WHITE PAPER

MICROBA

Microba's Community Profiler enables precise measurement of the gut microbiome

Donovan H. Parks, Fabio Rigato, Lutz Krause, Kaylyn Tousignant, Alena L. Pribyl, Philip Hugenholtz, Gene W. Tyson, David L.A. Wood

Microba's Community Profiler enables precise measurement of the gut microbiome

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

Introduction

Accurately identifying the microbial species present in biological samples is essential for understanding their role in clinical and environmental applications. In the context of the human microbiome, this is critical for establishing links between microbial species and health and disease. Our inability to culture most *in situ* microbial populations has severely limited our understanding of microbial ecosystems[1,2], and even highly studied habitats such as the human gut lack cultured representatives for ~70% of the component species[3]. The application of metagenomic sequencing – the recovery and analysis of all genomic DNA in a clinical or environmental sample – has transformed the study of microbial ecosystems by bypassing this cultivation bottleneck, providing an unbiased and comprehensive view of the taxonomic and functional composition of microbial communities and enabling the discovery of novel species within a sample.

The accurate analysis and interpretation of metagenomic datasets remains a computational challenge due to their complexity, the comparatively short read length of sequencing technologies, and incomplete genome reference databases[4,5]. There is, therefore, a critical need for taxonomic profiling tools to keep up with advancements in sequencing technologies. Several approaches are presently used to estimate the relative abundance of species in a metagenomic sample. These can be grouped into four categories based on how sequence similarity is established: i) genome alignment approaches such as Centrifuge[6], ii) protein alignment approaches such as Kaiju[7] and DIAMOND[8], iii) marker gene approaches such as MetaPhlAn[9] and mOTUs[10], and iv) composition or k-mer based approaches such as Kraken[11], Bracken[12], MetaCache[13], and Ganon[14].

Here we benchmark the Microba Community Profiler (MCP), which uses a genome alignment approach, against existing metagenomic classifiers using 140 *in silico* mock microbial communities comprised of varying numbers of bacterial and archaeal species. 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. Recognising that the quality of the reference database has a large influence on the performance of classifiers, we developed a single standardised reference database to compare classifier performance[5,15,16]. The mock communities were stratified by how close the genomes matched the representatives in the standardised reference database using the measure of average nucleotide identity (ANI). This resulted in mock communities with identical, high, moderate and low ANIs. For the purposes of this summary, the figures below used only the high and moderate ANI mock communities (80 total mock communities), as these ANIs are the most representative of gut microbiome samples.

1 Microba Life Sciences Limited, 12/388 Queen St, Brisbane, QLD 4000, Australia
* david.wood@microba.com

Superior Detection Limit

The *in silico* mock communities were used to establish detection limits for the different classifiers. The detection limit is the lowest abundance that a species in a sample can be identified before an unacceptable number of false positives are reported. It is important to minimise false positives 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. At an FDR of 0.1% (where an FDR of 0.1% indicates that 1 in 1000 species is expected to be a false positive), the MCP had the lowest overall mean detection limit at 0.007%. When comparing detection limits across multiple target FDRs, the MCP maintained the lowest detection limit, whereas there was substantial variation between other classifiers (**Figure 1**).

