nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-020-00480-3

Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit

Fernando Meyer[1], Till-Robin Lesker[1,2], David Koslicki[3], Adrian Fritz[1], Alexey Gurevich[4], Aaron E. Darling[5], Alexander Sczyrba[6], Andreas Bremges[1,2] and Alice C. McHardy[1]✉

Computational methods are key in microbiome research, and obtaining a quantitative and unbiased performance estimate is important for method developers and applied researchers. For meaningful comparisons between methods, to identify best practices and common use cases, and to reduce overhead in benchmarking, it is necessary to have standardized datasets, procedures and metrics for evaluation. In this tutorial, we describe emerging standards in computational metaomics benchmarking derived and agreed upon by a larger community of researchers. Specifically, we outline recent efforts by the Critical Assessment of Metagenome Interpretation (CAMI) initiative, which supplies method developers and applied researchers with exhaustive quantitative data about software performance in realistic scenarios and organizes community-driven benchmarking challenges. We explain the most relevant evaluation metrics for assessing metagenome assembly, binning and profiling results, and provide step-by-step instructions on how to generate them. The instructions use simulated mouse gut metagenome data released in preparation for the second round of CAMI challenges and showcase the use of a repository of tool results for CAMI datasets. This tutorial will serve as a reference for the community and facilitate informative and reproducible benchmarking in microbiome research.

Since the release of the first shotgun metagenome from the Sargasso Sea by metagenomics pioneer Craig Venter[1], the field has witnessed an explosive growth of data and methods. Microbiome data repositories[2,3] host hundreds of thousands of datasets, and numbers are still rising rapidly.

Metagenomics (see glossary; Table 1) created new computational challenges, such as the need to reconstruct the genomes of community members from a mixture of reads originating from potentially thousands of microbial, viral, and eukaryotic taxa[4]. These taxa differ in their relatedness to each other, are often absent from sequence databases, and present at varying abundances. Genomes can be reconstructed by metagenome assembly, which creates longer, contiguous sequence fragments, followed by binning, which is usually a clustering method that places fragments into genome bins. There have been spectacular successes in recovering thousands of metagenome-assembled genomes (MAGs) for uncultured taxa[5–7]. Identifying the taxa and their abundances for a community is known as taxonomic profiling; taxonomic 'binners' assign taxonomic labels to individual sequence fragments. Both tasks are challenging, particularly for the lower taxonomic ranks[8]. Another challenge is the de novo assembly of closely related genomes (>95% average nucleotide identity)[8]. Finally, fragmentary assemblies with many short contigs obtained from short read sequence data in metagenomics have required adaptation of gene-finding methods and complicate operon-level functional analyses of genes. The maturation of long-read sequencing technologies[9,10], which for many years were characterized by low throughput, high cost, and high error rates, has sparked further development and is expected to lead to better solutions for some of these challenges.

The relevance of standards for performance evaluation and benchmarking

Methodological development is oftentimes accompanied by performance evaluations. This has historically been done on an ad hoc basis by developers, often using different datasets and performance metrics, which are both critical choices in regard to performance evaluation. This practice made it difficult to compare results across publications and to identify suitable techniques for specific datasets and tasks. It also made performance benchmarking for developers tedious and ineffective. For instance, performance might differ substantially for reference-based methods using public databases across datasets, depending on evolutionary divergence between the sampled and the database taxa[8]. Similarly, organismal complexity, strain-level diversity, realistic community genome abundance distributions, the presence of non-bacterial genomic information, and sequencing error profiles of datasets are some of the

[1]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Braunschweig, Germany. [2]German Center for Infection Research (DZIF), Braunschweig, Germany. [3]Computer Science and Engineering, Biology, and The Huck Institutes of the Life Sciences, Penn State University, State College, PA, USA. [4]Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia. [5]The ithree institute, University of Technology Sydney, Sydney, Australia. [6]Faculty of Technology and Center for Biotechnology, Bielefeld University, Bielefeld, Germany. ✉e-mail: alice.mchardy@helmholtz-hzi.de

Table 1 | Glossary

Term	Definition
Assembly	Reconstruction of complete or partial genomes or DNA sequence fragments, often by merging sequence reads into longer pieces called contigs
Benchmarking	Systematic comparison of (computational) techniques using performance metrics in specific scenarios
Binning	Clustering or classification of sequences or contigs into bins representing genomes (genome binning) or taxa (taxonomic binning) of the underlying microbial community
Coverage	Number of sequence reads that cover a certain genomic position
Docker	A software tool designed to make it easy to distribute and run applications by using software packages (containers) and operating system-level virtualization
Metagenomics	A set of techniques for recovering and sequencing of the genetic material of microbial communities and their functional and taxonomic characterization
Profiling	Microbial community characterization from a metagenomic sample in terms of presence and absence of taxa and their relative abundances
Standard of truth/ground truth/ gold standard	The correct result, such as, for example, the correct taxon that a sequence originates from, which can be used to compare with and benchmark other methods' results

factors that may affect method performance. It became evident, as in other fields[11–13], that standards would greatly facilitate comparisons across methods and articles, as well as unequivocal determination of appropriate solutions and open challenges.

CAMI

To satisfy this need, CAMI, the community-driven initiative for the Critical Assessment of Metagenome Interpretation, was founded in 2014 by A. Sczyrba, T. Rattei, and A.C. McHardy (http://blogs.nature.com/methagora/2014/06/the-critical-assessment-of-metagenome-interpretation-cami-competition.html) during the metagenomics program at the Isaac Newton Institute in Cambridge (https://www.newton.ac.uk/event/mtgw01). CAMI design decisions are based on feedback gathered in community workshops, which ensures inclusion of a wide range of expert inputs and establishes a community consensus. By regularly interacting with scientists in workshops, during hackathons, and at conferences such as the Microbiome track of the international conference on Intelligent Systems for Molecular Biology (ISMB), CAMI aims to identify and implement best practices for benchmarking in microbiome research, including (i) key properties of benchmark datasets (see also refs. [14,15] for an overview of general benchmarking practices); (ii) appropriate performance metrics for different tasks; (iii) benchmarking procedures, that is, how to run benchmarking challenges; and (iv) performance evaluation procedures, to allow the most realistic, fair, and unbiased assessment. (v) Reproducibility/reusability has been identified as the fifth key criterion. We provide further details on these key aspects in the sections below.

The first CAMI challenge took place in 2015 and provided an extensive performance overview for commonly used data processing methods, namely those for assembly, genome and taxonomic binning, and taxonomic profiling[8]. The six benchmark datasets—reflecting a range of complexities—have since been used extensively for further benchmarking in the field. These include three 'toy' datasets created from public data and provided before a challenge, as well as three 'challenge' datasets derived exclusively from genomic data that were not publicly available at the time. These genomic data are now in public sequence repositories such as the National Center for Biotechnology Information (NCBI). All CAMI benchmark datasets are made available after the challenges with digital object identifiers (DOIs) (Table 2) and are also downloadable from the CAMI portal at https://data.cami-challenge.org/. Further benchmarking studies have also provided valuable insights into the performance of data processing methods[16–19]. The second CAMI challenge (CAMI II) was launched in 2019 and offered challenges for the same tasks on two large, multisample datasets reflecting specific environments (marine, rhizosphere) and an extremely high strain diversity dataset (strain madness). In addition, a clinical pathogen detection challenge was offered. The challenges on the marine, strain madness, and pathogen datasets closed in October 2019, whereas the challenge on the rhizosphere dataset, which was launched in 2020, will close in early 2021. The results are expected to provide insights into important questions such as the potential of long-read data for metagenomics[20].

Advantages of benchmarking challenges

Challenges provide insights into method performance, suggesting best practices as well as identifying open problems in the field. They can also further the development and adoption of standards, such as data input and output formats, or choice of reference datasets, such as the NCBI taxonomy. Once standards are realized, benchmarking competitions offer a low-effort opportunity for extensive benchmarking, as datasets, other method results, and evaluation methods do not have to be created by the developer of a new metagenome analysis method.

Some participants might worry about publishing poor performances, which is why CAMI challenge participants can opt out of results publication and use them only for their own benefit. Defining the evaluation metrics is also open for the field; thus, all labs participating in these discussions can

Table 2 | CAMI benchmark datasets and their respective DOIs

CAMI benchmark dataset	DOI
CAMI I: low, medium, high complexity, and 'toy' datasets	10.5524/100344
CAMI II: human microbiome project and mouse gut toy datasets	10.4126/FRL01-006425518 and 10.4126/FRL01-006421672
CAMI II: marine, strain madness, rhizosphere, and pathogen detection challenge datasets	10.4126/FRL01-006425521

All datasets are also downloadable from the CAMI portal at https://data.cami-challenge.org/.

contribute to the challenge evaluation. Participants can thus suggest and define metrics that highlight the expected benefits of their techniques, with these simultaneously being subjected to peer group review. To ensure a maximum of objectivity in these evaluations, CAMI challenges are performed blinded in two ways. The standard of truth (see glossary; Table 1) for the challenge dataset is provided only after the challenges end, preventing performance optimization in any way on these particular datasets. Challenge datasets include many genomes that will become publicly available only after the challenge. 'Toy' datasets, for which a standard of truth is made available at the outset, are provided before the actual challenges to enable teams to familiarize themselves with the data structure and its properties. The evaluation of the different challenge submissions is also performed blindly, such that the evaluation panel does not know the names of the submitters or information about the submitted techniques, to tackle evaluator biases. Evaluations are open to anyone wishing to participate, and a consensus is reached in a workshop with a group of experts.

