Nature Methods

Taxonomic metagenome sequence assignment with structured output models

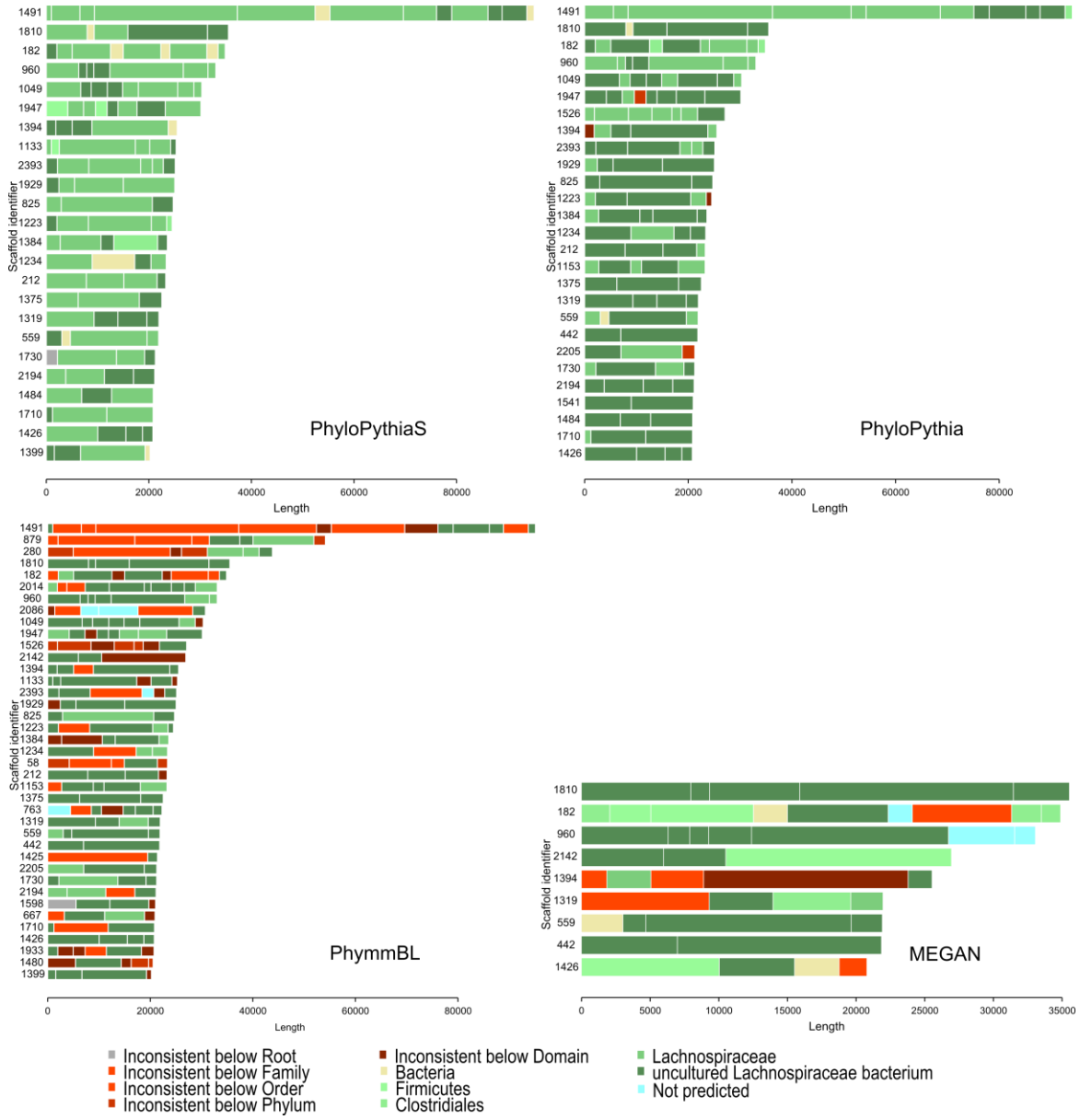
K R Patil, P Haider, P B Pope, P J Turnbaugh, M Morrison, T Scheffer & A C McHardy

Supplementary Figure 1	Evaluation on the simMC (simulated acid mine drainage) data set in four different settings.
Supplementary Figure 2	Scaffold-contig visualization of different binning methods for the WG-2 population in the Tammar wallaby metagenome sample.
Supplementary Figure 3	Evaluation of different binning methods on short fragments of varying lengths.
Supplementary Figure 4	Overlap between predictions of different methods on the TW sample for the three uncultured populations.
Supplementary Figure 5	Overlap between predictions of different methods on TW sample for dominant phyla.
Supplementary Table 1	Assignment accuracy of different binning methods on the simulated Acid Mine Drainage data set.
Supplementary Table 2	Performance of different binning methods for the abundant populations in the TW sample.
Supplementary Table 3	NUCmer analysis of the WG-1 assignments for TW sample.
Supplementary Table 4	Modeled clades for the TW sample.
Supplementary Table 5	Taxonomic assignments for abundant genera in the human gut metagenome samples.
Supplementary Table 6	Bin validation for the human gut metagenome samples using marker genes.
Supplementary Table 7	Validation for the human gut metagenome samples using CD-HIT (fraction matched).
Supplementary Table 8	Modeled clades for PhyloPythiaS for the human gut metagenome samples (TS28 and TS29).
Supplementary Table 9	Statistical comparison of the assignments of different methods on TW data set.
Supplementary Table 10	Number of contigs classified by different methods at different taxonomic ranks for the TW sample.
Supplementary Table 11	Effect of sample specific data on the assignment of the TW sample for PhyloPythiaS and PhymmBL.
Supplementary Table 12	Genomes used for simulated short fragment test data set.
Supplementary Table 13	Performance evaluation of the different binning methods on a simulated data set of short fragments of varying lengths.
Supplementary Table 14	Execution time comparison for different methods for characterization of the three real metagenome samples.
Supplementary Note	

