Assessing taxonomic metagenome profilers with OPAL

Fernando Meyer[1,2], Andreas Bremges[1,2,3], Peter Belmann[1,2,4], Stefan Janssen[1,3,5], Alice C. McHardy[1,2*], David Koslicki[6*]

1 Department of Computational Biology of Infection Research, Helmholtz Centre for Infection Research (HZI), Braunschweig, Germany
2 Braunschweig Integrated Centre of Systems Biology (BRICS), Braunschweig, Germany
3 German Center for Infection Research (DZIF), partner site Hannover-Braunschweig, Braunschweig, Germany
4 Faculty of Technology and Center for Biotechnology, Bielefeld University, Bielefeld, Germany
5 Department of Pediatrics, University of California San Diego, La Jolla, California, USA
6 Mathematics Department, Oregon State University, Corvallis, OR, USA

* ACM and DK are shared last authors.
alice.mchardy@helmholtz-hzi.de and david.koslicki@math.oregonstate.edu

Abstract

Taxonomic metagenome profilers predict the presence and relative abundance of microorganisms from shotgun sequence samples of DNA isolated directly from a microbial community. Over the past years, there has been an explosive growth of software and algorithms for this task, resulting in a need for more systematic comparisons of these methods based on relevant performance criteria. Here, we present OPAL, a software package implementing commonly used performance metrics, including those of the first challenge of the Initiative for the Critical Assessment of Metagenome Interpretation (CAMI), together with convenient visualizations. In addition, OPAL implements diversity metrics from microbial ecology, as well as run time and memory efficiency measurements. By allowing users to customize the relative importance of metrics, OPAL facilitates in-depth performance comparisons, as well as the development of new methods and data analysis workflows. To demonstrate the application, we compared seven profilers on benchmark datasets of the first and second CAMI challenges using all metrics and performance measurements available in OPAL. The software is implemented in Python 3 and available under the Apache 2.0 license on GitHub (https://github.com/CAMI-challenge/OPAL).

Author summary

There are many computational approaches for inferring the presence and relative abundance of taxa (i.e. taxonomic profiling) from shotgun metagenome samples of microbial communities, making systematic performance evaluations a very important task. However, there has yet to be introduced a computational framework in which profiler performances can be compared. This delays method development and applied studies, as researchers need to implement their own custom evaluation frameworks. Here, we present OPAL, a software package that facilitates standardized comparisons of taxonomic metagenome profilers. It implements a variety of performance metrics frequently employed in microbiome research, including runtime and memory usage, and generates comparison reports and visualizations. OPAL thus facilitates and accelerates benchmarking of taxonomic profiling techniques on ground truth data. This enables researchers to arrive at informed decisions about which computational techniques to use for specific datasets and research questions.

Introduction

Taxonomic metagenome profilers predict the taxonomic identities and relative abundances of microorganisms of a microbial community from shotgun sequence samples. In contrast to taxonomic binning, profiling does not result in assignments for individual sequences, but derives a summary of the presence and relative abundance of different taxa in microbial community. In some use-cases, such as pathogen identification for clinical diagnostics, accurate determination of the presence or absence of a particular taxon is important, while for comparative studies, such as quantifying the dynamics of a microbial community over an ecological gradient, accurately determining relative abundances of taxa is paramount.

Given the variety of use-cases, it is important to understand the benefits and drawbacks of the particular taxonomic profiler for different applications. While there has been much effort in developing taxonomic profiling methods [1–12], only recently have community efforts arisen to perform unbiased comparisons of such techniques and assess their strengths and weaknesses [13, 14]. Critical obstacles to such comparisons have been a lack of consensus on performance metrics and output formats by the community, as different taxonomic profilers report their results in a variety of formats and interested parties had to implement their own metrics for comparisons.

Here, we describe OPAL (Open-community Profiling Assessment tooL), a framework that directly addresses these issues. OPAL aggregates the results of multiple taxonomic profilers for one or more benchmark datasets, computes relevant metrics for different applications on them, and then presents the relative strengths and weaknesses of different tools in intuitive graphics. OPAL leverages the emerging standardized output format recently developed by the CAMI consortium [13, 15] to represent a taxonomic profile and which has been implemented for a variety of popular taxonomic profilers [2, 4–10, 12]. OPAL can also use the popular BIOM format [16]. The metrics that OPAL computes range from simple presence-absence metrics to more sophisticated comparative metrics such as UniFrac [17] and diversity metrics. The resulting metrics are displayed in graphics viewable in a browser and allow a user to dynamically rank taxonomic profilers based on the combination of metrics of their choice.

Similar efforts to provide comparative frameworks have recently been made for genome binners of metagenome samples (AMBER [18]) and metagenomic assemblers (QUAST [19, 20]). OPAL augments these efforts by addressing the issue of comparing and assessing taxonomic profilers. OPAL will assist future systematic benchmarking efforts. It will aid method developers to rapidly assess how their implemented taxonomic profilers perform in comparison to other techniques, and facilitate assessing profiler performance characteristics, such as clarifying when and where tool performance degrades (eg. performance at particular taxonomic ranks). Importantly, OPAL will help to decide which profiler is best suited to analyze particular datasets and biological research questions, which vary widely, depending on nature of the sampled microbial community, experimental setup and sequencing technology used [21].

Methods

Inputs

OPAL accepts as inputs one or several taxonomic profiles and benchmarks them at different taxonomic ranks against a given taxonomic gold standard profile.

Both the predicted and gold standard taxonomic profiles may contain information for multiple samples, such as for a time series, technical or biological replicates. A gold standard taxonomic profile can, for instance, be created with the CAMISIM metagenome simulator [21, 22]. The taxonomic profiles can be either in the Bioboxes profiling format [15, 23] or the BIOM (Biological Observation Matrix) format [16]. Examples are provided in the OPAL GitHub repository [24].

