Lightweight taxonomic profiling of long-read sequenced metagenomes with Lemur and Magnet

Nicolae Sapoval [1,*], Yunxi Liu [1], Kristen D. Curry [1], Bryce Kille [1], Wenyu Huang [1], Natalie Kokroko [1], Michael G. Nute [1], Alona Tyshaieva[3], Alexander Dilthey [3], Erin K. Molloy [2,†], and Todd J. Treangen [1,†,*]

[†]These authors have contributed equally to this work.

[1]Department of Computer Science, Rice University, Houston, TX 77005, USA

[2]Department of Computer Science, University of Maryland, College Park, MD 20742, USA

[3]Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[*]Corresponding authors: nsapoval@rice.edu, treangen@rice.edu

Abstract

Taxonomic profiling is a ubiquitous task in the analysis of clinical and environmental microbiomes. The advent of long-read sequencing of microbiomes necessitates the development of new taxonomic profilers tailored to long-read shotgun metagenomic datasets. Here, we introduce Lemur and Magnet, a pair of tools optimized for lightweight and accurate taxonomic profiling from long-read shotgun metagenomic datasets. Lemur is a marker-gene based method that leverages an EM algorithm to reduce false positive calls while preserving true positives; Magnet makes detailed presence/absence calls for bacterial genomes based on whole-genome read mapping. The tools work in sequence: Lemur estimates abundances conservatively, and Magnet operates on the genomes of identified organisms to filter out likely false positive taxa. The result is an increase in precision of as much as 70%, which far exceeds competing methods. By operating only on marker genes, Lemur is a comparatively lightweight software. We demonstrate that it can run in minutes to hours on a laptop with 32 GB of RAM, even for large inputs – a crucial feature given the portability of long-read sequencing machines. Furthermore, the marker gene database used by Lemur is only 4 GB and contains information from over 300,000 RefSeq genomes. The reference is available at `https://zenodo.org/records/10802546`, and the software is open-source and available at `https://github.com/treangenlab/lemur`.

1 Introduction

The democratization of long-read sequencing has arrived [23] and is now common practice in metagenomic studies due to a combination of higher accuracy, increased affordability, and greater genomic resolution provided by longer reads [1]. One of the most common tasks in metagenomics is to perform taxonomic profiling of a microbial community specific to a host microbiome or environmental microbiome. Existing taxonomic read classification tools such as Kraken 2 [39] have established themselves as a *de facto* standard approach for taxonomic read classification and taxonomic profiling with short-read data. Several new tools have recently been developed to leverage long-reads for taxonomic profiling. Common approaches taken by the developers consist of k-mer based methods (Kraken 2 [39], Sourmash [11]), read-mapping to a succinct index (Centrifuger [35], MetaMaps [7]), protein based (MEGAN-LR [10]) and marker gene [40] based approaches (Melon [41], PhyloSift [6], MetaPhyler [22]).

Prior benchmarking studies have evaluated the accuracy of some of these methods for long-read metagenome taxonomic profiling [29]; methods designed specifically for long reads tend to perform better. However, in the prior studies the focus of evaluation was primarily on precision and recall. Given that the long-read technologies offer potential for portable and streaming sequence analysis [30], it is also important to evaluate scalability and fitness for execution in low resource environments such as laptops and tablet computers. In particular, we note that depending on the database size and computational requirements, several of the currently available tools need dedicated server nodes with high RAM (> 32 GB) and/or high parallelization capacity to achieve time-to-answer below 2 hours.

To address these issues, we present a two-tool suite consisting of Lemur, a marker gene based long-read taxonomic profiler, and Magnet, a genome based validation tool for confirming the presence/absence of microbial genomes present in a sample. Combined, Lemur and Magnet can run in limited resource settings, such as individual laptops, and yield comparable or superior performance in terms of precision and recall. Our overall contribution hence is threefold: (1) proposing two novel methods for taxonomic profiling that achieve superior precision and portability, (2) extending the marker gene database to include fungal genes for fungi classification, and (3) facilitating taxonomic profiling to be run in lightweight compute environments.

2 Materials and methods

An overview of Lemur and Magnet is presented in Fig. 1. Both methods require a FASTQ file containing sequencing reads as input. Lemur additionally requires a marker gene (MG) database, whereas Magnet requires a (ideally small) set of genomes. In both cases, the first step is to map the reads to the database using minimap2 [21]. For our study, we built a new marker gene database with 43 markers for bacteria+archaea and 48 markers for fungi. These markers were validated by prior studies [24, 32, 41, 16]. For all marker genes, we downloaded the HMM profiles from both TIPP2 reference package and KofamKOALA database [3]. All HMM profiles used in the studies are provided in the source data. We then built the database using recent versions of NCBI RefSeq: version 221 for both bacteria (329,194 assemblies) and archaea (1,911) and version 222 for fungi (564). A total of 331,669 genome assemblies, including all of their annotated protein sequences and corresponding CDS sequences were downloaded from NCBI RefSeq. All of the pseudogenes were excluded. We used fetchMGs (v1.2) to extract the selected marker gene sequences with command `fetchMGs.pl -m extraction` [36], and created a mapping between the sequences and their corresponding species rank taxonomic ids with our custom Python script. We used the Emu database creation tool [5] for the final step in the database construction for individual marker genes with the command `emu build-database --ncbi-taxonomy`. Finally, individual marker gene databases were concatenated, and a single joint taxonomy mapping was generated for the combined database. The final database was 4.1 GB, containing 3,335,783 sequences.

2.1 Lemur

Lemur estimates a relative abundance taxonomic profile using an Expectation-Maximization (EM) framework, similar to the 16S profiler Emu [5]. The EM algorithm proceeds in two steps. First, in the E-step, we compute the probability $P(t|r)$ of taxon t being in the sample given an observed read r, which uses an abundance prior $F(t)$ and Bayes rule: $P(t|r) = \frac{P(r|t)F(t)}{\sum_{t \in T} P(r|t)F(t)}$, where $P(r|t)$ is the probability of observing read r originating from taxon t. Second, in the M-step, the abundance vector $F(t)$ is re-estimated using $P(t|r)$ for all reads r in the dataset as $F(t) = \frac{\sum_{r \in R} P(t|r)}{|R|}$, $\forall t \in T$. Finally, we compute the total log-likelihood

$$L = \sum_{r \in R} \sum_{t \in T} P(r|t)F(t)$$

and the change in log-likelihood ΔLL from the prior step L' as $\Delta LL = L - L'$. These steps are repeated

until $\Delta LL < \varepsilon$, where ε is a user defined threshold with a default value of 0.01 (see blue cycle of Fig. 1A). After convergence, Lemur removes any taxonomic assignments below a user-specified threshold (default: 1 / number of reads that map to marker genes) and then runs a final iteration of the EM algorithm to re-estimate $F(t)$.

The EM algorithm begins by initializing $F(t)$ to the uniform distribution and initializing $P(r|t)$ for each read and taxon pair (r, t). Lemur implements several options for calculating $P(r|t)$. The fastest and default option is to take the alignment score output by minimap2 and normalize it: $P(r|s) = AS(r, s)/2L$ where AS is the alignment score from the minimap2 SAM file and L is the alignment length. Lemur aggregates the scores for all sequences assigned the same taxonomic label retaining the score from teh most likely alignment, setting $P(r|t) = \max_{s \in t} P(r|s)$ (as there can be multiple genomes corresponding to the same taxon). Although not explored in this study, Lemur implements two other approaches for estimating $P(r|s)$, including the CIGAR-based approach from Emu [5] and a Markov chain model of sequence alignment.