Benchmark datasets

Benchmark datasets should be as realistic and representative of real meta-omics data as possible. For CAMI challenges, experimental groups contribute unpublished genomes, including some organisms from poorly characterized phyla without any publicly available genomes of close relatives. These genomes are used for benchmark data creation and are published only after the challenge. Because many taxa present in real environmental samples have unknown cultivation conditions and no isolate genomes are available in reference databases, measuring performance on novel organisms is essential. This is particularly true for a comprehensive evaluation of reference-based methods such as taxonomic profilers and binners, which perform best for genomes closely related to those in public databases[8]. The challenge datasets for CAMI I were created entirely from newly sequenced genomes that were unpublished at the time of the challenge with the CAMISIM microbial community and metagenome simulator[21]. For CAMI II, both unpublished and already-public high-quality genomes were used to allow a more comprehensive assessment of assembly qualities. CAMISIM allows the incorporation of many key properties into datasets, such as varying experimental designs (number of samples, sequencing depth, insert sizes, type of experiment, including differential abundances, time series) and sequencing technologies and community properties (organismal complexity, different genome abundance distributions, strain diversity, taxa from different domains of life, viruses, mobile circular elements). An alternative way to create benchmark data is to sequence lab-created DNA mixtures as in ref. [22], which would enable a more realistic assessment of technical variation and biases introduced in data generation. However, creating communities with realistic organismal complexities is currently impractical for many environments, which can have hundreds to thousands of genomes at highly varying abundances.

Metrics for performance evaluation

Choosing the appropriate (combination of) metrics for comparing method performances—such as fraction of correctly assembled genomes and number of contigs, or number of correctly identified taxa in taxonomic profiling and binning—is a key task in benchmarking that directly influences the ranking of methods. The metrics used in CAMI challenges[8] are decided on in public workshops and reassessed regularly. They should be easy to interpret and meaningful to both developers and applied scientists. A comprehensive assessment is achieved by including multiple metrics that highlight the strengths of different approaches (see 'Benchmarking demonstration' section). Furthermore, assessing properties such as runtime, disk space, and memory consumption is important.

Reproducibility and FAIR (Findable, Accessible, Interoperable, Reusable) principles

Imagine running a benchmarking contest and identifying the top-performing technique by key criteria, potentially representing the new state-of-the-art for future studies. However, the submitting team has unfortunately lost track of the software version and parameter settings used, and is unable to reproduce its own results. To avoid such issues, reproducibility has been selected as a core principle in CAMI for all steps of benchmarking, from data generation with CAMISIM[21] to running software benchmarked in the contest, and to evaluating results. Evaluation metrics are extensively tested and implemented in the MetaQUAST[23], AMBER[24], and OPAL[25] benchmarking packages (Table 3), available via Bioconda[26]. All software released by CAMI is available as open source under the appropriate licenses, such as Apache 2 or GPL. A key lesson learned from the first challenge was that parameter settings substantially affect program performance. A minimal requirement for public CAMI challenge results is therefore documenting the exact program versions and command-line calls or, even better, using a workflow manager such as GNU Make

Table 3 | CAMI benchmarking software packages

Software	Description
CAMISIM[21]	A microbial community and metagenome simulator that models different microbial abundance profiles, multisample time series, and differential abundance studies, as well as real and simulated strain-level diversity, and generates second- and third-generation sequencing data from taxonomic profiles or de novo. CAMISIM was used to generate several benchmark datasets for CAMI challenges
MetaQUAST[23]	A quality assessment tool for metagenome assembly evaluation. It computes various quality metrics on the basis of alignment of assemblies to a standard of truth or close reference genomes. A standard of truth is used in CAMI
AMBER[24]	Software for the comparative assessment of genome reconstructions and taxonomic assignments from metagenome benchmark datasets. It calculates performance metrics such as (rank-specific taxon) bin completeness and purity, average Rand index, assignment accuracy, and comparative visualizations used in CAMI challenges
OPAL[25]	A tool for computing performance metrics and creating visualizations for assessing taxonomic metagenome profilers. The metrics include presence–absence metrics (number of true and false positives, false negatives, completeness, purity, F1 score, Jaccard index) and abundance metrics such as UniFrac, L1 norm and the Bray–Curtis distance
Bioboxes[29]	Docker containers with standardized interfaces facilitating interchange of software in bioinformatics pipelines, distribution of specific software versions with predefined parameter settings, and therefore reproducibility of results and benchmarking. The Bioboxes standard was used to containerize the methods benchmarked in the CAMI I challenges and are continuously used along with BioContainers[30] and workflow and package managers such as Snakemake[27], Nextflow[28], and Bioconda[26]

(https://www.gnu.org/software/make/), Snakemake[27], Nextflow[28], or CWL (Figshare repository: https://doi.org/10.6084/M9.FIGSHARE.3115156.V2). The ideal, although time-consuming, approach is to containerize the program, for example, in Docker, Bioboxes[29], or BioContainers[30], as well as to document and bundle dependencies to facilitate installation with pip (https://pypi.org/project/pip/) or Bioconda[26].

To maximize the scientific value, not only the methods, but also all data required for reproducing and building on the results of a study should be made available. CAMI commits to the FAIR principles for scientific data management and stewardship[31]. CAMI benchmark and reference datasets, program results, and computed metrics are provided with DOIs on Zenodo (https://zenodo.org/communities/cami) and GigaDB (http://gigadb.org/dataset/100344). This improves reusability and sustainability of the efforts, as others can directly build on a study, for instance, by adding their own method's results to the existing results of a benchmarking effort or by adding calculation of new metrics to a benchmark study for more sophisticated interpretation.

CAMI benchmarking workflow

A schematic representation of CAMI's benchmarking workflow is shown in Fig. 1. In the following, we demonstrate this principle of convenient benchmarking by extending previous results for the four software categories (assembly, genome and taxonomic binning, and profiling) benchmarked on the CAMI II multisample mouse gut dataset, creating a flexible benchmarking resource for individual studies.

Benchmarking demonstration

We demonstrate how to benchmark in practice according to the standards developed in the context of CAMI—for example, in terms of benchmarking metrics and file formats—and realized in benchmarking software (Table 3) for different challenges in computational metagenomics, such as assembly, genome and taxon binning, and taxonomic profiling. We analyze the mouse gut metagenome 'toy' dataset[21] provided to prepare for CAMI II (Table 2), starting below with a description of its simulation. Analyses of this dataset with several taxonomic profiling and assembly methods were previously described[21,25]. The benchmarked assemblers, taxon and genome binners, and taxonomic profilers were chosen on the basis of popularity and performance in the first CAMI challenge[8]. All method results for this and other benchmark datasets can be obtained from a new resource on Zenodo at https://zenodo.org/communities/cami, and curated metadata are provided at https://github.com/CAMI-challenge/data. Users can continue to add results to these repositories, thus building a growing method result collection for benchmarking.

Simulation of benchmark dataset

The mouse gut metagenome toy dataset was generated with CAMISIM v.0.2 (ref. [21]; Table 3) using a microbial community genome abundance distribution modeled from 791 public prokaryotic genomes marked as at least 'scaffolds' in the NCBI RefSeq[32]. They comprise 8 phyla, 18 classes, 26 orders, 50 families, 157 genera, and 549 species. The community genome abundance distribution matches as closely as possible the 16S taxonomic profiles for 64 mouse gut samples. As such, this dataset enables us to assess how well sequenced community members can be characterized with different techniques from the metagenomes of similar communities. On average, within each of the 64 samples, 91.8 genomes are represented. Both long- (PacBio) and short-read (Illumina HiSeq 2000) metagenome sequencing data are available, with 5 Gb of sequences per sample, leading to an average genome coverage of 4.7× (ref. [21]). The runtime to generate these data was approximately 3 weeks using eight CPU cores of a computer with an AMD Opteron 6378 CPU and 968 GB of main memory.

