# Evaluation of the Microba Community Profiler (MCP) for taxonomic profiling of metagenomic datasets from the human gut microbiome

Donovan H. Parks[1], Fabio Rigato[1], Lutz Krause[1], Philip Hugenholtz[1], Gene W. Tyson[1], David L.A. Wood[1]*

A fundamental goal of microbial ecology is to accurately determine the species composition in a given microbial ecosystem. In the context of the human microbiome, this is important for establishing links between microbial species and disease states. Here we benchmark the Microba Community Profiler (MCP) against other metagenomic classifiers using 140 moderate to complex *in silico* microbial communities and a standardized reference genome database. MCP generated accurate relative abundance estimates and made substantially fewer false positive predictions than other classifiers while retaining a high recall rate. We further demonstrated that the accuracy of species classification was substantially increased using the Microba Genome Database, which is more comprehensive than reference datasets used by other classifiers and illustrates the importance of including genomes of uncultured taxa in reference databases. Consequently, MCP classifies appreciably more reads than other classifiers when using their recommended reference databases. These results establish MCP as best-in-class with the capability of producing comprehensive and accurate species profiles of human gastrointestinal samples.

Identifying the microbial species present in natural biological samples is essential for understanding their role in a range of applications including developing diagnostics and therapeutics for human health (Greenblum *et al.*, 2012; Lloyd-Price *et al.*, 2016; Gentile and Weir, 2018; Zmora, *et al*. 2019), refining agricultural practices (Kennedy and Smith, 1995; Orellana *et al.*, 2018), and gaining insights into biogeochemical cycles (Kuypers *et al.*, 2018; Evans *et al.*, 2019). Our inability to culture most *in situ* populations has severely limited our understanding of microbial ecosystems (Epstein 2013; Lloyd *et al.*, 2019), and it is estimated that even highly studied habitats such as the human gut lack cultured representatives for the majority of species (Almeida *et al.*, 2020). Metagenomics, the sequencing of DNA extracted directly from clinical and environmental ecosystems has emerged as a powerful approach to bypassing this cultivation bottleneck, providing a holistic view of both the taxonomic and functional diversity of microbial communities (Hugenholtz & Tyson, 2008). This approach has been driven by exponential increases in sequencing throughput and associated decreasing costs leading to the widespread adoption of metagenomics by environmental and clinical researchers.

Metagenomics provides a relatively unbiased sampling of all populations within a community, including bacteria, archaea, eukaryotes and viruses, and the ability to resolve strains along with genes of interest such as those conferring antimicrobial resistance or pathogenicity (Weinstock, 2012; Köser *et al.*, 2014; Jovel *et al.*, 2016). However, accurately establishing the composition of microbial communities from metagenomic data remains a challenge due to their complexity, the comparatively short read length of the most widely used sequencing technologies (typically 150 to 250 bp), and incomplete genome reference databases (Sczyrba *et al.*, 2017; Ye *et al.*, 2019). This latter limitation is being addressed by recent approaches that recover high quality metagenome-assembled genomes (MAGs) from metagenomic datasets resulting in the availability of tens of thousands of draft genomes of uncultured taxa, most notably from the human gastrointestinal tract (Pasolli *et al.*, 2019; Almeida *et al.*, 2019; Nayfach *et al.*, 2019).

Several approaches have been proposed for taxonomically classifying metagenomic data in order to estimate the relative abundance of species in a sample. Metagenomic reads are

[1] Microba Life Sciences Limited, 388 Queen St, Brisbane, QLD 4000, Australia
* david.wood@microba.com

classified on the basis of sequence similarity to a reference database of previously characterized sequence data, often whole-genome assemblies. Existing metagenomic classifiers can be divided into four groups based on how they establish sequence similarity; namely, i) genome nucleotide alignment approaches such as Centrifuge (Kim *et al*., 2016), ii) protein alignment approaches such as Kaiju (Menzel *et al*., 2016) and DIAMOND (Buchfink *et al*., 2015), iii) marker gene approaches such as MetaPhlAn (Segata *et al*., 2012) and mOTUs (Milanese *et al*., 2019), and iv) composition or k-mer based approaches such as Kraken (Wood *et al*., 2019), Bracken (Lu *et al*., 2017), MetaCache (Müller *et al*., 2017), and Ganon (Piro *et al*., 2019). In general, k-mer-based approaches are the most computationally efficient, although have high memory requirements. Marker-based approaches typically have lower memory requirements but at the cost of only classifying reads from a specific subset of genes or genomic regions. Alignment-based approaches favour the additional information provided from mapping reads to reference sequences at the cost of higher computational requirements than k-mer based approaches and higher memory requirements than marker-based approaches.

The Microba Community Profiler (MCP) was developed to be a highly accurate and specific tool to estimate the relative abundance of bacterial, archaeal, eukaryotic, and viral community members by aligning metagenomic reads to a high quality and comprehensive database of microbial reference MAGs and isolate genomes. Similar to other classifiers, MCP provides per-read classifications along with an estimate of the proportion of reads assigned to a species. MCP also explicitly indicates the species predicted to be present in a community profile, in contrast to the majority of classifiers considered in this study which report thousands of false positive species if profiles are not manually inspected and appropriately filtered. The community profiles produced by the MCP are based on the rank normalized taxa and comprehensive species clusters defined by the Genome Taxonomy Database (GTDB; Parks *et al*., 2018, 2020) which provides higher taxonomic resolution than the NCBI Taxonomy (Parks *et al*., 2020; Federhen, 2015). Here we benchmark MCP against a range of widely used academic metagenomic classifiers using 140 *in silico* mock communities of varying complexity. We demonstrate that MCP has superior recall and precision and maps a higher proportion of reads from gut metagenome datasets.

## Results

### *Metagenomic classifiers and standardized reference database*

We evaluated the performance of MCP and nine publicly available metagenomic classifiers (**Table 1**), which use a variety of approaches and have previously been shown to be amongst the best performing classifiers (Seppey *et al*., 2020; Ye *et al*., 2019; Lindgreen *et al*., 2016; Sczyrba *et al*., 2017). A single standardized reference database was used by all classifiers in order to evaluate classification performance independent of the reference database (Ye *et al*., 2019; Nasko *et al*. 2018; Méric *et al*., 2019), with the exception of MetaPhlAn2, which was used with its pre-built marker database because building a custom database was not practical. The standardized reference database is comprised of 15,555 quality filtered isolate genomes from 12,250 bacterial and archaeal species (*see Methods*; **Supp. Table 1**) estimated to have an average completeness and contamination of 99.2% and 0.73%, respectively. Only high-quality isolate genomes were included in the standardized reference database to ensure classification performance would not be adversely impacted by low genome quality and to reflect that most classifiers recommend the use of reference databases comprised solely of complete isolate genomes (*see Methods*). Species were limited to a maximum of five representative genomes in order to reserve a wide diversity of strains for simulating *in silico* mock communities. Species represented by >1 genome (1,474 of 12,250) had an average intraspecific ANI of 97.8 ± 0.96%. The standardized reference database and comparison of profilers was limited to bacterial and archaeal species as not all classifiers support the classification of eukaryotic or viral species.

Three parameter settings for the MCP were evaluated: i) MCP with the standardized reference database and default parameters used to filter out expected false positive predictions (referred to as MCP); ii) MCP without removing expected false positives (referred to as unfiltered MCP or uMCP); and iii) MCP with default filtering parameters using the Microba Genome Database (MGDB), which comprises 73,646 dereplicated genomes from 28,246 species and is the reference database used by MCP in practice (referred to as MCP-MGDB).

**Table 1.** Properties of classifiers compared in this study.

| Classifier | Version | Classifier Type | Base Type | Reference |
|---|---|---|---|---|
| MCP | 2.0.15 | genome | DNA | (this study) |
| Ganon | 0.1.5 | k-mer (k=19) | DNA | *Piro et al.*, 2019 |
| Kraken | 2.0.7 | k-mer (k=35) | DNA | *Wood et al.*, 2019 |
| Bracken | 2.5.0 | k-mer (k=35) | DNA | *Lu et al.*, 2017 |
| MetaCache | 0.9.0 | k-mer (k=16) | DNA | *Müller et al.*, 2017 |
| Centrifuge | 1.0.4 | genome | DNA | *Kim et al.*, 2016 |
| DIAMOND-LCA | 0.9.29 | protein | protein | *Buchfink et al.*, 2015 |
| Kaiju | 1.7.2 | protein | protein | *Menzel et al.*, 2016 |
| mOTUs | 2.5.1 | marker | DNA | *Milanese et al.*, 2019 |
| MetaPhlAn[#] | 2.96.1 | marker | DNA | *Truong et al.*, 2012 |

[#] evaluated using MetaPhlAn database v296 downloaded on Feb. 24, 2020

## *Simulation of* in silico *mock communities*

We simulated 140 *in silico* mock microbial communities with varying species diversity, intraspecific diversity, and genomic similarity to reference database genomes (**Table 2; Supp. Table 2**). Communities were comprised of bacterial and archaeal species and simulated with either medium (100 ± 25) or high (500 ± 100) species diversity relative to previously used mocks (Sczyrba *et al*., 2017), with each species comprised of either a single strain or up to 10 randomly selected strains (*see Methods*). The average nucleotide identity (ANI) to reference genomes was used to construct mock communities with high (ANI of 99% to 99.75%), moderate (ANI of 97% to 99%), and low (ANI of 95% to 97%) genomic similarity to the standardized reference database. A baseline of 95% ANI was selected to match the commonly used operational definition of a microbial species (Jain *et al*., 2018; Parks *et al*., 2020). Mock communities were simulated under all combinations of these parameters, with the exception of mocks with high species diversity and low ANI similarity, as there were insufficient species with available genomes within this lower ANI range. In addition, mock communities were simulated from the reference genomes in order to establish a baseline at 100% ANI similarity for examining the impact of increasing genomic dissimilarity from reference genomes on classifier performance. The 140 mock communities span 6,971 unique species from 2,268 genera and 50 phyla, and contain species ranging from 0.0000019 to 80.5% of the community (**Table 2**). Communities were simulated to a depth of 2.1 Gb using 2×150 bp paired-end reads with the abundance of strains following a log-normal distribution as this is commonly used for modelling microbial communities (Curtis *et al*., 2002; Fritz *et al*., 2019; *see Methods*).

## *Establishing detection limits of classifiers*

By default, many metagenomic classifiers report all species with any evidence of being present within a sample, down to a single mapped read, which can result in thousands of low abundance false positive species predictions, i.e. species not present in the sample (**Figs. 1A** and **1B**; **Supp. Table 3**). The implicit expectation is that researchers will filter low abundance predictions or only consider analyses which are insensitive to false positive predictions (Ye *et al*., 2019). Unfortunately, the former is challenging without explicit guidance and the latter is highly restrictive as it limits the ability to confidently assert the presence of low abundance species in a sample. MCP, mOTUs, and MetaPhlAn are exceptions as their predicted community profiles contain only those species with sufficient evidence to assert with high confidence that a species is present in a sample (**Figs. 1A** and **1B**). Consequently, even for the mock communities with high ANI similarity to reference database genomes the evaluated classifiers report a high proportion of false positives (average of 86.4 to 96.8% of predicted species), with the exceptions of MCP (0.18 ± 0.44%) and to a lesser extent mOTUs (3.6 ± 2.1%) and MetaPhlAn (7.6 ± 3.7%) (**Supp. Table 3**).