Additionally, Lemur can employ a uniform coverage (UC) filter prior to the EM-steps to eliminate taxa from the taxonomic profile if the read-coverage pattern across the marker genes significantly deviates from the expected value. The goal is to improve Lemur's precision when sequencing depth is high, and coverage is uniform (so taxa with non-uniform coverage may be false positives). Specifically, let G be the number of marker genes. Let N be the number of reads that map to a particular taxon. Let X denote the random variable corresponding to the number of distinct genes with at least one read aligned to them. Let X_i be the indicator variable for the event that a gene i is covered by a read. Then, the expected number of genes hit is

$$\begin{aligned}\mathbb{E}(X) = \mathbb{E}\left(\sum_{i \in \{1,\ldots,G\}} X_i\right) &= \mathbb{E} \sum_{i \in \{1,\ldots,G\}} (X_i) \\ &= G\mathbb{P}(X_1 = 1) = G\left(1 - \left(\frac{G-1}{G}\right)^N\right).\end{aligned}$$

The first equality is by linearity of expectation, and since the gene being hit by at least one read is a complementary event to all reads aligning to one of the other $G - 1$ genes, it follows that the desired equality holds (we treat individual read to gene alignments as independent events). Furthermore, we note that:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - \mu^2 \\ &= \mathbb{E}\left[\left(\sum_{i \in \{1,\ldots,G\}} X_i\right)^2\right] - \mu^2 \\ &= \sum_{i \in \{1,\ldots,G\}} \mathbb{E}(X_i^2) + \sum_{i,j \in \{1,\ldots,G\}} \mathbb{E}(X_i X_j) - \mu^2 \\ &= G\left(1 - \left(\frac{G-1}{G}\right)^N\right) + (G^2 - G)(1 - (1 - 2/G)^N) - G^2\left(1 - \left(\frac{G-1}{G}\right)^N\right)^2 \\ &= G(1 - 1/G)^N + G^2(1 - 1/G)(1 - 2/G)^N - G^2(1 - 1/G)^{2N}.\end{aligned}$$

Again, the first expansion is due to the linearity of expectation. Then we note that the random variable X_i^2 has the same expected value as X_i, while $\mathbb{P}(X_i X_j = 1)$ is given by $1 - \mathbb{P}(\vee_{k \neq i,j} X_k = 1)$ which is simply $1 - ((G-2)/G)^N$.

Finally, the actual number of genes hit, denoted G_{hit}, is also computed directly from the data. If $\mu - G_{hit} \leq 3\sigma^2$ and $G_{hit}/\mu \leq 0.7$, the *taxon* is removed (filtered) from further consideration.

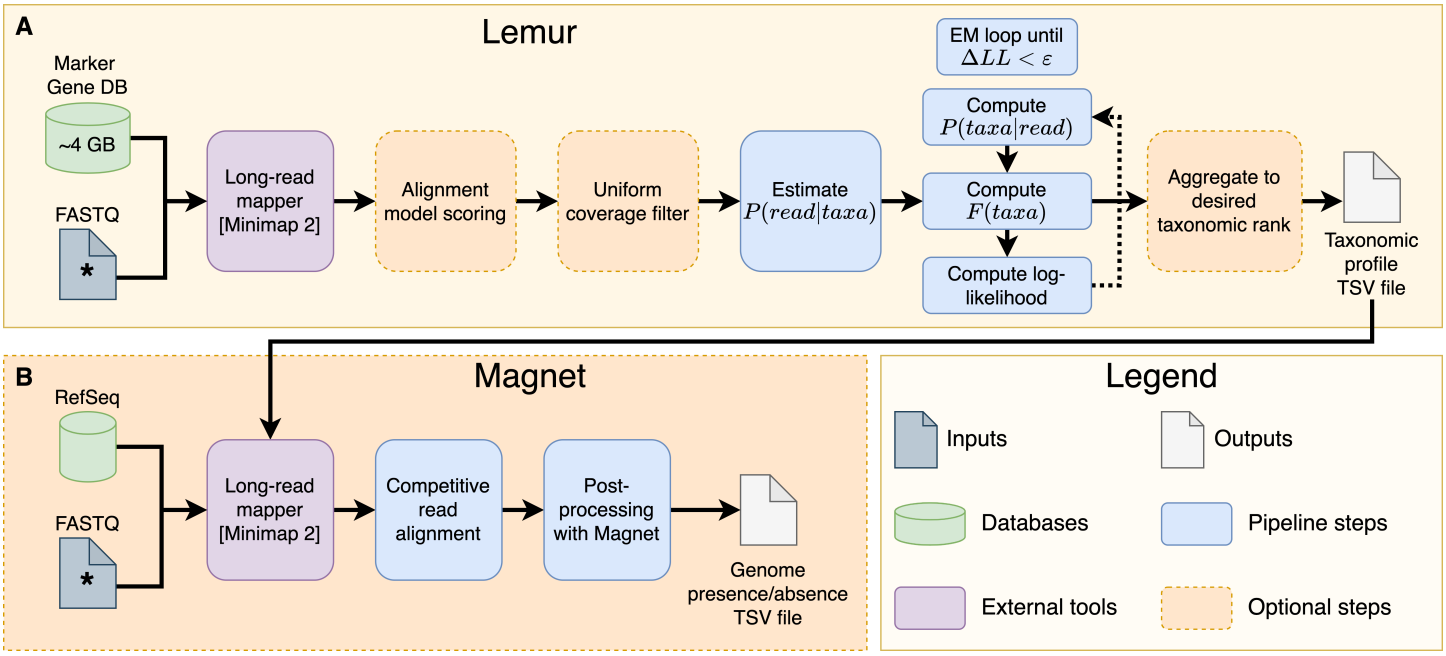

Figure 1: Overview of Lemur **(A)** and Magnet **(B)** pipelines. Input FASTQ files with the * symbol indicate same file. Taxonomic profile table provided to Magnet can come from any taxonomic profiler or classifier, as long as it respects formatting and uses the same set of taxonomic ids.

As a post-processing step, the taxonomic profile, along with the original reads, can be fed into Magnet for genome presence/absence evaluation. Magnet employs a competitive alignment strategy designed to differentiate closely related genomes.

2.2 Competitive read alignment with Magnet

The goal of Magnet is to detect and remove potential false positives by performing competitive read alignment leveraging all of the reads mapped against the entire reference genome (rather than just the marker gene reads and marker genes used by Lemur). As input, Magnet requires reads as well as a taxonomic abundance profile (estimated from the input reads *e.g.* using Lemur). For each candidate species identified in the profile, Magnet downloads the highest ranking genome from NCBI RefSeq and Genbank databases, where the ranking is determined by (1) representative genomes in RefSeq, (2) complete genomes in RefSeq, (3) other genomes sorted by assembly level (complete, chromosome, scaffold, contig).

Once the genomes are downloaded, Magnet constructs a dissimilarity matrix by calculating pairwise average nucleotide identity (ANI) between all candidate sequences using FastANI [12]. Magnet then performs agglomerative clustering with scikit-learn [27] using a threshold of 0.05. For each cluster, one representative genome is selected. If a cluster contains complete genomes, the complete genome with the greatest sum of ANI to other complete genomes within the cluster is selected as representative. Otherwise, a genome at any assembly level with the greatest sum of ANI to other cluster members is selected as the representative.

The reads from the input dataset are aligned to all cluster representative genomes at once using minimap2. Magnet calculates the observed breadth and depth of coverage for the alignment with `samtools coverage` for two cases: (1) including all alignments with MAPQ ≥ 1 and (2) including only primary alignments with MAPQ ≥ 20. For both cases, Magnet calculates the expected breadth of coverage based on the abundance and the coverage score, defined as the ratio of the observed breadth of coverage and expected breadth of coverage [4]. The coverage score is used to measure uniformity of the alignment distribution along the reference genome. Magnet generates the consensus genome for each cluster based on alignments and then estimates ANI between it and the

reference genome.

Lastly, Magnet marks species as present or absent. If the consensus vs reference ANI is greater than 0.97 and the breadth of coverage for high MAPQ primary alignments is greater than 0.2, the species is determined to be present. The species is also marked as present if ANI is greater than 0.7 and the *reduction* in coverage score and breadth of coverage between all alignments and high MAPQ primary alignments is less or equal to 0.1. Otherwise, the species is marked as absent. The Magnet presence/absence calling process is based on both the ANI between consensus and the aligned region reference, and coverage and uniformity decrease after excluding low MAPQ alignment. For species that are truly present in the sample, excluding low MAPQ alignments has a limited effect on breadth of coverage or coverage score. On the other hand, for likely false positives, excluding low MAPQ alignments usually causes a significant breadth of coverage or coverage score reduction, as alignments to the false positive species usually have a lower average MAPQ. The implemented thresholds were chosen based on empirical evidence and simulation-based testing.

2.3 Method Comparison

We compared the performance of Lemur to taxonomic classification tools Centrifuger [35], Kraken 2 [39], Melon [41], Metamaps [7], and Sourmash [11]. Lemur was evaluated both on its own and in combination with Magnet (Lemur + Magnet). Melon doesn't include any fungal references in its database and thus cannot classify fungi. Therefore, for two datasets that include fungi in the ground truth, we report Melon and Lemur results on bacterial species only in addition to results using complete ground truth. Exact commands used for running each tool are provided in Section A.1.1 of Supplement.

