nature portfolio

https://doi.org/10.1038/s41596-020-00480-3

Supplementary information

Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit

Supplementary information

Tutorial: Assessing metagenomics software with the CAMI benchmarking toolkit

---

Supplementary Table 1: Digital Object Identifiers (DOIs) for assembly results of the first 10 of 64 short read samples of the CAMI II mouse gut data set and associated links. MEGAHIT was run with default (df), meta-large (ml), and meta-sensitive (ms) settings.

Assembler	Results DOI	Links for commands used and download
MEGAHIT 1.0.3 df	10.5281/zenodo.3663885	https://doi.org/10.5281/zenodo.3663885 https://zenodo.org/record/3663885/files/megahit103-Sample0-9-df-final.contigs.fa.gz?download=1
MEGAHIT 1.1.3 df	10.5281/zenodo.3663885	https://doi.org/10.5281/zenodo.3663885 https://zenodo.org/record/3663885/files/megahit113-Sample0-9-df-final.contigs.fa.gz?download=1
MEGAHIT 1.1.3 ml	10.5281/zenodo.3663885	https://doi.org/10.5281/zenodo.3663885 https://zenodo.org/record/3663885/files/megahit113-Sample0-9-ml-final.contigs.fa.gz?download=1
MEGAHIT 1.1.3 ms	10.5281/zenodo.3663885	https://doi.org/10.5281/zenodo.3663885 https://zenodo.org/record/3663885/files/megahit113-Sample0-9-ms-final.contigs.fa.gz?download=1
MEGAHIT 1.2.9 df	10.5281/zenodo.3663885	https://doi.org/10.5281/zenodo.3663885 https://zenodo.org/record/3663885/files/megahit129-Sample0-9-df-final.contigs.fa.gz?download=1
metaSPAdes 3.13.0	10.5281/zenodo.3664090	https://doi.org/10.5281/zenodo.3664090 https://zenodo.org/record/3664090/files/metaSPAdes3130-Sample0-9-contigs.fasta.gz?download=1

Supplementary Table 2: Elapsed (wall clock) time (h:mm) of assembly methods on the first 10 of 64 short read samples of the CAMI II mouse gut data set. The best result is shown in bold. The assemblers were run on a computer with several Intel Xeon Gold 6142 CPUs, virtualized to 58 logical cores, and 1.4 TB of main memory.

Assembler	default (df)	meta-sensitive (ms)	meta-large (ml)
MEGAHIT 1.0.3	7:14	–	–
MEGAHIT 1.1.3	6:24	14:00	11:56
MEGAHIT 1.2.9	**4:11**	–	–
metaSPAdes 3.13.0	41:06	–	–

Supplementary Table 3: Maximum resident set size (kbytes of main computer memory, not considering disk storage) of assembly methods on the first 10 of 64 short read samples of the CAMI II mouse gut data set. The best results are shown in bold.

Assembler	default (df)	meta-sensitive (ms)	meta-large (ml)
MEGAHIT 1.0.3	127,403,256	–	–
MEGAHIT 1.1.3	**42,278,824**	186,245,380	179,530,964
MEGAHIT 1.2.9	**42,919,140**	–	–
metaSPAdes 3.13.0	189,646,496	–	–

Supplementary Table 4: Digital Object Identifiers (DOIs) and associated links of genome binning results and ground truth of the gold standard cross-sample assembly of the CAMI II mouse gut data set. Also given is the average coverage of the underlying genomes.

