 40 and Methods). Efficient methods capable of processing the entire datasets within minutes to a few hours were available in every method category, including some top ranked techniques with other metrics. Substantial differences were seen within categories and even between versions, ranging from methods executable on standard desktop machines to those requiring extensive hardware and heavy parallelization. MetaHipMer was the fastest assembler and required 2.1 h to process marine short reads, 3.3× less than the second fastest assembler, MEGAHIT. However, MetaHipMer used the most memory (1,961 gigabytes (GB)). MEGAHIT used the least memory (42 GB), followed by GATB (56.6 GB). On marine assemblies, genome binners on average required roughly three times less time than for the smaller strain-madness assemblies (29.2 versus 86.1 h), but used almost 4× more memory (69.9 versus 18.5 GB). MetaBAT v.2.13.33 was the fastest (1.07 and 0.05 h) and most memory efficient binner (maximum memory usage 2.66 and 1.5 GB) on both datasets. It was roughly 5× and 635× faster than the second fastest method, Vamb v.fa045c0, roughly 6× faster than LSHVec v.1dfe822 on marine and 765× faster than SolidBin v.1.3 on strain-madness data; roughly twice and 5× more memory efficient than the next ranking MaxBin v.2.0.2 and CONCOCT v.1.1.0 on marine data, respectively. Both MetaBAT and CONCOCT were substantially (roughly 11× and 4×) faster than their CAMI 1 versions. Like genome binners, taxonomic binners ran longer on the marine than the strain-madness assemblies, for example PhyloPythiaS+ with 287.3 versus 36 h, respectively, but had a similar or slightly