2.4 Synthetic and simulated datasets

2.4.1 Simulated data from [7]

Our first dataset is a simulated dataset from [7], with 96 bacterial strains. After validating the metadata, we confirmed that 94 out of the specified 96 strains have a corresponding species representative in the RefSeq database. The dataset contains 200,114 reads with an average length of 4,997bps and has a total size of 1.9 GB. Additional simulation details are provided in the original manuscript [7].

2.4.2 Zymo EVEN & Zymo LOG

The Zymo EVEN data sets were constructed from the ZymoBIOMICS Microbial Community Standard (D6300) sample, which consists of a DNA mixture of 8 bacterial species at an even total DNA abundance of 12% and 2 fungi at 2% DNA abundance. The Zymo LOG datasets were constructed from the ZymoBIOMICS Microbial Community Standard (D6310) sample, which consists of a DNA mixture of 8 bacterial species in a 10-fold dilution series by total DNA abundance and 2 fungi. Species used in D6310 are identical to those in D6300, although the abundances in the latter are intentionally given a heavily skewed distribution, with several low-abundance taxa present.

Original sequencing data from a Nanopore GridION device with R9.4 chemistry produced by Loman lab [25] was downloaded from GitHub. Experiments to understand the effect of coverage level were conducted by subsampling the reads with the `seqkit sample` command [34] by setting the proportion of sampled reads to 1% (Median # reads: 34.8k, avg. len.: 4006bp, avg. size: 272MB) and 75% (2.61m, 4012bp, 20GB) for Zymo EVEN and to 10% (367k, 4368bp, 3.1GB) and 75% (2.75m, 4372bp, 23GB) for Zymo LOG. For each sampling threshold, this process was repeated five times (setting the random seed parameter from 1 to 5) to create 5 replicate datasets.

2.4.3 Simulated metagenome

We simulated a challenging metagenome using 50 species with representative genomes deposited in RefSeq, with metagenome assembled genomes (MAGs) available and having an assembly-quality of at least 'scaffold' were selected. 5 replicate datasets where then created by randomly sampling 20 genomes spanning 20 distinct species, and 18-20 distinct genera. Reads were simulated from these genomes using pbsim3 [26] with a quality-score based ONT HMM profile and a mean read-length of 4,050bps (S.D: 1,000bp). Reads were then combined into simulated metagenomic samples (avg. size: 6GB). The exact pbsim3 command is provided in Section A.2 of the Supplement.

2.4.4 Zymo Fecal Reference with TruMatrix

Our final dataset is sequencing data generated on a Pacific Biosciences Sequel IIe device from the ZymoBIOMICS fecal reference samples. These data along with reference taxonomic abundance profiles were downloaded from ZymoBIOMICS and were based on a separate, curated analysis based on multiple rounds of Illumina shotgun sequencing. A total of 6 samples were analyzed ranging in size from 54 to 105GB (avg. read len.: 6973bps, avg. size: 75.8GB).

This dataset is challenging to evaluate because some community members are listed as unknown at the species and, in several cases, at the genus level. Additionally, several species are listed as 'uncultured (genus)' which is problematic because this is also a possible species-label that the tools can output, although in that case it would represent something less than a clear true positive. To avoid this ambiguity, for this dataset all such taxa were excluded from evaluation. The filtered abundance profile containing 180 bacterial species and 1 archaeal species spanning a total of 78 genera was used as the ground truth.

2.5 Gut sample from healthy donor (SRR17687125)

For a real metagenomic dataset, a gut metagenomic sample obtained from a Korean donor reportedly in good health and with an omnivorous diet was used [15]. The sample was sequenced on a Pacific Biosciences Sequel II device and contained 2.02M reads with an average read-length of 14,670bp. While no ground truth is available for this sample, the set of the assembled genomes from the original study [15] was used as a proxy for organisms present in the sample. False-positive calls were not assessed, only the recovery of taxa from assembled genomes as well as the general consensus among different taxonomic profilers in the study. This experiment was partly intended as a sanity-check for Lemur to see that it produces a plausible gut community without any highly-improbable microbes.

2.6 Chicken gut metagenome

Additionally, 6 chicken gut metagenome samples spanning different sections of the intestinal tract were analyzed. The samples were sequenced on a Pacific Biosciences Sequel II device and contained 521,403–4,282,198 reads (avg.: 2,615,037) with an average read-length of 15,989bp. Similarly to the human gut metagenome no ground truth is available for this dataset, but a set of high quality MAGs is available from the original study that produced the data [42].

2.7 Method evaluation

For all the tools and datasets with known ground truth we report recall, precision, and F1 score. Precision and recall are defined at the genus and at the species level, and in each case are defined over the set of taxa that are known or estimated to be present above a threshold of 1e-12 (with the exception of Zymo EVEN experiments, where the threshold was set to 1e-3). This was applied to convert abundance profiles for presence/absence.

Table 1: Performance on simulated data from [7]

Method	Recall	Precision	F1 score
Melon	0.963	0.929	**0.946**
Lemur + Magnet	0.927	**0.950**	0.938
Metamaps	**1.000**	0.862	0.926
Lemur	0.951	0.703	0.808
Kraken 2	0.976	0.055	0.104
Centrifuger	0.774	0.050	0.093

For tools that provide relative abundance estimates or can be used to derive relative abundance, we also evaluate normalized L1-loss and Spearman's rank correlation coefficient ρ. For datasets with < 20 ground truth species and even abundance distribution we use normalized L1-loss defined as $\frac{1}{2}\sum_{t \in T} |F(t) - \hat{F}(t)|$, where F is actual and $\hat{F}$ is estimated abundance. L1-loss is less infomrative for datasets with large numbers of taxa or with an uneven distribution, so we use Spearman's ρ as an alternative metric of correspondence between F and $\hat{F}$, and in this case it is defined over the set of species t such that both $F(t) > 0$ and $\hat{F}(t) > 0$ (so methods are not punished for false positives or false negatives in the Spearman's ρ).

All methods except for MetaMaps were ran on a server running Red Hat Enterprise Linux (v8.9) with an AMD EPYC 7742 64-core processor and 128 GB of RAM. All server runs were executed with 8 threads to emulate a laptop-like environment. No additional restrictions were placed on the RAM available to the tools. MetaMaps runs were performed utilizing 20 (for 1% and 10% subsamples) and 30 (for 75% subsamples) threads on servers with 4 Intel Xeon Gold 6150 (18-core) and 4 Intel Xeon Gold 6240 (18-core) processors and 1.5 TB of RAM running CentOS 7.9.2009 Linux. Separate experiments on laptop hardware are described later.

3 Results

We have conducted a series of evaluations of increased complexity on synthetic, simulated and real data. We compared performance of our proposed tool Lemur to Centrifuger, Kraken 2, Melon, MetaMaps, and Sourmash. For Lemur results we also show post-processing with Magnet for improved specificity (Lemur + Magnet). Additionally, because Lemur includes fungal sequences in its reference database, we illustrate the impact of this by evaluating both Melon and Lemur separately with all non-bacterial taxa excluded from evaluation.

3.1 Prior simulated data

We first evaluated the methods on simulated data from a prior study [7]. On the simulated community of 96 bacterial strains Metamaps shows perfect recall at the species level, followed by Kraken 2 with recall of 0.976. Table 1 contains the full precision/recall data for these five methods on this data. Kraken 2 has high recall but low precision. Melon, Lemur and Metamaps, all EM-algorithms, each have high recall and moderate to high precision.

For this analysis we used the Metamaps results provided in the original manuscript [7].

3.2 ZymoBIOMICS Microbial Standards

We evaluated our tools on the Zymo EVEN and LOG datasets. We evaluated tools on 5 replicate subsamples of EVEN community (Fig. 2A-B) at 1% and 75% sampling rates, and 5 replicate subsamples of LOG community (Fig. 2C-D) at 10% and 75% sampling rates.

