nature portfolio

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

Supplementary information

Critical Assessment of Metagenome Interpretation: the second round of challenges

In the format provided by the authors and unedited

Supplementary information

Critical Assessment of Metagenome Interpretation - the second round of challenges

CAMI II benchmark data generation

Marine and plant-associated datasets

To generate the CAMI II marine dataset, we used the CAMISIM[1] simulator to simulate a hybrid long and short-read shotgun metagenome dataset, including ten samples with a taxonomic composition matching an arctic marine microbiome[2] profile. The taxonomic profile was generated using USEARCH[3] for OTU clustering and taxonomic classification with the RDP[4] classifier based on pyrosequencing 16S data from an arctic marine environment (ERP003605). As input genomic data, 622 high quality genomes from the MarRef database[5] as well as 224 representative terrestrial genomes from proGenomes[6] were used, as well as 176 and 216 novel genomes from the respective environment as well as 598 short circular elements currently not in the public domain. The new genomes were assembled with the SPAdes[7] assembler, version 3.12, using the --careful flag and default arguments otherwise. For each genome, sequences of length 1 kb or less were removed and a taxonomic annotation with CAMITAX[8] was performed. CAMITAX includes a CheckM[9] run, whose contamination and completion values were used to remove the new as well as multi-contig MarRef genomes, if completeness was less than 90% or contamination higher than 5%.

For the database genomes, where a taxonomic classification was available, CAMITAX was used as a consistency check, and the lowest common ancestor of the CAMITAX classification and the original taxonomic classification was chosen. The resulting taxonomically annotated genome set (including the newly sequenced genomes) and their taxonomic classification along with a 16S rRNA profile were then used to generate shotgun metagenome samples with CAMISIM in the community design mode. Within CAMISIM, a community genome abundance profile was generated by mapping the taxa from 16S rRNA profile to the input genomes and abundances were assigned accordingly. Genomes not mapped to any taxa were subsequently randomly assigned to the most abundant, non-assigned OTU, such that all input genomes were included in the dataset. In the next step, plasmids were added to the generated community genome abundance profile. We used plasmids specifically sequenced and classified for CAMI. Since the plasmids are expected to be circular while fasta files only allow a linear representation, the original

fasta files were treated in the following way. Given a plasmid sequence of *n* bases, split the sequence in 10 equally long segments, *n_1, …, n_10*, where each *n_i* consists of the *i-th tenth* bases of the input sequence. Then create 10 permutations of the input sequence, where always the first segment is moved to the end of the previous permutation. When the original sequence *n_1, …, n_10* is the first permutation, then *n_2, …, n_10, n_1* is the second, *n_3, …, n_10, n_1, n_2* the third and so forth.

The plasmids were identified as either virus or plasmid (or unknown) using the following process[10]:

- Prodigal[11] complete gene prediction (including genes overlapping sequence ends)
- Annotation using hmmscan[12] (max e-value 1e-4) and PFAMv27[13]
- Any sequence with plasmid replication, mobilization, or stability was classified as plasmid
- Any sequence without these plasmid markers, but with a viral replication or capsid gene was classified as virus/phage.

Any sequence with neither of the features was designated unknown.

To add the plasmids treated this way, the desired number, 200 for the marine challenge, were selected and randomly assigned to the 200 highest abundant genomes to emulate the affiliation of the plasmids with the genomes. Since plasmids are highly abundant in metagenomic datasets, they were chosen to have ~15x the abundance of the input genomes, in particular this meant that every permutation was assigned roughly 1.5x of the affiliated genomes' abundance. The exact number was calculated by the formula: ab_plasmid = ab_genome * 1.5 * N(1,0.1).

Given all the genomes and plasmids with their respective abundance, CAMISIM could finally be run to create the challenge dataset.

The scripts and command line options are provided on Github under https://github.com/CAMI-challenge/second_challenge_evaluation

Strain madness dataset

For the strain madness dataset, 408 new genomes were sequenced, assembled, and taxonomically classified using CAMITAX. 395 have a closely related genome present with 180 *Streptococcus pneumoniae*, 97 *Escherichica coli*, 47 *Klebsiella pneumoniae*, 21 *Staphylococcus aureus*, 21 *Enterococcus faecium* and 27 further *Enterobacteriacaeae*. Additionally, 13 unique genomes from a mouse gut, mainly consisting of *Lactobacillus* and Bacteroides, were added. This dataset was simulated without a BIOM profile as input, but instead using the *differential* mode of CAMISIM.

CAMI metagenome assembly evaluation for strain-specific assemblies

To assess strain-specific assemblies, we engaged the computational metagenomics community in the CAMI II evaluation workshop and defined assembly properties and evaluation strategies. Assessments in CAMI are done according to these agreed strategies, which have been incorporated in MetaQUAST version 5.1.0rc and are available via the new `--unique-mapping` option (http://cab.cc.spbu.ru/quast/manual.html#unique_mapping). The following command shows how to apply the new option on the strain madness coassembly, as an example.

Synopsis: *quast-5.1.0rc1/metaquast.py --reuse-combined-alignments --no-icarus -o ../strmgCAMI2_co_assembly_metaquast-5.1.0rc1 -r `cat ../refs` -t 28 --unique-mapping ../strmgCAMI2_pooled/GS_**

We consider two genomes to be different strains of the same species if they have >95% average nucleotide identity.
A *consensus* (or *strain-unresolved*) assembly is one where each contig may correspond to 1 or many strains (*strain-unresolved* contig), and reciprocally (and importantly), any genomic region that has one or more homologs across different strains is represented in only one contig. Assemblers, e.g., MEGAHIT and metaSPAdes, that create such assemblies are *strain-oblivious* assemblers.
A *strain-resolved, or strain-specific* assembly is one where each contig either i) maps equally likely to >1 strains (*core* contigs), or ii) maps unambiguously to only 1 strain (*strain-specific* contigs). Assemblers that create such assemblies are strain-aware or strain-resolved assemblers. In addition, core contigs should be present in as many copies as there are genomes containing such regions (see example below).

Evaluation of strain-specific assemblies

Consider the following reference genomes:

R1 and R2 are two strains of the same species (%-identity higher or identical to our threshold set above). R3 is a different species, without homologous regions with R1 and R2.

Case 1: two (extreme) examples of assemblies

Assembly **A1**

A1 is a consensus assembly. Here, contig C1 corresponds to a consensus of R1 and R2, and contig C2 corresponds exactly to R3.

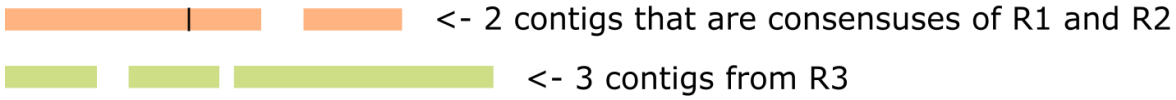
Assembly **A2**

The contigs of A2 correspond exactly to the reference genomes. A2 is a strain-specific assembly. Contigs C1 and C2 are >= 95% identical.

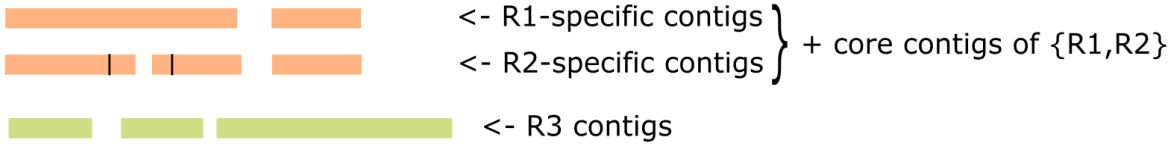
Between **A1** and **A2**, in the context of strain-aware evaluation, we favor assembly **A2** over **A1**. In particular, recovered genome fraction will be higher for **A2** than for **A1**.