Supplementary Figure 1: Evaluation on the simMC (simulated acid mine drainage) data set in four different settings.

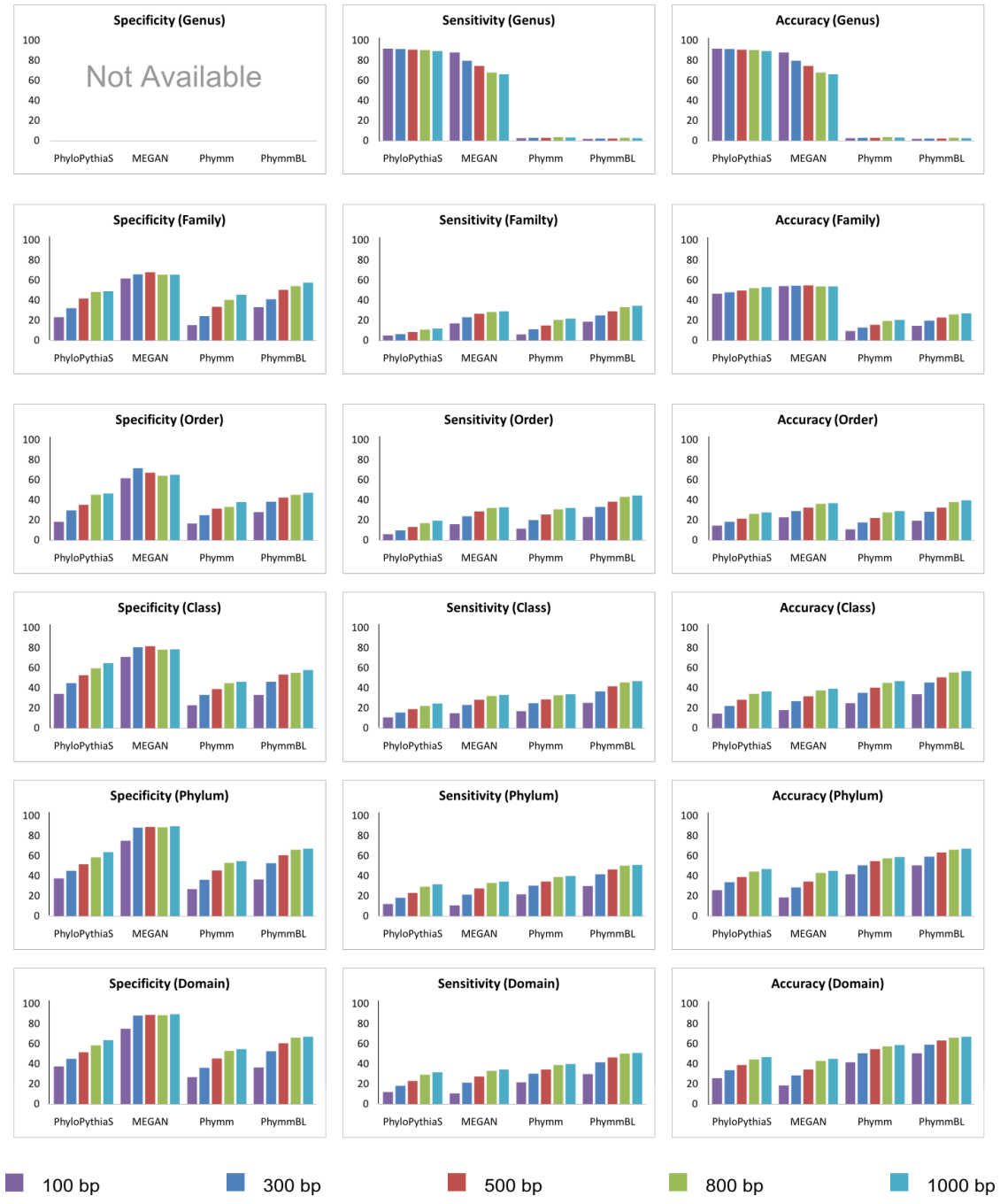
Known species: Genomes used in simulation were excluded from reference; *New genus:* Genomes used for simulation and genus level genomes for dominant populations were excluded from reference; *New order:* Genomes used for simulation and order level genomes for dominant populations were excluded from the reference and *New class:* Genomes used for simulation and class level genomes for dominant populations were excluded from reference.