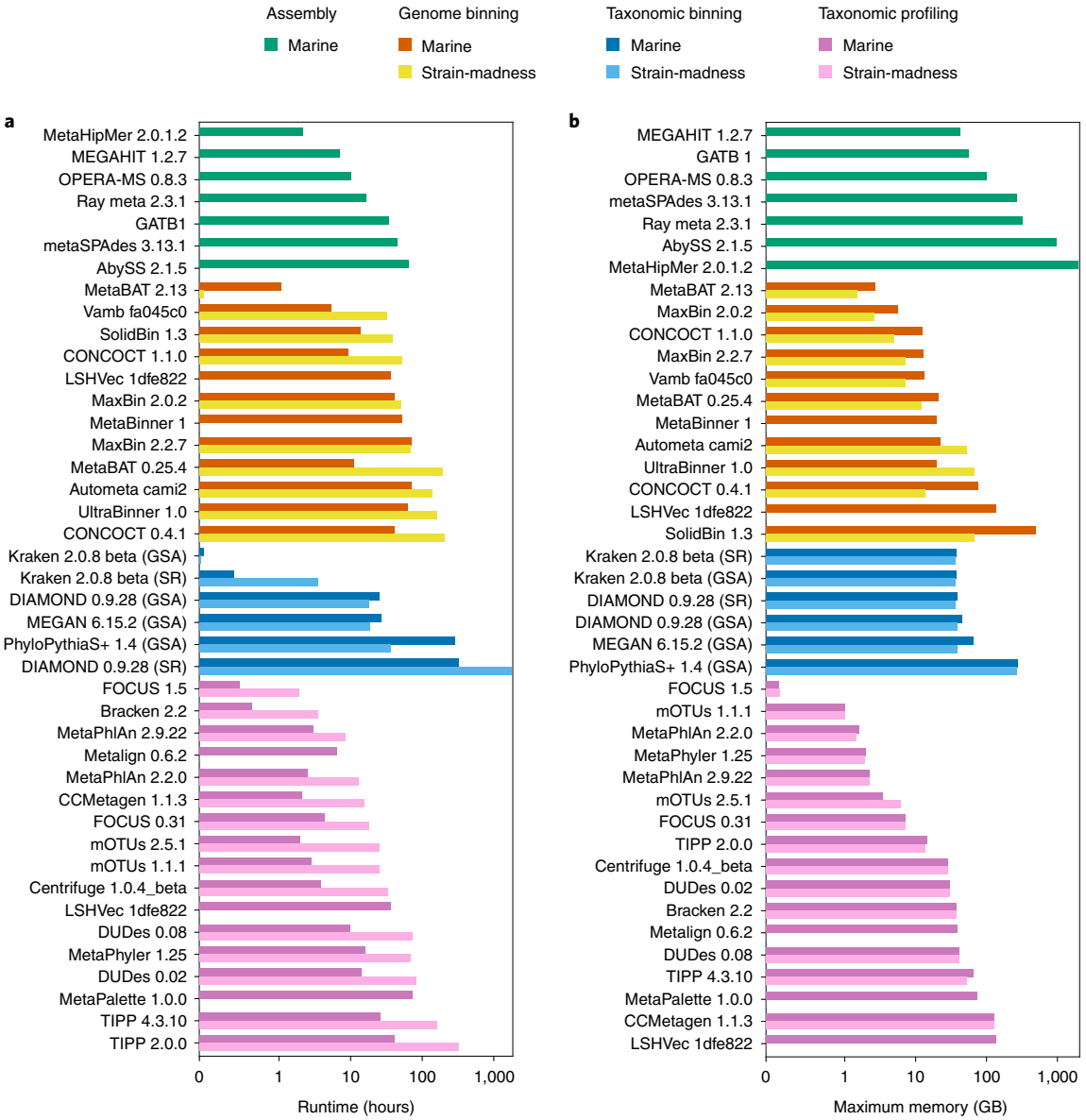

Fig. 5 | Computational requirements of software from all categories. a, Runtime. **b**, Maximum memory usage. Results are reported for the marine and strain-madness read data or GSAs (Supplementary Table 40). The *x* axes are log scaled and the numbers given are the software version numbers.

higher memory usage. On the marine read data, taxon profilers, however, were almost 4× faster on average (16.1 versus 60.8 h) than on the ten times larger strain-madness read dataset, but used more memory (38.1 versus 25 GB). The fastest and most memory efficient taxonomic binner was Kraken, requiring only 0.05 and 0.02 h, respectively, and roughly 37 GB memory on both datasets, for reads or contigs. It was followed by DIAMOND, which ran roughly 500× and 910× as long on the marine and strain-madness GSAs, respectively. FOCUS v.1.5 (ref. [58]) and Bracken v.2.2 were the fastest profilers on the marine (0.51, 0.66 h, respectively) and strain-madness (1.89, 3.45 h) data. FOCUS v.1.5 also required the least memory (0.16 GB for marine, 0.17 GB for strain-madness), followed by mOTUs v.1.1.1 and MetaPhlAn v.2.2.0.

Discussion

Assessing metagenomic analysis software thoroughly, comprehensively and with little bias is key for optimizing data processing strategies and tackling open challenges in the field. In its second round, CAMI offered a diverse set of benchmarking challenges across a comprehensive data collection reflecting recent technical developments. Overall, we analyzed 5,002 results of 76 program versions with different parameter settings across 131 long- and short-read metagenome samples from four datasets (marine, plant-associated, strain-madness, clinical pathogen challenge). This effort increased the number of results 22× and the number of benchmarked software versions 3× relative to the first challenge, delivering extensive new insights into software performances across a range of conditions. By systematically assessing runtime and memory requirements, we added two more key performance dimensions to the benchmark, which are important to consider given the ever-increasing dataset sizes.

In comparison to software assessed in the first challenges, assembler performances rose by up to 30%. Still, in the presence of closely related strains, assembly contiguity, genome fractions and strain recall decreased, suggesting that most assemblers, sometimes intentionally[19,26], did not resolve strain variation, resulting in more fragmented, less strain-specific assemblies. In addition, genome coverage, parameter settings and data preprocessing impacted assembly quality, while performances were similar across software

versions. Most submitted metagenome assemblies used only short reads, and long and hybrid assemblies had no higher overall quality. Hybrid assemblies, however, were better for difficult to assemble regions, such as the 16S rRNA gene, recovering more complete genes than most short-read submissions. Hybrid assemblers were also less affected by closely related strains in pooled samples, suggesting that long reads help to distinguish strains.