Case 2: two more realistic assemblies

A3

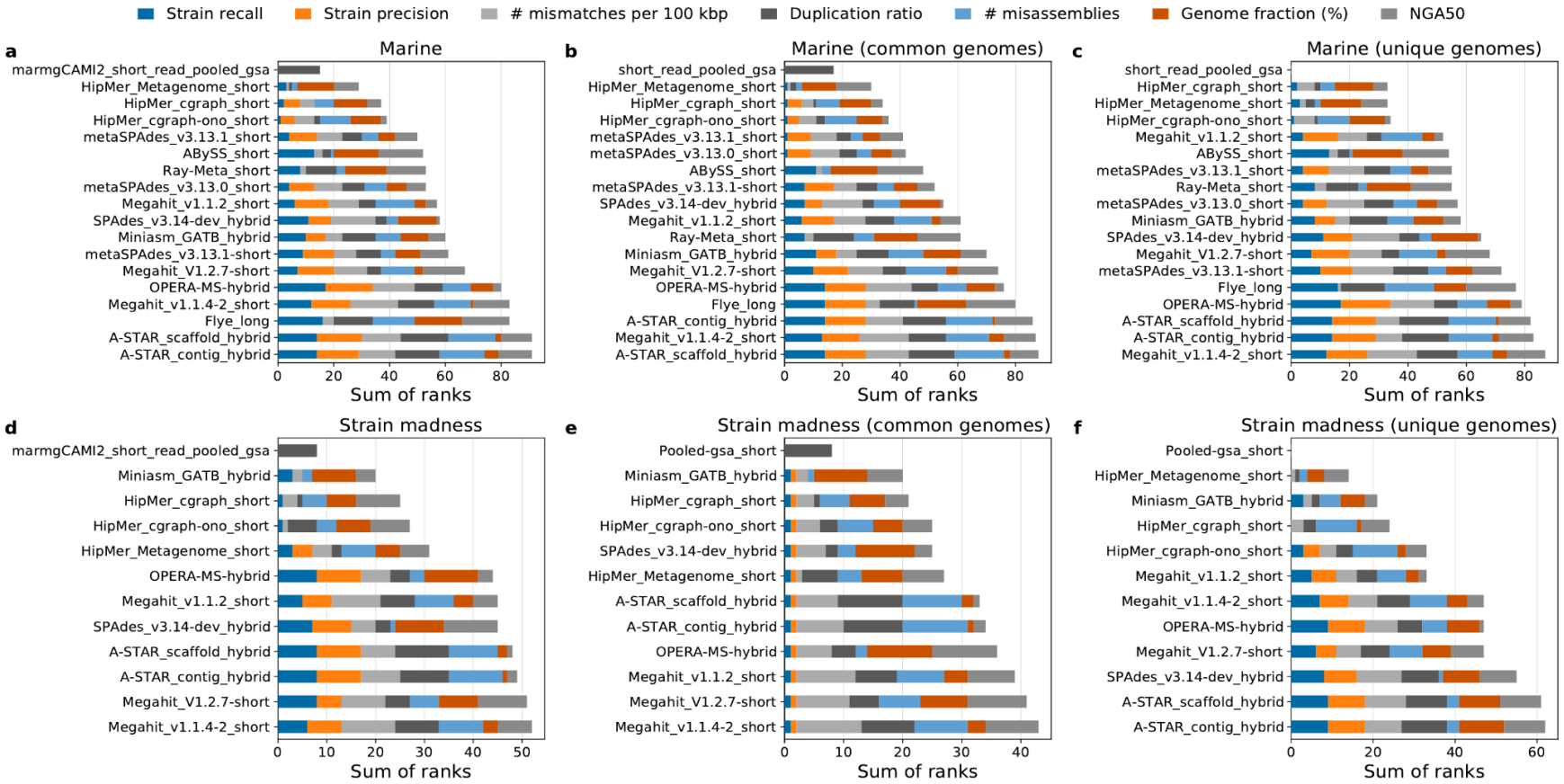


A4

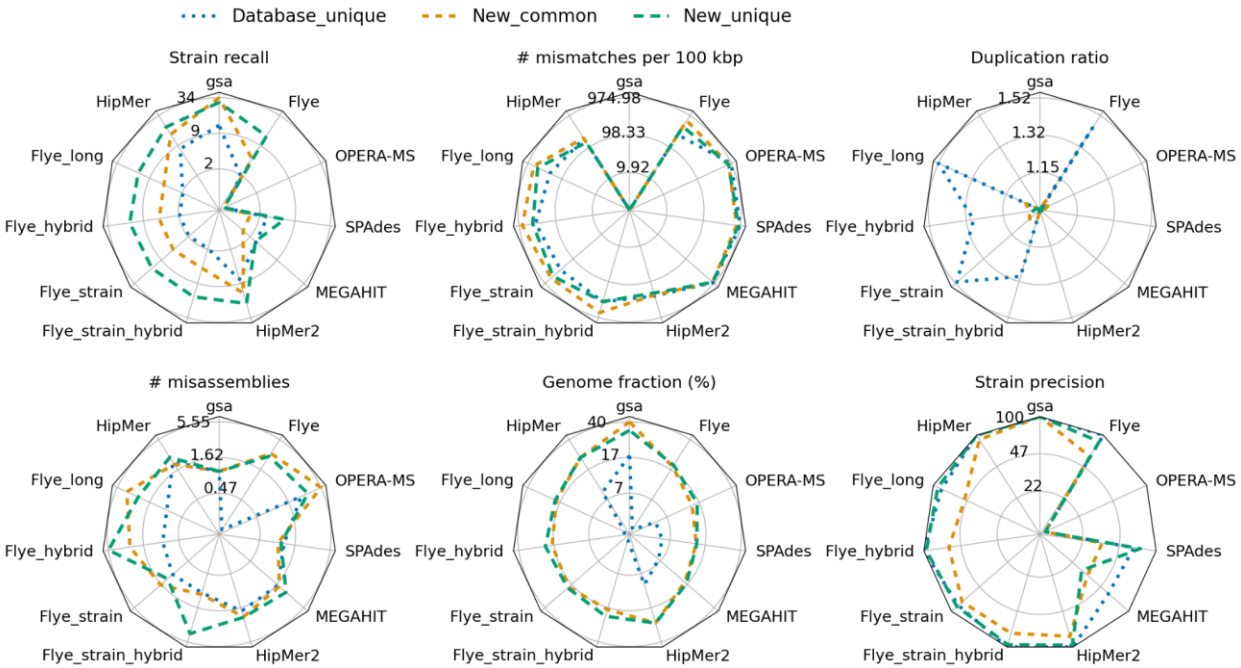


CAMI performed a strain-aware evaluation. Thus, in that context, we prefer assemblies like **A4** to assemblies like **A3** (despite the fragmentation in R2 contigs). We also prefer **A2** to **A4**.