Genome binner	Results DOI	Links for commands used and download
MaxBin 2.2.7	10.5281/zenodo.3629588	https://doi.org/10.5281/zenodo.3629588 https://zenodo.org/record/3629588/files/cami2_mouse_gut_maxbin2.2.7.binning?download=1
MetaBAT 2.12.1	10.5281/zenodo.3629590	https://doi.org/10.5281/zenodo.3629590 https://zenodo.org/record/3629590/files/cami2_mouse_gut_metabat2.12.1.binning?download=1
CONCOCT 1.0.0	10.5281/zenodo.3629592	https://doi.org/10.5281/zenodo.3629592 https://zenodo.org/record/3629592/files/cami2_mouse_gut_concoct1.0.0.binning?download=1
DAS Tool 1.1.2	10.5281/zenodo.3629594	https://doi.org/10.5281/zenodo.3629594 https://zenodo.org/record/3629594/files/cami2_mouse_gut_dastool1.1.2.binning?download=1
Binning ground truth	10.5281/zenodo.3632511	https://doi.org/10.5281/zenodo.3632511 https://zenodo.org/record/3632511/files/cami2_mouse_gut_gsa_pooled.binning?download=1
Average genome coverage	10.5281/zenodo.3667475	https://doi.org/10.5281/zenodo.3667475 https://zenodo.org/record/3667475/files/cami2_mouse_gut_average_genome_coverage.tsv?download=1

Supplementary Table 5: Elapsed (wall clock) time (h:mm) and maximum resident set size (kbytes of main computer memory, not considering disk storage) of genome binning methods on the cross-sample gold standard assembly of the CAMI II mouse gut data set. The best result is shown in bold. DAS Tool 1.1.2 (refinement only) is the time required to run only DAS Tool 1.1.2 using the output of MaxBin 2.2.7, MetaBAT 2.12.1, and CONCOCT 1.0.0. DAS Tool 1.1.2 (total) is the time required to run all these binners, including DAS Tool 1.1.2. The binners were run on a computer with an Intel Xeon E5-4650 v4 CPU (virtualized to 16 CPU cores, 1 thread per core) and 512 GB (536.870.912 kbytes) of main memory.

Genome binner	Time (hh:mm)	Memory (kbytes)
MaxBin 2.2.7	329:46	13,789,512
MetaBAT 2.12.1	**33:29**	**12,985,728**
CONCOCT 1.0.0	41:06	**12,985,728**
DAS Tool 1.1.2 (refinement only)	**00:37**	3,755,972
DAS Tool 1.1.2 (total)	404:58	13,789,512

Supplementary Table 6: CheckM and AMBER average purity assessment of genome binning results and ground truth of the gold standard cross-sample assembly of the CAMI II mouse gut data set. In parentheses is the average purity of the predicted bins with completeness > 70% and contamination < 10% according to AMBER's assessment. Also shown is the absolute difference between CheckM's and AMBER's assessments.

Average purity (completeness > 70% and contamination < 10%)

Genome binner	CheckM	AMBER	Difference (%)
Binning ground truth	0.984 (0.984)	1.000 (1.000)	1.564 (1.564)
MaxBin 2.2.7	0.939 (0.957)	0.774 (0.988)	21.346 (3.092)
MetaBAT 2.12.1	0.949 (0.961)	0.909 (0.994)	4.400 (3.324)
CONCOCT 1.0.0	0.659 (0.694)	0.594 (0.989)	10.909 (29.829)
DAS Tool 1.1.2	0.968 (0.988)	0.929 (0.989)	4.161 (0.122)

Supplementary Table 7: CheckM and AMBER average completeness assessment of genome binning results and ground truth of the gold standard cross-sample assembly of the CAMI II mouse gut data set. In parentheses is the average completeness of the predicted bins with completeness > 70% and contamination < 10% according to AMBER's assessment. Also shown is the absolute difference between CheckM's and AMBER's assessments.

Average completeness (completeness > 70% and contamination < 10%)

Genome binner	CheckM	AMBER	Difference (%)
Binning ground truth	0.927 (0.927)	1.000 (1.000)	7.270 (7.270)
MaxBin 2.2.7	0.692 (0.785)	0.641 (0.954)	7.949 (17.674)
MetaBAT 2.12.1	0.698 (0.788)	0.683 (0.939)	2.208 (16.105)
CONCOCT 1.0.0	0.868 (0.874)	0.848 (0.964)	2.291 (9.333)
DAS Tool 1.1.2	0.910 (0.967)	0.877 (0.949)	3.770 (1.964)