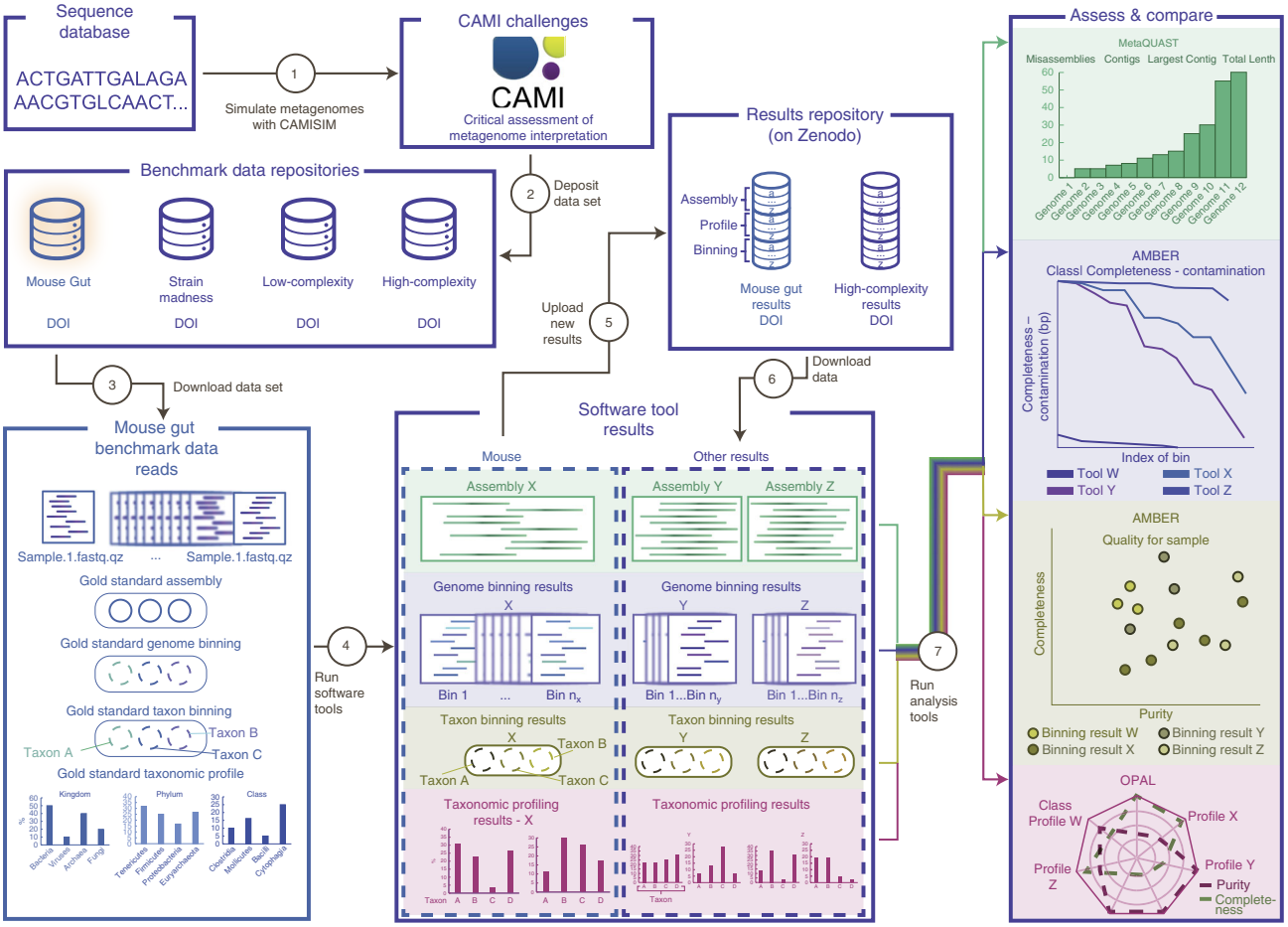


Fig. 1 | CAMI benchmarking workflow. The initial step is (1) the simulation of metagenome data from a sequence database with CAMISIM[21], which includes the microbial community design and generation of standards of truth. (2) The simulated metagenome data are stored in benchmark data repositories with digital object identifiers (DOIs) or temporarily without DOIs for ongoing CAMI challenges, as the standards of truth are provided only after the challenges. (3) The data can then be downloaded and (4) software tools such as metagenome assemblers, genome and taxonomic binners, and profilers can be run on the data. This leads to the creation of a pool of software tool results. (5) These results can be submitted to an ongoing challenge or uploaded to a public repository such as Zenodo. (6) Already-existing results can be downloaded and (7) integrated along with newly generated results in benchmark analyses with MetaQUAST[23], AMBER[24], and OPAL[25].

CAMISIM can be installed according to the instructions at https://github.com/CAMI-challenge/CAMISIM/ or by using Docker with the following command:

```
docker pull cami/camisim
```

To generate the mouse gut dataset, use the following command: `./metagenome_from_profile -p profile.biom -o out/`

`profile.biom` is a BIOM[33] file storing the microbial community genome abundance distribution for the 64 samples. It can be obtained together with the dataset (Table 2). Per default, CAMISIM simulates 5 Gb of sequences per sample.

If CAMI benchmark data generated with CAMISIM have been downloaded, the following files and folders should appear:

- One folder per sample
  - Reads (anonymized and shuffled) as FASTQ files
  - Contigs (gold-standard assembly) as FASTA files
  - Gold-standard mappings (binning) in BAM and CAMI formats (see format specifications at https://github.com/CAMI-challenge/file_formats)
- For multisample simulations:
  - File containing contigs (gold-standard assembly) as FASTA files
  - File containing gold-standard mappings (binning and profiling) in CAMI format
- Profiling gold standard per sample in CAMI format
- One folder (called 'source genomes') containing all reference genome sequences as FASTA files
- One folder (called 'distributions') containing files with the absolute abundances per genome for each sampled microbial community
- One folder (called 'internal') containing the input metadata and a list of unused genomes
- Metadata (CAMISIM.ini config file)

Assembly

Cross-sample co-assemblies of the first 10 of 64 metagenome samples were performed with MEGAHIT[34] v.1.0.3, v.1.1.3 (with

Worst Median Best

Genome statistics	MEGAHIT 1.0.3 df	MEGAHIT 1.1.3 df	MEGAHIT 1.1.3 ml	MEGAHIT 1.1.3 ms	MEGAHIT 1.2.9 df	metaSPAdes 3.13.0
Genome fraction (%)	23.507	26.164	26.039	26.292	26.691	23.262
Duplication ratio	1.023	1.037	1.046	1.05	1.034	1.017
Largest alignment	354703	904953	859640	753008	787657	1034619
Total aligned length	436725459	492514960	493969107	500306789	500856984	429280747
NGA50	...	...	...	...	...	...
LGA50	...	...	...	...	...	...
Misassemblies						
# misassemblies	5770	8685	5336	9381	8807	3488
Misassembled contigslength	10879967	43068359	34576388	56221107	50536067	25409676
Mismatches						
# mismatches per 100 kbp	542.07	580.14	887.27	945.71	585.26	405.65
# indels per 100 kbp	2.39	4.17	3.92	4.75	4.3	2.57
# N's per 100 kbp	0	0	0	0	0	0
Statistics without reference						
# contigs	225585	220757	278807	282136	225167	174693
Largest contig	354703	904953	859640	754056	788697	1034619
Total length	438032656	494653238	496722592	503491159	503073431	430847014
Total length (>= 1,000 bp)	342669622	399682035	368806791	372764886	405211262	354794894
Total length (>= 10,000 bp)	154362921	228640882	192818790	198110070	236255195	225930387
Total length (>= 50,000 bp)	38821616	102990325	82724532	83865010	106551070	119684054

Fig. 2 | MetaQUAST assembly benchmarking metrics. Genome fraction is the total number of aligned bases in the reference, divided by the genome size; # contigs is the number of contigs in the assembly; NG50 is the contig length, such that contigs of that length or longer cover half (50%) of the bases of the reference genome; NGA50 is NG50 such that the lengths of aligned blocks are counted instead of contig lengths; and LGA50 is the minimal number of alignment blocks covering half of the bases of the reference genome. NG50, NGA50 and LGA50 are shown per genome in the Supplementary Results.

default, meta-sensitive, and meta-large settings), and v.1.2.9, as well as metaSPAdes[35] v.3.13.0, as the computer main memory was insufficient to run metaSPAdes on more than 10 samples. The choice of the first 10 samples was analogous to the CAMI II challenge specifications. All results and commands used are available on Zenodo (Supplementary Table 1). The computer specifications, memory usage, and runtimes are available in Supplementary Tables 2 and 3.

Assemblies were evaluated by mapping them against the gold-standard assembly, defined as the fraction of the genome covered by at least one read in the set of analyzed samples, using MetaQUAST[23] v.5.0.2. The gold-standard genomes are known through the simulation with CAMISIM and provided to MetaQUAST for the evaluation. In the case that the underlying genomes are unknown, such as when assessing de novo assemblies from less studied environments, reference-free methods[36–38] can be considered.

MetaQUAST can be installed with Bioconda using the following command:

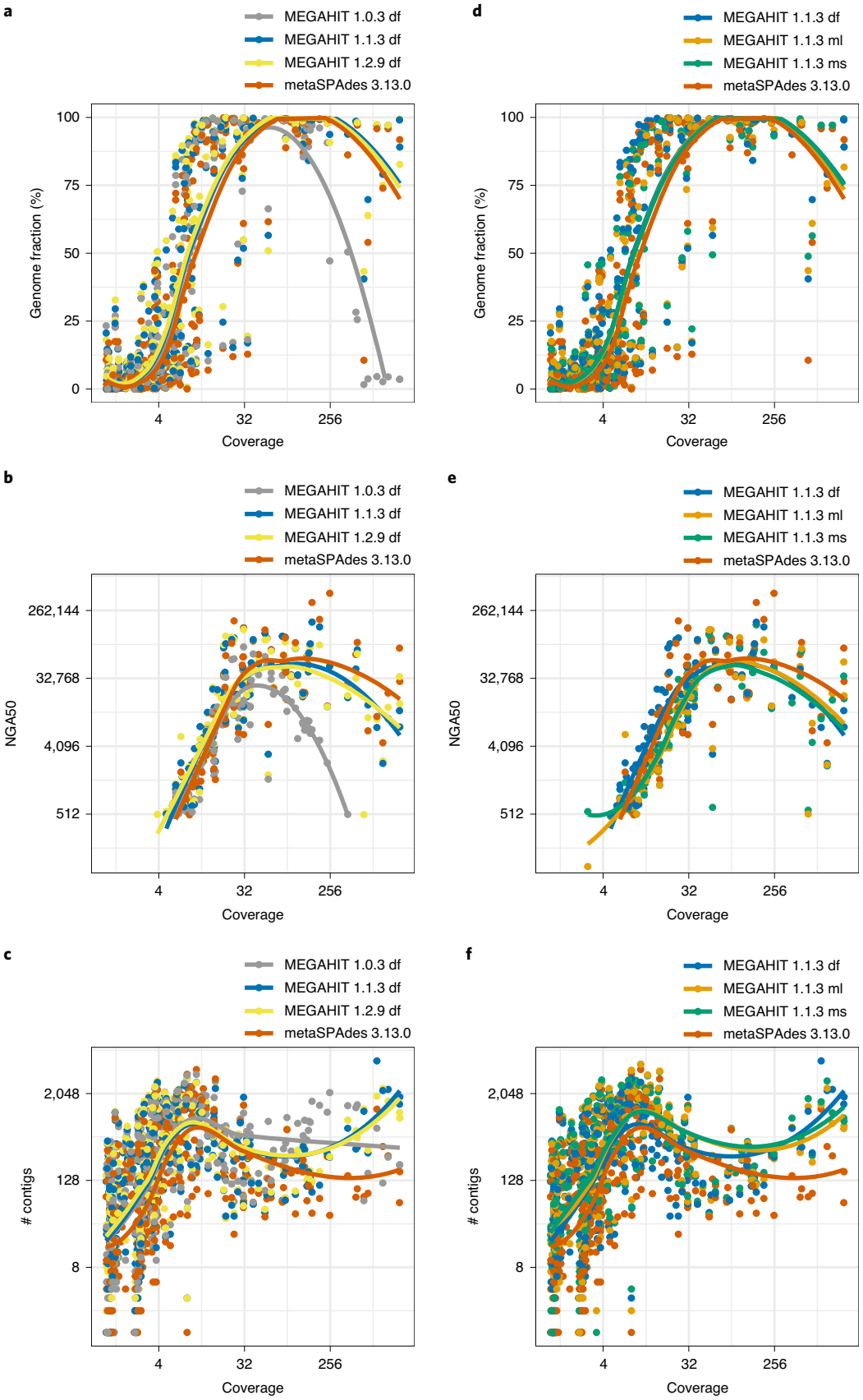
```
conda create --name quast quast
```

This requires Conda to be installed and the Bioconda channel configured; see https://bioconda.github.io/user/install.html for details. Other installation methods are described in the MetaQUAST GitHub repository at https://github.com/ablab/quast/. To run MetaQUAST, type:

```
conda activate quast
metaquast -r /path/to/set0-9/ref-genomes \
-t 24 --unique-mapping --no-icarus -o /path/
to/output_dir \
-l megahit-103-df,megahit-113-df,megahit-
113-ml,\
megahit-113-ms,megahit-129-df,metaSPAdes \
/path/to/megahit103-Sample0-9-default/
final.contigs.fa \
/path/to/megahit113-Sample0-9-default/
final.contigs.fa \
/path/to/megahit113-Sample0-9-meta-large/
final.contigs.fa \
/path/to/megahit113-Sample0-9-meta-
sensitive/final.contigs.fa \
/path/to/megahit129-Sample0-9-default/
final.contigs.fa \
/path/to/metaSPAdes3130-Sample0-9/contigs.fasta
```

For evaluating assembly quality, we rely on the metrics provided by MetaQUAST. Figure 2 shows the metrics we focus on here, whereas Supplementary Results shows all metrics computed by MetaQUAST. Performance values are calculated for the whole assembly versus the combined reference (i.e. concatenation of all provided references).