Here, the *in silico* mock communities were used to establish detection limits for the different classifiers. Intuitively, the detection limit of a classifier is the lowest abundance species in a sample that can be identified before an unacceptable number of false positive species are reported. While the tolerance for false positives

**Table 2.** Properties of the 140 *in silico* mock communities averaged over the 10 replicates from each class.

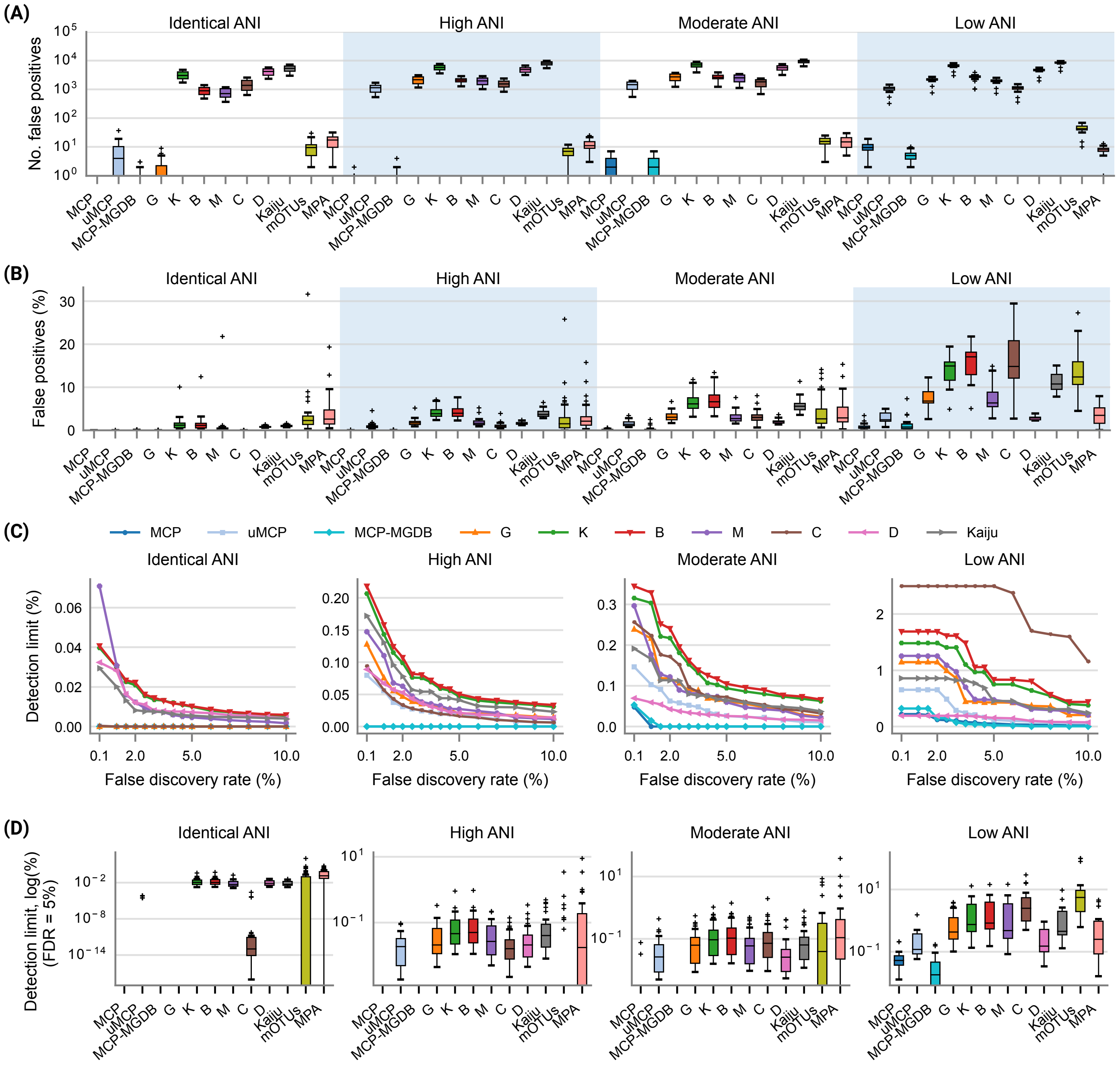
| ANI similarity | Species diversity | Strain diversity | ANI to closest reference genome (%) | AF to closest reference genome (%)[#] | No. species | Strains per species | Species abundance (%) |
|---|---|---|---|---|---|---|---|
| Identical: 100% | Medium | Single | 100 | 100 | 106 ± 15.8 | 1 | 26.8 to 3.3×10-4 |
| Identical | Medium | Multiple | 100 | 100 | 92 ± 22.5 | 2.6 ± 0.16 | 62.9 to 9.0×10-5 |
| Identical | High | Single | 100 | 100 | 490 ± 96.0 | 1 | 26.9 to 1.9×10-6 |
| Identical | High | Multiple | 100 | 100 | 505 ± 74.8 | 2.5 ± 0.07 | 13.2 to 6.1×10-6 |
| High: [99%, 99.75%] | Medium | Single | 99.4 ± 0.22 | 94.5 ± 3.02 | 99 ± 21.3 | 1 | 38.0 to 2.4×10-4 |
| High | Medium | Multiple | 99.3 ± 0.22 | 94.4 ± 3.06 | 106 ± 29.7 | 4.7 ± 0.33 | 39.4 to 3.2×10-4 |
| High | High | Single | 99.4 ± 0.22 | 94.5 ± 2.93 | 499 ± 86.1 | 1 | 60.3 to 1.6×10-5 |
| High | High | Multiple | 99.3 ± 0.22 | 94.4 ± 3.00 | 450 ± 116 | 4.0 ± 0.32 | 18.4 to 1.3×10-5 |
| Moderate: [97%, 99%) | Medium | Single | 98.3 ± 0.54 | 90.9 ± 4.41 | 104 ± 24.3 | 1 | 62.3 to 2.3×10-4 |
| Moderate | Medium | Multiple | 98.4 ± 0.52 | 91.2 ± 3.94 | 106 ± 19.6 | 4.7 ± 0.16 | 29.6 to 3.2×10-4 |
| Moderate | High | Single | 98.3 ± 0.54 | 90.8 ± 4.23 | 509 ± 58.6 | 1 | 23.2 to 2.8×10-5 |
| Moderate | High | Multiple | 98.3 ± 0.53 | 91.1 ± 4.19 | 532 ± 70.9 | 3.8 ± 0.26 | 10.0 to 9.2×10-6 |
| Low: [95%, 97%) | Medium | Single | 96.4 ± 0.50 | 87.9 ± 4.56 | 93 ± 32.9 | 1 | 80.5 to 2.8×10-4 |
| Low | Medium | Multiple | 96.3 ± 0.52 | 88.0 ± 4.33 | 109 ± 26.6 | 3.2 ± 0.23 | 36.6 to 1.4×10-4 |

[#] AF = alignment fraction, i.e. percentage of orthologous regions shared between two genomes

is subjective and application-specific, in general false positive predictions must be kept low in order to have confidence in the species reported by a classifier. We define the detection limit of each classifier as the lowest reported abundance at which a target false discovery rate (FDR) can be achieved. As expected, the detection limit increases as community members becoming increasingly divergent from genomes in the reference database (**Fig. 1C**). The detection limit also varies substantially between classifiers with MCP having the lowest detection limit regardless of the target FDR (**Fig. 1C; Table 3**). At an FDR of 0.1% (i.e. 1 in 1000 species expected to be false positives), the MCP had a mean detection limit of 0.0068%, 0.069%, and 0.52% on mock communities with high, moderate, and low ANI similarity to the reference database, respectively (**Table 3**). Examining the results at an FDR of 5%, illustrates that the detection limit varies substantially for individual mock communities at a specific level of ANI similarity (**Fig. 1D**). This highlights the challenge in specifying a fixed abundance threshold for filtering classification results which will reliably remove the majority of false positives species and, hence, the need for classifiers to directly address the issue of false positive predictions.

**Table 3.** Mean detection limit of classifiers at select false detection rates.

| | High ANI | | | | Moderate ANI | | | | Low ANI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | **0.1%** | **1%** | **5%** | **10%** | **0.1%** | **1%** | **5%** | **10%** | **0.1%** | **1%** | **5%** | **10%** |
| MCP | 0.0068 | 0.0016 | 0 | 0 | 0.069 | 0.048 | 0.0027 | 0 | 0.52 | 0.52 | 0.069 | 0.014 |
| Unfiltered MCP | 0.25 | 0.23 | 0.023 | 0.011 | 0.23 | 0.21 | 0.056 | 0.025 | 0.98 | 0.98 | 0.26 | 0.079 |
| MCP w/ MGDB | 0.014 | 0.00097 | 0 | 0 | 0.17 | 0.14 | 0 | 0 | 1.5 | 1.5 | 0.037 | 0.0039 |
| Ganon | 0.30 | 0.27 | 0.045 | 0.021 | 0.38 | 0.35 | 0.095 | 0.046 | 2.2 | 2.2 | 1.1 | 0.62 |
| Kraken | 0.39 | 0.37 | 0.097 | 0.038 | 0.51 | 0.47 | 0.18 | 0.10 | 2.6 | 2.6 | 2.2 | 0.9 |
| Bracken | 0.41 | 0.38 | 0.11 | 0.042 | 0.58 | 0.54 | 0.21 | 0.11 | 2.9 | 2.9 | 2.6 | 1.1 |
| MetaCache | 0.36 | 0.33 | 0.049 | 0.021 | 0.49 | 0.43 | 0.098 | 0.037 | 2.7 | 2.7 | 2.2 | 0.55 |
| Centrifuge | 0.25 | 0.22 | 0.026 | 0.012 | 0.49 | 0.45 | 0.16 | 0.061 | 5.5 | 5.5 | 5.3 | 4.6 |
| DIAMOND-LCA | 0.14 | 0.13 | 0.042 | 0.018 | 0.15 | 0.15 | 0.053 | 0.032 | 0.33 | 0.33 | 0.3 | 0.12 |
| Kaiju | 0.27 | 0.24 | 0.085 | 0.032 | 0.62 | 0.60 | 0.12 | 0.061 | 1.8 | 1.8 | 1.5 | 0.56 |
| mOTUs | 4.0 | 3.9 | 0.13 | 0 | 2.6 | 2.6 | 0.7 | 0.041 | 19 | 19 | 18 | 10 |
| MetaPhlAn | 2.3 | 2.3 | 0.43 | 0.0094 | 2.9 | 2.9 | 1.6 | 0.063 | 2.5 | 2.5 | 0.96 | 0.15 |



uMCP = unfiltered MCP; G = Ganon; K = Kraken; B = Bracken; M = MetaCache; C = Centrifuge; D = DIAMOND-LCA, MPA = MetaPhlAn

**Figure 1.** Metagenomic classifiers have different minimum species abundance limits at which species can be reliably detected. (**A**) Number of false positive species predictions made by each classifier on mock communities with decreasing ANI similarity to reference database genomes. (**B**) Percentage of predicted community profiles comprised of false positive predictions. The relatively low community percentages indicate that the majority of false positive predictions are low abundance species. With the exception of the MCP, mOTUs, and MetaPhlAn, these results illustrate that low abundance species must be filtered from the profiles predicted by metagenomic classifiers in order to reduce false positive predictions. (**C**) Median detection limit of each classifier over all mock communities at a given level of ANI similarity to the reference database for varying false discovery rates. MCP, uMCP, and Ganon report zero false positives for mock communities comprised of genomes in the reference database (identical ANI) and consequently have a median detection limit reported as 0% indicating all species could be identified without any false positives. Centrifuge and MCP-MGDB only report extremely low abundance false positives for the identical ANI mock communities resulting in median detection limits of 0.00036% and 0%, respectively. Results for mOTUs and MetaPhlAn are not shown as they have substantially higher detection limits than the other classifiers (**Supp. Fig. 1; Table 3**). (**D**) Detection limit of each classifier on each mock community resulting in a false discovery rate of 5%. MCP, MCP-MGDB, and Ganon have detection level at or near 0% across all samples at a number of ANI levels so do not produce visible box-and-whisker plots (see **Table 3**). The box-and-whisker plots show the lower and upper quartiles as a box, the median value as a line within the box, 1.5X the interquartile range as whiskers, and outliers as crosses.

## *Predicting the presence or absence of species*

In order to assess the accuracy of species predictions for the different classifiers, we conservatively removed low abundance populations at <0.01% as these have a high probability of being reported as false positives by all classifiers other than the MCP (**Fig. 1C**). Removing lower abundance species ensures more accurate results as it acknowledges that species comprising the "long tail" of microbial communities (Curtis *et al.*, 2002; Fritz *et al.*, 2019) cannot be identified by most metagenomic classifiers without reporting unacceptable numbers of false positives (**Fig. 1A** and **1B**). The mock communities contained an average of 271.4 ± 205.8 and 210.0 ± 141.9 species before and after removal of species at <0.01% abundance, respectively (**Supp. Table 2**).