Supplementary Note: Bin quality metrics for CheckM

The purity for CheckM was calculated as the number of marker genes inferred for the bin lineage divided by the number of markers identified in the bin. An example report from CheckM's output file `bin_stats_ext.tsv` showing the assessments for three bins predicted with CONCOCT (bin IDs 111, 105, and 1), with the respective purity computation shown in bold, is as follows:

```
111    {'marker lineage': 'root', '# genomes': 5656, '# markers': 56, '# marker sets':
24, '0': 56, '1': 0, '2': 0, '3': 0, '4': 0, '5+': 0, 'Completeness': 0.0,
'Contamination': 0.0, ...
Purity = 56 / (56 + 0 + 0 * 2 + 0 * 3 + 0 * 4 + 0 * 5) = 1.0

105    {'marker lineage': 'root', '# genomes': 5656, '# markers': 56, '# marker sets':
24, '0': 0, '1': 0, '2': 7, '3': 46, '4': 3, '5+': 0, 'Completeness': 100.0,
'Contamination': 194.69696969696972, ...
Purity = 56 / (0 + 0 + 7 * 2 + 46 * 3 + 3 * 4 + 0 + 0 * 5) = 0.341

1      {'marker lineage': 'root', '# genomes': 5656, '# markers': 56, '# marker sets':
24, '0': 0, '1': 0, '2': 1, '3': 52, '4': 3, '5+': 0, 'Completeness': 100.0,
'Contamination': 206.25, ...
Purity = 56 / (0 + 0 + 1 * 2 + 52 * 3 + 3 * 4 + 0 + 0 * 5) = 0.329
```

As CheckM reports metrics per bin, not per genome, we calculated the average purity and completeness for both CheckM and AMBER as a simple average of these metrics over all bins. For AMBER, the completeness for a bin was then determined as the fraction of bp from the gold standard genome most abundant in a bin.

Supplementary Table 8: Digital Object Identifiers (DOIs) and associated links of taxonomic binning results and binning ground truth of the gold standard cross-sample assembly of the CAMI II mouse gut data set.