Overall, the performance of the MEGAHIT and MetaSPAdes assemblers is quite similar. MEGAHIT v.1.0.3 shows poor performance for high coverage (i.e. high abundance) genomes. This effect has been described for earlier versions of MEGAHIT before[8]. The more recent versions of MEGAHIT (1.1.3 and 1.2.9) handle high coverage genomes much better and show similar performance to MetaSPAdes. For coverages of 16× and above, the fraction of the recovered genomes is above 75% with some outliers for coverage higher than 250x. The NGA50 metric shows similar performance for MEGAHIT and metaSPAdes, reaching 32 kb and more for coverage of 32x and above (Fig. 3a–c). MetaSPAdes delivers fewer fragmented assemblies (fewer contigs and higher NGA50, Fig. 3d,e) than the newer MEGAHIT versions with only slightly lower genome fraction (Fig. 3d).


a
MEGAHIT 1.0.3 df
MEGAHIT 1.1.3 df
MEGAHIT 1.2.9 df
metaSPAdes 3.13.0
Genome fraction (%)
100
75
50
25
0
4
32
256
Coverage
d
MEGAHIT 1.1.3 df
MEGAHIT 1.1.3 ml
MEGAHIT 1.1.3 ms
metaSPAdes 3.13.0
Genome fraction (%)
Coverage
b
NGA50
262,144
32,768
4,096
512
e
c
contigs
2,048
128
8
f

◀ **Fig. 3 | Assessing metagenome cross-sample assembly quality with MetaQUAST for the CAMI II mouse gut dataset. a–c**, Genome-wide MetaQUAST metrics (genome fraction (**a**), NGA50 (**b**), # contigs (**c**)) for assemblies generated with MEGAHIT versions 1.0.3, 1.1.3, 1.2.9 and metaSPAdes 3.13.0 versus sum of read coverages for individual genomes (dots) in ten cross-sample gold-standard assemblies. The higher the genome fraction and NGA50, the better the assembly quality. Higher values of # contigs can indicate a higher amount of assembled data but also more fragmented assemblies, whereas lower values of # contigs can indicate aggressive traversal of repeats by an assembler, leading to incorrect junctions of sequence fragments and thus misassemblies. **d–f**, MetaQUAST metrics (genome fraction (**d**), NGA50 (**e**), # contigs (**f**)) for assemblies generated with different settings of MEGAHIT 1.1.3 (default (df), meta-sensitive (ms) and meta-large (ml)) and metaSPAdes 3.13.0. All lines are fitted with local regression using the R stats::LOESS (locally estimated scatterplot smoothing) function.

When assessing different settings for MEGAHIT v.1.1.3 (Fig. 3d–f), smaller, but notable differences were found. For instance, the settings meta-sensitive (ms) and meta-large (ml) delivered higher genome fractions for low coverage genomes, at the cost of higher genome fragmentation rates (decreased NGA50 and more contigs).

Genome binning

Genome binning can be seen as a clustering problem, where sequences are grouped into bins without taxon labels. We reconstructed genome bins from the cross-sample gold-standard assembly with the popular binners MaxBin v.2.2.7 (ref. [39]), MetaBAT v.2.12.1 (ref. [40]), CONCOCT v.1.0.0 (ref. [41]), and DAS Tool v.1.1.2 (ref. [42]). DAS Tool combines the genome bins of individual methods to further improve bin quality. All results and commands used are available on Zenodo (Supplementary Table 4). Runtimes and memory usage are provided in Supplementary Table 5. Binning quality was evaluated with AMBER v.2.0.2 (ref. [24]) (Table 3), which computes binning performance metrics for metagenome data with a ground truth available, i.e., for which the correct assignment of sequences to genome bins is known (see glossary in Table 1). The binning file format used in CAMI is described at https://github.com/CAMI-challenge/file_formats. To reproduce the evaluation, the binning results must first be downloaded from Zenodo, then AMBER installed using Bioconda, as follows:

```
conda create --name amber cami-amber
```

Other installation methods are described at https://github.com/CAMI-challenge/AMBER/. To run AMBER, type:

```
conda activate amber
amber.py --gold_standard_file /path/to/
cami2_mouse_gut_gsa_pooled.binning \
/path/to/cami2_mouse_gut_maxbin2.2.7.
binning \
/path/to/cami2_mouse_gut_metabat2.12.1.
binning \
/path/to/cami2_mouse_gut_concoct1.0.0.
binning \
/path/to/cami2_mouse_gut_dastool1.1.2.
binning \
--labels "MaxBin 2.2.7, MetaBAT 2.12.1,
CONCOCT 1.0.0, DAS Tool 1.1.2" \
--genome_coverage /path/to/cami2_mouse_
gut_average_genome_coverage.tsv \
--output_dir /path/to/output_dir
```

The file `cami2_mouse_gut_average_genome_coverage.tsv` above contains the average coverage of the genomes in the CAMI II mouse gut dataset and is also available on Zenodo (Supplementary Table 4). This file is optional and is used by AMBER to generate performance plots relative to the average genome coverage (Fig. 4a,b).

In the evaluation of genome binning, several metrics are often jointly assessed. For each genome, *completeness*, or recall, is evaluated from the predicted bin containing the largest number of base pairs (bp) of the genome. It is the number of base pairs (or contigs) in the genome in that bin divided by the genome size (in base pairs or contigs). Sequences of that genome assigned to other bins are considered false positives for those bins. Completeness can be zero, in the case that no part of a genome has been binned by the respective binner. *Purity* denotes how 'clean' predicted bins are in terms of their assigned content. It is computed as the fraction of contigs, or base pairs, coming from one genome, for the most abundant genome in that bin. *Contamination* is defined as 100% minus purity. As genomes can differ in their abundances, it is also common to consider sample-wise metrics, such as the overall *percentage of assigned base pairs* and the *adjusted Rand index* (ARI) on that assigned fraction. The ARI reflects the overall resolution of the underlying ground truth genomes by a binner on the binned part of the sample. The ARI gives more importance to 'large' bins, that is, bins of large and/or abundant genomes, than do completeness or purity, where each gold-standard genome (for completeness) and predicted bin (for purity) contributes the same, irrespective of its size. In the following, all evaluations are based on base pair counts.

Completeness was high for all methods and was highest for CONCOCT. Binners recovered the abundant genomes better, with average completeness >90% for genomes at more than threefold coverage (Fig. 4a). Purity was also high (Fig. 4b), except for CONCOCT, and was highest for MetaBAT, which was further improved by DAS Tool. Completeness was >90% for genome bins with an average of 3.5 Mb for most binners (Fig. 4c). CONCOCT and MetaBAT predicted bins that were larger than their true sizes. This can be seen in Fig. 4d, as the pink and green lines exceed the blue (gold-standard) line on the x axis. Purity was >90% for predicted genomes bins, with an average of 2.6 million to 3.5 million bp (Fig. 4d). Both purity and completeness were much lower for smaller and larger bins. CONCOCT assigned the most base pairs (Fig. 4e), although to fewer bins. Low purity and fewer bins indicate 'underbinning', that is, multiple genomes being placed together in one bin. The other extreme, 'overbinning', occurs when genomes are split across multiple bins, resulting in low