Metrics and accompanying visualizations

OPAL calculates a range of relevant metrics commonly used in the field [13] for one or more taxonomic profiles of a given dataset by comparing to a gold standard taxonomic profile. Below, we give formal definitions of all metrics, together with an explanation of their biological meaning.

Preliminaries

For r, a particular taxonomic rank (or simply rank), let x_r be the true bacterial relative abundances at rank r given by the gold standard. That is, x_r is a vector indexed by all taxa at rank r, where entry $(x_r)_i$ is the relative abundance of taxon i in the sampled microbial community at rank r. With x_r^* we denote the vector of predicted bacterial relative abundances at rank r. Accordingly, $(x_r^*)_i$ is the predicted relative abundance of taxon i at rank r.

By default, OPAL normalizes all (predicted) abundances prior to computing metrics, such that the sum of all abundances equals 1 at each rank, i.e. $\sum_i (x_r)_i = 1$ and $\sum_i (x_r^*)_i = 1$. This is to avoid any bias towards profiling software that makes fewer predictions, say, for only 50% of the sample.

Assessing the presence or absence of taxa

The purity and completeness of taxonomic predictions are common measures for assessing profiling quality [25]. They assess how well a profiler correctly identifies the presence and absence of taxa in a sampled microbial community without considering how well their relative abundances were inferred. This can be relevant, for example, in an emergency situation in clinical diagnostics, when searching for a pathogen in a metagenomic sample taken from patient material. To define these measures, let the support of the vector x_r be

$$supp(x_r) = \{i | (x_r)_i > 0\}. \tag{1}$$

That is, $supp(x_r)$ is the set of indices of the taxa at rank r present in the sample. Analogously, $supp(x_r^*)$ is the set of indices of the taxa at rank r predicted to be in the sample. For each rank r, we define the true positives TP_r, false positives FP_r, and false negatives FN_r, respectively, as

$$TP_r = |supp(x_r) \cap supp(x_r^*)| \tag{2}$$

$$FP_r = |supp(x_r)^c \cap supp(x_r^*)| \tag{3}$$

$$FN_r = |supp(x_r) \cap supp(x_r^*)^c| \tag{4}$$

where $supp(x_r)^c$ and $supp(x_r^*)^c$ are the complement of the respective support vectors, and, thus, give the indices of the taxa at rank r absent or predicted as absent in the sample. Specifically, TP_r and FP_r are the number of taxa correctly and incorrectly predicted as present in the sample, respectively, and FN_r is the number of taxa incorrectly predicted as being absent in the sample.

The **purity** $\boldsymbol{p_r}$ at rank r, also known as precision or specificity, is the ratio of taxa correctly predicted as present in the sample and all predicted taxa at that rank. For each rank r, the purity is computed as

$$p_r = \frac{TP_r}{TP_r + FP_r}. \tag{5}$$

The **completeness** $\boldsymbol{s_r}$ at rank r, also known as recall or sensitivity, is the ratio of taxa correctly predicted as present and all taxa present in the sample at that rank. For each taxonomic rank r, the completeness is computed as

$$s_r = \frac{TP_r}{TP_r + FN_r}. \tag{6}$$

Purity and completeness range from 0 (worst) to 1 (best).

We combine purity and completeness into a single metric by computing their harmonic average, also known as the **F1 score**. It is defined for each rank r as

$$\mathrm{F1}_r = 2 * \frac{p_r * s_r}{p_r + s_r}. \tag{7}$$

The F1 score ranges from 0 to 1, being closer to 0 if at least one of the metrics purity or completeness has a low value, and closer to 1 if both the purity and completeness are high.

The **Jaccard index** $\boldsymbol{J}$ is a common metric to determine the percentage of organisms common to two populations or samples. We define it as an indicator of similarity between the sets of true and predicted taxa at each rank by computing the ratio of the number of taxa in the intersection of these sets to the number of taxa in their union. Formally, it is computed for each rank as

$$J_r = \frac{|supp(x_r) \cap supp(x_r^*)|}{|supp(x_r) \cup supp(x_r^*)|}. \tag{8}$$

The Jaccard index ranges from 0 (complete dissimilarity) to 1 (complete overlap).

Abundance estimates

The next category of metrics for assessing profiling quality not only considers whether taxa was predicted as present or absent in the sample, but also their abundances.

The L1 norm measures the accuracy of reconstructing the relative abundance of taxa in a sample at rank r. The L1 norm is given by

$$\mathrm{L1}_r = \sum_i |(x_r)_i - (x_r^*)_i|. \tag{9}$$

The **L1 norm** thus gives the total error between the true and predicted abundances of the taxa at rank r. It ranges from 0 to 2, where 0 indicates perfect reconstruction of the relative abundances of organisms in a sample and 2 indicates totally incorrect reconstruction of relative abundances.

Another metric, the **Bray-Curtis distance $\boldsymbol{d_r}$**, is derived from the L1 norm by dividing the sum of the absolute pairwise differences of taxa abundances by the sums of all abundances at the given rank. This bounds the Bray-Curtis distance between 0 and 1. For each rank r, it defined as

$$d_r = \frac{\sum_i |(x_r)_i - (x_r^*)_i|}{\sum_i (x_r)_i + \sum_i (x_r^*)_i}. \tag{10}$$

The **weighted UniFrac distance** is a tree-based measure of taxonomic similarity of microbial communities [17] measuring the similarity between true and predicted abundances. Instead of a phylogenetic tree as in [17], we use a taxonomic tree with nodes restricted to eight major ranks and store the true and predicted abundances on the appropriate nodes. In summary, the UniFrac distance is the total amount of predicted abundances that must be moved (along the edges of the taxonomic tree, with all branch lengths here set to 1) to cause them to overlap with the true relative abundances. We use the EMDUnifrac implementation of the UniFrac distance [26–28]. A low UniFrac distance indicates that a taxonomic profiling algorithm gives a prediction that is taxonomically similar to the actual profile of the sample. The weighted UniFrac distance ranges between 0 and twice the height of the taxonomic tree used. Because each level of the tree represents one of the ranks superkingdom, phylum, class, order, family, genus, species, and strain, the maximum weighted UniFrac distance is 16.