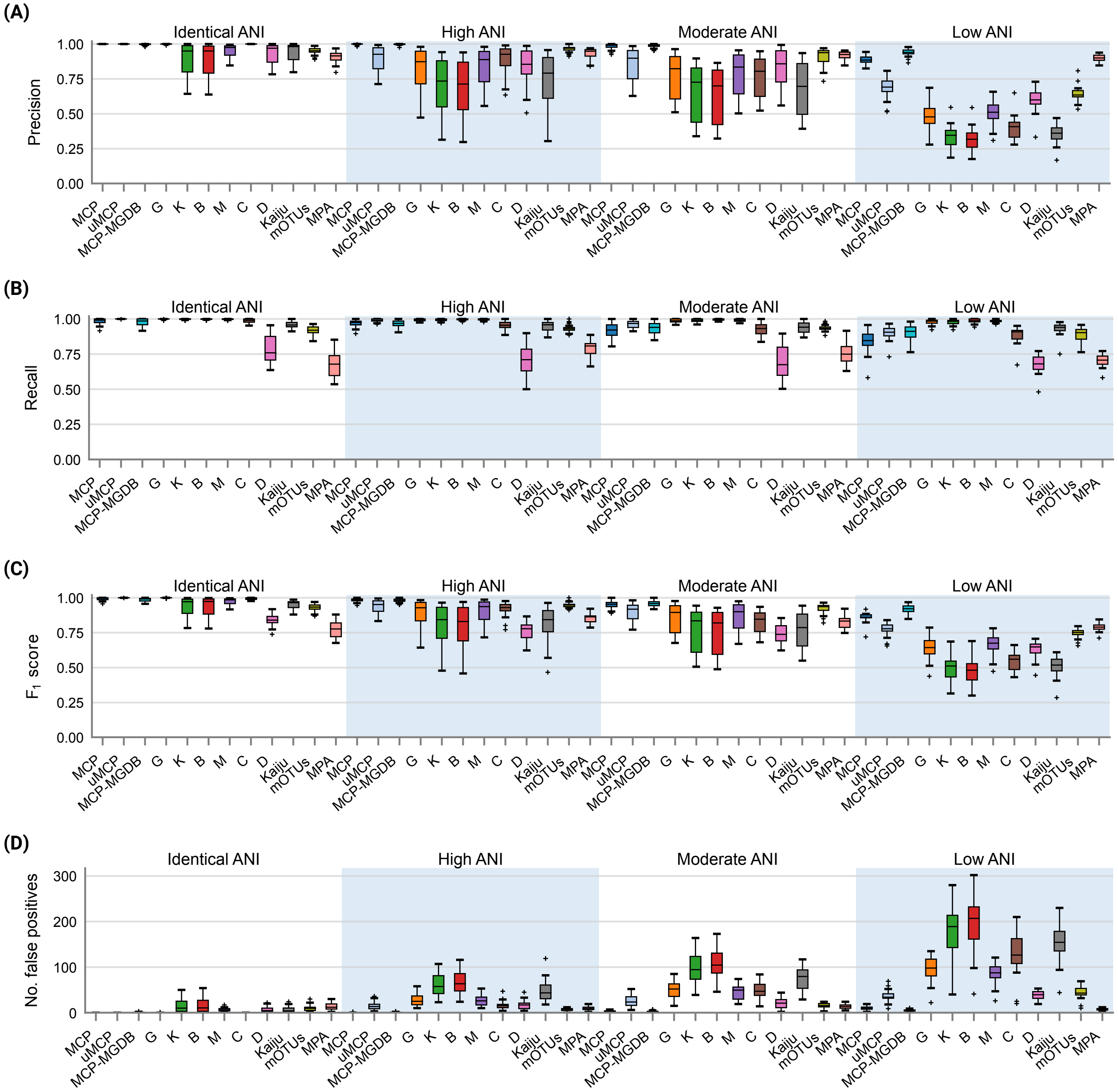
The performance of classifiers generally decreased with increasing ANI divergence from the reference database (**Fig. 2; Table 4; Supp. Table 4**), consistent with previous studies showing the importance of using a comprehensive reference database (Méric *et al.*, 2019; Piro *et al.*, 2019). MCP reported the lowest number of false positive species as indicated by its high precision (**Fig. 2A**). However, there is typically a trade-off between precision and recall, and this is reflected in MCP failing to identify some species whereas other classifiers such as MetaCache and Bracken have high recall with low relative precision (**Figs. 2A** and **2B**). With equal weight given to precision and recall the MCP using the MGDB database has the best overall performance ($F_1$ = 0.97 averaged across all mocks; **Fig. 2C; Table 4**), which demonstrates the positive impact of using a large, comprehensive reference database. Among the classifiers using the standardized reference database, MCP has the best performance across all mock communities ($F_1$ = 0.96) followed by the unfiltered MCP profiles ($F_1$ = 0.92), mOTUs ($F_1$ = 0.91), and MetaCache ($F_1$ = 0.88) (**Fig. 2C; Table 4**). MetaPhlAn performs relatively poorly ($F_1$ = 0.81) despite using a reference database built from nearly six times as many genomes as the standardized reference database illustrating that a comprehensive database is not sufficient in and of itself to provide good performance.

MCP provided the best overall performance without the need for manual thresholding because it automatically filters species profiles based on the number of stringently mapped reads being assigned to a species. By contrast, all other classifiers, with the exception of mOTUs and MetaPhlAn, report large numbers of false positives despite limiting results to species at ≥0.01% relative abundance (**Fig. 2D**). In order to further explore the performance of MCP relative to the other classifiers, profiles were filtered at the species abundance resulting in the highest $F_1$ score as determined independently for each classifier on each mock community (referred to as the optimized $F_1$ score). Notably, the average MCP $F_1$ score *without optimization* of 0.96 is higher than the optimized $F_1$ score of all other classifiers (**Table 4**). We note that establishing the $F_1$ optimized species abundance threshold is not possible on real data and any fixed abundance threshold will result in the same or worse performance than achieved with these optimized thresholds (**Fig. 1D**).

## *Estimating the relative abundance of species*

Based on mock community analysis (mocks filtered at 0.01%), the accuracy of relative abundance estimates decreased with increasing ANI divergence from the reference database (**Fig. 3; Table 5; Supp. Table 5**). Centrifuge, DIAMOND-LCA, Kaiju, MetaPhlAn, and to a lesser extent mOTUs deviate substantially from the expected species abundances (**Table 5**), consistent with prior benchmarking of these classifiers (Ye *et al.*, 2019). The other classifiers have similar overall accuracy in terms of L1 distance (i.e. absolute differences between profiles) with MetaCache (9.0%) performing the best followed by MCP-MGDB (10.0%), MCP (10.8%), and Bracken (11.1%) (**Fig. 3A; Table 5**). Results were similar for the relative absolute percent error with MetaCache having a 1 to 2% overall improvement over the other classifiers (**Fig. 3A; Table 5**). These results indicate that MCP, MCP-MGDB, Ganon, Kraken, Bracken, and MetaCache are all able to provide reasonably accurate species abundance estimates although the obtained accuracy depends heavily on the similarity of community members to genomes in the reference database. This is seen most clearly with the low ANI similarity mock communities where the abundance estimates are substantially less accurate and more variable (**Fig. 3A**).

The high precision of the MCP (**Fig. 2A**) resulted in only a small percentage of the predicted community being comprised of false positive species (0.18 ± 0.45%; **Table 5**). This is in contrast to the other classifiers which predict more false positive species at an appreciably higher percentage of the community, e.g. MetaCache at 1.97 ± 3.0% and Bracken at 3.92 ± 4.53% (**Fig. 3C; Table 5**). The low ANI similarity mock communities best highlight the tradeoff between the MCP and a more lenient classifier such as MetaCache where false positive species account for 1.0 ± 0.77% and 6.6 ± 3.2% of the reported communities, respectively (**Fig. 3C; Supp. Table 5**). Classifiers generally only fail to identify low abundance species with MCP showing slightly decreased performance as expected from its lower recall rate (**Figs. 2B** and **3D; Table 5**). This again highlights the trade-off between false negative and false positive predictions, and illustrates that the MCP favors a slight increase in the percentage of the community that is undetected (**Fig. 3D**) in order to substantially reduce the percentage of the reported community comprised of erroneously identified species (**Fig. 3C**).



uMCP = unfiltered MCP; G = Ganon; K = Kraken; B = Bracken; M = MetaCache; C = Centrifuge; D = DIAMOND-LCA, MPA = MetaPhlAn

**Figure 2.** Performance of metagenomic classifiers to predict the presence or absence of species measured using (**A**) precision, (**B**) recall, (**C**) $F_1$ score, and (**D**) number of false positive predictions on mock communities with decreasing ANI similarity to reference database genomes.

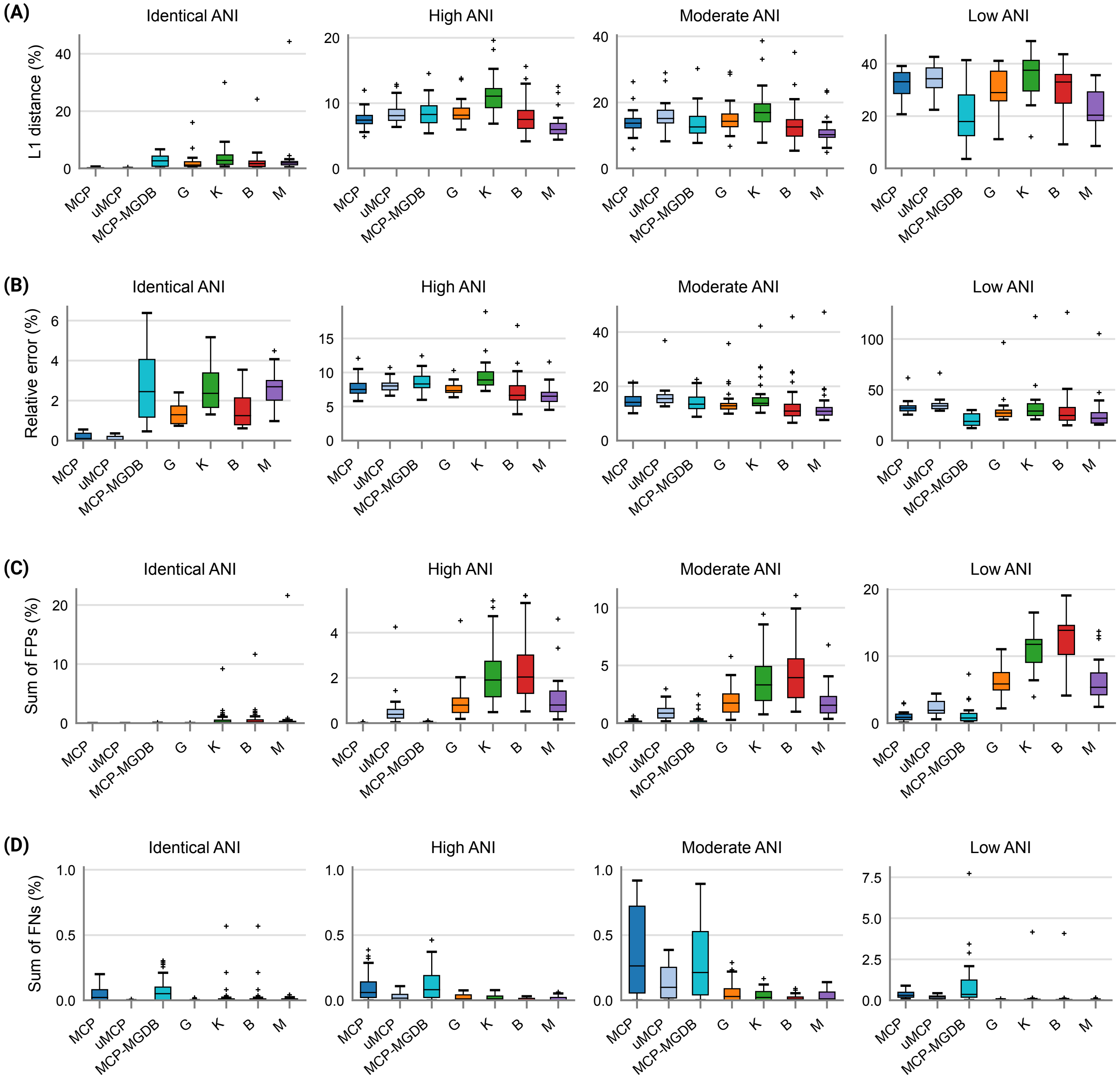
**Table 4.** Evaluation of classifiers to predict the presence or absence of species across the 140 mock communities with and without optimizing the $F_1$ score (mean ± std. dev.). MCP and MCP w/MGDB were run with default settings without $F_1$ optimization.