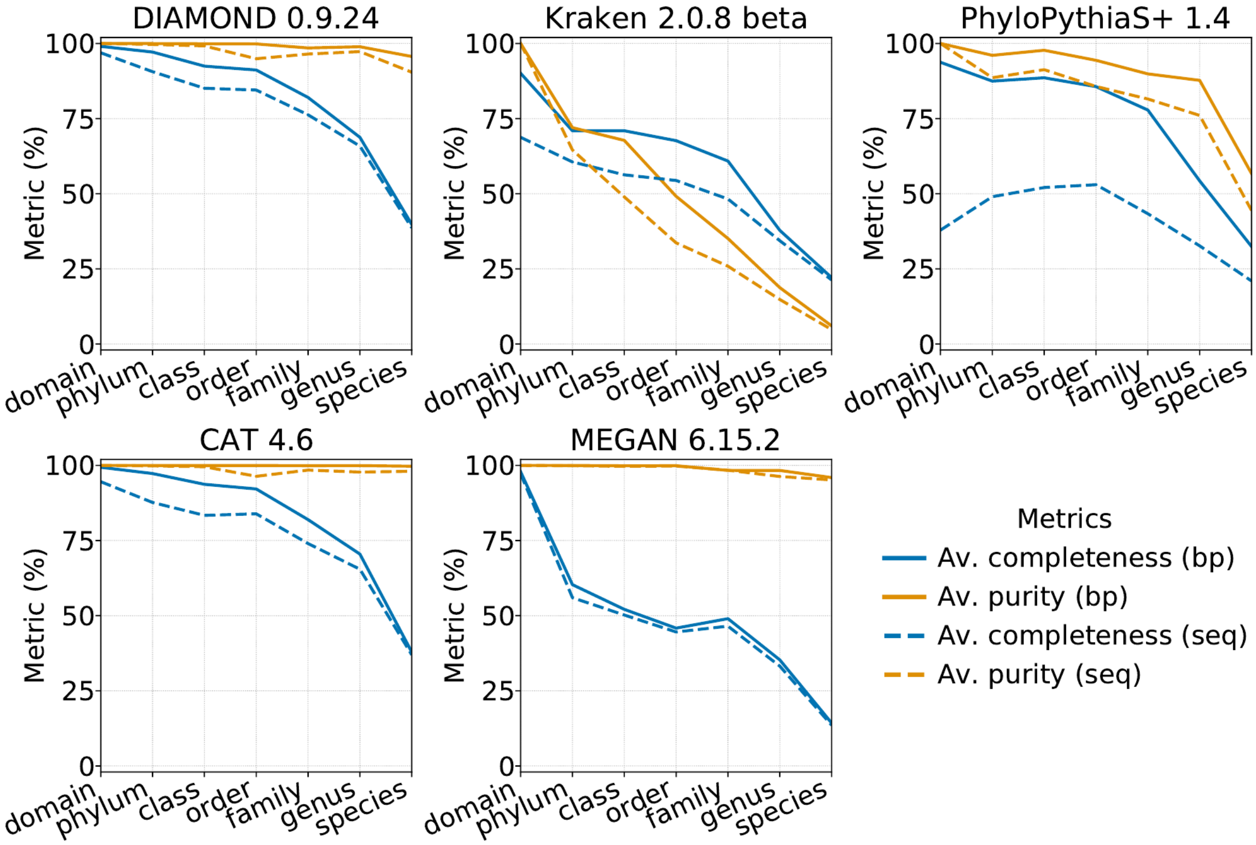
Taxonomic binner	Results DOI	Links for commands used and download
DIAMOND 0.9.24	10.5281/zenodo.3629598	http://doi.org/10.5281/zenodo.3629598 https://zenodo.org/record/3629598/files/cami2_mouse_gut_diamond0.9.24.binning?download=1
Kraken 2.0.8 beta	10.5281/zenodo.3629600	http://doi.org/10.5281/zenodo.3629600 https://zenodo.org/record/3629600/files/cami2_mouse_gut_kraken2.0.8beta.binning?download=1
PhyloPythiaS+ 1.4	10.5281/zenodo.3629602	http://doi.org/10.5281/zenodo.3629602 https://zenodo.org/record/3629602/files/cami2_mouse_gut_ppsp1.4.binning?download=1
CAT 4.6	10.5281/zenodo.3629604	http://doi.org/10.5281/zenodo.3629604 https://zenodo.org/record/3629604/files/cami2_mouse_gut_cat4.6.binning?download=1
MEGAN 6.15.2	10.5281/zenodo.3629606	http://doi.org/10.5281/zenodo.3629606 https://zenodo.org/record/3629606/files/cami2_mouse_gut_megan6.15.2.binning?download=1
Binning ground truth	10.5281/zenodo.3632511	http://doi.org/10.5281/zenodo.3632511 https://zenodo.org/record/3632511/files/cami2_mouse_gut_gsa_pooled.binning?download=1

Supplementary Table 9: Elapsed (wall clock) time (h:mm) and maximum resident set size (kbytes of main computer memory, not considering disk storage) of taxonomic binning methods on the cross-sample gold standard assembly of the CAMI II mouse gut data set. The best result is shown in bold. The time for MEGAN 6.15.2 is the sum of the time to run DIAMOND 0.9.24 and MEGAN's tool daa2rma, which uses DIAMOND's output. The binners were run on a computer with an Intel Xeon E5-4650 v4 CPU (virtualized to 16 CPU cores, 1 thread per core) and 512 GB (536.870.912 kbytes) of main memory.

Taxonomic binner	Time (hh:mm)	Memory (kbytes)
DIAMOND 0.9.24	21:58	43,350,528
Kraken 2.0.8 beta	**0:22**	39,439,795
PhyloPythiaS+ 1.4	206:38	285,949,912
CAT 4.6	49:17	**19,039,232**
MEGAN 6.15.2	23:14	97,196,924

Supplementary Table 10: Number of true and false positive (TP, FP) bins per taxonomic rank predicted from the CAMI II mouse gut data set, average bin size, and average precision. For the assessments (see main text), the smallest bins per method and rank were filtered out (overall 1% of the binned data in bp). The number of TP and FP bins removed and their average size are shown in the respective columns. TP bins are those labeled with a taxon that is present in the ground truth and whose precision is > 0, i.e. at least one sequence is correctly assigned to the bin. FP bins are those with zero precision.