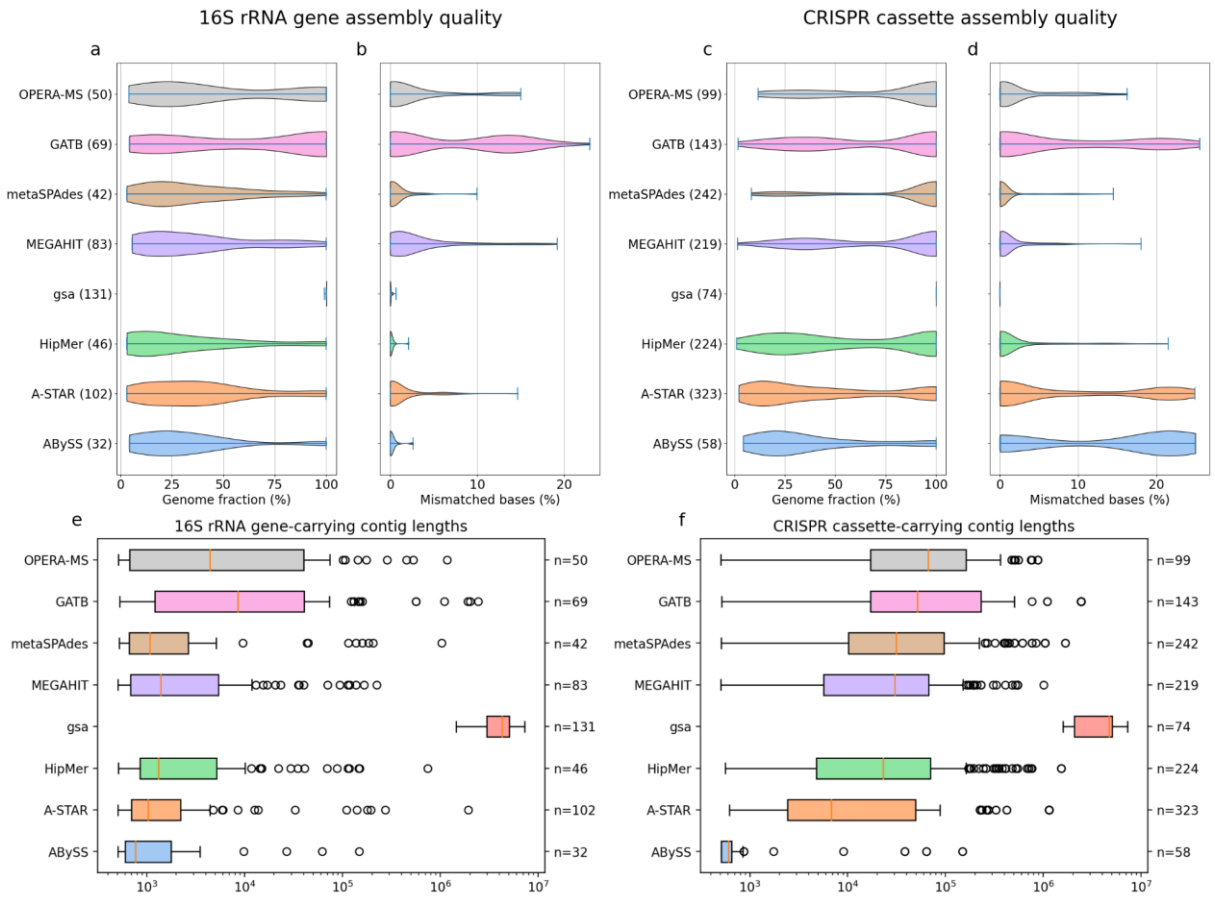
Supplementary figures



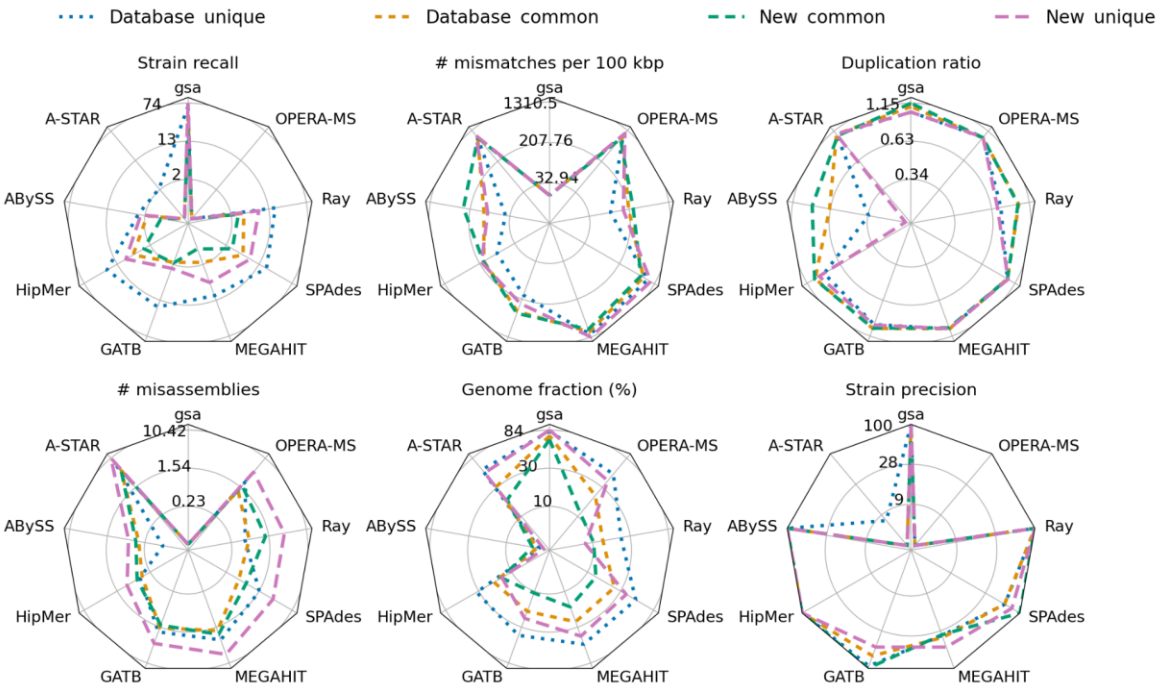
Supplementary Fig. 1: Sum of ranks per metric of assemblers on different datasets. a, b, c, All, common, and unique genomes of the marine dataset, respectively. **d, e, f,** All, common, and unique genomes of the strain madness dataset, respectively. The best assembler with a metric on a dataset gets a score of 0, the second best gets a score of 1, and so on, and are ranked accordingly, as computed in Supplementary Tables 3-7. The lower the rank of an assembler for a metric, the better the assembler performs with that metric compared to other assemblers.



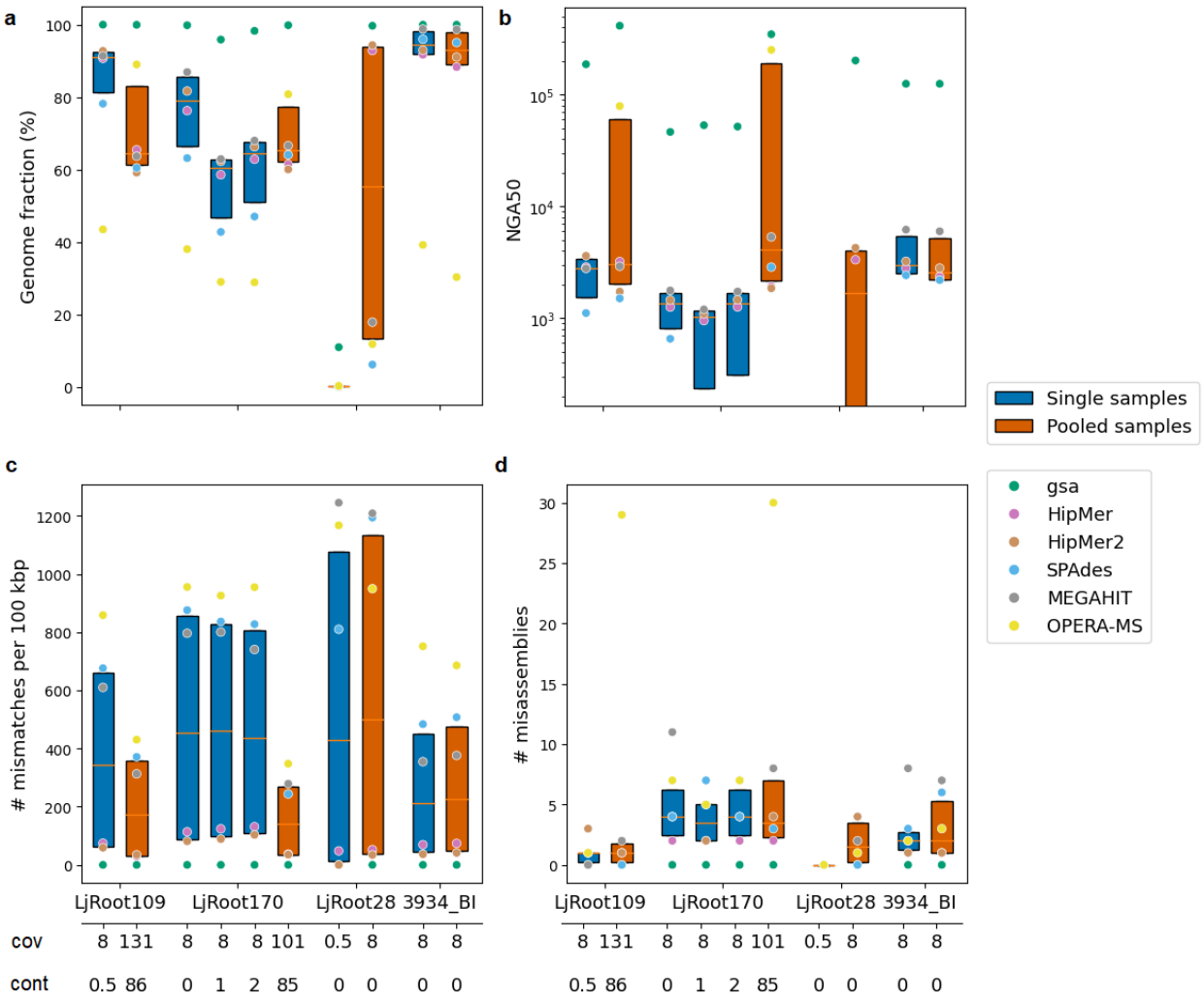
Supplementary Fig. 2: Metagenome assembler performances on the plant-associated dataset. Radar plots of strain recall, mismatches per 100 kb, duplication ratio, misassemblies, genome fraction, and strain precision.