| Classifier | Precision | Recall | $F_1$ score | Precision ($F_1$ optimized) | Recall ($F_1$ optimized) | Optimized $F_1$ score |
|---|---|---|---|---|---|---|
| MCP | 0.98 ± 0.04 | 0.94 ± 0.06 | 0.96 ± 0.05 | - | - | - |
| Unfiltered MCP | 0.88 ± 0.13 | 0.97 ± 0.04 | 0.92 ± 0.09 | 0.94 ± 0.07 | 0.95 ± 0.06 | 0.94 ± 0.06 |
| MCP w/ MGDB | 0.99 ± 0.02 | 0.95 ± 0.04 | 0.97 ± 0.03 | - | - | - |
| Ganon | 0.81 ± 0.20 | 0.99 ± 0.01 | 0.87 ± 0.14 | 0.91 ± 0.10 | 0.94 ± 0.07 | 0.92 ± 0.08 |
| Kraken | 0.69 ± 0.24 | 0.99 ± 0.01 | 0.79 ± 0.18 | 0.87 ± 0.11 | 0.91 ± 0.09 | 0.88 ± 0.09 |
| Bracken | 0.68 ± 0.24 | 1.00 ± 0.01 | 0.78 ± 0.19 | 0.87 ± 0.11 | 0.90 ± 0.10 | 0.88 ± 0.09 |
| MetaCache | 0.80 ± 0.18 | 0.99 ± 0.01 | 0.88 ± 0.12 | 0.90 ± 0.09 | 0.95 ± 0.05 | 0.92 ± 0.07 |
| Centrifuge | 0.82 ± 0.21 | 0.95 ± 0.05 | 0.86 ± 0.15 | 0.89 ± 0.13 | 0.90 ± 0.10 | 0.90 ± 0.11 |
| DIAMOND-LCA | 0.83 ± 0.15 | 0.72 ± 0.11 | 0.76 ± 0.09 | 0.87 ± 0.11 | 0.70 ± 0.10 | 0.77 ± 0.08 |
| Kaiju | 0.73 ± 0.23 | 0.95 ± 0.04 | 0.80 ± 0.17 | 0.87 ± 0.12 | 0.87 ± 0.09 | 0.87 ± 0.10 |
| mOTUs | 0.90 ± 0.11 | 0.92 ± 0.03 | 0.91 ± 0.07 | 0.91 ± 0.10 | 0.92 ± 0.04 | 0.91 ± 0.07 |
| MetaPhlAn | 0.92 ± 0.03 | 0.73 ± 0.08 | 0.81 ± 0.07 | 0.92 ± 0.03 | 0.73 ± 0.08 | 0.81 ± 0.05 |

### *Comparison of metagenomic classifiers on human gastrointestinal metagenomes*

Community profiles produced by MetaCache, Kraken, Bracken, mOTUs, and MetaPhlAn were compared to those obtained using the MCP on a set of 100 deidentified Australian fecal metagenomes. We focused on these classifiers as they were the strongest performing classifiers on the *in silico* mock communities and/or are widely used by the research community. Unlike the *in silico* mock community analysis, here each classifier was evaluated using its recommended reference database. MCP uses the MGDB which consists of 73,646 dereplicated genomes which span 28,246 species clusters (*see Properties of the Microba Genome Database*). MetaCache uses a reference database comprising 16,488 bacterial, 343 archaeal, and 8,999 viral genomes annotated as complete in RefSeq. Kraken and Bracken use a slightly expanded set of 18,871 bacterial, 360 archaeal, and 9,334 viral genomes along with a human reference genome and a collection of known vectors. mOTUs uses a pre-built database of marker genes obtained from ~25,000 bacterial and archaeal reference genomes which have been supplemented with additional marker genes obtained from public metagenomes. The MetaPhlAn database consists of ~1.5 million unique clade-specific marker genes obtained from ~100,000 bacterial, archaeal, and eukaryotic genomes. Species profiles for all classifiers are defined according to the NCBI Taxonomy (Federhen, 2015) with the exception of the MCP which uses the GTDB (Parks *et al*., 2020).

Since the community composition of the fecal samples are unknown, other measurable aspects of the community profiles produced by each metagenomic classifier were evaluated. The percentage of reads assigned to a species was substantially higher for the MCP (84.4% on average) than Kraken (49.0% on average), Bracken (58.8% on average), or MetaCache (50.5% on average; **Fig. 4A**). This was attributed to the large number of uncultured gut microbiome species represented in MGDB that are absent from the reference databases used by the other classifiers. By design, mOTUs and MetaPhlAn only classify the small subset of reads that map to marker genes and thus assessing percentage of mapped metagenomic reads is not a meaningful comparsion. As expected, Kraken, Bracken, and MetaCache report thousands of species (**Fig. 4B**), the vast majorityof which are likely low abundance false positives based on mock community results (**Fig. 1A** and **1B**). Consequently, species with an estimated abundance <0.01% were removed as these are expected to predominately be false positive predictions. The MCP reports the largest number of species with an abundance ≥0.01% (171.6 on average) followed by Bracken (164.6 on average), MetaCache (143.6 on average), and Kraken (136.7 on average). It is notable that mOTUs (131.9 on average) and MetaPhlAn (70.6 on average) report the fewest species in these samples, but were observed to produce far fewer false positives than Kraken, Bracken, and MetaCache on the *in silico* mock communities (**Fig. 2D**). This suggests that these latter classifiers may only be reporting greater numbers of species than mOTUs and MetaPhlAn as a result of increased numbers of false positive predictions.



uMCP = unfiltered MCP; G = Ganon; K = Kraken; B = Bracken; M = MetaCache

**Figure 3.** Performance of metagenomic classifiers to predict species abundances. (**A**) L1 distance (0% = identical to ground truth; 200% = no species in common with ground truth) between the ground truth and predicted species profiles. (**B**) Mean relative error of species present in both the ground truth and predicted species profiles. (**C**) Sum of false positive species abundances. (**D**) Sum of false negative species abundances. Lower values indicate better performance. Results are provided across mock communities of increasing ANI divergence from the reference database. Results for Centrifuge, DIAMOND-LCA, Kaiju, and mOTUs are not shown as they have substantially worse species abundance estimates than the other classifiers (**Supp. Fig. 2**).

**Table 5.** Performance statistics for species abundance estimates across the 140 mock communities (mean ± std. dev.).

| Classifier | L1 Distance | Relative Error (%) | Abundance of FPs (%) | Abundance of FNs (%) |
|---|---|---|---|---|
| MCP | 10.8 ± 10.47 | 11.2 ± 10.97 | 0.18 ± 0.45 | 0.20 ± 0.25 |
| Unfiltered MCP | 11.3 ± 10.98 | 11.9 ± 11.92 | 0.73 ± 0.96 | 0.08 ± 0.12 |
| MCP w/ MGDB | 10.0 ± 7.60 | 10.1 ± 6.69 | 0.27 ± 0.82 | 0.31 ± 0.78 |
| Ganon | 11.6 ± 9.90 | 10.8 ± 11.32 | 1.72 ± 2.32 | 0.03 ± 0.05 |
| Kraken | 14.4 ± 11.39 | 12.9 ± 13.53 | 3.42 ± 3.89 | 0.06 ± 0.35 |
| Bracken | 11.1 ± 10.33 | 10.5 ± 13.50 | 3.92 ± 4.53 | 0.05 ± 0.35 |
| MetaCache | 9.0 ± 8.05 | 10.0 ± 11.46 | 1.97 ± 3.00 | 0.02 ± 0.03 |
| Centrifuge | 49.0 ± 22.18 | 52.7 ± 24.92 | 2.97 ± 5.73 | 0.28 ± 0.32 |
| DIAMOND-LCA | 78.1 ± 9.54 | 68.6 ± 7.27 | 0.54 ± 0.53 | 3.05 ± 2.41 |
| Kaiju | 42.5 ± 13.48 | 34.6 ± 11.20 | 2.19 ± 2.59 | 0.20 ± 0.19 |
| mOTUs | 18.26 ± 11.86 | 37.2 ± 34.86 | 4.79 ± 5.76 | 4.58 ± 4.69 |
| MetaPhlAn | 43.3 ± 19.97 | 37.9 ± 40.37 | 3.58 ± 3.21 | 16.63 ± 10.82 |

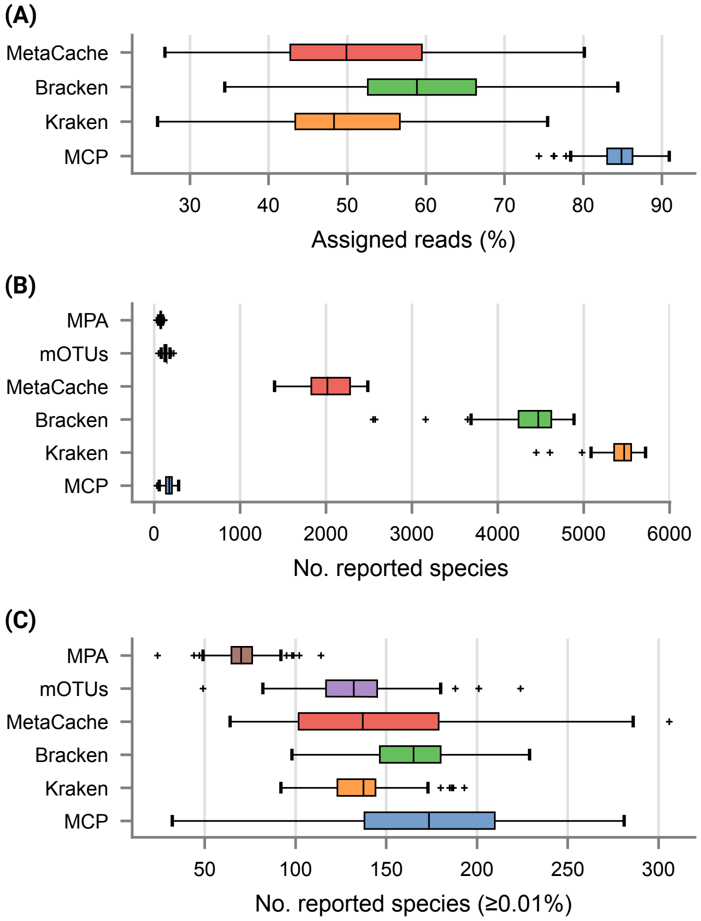


**Figure 4.** Comparison of metagenomic classifiers on 100 deidentified Australian fecal samples. Community profiles were produced by each classifier using their recommended reference database. (**A**) Percentage of reads assigned to a species in community profiles. (**B**) Number of species reported by each classifier. (**C**) Number of species reported by each classifier with an estimated abundance ≥0.01%.

### *Properties of the Microba Genome Database*

The Microba Genome Database (MGDB), the default reference database for the MCP, consists of 73,646 dereplicated genomes from 28,246 species clusters as defined by the Genome Taxonomy Database (GTDB; Parks *et al.*, 2019; Parks *et al.*, 2020). The 73,646 genomes in the MGDB were selected in order to provide comprehensive coverage of the genomic diversity within each species and with a specific focus on the human gastrointestinal tract. These genomes were obtained from a variety of sources including the NCBI Assembly database (52.4%), recent large-scale efforts to recover human gastrointestinal MAGs (35.7%; Almeida *et al.*, 2019; Nayfach *et al.*, 2019; Pasolli *et al.*, 2019 or isolates (1.2%; Forster *et al.*, 2019; Zou *et al.*, 2019), and Microba's own initiatives to obtain MAGs from customer samples (7.6%) and public metagenomes (3.2%; **Fig. 5A**). The 73,646 MGDB genomes are predominately MAGs (66.3%; **Fig. 5A**) in agreement with a recent estimate that ~70% of microbial species in the human gastrointestinal tract remain to be cultured (Almeida *et al.*, 2020). These MAGs have an average completeness of 89.5 ± 10.0% and contamination of 1.34 ± 1.48% with ~60% meeting the completeness and contamination criteria used to define high-quality MAGs (Bowers *et al.*, 2017).

Nearly 50% (13,673) of the 28,246 species in the MGDB are comprised solely of uncultured genomes (i.e. MAGs or single-amplified genomes) with 625 species being comprised exclusively of MAGs obtained by Microba (**Fig. 5B**), which is reflected in their taxonomic assignments. Only 36.6% of the 28,246 species clusters in the MGDB have a species assignment in the NCBI Taxonomy

(**Fig. 5B**). For this reason, the MGDB and by extension the MCP uses the GTDB as a taxonomic resource as it provides a substantial improvement in taxonomic resolution with 87.5% of the MGDB species having a GTDB species assignment. Furthermore, adoption of the quantitative criteria used by the GTDB to circumscribe taxa allowed the 625 species exclusive to Microba to be readily identified and given temporary placeholder names with appropriate higher taxonomic ranks as determined by the GTDB-Tk (Chaumeil *et al*., 2019). The lack of taxonomic resolution in the NCBI Taxonomy extends beyond the rank of species with only 56.8% and 62.5% of MGDB species clusters having an NCBI genus or family assignment, respectively. In contrast, 97.9% and 99.8% of MGDB species clusters have GTDB genus or family assignments, respectively.

## *Performance of MCP with the MGDB on mocks and gastrointestinal samples*

On *in silico* mock communities the MCP generally performs better using the more comprehensive MGDB than the relatively small standardized reference database. In particular, use of the MGDB results in an improvement in correct identification of species comprising the mock communities and in the accuracy of species abundance estimates (**Tables 4** and **5**). The proportion of the community resulting from false positive (0.18 to 0.27%) or false negative (0.20 to 0.31%) predictions increases slightly with the use of the MGDB (**Table 5**). We attribute this to challenges inherent in robustly distinguishing between highly similar species which are more prevalent in the MGDB (28,246 species) compared to the standardized reference database (12,250 species). Low levels of contamination in MAGs within the MGDB may also contribute to the small increase in false positives.

