**MetaPhlAn2 for enhanced metagenomic taxonomic profiling**

Duy Tin Truong[1], Eric Franzosa[2,3], Timothy L. Tickle[2,3], Matthias Scholz[1], George Weingart[2], Edoardo Pasolli[1], Adrian Tett[1], Curtis Huttenhower[2,3], and Nicola Segata[1]

[1] Centre for Integrative Biology, University of Trento, Trento 38123, Italy

[2] Biostatistics Department, Harvard School of Public Health, Boston, Massachusetts 02115, USA

[3] The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

Corresponding author: Nicola Segata, nicola.segata@unitn.it

**Supplementary Note 1. Description of the main MetaPhlAn2 additions compared to MetaPhlAn1**

- **Profiling of all domains of life**. Marker and quasi-marker genes are now identified not only for microbes (Bacteria and Archaea), but also for viruses and Eukaryotic microbes (Fungi, Protozoa) that are crucial components of microbial communities.
- **A 6-fold increase in the number of considered species**. Markers are now identified from >16,000 reference genomes and >7,000 unique species, dramatically expanding the comprehensiveness of the method. The new pipeline for identifying marker genes is also scalable to the quickly increasing number of reference genomes. See Supplementary Tables 1-3.
- **Introduction of the concept of quasi-markers, allowing more comprehensive and accurate profiling**. For species with less than 200 markers, MetaPhlAn2 adopts additional quasi-marker sequences (Supplementary Note 2) that are occasionally present in other genomes (because of vertical conservation or horizontal transfer). At profiling time, if no other markers of the potentially confounding species are detected, the corresponding quasi-local markers are used to improve the quality and accuracy of the profiling.
- **Addition of strain-specific barcoding for microbial strain tracking**. MetaPhlAn2 includes a completely new feature that exploits marker combinations to perform species-specific and genus-specific "barcoding" for strains in metagenomic samples (Supplementary Note 7). This feature can be used for culture-free pathogen tracking in epidemiology studies and strain tracking across microbiome samples. See Supplementary Figs. 12-20.
- **Strain-level identification for organisms with sequenced genomes**. For the case in which a microbiome includes strains that are very close to one of those already sequenced, MetaPhlAn2 is now able to identify such strains and readily reports their abundances. See Supplementary Note 7, Supplementary Table 13, and Supplementary Fig. 21.
- **Improvement of false positive and false negative rates**. Improvements in the underlying pipeline for identifying marker genes (including the increment of the adopted genomes and the use of quasi-markers) and the profiling procedure resulted in much improved quantitative performances (higher correlation with true abundances, lower false positive and false negative rates). See the validation on synthetic metagenomes in Supplementary Note 4.
- **Estimation of the percentage of reads mapped against known reference genomes.** MetaPhlAn2 is now able to estimate the number of reads that would map against genomes of each clade detected as present and for which an estimation of its relative abundance is provided by the default output. See Supplementary Note 3 for details.
- **Integration of MetaPhlAn with post-processing and visualization tools**. The MetaPhlAn2 package now includes a set of post-processing and visualization tools ("utils" subfolder of the MetaPhlAn2 repository). Multiple MetaPhlAn profiles can in fact be merged in an abundance table ("merge_metaphlan_tables.py"), exported as BIOM files, visualized as heatmap ("metaphlan_hclust_heatmap.py" or the integrated "hclust2" package), GraPhlAn plots ("export2graphlan.py" and the GraPhlAn package1), Krona2 plots ("metaphlan2krona.py"), and single microbe barplot across samples and conditions ("plot_bug.py").

- **Cloud and Galaxy implementation for integrating MetaPhlAn in metagenomic pipelines**. MetaPhlAn2 is now conveniently available online in the Galaxy platform (e.g. at http://huttenhower.sph.harvard.edu/galaxy) and in the Galaxy Tool Shed and the obtained results can thus be readily post-processed with other Galaxy modules. MetaPhlAn2 is also natively included in cloud-based infrastructures such as Illumina BaseSpace.
- **Use of a fast DNA aligner (BowTie2).** MetaPhlAn2 dropped the direct support of the Blast suite, and is now focused on current high-speed read aligners, in particular BowTie23. This contributed to substantially improve the computational performances (Supplementary Fig. 9).
- **Support for parallelization and external mapping.** MetaPhlAn2 can exploit multiple threads with an almost linear speed-up (Supplementary Fig. 9). The metagenome mapping step can also be performed externally (e.g. by BowTie2) and the result then fed to MetaPhlAn2 as SAM files.
- **Added support for FastQ input files for more accurate mapping.** The per-base quality score included in FastQ formatted files are now used in the mapping procedure to improve the precision of the process.
- **Extended documentation with step-by-step tutorials.** Improved documentation and step-by-step tutorial (http://segatalab.cibio.unitn.it/tools/metaphlan2/) are now available to guide the user.
- **Python3, multiple input type (e.g. SAM), and piping support**. Python 3.x is now supported (in addition to Python 2.x) as well as non FastQ input files such as mapped SAM/BAMs. MetaPhlAn2 also support its inclusion in complex pipeline by accepting the input on the standard input and the used of named pipes.

**Supplementary Note 2. Introduction of quasi-markers sequences in MetaPhlAn2**

The selection of markers is performed by processing the available reference genomes (see Supplementary Tables 1 and 3) with a two-step procedure. First, for each clade, core genes are identified; then, in the second step, core genes with nontrivial homology with genomes from other clades are screened out. For the core gene identification step, the original strategy described for MetaPhlAn14 has been extended to robustly account for misannotated genomes, noisy gene calls, and inconsistencies in the underlying taxonomy as we described elsewhere5,6. Additionally, we also now relax the uniqueness step by considering markers that show a minimal number of sequence hits in genomes outside the clade; such markers are called quasi-markers (see Supplementary Table 2), and their hits to external genomes are stored and used at profiling time. Specifically, a quasi-marker X for a clade A with an external hit to a genome of clade B is considered in estimating the relative abundance of A only if no other (strict) markers for clade B are present. Quasi-markers are ranked based on the number of external hits and are added to the marker set of a clade only if the number of (strict) markers is lower than 200. This allowed us to employ a larger number of markers for those clades with short genomes whose gene set is partially overlapping with other clades, and to be more robust to inconsistencies in the taxonomy or in the genome-associated information. Overall, the MetaPhlAn2 database includes 160,831 quasi-markers (18.3% of the total marker set) with avg 1.39 s.d. 17.2 external hits.

**Supplementary Note 3. Estimating the percentage of reads mapped against known reference genomes**

We introduced an estimation of the number of reads that would map against the genomes of clades with sequenced representatives. This estimation is enabled by the "-t rel_ab_w_read_stats" command line option. In brief, we estimate the RPKM (reads per kilo base per million mapped reads) assigned to each clade based on the (robustly computed) average RPKM of the markers using the core MetaPhlAn2 engine. Clade-specific RPKMs are then multiplied by the average genome length of the sequenced strains in the clade to obtain the average number of reads that would theoretically map against genomes of the clade. This is an interesting information that provides an estimate of the fraction of "microbial dark matter" in each sample without the need of an extensive and computationally unfeasible complete mapping of all reads against all available reference genomes. We illustrate the new this new MetaPhlAn2 features on the 763 HMP samples and 219 HMPII samples (See Supplementary Note 6). The resulting predicted percentage of reads mapped against known reference genomes, is summarized in the boxplot of Supplementary Fig. 22. The median value on the entire set of samples is equal to 47%. Vaginal samples have the highest mappability (median above 90% for posterior fornix) due to the very high abundance of one of four vaginal *Lactobacillus* species with many sequenced genomes. Samples from the oral cavity and the skin have medians above 50% with the exception of the buccal mucosa. Gut samples have a rather small median value (28%), which is lower than the value of 42% found by extensive reads-to-genome mapping7. This is likely due to the fact that in [7] relatively permissive mapping parameters have been used, and to the fact that many reads from uncharacterized species are still mapping against conserved genomic regions of false positive species.

**Supplementary Note 4. Validation of MetaPhlAn2 on synthetic metagenomes**

*Generation of synthetic metagenomes*

We generated 22 synthetic metagenomes datasets of 10 or 40 millions of paired-end reads using SynMetaP8 comprising, in total, 482 bacterial, 80 archaeal, 331 viral, and 88 eukaryotic species. The synthetic metagenome generation was set to simulate Illumina HiSeq 101nt long reads, as the large majority of available metagenomic datasets have these characteristics. Half of these datasets were generated with an even distribution of species abundance, whereas for the other half we adopted a log-normal distribution of the abundances. The synthetic metagenomes also comprised a total of 48 genomes from species not present in the MetaPhlAn2 marker database in order to test the scenario in which the metagenomes include organisms without closely related sequenced genomes. We also included in the validation two

synthetic datasets available in literature9. The characteristics of all the datasets used are presented in Supplementary Table 5 and all the synthetic metagenomes are available at http://goo.gl/5w9XTX.

*Comparative analysis of MetaPhlAn2, MetaPhlAn1, mOTU and Kraken*

We compared MetaPhlAn2 with four other methods: MetaPhlAn14, mOTU[10], Kraken[11] (mini-Kraken version), and Megan5[12]. All methods were evaluated on all the generated synthetic metagenomes except for Megan that was applied on one sample only due to its high computational load**.** All methods were run with their default parameters. We assessed the performances of each method on each dataset using Pearson and Spearman correlation for log-normally distributed datasets, and root mean squared error for evenly distributed ones. In detail, Pearson correlation measures the linear correlation between two variables *X* and *Y* with value interval between +1 and −1, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. Formally, Pearson correlation between two variables *X* and *Y* is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where $cov(X,Y)$ is the covariance between the two variables, and $\sigma_X$ , $\sigma_Y$ are the standard deviation of *X* and *Y*, respectively.
The Spearman correlation coefficient is the Pearson correlation between the ranked variables. For two variables *X* and *Y* with *n* raw scores $X_i, Y_i$, and the corresponding ranks $x_i, y_i$, the Spearman correlation is defined as:

$$\rho_{X,Y} = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)}$$

The root mean squared error measures the difference between the predicted and the true values of a variable. In detail, considered a variable *Y* and given *n* predictions $\hat{y}_i$ and corresponding true values $y_i$, the root mean squared error is computed as:

$$rmse_{Y,\hat{Y}} = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}}$$

Moreover, in this paper we counted as false positive (negative) the case in which the investigated method reports the presence (absence) of a species in the considered sample but is not really present (absent) based on the reference information. Supplementary Table 6 shows the average and standard deviation performance of four methods across 22 synthetic datasets. MetaPhlAn2 outperformed the other methods in terms of Pearson correlation, Spearman correlation and root mean squared error. Additionally, MetaPhlAn2 returned a smaller number of false positive and negative cases. A more detailed comparison is presented in Supplementary Tables 6-12 and Supplementary Figs. 1-8. The comparison with MEGAN5 (reported only for the sample Log_10M_1 in the Supplementary Table 9) showed how this tool is characterized by a low false negative rate at the price of a very high false positive rate and a prohibitive computational load.

**Supplementary Note 5. Metagenomic sequencing of four new elbow metagenomic samples and their profiling with MetaPhlAn2**

*Sample collection, DNA extraction, and Illumina shotgun sequencing*

Samples were collected by moistening cotton tip swabs (VWR, Milan, Italy) in SCF-1 sample buffer (50 mM Tris-HCl, pH 7.5; 1 mM EDTA, pH 8.0; 0.5% Tween-20) and swabbing the external elbow skin area for 30 seconds. To recover the sample the head of the swab was pushed against the side of sterile collection tube. Samples were pre-treated for 30 minutes at 37°C in a lysis solution (20 mM Tris-HCL, pH 8.0; 2 mM EDTA; 1 % Triton X-100) supplemented with Lysozyme (final concentration 20 mg/ml) (Sigma-Aldrich, Milan, Italy) before DNA was isolated with the Mo-Bio PowerSoil DNA

isolation kit (Mo Bio laboratories, Carlsband, CA,USA) as previously described[13]. Libraries were prepared using Illumina Nextera-XT DNA kit (Illumina inc, San Diego, CA, USA) as per manufacturer's instructions. Libraries were pooled and sequenced (101 bp paired end) on the Illumina HiSeq-2000 platform. The study was approved by the Ethical Committee of the University of Trento and informed written consent was obtained from all volunteers. The four samples have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession code 260277.

*Sample preprocessing and MetaPhlAn2 application*

The four generated raw metagenomes were processed with FastqMcf[14] by trimming positions with quality < 15, removing low-quality reads (mean quality < 25), and discarding reads shorter than 80 nt. We obtained 64 M reads (avg 16 M s.d. 6 M per sample) which were reduced to 44 M reads after human DNA and Bacteriophage phiX174 (Illumina spike-ins) removal (avg 11 M s.d. 7M per sample) by BowTie23 mapping against the reference genomes. MetaPhlAn2 was then run on these samples with default parameters.

**Supplementary Note 6. Application of MetaPhlAn2 on the extended HMP/HMPII dataset**

We applied MetaPhlAn2 with default parameters to 763 HMP samples and 219 HMPII samples. The resulting profiled dataset are available on the HMP DACC (http://hmpdacc.org/) and at http://cibiocm.bitbucket.org/data/hmp_hmpii.tar.gz and visually summarized in the heatmap of Supplementary Fig. 11. Briefly, as described in[15], stool samples were shotgun sequenced to a target depth of approximately 5-10 G nt per sample using 101nt Illumina GAIIx reads. Reads were quality trimmed, length filtered, and outlier samples quality controlled. See [13] for a full description of the methods.

**Supplementary Note 7. Strain tracking and identification with MetaPhlAn2**

Strain-level markers are those genes that are uniquely present in the genome of that strain and absent from any other available genome (in the same or other species). These markers can be used to identify strains in the metagenomic samples that are very closely related to sequenced genomes. MetaPhlAn2 performs this identification task by requiring that at least 70% of the markers available for a strain are present in the sample. When applied on real metagenomes, and specifically on the identification of *Bacteroides* strains in the human gut (from HMP samples), this resulted in the identification of strains that share at least 94% of the genome with the identified reference and with less than 0.5% SNP rates. As reported in Supplementary Table 13 and Supplementary Fig. 21, the other strains in the corresponding species have a much lower breadth of coverage (from 75% to 87% on average) and much higher SNP rates (above 1% in most cases). This confirmed the MetaPhlAn2' ability to identify strains in the metagenomic sample that are very closely related with already sequenced strains. For the identified strains, a much deeper analysis of their strain-specific SNPs and functional repertoires can thus be performed using the identified genome as a reference for post-processing analyses.