Supplementary Figure 2: Scaffold-contig visualization of different binning methods for the WG-2 population in the Tammar wallaby metagenome sample.



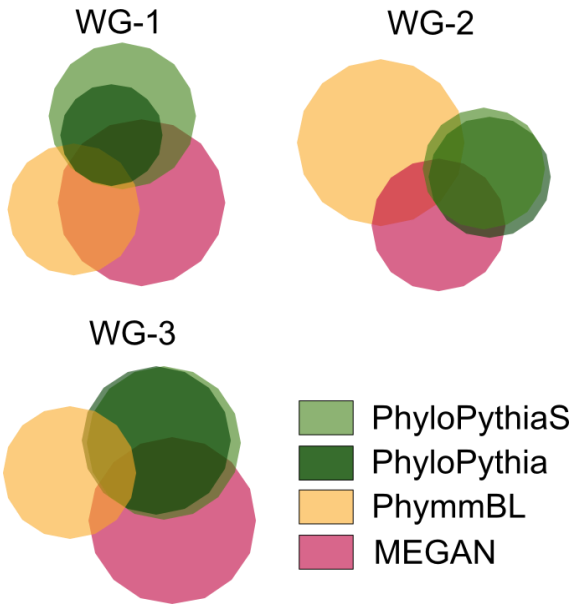
Every horizontal bar represents a scaffold and its constituent contigs. Every contig is color coded to represent its consistency with respect to the scaffold assignment. Only scaffolds >=20 kb in length are shown for clarity.

Supplementary Figure 3: Evaluation of different binning methods on short fragments of varying lengths.



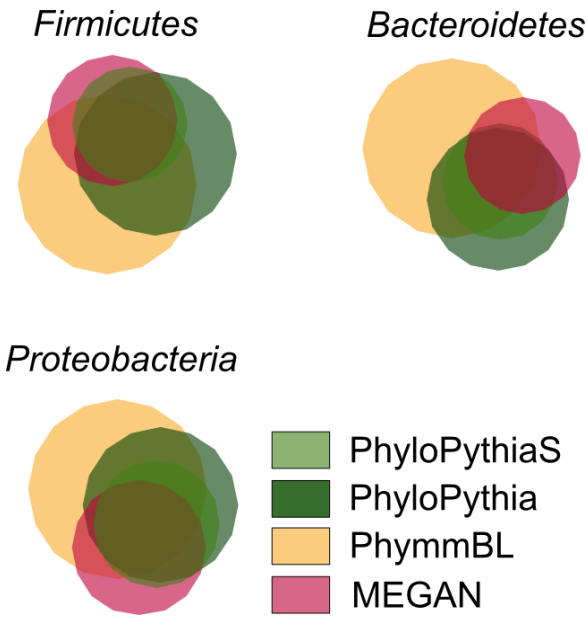
Short fragment data sets of varying fragment lengths were created using 100 whole genome assemblies (Supplementary Table 3) while complete genomes were used as reference/training sequences. Performance at genus rank quantifies over-binning.

Supplementary Figure 4: Overlap between predictions of different methods on the TW sample for the three uncultured populations.



The overlaps are represented as area proportional Euler diagrams[1]. Only exact predictions were taken into account for each population. The areas correspond to the predictions of the methods on the union of contigs predicted as a particular clade by at least one method. As can be seen, PhyloPythiaS and PhyloPythia have large overlaps for all populations.

Supplementary Figure 5: Overlap between predictions of different methods on TW sample for dominant phyla.



The overlaps are represented as area proportional Euler diagrams. The areas correspond to the predictions of the methods on the union of contigs predicted as a particular clade by at least one method. All the predictions were mapped to its corresponding phyla. As can be seen, PhyloPythiaS, PhyloPythia and MEGAN have large overlaps for all three phyla.

Supplementary Table 1: Assignment accuracy of different binning methods on the simulated Acid Mine Drainage data set. Experiments were performed using four different reference sets; A. Known species: Excluding the genomes used for simulation, B. New genus: 100 kb of the dominant strains available, which can be identified based on marker-gene analyses of the sample fragments or by fosmid sequencing, but no genome sequences of the same genera; C. New order: 100 kb of the dominant strains available but no genomes of the same order and D. New class: 100 kb of the dominant strains available but no genomes of the same classes.