As the MGDB is comprised of a large number of MAGs and isolates specific to the human gastrointestinal microbiome, we expect the benefits of the MGDB to be more pronounced on samples from this habitat than on the *in silico* mock communities. To illustrate this, we examined the species profiles produced by the MCP using the MGDB

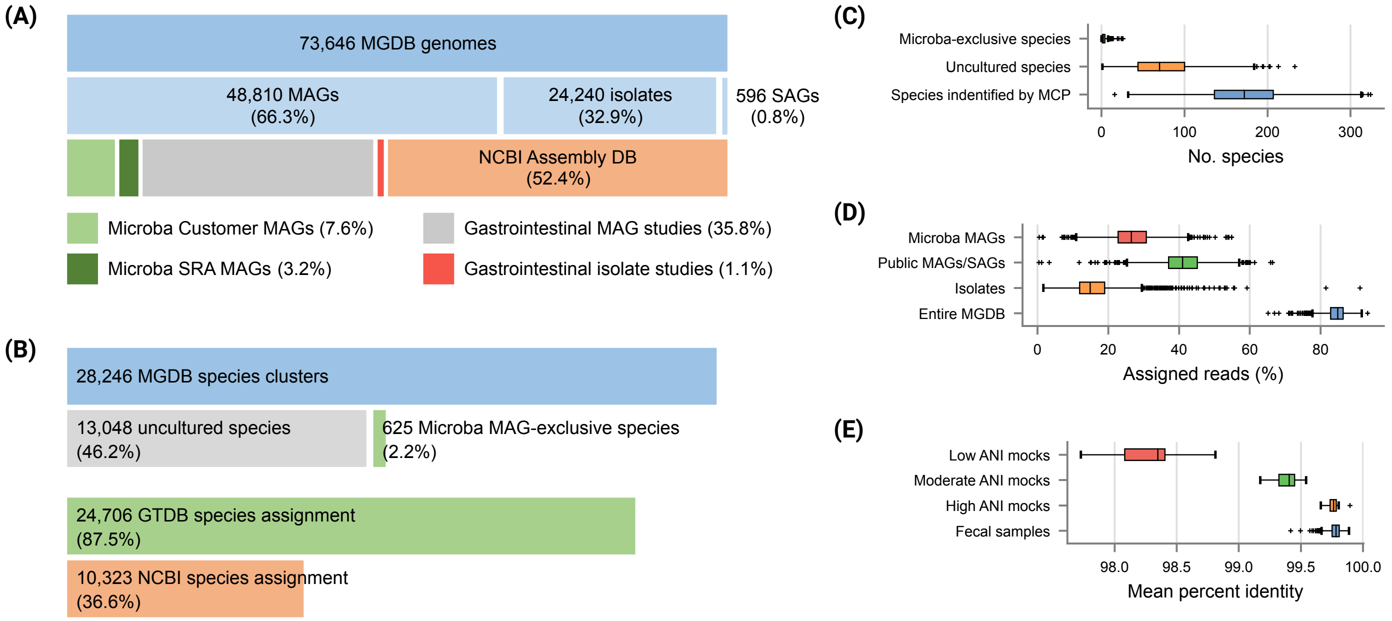


**Figure 5.** Properties of the Microba Genome Database (MGDB) and the results of profiling 2,000 deidentified Australian fecal samples using the MCP with the MGDB as a reference database. (**A**) Proportion of MAGs, isolates, and SAGs within the 73,646 genomes comprising the MGDB along with the source of these genome assemblies. Gastrointestinal MAGs were recovered from the studies of Almeida *et al*., 2019; Nayfach *et al*., 2019; and Pasolli *et al*., 2019 and gastrointestinal isolates from the studies of Forster *et al*., 2019; Zou *et al*., 2019. (**B**) Proportion of the 28,246 MGDB species clusters comprised exclusively of uncultured genomes (i.e. MAGs or single-amplified genomes) obtained from multiple sources or solely of MAGs recovered by Microba. This is followed by the proportion of MGDB species clusters that can be assigned a GTDB or NCBI species assignment. (**C**) Total number of species reported by the MCP for each sample and the number of these species which are uncultured species or Microba MAG-exclusive species. (**D**) Total percentage of reads assigned by the MCP to genomes in the MGDB and the percentage assigned to isolates, public MAGs/SAGs, or MAGs obtained by Microba. (**E**) Percent identity of reads mapped by the MCP for 2,000 Australian fecal samples and *in silico* mock communities with decreasing ANI similarity to genomes in the standardized reference database.

on 2,000 deidentified Australian fecal samples. The MCP reports an average of 171.8 ± 51.6 species per sample with 84.3 ± 3.1% of reads being mapped to a species in the MGDB (**Figs. 5C** and **5D**). The 7,950 unique MAGs obtained by Microba account for >10% of the genomes comprising the MGDB (**Fig. 5A**) and capture genomic variation within species not accounted for by publicly-available genomes. This is illustrated by MCP mapping reads to 6,035 ± 1,669 genomes on average and 1,534 ± 472 of these being to Microba recovered MAGs. Notably, 26.8 ± 6.6% of reads have a best mapping to a MAG obtained by Microba (**Fig. 5D**) and 34.3% of samples contain a Microba MAG which accounts for ≥5% of the mapped reads.This highlights the benefits of obtaining MAGs from the samples being profiled in order to build a reference database with strains specific to the habitat being studied.

We assessed the similarity of strains found in the human gastrointestinal tract to genomes comprising the MGDB by considering the percent identity (PI) and percent alignment length (PA) of reads mapped by the MCP. Mapped reads had a PI and PA of 99.78% and 99.99%, respectively, averaged over the 2,000 Australian fecal samples. Comparing these similarity values to the PI and PA observed for the *in silico* mock communities with known ANI to reference genomes suggests strains found in the human gastrointestinal tract generally have high ANI (i.e. >99%) to MGDB reference genomes (**Fig. 5E; Supp. Table 6**), indicating it is a comprehensive database for fecal microbiome profiling.

## Discussion

The Microba Community Profiler was developed to provide accurate metagenomic profiles of fecal microbiomes. Here we evaluated the performance of the MCP relative to nine metagenomic classifiers that are widely used and/or have been shown to be among the best performing classifiers (Seppey *et al*., 2020; Ye *et al*., 2019; Lindgreen *et al*., 2016; Sczyrba *et al*., 2017). Benchmarking was performed using 140 *in silico* mock communities with decreasing ANI similarity to genomes in a standardized reference database. To the best of our knowledge, this is the first benchmarking study to explicitly investigate the impact of genomic similarity to reference database genomes on classification performance. Our results demonstrate that the MCP has the highest combined precision and recall (i.e. $F_1$ score) among all evaluated classifiers indicating that the optimized trade-off between false positive and false negative predictions used by MCP provides the most accurate community profiles (**Fig. 2**). The strong performance of the MCP was observed across all mock communities indicating it can reliably identify species even when strains are up to 5% divergent at the nucleotide level from genomes in the reference database. This is in contrast to the other evaluated classifiers which showed a substantial reduction in performance on mock communities with low similarity to genomes in the standardized reference database (**Fig. 2C**).

MCP, Kraken, Bracken, and Ganon all provide reasonably accurate estimates of the relative abundance of species in moderate and high ANI mock communities with MetaCache showing slightly better performance (**Figs. 3A** and **3B**). An advantage of MCP is a smaller portion of false positive predictions (**Fig. 3C**) giving researchers confidence in the predicted community profile. All classifiers failed to provide accurate estimates of the abundance of species on the low ANI mock communities (**Figs. 3A** and **3B**) with the standard reference database. While this limitation warrants further investigation to improve classifier performance, inspection of community profiles of fecal samples produced by the MCP when using the MGDB as a reference database suggests that strains found in the human gastrointestinal tract typically have high ANI similarity to MGDB reference genomes (**Fig. 5E**). This is encouraging as the mock community results suggest that low abundance species (<0.01%) can be identified by the MCP with a low false discovery rate when using a reference database containing closely related strains (**Fig. 1C; Table 3**).

Our benchmarking analysis follows the recommendation that classifiers be evaluated independently of their reference database (Ye *et al*., 2019) as the specific composition of databases can have a considerable impact on classification performance (Nasko *et al*., 2018; Méric *et al*., 2019). This is evident from the higher number of reads from human fecal samples that were classified by MCP compared to MetaCache, Kraken, and Bracken using the default reference databases of each classifier (**Fig. 4A**). We attribute the substantially higher percentage of reads classified by MCP, in part, to the use of a more comprehensive human gut microbiome database (**Fig. 5A**), consistent with previous studies showing the benefit of including human gastrointestinal MAGs (Pasolli *et al*., 2019; Almeida *et al*., 2019; Nayfach *et al*., 2019). As the recovery of MAGs is outpacing our ability to culture new species, it is critical for metagenomic classifiers to make use of this additional source of information, including taxonomic frameworks that accommodate uncultivated species, such as GTDB (Parks *et al*., 2020; **Fig. 5B**).

The majority of evaluated classifiers provide only a partial solution to the goal of establishing which species are present within a community. This is exemplified by the large number of false positives reported by Ganon, Kraken, Bracken, MetaCache, DIAMOND-LCA, and Kaiju (**Fig. 5D**). Ultimately, these classifiers require researchers to investigate the resulting profiles to establish suitable criteria for establishing which species are likely true positives (Ye *et al*., 2019). This is in contrast to the MCP, mOTUs, and MetaPhlAn which explicitly aim to produce community profiles comprised solely of true positive predictions, without user input.

MCP is under ongoing development and MGDB is constantly updated with genomes of newly identified species. Current efforts are focused on improving the accuracy of species abundance estimates by expanding the genomic diversity of gut species captured by the MGDB and exploring if unclassified reads can be assigned to species without increasing false positive predictions. Future improvements to the detection limit of MCP include identifying and removing contamination in reference genomes which can result in low abundance false positive predictions. While there are opportunities to continue improving the performance of MCP, the results of this study illustrate that MCP is the best overall classifier.

## Material and Methods

### *Standardized reference database for classifiers*

A reference database of 15,555 genomes from 12,250 species was constructed from RefSeq release 97 (Kitts *et al*., 2016) obtained from NCBI on 22 November 2019 for use by all metagenomic classifiers (**Supp. Table 1**). Only isolate genomes estimated to be >90% complete with <5% contamination by CheckM v1.0.13 (Parks *et al*., 2015 and where the assembly meet the following criteria were considered for inclusion in the database: i) <500 contigs, ii) N50 >20kb, and iii) <10,000 undetermined bases. In addition, only genomes with species designations forming a 1-to-1 mapping between the GTDB R04-RS89 (Parks *et al*., 2018) and NCBI (Federhen 2015; downloaded 22 November 2019) taxonomies were considered to help ensure reference genomes had correct species assignments. This limited the genomes selected for the reference database to those in GTDB R04-RS89 (based on RefSeq release 89), in order to allow recently submitted genomes to be used for generating *in silico* mock communities. A maximum of 5 genomes were selected for each species in order of assembly quality as defined by Q = completeness – 5×contamination – 0.05* (no. contigs) – 0.00005* (no. undetermined bases), with an additional 100 added to the assembly quality if it was annotated as complete as determined by consulting the 'assembly level' annotation at NCBI. In order to avoid having highly similar genomes in the reference database, a genome was only included if it had an ANI <99% to all other intraspecific genomes as determined with Mash v2.1.1 (Ondov *et al*., 2016). The reference database contains 10,776 species with exactly 1 genome and 1,474 species represented by >1 genome, and these species have an average intraspecific ANI of 97.8 ± 0.96% as determined with FastANI v1.3 (Jain *et al*., 2018).

### *Generation of in silico mock communities*

*In silico* mock communities were constructed from RefSeq release 97 genomes which passed the same filtering criteria used for the standardized reference database, including the requirement of a 1-to-1 mapping between GTDB and NCBI species assignments (*see above*). The 67,299 genomes in RefSeq release 97 not covered by GTDB R04-R89 were assigned GTDB classifications using GTDB-Tk v0.3.3 (Chaumeil *et al*., 2019). Intraspecific ANI values between reference database genomes and potential mock community genomes were calculated with FastANI v1.3. These ANI values were used to generate mock communities comprised of genomes which were increasingly divergent from those in the standardized reference database at ANI intervals of [99%, 99.75%], [97%, 99%), and [95%, 97%) (**Table 2**). In addition, mock communities comprised of genomes in the reference database (ANI = 100%) were considered as these provide a useful point of comparison.