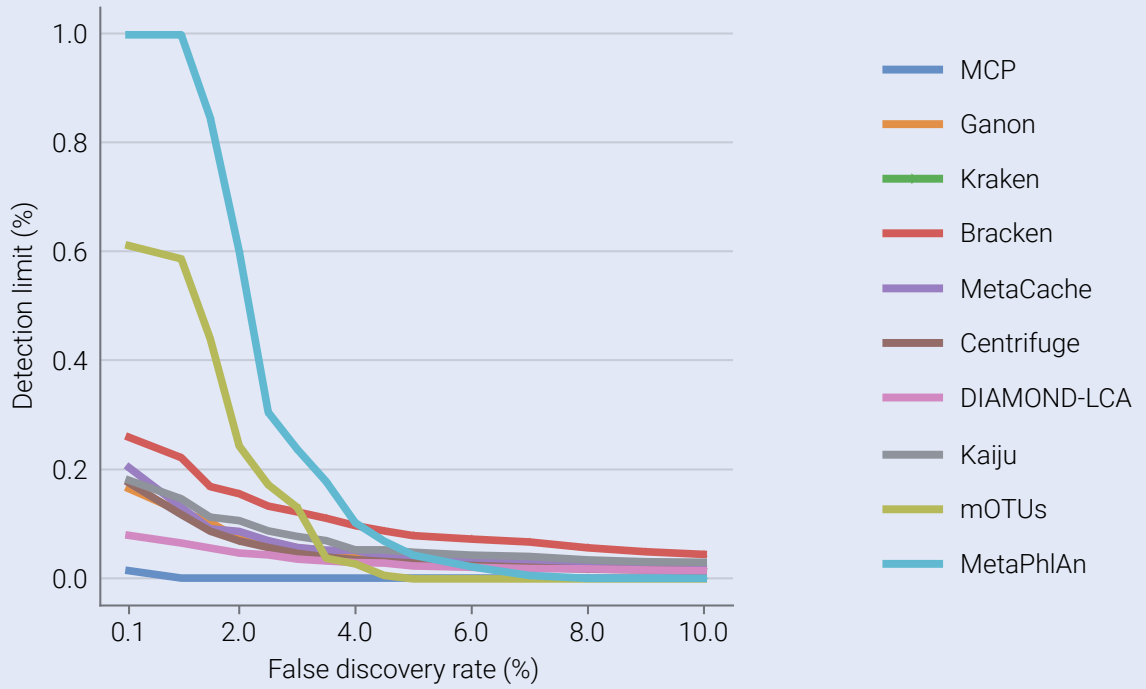

Figure 1. Comparison of minimum detection limits at which species can be reliably detected by classifiers

> With a detection limit of 0.007%, the MCP was 20-60 times more sensitive than other metagenomic classifiers.

Best Performance in Predicting the Presence or Absence of Species

Here, the *in silico* mock communities were used to determine the performance of species predictions (presence or absence of species) for the different metagenomic classifiers using an F_1 score (see below). For this analysis, we first removed low abundance populations <0.01%, which ensures more accurate results by acknowledging that species comprising the "long tail" of microbial communities[17,18] cannot be identified by most metagenomic classifiers without reporting unacceptable numbers of false positives (**Figure 1**).

In computational prediction models, an F_1 score is a measure of a test's performance, with the highest possible value of 1. The F_1 score is calculated based on the precision and recall of the classifier, where precision is the proportion of correctly identified species out of all species identified, and recall is the proportion of correctly identified species out of all species that should have been identified. There is typically a trade-off between precision and recall, and it is important to optimise this balance to increase the overall performance of the classifier. Giving equal weight to precision and recall, this analysis shows the MCP is the most accurate (F_1 = 0.97) at predicting the presence or absence of species (**Figure 2**).

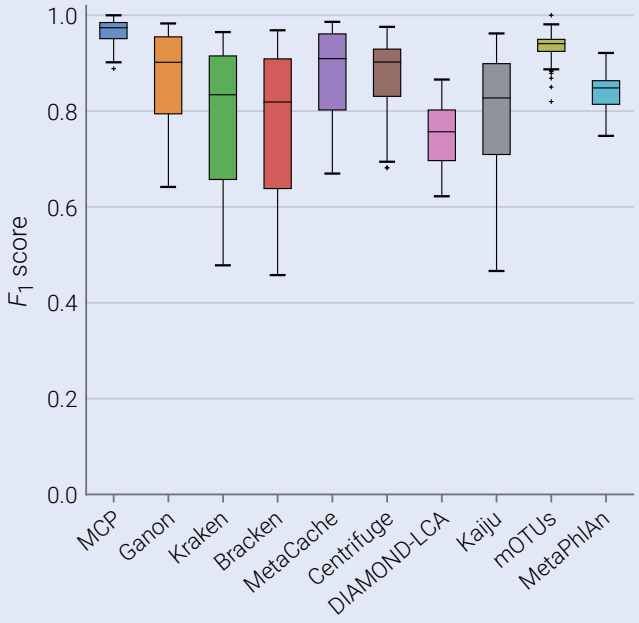

Figure 2. Comparison of F_1 score amongst metagenomic classifiers

> With an F_1 score of 0.97, the MCP has superior recall and precision, outperforming other classifiers by 5-20%

Equivalent or Superior Relative Abundance Estimates

In addition to accurately identifying the presence or absence of a species, it is also important that a classifier can accurately estimate the relative abundance of a species. This was assessed by calculating the L1 distance, which is the absolute difference between predicted and mock profiles, for all 140 *in silico* mock communities (with mock communities filtered to remove species present at <0.01% abundance). These results indicate that MCP, Ganon, Kraken, Bracken, and MetaCache are all able to provide reasonably accurate species abundance estimates (**Figure 3**).