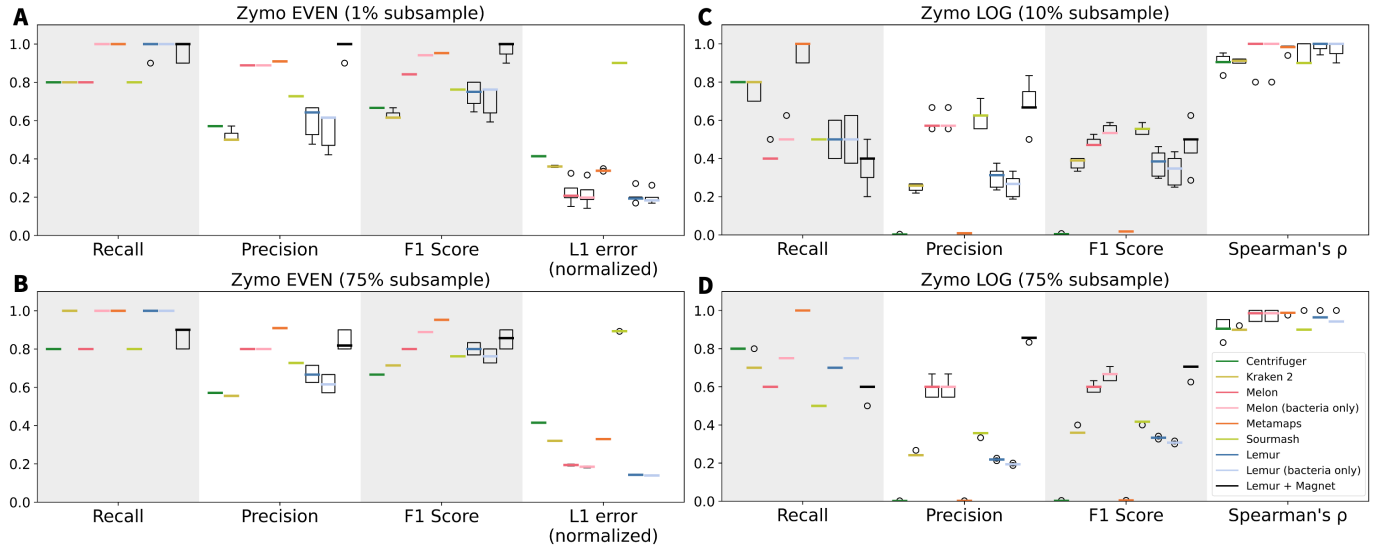

Figure 2: Performance of all methods on Zymo EVEN (**A**, **B**) and Zymo LOG (**C**, **D**). Panels show recall, precision, F1 score and normalized L1 error (**A**, **B**) or Spearman's ρ (**C**, **D**) for each of the tools.

On the EVEN community, Lemur showed high recall matched by MetaMaps. We note that Lemur can achieve full recall without restricting evaluation to bacteria-only, suggesting that it is capable of accurately evaluating both kingdoms jointly. Additionally, the benefit of polishing the abundance profiles with Magnet is supported by an increase in precision particularly in low-coverage scenarios where false-positives are more likely. Finally, in both low and high-coverage settings Lemur showed lowest L1 error.

On the LOG community, recall of all tools except MetaMaps drops as some species are present at such a low level that their expected genome copy-number is less than 0.05. In particular, marker gene based tools Melon and Lemur have recall lower than MetaMaps, Centrifuger and Kraken 2 which can utilize the information from across the whole genome (Fig. 2C-D). By contrast, the combination of many low-abundance taxa and low-coverage creates a large number of false positives as shown by the lower precision of most tools (Fig. 2C-D). However, Magnet does well with identifying false positives and improves the precision by nearly 0.6 over Lemur alone. This indicates that the tools retain ability to make confident species level calls in extremely low-abundance settings (Fig. 2C-D).

Additionally, Lemur shows strong performance on Spearman's ρ indicating that it's abundance estimates are broadly accurate vis-a-vis ordering of the microbes within a sample. Further, the reduction in variance of ρ in the higher-coverage sample is reassuring since this is an expected outcome of the additional data (Fig. 2C-D, rightmost panel).

3.3 Simulated metagenome

The simulated metagenome data set is intended to present the tools with a more challenging setting, with more diverse taxa present, but where ground truth is *known*. At species level, Sourmash and Melon show the highest recall closely followed by Lemur (Fig. 3A). Post-processing with Magnet reduces recall, but improves precision, in particular achieving highest precision among all tools at both species and genus levels (Fig. 2A-B). We also note that Sourmash and Lemur have high precision at the genus level (Fig. 2B), indicating that most false positive species-level calls made by these tools come from the correct genus.

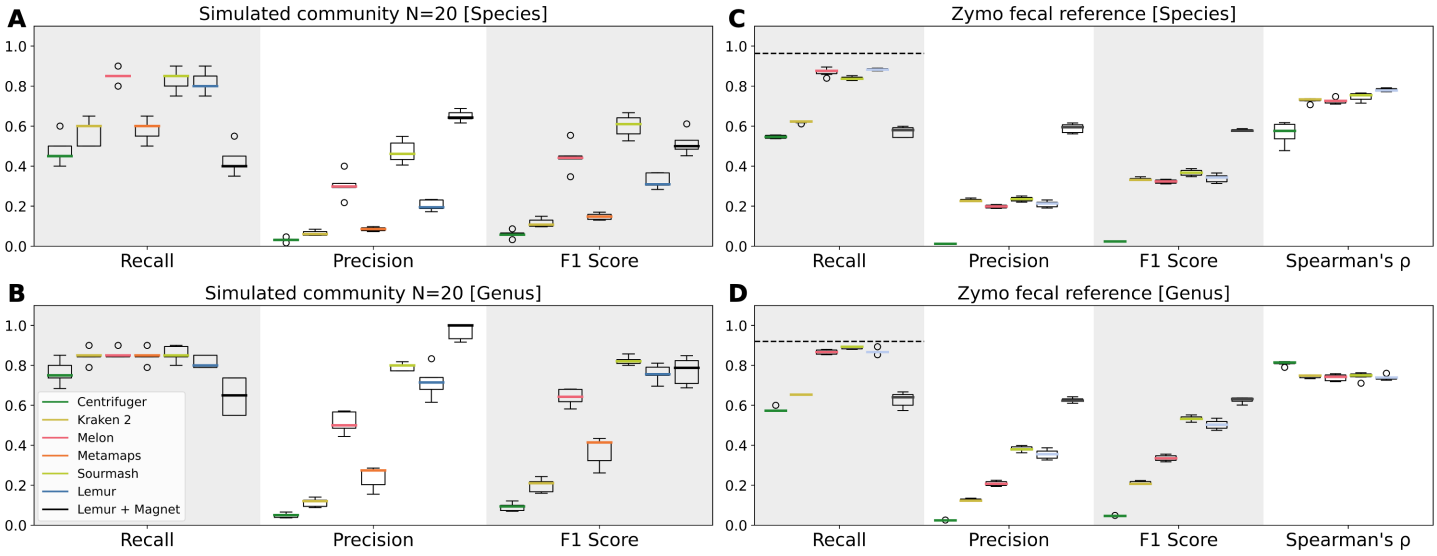

Figure 3: Performance of all methods on a simulated metagenomic community (A, B) and Zymo fecal reference (C, D). Panels show recall, precision, F1 score and Spearman's ρ (**C**, **D**) for each of the tools.

3.4 ZymoBIOMICS Fecal Reference

Next, we evaluated performance of all tools on the ZymoBIOMICS fecal reference samples (Fig. 3C-D). We did not include MetaMaps in this evaluation due to computational constraints. Lemur and Melon are the most sensitive tools at the species level, closely followed by Sourmash (Fig. 3C), at the genus level Sourmash has the highest recall with Lemur and Melon matching closely (Fig. 3D). However, the precision of Sourmash and Lemur is higher than Melon at both ranks. As with the previous datasets, post-processing with Magnet improves precision, although it imposes a penalty on recall. The tool combination maintains the highest F1 score (Fig. 3C-D). Lemur also has the highest ρ at the species level. At the genus level, Centrifuger has the highest Spearman's ρ, although as noted in section 2.7, the calculation in this case ignores false positives and false negatives (Fig. 3C-D, rightmost panels).

3.5 Human gut metagenome

Finally, species and genus level calls made by Lemur were evaluated on stool sample from a healthy donor [15]. We note that the most abundant phyla identified by Lemur (Fig. 4) match those in the prior literature. In particular *Actinomycetota*, *Bactriodota*, *Bacillota*, and *Pseudomonadota* are well-represented phyla (Fig. 4P). Additionally, *Faecalibacterium prausnitzii*, *Bacteroides ovatus*, *Parabacteroides distasonis*, and *Alistipes onderdonkii* (Fig. 4S) are all supported by the co-assemblies from the original study. Other bacterial species identified by Lemur such as *Roseburia faecis*, *Phocaeicola vulgatus*, and *Lachnospira Pectinoschiza* have all been previously isolated from human feces and can be considered putative true-positives (Fig. 4S).