Taxonomic binner	Taxonomic rank	TP bins	FP bins	TP bins removed	FP bins removed	Average bin size in Mb (average precision %)	Average bin size in Mb (average precision %) of removed bins
Gold standard	superkingdom	1	0	-	-	2,711.00 (100.0)	- -
	phylum	8	0	-	-	338.87 (100.0)	- -
	class	18	0	-	-	150.61 (100.0)	- -
	order	26	0	-	-	104.27 (100.0)	- -
	family	50	0	-	-	53.58 (100.0)	- -
	genus	157	0	-	-	15.62 (100.0)	- -
	species	549	0	-	-	4.94 (100.0)	- -
DIAMOND 0.9.24	superkingdom	1	2	0	2	896.27 (33.3)	1.91 (0.0)
	phylum	8	20	3	20	93.73 (28.6)	0.36 (13.0)
	class	18	26	6	26	57.45 (40.9)	0.61 (18.7)
	order	26	59	7	59	29.49 (30.5)	0.33 (10.6)
	family	50	93	7	93	14.18 (34.5)	0.19 (7.0)
	genus	157	186	31	186	5.08 (44.9)	0.08 (13.5)
	species	523	693	88	687	0.79 (41.0)	0.01 (9.8)
Kraken 2.0.8 beta	superkingdom	1	3	0	3	614.75 (25.0)	5.61 (0.0)
	phylum	8	33	3	32	59.27 (13.7)	0.34 (3.7)
	class	18	59	5	56	31.31 (17.0)	0.33 (3.7)
	order	24	150	2	138	13.83 (10.2)	0.17 (0.7)
	family	43	343	1	294	6.08 (8.4)	0.08 (0.1)
	genus	67	1160	0	945	1.75 (4.3)	0.02 (0.0)
	species	129	3895	0	2611	0.57 (2.1)	0.01 (0.0)
PhyloPythiaS+ 1.4	superkingdom	1	1	0	1	1,270.71 (50.0)	0.22 (0.0)
	phylum	8	1	4	1	272.40 (84.5)	4.49 (75.3)
	class	18	1	6	1	127.41 (91.6)	2.91 (80.9)
	order	26	4	7	4	80.00 (82.4)	1.92 (61.7)
	family	49	7	10	6	40.44 (81.6)	1.30 (60.9)
	genus	130	23	24	18	12.22 (77.6)	0.44 (50.9)
	species	341	125	72	54	4.32 (52.2)	0.16 (39.9)
CAT 4.6	superkingdom	1	3	0	3	673.67 (25.0)	0.22 (0.0)
	phylum	8	12	3	12	130.57 (40.0)	0.35 (20.0)
	class	18	14	6	14	79.77 (56.2)	0.84 (30.0)
	order	26	37	8	37	40.23 (41.3)	0.49 (17.8)
	family	50	58	10	58	18.94 (46.2)	0.29 (14.6)
	genus	156	111	41	111	6.62 (58.1)	0.11 (26.4)
	species	495	465	135	465	0.89 (50.3)	0.01 (20.6)
MEGAN 6.15.2	superkingdom	1	2	0	2	885.89 (33.3)	1.98 (0.0)
	phylum	5	0	1	0	523.05 (100.0)	21.81 (99.8)
	class	10	0	1	0	251.24 (99.9)	9.46 (100.0)
	order	13	0	3	0	189.28 (99.9)	4.84 (100.0)
	family	28	2	4	2	66.92 (92.0)	3.13 (66.7)
	genus	73	1	13	1	21.54 (97.3)	1.14 (92.7)
	species	139	2	20	2	4.04 (94.3)	0.26 (85.3)
Average	all ranks	82.8	180.7	14.9	169.8	235.80 (59.2)	1.90 (32.1)