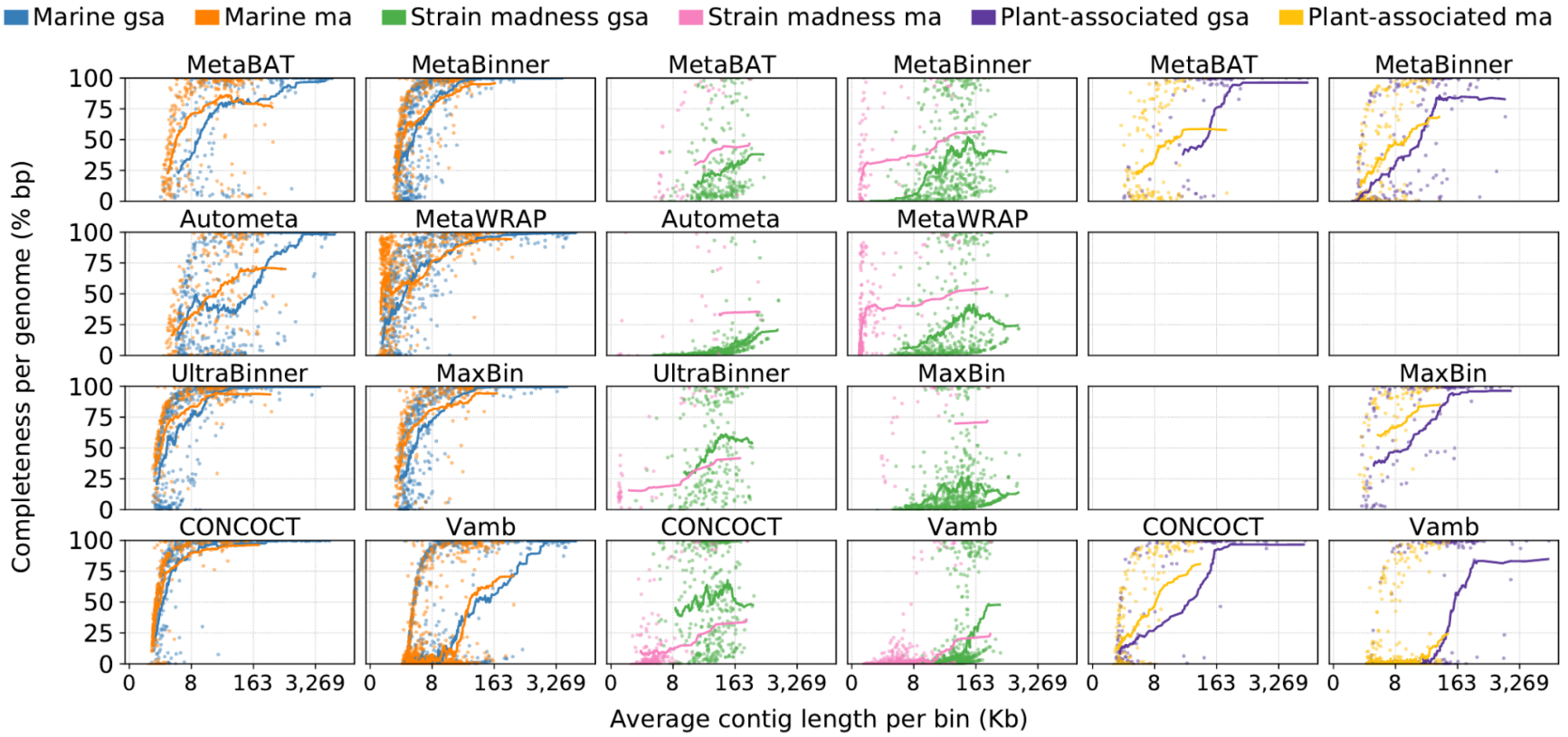
Supplementary Fig. 3: **Assembly quality of 16S rRNA gene and CRISPR cassette sequences. a, c** Genome fraction. **b, d** Divergence of assemblies on the 16S rRNA gene of 50 high-quality database genomes and for CRISPR cassettes of 30 out of these for which a cassette could be found. Evaluations were done using MetaQUAST[14,15] against the 16S rRNA gene sequence of high-quality genomes, extracted from NCBI. To avoid mappings from other genomes, the contigs were aligned to the high-quality genomes first and then all mapped contigs were evaluated individually against the corresponding genome using MetaQUAST. Completeness describes the genome fraction, divergence the gap-compressed divergence, counting consecutive gaps as single error, and blue bars show the standard deviation. The number in brackets denotes the total number of reconstructed 16S rRNA and CRISPR cassette sequences. For example, the gold standard (denoted by gsa) contained 131 16S rRNA gene sequences in the 58 genomes and A-STAR recovered 102 of them. **e, f** Lengths of the contigs aligned to the 16S rRNA gene and CRISPR cassettes in log scale. The orange line is the median, the box size is the interquartile range (IQR), and the whiskers extend to 1.5×IQR or to the maximum and minimum if there are no outliers. Outliers are contig lengths represented as points outside 1.5×IQR above the upper quartile and below the lower quartile.



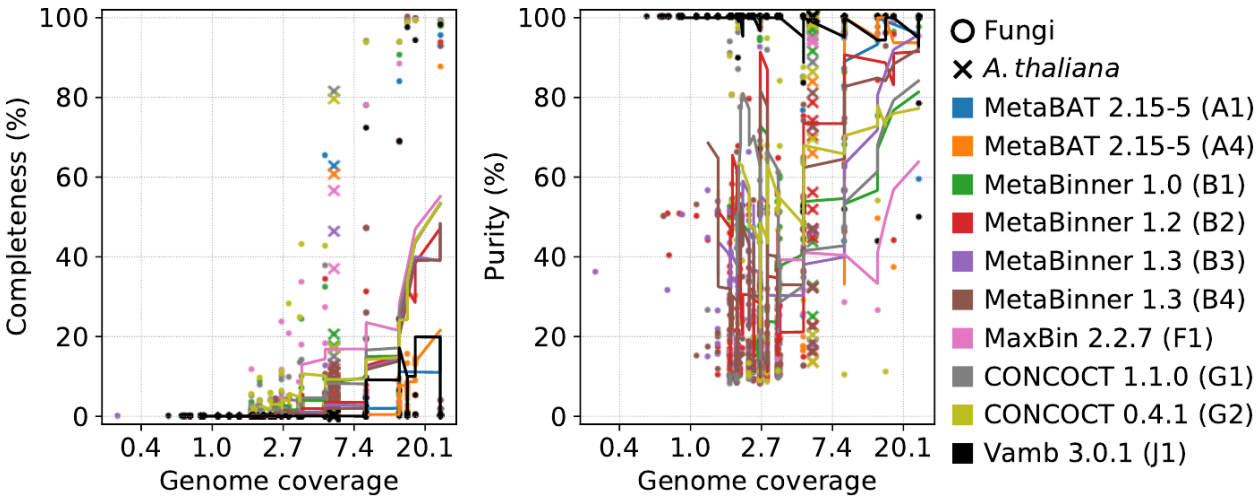
Supplementary Fig. 4: Radar plots comparing the assembly quality of new and database genomes and common and unique genomes of the marine dataset.



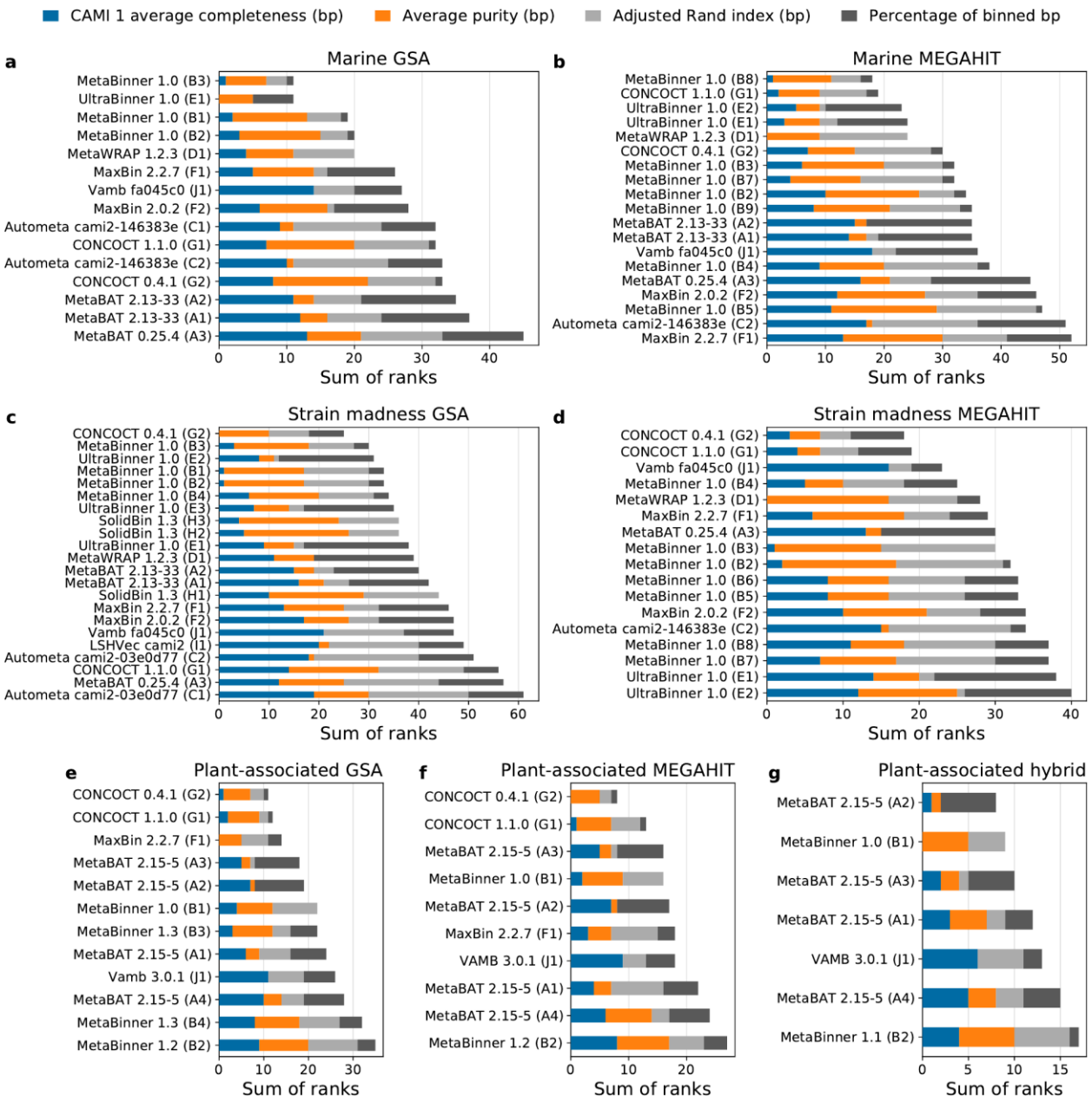
Supplementary Fig. 5: Assembly quality of single and pooled samples of the plant-associated dataset. Boxplots of **(a)** genome fraction, **(b)** NGA50, **(c)** number of mismatches per 100 kb and **(d)** number of misassemblies of four spiked genomes for single (blue) and pooled samples (orange). NGA50 is shown with log-scale; the individual assemblers for which single sample and pooled assemblies were available are color-coded; the gold standard (green) is denoted with gsa. “cov” denotes the coverage of the corresponding genome in the single or pooled samples and “cont” (contamination) the total coverage of closely related genomes present in the sample. Boxes are interquartile range, the pink line the mean, and all individual points are shown.