The **unweighted UniFrac distance** is similar to the weighted UniFrac distance, but instead of storing the *relative abundances* for the appropriate nodes, a 1 is placed on the node if the profile indicates a non-zero relative abundance at that node and a 0 otherwise. Hence, it can be considered a measure of how well (in terms of taxonomic similarity) a profiler correctly identified the presence and absence of taxa in a sample. The maximum unweighted UniFrac distance is equal to

$$(|R| - 1) * \sum_{r \in R} |supp(x_r)|. \tag{11}$$

where R is the set of all taxonomic ranks.

Alpha diversity metrics

Unlike the metrics above, alpha diversity metrics are computed from a single profile of (predicted) abundances at each rank, without a comparison to e.g. a gold standard profile. Alpha diversity metrics summarize the variety (or richness) and distribution of taxa present in a profile [29] and, among other uses, is commonly used to observe global shifts in community structure as a result of some environmental parameter [30–33].

The simplest alpha diversity metric is the number of taxa present in a given environment. We measure this at each rank individually for a given profiler, allowing a comparison to the underlying gold standard. For a given profile x_r (or x_r^*), we denote the number of taxa at rank r as $S_r = |supp(x_r)|$.

As a measure of diversity also considering the relative taxon abundances, we combine S_r and all abundances $(x_r)_i$ (or $(x_r^*)_i$) using the **Shannon diversity index $\boldsymbol{H_r}$** [34]. For each rank r, it is calculated as

$$H_r = \sum_{i=1}^{S_r} (x_r)_i \ln(x_r)_i. \tag{12}$$

H_r ranges from 0 to $\ln(S_r)$, where $\ln(S_r)$ represents the maximal possible diversity, with all taxa being evenly represented. We note that the Shannon diversity index traditionally assumes that all taxa are represented in the sample. However, because some profilers may not predict abundances for all taxa, we ignore such taxa in the sum (where $(x_r^*)_i = 0$ or $(x_r)_i = 0$).

While H_r accounts for diversity and evenness, the **Shannon equitability index E_r** is a measure of evenness. It is a normalized form of the Shannon diversity index obtained by dividing H_r by its maximum value $\ln(S_r)$, i.e.

$$E_r = \frac{H_r}{\ln(S_r)}. \tag{13}$$

Thus, E_r ranges from 0 to 1 with 1 indicating complete evenness.

Beta diversity metrics

In contrast to alpha diversity, beta diversity metrics give an indication of taxa distribution similarity between a pair of profiles [29]. If beta diversity is small, not only is the diversity similar between the profiles, but the actual *distribution* of relative abundances between profiles are similar. To compare the similarity of beta diversity predictions for each profiler versus the gold standard, we display the following information in a scatter plot. Each point corresponds to a pair of input samples with the x-coordinate being the Bray-Curtis distance between the taxonomic profilers predictions on the pair of samples. The y-coordinate is the Bray-Curtis distance between the gold standards corresponding to the pair of samples. The closer this scatter plot is to the line $y = x$, the more closely the taxonomic profiler results in taxa distributions similar to the gold standard. These plots are shown at each taxonomic rank.

Rankings

To indicate a global sense of relative performance, we also rank profilers by their relative performance across each sample, taxonomic rank, and metric. In particular, each profiler is assigned a score for its performance for each metric within a taxonomic rank and sample. The best performing profiler gets 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. Also, an overall score of each profiler is computed by summing up all its scores per metric. The resulting scores are displayed in an interactive table of an HTML page, with a row per profiler, a column per metric, and an additional column for the overall scores. The columns can be sorted by the user and, therefore, yield a ranking of the profilers over all metrics or for a specific one. Optionally, the overall score of each profiler can be computed by summing up its score per metric in a weighted fashion, i.e., a user can interactively select custom weighting on the HTML page, depending on the combination of metrics that most suits their needs. The default weight of each metric is 1 and can vary between 0 and 10, in steps of 0.1. For example, if a user is interested in profilers that are highly precise and accurately reconstruct the exact relative abundance of predicted taxa, they can emphasize purity and L1 norm (e.g. giving each weight 3) over UniFrac error and completeness (e.g. giving each weight 1). The resulting rankings are dynamically updated in real time and graphically presented to the user.

Output and visualizations

OPAL outputs the assessment of the predictions of multiple profilers in several formats: flat files, tables (per profiling program, taxonomic rank, and in tidy format [35]), plots, and in an interactive HTML visualization. An example page is available at [36]. The visualizations created include:

- **Absolute performance plots**: To visually compare the relative performance of multiple profilers, spider plots (also known as radar plots) of completeness and purity are created, with the spokes labeled with the corresponding profiler name. At least three profilers are required for these plots. The completeness and purity metrics are shown as colored lines connecting the spokes, with the scale on the spokes indicating the value of the error metric. One such spider plot is created at each taxonomic rank to give an indication of performance versus rank.

- **Relative performance plots**: Similarly, spider plots are created for the completeness, purity, false positives, weighted UniFrac, and L1 norm for three or more profilers. Since the values of these metrics have very different scales, they are each normalized by the maximum value attained by any input profiler. Hence, these plots indicate the relative performance of each profiler with respect to the different metrics. For example, one profiler having the largest value of the purity metric would indicate that, amongst the compared profilers, it is the most precise (without indicating what the exact value of the purity metric is). These plots are also shown at each taxonomic rank.