A. Known species

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	97.37	53.36	44.59
	MEGAN	79.5	65.22	81.21
	Phymm	65.551	44.097	52.812
	PhymmBL	79.5	65.22	81.21
Family	PhyloPythiaS	99.62	65.48	65.91
	MEGAN	92.57	67	84.52
	Phymm	96.717	61.234	65.307
	PhymmBL	97.55	77.52	88.83
Order	PhyloPythiaS	99.56	69.91	76.39
	MEGAN	84.16	68.17	87.11
	Phymm	78.091	66.93	78.746
	PhymmBL	92.22	80.82	94.33
Class	PhyloPythiaS	98.8	75.66	82.48
	MEGAN	99.99	92.29	89.16
	Phymm	95.866	87.647	87.012
	PhymmBL	97.88	97.63	96.09
Phylum	PhyloPythiaS	100	92.01	91.78
	MEGAN	91.8	99.97	92.3
	Phymm	99.913	95.014	94.868
	PhymmBL	99.92	98.61	98.51
Domain	PhyloPythiaS	99.62	99.99	99.63
	MEGAN	100	95.8	95.79
	Phymm	99.986	98.96	98.946
	PhymmBL	99.99	99.59	99.58

B. New genus

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	92.234	31.669	41.919
	MEGAN	65.172	3.209	2.477
	Phymm	65.357	17.733	18.722
	PhymmBL	67.049	10.225	8.17
Family	PhyloPythiaS	98.537	45.807	67.114

	MEGAN	77.883	34.96	34.186
	Phymm	93.579	36.608	27.645
	PhymmBL	86.752	43.692	40.167
Order	PhyloPythiaS	92.572	48.489	72.437
	MEGAN	73.003	43.893	55.289
	Phymm	69.735	48.046	55.07
	PhymmBL	74.495	56.876	69.194
Class	PhyloPythiaS	99.367	73.169	82.154
	MEGAN	96.564	61.151	63.597
	Phymm	93.476	74.413	74.887
	PhymmBL	94.319	80.817	81.415
Phylum	PhyloPythiaS	99.954	88.968	88.668
	MEGAN	99.983	78.984	78.582
	Phymm	99.91	91.626	91.501
	PhymmBL	99.913	94.863	94.772
Domain	PhyloPythiaS	99.986	99.685	99.672
	MEGAN	100	91.623	91.611
	Phymm	99.986	98.494	98.481
	PhymmBL	99.986	98.987	98.974

C. New order

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	76.2	28.787	38.155
	MEGAN	61.161	3.654	2.942
	Phymm	63.283	21.365	24.62
	PhymmBL	63.845	13.566	12.44
Family	PhyloPythiaS	98.342	45.134	54.961
	MEGAN	70.381	11.51	4.461
	Phymm	92.832	39.47	31.326
	PhymmBL	75.222	32.713	15.738
Order	PhyloPythiaS	86.092	47.797	54.961
	MEGAN	69.091	12.843	4.407
	Phymm	70.221	39.69	31.422
	PhymmBL	69.601	34.713	15.944
Class	PhyloPythiaS	99.486	64.168	75.667
	MEGAN	76.498	27.257	33.037
	Phymm	90.652	66.357	61.53
	PhymmBL	88.383	64.532	56.152
Phylum	PhyloPythiaS	99.969	88.61	88.354
	MEGAN	99.979	65.693	65.376
	Phymm	99.904	86.159	86.068
	PhymmBL	99.907	88.569	88.518

Domain	PhyloPythiaS	99.986	99.576	99.562
	MEGAN	100	88.626	88.614
	Phymm	99.986	96.072	96.059
	PhymmBL	99.986	96.606	96.592

D. New class

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	89.988	30.105	42.398
	MEGAN	80.309	3.002	3.394
	Phymm	82.975	27.743	33.242
	PhymmBL	82.832	17.216	17.203
Family	PhyloPythiaS	99.445	36.593	67.702
	MEGAN	97.936	3.313	4.557
	Phymm	99.94	30.083	41.946
	PhymmBL	99.849	20.211	20.857
Order	PhyloPythiaS	99.445	36.589	67.634
	MEGAN	97.936	3.313	4.42
	Phymm	99.94	30.08	42.028
	PhymmBL	99.849	20.21	20.939
Class	PhyloPythiaS	99.978	54.028	68.551
	MEGAN	98.551	4.888	4.612
	Phymm	100	44.171	42.233
	PhymmBL	100	29.718	21.076
Phylum	PhyloPythiaS	99.966	81.077	80.936
	MEGAN	99.972	48.396	48.187
	Phymm	99.911	77.717	77.693
	PhymmBL	99.912	78.391	78.391
Domain	PhyloPythiaS	99.986	99.453	99.439
	MEGAN	100	84.848	84.836
	Phymm	99.986	94.676	94.663
	PhymmBL	99.986	94.552	94.539

Supplementary Table 2: Performance of different binning methods for the abundant populations in the TW sample. Assignment accuracy is evaluated based on the consistency of taxonomic assignment for contigs of the same scaffold (see Supplementary notes). Sample specific data was used for all methods.