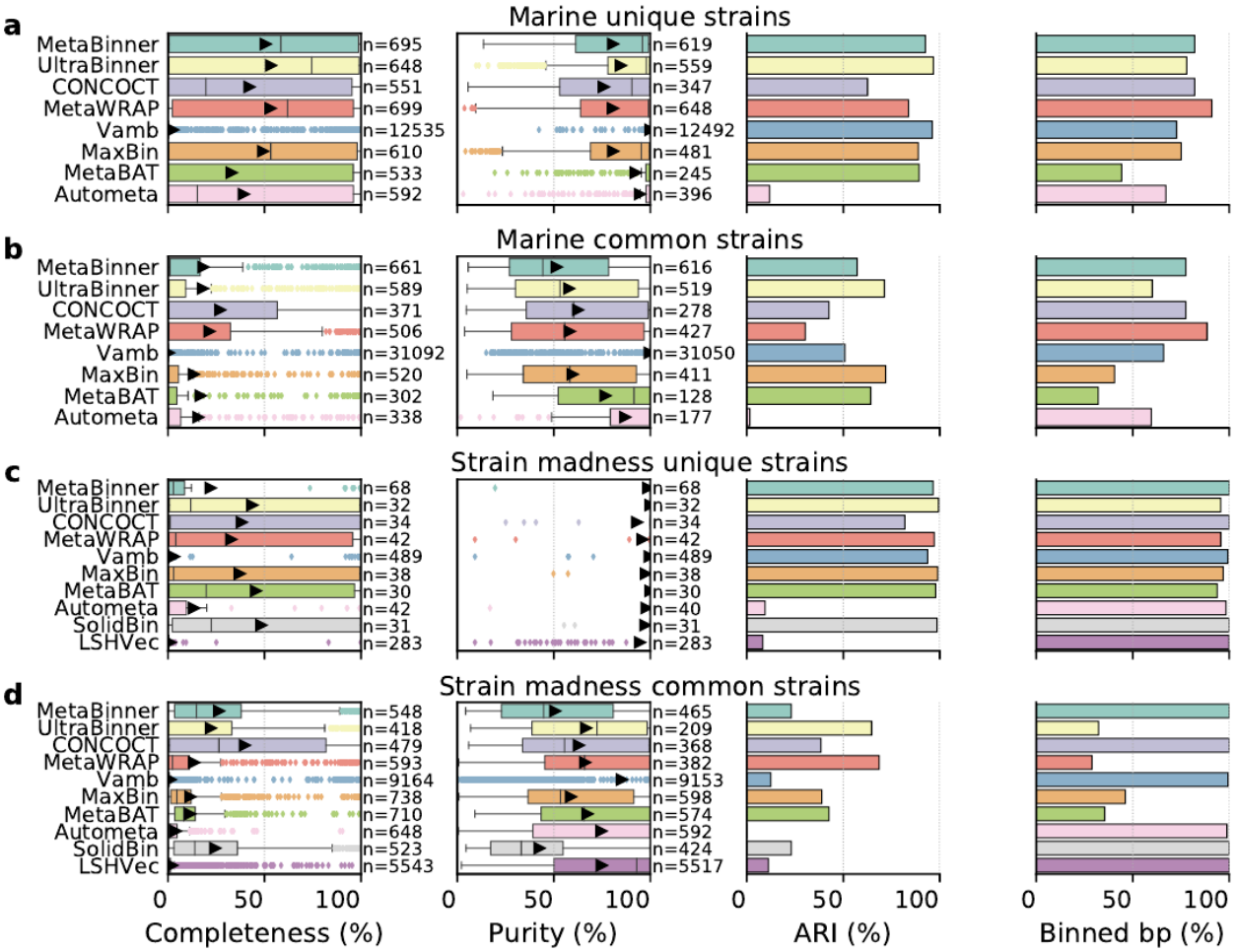
Supplementary Fig. 6: **Average length of the contigs per genome bin vs. the completeness of the corresponding genome on different datasets per genome binner.** “gsa” and “ma” are the binnings of the gold standard and MEGAHIT assemblies of a dataset, respectively. Lines give the running average completeness of 50 consecutive genomes ordered by the average contig length in the bins. For each method, the best-performing submission in terms of F1-score is shown.



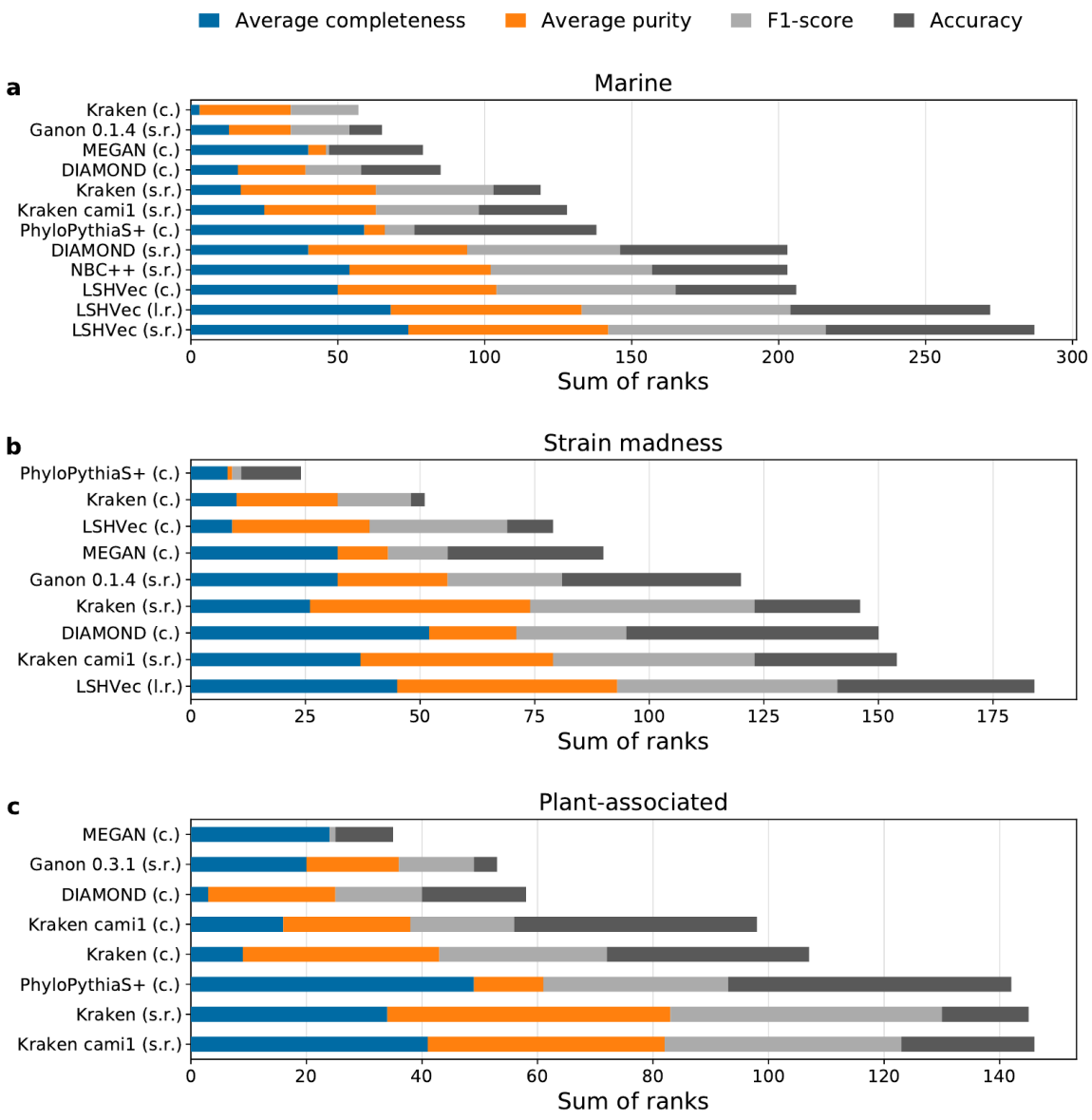
Supplementary Fig. 7: **Coverage of fungal genomes (circles) and the *A. thaliana* genome (cross marks) in the plant-associated dataset recovered by genome binners vs. completeness per genome (left) and purity per bin (right).** Lines indicate the running average completeness or purity of 10 consecutive fungal genomes or bins ordered by coverage. Raw data is available at https://github.com/CAMI-challenge/second_challenge_evaluation/blob/master/assembly/scripts/data/funghi_coverage.tsv.