- **Shannon equitability**: The Shannon equitability index is plotted against taxonomic ranks for each input profile along with the gold standard. This results in a visual indication of how closely a taxonomic profile reflects the actual alpha diversity of the gold standard.

- **Bray-Curtis distances**: For each profiler, a scatter plot of Bray-Curtis distances is created to compare the similarity of beta diversity of the profiler predictions versus the gold standard. For details, see the section above on beta diversity metrics.

- **Ranking**: In a bar chart shown on the created HTML page, each bar corresponds to the sum of scores obtained by a profiler as a result of its ranking for the metrics completeness, purity, L1 norm, and weighted UniFrac over all major taxonomic ranks. The bar chart is dynamically updated in real time according to the weight assigned to each metric by the user. For details of the computation of the scores, see the above section on rankings.

- **Taxa proportions**: For each taxonomic rank, a stacked bar chart shows the taxa proportions in each sample of the gold standard, with each bar corresponding to a sample and each color to a taxon. This gives a visual indication of the taxa abundances and variations among the samples. On the HTML page, the user may opt to see a legend of the colors and corresponding taxa. The legend is only optionally displayed since the number of taxa can vary between a few superkingdoms to hundreds or thousands of species or strains, and these cannot all be reasonably displayed on a single image.

- **Rarefaction and accumulation curves**: A plot simultaneously shows rarefaction and accumulation curves for all the major taxonomic ranks. To ease the visualization at different ranks, another plot shows the curves in logarithmic scale with base 10.

Results and discussion

To demonstrate an application, we evaluated taxonomic profilers on two datasets. First, we evaluated taxonomic profiling submissions to the first CAMI challenge [13] on the dataset with the highest microbial complexity in the challenge. We will call this dataset

CAMI I HC for short. This is a simulated time series benchmark dataset with 5 samples, each with size 15 Gbp, and a total of 596 genomes. It includes Bacteria, Archaea, and high-copy circular elements (plasmids and viruses) with substantial real and simulated strain-level diversity. We reproduce and extend the results for this dataset from [13] with alpha and beta diversity metrics implemented in OPAL and measure the run time and memory usage of profiling methods. The second dataset that we evaluated taxonomic profilers on were the short-read data of a new *practice* dataset of the second CAMI challenge (CAMI II MG, for short). This consists of 64 samples with a total size of 320 Gbp, and was simulated from taxonomic profiles for microbial communities from the guts of different mice [21]. This resulted in the inclusion of 791 genomes as *meta-community* members from public databases. The samples in both CAMI I HC and CAMI II MG are paired-end 150 bp Illumina reads and are available at [37, 38].

To visualize the taxonomic composition and properties of these datasets, we produced plots of the taxa proportions at all major taxonomic ranks for all samples with OPAL (Figs A and B in S1 Supporting Information for CAMI II MG and CAMI I HC, respectively) and calculated rarefaction curves (Fig C in S1 Supporting Information).

The assessed profilers were CommonKmers (an early version of MetaPalette) [2], a combination of Quikr [8], ARK [9], and SEK [10] (abbreviated Quikr), TIPP 2.0.0 [12], Metaphlan 2.2.0 [5], MetaPhyler 1.25 [6], mOTU 1.1 [7], and FOCUS (CAMI version) [4]. The software parameters used are available in Table A in S1 Supporting Information. The reference databases used by each profiler precede the release of the genomes used for generating the first CAMI challenge datasets. Thus the metagenomic information of the CAMI I HC dataset was completely new for these profilers and at different taxonomic distances to available reference genomes, differently from the metagenome data of the CAMI II MG practice dataset. The profilers were run as Bioboxes docker containers on a computer with an Intel Xeon E5-4650 v4 CPU (virtualized to 16 CPU cores, 1 thread per core) and 512 GB of main memory. Metaphlan was the fastest method on CAMI II MG with a run time of 12.5 hours, whereas on CAMI I HC, Metaphlan and Quikr were the fastest methods, requiring roughly the same execution time of 2.12 hours (Fig 1, Table B in S1 Supporting Information). mOTU was the most memory efficient method on both datasets (1.19 GB of maximum main memory usage), closely followed by Metaphlan (1.44 and 1.66 GB maximum main memory usage on CAMI II MG and CAMI I HC respectively).

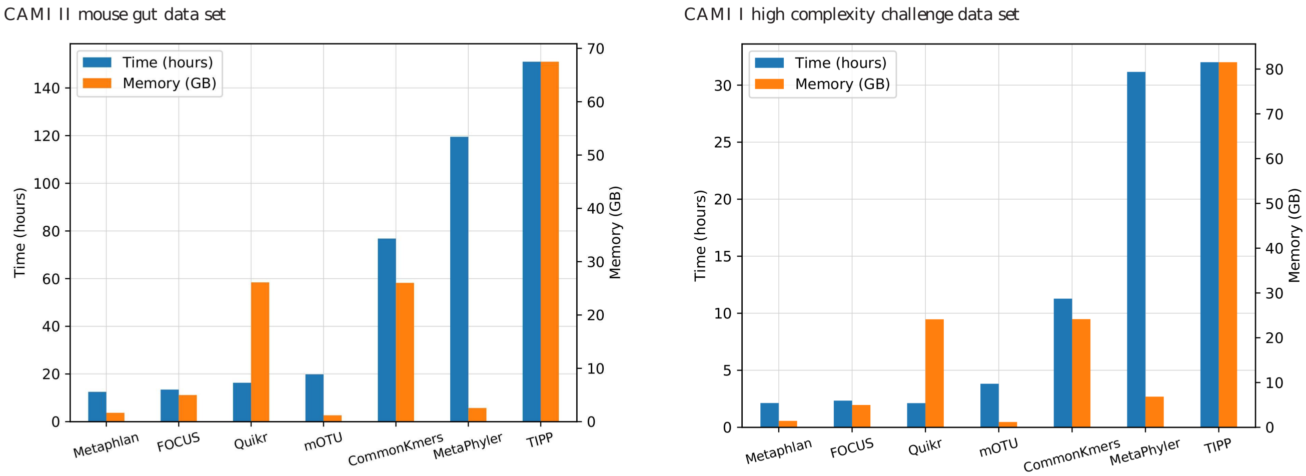


Fig 1. Computing efficiency. Run time in hours and maximum main memory usage in gigabytes required the by profilers to process the CAMI II mouse gut (left) and the CAMI I high complexity (right) datasets.