To perform a deeper dive, we also investigated the concordance between the species and genus level calls across the tools. One genus level call that is uniquely inferred by Lemur corresponds to *Massilimicrobiota*. A recent isolate from human gut belonging to the *Massilimicrobiota* genus has been identified [37] indicating that this call is a plausible true positive. Conversely, 9 genera were identified by Kraken 2 and Melon but *not* Lemur. Those include several genera whose reported isolation sources are distant from the human gut, including a biogas fermenter, two from marine thermal vents, marine sediment, and wastewater treatment plant. The remaining four genera were previously found in the human gut. The list of genera, their isolation source and relevant citations is contained in Section A.3.1 of the Supplement.

While this is not a conclusive experiment to assess false positives, it represents a sensible evaluation of a new

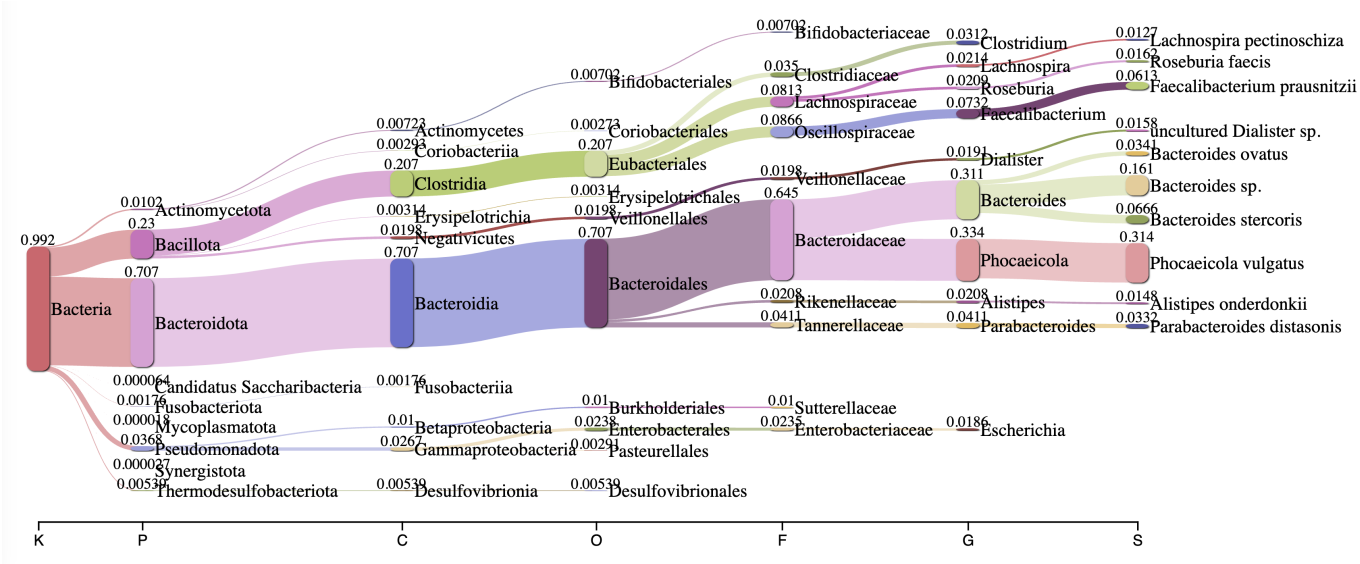

Figure 4: Sankey plot of major taxonomic groups inferred by Lemur from the real gut metagenome.

method and a step towards improved precision in challenging settings.

3.6 Chicken gut metagenome

Additionally, we have analyzed a chicken gut metagenome spanning four major sections of the intestinal tract: duodenum, jejunum, ileum, and colorectum (last two shown in Fig. 5A, B). Similarly to prior studies we have noticed an increase in microbial diversity along the intestinal tract [42, 9]. Furthermore, we noted that the samples from foregut (duodenum, jejunum, ileum) were dominated by a few abundant taxa from the *Lactobacillacea* family (Fig. 5A), while the hindgut showed a more even abundance distribution across a wider set of taxa. These observations match results obtained from metagenomic assembly in the original study [42].

The colorectum shows the most diverse set of taxa (Fig. 5B) including a single archaeal species in class *Methanobacteriota*. Notably, among the 337 assemblies from the original study [42], there is exactly 1 archaeal genome. While, analogously to the human gut metagenome example this data does not definitively assess accuracy of Lemur, it suggests that in practical settings the inferred taxonomic profiles match closely with the prior literature [9, 42].

3.7 Computational performance

Computational performance of Centrifuger, Kraken 2, Melon, MetaMaps, Sourmash and Lemur for each of the four main datasets is captured in Fig. 6. Sourmash, Melon and Lemur have their total RAM requirement consistently under 32 GB, with Lemur requiring more RAM than Melon and Sourmash (Fig. 6G-J). The RAM requirement for MetaMaps exceeds 300 GB, as indicated by an asterisk (*) on the plot. CPU time requirements for the tools are consistent on small and medium-sized datasets (Fig. 6A-E) with Kraken 2 requiring the least time, followed by Lemur. Notably Sourmash and Metamaps require at least an order of magnitude more CPU time (Fig. 6A-E, log-scale y-axis). Finally, post-processing with Magnet doesn't incur much additional cost for small and medium datasets (Fig. 6A-E, time shown for serial execution of Lemur and Magnet). On the large and diverse data (Fig. 6F) the overall trend remains the same with Kraken 2 being the fastest tool, followed by Lemur. However, in this case post-processing with Magnet takes a significant portion of total time due to richer microbial composition of the dataset.

Overall, Kraken 2 is well-known for its high speed processing of metagenomic samples but at the cost of high RAM requirements for larger databases. Lemur has the second lowest CPU time requirement, and performs

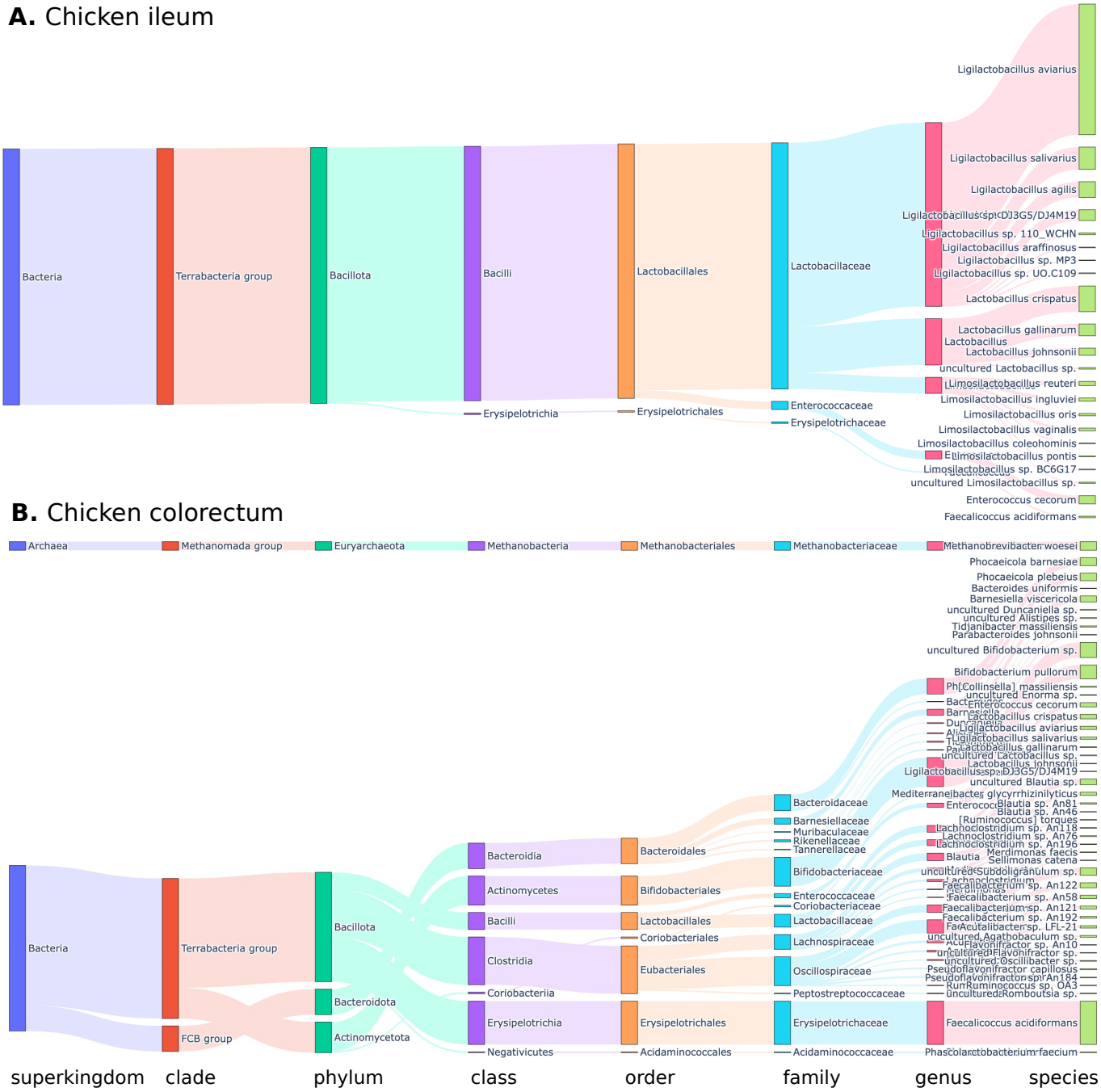

Figure 5: Sankey plot of major taxonomic groups inferred by Lemur from a chicken gut metagenome of two intestinal compartments. Top panel corresponds to the ileum (**A**) and bottom panel represents the colorectum (**B**).