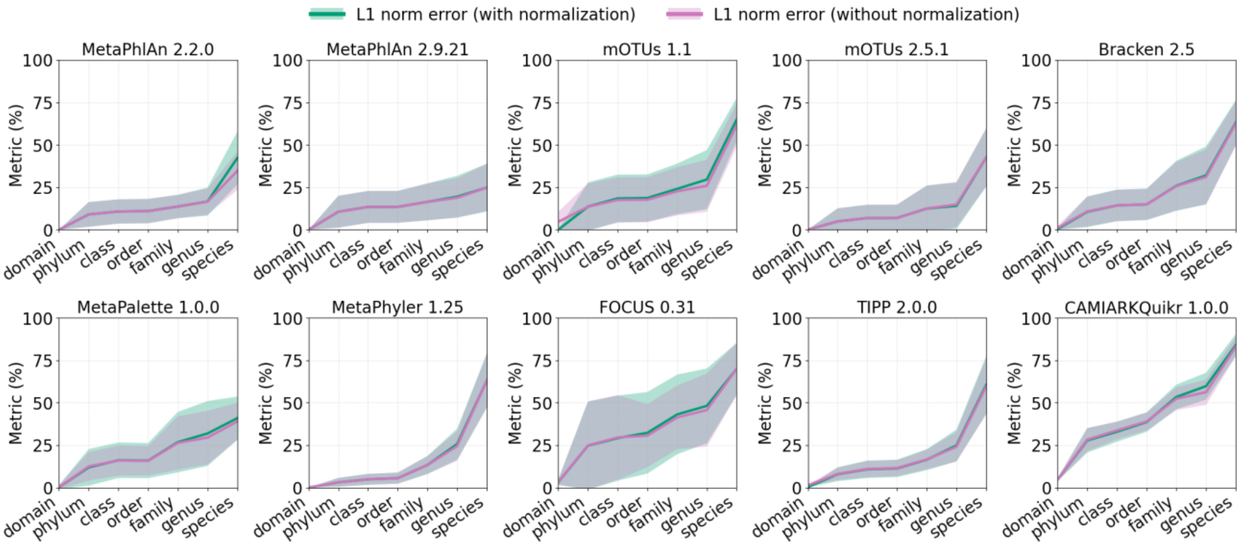
Supplementary Fig. 1: Average completeness and purity based on base pair (solid lines) and contig (dashed lines) counts assessed for taxonomic bins predicted from the CAMI II mouse gut data set. In the metrics based on bp counts, longer contigs have higher weight within a bin than shorter contigs, and better performance measured using these counts instead of contig counts indicates more accurate binning of longer contigs than shorter ones. Bp counts are used in the assessments in the main document (Fig. 4) and reproduced here as Average completeness (bp) and Average purity (bp). All bins contribute equally in the average computation. In this assessment, the smallest bins per method and rank were filtered out (overall 1% of the binned data in bp).

Supplementary Table 11: Digital Object Identifiers (DOIs) and associated links of taxonomic profiling results and profiling ground truth for the 64 short read samples of the CAMI II mouse gut data set.

Taxonomic profiler	Results DOI	Links for commands used and download
MetaPhlAn 2.9.21	10.5281/zenodo.3629610	https://doi.org/10.5281/zenodo.3629610 https://zenodo.org/record/3629610/files/cami2_mouse_gut_metaphlan2.9.21.profile?download=1
MetaPhlAn 2.2.0	10.5281/zenodo.3629612	https://doi.org/10.5281/zenodo.3629612 https://zenodo.org/record/3629612/files/cami2_mouse_gut_metaphlan2.2.0.profile?download=1
Bracken 2.5	10.5281/zenodo.3629614	https://doi.org/10.5281/zenodo.3629614 https://zenodo.org/record/3629614/files/cami2_mouse_gut_bracken2.5.profile?download=1
FOCUS 0.31	10.5281/zenodo.3629620	https://doi.org/10.5281/zenodo.3629620 https://zenodo.org/record/3629620/files/cami2_mouse_gut_focus0.31.profile?download=1
CAMIARKQuikr 1.0.0	10.5281/zenodo.3629622	https://doi.org/10.5281/zenodo.3629622 https://zenodo.org/record/3629622/files/cami2_mouse_gut_camiarkquikr1.0.0.profile?download=1
mOTUs 1.1	10.5281/zenodo.3629624	https://doi.org/10.5281/zenodo.3629624 https://zenodo.org/record/3629624/files/cami2_mouse_gut_motus1.1.profile?download=1
mOTUs 2.5.1	10.5281/zenodo.3629626	https://doi.org/10.5281/zenodo.3629626 https://zenodo.org/record/3629626/files/cami2_mouse_gut_motus2.5.1.profile?download=1
MetaPalette 1.0.0	10.5281/zenodo.3629628	https://doi.org/10.5281/zenodo.3629628 https://zenodo.org/record/3629628/files/cami2_mouse_gut_metapalette1.0.0.profile?download=1
TIPP 2.0.0	10.5281/zenodo.3629630	https://doi.org/10.5281/zenodo.3629630 https://zenodo.org/record/3629630/files/cami2_mouse_gut_tipp2.0.0.profile?download=1
MetaPhyler 1.25	10.5281/zenodo.3629632	https://doi.org/10.5281/zenodo.3629632 https://zenodo.org/record/3629632/files/cami2_mouse_gut_metaphyler1.25.profile?download=1
Profiling ground truth	10.5281/zenodo.3632528	https://doi.org/10.5281/zenodo.3632528 https://zenodo.org/record/3632528/files/cami2_mouse_gut_gs.profile?download=1