The number of species in a mock community was modeled on a normal distribution with μ (mean number of species) =100 and σ (standard deviation in number of species) =25, or μ=500 and σ=100, in order to generate medium and high complexity communities, respectively. Communities were constructed with either a single genome selected from each species, or with 2 to 10 genomes randomly selected from each species. The relative abundance of genomes comprising mock communities were drawn from a log-normal distribution with a mean of 1 and a standard deviation of 2 as commonly used for modelling microbial communities (Curtis *et al*., 2002; Fritz *et al*., 2019).

The number of paired reads generated for each genome was $n_i = N \times (a_i \;\; s_i \,/\, \sum_j a_j \;\; s_j)$, where $s_i$ is the size of genome *i*, $a_i$ is the relative abundance of genome *i*, and *N* is the total number of paired reads comprising the *in silico* community. All *in silico* communities were simulated to a depth of 2.1 Gb by randomly sampling 2×150 bp paired-end reads with an insert size of 200 ± 25 bp across each genome in the mock community.

### *Building custom databases for metagenomic classifiers*

The genomes comprising the standardized reference database were used to build a custom database for each classifier using recommended default parameters. Genomes comprising the standardized reference database were contained in individual FASTA files in a single directory (db_genomes) and concatenated into a single FASTA file (db_genomes_all.fna) in order to facilitate the requirements of the different metagenomic classifiers. The custom databases were built using the same NCBI Taxonomy data files used while constructing the standardized reference database which were obtained from NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy) on 22 November 2019 and consist of the files nodes.dmp, names.dmp, merged.dmp, nucl_gb.accession2taxid, and nucl_wgs accession2taxid. DIAMOND and Kaiju require protein sequences which were called for each reference genomes using Prodigal v2.6.3 (Hyatt *et al*., 2010) and the translation table specified at NCBI: prodigal -c -m -q -f gff -p single -g <trans_table> -i <ref_genome> -a <aa_output>. Prodigal was used to predict protein sequences as NCBI does not provide protein sequences for all genomes comprising the standardized reference database. A mapping file indicating the NCBI species ID for each predicted protein (db_proteins_all.taxid_map.tsv) and a FASTA file containing all proteins (db_proteins_all.faa) were created to facilitate building the DIAMOND and Kaiju databases. The commands executed to build custom databases for each classifier are given in **Table 6**.

**Table 6.** Commands for building custom databases for each of the metagenomic classifiers.

| Classifier | Command(s) |
|---|---|
| Kraken | NCBI Taxonomy data files were placed in kraken2_db/taxonomy and the database built with:<br>> kraken2-build --threads 4 --add-to-library db_genomes_all.fna --db kraken2_db<br>> kraken2-build --threads 64 --build --db kraken2_db |
| Bracken | Built from Kraken 2 database using:<br>> bracken-build -d kraken2_db -t 60 -k 35 -l 150 |
| Centrifuge | > centrifuge-build -p 96 --conversion-table nucl_wgs_gb.accession2taxid --name-table names.dmp --taxonomy-tree nodes.dmp db_genomes_all.fna centrifuge_db |
| Ganon | > ganon build -d ganon_db --input-files db_genomes_all.fna --taxdump-file nodes.dmp names.dmp merged.dmp -t 48 |
| DIAMOND-LCA | > diamond makedb -p 40 --db db_proteins_all.faa --in db_proteins_all.faa --taxonmap db_proteins_all.taxid_map.tsv --taxonnodes nodes.dmp |
| Kaiju | Sequence headers in db_proteins_all.faa were formatted to contain NCBI TaxIds and the database built with:<br>> kaiju-mkbwt -n 20 -o kaiju_db db_proteins_all.faa<br>> kaiju-mkfmi kaiju_db |
| MetaCache | NCBI Taxonomy data files were placed in the directory ncbi_taxonomy and the database built with:<br>> metacache build metacache_db db_genomes -taxonomy ncbi_taxonomy |
| mOTUs | A file, genomes.list, containing the accession of all standardized database genomes was created along with a mOTUs 2 formatted taxonomy file, taxonomy_file.txt. Since mOTUs 2 only supports extending an existing database, empty mOTU 2 data files were created in the directory clean_db. The database was then built with:<br>> parallel -j 64 -a genomes.list "extend_mOTUs_addGenome.sh db_genomes/{}.fasta {} standard_db extend_mOTUs_DB/SCRIPTS/clean_db"<br>> extend_mOTUs_generateDB.sh genomes.list STANDARD_DB taxonomy_file.txt standard_db extend_mOTUs_DB/SCRIPTS/clean_db |
| MetaPhlAn | MetaPhlAn 2 was run using the marker dataset v296_CHOCOPhlAn_201901 downloaded on Feb. 25, 2020. |

### *Species-level community profiling with metagenomic classifiers*

Community profiles were generated for mock communities using each of the metagenomic classifiers run with default parameters (**Table 7**). DIAMOND indicates the lowest common ancestor (LCA) for each query read, but does not produce a profile indicating the proportion of reads assigned to each species. A custom script was used to tabulate the proportion of reads assigned to each species. Reads with an LCA above the rank of species were considered unclassified for the purposes of creating a species profile for each mock community.

MetaPhlAn results were obtained using the v296_CHOCOPhlAn_201901 marker set which may have species assignments that differ from those defined for the *in silico* mock communities due to reclassifications at NCBI. To account for this, the NCBI TaxIds produced by MetaPhlAn were used to establish species names as defined in the 22 November 2019 NCBI Taxonomy data files, the same files used to construct the mock communities.

### *Microba Genome Database*

The Microba Genome Database (MGDB) v2 was built from genomes in GTDB R04-RS89, MAGs obtained from Australian fecal samples, MAGs mined from SRA samples by Microba, and MAGs and isolate genomes from the Almeida *et al*. (2019), Forster *et al*. (2019), Nayfach *et al*. (2019), Pasolli *et al*. (2019), and Zou *et al*. (2019). Together these sources span 411,415 genomes after removing lower quality assemblies as defined by having a completeness estimate <80%, a contamination estimate >5%, being comprised of >1000 contigs, or having an N50 <5Kb. These genomes were dereplicated based on ANI similarity to obtain a final database consisting of 73,646 genomes from 28,246 species. Completeness and contamination estimates for genomes within the MGDB were determined using CheckM v1.1.2 (Parks *et al*., 2015). Genomes without taxonomic assignments in GTDB R04-RS89 were assigned a GTDB classification using GTDB-Tk v0.3.3 (Chaumeil *et al*., 2019) and additional species clusters defined using the ANI criteria used by the GTDB (Parks *et al*., 2020).

### *Classifier performance metrics*

Precision and recall can be defined in terms of the number of species correctly (true positives; *TP*) and incorrectly (false positive; *FP*) identified by a classifier along with the number of unidentified species present in a sample (false negative; *FN*). Precision, $P=TP/(TP+FP)$, is the fraction of species identified by a classifier that are correct, while recall, $R=TP/(TP+FN)$, is the fraction of correctly identified species within a sample. The $F_1$ score is the harmonic mean of precision and recall, $(2 \times P \times R)/(P+R)$, which weights these terms equally in a single metric.

Absolute and relative percent error for each species within a sample are defined in terms of the true, *T*, and estimated, *E*, abundance of a species. Absolute error, $A=|T-E|$, indicates how close abundances estimates are to the true abundance of a species, while relative percent error, $R=100 \times A/T$, expresses how large the absolute error is compared to the true abundance which highlights poor estimates of low abundance species. The L1 (Manhattan) distance is the sum of absolute errors across all ground

**Table 7.** Commands for profiling mock communities with each of the metagenomic classifiers.

| Classifier | Command(s) |
|---|---|
| Kraken | > kraken2 --db {kraken2_db} --report {sample_id}.kreport2 --output {sample_id} --paired {left_reads} {right_reads} |
| Bracken | > est_abundance.py -i {sample_id}.kreport2 -k {bracken2_db} -o {sample_id}.bracken |
| Centrifuge | > centrifuge -x {centrifuge_db} -1 {left_reads} -2 {right_reads} --report-file {profile_file} -S {per_read_file} |
| Ganon | > ganon classify -d {ganon_db} -p {left_reads} {right_reads} -o {profile_file} |
| DIAMOND-LCA | > diamond blastx --query-gencode 11 -f 102 --top 10 --min-score 50 -d {diamond_db} -q {left_reads} -o {profile_file} |
| Kaiju | > kaiju -v -t {ncbi_nodes} -f {kaiju_db} -i {left_reads} -j {right_reads} -o {report}<br>> kaiju2table -v -t {ncbi_nodes} -n {ncbi_names} -r species -o { profile_file} {report} |
| MetaCache | > metacache query {metacache_db} {left_reads} {right_reads} -pairfiles -out {report} -abundances {profile_file} -abundance-per species |
| mOTUs | > motus profile -f {left_reads} -r {right_reads} -db {motu_db} -o {profile_file} |
| MetaPhlAn | > metaphlan2.py {left_reads},{right_reads} {profile_file} --bowtie2out {sample_id}.bowtie2.bz2 --index v296_CHOCOPhlAn_201901 --ignore_eukaryotes --input_type fastq |

truth *and* predicted species which provides a measure that incorporates false positive predictions (Ye *et al*., 2019). The mean relative percent error across all ground truth species in a sample was used as for assessing classifier performance. Different ground truth abundances were used for classifiers that estimate i) the relative proportion of reads from each species (Ganon, Kraken, Bracken, MetaCache, DIAMOND-LCA, Kaiju) and ii) the relative proportion of reads normalized by genome size (MCP, Centrifuge, mOTUs, MetaPhlAn).

Previous benchmarking studies have suggested the use of the Euclidean distance and the area under the precision-recall curve (AUPR) for evaluating classifier performance (Ye *et al*., 2019). We elected to use the L1 distance as it does not give additional weight to high abundance species and report precision and recall independently as the AUPR is known to be biased toward low-precision, high-recall classifiers (Ye *et al*., 2019). This is a notable limitation as many classifiers fall into this categorization.

### *Establishing classifier detection limits*

The detection limit for classifiers was defined as the lowest abundance species in a sample that achieved a specified false discovery rates, *FDR=FP/(TP+FP)*. This was determined by ordering identified species in ascending order of abundance and calculating the FDR after filtering species below each abundance level. The detection limit for a sample is the lowest abundance at which the desired FDR could be achieved.

### *Community profiles for human gastrointestinal metagenomes*

Community profiles for deidentified Australian fecal samples were produced for selected metagenomic classifiers using recommended reference databases. Reference databases for MetaCache and Kraken were obtained using the scripts and recommended parameters suggested by these classifiers (**Tables 6** and **7**). These databases were built on March 3, 2020. Kraken v2.0.8 was used for this analysis as opposed to v2.0.7 as changes to NCBI data formats required the use of this later version. MetaCache and Kraken differ in the set of included reference genomes as MetaCache only considered genomes annotated as a “Complete Genome” at NCBI while Kraken also includes genomes annotated as “Chromosome”. Bracken results are derived from the mapping information produced by Kraken. mOTUs and MetaPhlAn results were obtained using pre-built marker databases. Profiling was performed as previous described (**Table 8**).

**Table 8.** Commands for building recommended reference databases.

| Classifier | Command(s) |
|---|---|
| Kraken | > kraken2-build --standard --db standard_db |
| Bracken | Results derived from information produced by Kraken |
| MetaCache | > metacache-build-refseq |
| mOTUs | Marker database is pre-built and provided with software |
| MetaPhlAn | MetaPhlAn 2 was run using the marker dataset v296_CHOCOPhlAn_201901 |

## Data availability

The *in silico* paired-end reads and ground truth data for the 140 mock communities are available upon request. Genomes used to build the standardized reference database are given in **Supp. Table 1** and can be obtained from the NCBI Assembly database (Kitts *et al*., 2016). The metagenomic classifiers can be obtained from their respective websites as indicated in the cited literature with the exception of MCP which is proprietary software developed by Microba Life Sciences Limited.

## References

Almeida A, *et al*. (2019). A new genomic blueprint of the human gut microbiota. *Nature* **568**: 499-504.

Almeida A, *et al*. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*, doi: https://doi.org/10.1038/s41587-020-0603-3.

Bolyen E, *et al*. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852-857.

Bowers RM, *et al*. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725-31.

Breitwieser FP, *et al*. (2019). A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* **20**: 1125-1136.

Burrows M, Wheeler DJ. (1994). A block-sorting lossless data compression algorithm. *Technical report* 124. Palo Alto, CA: Digital Equipment Corporation.

Buchfink B, Xie C, Huson DH. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Bundgaard-Nielsen C, *et al*. (2018). Interpersonal variations in gut microbiota profiles supersedes the effects of differing fecal storage conditions. *Sci Rep* **8**: 17367.