Strain-level markers also provide a powerful means to perform strain barcoding and tracking. The set of strain-level markers belonging to a species can be seen as the subset of pan-genome genes for that species with the property of not being present in other species. Strains from the same species without sequenced representatives possess a combination of these species-specific pangenome markers that can thus be used to fingerprint them. By focusing on those markers that show variable patterns of presence/absence in different samples, it is then possible to estimate the genomic variability of sample-specific markers and, if patterns in different samples are extremely similar, to hypothesize microbial transmission (if samples belong to different subjects) or microbial retention (if samples belong to the same subject but collected at different time points).

To test this feature of MetaPhlAn2, we generated twelve synthetic read sets (with the noise parameters described above) from six genomes (two sets for each genome at different coverage levels). The genomes were selected from two species (*Bacteroides fragilis and Bacteroides vulgatus*) and represented strains not already included in the generation of

the MetaPhlAn2 database and were thus ideal to test the ability of MetaPhlAn2 in typing and tracking previously unseen organisms. The twelve generated read sets were added to randomly-selected samples (Log_10M_1 to Log_10M_7). These combined samples were then profiled by MetaPhlAn2 and their barcoding strains are plotted as in Supplementary Figs. 19 and 20. MetaPhlAn2 successfully found (almost) identical barcodes for the same (unknown) strains shredded at different coverages and merged with different synthetic metagenomes. At the same time, different strains were clearly discriminated based on the presence/absence of their markers as highlighted by the dendrograms built using simple hierarchical clustering. The only case in which different strains were not clearly distinct in their barcodes occured for two strains (Supplementary Fig. 20) that were indeed almost identical even looking at their original genome directly.

To further illustrate the strain tracking ability of MetaPhlAn2 on the extended HMP dataset, we plotted the presence/absence patterns of the markers from *Prevotella copri* (a known key member in the human gut, Supplementary Fig. 12) and from the three non-*Bacteroides* species and the three *Bacteroides* species most abundant in the human gut: *Alistipes putredinis* (Supplementary Fig. 13), *Eubacterium rectale* (Supplementary Fig. 14), *Parabacteroides merdae* (Supplementary Fig. 15), *Bacteroides ovatus* (Supplementary Fig. 16), *Bacteroides uniformis* (Supplementary Fig. 17), *and Bacteroides vulgatus* (Supplementary Fig. 18). By performing a hierarchical clustering (average linkage, hamming distance) and associating samples with subject IDs and collection time points, strong patterns of strain retention were evident (non-variable markers are removed from the visualization for improving readability).

**Supplementary Table 1. Number of genomes in each domain of life used for MetaPhlAn2 marker definition**

| Kingdom | Number of genomes |
|---|---|
| Archaea | 300 |
| Bacteria | 12926 |
| Viruses | 3565 |
| Eukaryotes | 112 |

**Supplementary Table 2. Marker statistics in each domain of life considered in the MetaPhlAn2 database**

| Kingdom | Markers | | Quasi markers | | |
|---|---|---|---|---|---|
| | Total number | Avg. marker lengths (with s.d.) | Total number | Avg. marker lengths (with s.d.) | Avg. # of genomes violating the uniqueness property (with s.d.) |
| Archaea | 46649 | 669.30 (526.78) | 5613 | 728.81 (605.64) | 9.34 (17.33) |
| Bacteria | 767167 | 661.65 (541.89) | 129614 | 689.03 (643.88) | 6.62 (12.68) |
| Viruses | 38809 | 688.86 (832.05) | 23081 | 1028.60 (1454.94) | 22.50 (107.54) |
| Euk. | 22371 | 1156.54 (962.16) | 2523 | 1141.27 (1336.81) | 4.10 (9.21) |

**Supplementary Table 3. Number of distinct clades at different taxonomic levels considered in the MetaPhlAn2 database**

| Taxonomic levels | Number of different clades |
|---|---|
| Phyla | 50 |
| Classes | 100 |
| Orders | 197 |
| Families | 481 |
| Genera | 1670 |
| Species | 7677 |
| Species (excluding "spp.") | 6500 |
| Strains | 16903 |

**Supplementary Table 4. The number of reads in skin samples sequenced from three subjects**

| Skin samples | Total number of reads | Number of reads after quality control and removal of human DNA and Bacteriophage phiX174 |
|---|---|---|
| Skin_1 | 11,781,066 | 2,878,998 |
| Skin_2 | 9,481,814 | 5,643,960 |
| Skin_3 | 25,653,734 | 21,789,989 |
| Skin_4 | 16,751,638 | 13,486,694 |

**Supplementary Table 5. The list and characteristics of the synthetic metagenomes used in this work**

| Synthetic metagenomes | Number of reads | Read length | Number of species | Abundance distribution |
|---|---|---|---|---|
| Even_10M_1 | 10 M | 101 | 84 | Evenly |
| Even_10M_2 | 10 M | 101 | 87 | Evenly |
| Even_10M_3 | 10 M | 101 | 86 | Evenly |
| Even_10M_4 | 10 M | 101 | 89 | Evenly |
| Even_10M_5 | 10 M | 101 | 80 | Evenly |
| Even_10M_6 | 10 M | 101 | 88 | Evenly |
| Even_10M_7 | 10 M | 101 | 100 | Evenly |
| Even_40M_1 | 40 M | 101 | 150 | Evenly |
| Even_40M_2 | 40 M | 101 | 150 | Evenly |
| Even_40M_3 | 40 M | 101 | 150 | Evenly |
| Even_40M_4 | 40 M | 101 | 150 | Evenly |
| Log_10M_1 | 10 M | 101 | 85 | log-normally |
| Log_10M_2 | 10 M | 101 | 85 | log-normally |
| Log_10M_3 | 10 M | 101 | 85 | log-normally |
| Log_10M_4 | 10 M | 101 | 88 | log-normally |
| Log_10M_5 | 10 M | 101 | 92 | log-normally |
| Log_10M_6 | 10 M | 101 | 85 | log-normally |
| Log_10M_7 | 10 M | 101 | 100 | log-normally |
| Log_40M_1 | 40 M | 101 | 150 | log-normally |
| Log_40M_2 | 40 M | 101 | 150 | log-normally |
| Log_40M_3 | 40 M | 101 | 150 | log-normally |
| Log_40M_4 | 40 M | 101 | 150 | log-normally |

**Supplementary Table 6. Average and standard deviation of the performances achieved by MetaPhlAn2, MetaPhlAn1, mOTUS and Kraken on the log-normally and evenly distributed datasets at the species level. The performance of MetaPhlAn2 is computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes) whereas the other methods are scored on Archaea and Bacteria only**

| | Method \ Dataset | Log datasets | | | Method \ Dataset | Even datasets | |
|---|---|---|---|---|---|---|---|
| | | Average | S.d. | | | Average | S.d. |
| Pearson correlation | MetaPhlAn2 | 0.95 | 0.05 | Root mean squared error | MetaPhlAn2 | 0.34 | 0.08 |
| | MetaPhlAn1 | 0.80 | 0.21 | | MetaPhlAn1 | 1.20 | 0.25 |
| | mOTUs | 0.80 | 0.21 | | mOTUs | 1.10 | 0.24 |
| | Kraken | 0.75 | 0.22 | | Kraken | 1.61 | 0.44 |
| Spearman correlation | MetaPhlAn2 | 0.68 | 0.11 | | | | |
| | MetaPhlAn1 | 0.18 | 0.18 | | | | |
| | mOTUs | 0.30 | 0.19 | | | | |
| | Kraken | 0.22 | 0.16 | | | | |
| False positive | MetaPhlAn2 | 13 | 7 | False positive | MetaPhlAn2 | 10 | 3 |
| | MetaPhlAn1 | 21 | 14 | | MetaPhlAn1 | 25 | 9 |
| | mOTUs | 13 | 10 | | mOTUs | 22 | 15 |
| | Kraken | 20 | 12 | | Kraken | 23 | 10 |
| False positive excluding “unclassified” | MetaPhlAn2 | 5 | 4 | False positive excluding “unclassified” | MetaPhlAn2 | 11 | 10 |
| | MetaPhlAn1 | 12 | 10 | | MetaPhlAn1 | 24 | 19 |
| | mOTUs | 13 | 10 | | mOTUs | 22 | 15 |
| | Kraken | 20 | 12 | | Kraken | 23 | 10 |
| False negative | MetaPhlAn2 | 33 | 10 | False negative | MetaPhlAn2 | 12 | 10 |
| | MetaPhlAn1 | 35 | 15 | | MetaPhlAn1 | 29 | 16 |
| | mOTUs | 33 | 13 | | mOTUs | 27 | 14 |
| | Kraken | 33 | 15 | | Kraken | 27 | 13 |

**Supplementary Table 7. Comparative results of the application of MetaPhlAn2, MetaPhlAn1, mOTUs, and Kraken on the log-normally distributed metagenomes in profiling the archaeal and bacterial organisms at the species level**

| | Method \ Dataset | Log 10M_1 | Log 10M_2 | Log 10M_3 | Log 10M_4 | Log 10M_5 | Log 10M_6 | Log 10M_7 | Log 40M_1 | Log 40M_2 | Log 40M_3 | Log 40M_4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pearson correlation | MetaPhlAn2 | 0.99 | 1.00 | 0.99 | 0.98 | 0.90 | 0.97 | 1.00 | 0.97 | 0.96 | 0.79 | 1.00 |
| | MetaPhlAn1 | 0.92 | 0.95 | 1.00 | 0.86 | 0.86 | 0.79 | 0.95 | 0.78 | 0.65 | 0.87 | <0.50 |
| | mOTUs | 0.93 | 0.95 | 1.00 | 0.86 | 0.92 | 0.79 | 0.95 | 0.83 | 0.69 | 0.74 | <0.50 |
| | Kraken | 0.84 | 0.93 | 1.00 | 0.81 | 0.88 | 0.76 | 0.67 | 0.59 | 0.82 | 0.82 | <0.50 |
| Spearman correlation | MetaPhlAn2 | 0.73 | 0.92 | 0.73 | 0.59 | 0.63 | 0.77 | 0.78 | <0.50 | 0.55 | 0.86 | 0.88 |
| | MetaPhlAn1 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 |
| | mOTUs | <0.50 | 0.66 | 0.52 | <0.50 | 0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 |
| | Kraken | <0.50 | 0.51 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 |
| False positive | MetaPhlAn2 | 4 | 2 | 4 | 7 | 4 | 5 | 11 | 23 | 16 | 9 | 19 |
| | MetaPhlAn1 | 12 | 11 | 6 | 7 | 11 | 9 | 23 | 47 | 30 | 31 | 41 |
| | mOTUs | 3 | 2 | 1 | 5 | 7 | 10 | 19 | 36 | 15 | 19 | 24 |
| | Kraken | 10 | 11 | 8 | 3 | 10 | 32 | 42 | 28 | 14 | 23 | 35 |
| False positive excluding “unclassified” | MetaPhlAn2 | 0 | 0 | 0 | 0 | 1 | 1 | 4 | 13 | 5 | 2 | 7 |
| | MetaPhlAn1 | 8 | 4 | 1 | 3 | 8 | 4 | 11 | 34 | 16 | 17 | 27 |
| | mOTUs | 3 | 2 | 1 | 5 | 7 | 10 | 19 | 36 | 15 | 19 | 24 |
| | Kraken | 10 | 11 | 8 | 3 | 10 | 32 | 42 | 28 | 14 | 23 | 35 |
| False negative | MetaPhlAn2 | 17 | 12 | 14 | 21 | 13 | 16 | 22 | 27 | 33 | 5 | 6 |
| | MetaPhlAn1 | 23 | 18 | 23 | 28 | 16 | 24 | 44 | 55 | 63 | 47 | 40 |
| | mOTUs | 26 | 16 | 26 | 28 | 19 | 24 | 42 | 52 | 58 | 42 | 31 |
| | Kraken | 21 | 17 | 22 | 26 | 19 | 21 | 44 | 54 | 59 | 48 | 37 |

**Supplementary Table 8. Comparative results of the application of MetaPhlAn2 and Kraken on the log-normally distributed metagenomes in profiling the viral and eukaryotic organisms at the species level (the other methods are not able to detect viruses and eukaryote)**

| | Method \ Dataset | Log 10M_1 | Log 10M_2 | Log 10M_3 | Log 10M_4 | Log 10M_5 | Log 10M_6 | Log 10M_7 |
|---|---|---|---|---|---|---|---|---|
| Pearson correlation | MetaPhlAn2 | 0.98 | 0.95 | 0.95 | 1.00 | 0.90 | 0.94 | 1.00 |
| | Kraken | -0.02 | 0.89 | 0.00 | 0.98 | 0.75 | 0.72 | 0.89 |
| Spearman correlation | MetaPhlAn2 | 0.37 | 0.71 | 0.57 | 0.74 | 0.38 | 0.73 | 0.76 |
| | Kraken | 0.17 | 0.38 | -0.16 | 0.24 | 0.35 | 0.33 | 0.51 |
| False positive | MetaPhlAn2 | 7 | 3 | 4 | 1 | 5 | 2 | 2 |
| | Kraken | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| False positive excluding “unclassified” | MetaPhlAn2 | 5 | 1 | 2 | 0 | 2 | 1 | 1 |
| | Kraken | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| False negative | MetaPhlAn2 | 22 | 18 | 25 | 22 | 24 | 21 | 5 |
| | Kraken | 39 | 39 | 45 | 42 | 45 | 41 | 19 |