On the CAMI I HC data, Quikr, TIPP, and MetaPhyler, in this order, achieved the overall highest completeness (Figs D(a-c) and E-G(h-n) in S1 Supporting Information). However, these profilers obtained the lowest purity. In this metric, CommonKmers and Metaphlan performed best. In terms of the F1 score, computed from completeness and purity, Metaphlan was the best method. This indicates that Metaphlan performed particularly well in determining the presence or absence of taxa. However, it could not accurately predict their relative abundances, as indicated by the high L1 norm error. In this metric, MetaPhyler did well, followed by FOCUS and CommonKmers.

When ranking methods over all taxonomic ranks using completeness, purity, L1 norm, and weighted UniFrac with equal weights (Figs D(c) and H(b) in S1 Supporting Information), TIPP performed best with total score 184. TIPP ranked second for completeness and weighted UniFrac (scores 31 and 5, respectively), third for L1 norm (score 52), and only for purity it did not do so well and was ranked fifth (score 96). When considering the performance of the profilers at different taxonomic ranks, we found that most profilers performed well until the family level. For example, TIPP and MetaPhyler achieved a 0.92 completeness at the family level, but this decreased to 0.43 at the genus level. Similarly, the purity of CommonKmers decreased from 0.96 at the family level to 0.77 and 0.08 at the genus and species levels, respectively.

In terms of alpha diversity, no profiler estimated taxon counts well. Most programs overestimated diversity at all taxonomic ranks. Quikr, FOCUS, and CommonKmers predicted taxon abundances that better reflect the the Shannon equitability of the gold standard. However, Quikr, mOTU, and TIPP made no predictions at the strain level. The predicted abundance distributions of CommonKmers and mOTU across all samples at the species level best reflect the gold standard, as visualized with the scatter plots of Bray-Curtis distances (Fig J in S1 Supporting Information). Taken together, the OPAL results fully reproduce the results from [13], where performance was summarized in three categories of profilers: profilers that correctly predicted relative abundances, profilers with high purity, and those with high completeness. OPAL extends the overall performance view by providing analysis of computing efficiency and microbial diversity predictors.

On the CAMI II MG data, Metaphlan obtained the overall best ranking over all taxonomic ranks, using the equally weighted metrics completeness, purity, L1 norm, and weighted UniFrac (Fig 2(d) and Fig H(a) in S1 Supporting Information). MetaPhyler achieved the highest completeness at most taxonomic ranks, followed by TIPP and Metaphlan (Figs E-G(a-g) in S1 Supporting Information), whereas CommonKmers achieved the highest completeness at the species level (Fig 2(c)). Metaphlan was not only among the profilers with the highest completeness, but it also maintained a high purity throughout all taxonomic ranks, with only a small decrease from genus (0.94) to species (0.89). This can be explained by a high coverage of CAMI II MG by the reference genomes used by Metaphlan. It also contrasts with the results in [13], showing that a profiler can be precise while achieving a relative high completeness, but with this being very dependent on the input data. Metaphlan also predicted taxon distributions across the samples well. MetaPhyler and TIPP could not identify well differences in taxa abundances for the samples and tended to predict similar abundances, which is reflected in many points in the plots being above the line $x = y$ (Fig 3(b-h)).

In terms of alpha diversity, Metaphlan, CommonKmers, and mOTU predicted taxon counts similar to the gold standard for most taxonomic ranks, whereas the other profilers mostly overestimated the counts. On the other hand, TIPP, MetaPhyler, and mOTU predicted taxon abundances that more closely reflect their evenness, i.e., Shannon equitability, in the gold standard (Fig 3(a) and Fig I(a-b) in S1 Supporting Information). As on the CAMI I HC data, Quikr, mOTU, and TIPP made no strain-level predictions on this dataset.

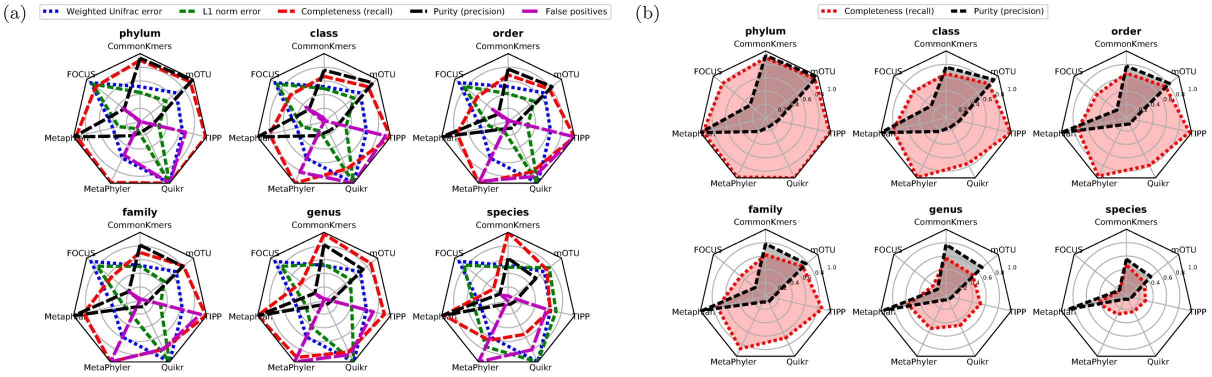
Conclusions

OPAL facilitates performance assessment and interpretation for taxonomic profilers using shotgun metagenome datasets as input. It implements commonly used performance metrics, including diversity metrics from microbial ecology, and outputs the assessment results in a convenient HTML page, in tables, and plots. By providing rankings and the possibility to give different weights to the metrics, OPAL enables the selection of the best profiler suitable for a researchers particular biological interest. In addition, computational efficiency results that OPAL returns can guide users on the choice of a profiler under time and memory constraints. We plan to continually extend the metrics and visualizations of OPAL according to community requirements and suggestions.