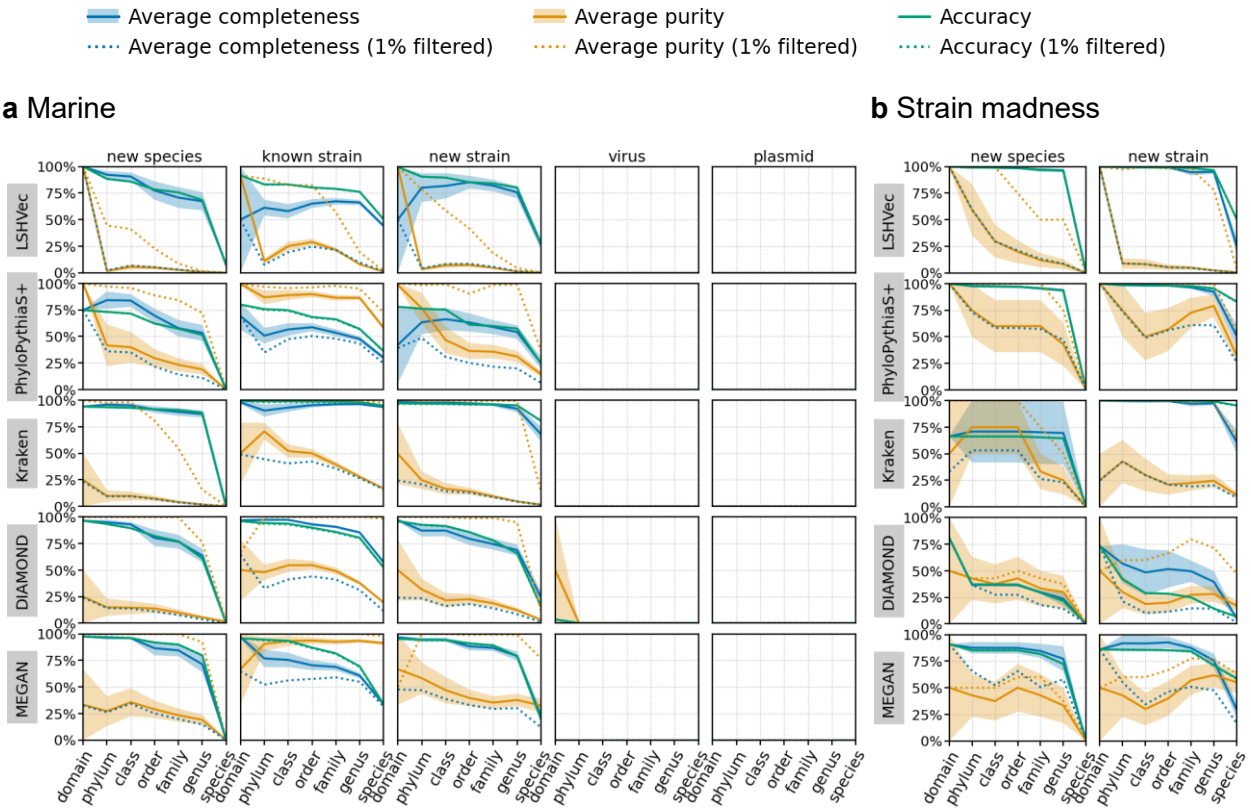
Supplementary Fig. 8: Overall summary ranking score, or sum of ranks per metric, of genome binners on different data. **a, b,** Marine gold standard assembly (GSA) and MEGAHIT assembly, respectively. **c, d,** Strain madness GSA and MEGAHIT assemblies, respectively. **e-g,** Plant-associated GSA, MEGAHIT, and hybrid assemblies, respectively. The best genome binner with a metric on a dataset gets a score of 0, the second best gets a score of 1, and so on, and are ranked accordingly, as computed in Supplementary Tables 17-19. The lower the rank of a genome binner for a metric, the better it performs with that metric compared to other genome binners.



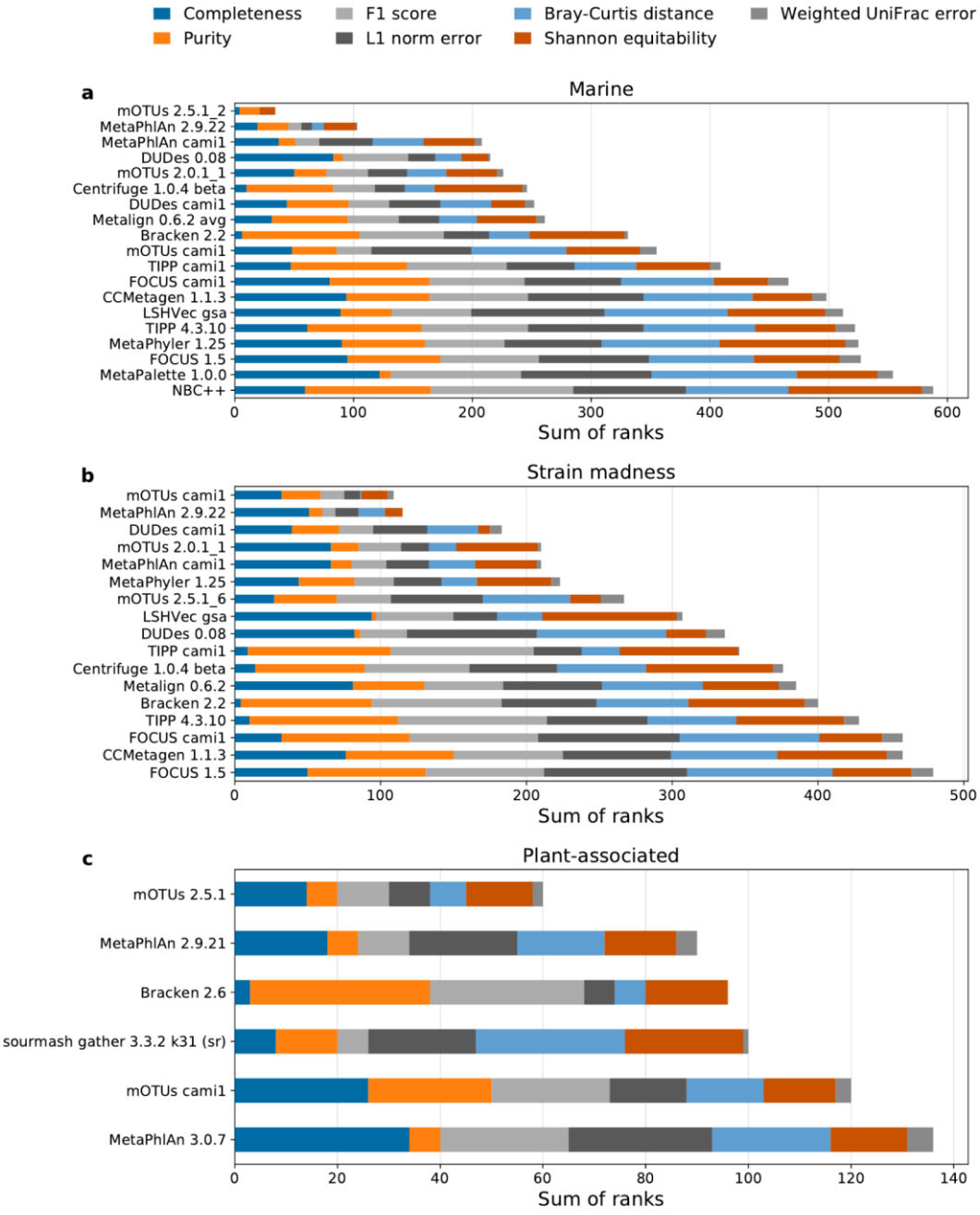
Supplementary Fig. 9: Effect of strain diversity on average completeness, purity, adjusted Rand index (ARI), and percentage of binned bp for genome binning of unique and common strains of the marine and strain madness short-read GSAs. Genomes were reconstructed by genome binners for genomes of unique strains with ANI < 95% to others and common strains with ANI ≥ 95% to each other. Boxes in boxplots indicate the interquartile range (IQR) of *n* results, the center line the median, and arrows the average. Whiskers extend to 1.5×IQR or to the maximum and minimum if there is no outlier. Outliers are results represented as data points outside 1.5 * IQR above the upper quartile and below the lower quartile.