**Supplementary Table 9. Comparative results of the application of MetaPhlAn2, MetaPhlAn1, mOTUs, Kraken, and MEGAN5 on the log-normally distributed metagenomes in profiling the archaeal, bacterial, viral and eukaryotic organisms at the species level**

| | Method \ Dataset | Log 10M_1 | Log 10M_2 | Log 10M_3 | Log 10M_4 | Log 10M_5 | Log 10M_6 | Log 10M_7 | Log 40M_1 | Log 40M_2 | Log 40M_3 | Log 40M_4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pearson correlation | MetaPhlAn2 | 0.98 | 0.95 | 0.99 | 1.00 | 0.90 | 0.94 | 1.00 | 0.97 | 0.97 | 0.81 | 0.99 |
| | MetaPhlAn1 | <0.50 | <0.50 | 1.00 | <0.50 | 0.86 | 0.73 | 0.91 | 0.72 | <0.50 | 0.86 | <0.50 |
| | mOTUs | <0.50 | <0.50 | 1.00 | <0.50 | 0.91 | 0.74 | 0.91 | 0.77 | <0.50 | 0.73 | <0.50 |
| | Kraken | <0.50 | <0.50 | 1.00 | 0.61 | 0.88 | 0.70 | 0.64 | 0.55 | <0.50 | 0.80 | <0.50 |
| | MEGAN5 | 0.84 | - | - | - | - | - | - | - | - | - | - |
| Spearman correlation | MetaPhlAn2 | 0.57 | 0.80 | 0.64 | 0.66 | 0.53 | 0.75 | 0.78 | 0.56 | 0.61 | 0.78 | 0.85 |
| | MetaPhlAn1 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 |
| | mOTUs | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 |
| | Kraken | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 | <0.50 |
| | MEGAN5 | <0.50 | - | - | - | - | - | - | - | - | - | - |
| False positive | MetaPhlAn2 | 11 | 5 | 8 | 8 | 9 | 7 | 13 | 26 | 20 | 14 | 22 |
| | MetaPhlAn1 | 12 | 11 | 6 | 7 | 11 | 9 | 23 | 47 | 30 | 31 | 41 |
| | mOTUs | 3 | 2 | 1 | 5 | 7 | 10 | 19 | 36 | 15 | 19 | 24 |
| | Kraken | 10 | 11 | 10 | 4 | 10 | 32 | 42 | 29 | 14 | 23 | 36 |
| | MEGAN5 | >100 | - | - | - | - | - | - | - | - | - | - |
| False positive excluding “unclassified” | MetaPhlAn2 | 5 | 1 | 2 | 0 | 3 | 2 | 5 | 13 | 5 | 5 | 9 |
| | MetaPhlAn1 | 8 | 4 | 1 | 3 | 8 | 4 | 11 | 34 | 16 | 17 | 27 |
| | mOTUs | 3 | 2 | 1 | 5 | 7 | 10 | 19 | 36 | 15 | 19 | 24 |
| | Kraken | 10 | 11 | 10 | 4 | 10 | 32 | 42 | 29 | 14 | 23 | 36 |
| | MEGAN5 | >100 | - | - | - | - | - | - | - | - | - | - |
| False negative | MetaPhlAn2 | 39 | 30 | 39 | 43 | 37 | 37 | 27 | 35 | 47 | 22 | 10 |
| | MetaPhlAn1 | 63 | 61 | 69 | 73 | 64 | 67 | 65 | 105 | 113 | 97 | 90 |
| | mOTUs | 66 | 59 | 72 | 73 | 67 | 67 | 63 | 102 | 108 | 92 | 81 |
| | Kraken | 60 | 56 | 67 | 68 | 64 | 62 | 63 | 103 | 106 | 96 | 83 |
| | MEGAN5 | 8 | - | - | - | - | - | - | - | - | - | - |

**Supplementary Table 10. Comparative results of the application of MetaPhlAn2, MetaPhlAn1, mOTUs, and Kraken on the evenly distributed metagenomes in profiling the archaeal and bacterial organisms at the species level**

| | Method \ Dataset | Even 10M_1 | Even 10M_2 | Even 10M_3 | Even 10M_4 | Even 10M_5 | Even 10M_6 | Even 10M_7 | Even 40M_1 | Even 40M_2 | Even 40M_3 | Even 40M_4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Root mean square error | MetaPhlAn2 | 0.59 | 0.39 | 0.51 | 0.63 | 0.40 | 0.35 | 0.30 | 0.54 | 0.56 | 0.19 | 0.27 |
| | MetaPhlAn1 | 1.39 | 1.43 | 1.47 | 1.40 | 1.37 | 1.44 | 1.09 | 0.96 | 0.98 | 0.85 | 0.79 |
| | mOTUs | 1.30 | 1.43 | 1.35 | 1.21 | 1.28 | 1.26 | 0.93 | 0.90 | 0.92 | 0.78 | 0.72 |
| | Kraken | 2.07 | 2.02 | 2.02 | 1.93 | 1.85 | 2.12 | 1.42 | 1.20 | 1.11 | 0.99 | 1.03 |
| False positive | MetaPhlAn2 | 6 | 7 | 6 | 5 | 8 | 6 | 11 | 17 | 11 | 3 | 7 |
| | MetaPhlAn1 | 13 | 25 | 16 | 18 | 18 | 27 | 23 | 44 | 26 | 27 | 35 |
| | mOTUs | 6 | 20 | 6 | 8 | 15 | 13 | 21 | 56 | 43 | 28 | 27 |
| | Kraken | 12 | 30 | 9 | 12 | 14 | 23 | 34 | 39 | 35 | 20 | 24 |
| False positive excluding “unclassified” | MetaPhlAn2 | 3 | 2 | 2 | 1 | 3 | 2 | 6 | 29 | 22 | 11 | 17 |
| | MetaPhlAn1 | 6 | 15 | 5 | 9 | 9 | 16 | 13 | 61 | 43 | 41 | 50 |
| | mOTUs | 6 | 20 | 6 | 8 | 15 | 13 | 21 | 57 | 44 | 28 | 27 |
| | Kraken | 12 | 30 | 9 | 12 | 14 | 23 | 34 | 39 | 35 | 20 | 24 |
| False negative | MetaPhlAn2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 26 | 26 | 2 | 3 |
| | MetaPhlAn1 | 14 | 19 | 13 | 15 | 14 | 22 | 25 | 54 | 56 | 44 | 38 |
| | mOTUs | 13 | 19 | 16 | 11 | 15 | 20 | 26 | 51 | 52 | 39 | 30 |
| | Kraken | 11 | 20 | 14 | 16 | 13 | 22 | 30 | 48 | 50 | 35 | 34 |

**Supplementary Table 11. Comparative results of the application of MetaPhlAn2 and Kraken on the evenly distributed metagenomes in profiling the viral and eukaryotic organisms at the species level**

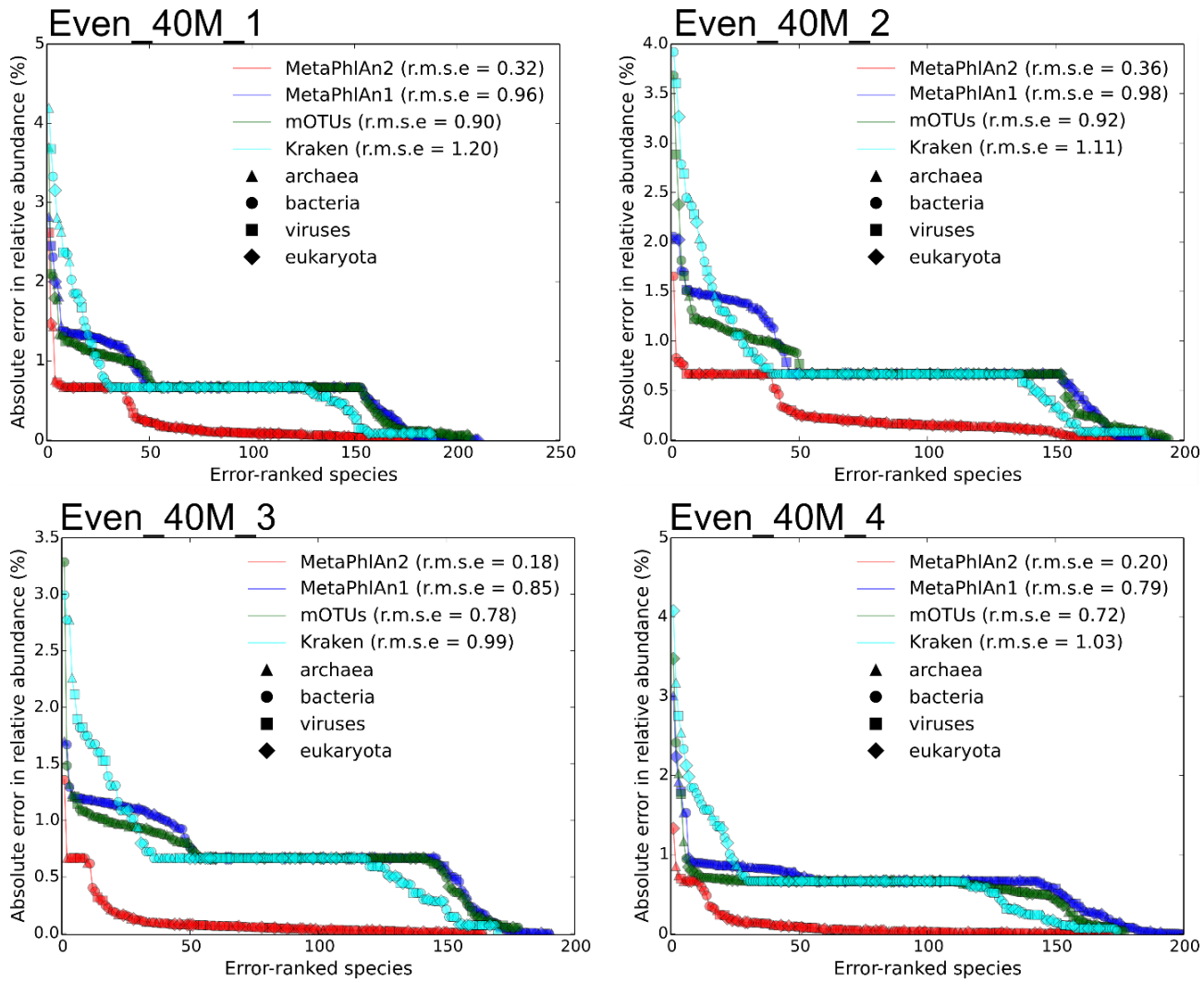
| | Method \ Dataset | Even 10M_1 | Even 10M_2 | Even 10M_3 | Even 10M_4 | Even 10M_5 | Even 10M_6 | Even 10M_7 |
|---|---|---|---|---|---|---|---|---|
| Root mean square error | MetaPhlAn2 | 1.20 | 1.22 | 1.13 | 1.04 | 1.32 | 0.98 | 1.39 |
| | Kraken | 11.13 | 15.43 | 7.67 | 14.58 | 11.66 | 15.81 | 12.67 |
| False positive | MetaPhlAn2 | 5 | 2 | 4 | 4 | 4 | 2 | 3 |
| | Kraken | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| False positive excluding "unclassified" | MetaPhlAn2 | 1 | 1 | 1 | 1 | 2 | 0 | 0 |
| | Kraken | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| False negative | MetaPhlAn2 | 8 | 6 | 7 | 8 | 7 | 4 | 2 |
| | Kraken | 41 | 40 | 40 | 45 | 37 | 38 | 27 |

**Supplementary Table 12. Comparative results of the application of MetaPhlAn2, MetaPhlAn1, mOTUs, and Kraken on the evenly distributed metagenomes in profiling the archaeal, bacterial, viral and eukaryotic organisms at the species level**

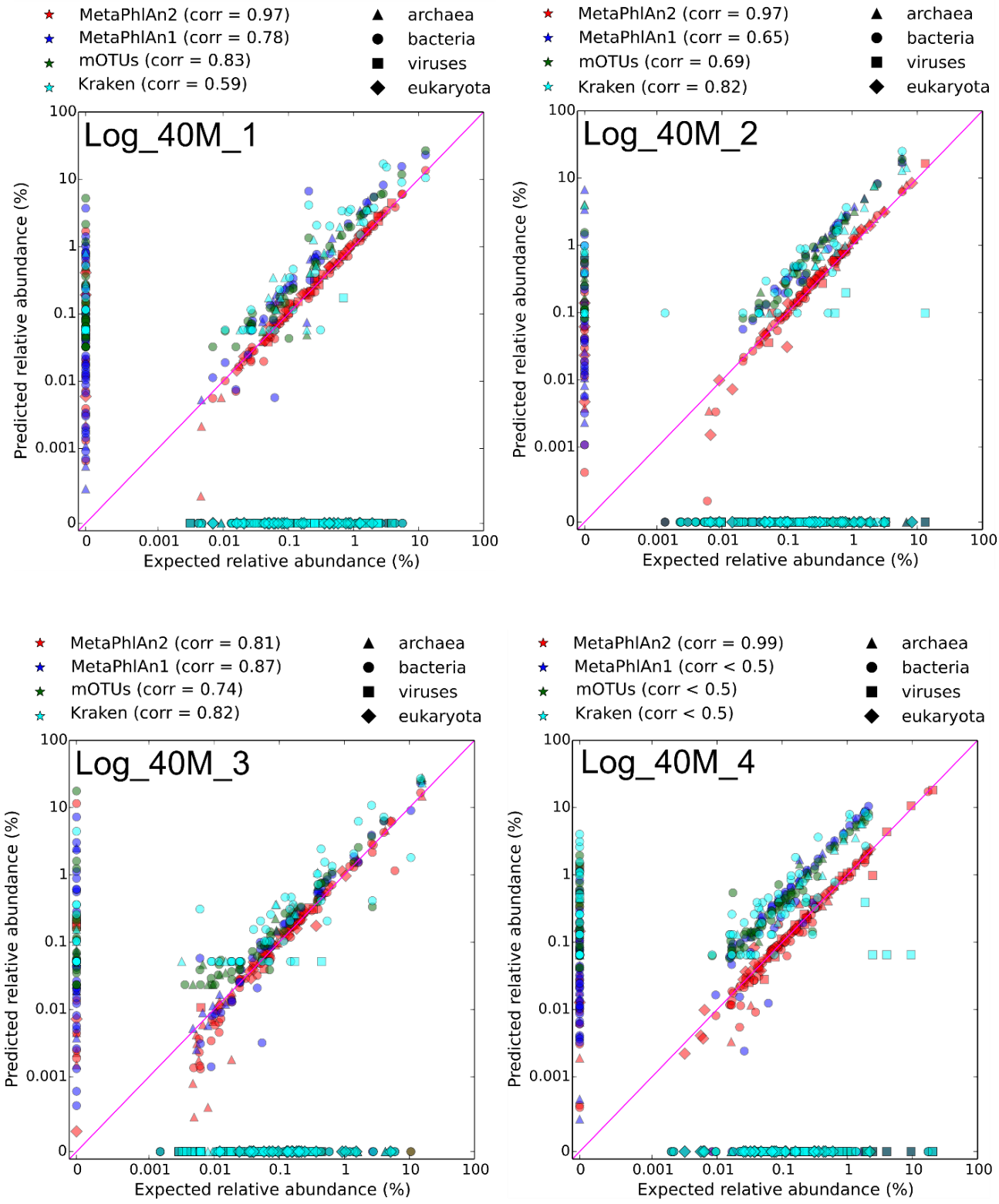
| | Method \ Dataset | Even 10M_1 | Even 10M_2 | Even 10M_3 | Even 10M_4 | Even 10M_5 | Even 10M_6 | Even 10M_7 | Even 40M_1 | Even 40M_2 | Even 40M_3 | Even 40M_4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Root mean square error | MetaPhlAn2 | 0.44 | 0.38 | 0.41 | 0.42 | 0.39 | 0.32 | 0.32 | 0.32 | 0.36 | 0.18 | 0.20 |
| | MetaPhlAn1 | 1.75 | 1.52 | 1.59 | 1.48 | 1.73 | 1.52 | 1.07 | 0.86 | 0.89 | 0.81 | 0.75 |
| | mOTUs | 1.77 | 1.42 | 1.53 | 1.45 | 1.70 | 1.38 | 0.92 | 0.82 | 0.84 | 0.77 | 0.73 |
| | Kraken | 1.79 | 1.79 | 1.88 | 1.84 | 1.84 | 1.96 | 1.32 | 1.06 | 0.99 | 0.92 | 0.95 |
| False positive | MetaPhlAn2 | 11 | 9 | 10 | 9 | 12 | 8 | 14 | 17 | 12 | 4 | 9 |
| | MetaPhlAn1 | 13 | 25 | 16 | 18 | 18 | 27 | 23 | 44 | 26 | 27 | 35 |
| | mOTUs | 6 | 20 | 6 | 8 | 15 | 13 | 21 | 56 | 43 | 28 | 27 |
| | Kraken | 12 | 30 | 9 | 12 | 14 | 23 | 34 | 39 | 35 | 20 | 24 |
| False positive excluding "unclassified" | MetaPhlAn2 | 4 | 3 | 3 | 2 | 5 | 2 | 6 | 32 | 26 | 14 | 21 |
| | MetaPhlAn1 | 6 | 15 | 5 | 9 | 9 | 16 | 13 | 61 | 43 | 41 | 50 |
| | mOTUs | 6 | 20 | 6 | 8 | 15 | 13 | 21 | 57 | 44 | 28 | 27 |
| | Kraken | 12 | 30 | 9 | 12 | 14 | 23 | 34 | 39 | 35 | 20 | 24 |
| False negative | MetaPhlAn2 | 10 | 7 | 8 | 9 | 8 | 5 | 4 | 30 | 34 | 8 | 6 |
| | MetaPhlAn1 | 57 | 60 | 57 | 61 | 53 | 61 | 54 | 104 | 106 | 94 | 88 |
| | mOTUs | 56 | 60 | 60 | 57 | 54 | 59 | 55 | 101 | 102 | 89 | 80 |
| | Kraken | 52 | 60 | 54 | 61 | 50 | 60 | 57 | 96 | 98 | 83 | 82 |