Supplementary Table 12: Elapsed (wall clock) time (h:mm) and maximum resident set size (kbytes of main computer memory, not considering disk storage) of taxonomic profiling methods on the 64 short read samples of the CAMI II mouse gut data set. The best results are shown in bold. Bracken requires to run Kraken, hence the times required to run Bracken and both tools are shown. The taxonomic profilers were run on a computer with an Intel Xeon E5-4650 v4 CPU (virtualized to 16 CPU cores, 1 thread per core) and 512 GB (536.870.912 kbytes) of main memory.

Taxonomic binner	**Time (hh:mm)**	**Memory (kbytes)**
MetaPhlAn 2.9.21	18:44	5,139,172
MetaPhlAn 2.2.0	12:30	1,741,304
Bracken 2.5 (only Bracken)	**0:01**	**24,472**
Bracken 2.5 (Kraken and Bracken)	**3:03**	39,439,796
FOCUS 0.31	13:27	5,236,199
CAMIARKQuikr 1.0.0	16:19	27,391,555
mOTUs 1.1	19:50	**1,251,296**
mOTUs 2.5.1	14:29	3,922,448
MetaPalette 1.0.0	76:49	27,297,132
TIPP 2.0.0	151:01	70,789,939
MetaPhyler 1.25	119:30	2,684,720



Supplementary Fig. 2: L1 norm error computed with and without prior normalization of the predicted organismal relative abundances per method and taxonomic rank (such that predicted taxon abundances sum to 100%, dismissing abundances of unassigned taxa at that rank) on the 64 short read samples of the CAMI II mouse gut data set. The shaded bands show the standard deviation of a metric.

Supplementary Table 13: Average sum of predicted taxon abundances per taxonomic profiler and rank for the 64 short read samples of the CAMI II mouse gut data set.

Taxonomic profiler	superkingdom	phylum	class	order	family	genus	species
MetaPhlAn 2.2.0	99.99	99.99	99.99	99.99	99.98	97.49	84.43
MetaPhlAn 2.9.21	100.00	99.99	99.99	99.99	99.92	95.84	100.00
mOTUs 1.1	90.66	90.66	90.66	90.66	90.63	86.39	90.66
mOTUs 2.5.1	100.00	99.99	99.99	99.99	99.99	99.99	99.99
Bracken 2.5	98.69	98.66	98.53	98.63	98.26	97.58	98.69
MetaPalette 1.0.0	99.99	92.12	92.11	92.12	91.25	86.10	90.98
MetaPhyler 1.25	100.00	100.00	99.85	99.96	98.67	97.84	99.99
FOCUS 0.31	99.95	99.40	98.13	94.99	92.60	94.40	99.99
TIPP 2.0.0	97.83	97.83	97.60	97.71	96.58	96.35	97.83
CAMIARKQuikr 1.0.0	99.99	95.01	90.71	92.18	90.69	91.87	97.71