Method	Population	Kilo-bases assigned	Scaffold-contig consistency (% bp)	Scaffold-contig consistency (average taxonomic distance)
PhyloPythiaS	WG-1	2,669.60	97.71	0.38
	WG-2	2,512.93	97.24	0.34
	WG-3	892.65	94.11	0.43
	Total	13,552.86	78.54	0.44
PhyloPythia	WG-1	2,674.70	97.94	0.29
	WG-2	2,326.76	89.75	0.53
	WG-3	870.60	94.70	0.35
	Total	12,830.05	82.90	0.43
PhymmBL	WG-1	3,542.94	69.90	0.72
	WG-2	2,809.81	56.69	1.12
	WG-3	1,005.99	64.59	1.12
	Total	13,286.18	60.78	1.01
MEGAN	WG-1	1,100.20	90.28	0.44
	WG-2	646.19	81.99	0.46
	WG-3	142.69	95.27	0.27
	Total	8,604.92	86.91	0.41

Supplementary Table 3: NUCmer analysis of the WG-1 assignments for TW sample. PhyloPythia and PhyloPythiaS assigned 344 and 604 contigs respectively to WG-1. The recovered genome size is 2,952,624 bp. The contigs assigned by PhyloPythia and PhyloPythiaS were mapped on the WG-1 genome with NUCmer, which corresponded to 1,995,748 bp and 2,104,034 bp respectively. See supplementary notes for details.

	PhyloPythia filtered	**PhyloPythia unfiltered**	**PhyloPythiaS filtered**	**PhyloPythiaS unfiltered**
# contigs aligned	357 (98%)	359 (98%)	525 (87%)	543 (90%)
Length match (bp)	1,798,591	1,941,532	1,803,892	1,972,064
Coverage (%)	90.09	97.28	85.77	93.7
Average IDY (%)	98.92	95.14	98.90	95.50

Supplementary Table 4: Modeled clades for the TW sample. Only the leaf clades are shown, all the clades at more general taxonomic ranks were included in the modeled taxonomy.

NCBI scientific name	NCBI taxonomic identifier	Sample-specific data (kb)
Acinetobacter	469	--
Actinobacteria (class)	1760	--
Bradyrhizobiaceae	41294	--
Campylobacter	194	--
Desulfovibrionaceae	194924	--
Enterobacteriaceae	543	--
Eubacteriaceae	186806	--
Fusobacteriaceae	203492	--
Methanomicrobiales	2191	--
Methanosarcina	2207	--
Pasteurellaceae	712	--
Prevotellaceae	171552	--
Psychrobacter	497	--
Ruminococcaceae	541000	--
Selenomonas	970	--
Staphylococcus	1279	--
Thermoplasma	2302	--
uncultured Erysipelotrichaceae bacterium (WG-3)	331630	5.7
uncultured Lachnospiraceae bacterium (WG-2)	297314	143
uncultured Succinivibrionaceae bacterium (WG-1)	538960	257

Supplementary Table 5: Taxonomic assignments for abundant genera in the human gut metagenome samples. Assignment accuracy is evaluated based on the consistency of taxonomic assignment for contigs of the same scaffold (see Supplementary notes).

Method	Genus-level bin / Population	Kilo-bases assigned		Scaffold-contig consistency (% bp)		Scaffold-contig consistency (average taxonomic distance)	
		TS28	TS29	TS28	TS29	TS28	TS29
PhyloPythiaS	*Ruminococcus*	13,787.33	13,016.96	95.10	94.68	0.16	0.20
	Faecalibacterium	17,049.71	8,490.69	93.44	90.75	0.18	0.16
	Clostridium	8296.77	3376.53	89.41	95.74	0.24	0.22
	Eubacterium	8840.37	2515.17	98.05	76.63	0.10	0.30
	Dorea	2,443.36	1,323.47	98.75	96.05	0.11	0.30
	Bifidobacterium	4,948.32	4,760.12	98.51	99.97	0.08	0.05
PhyloPythia	*Ruminococcus*	16,879.06	14,918.45	94.78	90.18	0.15	0.29
	Faecalibacterium	19,962.39	9,372.68	94.80	85.72	0.28	0.25
	Clostridium	11,797.44	4,097.59	77.42	85.62	0.39	0.45
	Eubacterium	10,138.96	1,859.18	97.12	89.78	0.16	0.51
	Dorea	3,412.84	1,511.66	97.21	82.30	0.11	0.49
	Bifidobacterium	4,946.77	4,767.18	98.40	99.78	0.06	0.03
PhymmBL	*Ruminococcus*	6,613.42	5,694.06	96.11	94.87	0.10	0.09
	Faecalibacterium	15,302.09	6,423.28	94.09	93.96	0.12	0.07
	Clostridium	13,246.25	4,917.47	87.30	92.22	0.22	0.19
	Eubacterium	5,624.48	1,337.88	98.01	85.77	0.08	0.26
	Dorea	3,118.58	1,381.38	97.61	82.95	0.05	0.21
	Bifidobacterium	5,057.49	4,757.60	97.96	99.93	0.11	0.03

Supplementary Table 6: Bin validation for the human gut metagenome samples using marker genes.
See supplementary notes for details.

Method	Sample	Domain	Phylum	Class	Family	Genus
PhyloPythia	TS28	99.21	89.57	87.20	69.49	55.51
	TS29	100.00	95.98	95.44	81.23	72.07
--	Controls	99.80	98.16	97.74	80.01	71.69
PhyloPythiaS	TS28	99.40	91.33	88.98	75.96	61.07
	TS29	100.00	95.11	95.11	86.18	76.41

Supplementary Table 7: Validation for the human gut metagenome samples using CD-HIT (fraction matched). See supplementary notes for details.