**Supplementary Table 13. An example of strain identification for *Bacteroides* strains on the gut HMP samples. The samples in which MetaPhlAn2 consistently detects a given *Bacteroides* strain are reported and complemented for validation purposes with their breadth of coverage (i.e., percentage of the strain genome covered by reads) and the best and average (with s.d.) breadth of coverage for all the other available genomes in the same species. The number of single nucleotide polymorphism (SNPs, by comparison of mapping consensus with the reference genome) are also reported for the detected strain and all the other strains in the species**

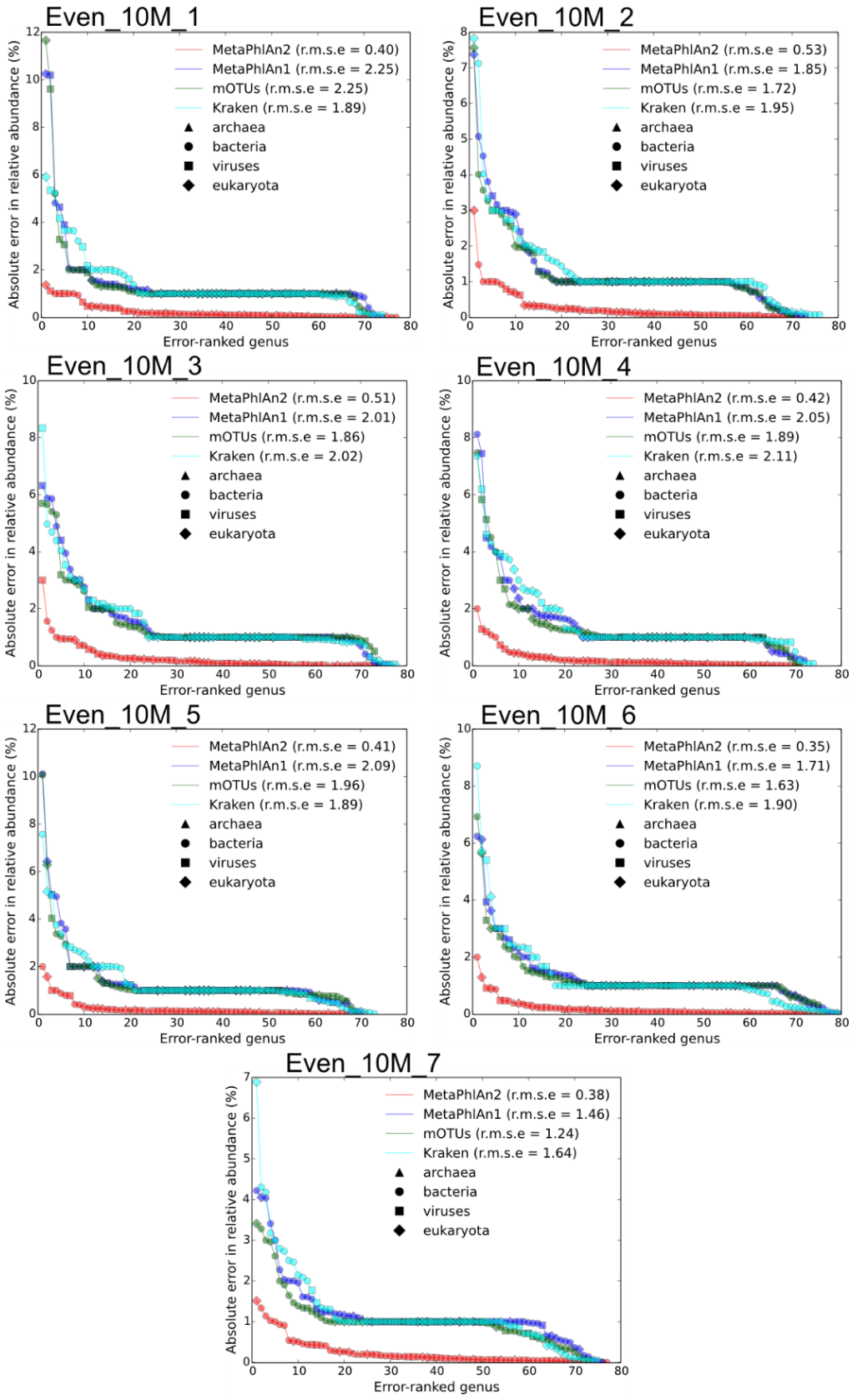
| Sample ID | Identified Strain | Cov. breadth | Best breadth (other strains) | Avg. breadth (other strains) | # of SNPs | Min # SNPs (other strains) | Avg. # of SNPs (other strains) |
|---|---|---|---|---|---|---|---|
| SRS011586 | *B. finegoldii* DSM 17565 | 0.94 | 0.75 | 0.75±0.00 | 22131 | 106486 | 106829±343 |
| SRS011586 | *B. fragilis* HMW 616 | 0.94 | 0.85 | 0.57±0.11 | 33002 | 64117 | 186206±52058 |
| SRS011586 | *B. uniformis* ATCC 8492 | 0.95 | 0.81 | 0.79±0.02 | 8700 | 52449 | 56640±5548 |
| SRS012273 | *B. ovatus* CL03T12C18 | 0.92 | 0.78 | 0.76±0.02 | 80767 | 148729 | 160700±9318 |
| SRS012273 | *B. uniformis* ATCC 8492 | 0.94 | 0.79 | 0.78±0.01 | 17182 | 62025 | 65538±4482 |
| SRS014683 | *B. ovatus* CL03T12C18 | 0.94 | 0.85 | 0.83±0.02 | 14223 | 101081 | 114476±12741 |
| SRS015133 | *B. fragilis* HMW 615 | 0.95 | 0.87 | 0.76±0.18 | 15343 | 41570 | 70400±49714 |
| SRS016203 | *B. uniformis* ATCC 8492 | 0.98 | 0.81 | 0.79±0.03 | 10647 | 53407 | 57115±4348 |
| SRS016267 | *B. ovatus* SD CMC 3f | 0.97 | 0.79 | 0.76±0.03 | 11576 | 102563 | 111011±6418 |
| SRS019030 | *B. ovatus* CL03T12C18 | 0.94 | 0.87 | 0.85±0.02 | 17440 | 104773 | 117105±11462 |
| SRS019787 | *B. ovatus* SD CMC 3f | 0.97 | 0.79 | 0.76±0.03 | 8224 | 110630 | 117565±6295 |
| SRS049900 | *B. uniformis* ATCC 8492 | 0.96 | 0.79 | 0.76±0.04 | 28353 | 67084 | 70203±2270 |
| SRS064645 | *B. uniformis* ATCC 8492 | 0.96 | 0.78 | 0.77±0.02 | 15463 | 57678 | 60234±2765 |



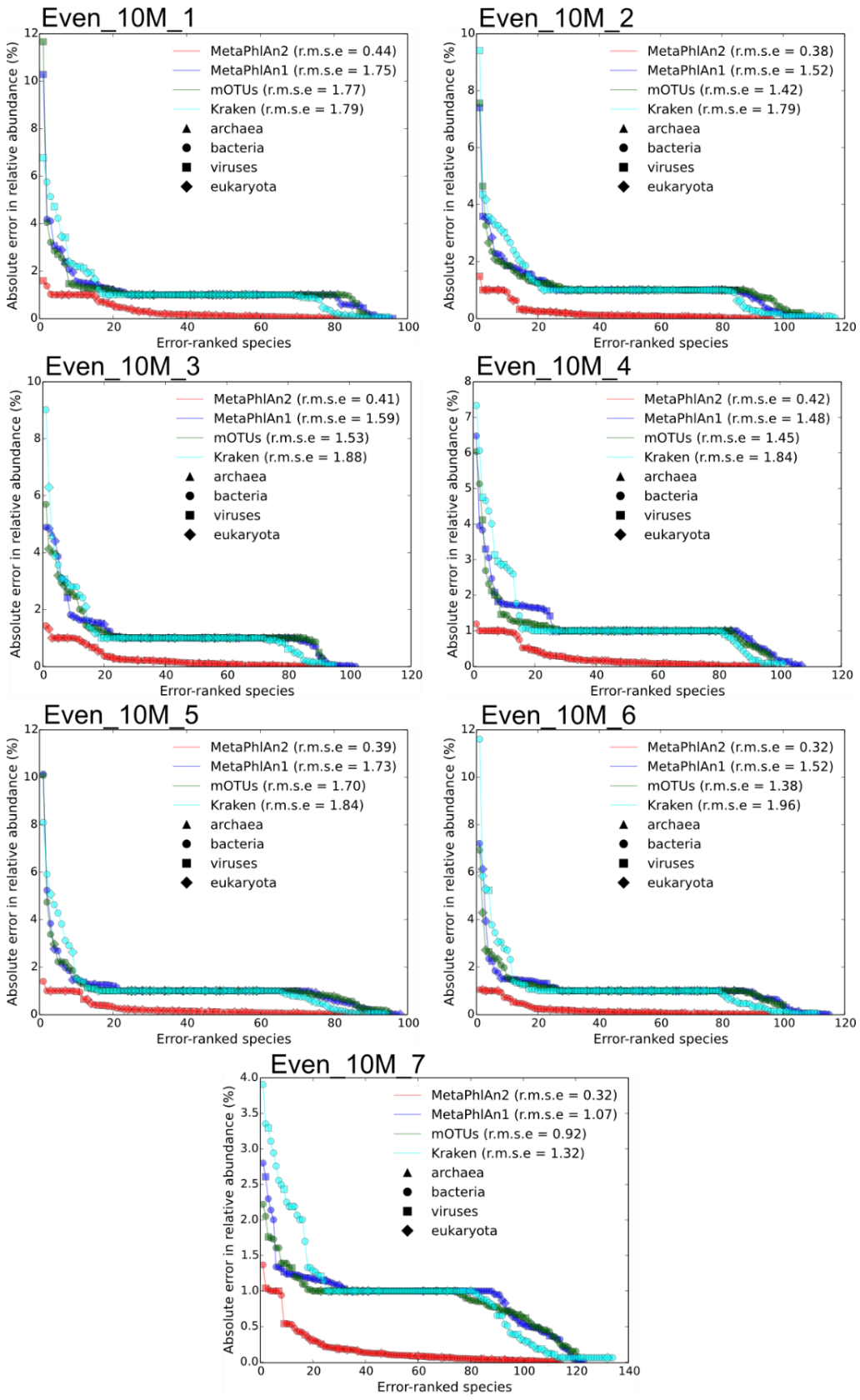

**Supplementary Fig. 1. Performance comparison of the four tested methods on evenly distributed 40M-read datasets at the species level based on the ranked root mean squared error (r.m.s.e). The performance of MetaPhlAn2 is computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes) whereas the other methods are scored on Archaea and Bacteria only**