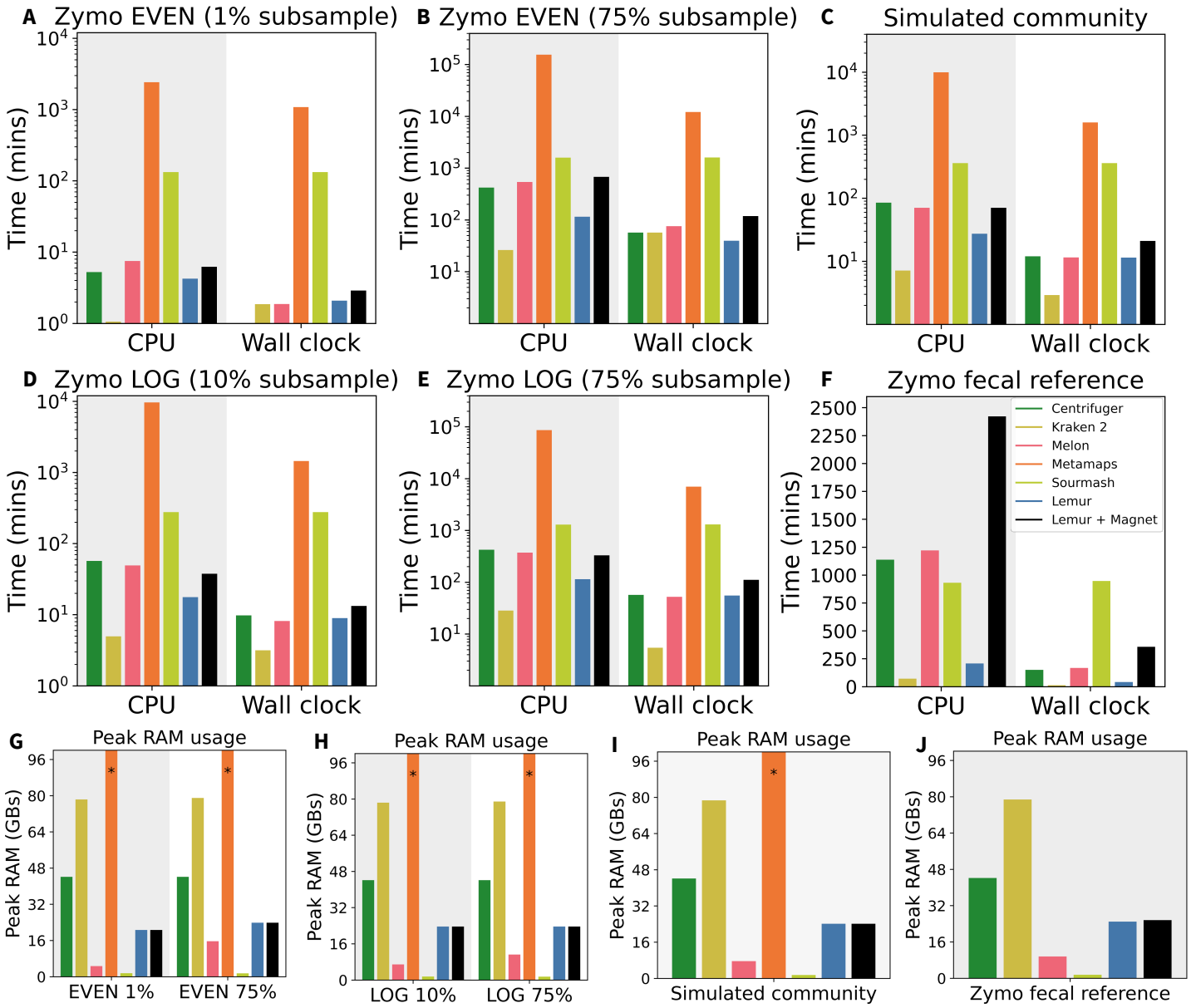

Figure 6: Computational performance metrics evaluated on Zymo EVEN (**A**, **B**, **G**), Zymo LOG (**D**, **E**, **H**), simulated metagenome (**C**, **I**), and Zymo fecal reference (**F**, **J**) datasets. Top six panels show CPU and wall clock times (**A**-**F**), and bottom four panels show peak RAM usage (**G**-**J**).

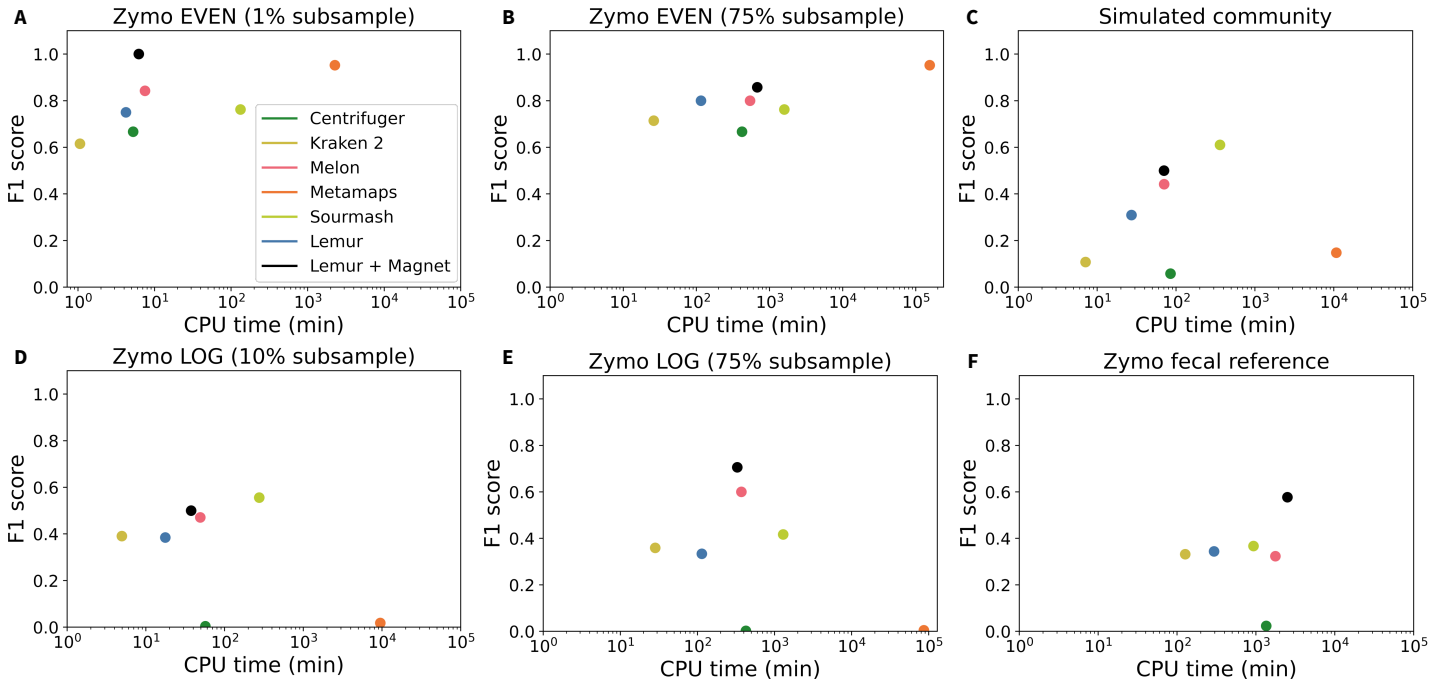

Figure 7: Median F1 score achieved by the tools as the function of required CPU time for processing (on logarithmic scale) for the datasets used in the study.

comparably to Melon in terms of wall clock time for small and medium-sized datasets. For large datasets, Lemur appears to have a larger CPU time advantage over the other methods. Magnet incurs a moderate additional cost for post-processing of the results, and hence is an attractive option for improved precision of the analyses.