Method	Comparison	Domain	Phylum	Class	Family	Genus
PhyloPythia	TS28vTS29	99.85	95.97	95.96	71.99	72.34
	Genomes	98.07	96.02	94.18	77.58	68.97
	TS29vGenomes	98.74	94.89	92.07	58.73	56.46
	TS28vGenomes	98.66	93.43	90.54	55.79	51.57
PhyloPythiaS	TS28vTS29	99.82	96.32	96.32	79.03	77.17
	Genomes	97.98	95.84	93.92	76.57	67.58
	TS29vGenomes	98.75	95.17	92.62	63.10	59.49
	TS28vGenomes	98.74	93.73	91.22	60.11	54.27

Supplementary Table 8: Modeled clades for PhyloPythiaS for the human gut metagenome samples (TS28 and TS29). Only the leaf clades are shown, all the clades at more general taxonomic ranks were included in the modeled taxonomy. Only part of the sample-specific data was used to learn PhyloPythia and PhyloPythiaS models (see Supplementary notes).

NCBI scientific name	NCBI taxonomic identifier	Sample-specific data (kb)
Alistipes	239,759	198
Anaerococcus	165,779	1,300
Anaerotruncus	244,127	74
Atopobium	1,380	--
Bacteroides	816	23,600
Bifidobacterium	1,678	3,800
Blautia	572,511	13
Butyrivibrio	830	6.2
Clostridium	1,485	7,200
Collinsella	102,106	512
Coprococcus	33,042	29
Dorea	189,330	1,500
Escherichia	561	--
Eubacterium	1,730	600
Faecalibacterium	216,851	2,300
Finegoldia	150,022	--
Holdemania	61,170	7.7
Lactococcus	1,357	--
Methanobrevibacter	2,172	1,300
Methanothermobacter	145,260	--
Parabacteroides	375,288	1,600
Porphyromonas	836	--
Providencia	586	--
Roseburia	841	31
Ruminococcus	1,263	4,000
Streptococcus	1,301	--

Supplementary Table 9: Statistical comparison of the assignments of different methods on TW data set. P-values obtained with two-tailed Wilcoxon paired sum-ranks test for different methods on the scaffold-contig consistency and kilo-bases assigned for 230 clades (union of all predicted clades). The bold values indicate pairs where the null hypothesis is rejected at 95% confidence. This table shows that PhymmBL is significantly different than other methods in both kilo-bases assigned and scaffold-contig consistency.

Methods	Scaffold-contig consistency	Kilo-bases assigned
PhyloPythiaS – PhyloPythia	**0.0338**	0.4242
PhyloPythiaS – PhymmBL	**5.5454e-09**	**1.7678e-07**
PhyloPythiaS – MEGAN	0.5720	0.8605
PhyloPythia - PhymmBL	**1.1306e-11**	**6.2198e-11**
PhyloPythia – MEGAN	0.0591	0.5781
PhymmBL – MEGAN	**2.0417e-12**	**8.0705e-06**

Supplementary Table 10: Number of contigs classified by different methods at different taxonomic ranks for the TW sample. There are 5,995 contigs in total for this metagenome sample. All numbers indicate the raw output of every method. PhyloPythia does not classify fragments shorter than 1,000 bp so the total number of contigs is less (5,245).

Taxonomic rank	PhyloPythiaS	PhyloPythia	PhymmBL	MEGAN
Domain	1,206	1,579	--	630
Phylum	503	485	--	191
Class	214	261	92	85
Order	1,748	801	1,086	401
Family	997	1,012	250	288
Genus	71	--	2,899	1,446
Species	1,255	1,062	1,525	277
Not assigned	1	45	143	2,677

Supplementary Table 11: Effect of sample specific data on the assignment of the TW sample for PhyloPythiaS and PhymmBL. The "#predictions" columns shows number of predictions obtained using the sample specific models and for both the sample specific and the non-sample specific models. The "#consistent predictions" column shows how many of these predictions are taxonomically consistent with the respective population. The Average distance column shows the average taxonomic distance between the predictions of the sample specific and non-sample specific models. For WG-2 PhymmBL without sample specific data made the specified number of consistent assignments to *Lachnospiraceae* due to relabeled *Ruminococcus*.

Population	Method	#predictions (sample-specific)	#predictions (joint)	#consistent predictions	Average taxonomic distance
WG-1	PhymmBL	530	434	0	8.93
	PhyloPythiaS	477	477	361	5.13
WG-2	PhymmBL	708	690	205	5.37
	PhyloPythiaS	482	482	419	2.05
WG-3	PhymmBL	286	201	0	8.59
	PhyloPythiaS	296	296	266	3.29

Supplementary Table 12: Genomes used for simulated short fragment test data set.The "parent" columns show the lowest parent available in the reference taxonomy. The genomes were chosen such that their parent at genus level was not represented in the reference data.