We used OPAL to analyze the CAMI I HC data, demonstrating how it enables reproduction of the results of this study [13]. We also used it for the analysis of a new large dataset, the CAMI II MG data, revealing consistency across many metrics and softwares analysed, and also a few striking differences. Specifically, while on the CAMI I HC data Quikr had the highest completeness by a wide margin, on the CAMI II MG data MetaPhyler performed best with this metric and Quikr was among the least complete profiling tools. Similarly, the Metaphlan results changed from the lowest to the highest weighted UniFrac score. Results such as these indicate the importance of choosing a program suitable for the particular properties of the microbial community analyzed and considering variables such as the availability of reference genome sequences of closely related organisms to those in the sample. Given the wide variety of environments from which metagenome data are obtained, this further demonstrates the relevance of OPAL.

Supporting information

S1 Supporting Information. Additional tables and figures.



(c)

Worst Median Best

Presence/absence of taxa

Tool	Gold standard	CommonKmers	FOCUS	Metaphlan	MetaPhyler	Quikr	TIPP	mOTU
Completeness (recall)	1 (0)	0.467 (0.0132)	0.128 (0.00801)	0.452 (0.0062)	0.303 (0.0059)	0.257 (0.00602)	0.295 (0.00566)	0.339 (0.00577)
Purity (precision)	1 (0)	0.552 (0.0199)	0.138 (0.0073)	0.891 (0.00416)	0.0505 (0.00175)	0.0547 (0.00262)	0.081 (0.0026)	0.466 (0.0101)
F1 score	1 (0)	0.484 (0.0109)	0.141 (0.00457)	0.598 (0.00558)	0.0852 (0.0025)	0.0886 (0.00374)	0.124 (0.00306)	0.386 (0.00427)
True positives	149 (6.86)	68.4 (3.36)	18.3 (1.25)	65.4 (2.66)	43.1 (1.56)	38.4 (1.97)	42 (1.56)	48.6 (1.86)
False positives	0 (0)	56.9 (3.32)	99.2 (4.87)	7.92 (0.384)	816 (15.7)	655 (8.72)	474 (6.12)	54.8 (1.7)
False negatives	0 (0)	80.5 (4.2)	131 (6.2)	83.5 (4.32)	106 (5.39)	110 (5.12)	107 (5.39)	100 (5.09)
Jaccard index	1 (0)	0.324 (0.00914)	0.0689 (0.00365)	0.428 (0.00573)	0.0446 (0.00136)	0.0466 (0.00205)	0.0665 (0.00173)	0.239 (0.00324)

Abundance estimates

Weighted Unifrac error	0 (0)	3.75 (0.184)	6.69 (0.364)	2.98 (0.124)	4.32 (0.0691)	7.27 (0.0765)	4.65 (0.078)	4.95 (0.196)
Unweighted Unifrac error	0 (0)	1.59e+03 (71.8)	2.47e+03 (98.2)	1.78e+03 (86)	7.86e+03 (154)	7.15e+03 (81.4)	5.36e+03 (88.9)	1.96e+03 (76.7)
L1 norm error	0 (0)	0.819 (0.0318)	1.4 (0.0385)	0.845 (0.0387)	1.26 (0.0399)	1.68 (0.0162)	1.21 (0.0425)	1.3 (0.0317)
Bray-Curtis distance	0 (0)	0.409 (0.0159)	0.699 (0.0192)	0.422 (0.0194)	0.632 (0.0199)	0.84 (0.00809)	0.605 (0.0212)	0.648 (0.0159)

Alpha diversity

Taxon counts	149 (6.86)	125 (4.43)	117 (5.95)	73.3 (2.93)	860 (16.3)	693 (9.5)	516 (6.87)	103 (2.99)
Shannon diversity	2.96 (0.0928)	3.09 (0.101)	3.59 (0.114)	2.39 (0.0904)	3.82 (0.078)	5.52 (0.0402)	3.56 (0.0794)	2.46 (0.0823)
Shannon equitability	0.469 (0.0147)	0.64 (0.0178)	0.795 (0.0109)	0.561 (0.019)	0.565 (0.0105)	0.845 (0.00492)	0.511 (0.0113)	0.529 (0.0157)

(d)

	Software (score)				
Ranking per metric	**Completeness**	**Purity**	**L1 norm**	**Weighted Unifrac**	**Sum of scores**
1st	MetaPhyler (211)	Metaphlan (75)	MetaPhyler (275)	Metaphlan (22)	Metaphlan (1,392)
2nd	TIPP (679)	CommonKmers (401)	Metaphlan (519)	CommonKmers (100)	MetaPhyler (2,115)
3rd	Metaphlan (776)	mOTU (484)	TIPP (691)	MetaPhyler (149)	CommonKmers (2,557)
4th	CommonKmers (1,079)	FOCUS (1,036)	mOTU (911)	mOTU (187)	mOTU (2,876)
5th	Quikr (1,084)	MetaPhyler (1,480)	CommonKmers (977)	TIPP (209)	TIPP (3,065)
6th	mOTU (1,294)	TIPP (1,486)	FOCUS (1,516)	FOCUS (306)	FOCUS (4,455)
7th	FOCUS (1,597)	Quikr (1,758)	Quikr (1,831)	Quikr (371)	Quikr (5,044)