Lastly, the feasibility of using Lemur for taxonomic profiling directly on a user laptop was evaluated using two datasets of different sizes. Both were run with Lemur on a MacBook Pro (macOS Sonoma 14.2.1) having a 2.3 GHz Intel Core i9 and 32 GB of RAM using 4 threads. The first dataset was a 3 GB Zymo LOG sample which ran in 13.56 mins (CPU time: 21.99 mins) with peak RAM usage of 17.09 GB. The second was a 77 GB Zymo Fecal Reference sample, which ran in 73.45 mins (CPU time: 206.66 mins) with peak RAM usage of 23.05 GB. Given the real-time and portable nature of long-read sequencing devices, this demonstrates an important element of compatibility with many envisioned use cases in the future.

3.8 F1 score per CPU time unit

For the combination of the datasets and tools evaluated we have also investigated the median achieved F1 score as a function of the CPU time required to analyze the data. We note that for Zymo EVEN 1%, Zymo LOG 75%, and Zymo fecal reference data Lemur + Magnet achieve the highest total F1 score while requiring an amount of CPU time comparable to other tools (Fig. 7A, E, F). For the cases where Lemur + Magnet do not achieve the highest F1 score (Fig. 7B-D) the tools that do require at least an order of magnitude more CPU time.

4 Discussion

Lemur and Magnet represent a novel computational tandem tailored for long-read taxonomic profiling of metagenomes. Lemur exhibits competitive performance by most standard metrics; when it is paired with Magnet, particularly in the presence of low-abundance or low-coverage data, it can improve precision by detecting and filtering out a large number of false positive calls. Our results indicate that Lemur can efficiently process large datasets within minutes to hours in limited resource settings. Furthermore, consistent RAM usage below 32 GB makes it an

attractive software for laptops and lightweight workstations. This reduced resource need does not come at the cost of accuracy; Lemur and Magnet combined offer high recall and precision on the experimental datasets included in our evaluation. Our comparative analysis of real human gut microbiomes indicates that Lemur produces a taxonomic profile consistent with the gut microbiome while avoiding outlier species calls.

Relying on a wide pool of single-copy universal marker genes allows Lemur to achieve high recall and relative abundance estimation accuracy while using only a small portion of the input data. These markers cover all bacteria but only a fraction of any given genome, Magnet does exactly the opposite: it starts with the set of genomes identified by Lemur and tests read-alignment quality and coverage distributions across the genome to make a maximally informed call about whether the putative genome is actually in the sample. Additionally, the inclusion of fungi into the database broadens the scope of the tool and allows for a more comprehensive characterization of metagenomes.

However, as with any method, Lemur has limitations. Reliance on marker genes implies that it cannot generalize to viral genome classification. Also, while Lemur + Magnet can filter out false positives in low-abundance/low-coverage settings, the reliance on the marker genes makes it less sensitive than alternatives like Kraken 2 or MetaMaps that use all reads and the full genomes. Third, the nature of the EM algorithm employed means that it is by necessity a closed-reference method, and thus, a bacteria from a novel, *i.e. out-of-database* family will necessarily be missed by Lemur. Finally, marker gene methods inherently lack the resolution to perform strain-level classification or other sub-species analysis, although the pairing with Magnet suggests one approach for sub-species inference that could change this in the future.

Overall, our results on simulated, synthetic, and real datasets provide experimental support that the combination of Lemur and Magnet represents an efficient and accurate taxonomic profiling workflow explicitly designed for long-read metagenomic datasets. In future work, we intend to expand our benchmark to include additional tools [17, 33, 28] and datasets, thus creating a more complete landscape of long-read metagenomic profilers and classifiers. Furthermore, we acknowledge that while our study focused on taxonomic profiling and binary presence/absence metrics for taxa several of the considered methods (Kraken 2, Centrifuger, MetaMaps) are in fact *metagenomic read classifiers*. Thus, a separate benchmark that focuses on the percentage of classified reads and the proportion of correctly classified reads is warranted to properly assess the accuracy of these methods for their specific tasks.

5 Acknowledgements

Authors would like to thank Austin Marshall for his feedback on the manuscript, and Daniel Portik and Dan Nasko for dataset recommendations. This work is supported in part by funds from the National Science Foundation (NSF: EF-2126387, IIS-2239114, and CNS-1338099), National Institutes of Health (NIH P01-AI152999), and a Centers for Disease Control (CDC) contract 75D30121C11180. K.C. was also supported by Ken Kennedy Institute Computational Science & Engineering Recruiting Fellowship. B.K. was also supported by the National Library of Medicine Training Program in Biomedical Informatics and Data Science (T15LM007093). E.K.M. was supported by the State of Maryland.

References

[1] Daniel P Agustinho, Yilei Fu, Vipin K Menon, Ginger A Metcalf, Todd J Treangen, and Fritz J Sedlazeck. Unveiling microbial diversity: harnessing long-read sequencing technology. *Nature Methods*, pages 1–13, 2024.

[2] Karine Alain, Patricia Pignet, Magali Zbinden, Morgane Quillevere, Francis Duchiron, Jean-Pierre Donval, Françoise Lesongeur, Gérard Raguenes, Philippe Crassous, Joël Querellou, et al. Caminicella sporogenes gen. nov., sp. nov., a novel thermophilic spore-forming bacterium isolated from an East-Pacific Rise

hydrothermal vent. *International journal of systematic and evolutionary microbiology*, 52(5):1621–1628, 2002.

[3] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020.

[4] Advait Balaji, Yunxi Liu, Michael G Nute, Bingbing Hu, Anthony D. Kappell, Danielle S. Lesassier, Gene D. Godbold, Krista Ternus, and Todd Treangen. Seqscreen-nano: a computational platform for streaming, in-field characterization of microbial pathogens. In *Proceedings of the 14th ACM BCB*, pages 1–10, 2023.

[5] Kristen D Curry, Qi Wang, Michael G Nute, Alona Tyshaieva, Elizabeth Reeves, Sirena Soriano, Qinglong Wu, Enid Graeber, Patrick Finzer, Werner Mendling, et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nature Methods*, 19(7):845–853, 2022.

[6] Aaron E Darling, Guillaume Jospin, Eric Lowe, Frederick A Matsen IV, Holly M Bik, and Jonathan A Eisen. Phylosift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243, 2014.

[7] Alexander T Dilthey, Chirag Jain, Sergey Koren, and Adam M Phillippy. Strain-level metagenomic assignment and compositional estimation for long reads with metamaps. *Nature Communications*, 10(1):3066, 2019.

[8] Maria E Hedberg, Edward RB Moore, Liselott Svensson-Stadler, Per Hörstedt, Vladimir Baranov, Olle Hernell, Sun Nyunt Wai, Sten Hammarström, and Marie-Louise Hammarström. Lachnoanaerobaculum gen. nov., a new genus in the lachnospiraceae: characterization of Lachnoanaerobaculum umeaense gen. nov., sp. nov., isolated from the human small intestine, and lachnoanaerobaculum orale sp. nov., isolated from saliva, and reclassification of Eubacterium saburreum (Prevot 1966) Holdeman and Moore 1970 as Lachnoanaerobaculum saburreum comb. nov. *International journal of systematic and evolutionary microbiology*, 62(Pt_11):2685–2690, 2012.

[9] Peng Huang, Yan Zhang, Kangpeng Xiao, Fan Jiang, Hengchao Wang, Dazhi Tang, Dan Liu, Bo Liu, Yisong Liu, Xi He, et al. The chicken gut metagenome and the modulatory effects of plant-derived benzylisoquinoline alkaloids. *Microbiome*, 6:1–17, 2018.

[10] Daniel H Huson, Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Gorska, Dino Jolic, and Rohan BH Williams. Megan-lr: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13:1–17, 2018.

[11] Luiz Irber, Phillip T Brooks, Taylor Reiter, N Tessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, and C Titus Brown. Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *bioRxiv*, pages 2022–01, 2022.

[12] Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114, 2018.