Organism (Taxonomic identifier)	Organism (Scientific name)	Parent (Taxonomic identifier)	Parent (Scientific name)	Parent (Taxonomic rank)
313624	*Nodularia spumigena* CCY 9414	1162	Nostocaceae	family
59196	*Rickettsiella grylli*	118968	Coxiellaceae	family
391597	*Limnobacter* sp. MED105	119060	Burkholderiaceae	family
214688	*Gemmata obscuriglobus* UQM 2246	126	Planctomycetaceae	family
314230	*Blastopirellula marina* DSM 3645	126	Planctomycetaceae	family
344747	*Planctomyces maris* DSM 8797	126	Planctomycetaceae	family
278957	*Opitutaceae bacterium* TAV2	134623	Opitutaceae	family
392484	*Methylophaga thiooxidans*	135616	Piscirickettsiaceae	family
207949	*Bermanella marisrubri*	135620	Oceanospirillaceae	family
207954	*Neptuniibacter caesariensis*	135620	Oceanospirillaceae	family
391606	*Carboxydibrachium pacificum* DSM 12653	186814	Thermoanaerobacteraceae	family
324057	*Paenibacillus* sp. JDR-2	186822	Paenibacillaceae	family
392917	*Paenibacillus larvae* subsp. larvae BRL-230010	186822	Paenibacillaceae	family
240016	*Verrucomicrobium spinosum* DSM 4136	203557	Verrucomicrobiaceae	family
391592	*Caminibacter mediatlanticus* TB-2	224467	Nautiliaceae	family
219305	*Micromonospora sp. ATCC* 39149	28056	Micromonosporaceae	family
244592	*Labrenzia alexandrii* DFL-11	31989	Rhodobacteraceae	family
252305	*Oceanicola batsensis* HTCC2597	31989	Rhodobacteraceae	family
314232	*Loktanella vestfoldensis* SKA53	31989	Rhodobacteraceae	family
314256	*Oceanicola granulosus* HTCC2516	31989	Rhodobacteraceae	family
314264	*Roseovarius* sp. 217	31989	Rhodobacteraceae	family
314265	*Roseovarius* sp. HTCC2601	31989	Rhodobacteraceae	family
314267	*Sulfitobacter* sp. NAS-14.1	31989	Rhodobacteraceae	family
383629	*Phaeobacter gallaeciensis* 2.10	31989	Rhodobacteraceae	family
384765	*Labrenzia aggregata* IAM 12614	31989	Rhodobacteraceae	family
388399	*Sagittula stellata* E-37	31989	Rhodobacteraceae	family
391613	*Roseovarius* sp. TM1035	31989	Rhodobacteraceae	family
391616	*Octadecabacter antarcticus*	31989	Rhodobacteraceae	family

	238			
391619	*Phaeobacter gallaeciensis* BS107	31989	Rhodobacteraceae	family
391624	*Oceanibulbus indolifex* HEL-45	31989	Rhodobacteraceae	family
391626	*Octadecabacter antarcticus* 307	31989	Rhodobacteraceae	family
52598	*Sulfitobacter* sp. EE-36	31989	Rhodobacteraceae	family
89187	*Roseovarius nubinhibens* ISM	31989	Rhodobacteraceae	family
279714	*Lutiella nitroferrum* 2002	481	Neisseriaceae	family
216432	*Croceibacter atlanticus* HTCC2559	49546	Flavobacteriaceae	family
313590	*Dokdonia donghaensis* MED134	49546	Flavobacteriaceae	family
313594	*Polaribacter irgensii* 23-P	49546	Flavobacteriaceae	family
313595	*Psychroflexus torquis* ATCC 700755	49546	Flavobacteriaceae	family
313596	*Robiginitalea biformata* HTCC2501	49546	Flavobacteriaceae	family
313598	*Polaribacter sp.* MED152	49546	Flavobacteriaceae	family
391587	*Kordia algicida* OT-1	49546	Flavobacteriaceae	family
411465	*Parvimonas micra* ATCC 33270	543310	Clostridiales Family XI. Incertae Sedis	family
392423	*Hydrogenivirga* sp. 128-5-R1-1	64898	Aquificaceae	family
314254	*Oceanicaulis alexandrii* HTCC2633	69657	Hyphomonadaceae	family
314278	*Nitrococcus mobilis* Nb-231	72276	Ectothiorhodospiraceae	family
391600	*Brevundimonas* sp. BAL3	76892	Caulobacteraceae	family
399795	*Comamonas testosteroni* KF-1	80864	Comamonadaceae	family
313606	*Microscilla marina* ATCC 23134	89373	Flexibacteraceae	family
165597	*Crocosphaera watsonii* WH 8501	1118	Chroococcales	order
180281	*Cyanobium* sp. PCC 7001	1118	Chroococcales	order
118168	*Microcoleus chthonoplastes* PCC 7420	1150	Oscillatoriales	order
313612	*Lyngbya* sp. PCC 8106	1150	Oscillatoriales	order
322866	*Leptolyngbya valderiana* BDU 20041	1150	Oscillatoriales	order
156578	*Alteromonadales bacterium* TW-7	135622	Alteromonadales	order
58051	*Moritella* sp. PE36	135622	Alteromonadales	order
391574	*Vibrionales bacterium* SWAT-3	135623	Vibrionales	order
401526	*Thermosinus carboxydivorans* Nor1	186802	Clostridiales	order
411459	*Ruminococcus obeum* ATCC 29174	186802	Clostridiales	order