Fig 2. Assessment results on the CAMI II mouse gut dataset. (a) Relative performance plots with results for the metrics: weighted UniFrac, L1 norm, completeness, purity, and number of false positives at different taxonomic ranks. The values of the metrics in these plots are normalized by the maximum value attained by any profiler at a certain rank. (b) Absolute performance plots with results for the metrics completeness and recall, ranging between 0 and 1. (c) Results at the species level for all computed metrics, as output by OPAL in the produced HTML page. The values are averaged over the results for all 64 samples of the dataset, with the standard error being shown in parentheses. The colors indicate the quality of the prediction by a profiler with respect to a metric, from best (dark blue) to worst (dark red). (d) Rankings of the profilers according to their performance and scores for different metrics computed over all samples and taxonomic ranks.

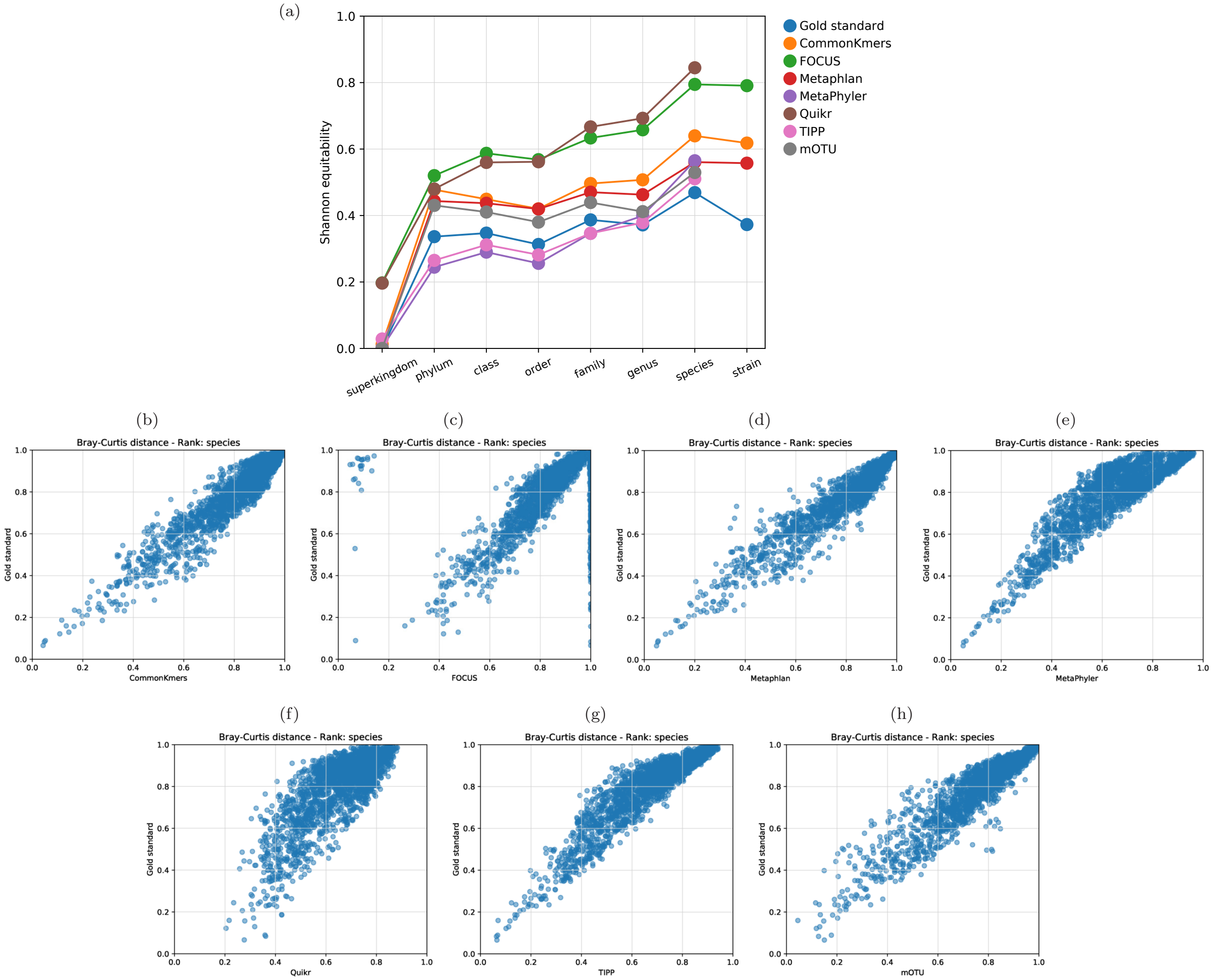


Fig 3. Examples of alpha and beta diversity plots from the results on the CAMI II mouse gut dataset. (a) Shannon equitability at different taxonomic ranks as a measure of alpha diversity. The closer the Shannon equitability of the predicted profile by a method to the gold standard, the better it reflects the actual alpha diversity in the gold standard in terms of evenness of the taxa abundances. (b-h) Scatter plots of Bray-Curtis distances visualizing beta diversity at the species level. For each profiling method and plot, a point corresponds to the Bray-Curtis distance between the abundance predictions for a pair of input samples by the method (x-axis) and the Bray-Curtis distance computed for the gold standard for the same pair of samples (y-axis). The closer a point is to the line $x = y$, the more similar the predicted taxa distributions are to the gold standard.

References

1. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):236+. doi:10.1186/s12864-015-1419-2.