Case RJ, *et al*. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* **73**: 278-288.

Chaumeil PA, *et al*. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics btz848: doi: 10.1093/*bioinformatics*/**btz848**.

Clarridge JE. (2004). Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clin Microbiol Rev* **17**: 840-862.

Costello EK, *et al*. (2013). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694-1697.

Curtis TP, *et al*. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci* **99**: 10494–9.

Eloe-Fadrosh EA, *et al*. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* **1**: 15032.

Engelbrektson A, *et al*. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**: 642-7.

Epstein SS. The phenomenon of microbial uncultivability. *Curr Opin Microbiol* **16**: 636-42.

Evans PN, *et al*. (2019). An evolving view of methan metabolism in the Archaea. *Nat. Rev. Microbiol.* **17**: 219-232.

Federhen S. (2015). Type material in the NCBI Taxonomy Database. *Nucleic Acids Res* **43**: D1086-98.

Forster SC, *et al*. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* **37**: 186-192.

Fritz A, *et al*. (2019). CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**:17.

Gentile CL and Weir TL. The gut microbiota at the intersection of diet and human health. *Science* **362**: 776-780.

Greenblum S, Turnbaugh PJ, Borenstein E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* **109**: 594-9.

Hugenholtz P and Tyson GW. (2008). Metagenomics. *Nature* **455**: 481-483.

Huttenhower C, *et al*. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-14.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. B*MC Bioinformatics* **11**: 119.

Jain C, *et al*. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 5114.

Janda JM and Abbott SL. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* **45**: 2761-2764.

Johnson JS, *et al*. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* **10**: 5029.

Jovel J, *et al*. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol* **7**: 459.

Karst SM, *et al*. (2018). Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat Biotechnol* **36**: 190-195.

Kennedy AC and Smith KL. (1995). Soil microbial diversity and the sustainability of agricultural soil. *Plant soil* **170**:75–86.

Kim D, *et al*. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26**: 1721-1729.

Kitts PA, *et al*. (2016). Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* **44**: D73-80.

Köser CU, *et al*. (2014). Whole-genome sequencing to control antimicrobial resistance. *Trends Genet* **30**: 401-7.

Kuypers MMM, *et al*. (2018). The microbial nitrogen-cycling network. *Nat Rev Microbiol* **16**: 263-276.

Laforest-Lapointe I and Arrieta MC. (2018). Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems* **13**: e00201-17.

Lander and Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231–239.

Li H and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**:1754-60.

Li H, *et al*. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-9.

Lindgreen S, *et al*. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep* **6**: 19233.

Lloyd KG, *et al*. (2019). Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *mSystems* **3**: e00055-18.

Lloyd-Price J, *et al*. (2016). The health human microbiome. *Genome Med*. **8**: 51.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**: e104.

McIntyre ABR, *et al*. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* **18**: 182.

Menzel P, *et al*. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**: 11257.

Méric G, *et al*. (2019). Correcting index databases improves metagenomic studies. *bioRxiv*: https://doi.org/10.1101/712166.

Milanese A, *et al*. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**: 1014.

Mitra A, *et al*. (2014). The role of mixotrophic protists in the biological carbon pump. *Biogeosciences* **11**: 995–1005.

Müller A, *et al*. (2017). MetaCache: context-aware classification of metagenomic reads using minhashing. *Bioinformatics* **33**: 3740-3748.

Nasko DJ, *et al*. (2018). RefSeq database growth influences the 696 accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol* **19**:165.

Nayfach S, *et al*. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**: 505-510.

Nilsson RH, *et al*. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* **47**: D259-D264.

Ondov BD, *et al*. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**: 132.

Orellana LH, *et al*. (2018). Year-Round Shotgun Metagenomes Reveal Stable Microbial Communities in Agricultural Soils and Novel Ammonia Oxidizers Responding to Fertilization. *Appl Environ Microbiol* **84**: e01646-17.

Parks DH, *et al*. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043-55.

Parks DH, *et al*. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**: 996-1004.

Parks DH, *et al*., (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* doi: 10.1038/s41587-020-0501-8.

Parfrey LW, *et al*. (2011). Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front Microbiol* **2**: 153.

Pasolli E, *et al*. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**: 649-662.

Piro VC, *et al*. (2019). ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *bioRxiv* doi: https://doi.org/10.1101/406017.

Quast C, *et al*. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl Acids Res* **41**: D590-D596.

Quince C, *et al*. (2017). Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**: 833-844.

Rohwer F, *et al*. (2009). Roles of viruses in the environment. *Environ Microbiol* **11**: 2771-4.

Scarpellini E, *et al*. (2015). The human gut microbiota and virome: Potential therapeutic implications. *Dig Liver Dis* **47**: 1007-12.

Schloss PD, *et al*. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537-41.

Schoch CL, *et al*. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A* **109**: 6241-6.

Schulz F, *et al*. (2017). Towards a balanced view of the bacterial tree of life. *Microbiome* **5**: 140.

Sczyrba A, *et al*. (2017). Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods* **14**: 1063-1071.

Segata N, *et al*. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**: 811-4.

Seppey M, *et al*. (2020). LEMMI: a continuous benchmarking platform for metagenomics classifiers. *Genome Res* **30**: 1208-1216.

Truong DT, *et al*. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**: 902-903.

Weinstock GM. (2012). Genomic approaches to studying the human microbiota. *Nature* **489**: 250-6.

Woese CR, *et al*. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* **87**: 4576-9.

Wood DE, *et al*. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**: 257.

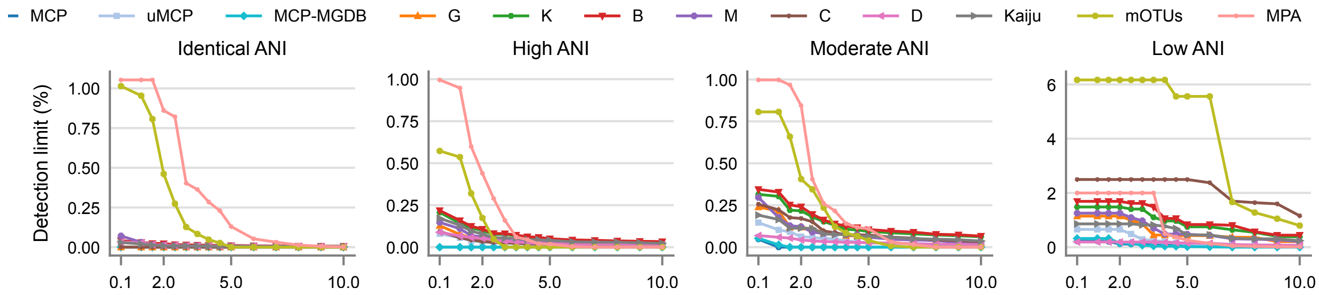
Yarze P, *et al*. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Net Rev Microbiol* **12**: 635-45.

Ye SH, *et al*. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**: 779-794.

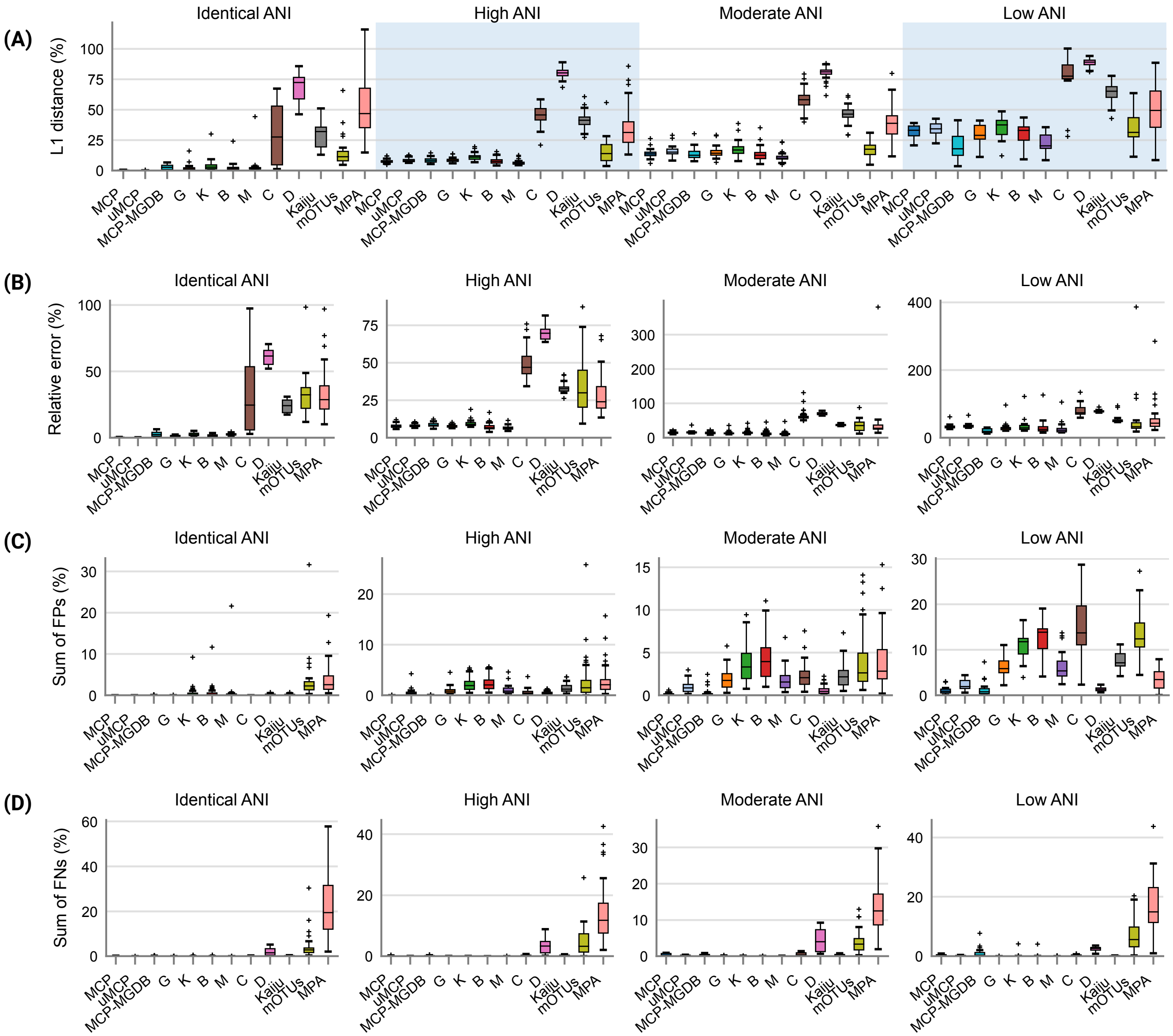
Zmora N, *et al*. (2019). You are what you eat: diet, health and the gut microbiota. *Nat Rev Gastroenterol Hepatol.* **16**: 35-56.

Zou Y, *et al*. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol.* **37**, 179-185.

## Supplemental Figures



**Supp. Fig. 1.** Median detection limit of each classifier, including mOTUs and MPA, over all mock communities at a given level of ANI similarity to the reference database for varying false discovery rates. uMCP = unfiltered MCP; G = Ganon; K = Kraken; B = Bracken; M = MetaCache; C = Centrifuge; D = DIAMOND-LCA, MPA = MetaPhlAn.



uMCP = unfiltered MCP; G = Ganon; K = Kraken; B = Bracken; M = MetaCache; C = Centrifuge; D = DIAMOND-LCA, MPA = MetaPhlAn

**Supp. Fig. 2.** Performance of metagenomic classifiers to predict species abundances. (**A**) L1 distance (0% = identical to ground truth; 200% = no species in common with ground truth) between the ground truth and predicted species profiles. (**B**) Mean relative error of species present in both the ground truth and predicted species profiles. (**C**) Sum of false positive species abundances. (**D**) Sum of false negative species abundances. Identical to **Figure 3** except results for Centrifuge, DIAMOND-LCA, Kaiju, mOTUs, and MetaPhlAn are included.

## Supplemental Tables

**Supp. Table 1.** Metadata for the 15,555 isolate genomes comprising the standardized reference database (see Excel file).