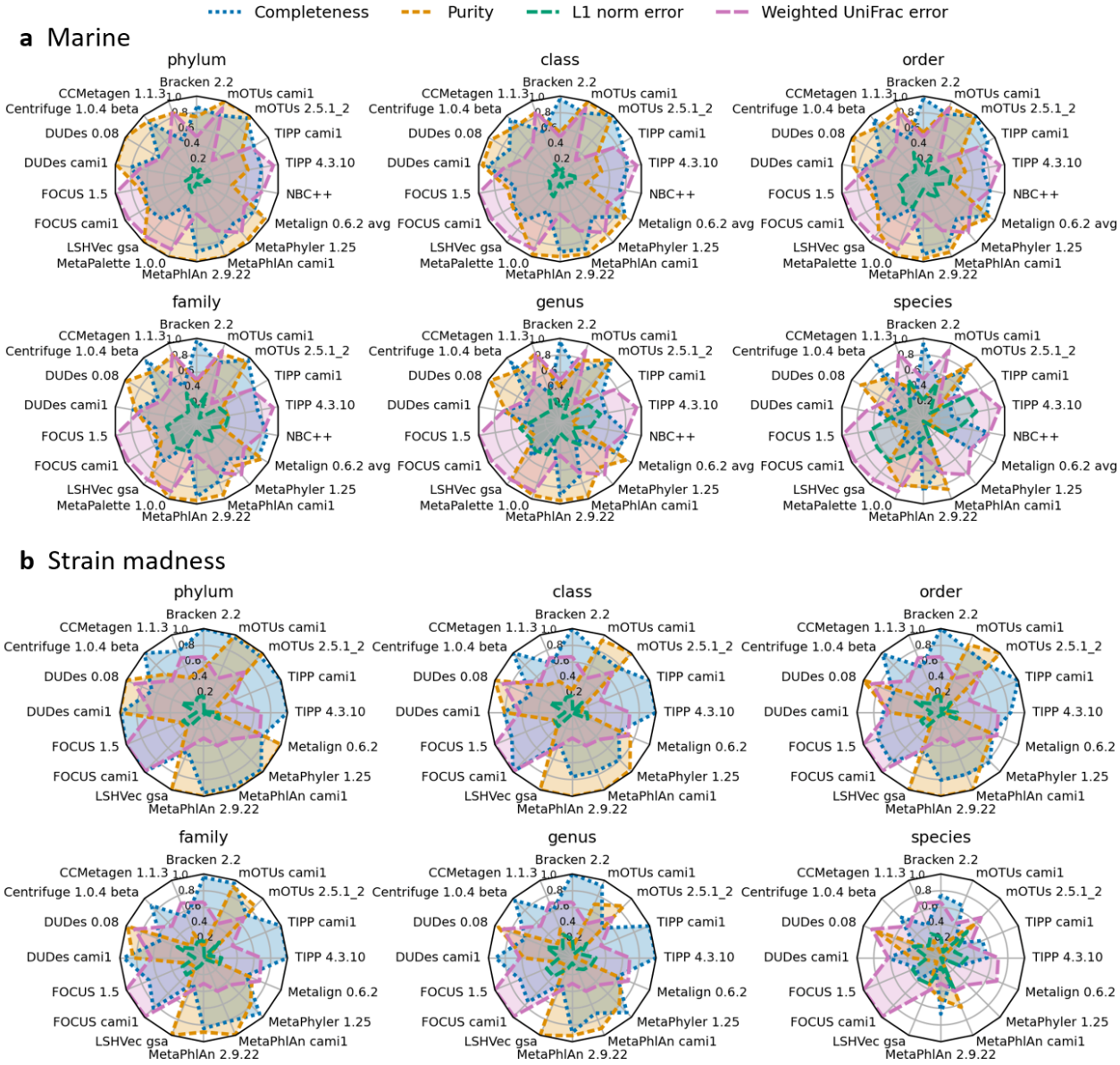
Supplementary Fig. 10: Overall summary ranking score, or sum of ranks per metric, across taxonomic levels (from domain to species) of taxonomic binners on different datasets. a, Marine. **b,** Strain madness. **c,** Plant-associated. The best taxonomic binner with a metric on a dataset and taxonomic level gets a score of 0, the second best gets a score of 1, and so on. The scores are then summed over the taxonomic levels for each metric, as computed in Supplementary Tables 27-29. The lower the rank of a taxonomic binner for a metric, the better it performs with that metric compared to other binners. Abbreviations of target data: c.: contigs, s.r.: short reads, l.r.: long reads.



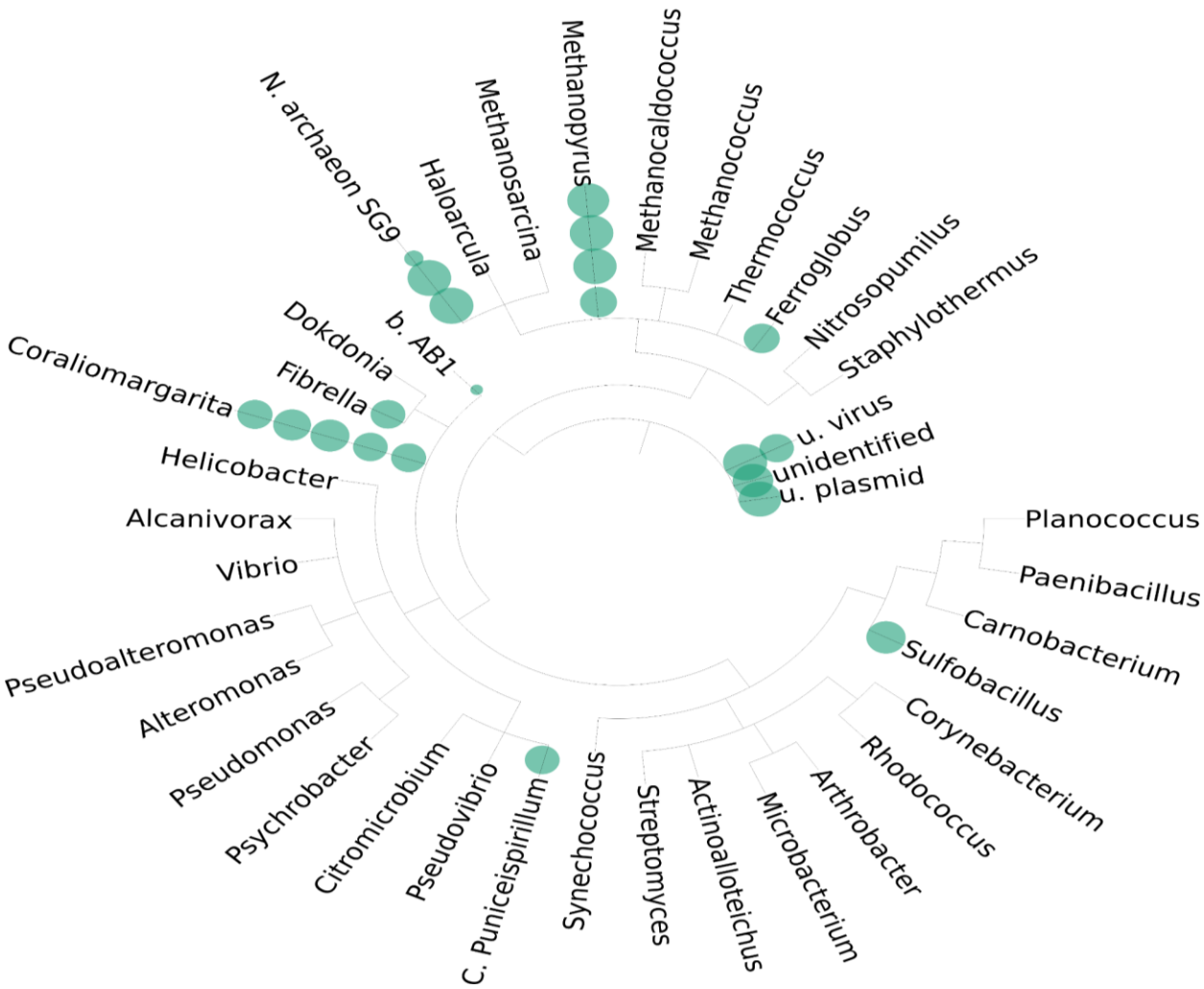
Supplementary Fig. 11: Taxonomic binning performance across ranks for the (a) marine and (b) strain madness datasets split by their taxonomic distances to public genomes as new species or strains, known species or known strains, viruses, or plasmids. A genome is classified as “new species” if no genome of that species is present in the NCBI RefSeq database. Metrics are computed over unfiltered and 1% filtered predicted bins (see main text). Shaded bands show the standard error across bins.



Supplementary Fig. 12: Overall summary ranking score, or sum of ranks per metric, across taxonomic levels (from domain to species) of taxonomic profilers on different datasets. a, Marine. **b,** Strain madness. **c,** Plant-associated. The best taxonomic profiler with a metric on a dataset and taxonomic level gets a score of 0, the second best gets a score of 1, and so on. The scores are then summed over the taxonomic levels for each metric, as computed in Supplementary Tables 33, 35, and 37. The lower the rank of a taxonomic profiler for a metric, the better it performs with that metric compared to other profilers.