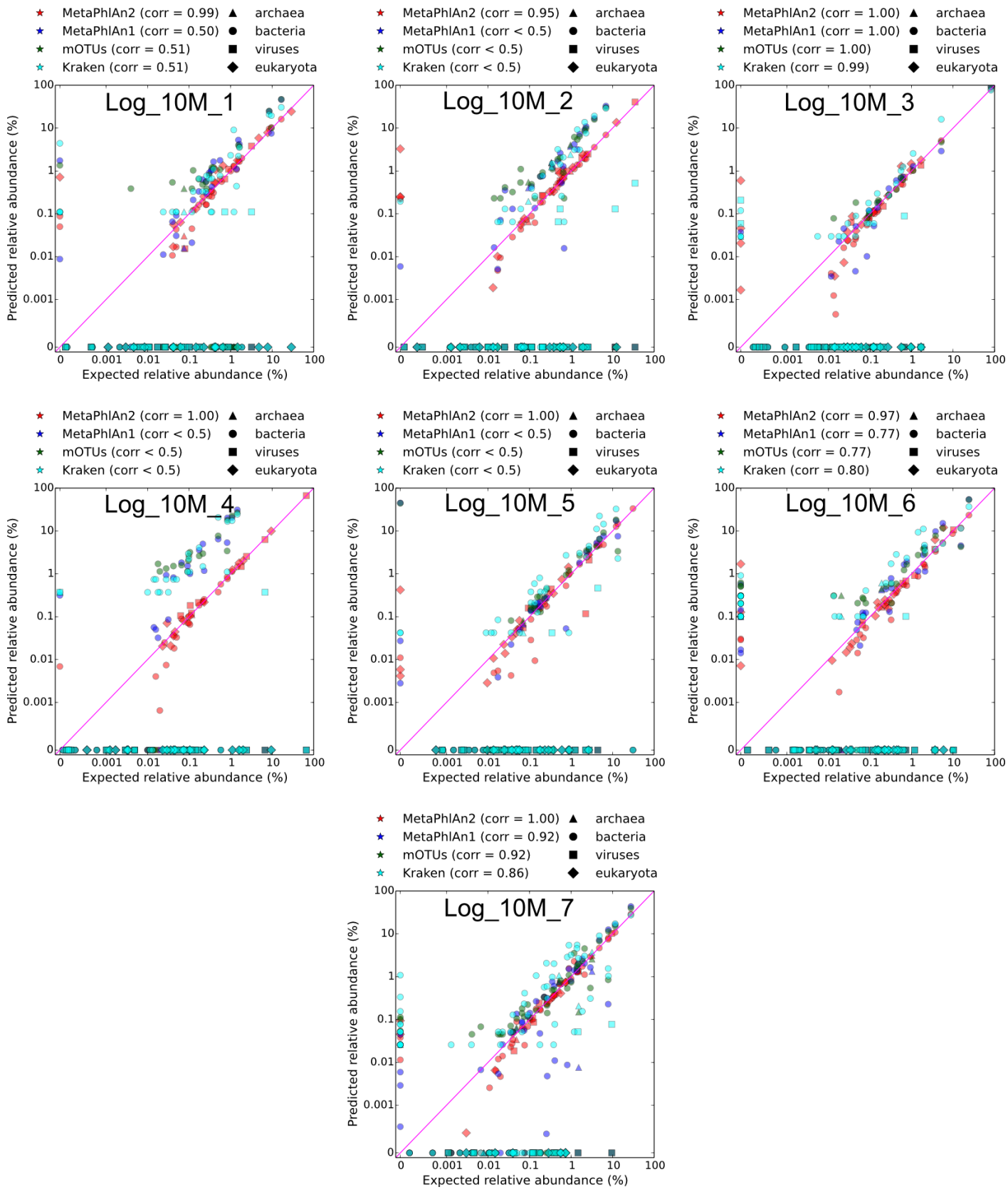

**Supplementary Fig. 2. Performance comparison of the four tested methods on log-normally distributed 40M-read datasets at the species level based on the Pearson correlation (corr). The performance of MetaPhlAn2 is computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes) whereas the other methods are scored on Archaea and Bacteria only**



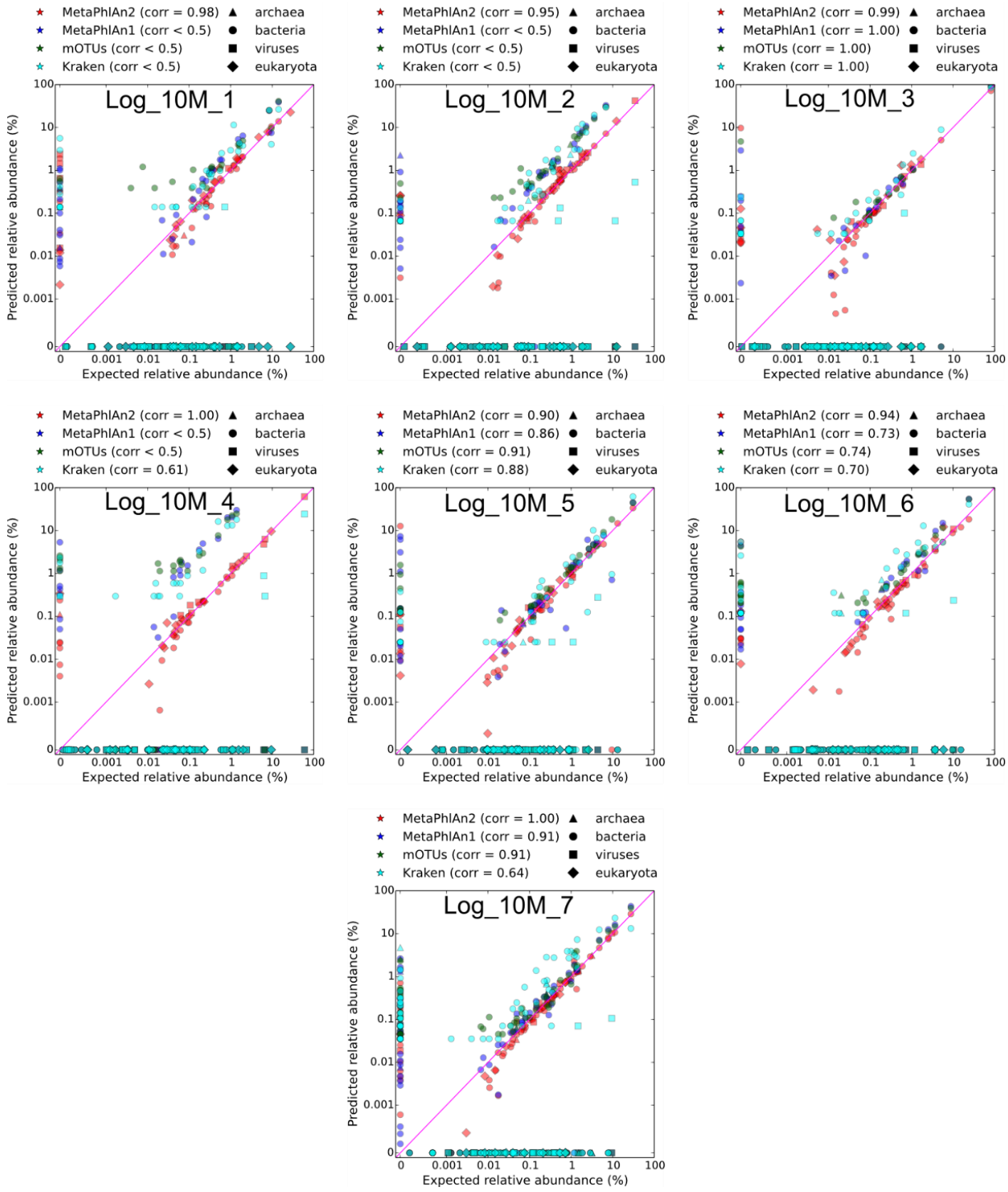

**Supplementary Fig. 3. Performance comparison of the four tested methods on evenly distributed 10M-read datasets at the genus level based on the ranked root mean squared error (r.m.s.e). The performance of all methods are computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes)**



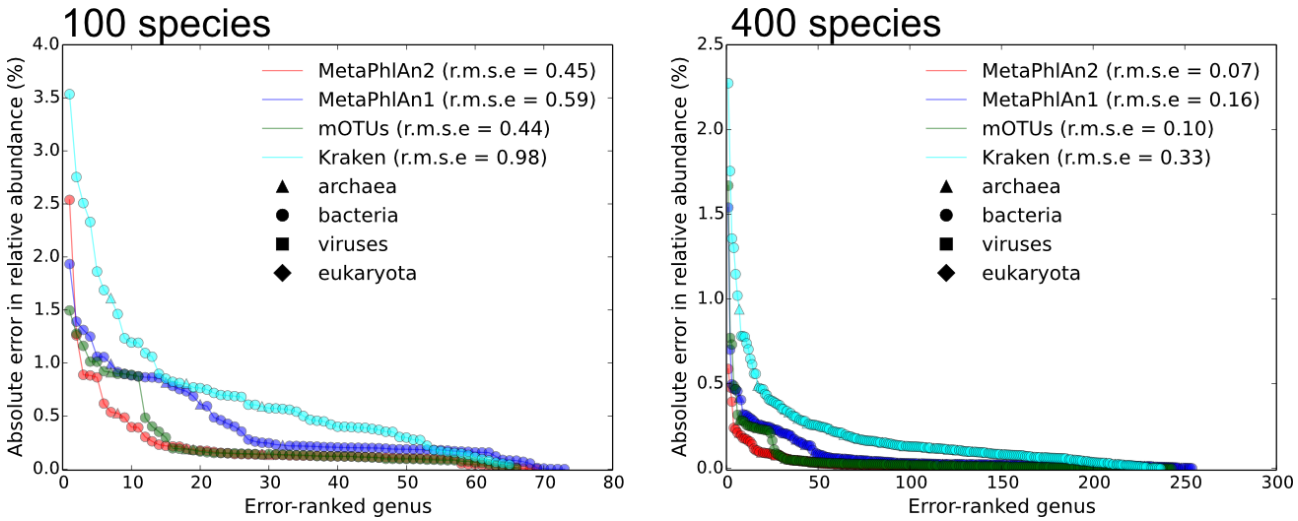

**Supplementary Fig. 4. Performance comparison of the four tested methods on evenly distributed 10M-read datasets at the species level based on the ranked root mean squared error (r.m.s.e). The performance of all methods are computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes)**



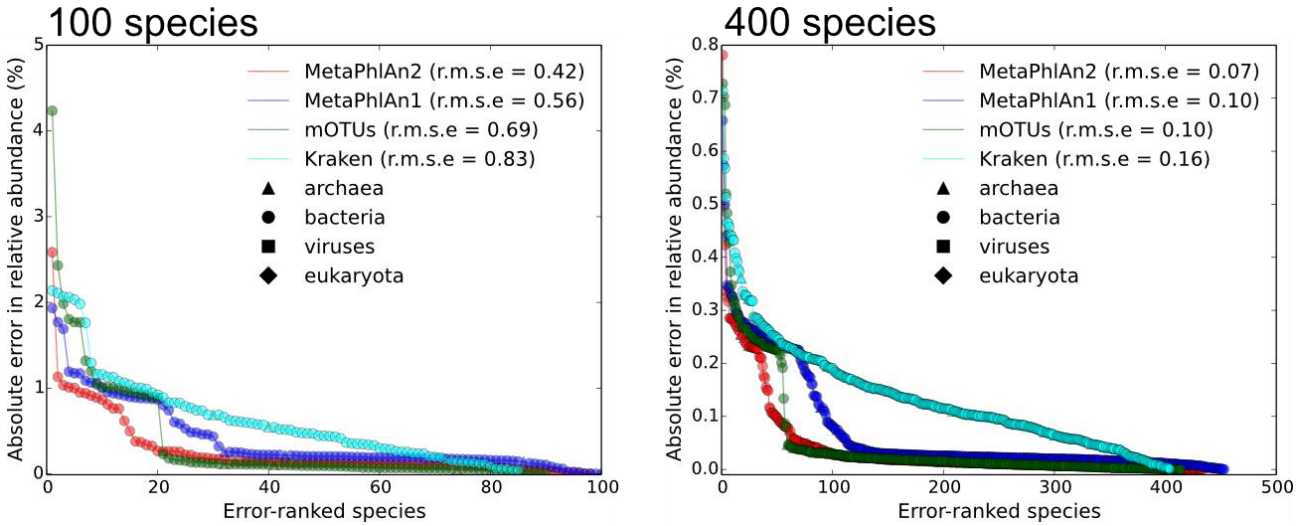

**Supplementary Fig. 5. Performance comparison of the four tested methods on log-normally distributed 10M-read datasets at the genus level based on the Pearson correlation (corr). The performance of all methods are computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes)**



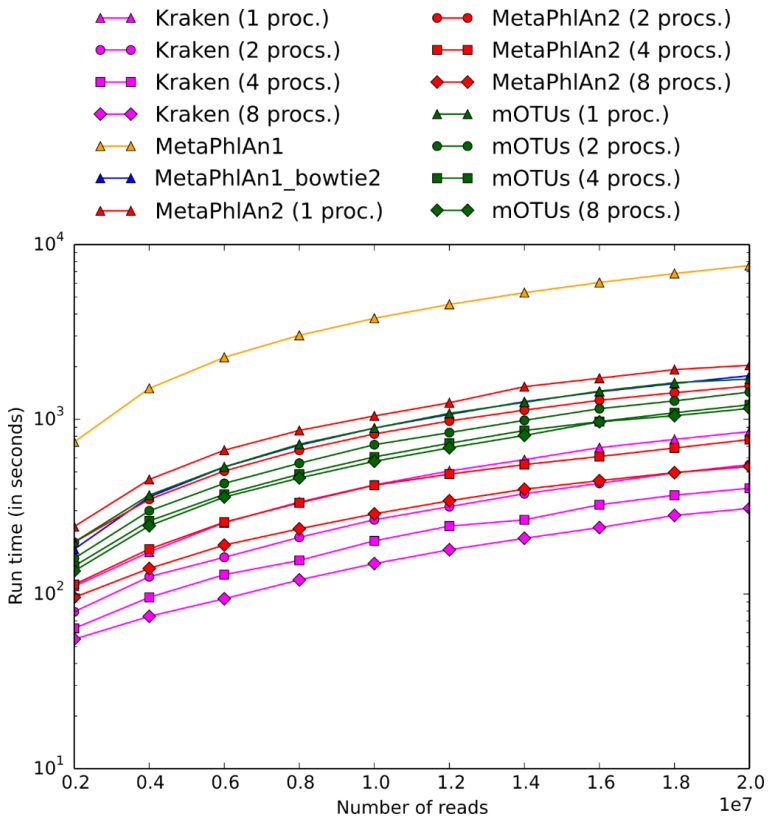

**Supplementary Fig. 6. Performance comparison of the four tested methods on log-normally distributed 10M-read datasets at the species level based on the Pearson correlation (corr). The performance of all methods are computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes)**