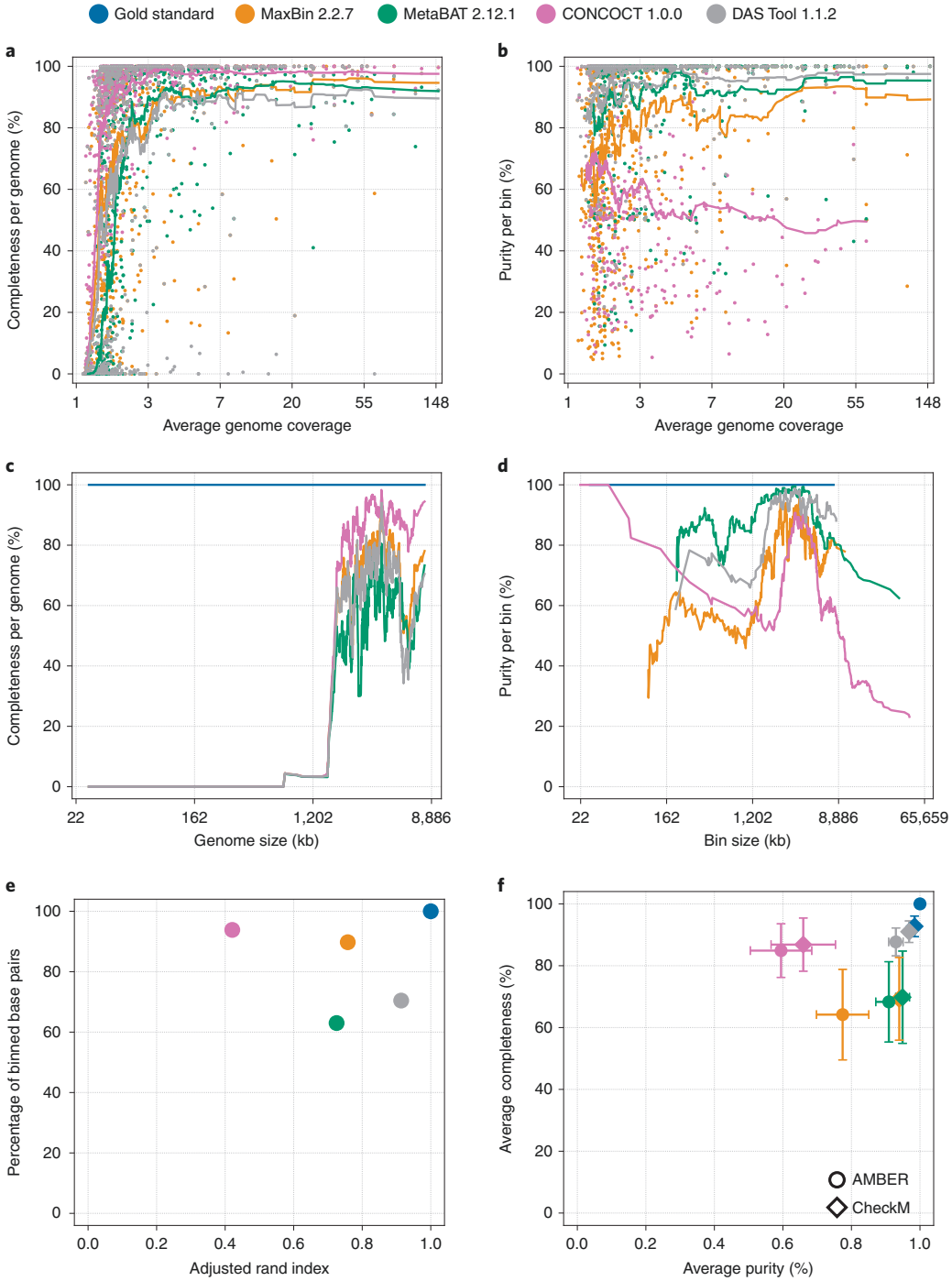


Fig. 4 | Assessing genome binners on the gold-standard assembly of the CAMI II mouse gut dataset. a, Average genome coverage (*x* axis) versus completeness per genome (*y* axis). **b**, Average genome coverage (*x* axis) versus purity per bin (*y* axis). The lines in **a** and **b** show the rolling average completeness or purity over 50 bins. **c**, Genome size in thousands of bp (*x* axis) versus completeness per genome (*y* axis). **d**, Bin size in kilobase pairs (kb (*x* axis) versus purity per bin (*y* axis). **e**, Adjusted Rand index (*x* axis) versus percentage of assigned base pairs (*y* axis). **f**, Average purity (*x* axis) versus average completeness (*y* axis) of all predicted bins per method assessed with AMBER (circles) and CheckM (diamonds), with the whiskers showing the variance. For all metrics, except genome and bin sizes, the range is between 0% (worst) and 100% (best).

completeness. After DAS Tool, MaxBin predictions had the highest ARI, followed by MetaBAT. DAS Tool substantially improved bin purity and ARI relative to the individual methods—at the cost of completeness and binning a lower percentage of base pairs than CONCOCT or MaxBin. MaxBin and DAS Tool recovered the most high-quality genomes, defined as genomes with >50% completeness and <10% contamination (Table 4). The total number of predicted bins per method was 867 (MaxBin), 592 (MetaBAT), 344 (CONCOCT), and 577 (DAS Tool).

Table 4 | Number of high-quality genomes and corresponding percentages recovered by genome binners from the gold-standard assembly of the CAMI II mouse gut dataset

Genome binner	% Contamination	Predicted bins (% completeness)		
		>50%	>70%	>90%
Gold standard		791 (100)	791 (100)	791 (100)
MaxBin 2.2.7	<10	**439 (55)**	**419 (53)**	**342 (43)**
	<5	**401 (51)**	**386 (49)**	**319 (40)**
MetaBAT 2.12.1	<10	353 (45)	318 (40)	240 (30)
	<5	339 (43)	309 (39)	236 (30)
CONCOCT 1.0.0	<10	95 (12)	95 (12)	84 (11)
	<5	88 (11)	88 (11)	79 (10)
DAS Tool 1.1.2 (ensemble method)	<10	**460 (58)**	**449 (57%)**	**354 (45)**
	<5	**422 (53)**	**416 (53)**	**334 (45)**

The best result per bin completeness of an individual (non-ensemble) method and among all methods is indicated in bold. In the gold standard, all 791 bins have 100% completeness and 0% contamination.

We compared the bin quality metrics obtained with AMBER with those returned by the commonly used CheckM software v.1.1.2, which assesses bin quality on the basis of the presence of lineage-specific marker genes[43] (Fig. 4f, Supplementary Note). The results were largely consistent. CheckM overestimated purity by 4% (MetaBAT and DAS Tool) to 21% (MaxBin) and completeness by 2% (MetaBAT and CONCOCT) to 7% (MaxBin) (Fig. 4f, Supplementary Tables 6 and 7). Because of CheckM's known bias of overestimating completeness and underestimating contamination[43], we also computed the averages of only those bins with >90% completeness and <10% contamination according to AMBER's assessment. In this case, CheckM's purity overestimates dropped to only up to 3% for all methods except CONCOCT, for which it increased to 29%. On the other hand, completeness was underestimated for most methods, by 9% (CONCOCT) to 17% (MaxBin).

Taxonomic binning

A taxon bin is a set of sequences, either contigs or reads, with the same taxonomic label. Taxonomic binning can be evaluated as a multiclass classification problem at individual taxonomic ranks, where one of many possible taxon labels from a reference taxonomy is assigned to each metagenomic sequence. The quality of a taxon binning is assessed by comparing predicted and ground truth taxon bins with each other.

We predicted taxon bins from the cross-sample gold-standard assembly with DIAMOND v.0.9.24 (ref. [44]), Kraken v.2.0.8 beta (ref. [45]), PhyloPythiaS+ v.1.4 (ref. [46]), CAT v.4.6 (ref. [47]), and MEGAN v.6.15.2 (ref. [48]). All results and commands used are available on Zenodo (Supplementary Table 8). Runtimes and memory usage are given in Supplementary Table 9. The release date of the NCBI taxonomy used by each method is indicated on Zenodo and can vary slightly, depending on the reference database of the method. Method performances were assessed with AMBER v.2.0.2, for all major taxonomic ranks (Figs. 5 and 6), using the NCBI taxonomy database from 2018/02/26. This reference taxonomy is provided with the mouse gut dataset of the CAMI II challenge (Table 2). To run AMBER, type the following command:

```
amber.py --gold_standard_file /path/to/
cami2_mouse_gut_gsa_pooled.binning \
--desc "CAMI 2 toy mouse gut data set" \
/path/to/cami2_mouse_gut_diamond0.9.24.
binning \
/path/to/cami2_mouse_gut_kraken2.0.8beta.
binning \
/path/to/cami2_mouse_gut_ppsp1.4.binning \
/path/to/cami2_mouse_gut_cat4.6.binning \
/path/to/cami2_mouse_gut_megan6.15.2.
binning \
--labels "DIAMOND 0.9.24, Kraken 2.0.8 beta,
PhyloPythiaS+ 1.4, CAT 4.6, MEGAN 6.15.2" \
--ncbi_nodes_file /path/to/nodes.dmp \
--ncbi_names_file /path/to/names.dmp \
--ncbi_merged_file /path/to/merged.dmp \
--filter 1 \
--output_dir /path/to/output_dir
```

For comparing predicted taxon bins to the ground truth, completeness and purity can be calculated. The completeness, or recall, for a taxon bin found in the ground truth is the fraction of ground truth contigs, or base pairs, that have been assigned to that taxon by a method. Completeness is averaged over all ground truth taxon bins at a particular rank and undefined for predicted taxon bins not present in the ground truth. The purity of a predicted taxon bin is the fraction of contigs, or base pairs, belonging to that taxon in the ground truth. Taxon bins without any correctly assigned sequences accordingly have a purity of zero. Purity is averaged over all predicted taxon bins at a particular rank. Contamination is defined as 100% minus purity. Finally, the *accuracy* is the fraction of contigs, or base pairs, that have been assigned by a method to the correct taxa for a taxonomic rank. Accuracy is a