**Supp. Table 2.** Metadata for the 140 *in silico* mock communities (see Excel file).

**Supp. Table 3.** Number of true positive (TP) and false positive (FP) species predictions along with the false discovery rate (FDR) of metagenomic classifiers on mock communities without filtering of low abundance species and with varying ANI to reference database genomes (mean ± std. dev.).

| | High ANI | | | Moderate ANI | | | Low ANI | | |
|---|---|---|---|---|---|---|---|---|---|
| Classifier | TP | FP | FDR | TP | FP | FDR | TP | FP | FDR |
| MCP | 214 ± 129.6 | 0.20 ± 0.46 | 0.18 ± 0.44 | 217 ± 130.1 | 2.52 ± 1.83 | 1.75 ± 1.89 | 75 ± 21.88 | 9.90 ± 4.45 | 11.34 ± 3.40 |
| Unfiltered MCP | 288 ± 201.0 | 1105 ± 329.5 | 81.7 ± 7.88 | 312 ± 213.2 | 1407 ± 425.2 | 84.0 ± 6.77 | 101 ± 30.8 | 1049 ± 263.2 | 91.2 ± 1.66 |
| MCP w/ MGDB | 213 ± 128.4 | 0.47 ± 0.77 | 0.27 ± 0.50 | 222 ± 133.7 | 2.42 ± 1.87 | 1.54 ± 1.40 | 81 ± 23.98 | 5.10 ± 2.23 | 6.14 ± 2.66 |
| Ganon | 288 ± 201.1 | 2148 ± 599.3 | 89.6 ± 4.81 | 312 ± 213.2 | 2673 ± 754.7 | 90.8 ± 4.19 | 101 ± 30.8 | 2154 ± 439.2 | 95.6 ± 0.84 |
| Kraken | 288 ± 201.1 | 5756 ± 1188.2 | 95.0 ± 2.36 | 312 ± 213.2 | 7003 ± 1338.6 | 96.1 ± 2.12 | 100 ± 30.8 | 6470 ± 1128.8 | 98.5 ± 0.33 |
| Bracken | 287 ± 200.1 | 2057 ± 367.0 | 88.7 ± 6.34 | 312 ± 212.6 | 2677 ± 559.4 | 90.6 ± 5.10 | 100 ± 30.8 | 2618 ± 638.5 | 96.2 ± 1.18 |
| MetaCache | 288 ± 201.0 | 1968 ± 533.8 | 88.7 ± 5.27 | 312 ± 213.2 | 2457 ± 676.9 | 90.1 ± 4.57 | 101 ± 30.8 | 1942 ± 406.6 | 95.1 ± 0.96 |
| Centrifuge | 288 ± 201.0 | 1584 ± 451.7 | 86.4 ± 6.26 | 312 ± 212.7 | 1716 ± 500.9 | 86.4 ± 5.97 | 100 ± 30.6 | 1097 ± 279.7 | 91.6 ± 1.77 |
| DIAMOND-LCA | 288 ± 200.9 | 4828 ± 953.3 | 94.9 ± 2.82 | 312 ± 213.0 | 5617 ± 1103.3 | 95.3 ± 2.58 | 101 ± 30.9 | 4611 ± 868.0 | 97.9 ± 0.48 |
| Kaiju | 288 ± 201.1 | 7943 ± 1177.9 | 96.8 ± 1.92 | 312 ± 213.2 | 9161 ± 1094.7 | 96.9 ± 1.85 | 101 ± 30.8 | 8385 ± 1247.4 | 98.8 ± 0.27 |
| mOTUs | 218 ± 136.0 | 6.67 ± 2.94 | 3.64 ± 2.05 | 240 ± 151.5 | 15.9 ± 5.30 | 8.58 ± 5.47 | 82 ± 25.3 | 44.65 ± 14.37 | 35.1 ± 6.00 |
| MetaPhlAn | 187 ± 113.6 | 12.2 ± 5.25 | 7.55 ± 3.72 | 193 ± 117.5 | 15.5 ± 6.58 | 8.72 ± 2.92 | 65 ± 19.6 | 7.85 ± 2.50 | 10.7 ± 2.63 |

**Supp. Table 4.** Evaluation of classifiers to predict the presence or absence of species in mock communities with varying ANI to reference database genomes (mean ± std. dev.).

| | High ANI | | | Moderate ANI | | | Low ANI | | |
|---|---|---|---|---|---|---|---|---|---|
| Classifier | Precision | Recall | F$_1$ score | Precision | Recall | F$_1$ score | Precision | Recall | F$_1$ score |
| MCP | 1.00 ± 0.00 | 0.97 ± 0.02 | 0.98 ± 0.01 | 0.98 ± 0.02 | 0.92 ± 0.05 | 0.95 ± 0.02 | 0.89 ± 0.03 | 0.84 ± 0.08 | 0.86 ± 0.04 |
| Unfiltered MCP | 0.90 ± 0.09 | 0.99 ± 0.01 | 0.94 ± 0.05 | 0.85 ± 0.11 | 0.96 ± 0.02 | 0.90 ± 0.06 | 0.67 ± 0.08 | 0.89 ± 0.06 | 0.76 ± 0.04 |
| MCP w/ MGDB | 1.00 ± 0.00 | 0.97 ± 0.02 | 0.98 ± 0.01 | 0.99 ± 0.01 | 0.93 ± 0.04 | 0.96 ± 0.02 | 0.94 ± 0.03 | 0.90 ± 0.06 | 0.92 ± 0.03 |
| Ganon | 0.83 ± 0.13 | 0.99 ± 0.01 | 0.90 ± 0.08 | 0.76 ± 0.16 | 0.99 ± 0.01 | 0.85 ± 0.10 | 0.48 ± 0.09 | 0.98 ± 0.02 | 0.64 ± 0.08 |
| Kraken | 0.71 ± 0.18 | 0.99 ± 0.01 | 0.82 ± 0.13 | 0.64 ± 0.20 | 0.99 ± 0.01 | 0.76 ± 0.15 | 0.34 ± 0.08 | 0.98 ± 0.02 | 0.50 ± 0.09 |
| Bracken | 0.70 ± 0.18 | 1.00 ± 0.00 | 0.81 ± 0.14 | 0.62 ± 0.20 | 0.99 ± 0.01 | 0.74 ± 0.16 | 0.32 ± 0.08 | 0.99 ± 0.01 | 0.48 ± 0.09 |
| MetaCache | 0.83 ± 0.13 | 1.00 ± 0.01 | 0.90 ± 0.08 | 0.78 ± 0.15 | 0.99 ± 0.01 | 0.86 ± 0.09 | 0.50 ± 0.08 | 0.99 ± 0.01 | 0.66 ± 0.07 |
| Centrifuge | 0.89 ± 0.09 | 0.96 ± 0.03 | 0.92 ± 0.04 | 0.76 ± 0.14 | 0.93 ± 0.04 | 0.83 ± 0.07 | 0.40 ± 0.08 | 0.88 ± 0.06 | 0.54 ± 0.06 |
| Diamond | 0.85 ± 0.12 | 0.70 ± 0.11 | 0.76 ± 0.06 | 0.83 ± 0.13 | 0.70 ± 0.12 | 0.74 ± 0.06 | 0.60 ± 0.09 | 0.68 ± 0.07 | 0.63 ± 0.06 |
| Kaiju | 0.75 ± 0.17 | 0.95 ± 0.04 | 0.82 ± 0.11 | 0.68 ± 0.19 | 0.94 ± 0.04 | 0.77 ± 0.12 | 0.35 ± 0.07 | 0.93 ± 0.05 | 0.51 ± 0.07 |
| mOTUs | 0.96 ± 0.02 | 0.93 ± 0.02 | 0.95 ± 0.02 | 0.91 ± 0.05 | 0.94 ± 0.02 | 0.92 ± 0.03 | 0.64 ± 0.06 | 0.89 ± 0.05 | 0.75 ± 0.04 |
| MetaPhlAn | 0.93 ± 0.03 | 0.79 ± 0.05 | 0.85 ± 0.03 | 0.92 ± 0.03 | 0.75 ± 0.07 | 0.82 ± 0.04 | 0.90 ± 0.03 | 0.70 ± 0.05 | 0.79 ± 0.03 |

**Supp. Table 5.** Performance statistics for species abundance estimates on mock communities with varying ANI to reference database genomes (mean ± std. dev.).

| | **High ANI** | | | **Moderate ANI** | | | **Low ANI** | | |
|---|---|---|---|---|---|---|---|---|---|
| Classifier | L1 distance | Relative error | Sum of FPs | L1 distance | Relative error | Sum of FPs | L1 distance | Relative error | Sum of FPs |
| MCP | 7.63 ± 1.43 | 7.84 ± 1.32 | 0.01 ± 0.02 | 14.2 ± 3.24 | 14.6 ± 2.60 | 0.12 ± 0.13 | 31.9 ± 5.31 | 33.2 ± 7.36 | 1.01 ± 0.77 |
| Unfiltered MCP | 8.41 ± 1.49 | 8.04 ± 0.85 | 0.53 ± 0.66 | 15.9 ± 3.64 | 15.8 ± 3.74 | 0.94 ± 0.63 | 34.5 ± 5.48 | 35.6 ± 7.73 | 2.31 ± 1.14 |
| MCP w/ MGDB | 8.39 ± 1.88 | 8.59 ± 1.35 | 0.02 ± 0.03 | 13.6 ± 4.17 | 14.1 ± 3.10 | 0.22 ± 0.47 | 20.2 ± 10.58 | 20.2 ± 5.98 | 1.35 ± 1.69 |
| Ganon | 8.53 ± 1.66 | 7.56 ± 0.80 | 0.93 ± 0.74 | 15.1 ± 4.17 | 13.7 ± 4.26 | 1.90 ± 1.14 | 30.3 ± 7.36 | 30.5 ± 15.89 | 6.38 ± 2.27 |
| Kraken | 11.03 ± 2.60 | 9.41 ± 1.99 | 2.05 ± 1.20 | 17.7 ± 5.48 | 15.6 ± 5.70 | 3.70 ± 2.08 | 35.6 ± 8.81 | 35.1 ± 21.44 | 11.2 ± 3.28 |
| Bracken | 7.66 ± 2.38 | 7.09 ± 2.31 | 2.27 ± 1.30 | 13.2 ± 5.20 | 12.6 ± 6.86 | 4.24 ± 2.38 | 31.0 ± 8.47 | 31.2 ± 23.43 | 12.9 ± 3.87 |
| MetaCache | 6.41 ± 1.65 | 6.48 ± 1.32 | 1.04 ± 0.84 | 11.0 ± 3.43 | 12.0 ± 6.27 | 1.81 ± 1.21 | 22.6 ± 7.35 | 27.7 ± 19.52 | 6.57 ± 3.24 |
| Centrifuge | 45.8 ± 7.40 | 49.0 ± 9.32 | 0.59 ± 0.60 | 57.6 ± 8.20 | 63.1 ± 14.50 | 2.32 ± 1.41 | 77.3 ± 17.34 | 80.7 ± 17.49 | 15.0 ± 7.20 |
| DIAMOND-LCA | 80.3 ± 4.11 | 69.8 ± 4.15 | 0.47 ± 0.31 | 80.5 ± 4.85 | 70.2 ± 3.53 | 0.60 ± 0.51 | 88.4 ± 3.52 | 78.7 ± 3.56 | 1.35 ± 0.46 |
| Kaiju | 41.6 ± 6.44 | 33.1 ± 3.11 | 1.42 ± 0.84 | 46.5 ± 6.35 | 37.2 ± 2.29 | 2.34 ± 1.46 | 63.4 ± 8.20 | 54.2 ± 13.01 | 7.51 ± 1.85 |
| mOTUs | 15.1 ± 9.58 | 33.3 ± 17.48 | 2.74 ± 4.39 | 17.3 ± 6.22 | 35.1 ± 15.14 | 4.07 ± 3.54 | 35.0 ± 12.41 | 59.4 ± 80.10 | 13.5 ± 5.49 |
| MetaPhlAn | 34.6 ± 16.25 | 28.9 ± 13.52 | 2.91 ± 3.03 | 40.0 ± 13.21 | 39.1 ± 55.33 | 3.98 ± 3.19 | 49.5 ± 19.46 | 63.1 ± 58.48 | 3.53 ± 2.47 |

**Supp. Table 6.** Percent identity (PI) and percent alignment length (PA) of reads mapped to reference databases by the MCP (mean ± std. dev.).

| Samples | No. samples | ANI to reference genomes | PI | PA |
|---|---|---|---|---|
| Australian fecal samples | 2,000 | (unknown) | 99.78 ± 0.56 | 99.99 ± 0.14 |
| High ANI mock samples | 40 | 99 to 99.75% | 99.76 ± 0.67 | 99.98 ± 0.27 |
| Moderate ANI mock samples | 40 | 97 to 99% | 99.39 ± 1.07 | 99.97 ± 0.37 |
| Low ANI mock samples | 20 | 95 to 97% | 98.26 ± 1.76 | 99.91 ± 0.66 |

Website: **www.microba.com**

Email: **info@microba.com**

Address: Microba Life Sciences, Ltd.,
Level 12, 388 Queen Street, Brisbane City,
QLD 4000, Australia