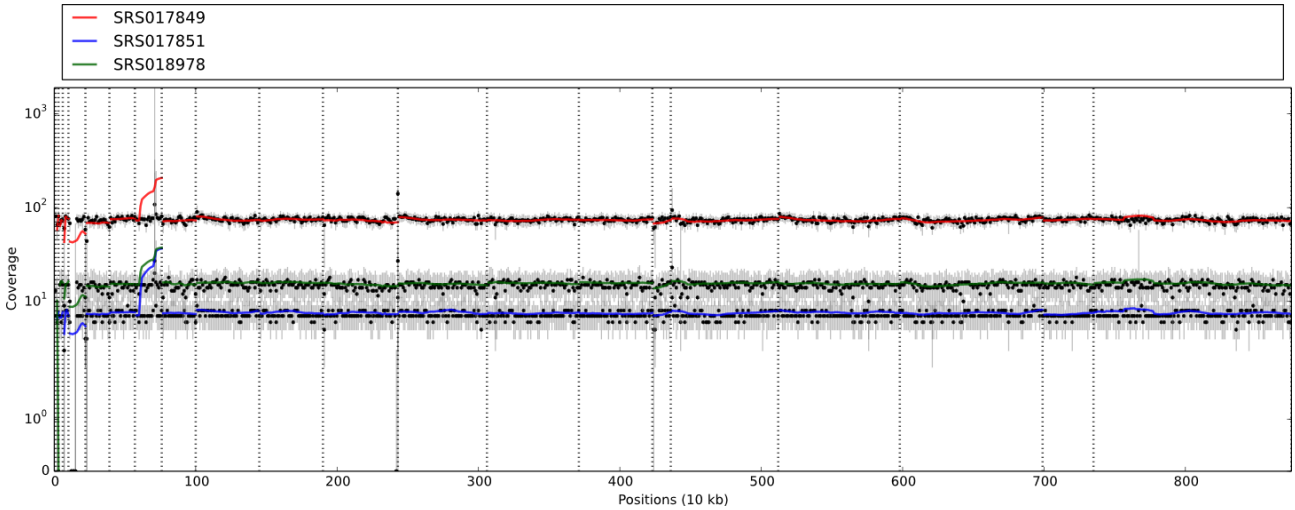

**Supplementary Fig. 7. Performance comparison of the four tested methods on the Mende *et al.*'s datasets**9 **at the genus level based on the ranked root mean squared error (r.m.s.e). The performance of all methods are computed on four kingdoms (Archaeal, Bacterial, Viruses and Eukaryotic microbes)**



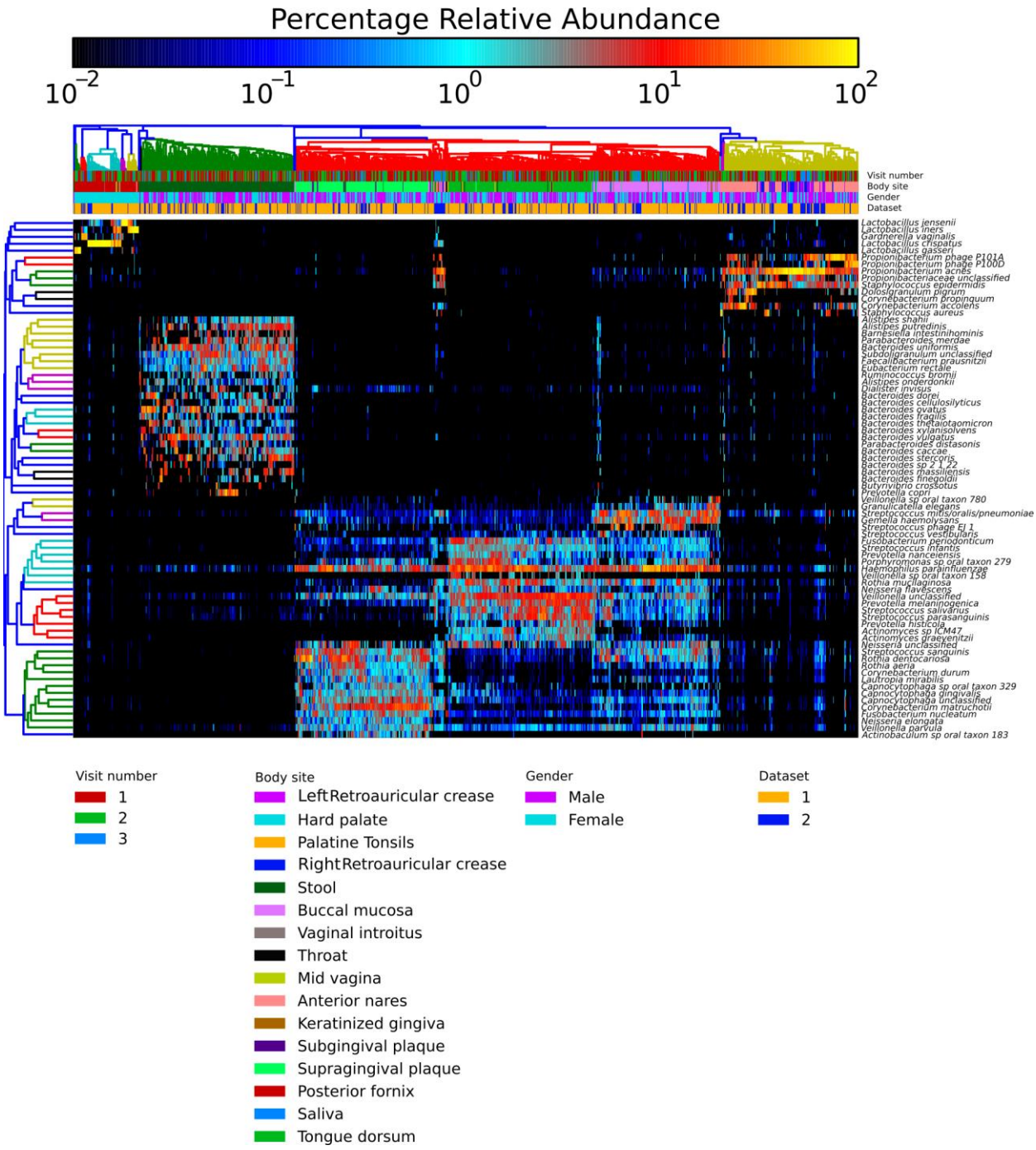

**Supplementary Fig. 8. Performance comparison of the four tested methods on the Mende *et al.*'s datasets**9 **at the species level based on the ranked root mean squared error (r.m.s.e). The performance of all methods are computed on four kingdoms (Archaea, Bacteria, Viruses and Eukaryotes)**



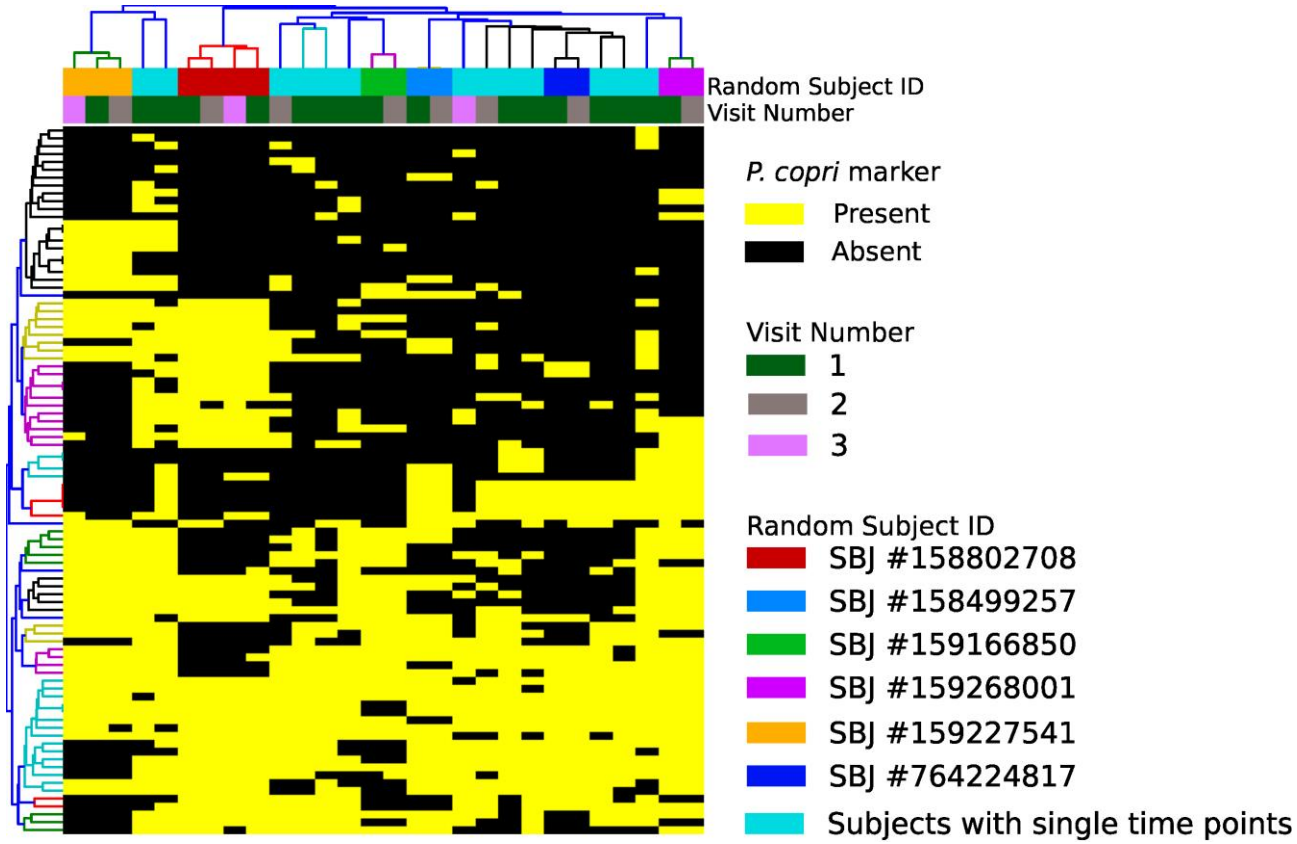

**Supplementary Fig. 9. Run-time comparison between the validated methods. The original implementation of MetaPhlAn1 was based on Blastn[16], but we evaluate here also its extension based on BowTie2. MetaPhlAn2, mOTUS, and Kraken are evaluated at increasing number of processors (from 1 to 8)**



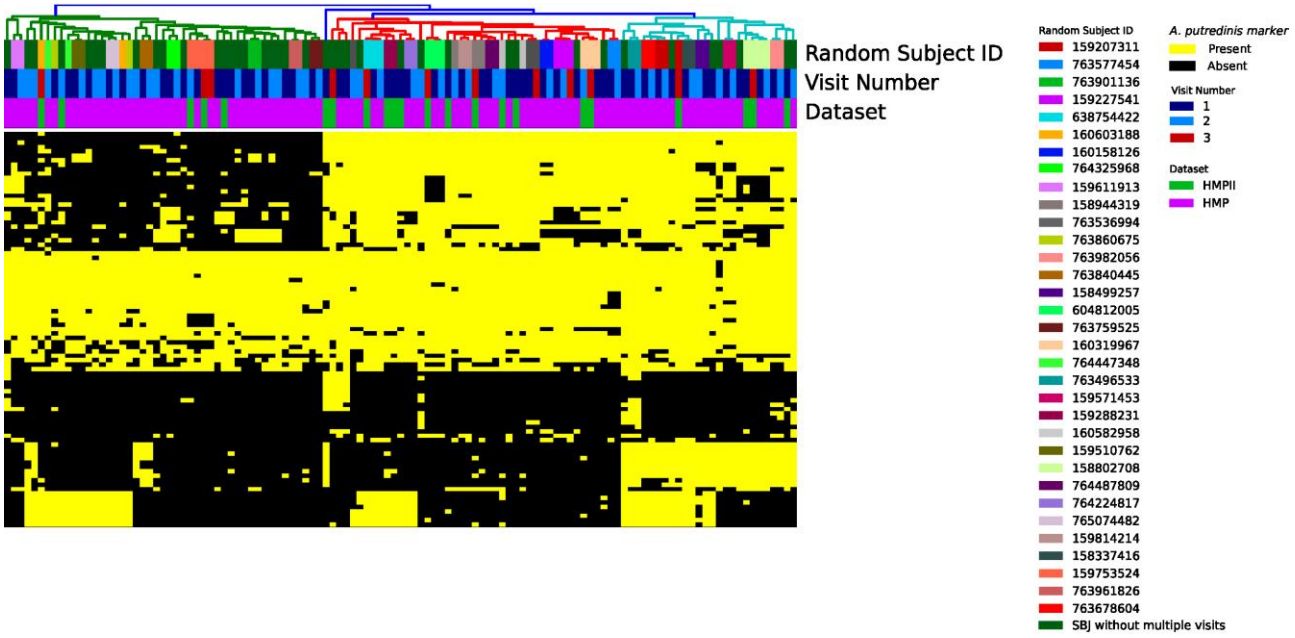

**Supplementary Fig. 10. Genome coverage plots of three HMP samples against the reference genome of *Malassezia globosa* (GCA_000181695) confirm the presence of this eukaryotic microbe on the human skin. Each point reports the average coverage on 10 kb windows, whereas the gray bars display the interquartile ranges**