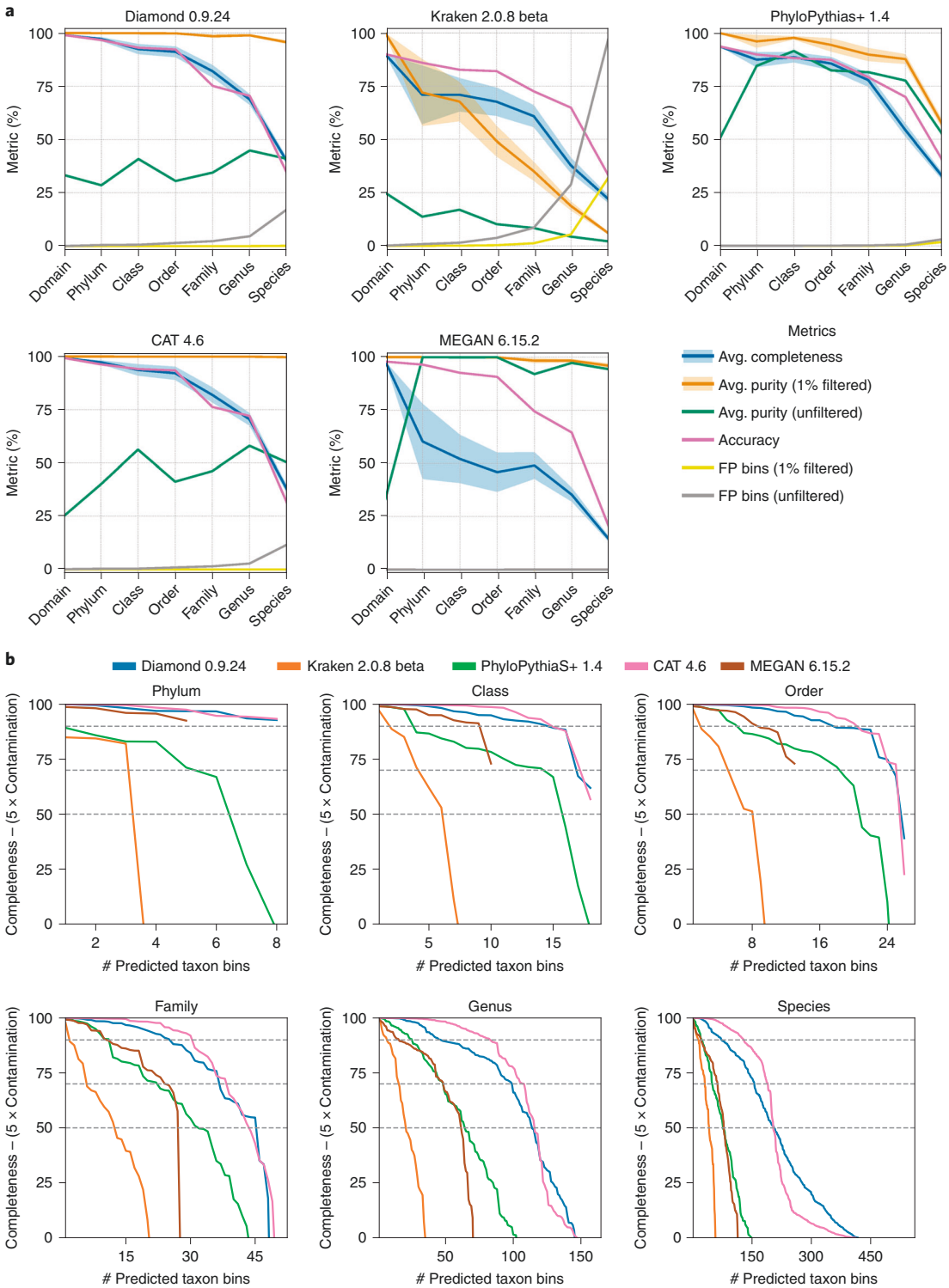


Fig. 5 | Assessing taxonomic binning results on the CAMI II mouse gut dataset. a, Average completeness and purity (1% filtered and unfiltered; see text), accuracy, and percentage of false-positive bins (number of bins with zero precision, normalized by the maximum bin number among all binners and ranks) per taxonomic rank for each binner. The shaded bands show the standard error of the metrics. **b**, Score (i.e., Completeness – (5 × Contamination); *y* axis) and number of predicted taxon bins (*x* axis) for the phylum to species ranks. The higher the number of high-scoring bins, the better is the binning performance. Only positive scores are shown. The dotted lines indicate the 90, 70, and 50 score thresholds. FP, false-positive.

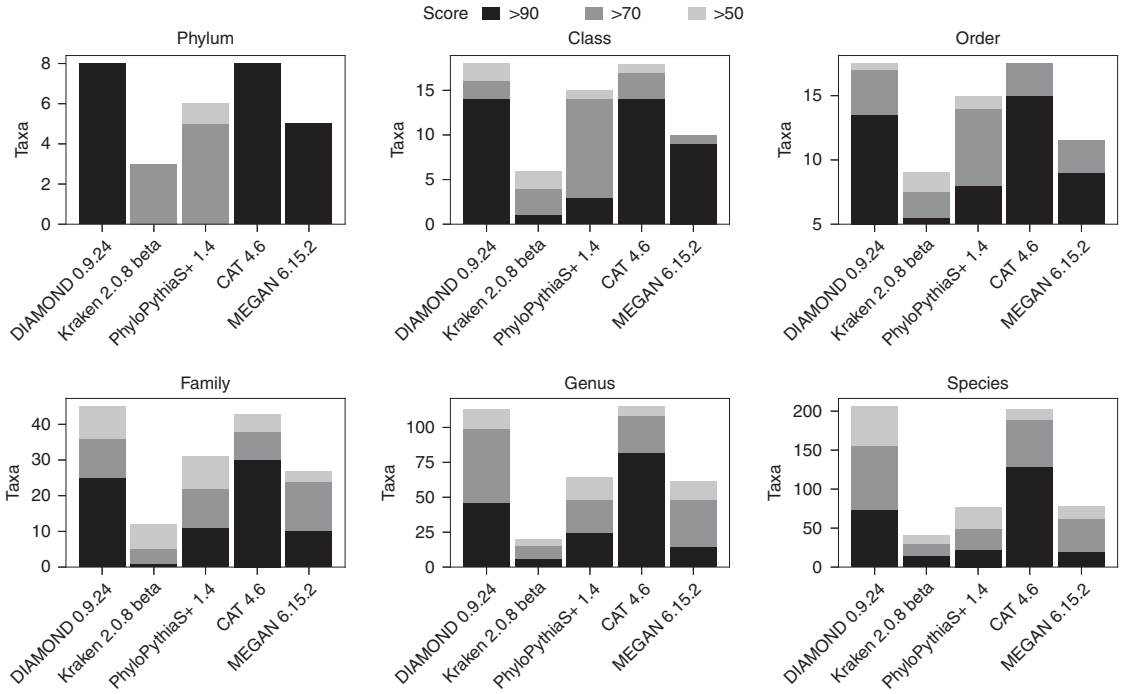


Fig. 6 | Number of high-quality taxon bins predicted from the CAMI II mouse gut dataset for the phylum to species ranks. Counted are the bins with score (i.e., Completeness – (5 × contamination)) higher than 90, 70, and 50. A number of bins closer to the number of taxa per rank in the gold standard (i.e., 8 phyla, 18 classes, 26 orders, 50 families, 157 genera, and 549 species) is better.

sample-specific metric to which larger taxon bins contribute more strongly than small ones, which is different from average completeness and purity.

DIAMOND and CAT, which relies on DIAMOND's output, obtained the highest average completeness for all ranks. This was >90% from superkingdom to order and continuously dropped at lower ranks (Fig. 5a). MEGAN, which also uses DIAMOND, achieved lower completeness for phylum level and below, but had the highest average purity at all ranks, except for superkingdom, at which PhyloPythiaS+ performed best. As purity can be reduced for small bins, we filtered out the smallest predicted bins per method and rank, removing overall 1% of the binned data in base pairs. This can be done with AMBER (using the `--filter 1` option) on the predicted bins, requiring no knowledge of the underlying gold standard. Across all ranks, the average size of the removed taxon bins was 1.9 Mb, whereas the average size of all bins was 235.8 Mb (Supplementary Table 10), with larger bins accumulating at higher ranks. DIAMOND and CAT profited most from this, with CAT reaching almost 100% filtered purity at all ranks. Researchers interested in taxa with small genomes, such as viruses, should keep in mind that filtering could remove these along with false-positive bins. Purity and completeness were also influenced by contig length and were higher overall for longer contigs (Supplementary Fig. 1). In terms of accuracy, all methods performed similarly well, with PhyloPythiaS+ being the most accurate at the species level.

Based on a quality score defined as completeness – (5 × contamination), as in refs. [7,49], we determined the number of high-quality bins found by each method with a score of >90, >70, and >50 at different taxonomic ranks (Fig. 6). DIAMOND, CAT, and PhyloPythiaS+, in this order, identified the most high-quality bins (>50) at all taxonomic ranks. CAT, followed by DIAMOND, found the most bins with a score >90.

Taxonomic profiling

Taxonomic profiling can be considered a multilabel problem at a given rank, where multiple taxon labels are assigned to a single sample and the relative taxon abundances are estimated. Profiling differs from binning in that individual reads are not necessarily assigned taxon labels. We predicted taxonomic identities and relative abundances of microbial community members for the 64 short-read samples of the mouse gut dataset with MetaPhlAn v.2.9.21 (ref. [50]), mOTUs v.2.5.1 (ref. [51]), and Bracken v.2.5 (ref. [52]). We assessed these together with results for MetaPhlAn v.2.2.0, mOTUs v.1.1, MetaPalette v.1.0.0, MetaPhyler v.1.25, FOCUS v.0.31, TIPP v.2.0.0, and CAMIARKQuikr v.1.0.0 from ref. [25]. The profiling results and commands used can be obtained from Zenodo (Supplementary Table 11). Runtimes and memory usage are given in Supplementary Table 12. Performance metrics and results visualizations were calculated with OPAL v.1.0.9 (ref. [25]) (Table 3), which uses the CAMI file format described at https://github.com/CAMI-challenge/file_formats. It can be installed with the following command, if Bioconda is configured:

```
conda create --name opal cami-opal
```

Other installation methods are described in the OPAL GitHub repository at https://github.com/CAMI-challenge/OPAL/. We then ran OPAL as:

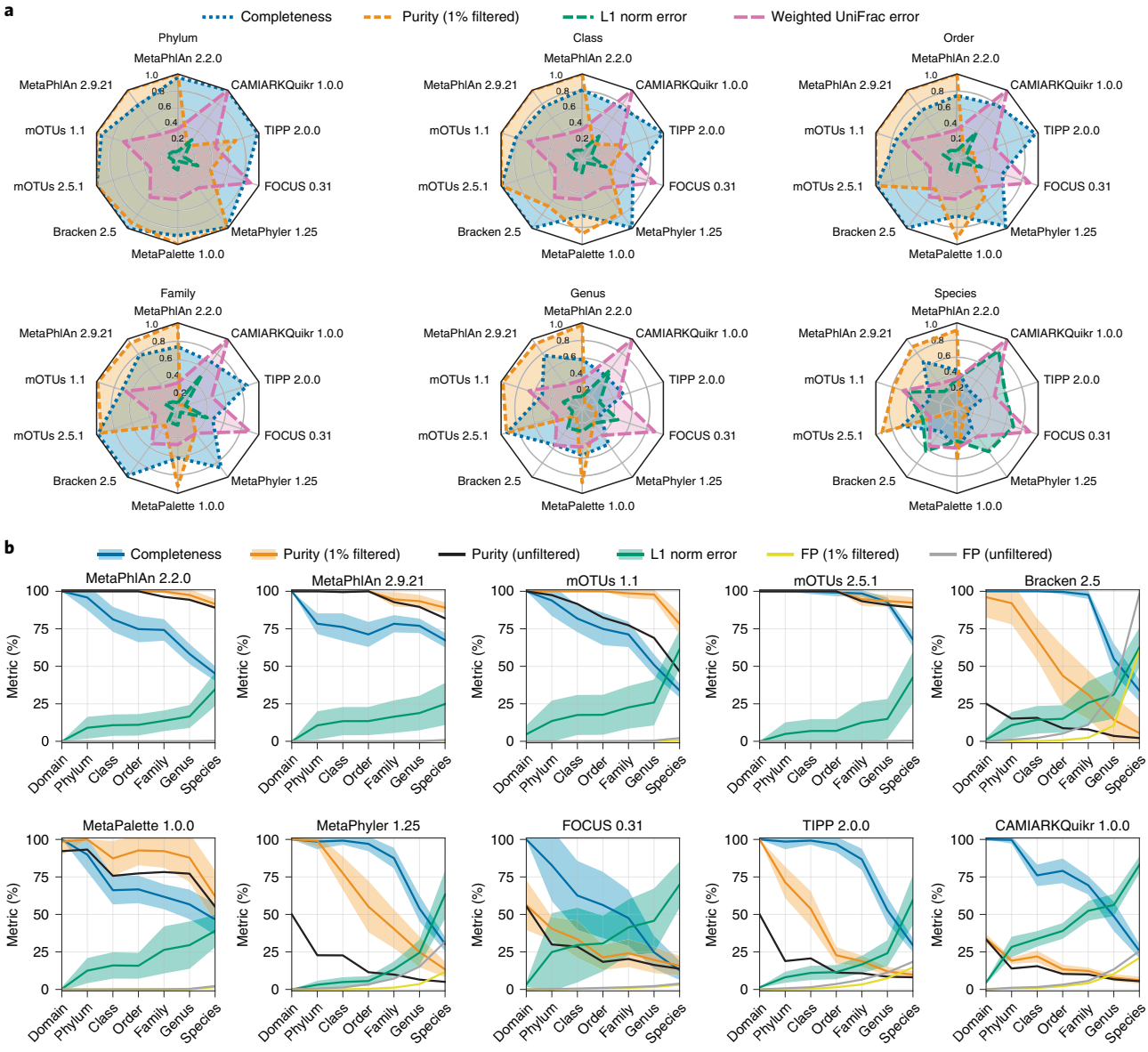
```
conda activate opal
opal.py --gold_standard_file /path/to/
cami2_mouse_gut_gs.profile \
/path/to/cami2_mouse_gut_metaphlan2.2.0.
profile \
/path/to/cami2_mouse_gut_metaphlan2.9.21.
profile \
/path/to/cami2_mouse_gut_motus1.1.profile \
/path/to/cami2_mouse_gut_motus2.5.1.profile \
/path/to/cami2_mouse_gut_bracken2.5.profile \
/path/to/cami2_mouse_gut_metapalette1.0.0.
profile \
/path/to/cami2_mouse_gut_metaphyler1.25.
profile \
/path/to/cami2_mouse_gut_focus0.31.profile \
/path/to/cami2_mouse_gut_tipp2.0.0.profile \
/path/to/cami2_mouse_gut_camiarkquikr1.0.0.
profile \
--labels "MetaPhlAn 2.2.0, MetaPhlAn 2.9.21,
mOTUs 1.1, mOTUs 2.5.1, Bracken 2.5, MetaPal-
ette 1.0.0, MetaPhyler 1.25, FOCUS 0.31, TIPP
2.0.0, CAMIARKQuikr 1.0.0" \
-d "2nd CAMI Challenge Mouse Gut Toy Dataset" \
--filter 1 \
--output_dir /path/to/output_dir
```

OPAL computes performance metrics and creates visualizations for profiling results on a benchmark dataset. It also generates weighted summary scores for ranking methods based on these metrics (see ref. [25] for a complete overview and formal definitions). For a taxonomic rank, the purity and completeness assess how well a profiler identified the presence and absence of taxa, without considering relative abundances. Purity, or precision, denotes the ratio of correctly predicted taxa to all taxa predicted at a taxonomic rank, whereas completeness, or recall, is the ratio of correctly identified taxa to all ground truth taxa at a taxonomic rank. A commonly used, heuristic approach designed to increase purity is to filter out low-abundance predictions on the basis of some threshold. To explore the effect of such heuristic post-processing of predictions on purity, we filtered low-abundance taxon predictions as we did for taxonomic binners (ref. [8]): by removing predictions with the lowest relative abundances, summing up to 1% of the total predicted organismal abundances per taxonomic rank.

For quantifying relative abundance estimates, the *L1 norm* and *weighted UniFrac error* are determined. The L1 norm assesses relative abundance estimates of taxa at a taxonomic rank, on the basis of the sum of the absolute differences between the true and predicted abundances across all taxa. The weighted UniFrac error computed by OPAL uses a taxonomic tree storing the predicted abundances at the appropriate nodes for eight major taxonomic ranks. The UniFrac error is the total amount of predicted abundances that must be moved along the edges of the tree to cause them to overlap with the true relative abundances. Branch lengths in the taxonomic tree can be set to 1 or to any function of the depth of the edge in the taxonomic tree. This choice is motivated by the fact that harmonizing phylogenetic trees (which express evolutionary distance with branch lengths) and taxonomic trees (which do not inherently have branch length information) remains an open problem under active investigation[53–56]. A low UniFrac error indicates good accuracy of abundance estimates. L1 norm and weighted UniFrac error are computed using unnormalized relative abundances; that is, their sum may be <1 if some data remain taxonomically unassigned. Normalization (optional in OPAL) can simplify the comparison of the L1 norm between methods (https://github.com/CAMI-challenge/firstchallenge_evaluation/tree/master/profiling); however, it may skew results for profilers with low recall that left many taxa unassigned. Assessment results with normalized relative abundance estimates are available in the OPAL GitHub repository at https://cami-challenge.github.io/OPAL/cami_ii_mg_filter1_normalized.

Using all these metrics, OPAL ranks the assessed profilers by their relative performance. For each metric, sample, and major taxonomic rank (from superkingdom to species), the best-performing profiler is assigned score 0; the second best, 1; and so on. These scores are then added over the taxonomic ranks and samples to produce a single score per metric for each profiler. OPAL can also assign different weights to the metrics, such that the importance of a metric, defined by the user, is reflected in the overall score and rank of a profiler. In our assessment, all metrics were weighted equally.

mOTUs v.2.5.1, Bracken v.2.5, MetaPhyler v.1.25, and TIPP v.2.0.0, in this order, achieved the overall highest completeness (Fig. 7c). mOTUs v.2.5.1 achieved high completeness up to genus level, whereas the other profilers performed well with this metric up to family level (Fig. 7a,b). Along with completeness, purity also drops for lower taxonomic ranks. Filtering low-abundant taxon predictions greatly improved purity, most strongly for MetaPhyler and Bracken v.2.5, which was ranked seventh instead of last with this metric. MetaPhlAn v.2.2.0 and mOTUs v.1.1 had the highest filtered purity across ranks, followed by mOTUs v.2.5.1 and MetaPhlAn v.2.9.21. mOTUs v.2.5.1 showed both high (filtered and unfiltered) purity and completeness and improved considerably in terms of completeness compared with its previous version. mOTUs v.2.5.1, MetaPhlAn v.2.9.21, MetaPhyler v.1.25, and MetaPhlAn v.2.2.0, in this order, best estimated the relative abundances measured with the L1 norm, with MetaPhlAn v.2.9.21 outperforming all methods at the species level. MetaPhlAn v.2.9.21also obtained the lowest UniFrac error, followed by mOTUs v.2.5.1 and MetaPhlAn v.2.2.0. Considering the sum of scores of all metrics, mOTUs v.2.5.1 ranked first, followed by MetaPhlAn v.2.2.0 and v.2.9.21. Notably, normalization of abundance estimates had almost no effect on the L1 norm error of the methods (Supplementary Fig. 2), as the estimates covered almost 100% of the data (Supplementary Table 13). This may differ for metagenome data with many taxa being distant from those found in genome sequence repositories.



c

Ranking permetric	Software (score)				
	Completeness	Purity (1% filtered)	L1 norm error	Weighted UniFrac error	Sum of scores
1st	mOTUs 2.5.1 (62)	MetaPhlAn 2.2.0 (106)	mOTUs 2.5.1 (652)	MetaPhlAn 2.9.21 (104)	mOTUs 2.5.1 (1,129)
2nd	Bracken 2.5 (608)	mOTUs 1.1 (277)	MetaPhlAn 2.9.21 (1,140)	mOTUs 2.5.1 (105)	MetaPhlAn 2.2.0 (2,860)
3rd	MetaPhyler 1.25 (846)	mOTUs 2.5.1 (310)	MetaPhyler 1.25 (1,244)	MetaPhlAn 2.2.0 (130)	MetaPhlAn 2.9.21 (3,191)
4th	TIPP 2.0.0 (969)	MetaPhlAn 2.9.21 (400)	MetaPhlAn 2.2.0 (1,246)	MetaPhyler 1.25 (188)	MetaPhyler 1.25 (3,979)
5th	MetaPhlAn 2.2.0 (1,378)	MetaPalette 1.0.0 (795)	TIPP 2.0.0 (1,991)	MetaPalette 1.0.0 (274)	mOTUs 1.1 (4,732)
6th	MetaPhlAn 2.9.21 (1,547)	MetaPhyler 1.25 (1,701)	MetaPalette 1.0.0 (2,051)	TIPP 2.0.0 (275)	MetaPalette 1.0.0 (5,103)
7th	mOTUs 1.1 (1,736)	Bracken 2.5 (2,292)	Bracken 2.5 (2,291)	Bracken 2.5 (319)	Bracken 2.5 (5,510)
8th	CAMIARKQuikr 1.0.0 (1,823)	TIPP 2.0.0 (2,728)	mOTUs 1.1 (2,294)	mOTUs 1.1 (425)	TIPP 2.0.0 (5,963)
9th	MetaPalette 1.0.0 (1,983)	FOCUS 0.31 (3,134)	FOCUS 0.31 (3,265)	FOCUS 0.31 (497)	FOCUS 0.31 (9,785)
10th	FOCUS 0.31 (2,889)	CAMIARKQuikr 1.0.0 (3,920)	CAMIARKQuikr 1.0.0 (3,858)	CAMIARKQuikr 1.0.0 (563)	CAMIARKQuikr 1.0.0 (10,164)