We further assessed the percent of the predicted community that was comprised of false positive and false negative species using the *in silico* mock communities. The MCP predicted a substantially lower overall percent abundance of false positive species compared to the other classifiers, while still maintaining a low overall percent abundance of false negatives (**Table 1**). This highlights the trade-off between false negative and false positive predictions, and shows the MCP favours a slight increase in the percent of the community that is not detected (false negatives) in order to substantially reduce the percent of the community comprised of erroneously reported species (false positives).

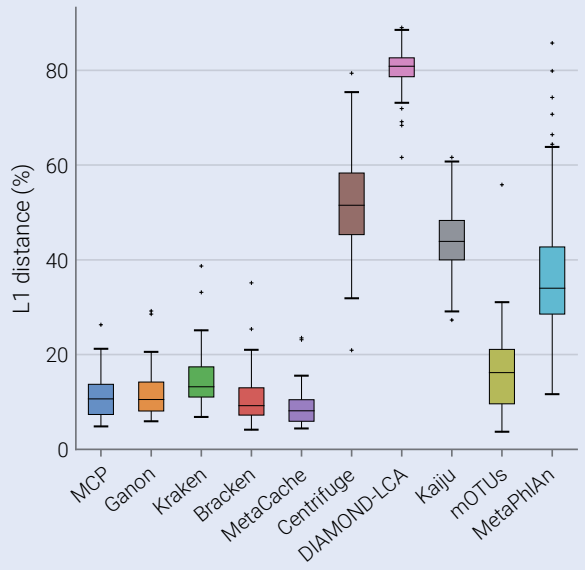

Figure 3. Comparison of L1 distance amongst metagenomic classifiers

Table 1. False positive (FP) and false negative (FN) species abundance estimates across the 140 mock communities (mean ± std. dev)

Classifier	Abundance of FPs (%)	Abundance of FNs (%)
MCP	0.18 ± 0.45	0.20 ± 0.25
Ganon	1.72 ± 2.32	0.03 ± 0.05
Kraken	3.42 ± 3.89	0.06 ± 0.35
Bracken	3.92 ± 4.53	0.05 ± 0.35
MetaCache	1.97 ± 3.00	0.02 ± 0.03
Centrifuge	2.97 ± 5.73	0.28 ± 0.32
DIAMOND-LCA	0.54 ± 0.53	3.05 ± 2.41
Kaiju	2.19 ± 2.59	0.20 ± 0.19
mOTUs	4.79 ± 5.76	4.58 ± 4.69
MetaPhlAn	3.58 ± 3.21	16.63 ± 10.82

> The MCP can accurately predict the relative abundance of species and the abundance of erroneously identified species is 4 -16x's less than other classifiers.

The MCP Identifies More Species

In recent years, the application of metagenomic sequencing has shed light on the numerous links between the gut microbiome and human health. The gut microbiome has been associated with a range of diseases including metabolic diseases[19], gastrointestinal disorders[20] and even cancer[21], making it critical to have accurate and comprehensive characterisation of communities to help guide diagnostics and therapeutic strategies.

Therefore, we evaluated the performance of the strongest performing classifiers on human faecal metagenomes. Community profiles produced by MetaCache, Kraken, and Bracken were compared to those obtained using the MCP on a set of 100 de-identified Australian faecal metagenomes. Unlike the *in silico* mock community analysis, here each classifier was evaluated using its recommended reference database. MCP uses the Microba Genome Database (MGDB) which consists of 73,646 dereplicated genomes which span the 24,706 species clusters defined by the Genome Taxonomy Database[22,23] and 3,540 additional species clusters consisting of MAGs and isolates from recent human gastrointestinal studies and Microba's own initiatives to obtain MAGs from research participant samples and public metagenomes.

Since the community composition of the faecal samples are unknown, we used the percent of assigned DNA sequencing reads to evaluate classifier performance. Across the 100 faecal metagenomes, the MCP could assign an average of 84.4% of the DNA reads to a species, which was at least 25% more than the next closest performing classifier (**Figure 4**). 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, highlighting the need for a comprehensive reference database to achieve improved species identification by metagenomic classifiers.