[13] Byung-Chun Kim, Byoung Seung Jeon, Seil Kim, Hyunook Kim, Youngsoon Um, and Byoung-In Sang. Caproiciproducens galactitolivorans gen. nov., sp. nov., a bacterium capable of producing caproic acid from galactitol, isolated from a wastewater treatment plant. *International journal of systematic and evolutionary microbiology*, 65(Pt_12):4902–4908, 2015.

[14] Caroline C Kim, William J Kelly, Mark L Patchett, Gerald W Tannock, Z Jordens, Halina M Stoklosinski, Jordan W Taylor, Ian M Sims, Tracey J Bell, and Douglas I Rosendale. Monoglobus pectinilyticus gen. nov., sp. nov., a pectinolytic bacterium isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, 67(12):4992–4998, 2017.

[15] Chan Yeong Kim, Junyeong Ma, and Insuk Lee. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nature Communications*, 13(1):6367, 2022.

[16] Dongwook Kim, Cameron LM Gilchrist, Jongsik Chun, and Martin Steinegger. UFCG: database of universal fungal core genes and pipeline for genome-wide phylogenetic analysis of fungi. *Nucleic Acids Research*, 51(D1):D777–D784, 2023.

[17] Jaebeom Kim and Martin Steinegger. Metabuli: sensitive and specific metagenomic classification via joint analysis of amino acid and dna. *Nature Methods*, pages 1–3, 2024.

[18] Maki Kitahara, Yuko Shigeno, Mari Shime, Yuki Matsumoto, Shota Nakamura, Daisuke Motooka, Shota Fukuoka, Hiroyoshi Nishikawa, and Yoshimi Benno. Vescimonas gen. nov., Vescimonas coprocola sp. nov., Vescimonas fastidiosa sp. nov., Pusillimonas gen. nov. and Pusillimonas faecalis sp. nov. isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, 71(11):005066, 2021.

[19] Nora Köller, Sarah Hahnke, Vladimir Zverlov, Daniel Wibberg, Andreas Klingl, Tobias Busche, Michael Klocke, Alfred Pühler, Andreas Schlüter, Wolfgang Liebl, et al. Anaeropeptidivorans aminofermentans gen. nov., sp. nov., a mesophilic proteolytic salt-tolerant bacterium isolated from a laboratory-scale biogas fermenter, and emended description of Clostridium colinum. *International Journal of Systematic and Evolutionary Microbiology*, 72(12):005668, 2022.

[20] Raja Lakhal, Nathalie Pradel, Anne Postec, Moktar Hamdi, Bernard Ollivier, Anne Godfroy, and Marie-Laure Fardeau. Vallitalea guaymasensis gen. nov., sp. nov., isolated from marine sediment. *International journal of systematic and evolutionary microbiology*, 63(Pt_8):3019–3023, 2013.

[21] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. 34(18):3094–3100, 2018.

[22] Bo Liu, Theodore Gibbons, Mohammad Ghodsi, Todd Treangen, and Mihai Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 12:1–27, 2011.

[23] Vivien Marx. Method of the year: long-read sequencing. *Nature Methods*, 20(1):6–11, 2023.

[24] Nam-phuong Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. Tipp: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.

[25] Samuel M Nicholls, Joshua C Quick, Shuiquan Tang, and Nicholas J Loman. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience*, 8(5):giz043, 2019.

[26] Yukiteru Ono, Michiaki Hamada, and Kiyoshi Asai. Pbsim3: a simulator for all types of pacbio and ont long reads. *NAR Genomics and Bioinformatics*, 4(4):lqac092, 2022.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

[28] Rafael Peres da Silva, Chayaporn Suphavilai, and Niranjan Nagarajan. Metagenn: a memory-efficient neural network taxonomic classifier robust to sequencing errors and missing genomes. *BMC bioinformatics*, 25(Suppl 1):153, 2024.

[29] Daniel M Portik, C Titus Brown, and N Tessa Pierce-Ward. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*, 23(1):541, 2022.

[30] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[31] Mitsuo Sakamoto, Nao Ikeyama, Atsushi Toyoda, Takumi Murakami, Hiroshi Mori, Sho Morohoshi, Tadao Kunihiro, Takao Iino, and Moriya Ohkuma. Coprobacter secundus subsp. similis subsp. nov. and Solibaculum mannosilyticum gen. nov., sp. nov., isolated from human feces. *Microbiology and Immunology*, 65(6):245–256, 2021.

[32] Nidhi Shah, Erin K Molloy, Mihai Pop, and Tandy Warnow. TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics*, 37(13):1839–1845, 2021.

[33] Jim Shaw and Yun William Yu. Metagenome profiling and containment estimation through abundance-corrected k-mer sketching with sylph. *bioRxiv*, pages 2023–11, 2023.

[34] Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. Seqkit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE*, 11(10):1–10, 2016.

[35] Li Song and Ben Langmead. Centrifuger: lossless compression of microbial genomes for efficient and accurate metagenomic sequence classification. In *RECOMB*, 2024.

[36] Shinichi Sunagawa, Daniel R Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12):1196–1199, 2013.

[37] Mamadou Lamine Tall, Sokhna Ndongo, Issa Isaac Ngom, Jeremy Delerce, S Khelaifia, Didier Raoult, P-E Fournier, and A Levasseur. Massilimicrobiota timonensis gen. nov., sp. nov., a new bacterium isolated from the human gut microbiota. *New Microbes and New Infections*, 31:100574, 2019.

[38] Nathalie Wery, Jean-Marc Moricet, Valerie Cueff, Joelle Jean, Patricia Pignet, Francoise Lesongeur, Marie-Anne Cambon-Bonavita, and Georges Barbier. Caloranaerobacter azorensis gen. nov., sp. nov., an anaerobic thermophilic bacterium isolated from a deep-sea hydrothermal vent. *International journal of systematic and evolutionary microbiology*, 51(5):1789–1796, 2001.

[39] Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20:1–13, 2019.

[40] Martin Wu and Jonathan A Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome biology*, 9:1–11, 2008.

[41] Chen Xi, Xiaole Yin, Xianghui Shi, Weifu Yan, Yu Yang, Lei Liu, and Tong Zhang. Melon: metagenomic long-read-based taxonomic identification and quantification using marker genes. *bioRxiv*, pages 2023–12, 2023.

[42] Yan Zhang, Fan Jiang, Boyuan Yang, Sen Wang, Hengchao Wang, Anqi Wang, Dong Xu, and Wei Fan. Improved microbial genomes and gene catalog of the chicken gut from metagenomic sequencing of high-fidelity long reads. *GigaScience*, 11:giac116, 2022.

A Supplemental materials

A.1 Details of Experimental Study

A.1.1 Benchmarking commands

Lemur was run as:

```
python lemur.py -d path-to-db --tax-path path-to-taxonomy
                -t 8 --minimap2-AS -r species r
                -i path-to-input.fastq --nof -o outdir
```

Kraken 2 was run as:

```
kraken2 --db kraken2-db/k2_fungistd --threads 8
        --output path-to-out.txt --report path-to-out.report.txt
        path-to-input.fastq
```

Melon was run as:

```
melon -d path-to-db -t 8 -o outdir path-to-input.fastq
```

Centrifuger was run as:

```
centrifuger -u path-to-input.fastq
            -x centrifuger-db/cfr_hpv+gbsarscov2
            -t 8 -o outdir
```

A.2 pbsim3 settings

The following command was used for generating reads with pbsim3:

```
pbsim3 --seed 42 --strategy wgs --depth 50.0 --method qshmm
       --qshmm pbsim3/data/QSHMM-ONT.model --length-mean 4050
       --length-sd 1000 --difference-ratio 39:24:36 --accuracy-mean 0.95
       --id-prefix {accession}
        --prefix ../Data/Simulated/Base-reads-MAGs/{accession}/{accession}
       --genome ../Data/Genomes/ncbi_datasets-MAGs/data/{accession}/*.fna
```

A.3 Details of Experimental Results

A.3.1 Additional Analysis of Human Gut Microbes

Table 2: 9 Genera predicted by Kraken2/Melon.

Genus	Isolation Source	Citation
Anaeropeptidivorans	Biogas fermenter	[19]
Caloranaerobacter	Marine thermal vents	[38]
Caminicella	Marine thermal vents	[2]
Vallitalea	Marine sediment	[20]
Caproiciproducens	WW treatment plant	[13]
Lachnoanaerobaculum	Human/Gut feces	[8]
Monoglobus	Human/Gut feces	[14]
Solibaculum	Human/Gut feces	[31]
Vescimonas	Human/Gut feces	[18]