Fig. 7 | Assessing taxonomic profiling results on the CAMI II mouse gut dataset. a, Comparison per taxonomic rank of methods in terms of completeness, purity (1% filtered; see main text), L1 norm error, and weighted UniFrac error. **b**, Performance per method at all major taxonomic ranks, with the shaded bands showing the standard deviation of a metric, and percentage of false-positive taxa (normalized by the maximum taxon number among all profilers and ranks). In **a** and **b**, completeness and purity range between 0 and 1. The L1 norm error is normalized to this range, and the weighted UniFrac error is rank independent and normalized by the maximum value obtained by the profilers. The higher the completeness and purity, and the lower the L1 norm and weighted UniFrac error, the better the profiling performance. **c**, Method rankings and scores obtained for the different metrics over all samples and taxonomic ranks. For score calculation, all metrics were weighted equally. FP, false-positive.

We note that performance estimates may differ strongly, depending on metric definitions. For instance, contrary to the results reported here, mOTUs and MetaPhlAN performed poorly in terms of the fraction of sample reads that they classified[19], which is a task that they were not designed for.

Summary and conclusions

Microbiome research using meta-omics technologies is a rapidly progressing field producing highly complex and heterogeneous data. For developing and assessing data processing techniques, adoption of benchmarking standards in the field is essential. We here outlined key elements of benchmarking and best practices developed by a larger group of scientists within CAMI for common computational analyses in metagenomics. Community-driven benchmarking challenges are a key component of unbiased performance evaluations, in addition to the assessments by individual developers that are commonly done. To facilitate the latter, we describe a benchmarking tool resource and the mechanisms to use and add to this resource, as indicated in ref. [8], in a flexible way. We show how to apply the best practices for benchmarking defined by the community within CAMI using the CAMI benchmarking toolkit and benchmark datasets. For profiling methods, we demonstrated the value of incremental benchmarking by reusing and combining tool results from different studies and saving these in the CAMI tool result repositories on Zenodo (https://zenodo.org/communities/cami). Curated metadata and instructions on how to contribute reproducible results are provided at https://github.com/CAMI-challenge/data. As these new resources grow, individual benchmarks of meta-omics software will become increasingly more efficient, informative, and reproducible.

Using the 64-sample simulated metagenome dataset from mouse guts as an example, we performed a comparative evaluation of metagenome assembly (for the first 10 samples), genome binning, and taxonomic binning and profiling on these data. Overall, the evaluation included 25 results for 19 computational methods: 2 assemblers, with 6 different settings and versions evaluated, 4 genome and 5 taxon binners, as well as 8 profilers, including 2 different versions. Seven of the profiling results originate from a previous evaluation study on the data, demonstrating the value of incremental data analysis. Notably, as the dataset was generated from genomes included in public databases, the results for reference-based methods, such as taxonomic binning and profiling techniques, are to be taken as representative only for microbial community members represented by close relatives in public database content. This is only true for a fraction of most microbial communities, if not considering computationally reconstructed MAGs as a reference. Accordingly, for reference-based techniques, that is, taxonomic binners and profilers, results were consistent with prior studies on data generated from publicly available genomes[25] and less congruent with performances on benchmark data including genomes more distantly related to public database content[8]. Performance on species that are distantly related to those with genomes in public databases continues to be an important point to keep in mind when selecting the most suitable method for analysis.

With the CAMI benchmarking resources in place, we invite researchers to make full use of these for tackling the big challenges in the field[57]. These include developing strain-resolved assembly; binning and profiling techniques for strain-specific genome reconstructions[58,59]; making use of long-read metagenomic sequencing data[60]; evaluating methods for other meta-omics, for example, metatranscriptomics, metaproteomics[61]; and metametabolomics. The applications of metagenomics are diverse and growing, and the best way to tackle them is via a large collaborative framework supported by good collaborative infrastructure, which CAMI aims to provide.

Data availability

The results of all benchmarked methods and gold standards are available at https://zenodo.org/communities/cami. Links to individual results and DOIs are available in Supplementary Tables 1, 4, 8, and 11. The gold-standard assembly is provided with the CAMI II mouse gut dataset (Table 2). Assembly results and code used to generate Fig. 3 are available at https://github.com/CAMI-challenge/BenchmarkingToolkitTutorial. Genome and taxonomic binning, and taxonomic profiling results used in Figs. 4–7 are available, respectively, in the AMBER and OPAL GitHub repositories at https://github.com/CAMI-challenge/AMBER and https://github.com/CAMI-challenge/OPAL. The code in this paper has been peer-reviewed.

References

1. Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
2. Mitchell, A. L. et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res* **46**, D726–D735 (2018).
3. Chen, I.-M. A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **47**, D666–D677 (2019).
4. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
5. 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).
6. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
7. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
8. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
9. Bansal, V. & Boucher, C. Sequencing technologies and analyses: where have we been and where are we going? *iScience* **18**, 37–41 (2019).
10. Mantere, T., Kersten, S. & Hoischen, A. Long-read sequencing emerging in medical genetics. *Front. Genet.* **10**, 426 (2019).
11. Mosimann, S., Meleshko, R. & James, M. N. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins* **23**, 301–317 (1995).
12. Andreoletti, G., Pal, L. R., Moult, J. & Brenner, S. E. Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. *Hum. Mutat.* **40**, 1197–1201 (2019).
13. Dessimoz, C., Škunca, N. & Thomas, P. D. CAFA and the open world of protein function predictions. *Trends Genet* **29**, 609–610 (2013).
14. Weber, L. M. et al. Essential guidelines for computational method benchmarking. *Genome Biol.* **20**, 125 (2019).

15. Mangul, S. et al. Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 1393 (2019).
16. Mavromatis, K. et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* **4**, 495–500 (2007).
17. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6**, 19233 (2016).
18. McIntyre, A. B. R. et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18**, 182 (2017).
19. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**, 779–794 (2019).
20. Bremges, A. & McHardy, A. C. Critical Assessment of Metagenome Interpretation enters the second round. *mSystems* 3, e00103-18 (2018).
21. Fritz, A. et al. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
22. Singer, E. et al. Next generation sequencing data of a defined microbial mock community. *Sci. Data* **3**, 160081 (2016).
23. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).
24. Meyer, F. et al. AMBER: Assessment of Metagenome BinnERs. *GigaScience* 7, giy069 (2018).
25. Meyer, F. et al. Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* **20**, 51 (2019).
26. Grüning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).
27. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
28. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
29. Belmann, P. et al. Bioboxes: standardised containers for interchangeable bioinformatics software. *Gigascience* **4**, 47 (2015).
30. da Veiga Leprevost, F. et al. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **33**, 2580–2582 (2017).
31. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
32. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–D65 (2007).
33. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* **1**, 7 (2012).
34. Li, D. et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
35. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824–834 (2017).
36. Mineeva, O., Rojas-Carulla, M., Ley, R. E., Schölkopf, B. & Youngblut, N. D. DeepMAsED: evaluating the quality of metagenomic assemblies. *Bioinformatics* **36**, 3011–3017 (2020).
37. Clark, S. C., Egan, R., Frazier, P. I. & Wang, Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* **29**, 435–443 (2013).
38. Kuhring, M., Dabrowski, P. W., Piro, V. C., Nitsche, A. & Renard, B. Y. SuRankCo: supervised ranking of contigs in de novo assemblies. *BMC Bioinforma.* **16**, 240 (2015).
39. 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).
40. 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).
41. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
42. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
43. 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).
44. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
45. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
46. Gregor, I., Dröge, J., Schirmer, M., Quince, C. & McHardy, A. C. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* **4**, e1603 (2016).
47. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
48. Huson, D. H. et al. MEGAN Community Edition – interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* **12**, e1004957 (2016).
49. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2020).
50. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
51. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
52. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
53. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
54. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
55. Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**, 6258–6264 (2005).
56. McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
57. Segata, N. On the road to strain-resolved comparative metagenomics. *mSystems* 3, e00190-17 (2018).
58. Quince, C. et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).
59. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* **27**, 626–638 (2017).
60. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020).
61. Sajulga, R. et al. Survey of metaproteomics software tools for functional microbiome analysis. *PLoS ONE* **15**, e0241503 (2020).

Acknowledgements

The authors thank P. B. Pope for helpful comments. A.E.D.'s contribution was facilitated in part by the Australian Research Council's Discovery Projects funding scheme (project DP180101506). A.G.'s contribution was facilitated by St. Petersburg State University, Russia (grant ID PURE 51555639).

Author contributions

F.M. and T.-R.L. performed the experiments; F.M., A.F., T.-R.L., and A.S. prepared the data; A.C.M., A.B., and A.S. conceived the experiments; A.C.M., F.M., and A.B. wrote the manuscript with comments by others; F.M., T.-R.L., D.K., A.F., A.G., A.E.D., A.S., A.B., and A.C.M. interpreted the results, and read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-020-00480-3.

Correspondence and requests for materials should be addressed to A.C.MH.

Peer review information *Nature Protocols* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 10 March 2020; Accepted: 26 November 2020;
Published online: 1 March 2021