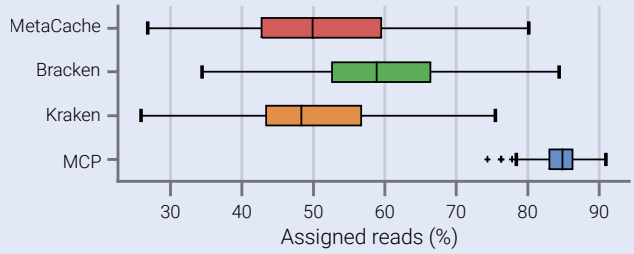

Figure 4. Comparison of the percentage of DNA reads assigned to a species by metagenomic classifiers on 100 de-identified Australian faecal samples

> Using Microba's Genome Database, the MCP can assign at least 25% more DNA reads per sample than other classifiers.

Conclusions

The Microba Community Profiler was developed to provide accurate metagenomic profiles of faecal microbiomes. Here we demonstrate that the MCP has a superior detection limit, the most reliable species predictions, equivalent or superior relative abundance estimates, and identifies at least 25% more species among all evaluated classifiers. Importantly, this analysis indicates that the MCP outperforms the other classifiers even when all classifiers use the same reference genome database.

The MCP is under ongoing development to further improve its detection limit and accuracy of species relative abundance estimates. In addition, the MGDB is constantly updated with the genomes of newly identified species to improve the percent of assigned DNA sequencing reads. While we will continually improve the performance of the MCP, the results of this study illustrate that the MCP is overall the best performing metagenomic classifier.

References

1. **Epstein, S. S.** The phenomenon of microbial uncultivability. *Current Opinion in Microbiology* **16**, 636-642, doi:https://doi.org/10.1016/j.mib.2013.08.003 (2013).
2. **Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L**. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, e00055-00018, doi:10.1128/mSystems.00055-18 (2018).
3. **Almeida, A. *et al.*** A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).
4. **Sczyrba, A. *et al.*** Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods* **14**, 1063-1071, doi:10.1038/nmeth.4458 (2017).
5. **Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C.** Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178**, 779-794, doi:10.1016/j.cell.2019.07.010 (2019).
6. **Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L.** Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26,** 1721-1729, doi:10.1101/gr.210641.116 (2016).
7. **Menzel, P., Ng, K. L. & Krogh, A.** Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* **7**, 11257, doi:10.1038/ncomms11257 (2016).
8. **Buchfink, B., Xie, C. & Huson, D. H.** Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).
9. **Segata, N. *et al.*** Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811-814, doi:10.1038/nmeth.2066 (2012).
10. **Milanese, A. *et al*.** Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications* **10**, 1014, doi:10.1038/s41467-019-08844-4 (2019).
11. **Wood, D. E., Lu, J. & Langmead, B.** Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).
12. **Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L.** Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**, e104, doi:10.7717/peerj-cs.104 (2017).
13. **Müller, A., Hundt, C., Hildebrandt, A., Hankeln, T. & Schmidt, B.** MetaCache: context-aware classification of metagenomic reads using minhashing. *Bioinformatics* **33**, 3740-3748, doi:10.1093/bioinformatics/btx520 (2017).
14. **Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K. & Renard, B. Y.** Ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *bioRxiv*, 406017, doi:10.1101/406017 (2019).
15. **Nasko, D. J., Koren, S., Phillippy, A. M. & Treangen, T. J.** RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biology* **19**, 165, doi:10.1186/s13059-018-1554-6 (2018).
16. **Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M.** Correcting index databases improves metagenomic studies. *bioRxiv*, 712166, doi:10.1101/712166 (2019).
17. **Curtis, T. P., Sloan, W. T. & Scannell, J. W.** Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences **99**, 10494-10499, doi:10.1073/pnas.142680199 (2002).
18. **Fritz, A. *et al.*** CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17, doi:10.1186/s40168-019-0633-6 (2019).
19. **Fan, Y. & Pedersen, O.** Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, doi:10.1038/s41579-020-0433-9 (2020).
20. **Caruso, R., Lo, B. C. & Núñez, G.** Host-microbiota interactions in inflammatory bowel disease. *Nat Rev Immunol,* doi:10.1038/s41577-019-0268-7 (2020).
21. **Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A. & Wargo, J. A.** The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. *Cancer Cell* **33**, 570-580, doi:10.1016/j.ccell.2018.03.015 (2018).
22. **Parks, D. H. *et al.*** A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* **38**, 1079-1086, doi:10.1038/s41587-020-0501-8 (2020).
23. **Parks, D. H. *et al*.** A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004, doi:10.1038/nbt.4229 (2018).

MICROBA

www.microba.com - info@microba.com

Microba Life Sciences Ltd., Level 12, 388 Queen Street,

Brisbane City, QLD 4000, Australia