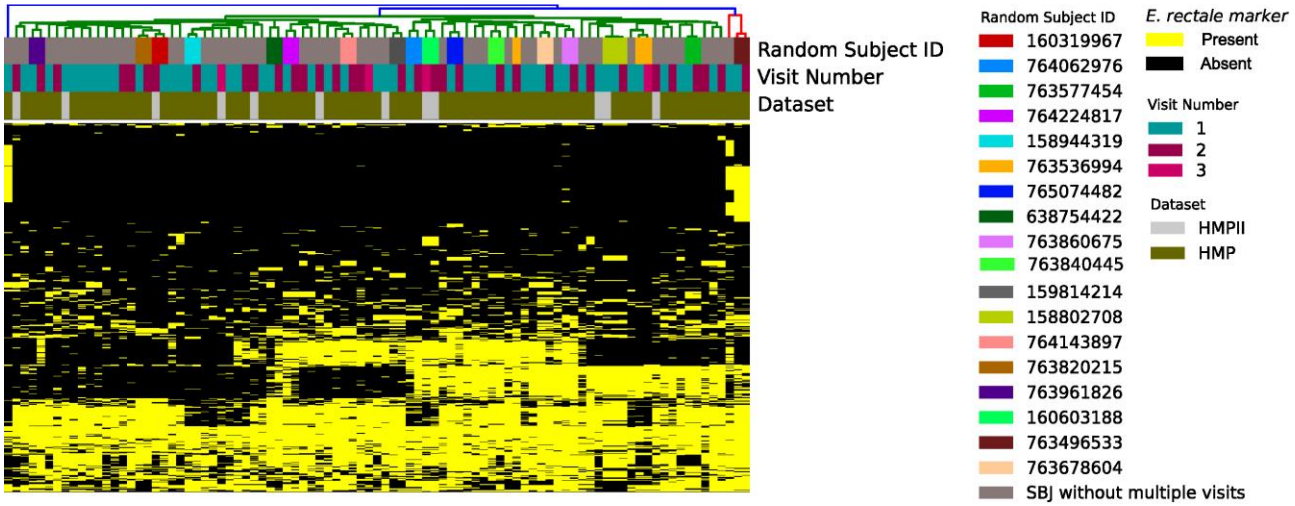

**Supplementary Fig. 11. MetaPhlAn2 profiling of HMP and HMPII samples. Only the 75 most abundant species (according to the 99$^{th}$ percentile ranking) are reported. Microbial species and samples are hierarchically clustered (average linkage) using correlation and Bray-Curtis distance (in root square abundance spaces) as similarity functions respectively**



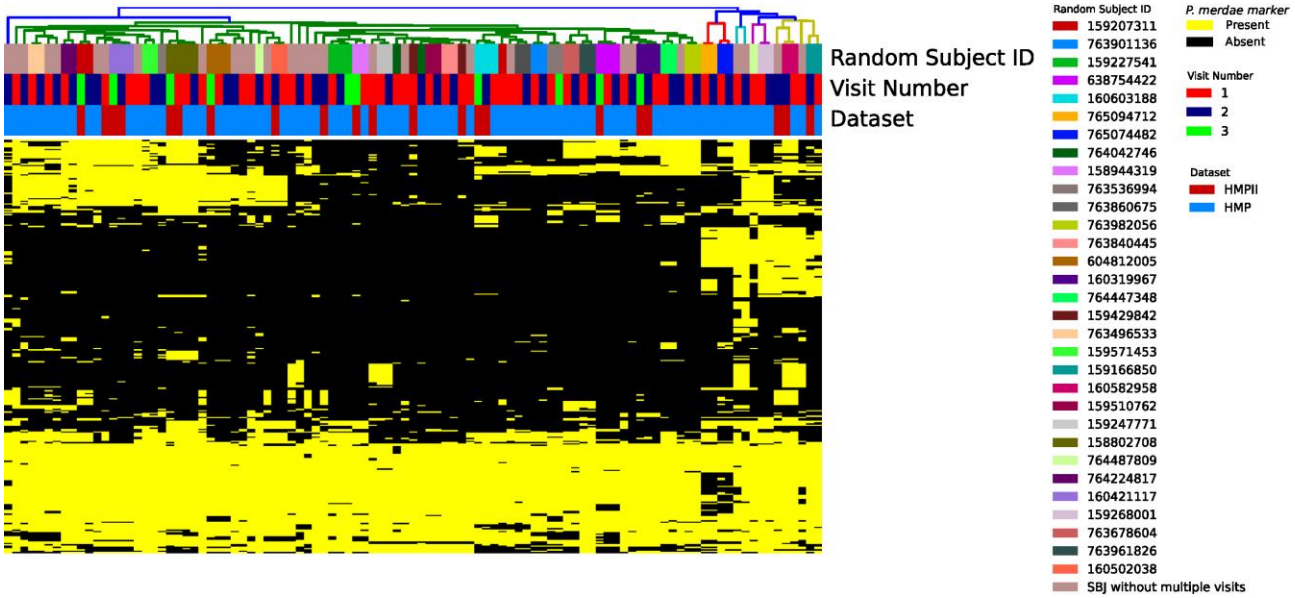

**Supplementary Fig. 12. Strain level fingerprinting of *Prevotella copri* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



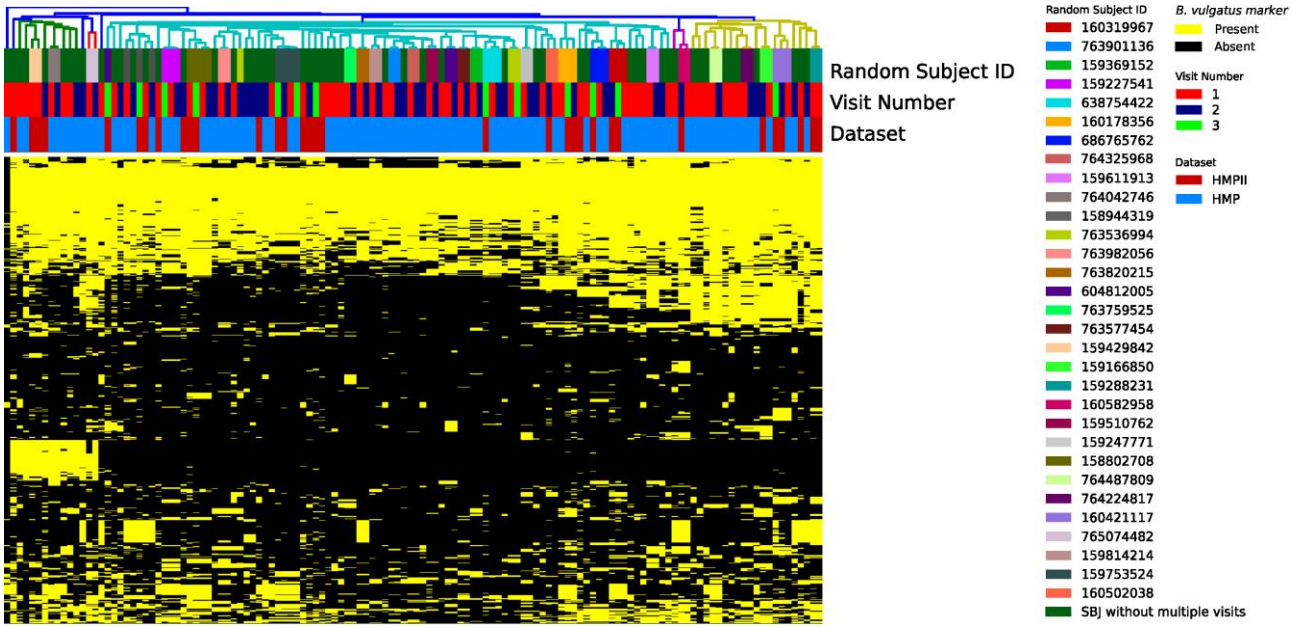

**Supplementary Fig. 13. Strain level fingerprinting of *Alistipes putredinis* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



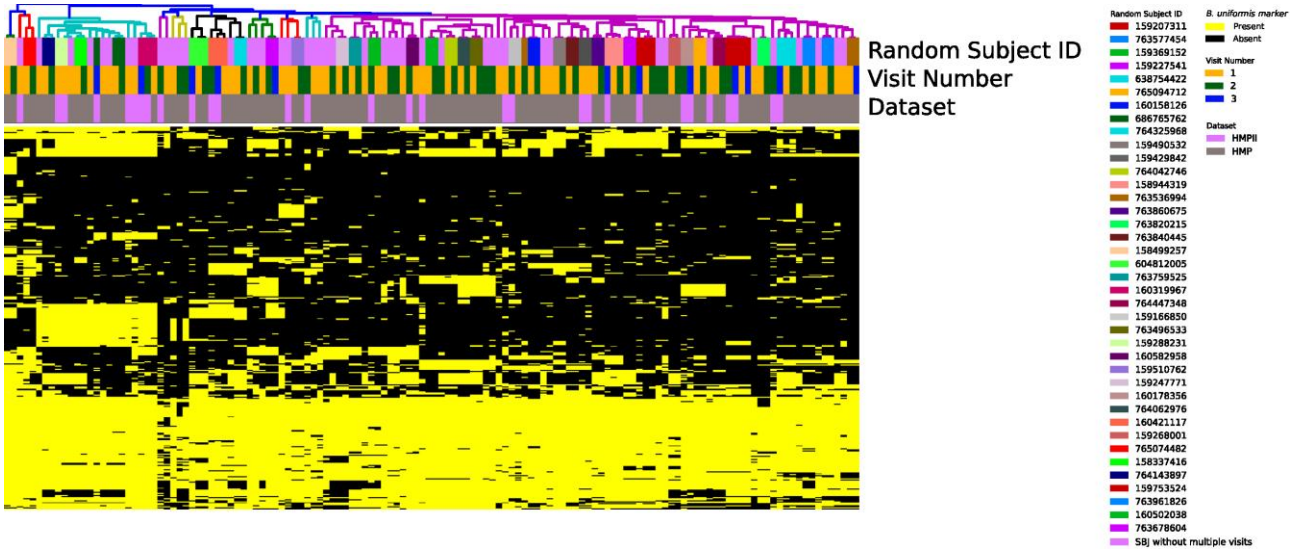

**Supplementary Fig. 14. Strain level fingerprinting of *Eubacterium rectale* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



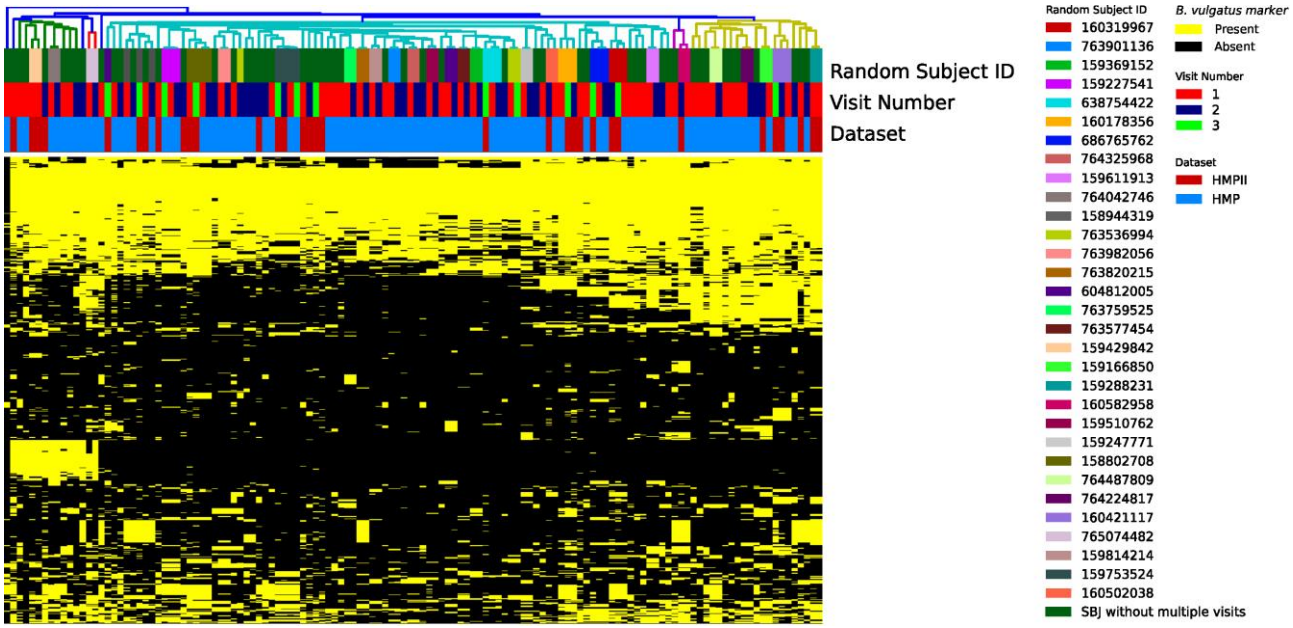

**Supplementary Fig. 15. Strain level fingerprinting of *Parabacteroides merdae* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



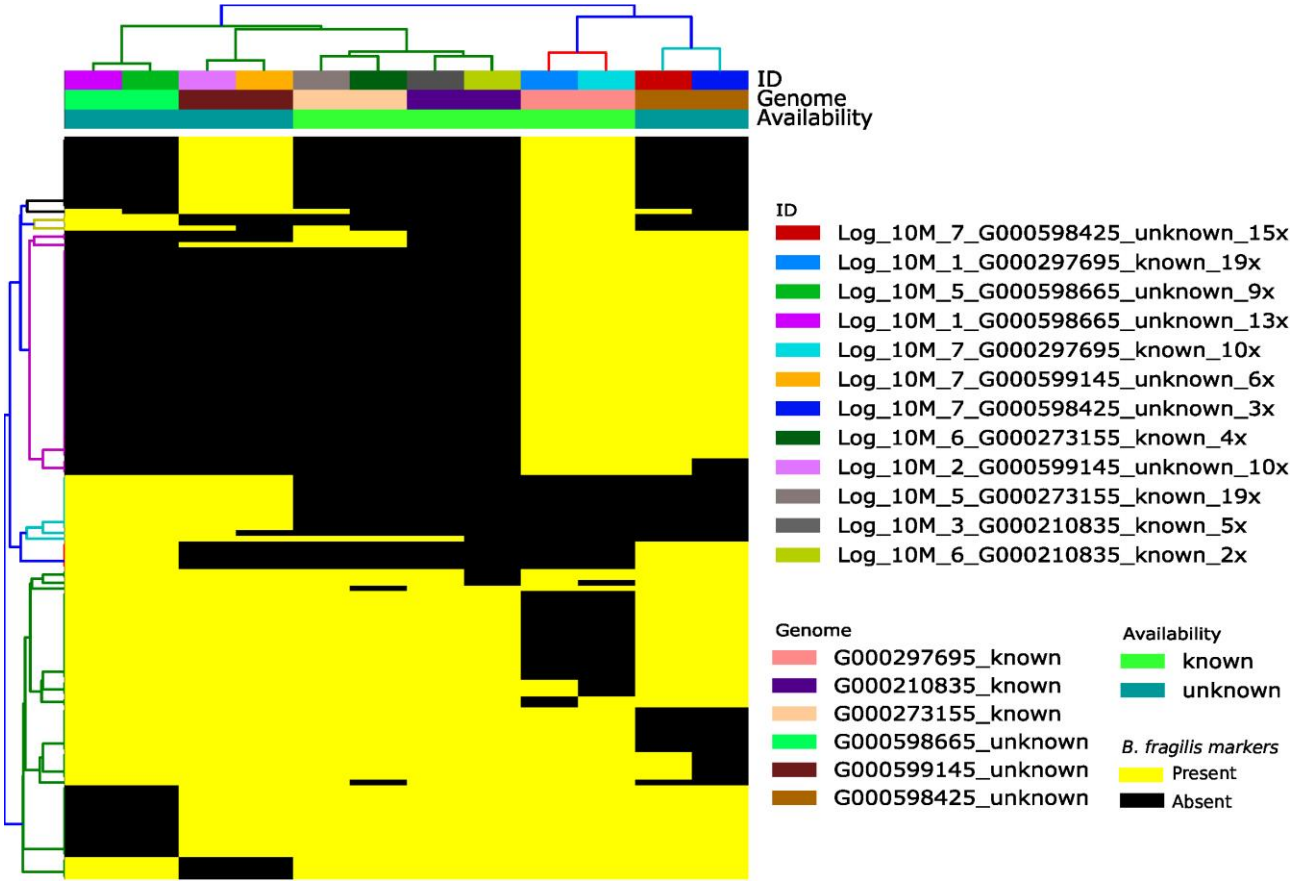

**Supplementary Fig. 16. Strain level fingerprinting of *Bacteroides ovatus* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