In comparison to the first CAMI challenges, ensemble binners presented a development showing substantial improvements across metrics compared to most individual methods. Overall, genome binners demonstrated variable performances across metrics and dataset types, with strain diversity and lower assembly quality presenting challenges that substantially reduced performances, even for the large sample number of the strain-madness dataset. For the plant host and 55 fungal genomes with sufficient coverage in the plant-associated data, high-quality bins were also obtained.

For taxonomic binners and profilers, highly performant and computationally efficient software was available, performing well across a range of conditions and metrics. Particularly profilers have matured since the first challenges, with less variance in top performers across taxon identification, abundance and diversity estimates. Performance was high for genus rank and above, with a substantial drop for bacterial species. As the second challenge data include high-quality public genomes, the data are less divergent from publicly available data than for the first challenges, on which method performances had already declined going from family to genus rank. It was also low for Archaea and viruses, suggesting a need for developers to extend their reference sequence collections and model development. Another encouraging result is that in the clinical pathogen challenge, several submissions identified the causal pathogen. However, due to manual curation, none was reproducible, indicating that these methods still require improvements, as well as assessment on large data collections. Although there is great potential of clinical metagenomics for pathogen diagnostics and characterization[57], multiple challenges still prevent its application in routine diagnostics[59].

In its second challenge, CAMI identified key advances for common metagenomics software categories as well as current challenges. As the state-of-the-art in methods and data generation progresses, it will be important to continuously re-evaluate these questions. In addition, computational methods for other microbiome data modalities[6] and multi-omics data integration could be jointly assessed. Most importantly, CAMI is a community-driven effort and we encourage everyone interested in benchmarking in microbiome research to join us.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01431-4.

Received: 22 July 2021; Accepted: 14 February 2022;
Published online: 8 April 2022

References

1. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic assembly: overview, challenges and applications. *Yale J. Biol. Med.* **89**, 353–362 (2016).
2. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, 1125–1136 (2019).
3. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016).
4. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation: a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
5. McIntyre, A. B. R. et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18**, 182 (2017).
6. Van Den Bossche, T. et al. Critical Assessment of Metaproteome Investigation (CAMPI): a multi-lab comparison of established workflows. *Nat. Commun.* **12**, 7305 (2021).
7. Commichaux, S. et al. A critical assessment of gene catalogs for metagenomic analysis. *Bioinformatics* **37**, 2848–2857 (2021).
8. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
9. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
10. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
11. Bremges, A. & McHardy, A. C. Critical assessment of metagenome interpretation enters the second round. *mSystems* **3**, e00103–e00118 (2018).
12. Turnbaugh, P. J. et al. The human microbiome project. *Nature* **449**, 804–810 (2007).
13. Meyer, F. et al. Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nat. Protoc.* **16**, 1785–1801 (2021).
14. Nawy, T. Microbiology: the strain in metagenomics. *Nat. Methods* **12**, 1005 (2015).
15. Segata, N. On the road to strain-resolved comparative metagenomics. *mSystems* **3**, e00190–17 (2018).
16. Fritz, A. et al. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
17. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).
18. Fritz, A. et al. Haploflow: strain-resolved de novo assembly of viral genomes. *Genome Biol.* **22**, 212 (2021).
19. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
20. Hofmeyr, S. et al. Terabase-scale metagenome coassembly with MetaHipMer. *Sci. Rep.* **10**, 10689 (2020).
21. Drezen, E. et al. GATB: genome assembly & analysis tool box. *Bioinformatics* **30**, 2959–2961 (2014).
22. Chikhi, R. & Rizk, G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms Mol. Biol.* **8**, 22 (2013).
23. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
24. Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
25. Bertrand, D. et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).
26. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
28. Li, M., Copeland, A. & Han, J. *DUK – A Fast and Efficient Kmer Based Sequence Matching Tool, Lawrence Berkeley National Laboratory. LBNL Report #: LBNL-4516E-Poster* (2011).
29. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).
30. Maguire, F. et al. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Micro. Genom.* **6**, mgen000436 (2020).
31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
32. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinf.* **8**, 209 (2007).
33. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).
34. Mreches, R. et al. GenomeNet/deepG: DeepG pre-release version. *Zenodo* https://doi.org/10.5281/zenodo.5561229 (2021).
35. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
36. Meyer, F. et al. AMBER: assessment of metagenome BinnERs. *Gigascience* **7**, giy069 (2018).
37. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
38. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
39. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
40. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