2. Koslicki D, Falush D. MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. mSystems. 2016;1(3):e00020–16+. doi:10.1128/msystems.00020-16.

3. Piro VC, Lindner MS, Renard BY. DUDes: a top-down taxonomic profiler for metagenomics. Bioinformatics. 2016;32(15):2272–2280. doi:10.1093/bioinformatics/btw150.

4. Silva GG, Cuevas DA, Dutilh BE, Edwards RA. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ. 2014;2. doi:10.7717/peerj.425.

5. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012;9(8):811–814. doi:10.1038/nmeth.2066.

6. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011;12(Suppl 2):S4+. doi:10.1186/1471-2164-12-s2-s4.

7. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nature Methods. 2013;10(12):1196–1199. doi:10.1038/nmeth.2693.

8. Koslicki D, Foucart S, Rosen G. Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. Bioinformatics. 2013;29(17):2096–2102. doi:10.1093/bioinformatics/btt336.

9. Koslicki D, Chatterjee S, Shahrivar D, Walker AW, Francis SC, Fraser LJ, et al. ARK: Aggregation of Reads by K-Means for Estimation of Bacterial Community Composition. PLOS ONE. 2015;10(10):1–16. doi:10.1371/journal.pone.0140644.

10. Chatterjee S, Koslicki D, Dong S, Innocenti N, Cheng L, Lan Y, et al. SEK: sparsity exploiting k-mer-based estimation of bacterial community composition. Bioinformatics. 2014;30(17):2423–2431. doi:10.1093/bioinformatics/btu320.

11. Klingenberg H, Aßhauer KP, Lingner T, Meinicke P. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinformatics. 2013;29(8):973–980. doi:10.1093/bioinformatics/btt077.

12. Nguyen Np, Mirarab S, Liu B, Pop M, Warnow T. TIPP: taxonomic identification and phylogenetic profiling. Bioinformatics. 2014;30(24):3548–3555. doi:10.1093/bioinformatics/btu721.

13. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nature Methods. 2017;14(11):1063–1071.

14. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports. 2016;6(1). doi:10.1038/srep19233.

15. Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A, Barton MD. Bioboxes: standardised containers for interchangeable bioinformatics software. GigaScience. 2015;4(1). doi:10.1186/s13742-015-0087-0.

16. McDonald D, Clemente JC, Kuczynski J, Rideout J, Stombaugh J, Wendel D, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience. 2012;1(1):7+. doi:10.1186/2047-217x-1-7.

17. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology. 2005;71(12):8228–8235. doi:10.1128/aem.71.12.8228.

18. Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, et al. AMBER: Assessment of Metagenome BinnERs. GigaScience. 2018;7(6). doi:10.1093/gigascience/giy069.

19. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics (Oxford, England). 2016;32(7):1088–1090. doi:10.1093/bioinformatics/btv697.

20. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England). 2013;29(8):1072–1075. doi:10.1093/bioinformatics/btt086.

21. Fritz A, Hofmann P, Majda S, Dahms E, Droege J, Fiedler J, et al. CAMISIM: Simulating metagenomes and microbial communities. bioRxiv. 2018;doi:10.1101/300970.

22. Fritz A, Hofmann P, Majda S, Dahms E, Droege J, Fiedler J, et al.. CAMISIM: Simulating metagenomes and microbial communities; 2018. Available from: `https://github.com/CAMI-challenge/CAMISIM/`.

23. Bioboxes profiling format; 2018. Available from: `https://github.com/bioboxes/rfc/tree/master/data-format`.

24. OPAL GitHub repository; 2018. Available from: `https://github.com/CAMI-challenge/OPAL`.

25. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16(5):412–424. doi:10.1093/bioinformatics/16.5.412.

26. Evans SN, Matsen FA. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012;74(3):569–592. doi:10.1111/j.1467-9868.2011.01018.x.

27. McClelland J, Koslicki D. EMDUniFrac: exact linear time computation of the UniFrac metric and identification of differentially abundant organisms. Journal of Mathematical Biology. 2018;doi:10.1007/s00285-018-1235-9.

28. EMDUnifrac GitHub repository; 2018. Available from: `https://github.com/dkoslicki/EMDUnifrac`.

29. Whittaker RH. Evolution and Measurement of Species Diversity. Taxon. 1972;21(2):213–251.

30. Menni C, Jackson MA, Pallister T, Steves CJ, Spector TD, Valdes AM. Gut microbiome diversity and high-fibre intake are related to lower long-term weight gain. International Journal Of Obesity. 2017;41:1099–1105. doi:10.1038/ijo.2017.66.

31. Menni C, Zierer J, Pallister T, Jackson MA, Long T, Mohney RP, et al. Omega-3 fatty acids correlate with gut microbiome diversity and production of N-carbamylglutamate in middle aged and elderly women. Scientific Reports. 2017;7(1):2045–2322. doi:10.1038/s41598-017-10382-2.

32. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences. 2012;109(52):21390–21395. doi:10.1073/pnas.1215210110.

33. Mendes LW, Tsai SM, Navarrete AA, de Hollander M, van Veen JA, Kuramae EE. Soil-Borne Microbiome: Linking Diversity to Function. Microbial Ecology. 2015;70(1):255–265. doi:10.1007/s00248-014-0559-2.

34. Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948;27:379–423. doi:10.1145/584091.584093.

35. Wickham H. Tidy data. The Journal of Statistical Software. 2014;59(10).

36. OPAL example page; 2018. Available from: `https://cami-challenge.github.io/OPAL/`.

37. CAMI datasets download page; 2018. Available from: `https://data.cami-challenge.org/participate`.

38. Belmann P, Bremges A, Dahms E, Dröge J, Fiedler J, Fritz A, et al. Benchmark datasets, software results and reference data for the first CAMI challenge. GigaDB. 2017;doi:10.5524/100344.