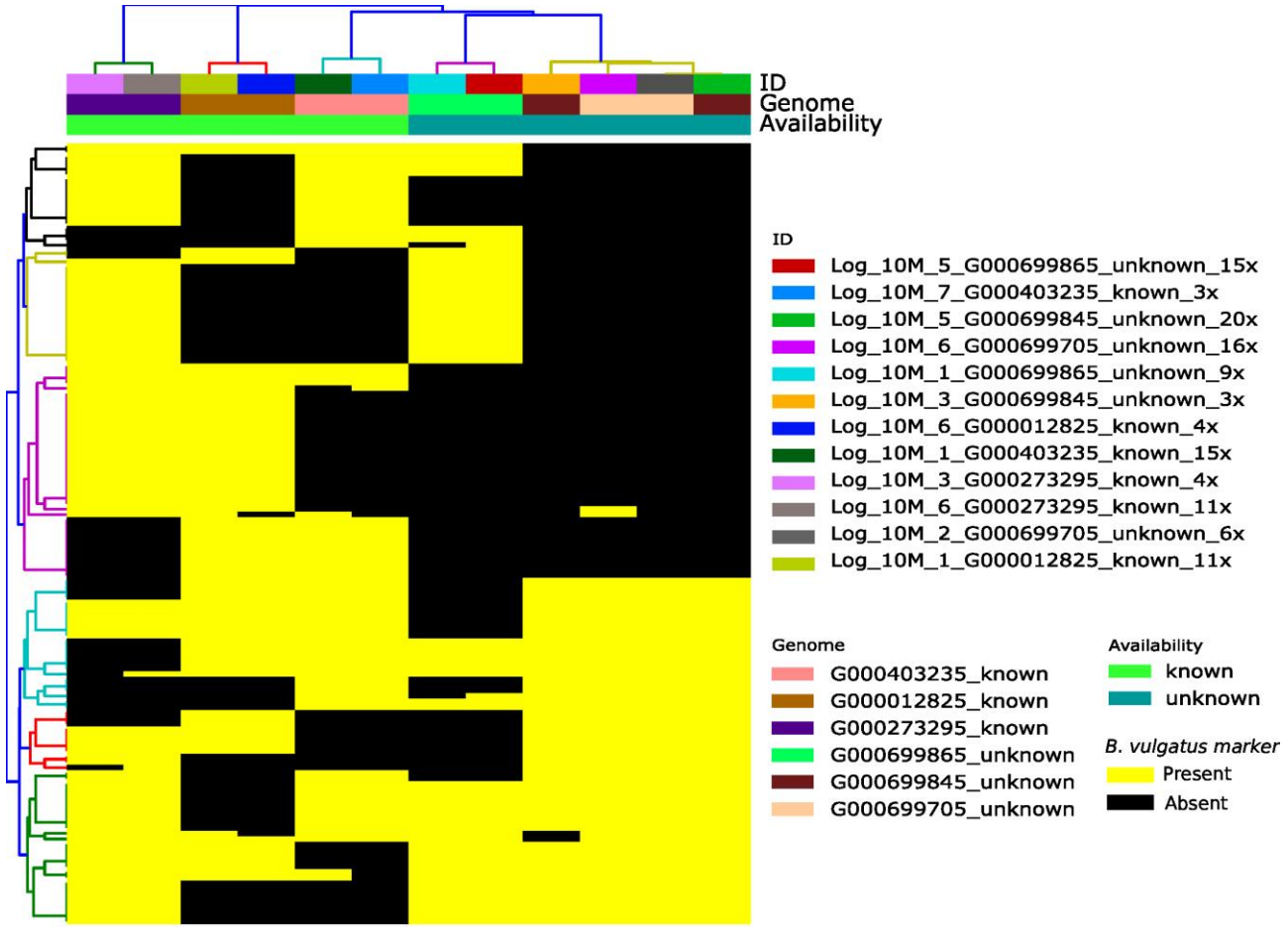

**Supplementary Fig. 17. Strain level fingerprinting of *Bacteroides uniformis* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



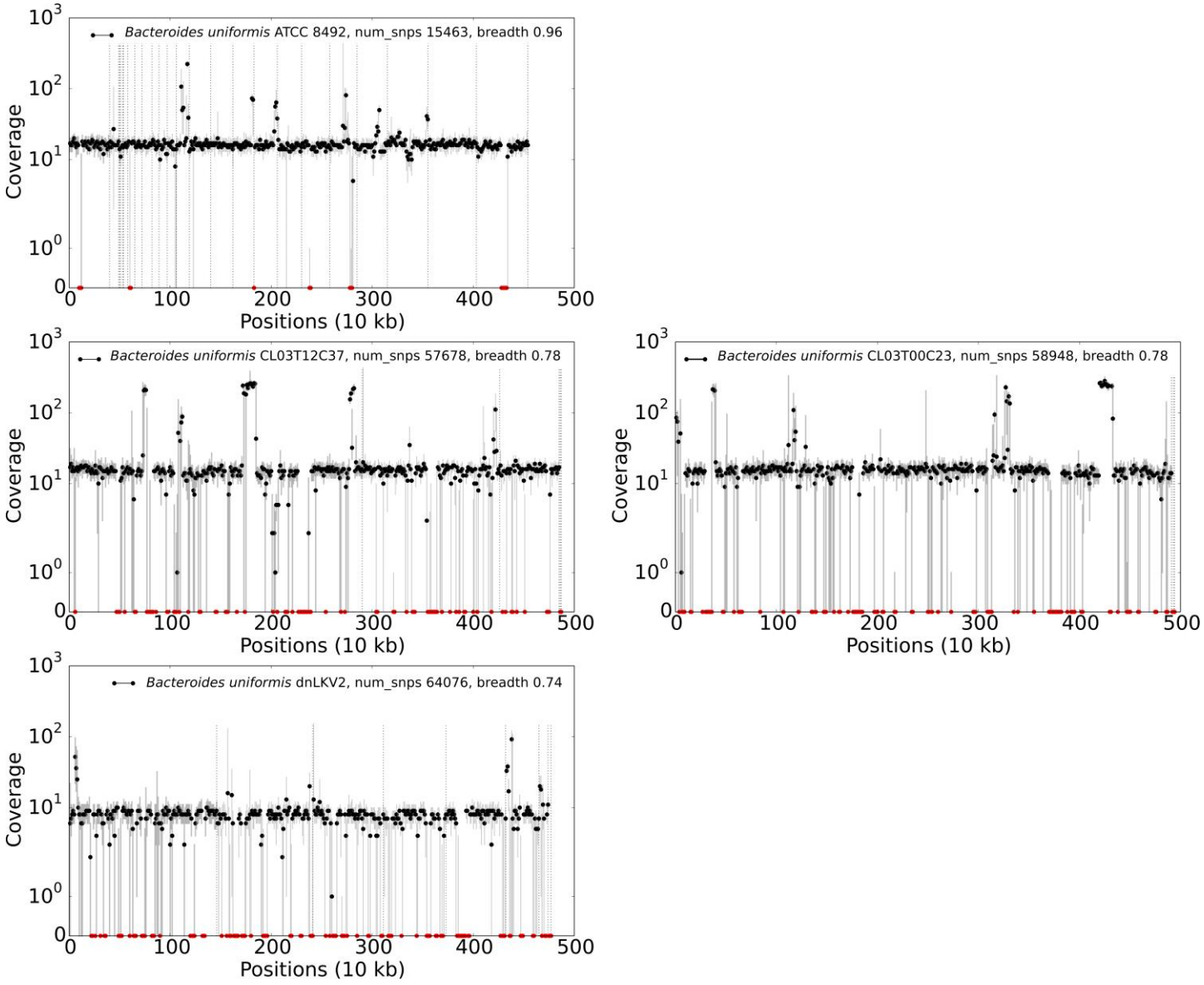

**Supplementary Fig. 18. Strain level fingerprinting of *Bacteroides vulgatus* in HMP/HMPII gut samples at multiple time points. The clustering step was performed based on Hamming distance**



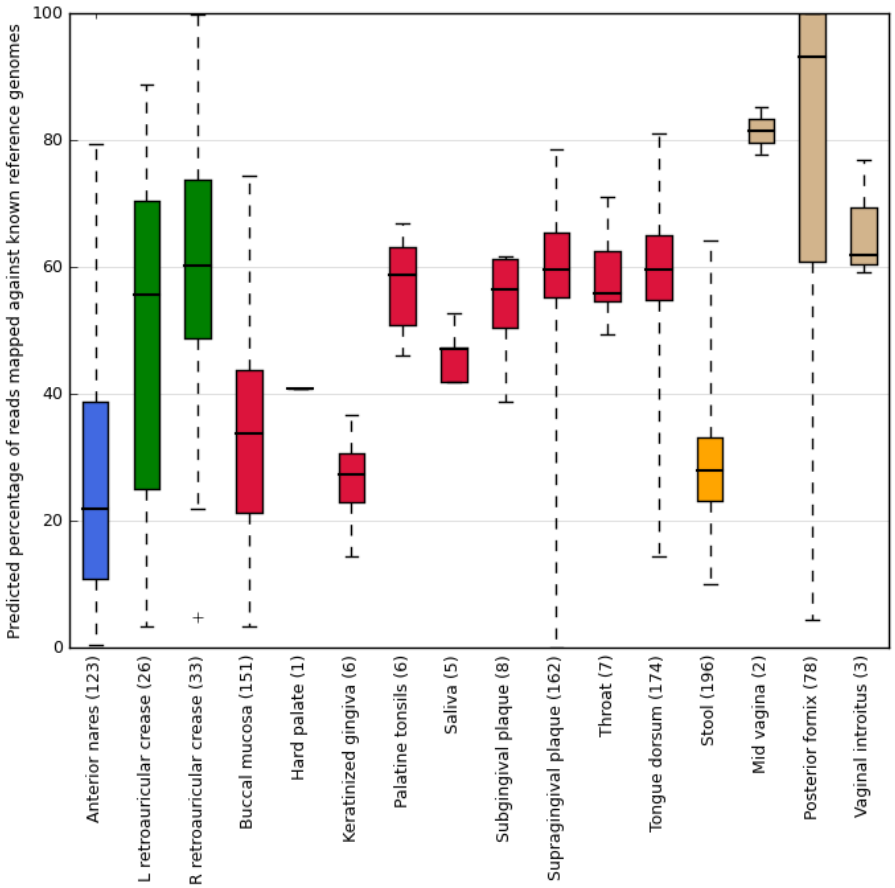

**Supplementary Fig. 19. Strain-level fingerprinting of *Bacteroides fragilis* in twelve synthetic samples generated from its six genomes sampled at different coverage (3 unknown and 3 known genomes) and merged with different synthetic metagenomes**




**Supplementary Fig. 20. Strain level fingerprinting of *Bacteroides vulgatus* in twelve synthetic samples generated from its six genomes sampled at different coverage (3 unknown and 3 known genomes) and merged with different synthetic metagenomes**




**Supplementary Fig. 21. An example of strain identification for *Bacteroides uniformis* strains. The top left panel shows the genome coverage of the strain *Bacteroides uniformis ATCC 8492* detected by MetaPhlAn2 while the remaining panels depict the coverages of the other sequenced *Bacteroides uniformis* strains. Windows of 10kb with zero coverage are highlighted in red**




**Supplementary Fig. 22. Predicted percentage of reads mapped against known reference genomes for the HMP/HMPII samples.**

**References**


1 Asnicar, F. *et al. PeerJ* **3**, e1029 (2015).
2 Ondov, B. *et al. BMC bioinformatics* **12** (2011).
3 Langmead, B. *et al. Nature methods* **9**, 357–359 (2012).
4 Segata, N. *et al*. *Nature methods* **9**, 811–814 (2012).
5 Segata, N. *et al*. *Nature communications* **4** (2013).
6 Huang, K. *et al*. *Nucleic acids research*, gkt1078 (2014).
7 Schloissing, S. *et al*. *Nature* **493**, 45–50 (2012).
8 Ren, B. *et al*. SynMetaP: a tool for simulating shotgun metagenomic sequencing data (https://bitbucket.org/Boyur/synmetap) (2014).
9 Mende, D.R. *et al*. *PloS one* **7**, e31386 (2012).
10 Sunagawa, S. *et al*. *Nature methods* **10**, 1196–1199 (2013).
11 Wood, D. *et al*. *Genome biology* **15** (2014).
12 Huson, D. H. *et al*. *Genome research* **21**, 1552–1560 (2011).
13 The Human Microbiome Project Consortium. *Nature* **486**, 215–221 (2012).
14 Aronesty, E. *Open bioinformatics journal* **7**, 1–8 (2013).
15 The Human Microbiome Project Consortium. *Nature* **486**, 207–214 (2012).
16 Altschuol, S.F. *et al*. *Journal of molecular biology* **215**, 403–410 (1990).