41. Sun, Z. et al. Challenges in benchmarking metagenomic profilers. *Nat. Methods* **18**, 618–626 (2021).
42. Meyer, F. et al. Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* **20**, 51 (2019).
43. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
44. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
45. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
46. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
47. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Sci.* **3**, e104 (2017).
48. Zhao, Z., Cristian, A. & Rosen, G. Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life. *BMC Bioinf.* **21**, 412 (2020).
49. Marcelino, V. R. et al. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol.* **21**, 103 (2020).
50. Piro, V. C., Lindner, M. S. & Renard, B. Y. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics* **32**, 2272–2280 (2016).
51. Shi, L. & Chen, B. LSHvec: a vector representation of DNA sequences using locality sensitive hashing and FastText word embeddings. In *Proc. 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (ed. Chairs, G., Jiang, H., Huang, X., Zhang, J. & Florida, G.) 1–10 (Association for Computing Machinery, 2021).
52. LaPierre, N., Alser, M., Eskin, E., Koslicki, D. & Mangul, S. Metalign: efficient alignment-based metagenomic profiling via containment min hash. *Genome Biol.* **21**, 242 (2020).
53. Koslicki, D. & Falush, D. MetaPalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. *mSystems* **1**, e00020–16 (2016).
54. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12**, S4 (2011).
55. Shah, N., Molloy, E. K., Pop, M. & Warnow, T. TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics* **37**, 1839–1845 (2021).
56. Pierce, N. T., Irber, L., Reiter, T., Brooks, P. & Brown, C. T. Large-scale sequence comparisons with sourmash. *F1000Res.* **8**, 1006 (2019).
57. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
58. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E. & Edwards, R. A. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* **2**, e425 (2014).
59. Dulanto Chiang, A. & Dekker, J. P. From the pipeline to the bedside: advances and challenges in clinical metagenomics. *J. Infect. Dis.* **221**, S331–S340 (2020).