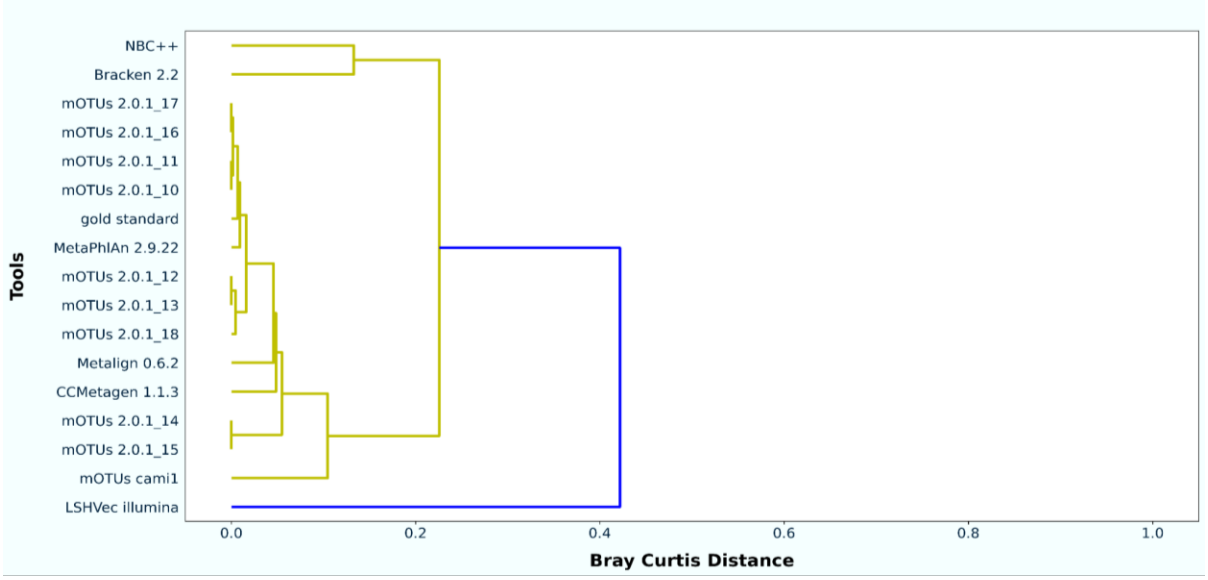


Supplementary Fig. 13: Taxonomic profiling results for the marine (a) and strain madness (b) datasets from phylum to species ranks.
L1 norm is divided by 2 to be in the range between 0 and 1, as completeness and purity. Weighted UniFrac error is normalized by the maximum value obtained by a method.

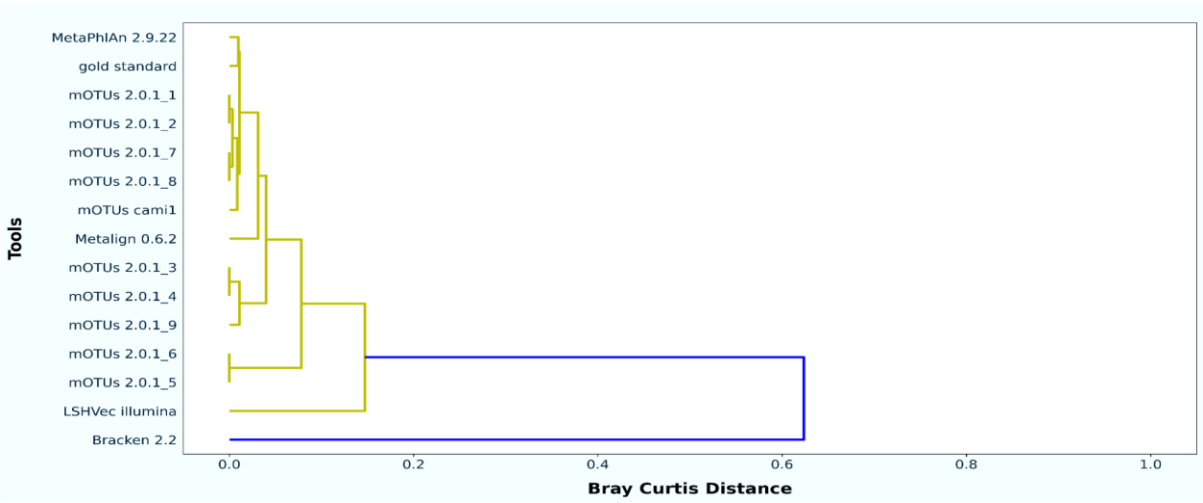


Supplementary Fig. 14: Taxonomic tree depicting most difficult to detect taxa in the marine dataset.
MeTEA (http://github.com/EESI/TEA) was utilized to identify the gold standard taxa most frequently missed by all tools. TAMPA (https://github.com/dkoslicki/TAMPA) was then used to depict these taxa on a tree. Green discs represent abundance in the ground truth, averaged over all marine datasets. Multiple discs represent different taxonomic ranks.

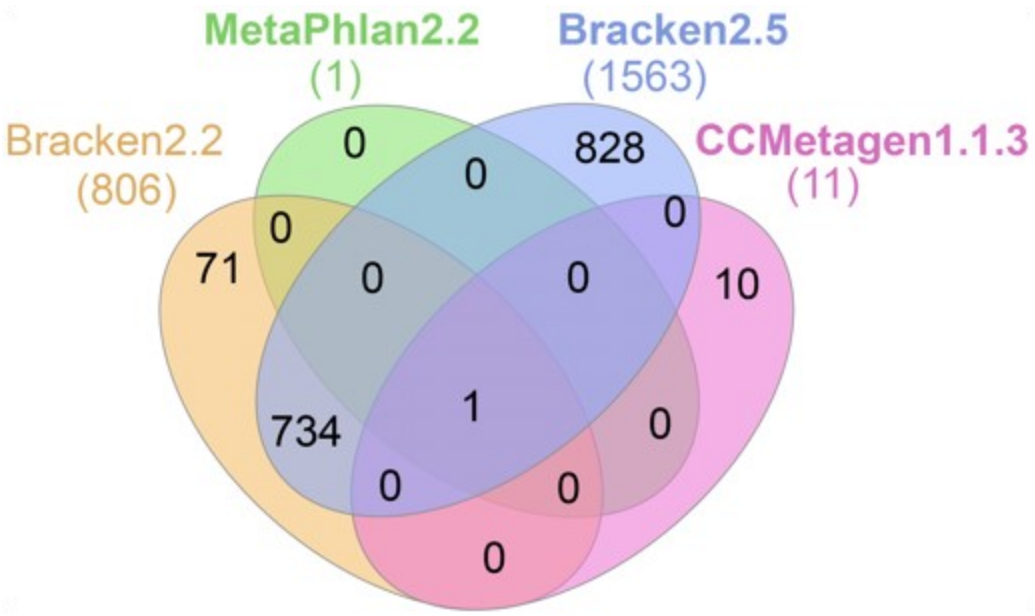
a



b



Supplementary Fig. 15: Tool performance clustering.
MeTEA (http://github.com/EESI/TEA) was utilized to cluster tool performance in the following way: at each taxonomic rank, MeTEA computed the F1 score for each tool and each taxon in the gold standard and then averaged over all ranks. The Bray-Curtis dissimilarity was used for the hierarchical clustering. Part **a** shows tool similarity for short-read marine dataset submissions and **b** depicts tool similarity for those that submitted results for the short-read strain madness dataset.
Methods using similar information types, e.g., k-mer based (NBC++, Bracken), alignment (CCMetagen, Metalign), and marker gene approaches (mOTUs, MetaPhlAn) tended to cluster; for example, the two alignment-based approaches are more similar to each other than to other methods. Interestingly, the marker gene approaches are most similar to the gold standard, suggesting this class of methods is particularly well suited to infer taxonomic profiles.



Supplementary Fig. 16: Venn diagram of taxa predicted by different submissions for the clinical pathogen detection challenge. Shown are methods that included the causal pathogen among the predicted taxa (total number in brackets) submitted to the challenge.

References

1. Fritz, A. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
2. Nguyen, T. T. & Landfald, B. Polar front associated variation in prokaryotic community structure in Arctic shelf seafloor. *Front. Microbiol.* **6**, 17 (2015).
3. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
4. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
5. Klemetsen, T. *et al.* The MAR databases: development and implementation of databases

specific for marine metagenomics. *Nucleic Acids Res.* **46**, D692–D699 (2018).

6. Mende, D. R. *et al.* proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res.* **48**, D621–D625 (2020).
7. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
8. Bremges, A., Fritz, A. & McHardy, A. C. CAMITAX: Taxon labels for microbial genomes. *Gigascience* **9**, (2020).
9. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
10. Jørgensen, T. S. *et al.* Plasmids, Viruses, And Other Circular Elements In Rat Gut. *bioRxiv* 143420 (2017) doi:10.1101/143420.
11. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
12. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
13. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–22 (2010).
14. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
15. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).