[1]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Braunschweig, Germany. [2]Braunschweig Integrated Centre of Systems Biology (BRICS), Technische Universität Braunschweig, Braunschweig, Germany. [3]German Center for Infection Research (DZIF), Hannover-Braunschweig Site, Braunschweig, Germany. [4]Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany. [5]Pennsylvania State University, State College, PA, USA. [6]Helmholtz Centre for Infection Research, Braunschweig, Germany. [7]Saint Petersburg State University, Saint Petersburg, Russia. [8]Department of Information Technology and Electrical Engineering, ETH Zürich, Zurich, Switzerland. [9]Center for Algorithmic Biotechnology, Saint Petersburg State University, Saint Petersburg, Russia. [10]Department CIBIO, University of Trento, Trento, Italy. [11]Genome Institute of Singapore, Singapore, Singapore. [12]University of Southern California, Los Angeles, CA, USA. [13]University of California, Davis, Davis, CA, USA. [14]Institute for Biological Data Science, Heinrich-Heine-University, Düsseldorf, Germany. [15]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [16]University of California, Berkeley, Berkeley, CA, USA. [17]Institut Pasteur, Paris, France. [18]National Food Institute, Division of Global Surveillance, Technical University of Denmark, Lyngby, Denmark. [19]Drexel University, Philadelphia, PA, USA. [20]Google Inc., Philadelphia, PA, USA. [21]Robert Koch-Institut, Berlin, Germany. [22]Hochschule für Technik und Wirtschaft Berlin, Berlin, Germany. [23]University of Technology Sydney, Sydney, Australia. [24]DOE Joint Genome Institute, Berkeley, CA, USA. [25]Lawrence Berkeley National Laboratories, Berkeley, CA, USA. [26]University of California, Los Angeles, Los Angeles, CA, USA. [27]Intel Corporation, Santa Clara, CA, USA. [28]Ecological and Evolutionary Signal-Processing and Informatics Laboratory, Philadelphia, PA, USA. [29]University of Copenhagen, Department of Plant and Environmental Science, Frederiksberg, Denmark. [30]School of Computer Science, Fudan University, Shanghai, China. [31]BGI-Shenzhen, Shenzhen, China. [32]Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI-Shenzhen, Shenzhen, China. [33]Technical University of Denmark, Novo Nordisk Foundation Center for Biosustainability, Lyngby, Denmark. [34]Aarhus University, Department of Environmental Science, Roskilde, Denmark. [35]Department of Cell Physiology and Metabolism, Faculty of Medicine, University of Geneva, Geneva, Switzerland. [36]Swiss Institute of Bioinformatics, Geneva, Switzerland. [37]The Arctic University of Norway, Tromsø, Norway. [38]Charité—Universitätsmedizin Berlin, Berlin, Germany. [39]Department of Computer Science and Engineering, University of California San Diego, San Diego, CA, USA. [40]Department of Statistical Modelling, Saint Petersburg State University, Saint Petersburg, Russia. [41]University of Wisconsin—Madison, Madison, WI, USA. [42]Univ. Rennes, Inria, CNRS, IRISA, Rennes, France. [43]Université Lille, CNRS, CRIStAL, Lille, France. [44]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany. [45]Sydney Medical School, The University of Sydney, Sydney, Australia. [46]Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Australia. [47]Department of Computer Science, Inria, University of Lille, CNRS, Lille, France. [48]Amsterdam University Medical Center, Amsterdam, the Netherlands. [49]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich, Switzerland. [50]Structural and Computational Biology Unit, EMBL, Heidelberg, Germany. [51]Genome Institute of Singapore, A*STAR, Singapore, Singapore. [52]National University of Singapore, Singapore, Singapore. [53]DTU Health Tech, Kongens, Lyngby, Denmark. [54]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [55]University of Virginia, Charlottesville, VA, USA. [56]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [57]Institute for Bioinformatics, FU Berlin, Berlin, Germany. [58]Bioinformatics Unit (MF1), Robert Koch Institute, Berlin, Germany. [59]Center for Biological Discovery from Big Data, Philadelphia, PA, USA. [60]Florida Polytechnic University, Lakeland, FL, USA. [61]Quantitative and Computational Biology Department, University of Southern California, Los Angeles, CA, USA. [62]University of Copenhagen, Copenhagen, Denmark. [63]University of British Columbia, Vancouver, British Columbia, Canada. [64]Diabetes Center, Faculty of Medicine, University of Geneva, Geneva, Switzerland. [65]Energy, Mining and Environment, National Research Council Canada, Montreal, Quebec, Canada. [66]Phase Genomics, Seattle, WA, USA. [67]School of Mathematical Sciences, Fudan University, Shanghai, China. [68]Department of Energy Joint Genome Institute, Berkeley, CA, USA. [69]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [70]School of Natural Sciences, University of California at Merced, Merced, CA, USA. [71]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. [72]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China. [73]Max Planck Institute for Plant Breeding Research, Köln, Germany. [74]Aarhus University, Aarhus, Denmark. [75]Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany. [76]These authors contributed equally: F. Meyer, A. Fritz. ✉e-mail: amc14@helmholtz-hzi.de