411460	*Ruminococcus torques* ATCC 27756	186802	Clostridiales	order
411461	*Dorea formicigenerans* ATCC 27755	186802	Clostridiales	order
411462	*Dorea longicatena* DSM 13814	186802	Clostridiales	order
411463	*Eubacterium ventriosum* ATCC 27560	186802	Clostridiales	order
411469	*Eubacterium hallii* DSM 3353	186802	Clostridiales	order
411470	*Ruminococcus gnavus* ATCC 29149	186802	Clostridiales	order
411471	*Subdoligranulum variabile DSM 15176*	186802	Clostridiales	order
411474	*Coprococcus eutactus* ATCC 27759	186802	Clostridiales	order
411485	*Faecalibacterium prausnitzii* M21/2	186802	Clostridiales	order
333990	*Carnobacterium* sp. AT7	186826	Lactobacillales	order
313603	*Flavobacteriales bacterium* HTCC2170	200644	Flavobacteriales	order
391598	*Flavobacteria bacterium* BAL38	200644	Flavobacteriales	order
391603	*Flavobacteriales bacterium* ALC-1	200644	Flavobacteriales	order
388413	*Algoriphagus* sp. PR1	200666	Sphingobacteriales	order
391596	*Pedobacter* sp. BAL39	200666	Sphingobacteriales	order
313589	*Janibacter* sp. HTCC2649	2037	Actinomycetales	order
321955	*Brevibacterium linens* BL2	2037	Actinomycetales	order
411466	*Actinomyces odontolyticus* ATCC 17982	2037	Actinomycetales	order
314270	*Rhodobacterales bacterium* HTCC2083	204455	Rhodobacterales	order
314271	*Rhodobacterales bacterium* HTCC2654	204455	Rhodobacterales	order
388401	*Rhodobacterales bacterium* HTCC2150	204455	Rhodobacterales	order
383631	*Methylophilales bacterium* HTCC2181	206350	Methylophilales	order
333146	*Ferroplasma acidarmanus* fer1	2301	Thermoplasmatales	order
378806	*Stigmatella aurantiaca* DW4/3-1	29	Myxococcales	order
391625	*Plesiocystis pacifica* SIR-1	29	Myxococcales	order
314231	*Fulvimarina pelagi* HTCC2506	356	Rhizobiales	order
320771	*bacterium Ellin*514	48461	Verrucomicrobiales	order
382464	*Verrucomicrobiae bacterium* DG1235	48461	Verrucomicrobiales	order
281689	*Desulfuromonas acetoxidans* DSM 684	69541	Desulfuromonadales	order
156586	*Flavobacteria bacterium*	117743	Flavobacteria	class

	BBFL7			
247633	*marine gamma proteobacterium* HTCC2143	1236	Gammaproteobacteria	class
247634	*marine gamma proteobacterium* HTCC2148	1236	Gammaproteobacteria	class
247639	*marine gamma proteobacterium* HTCC2080	1236	Gammaproteobacteria	class
314283	*Reinekea* sp. MED297	1236	Gammaproteobacteria	class
314285	*Congregibacter litoralis* KT71	1236	Gammaproteobacteria	class
391615	gamma proteobacterium HTCC5015	1236	Gammaproteobacteria	class
394104	*Endoriftia persephone* 'Hot96_1+Hot96_2'	1236	Gammaproteobacteria	class
312284	marine actinobacterium PHSC20C1	1760	Actinobacteria (class)	class
314260	Parvularcula bermudensis HTCC2503	28211	Alphaproteobacteria	class
331869	alpha proteobacterium BAL199	28211	Alphaproteobacteria	class
314607	beta proteobacterium KB13	28216	Betaproteobacteria	class
262489	delta proteobacterium MLMS-1	28221	Deltaproteobacteria	class

Supplementary Table 13: Performance evaluation of the different binning methods on a simulated data set of short fragments of varying lengths. Note that the performance measures are not monotonic with respect to the taxonomic ranks; as the lowest-level clades representing the test fragments belong to different taxonomic ranks for different organisms of the test data (Supplementary Table 3 and supplementary notes online), thus performance may differ across taxonomic ranks. The evaluation at genus rank quantifies over-binning. "PhyloPythiaS-3fold" rows show an evaluation in a 3-fold validation setup, in which complete genome sequences and whole genome assemblies were pooled and randomly split into 3 partitions.

A. 100 bp

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	--	91.60	91.60
	MEGAN	--	87.81	87.81
	Phymm	--	2.31	2.31
	PhymmBL	--	1.67	1.67
	PhyloPythiaS-3fold	--	74.54	74.54
Family	PhyloPythiaS	23.04	4.93	46.58
	MEGAN	61.71	17.06	54.27
	Phymm	15.19	5.74	9.50
	PhymmBL	33.30	18.57	14.58
	PhyloPythiaS-3fold	22.76	11.77	47.62
Order	PhyloPythiaS	18.20	6.02	14.51
	MEGAN	61.66	15.82	22.66
	Phymm	16.63	11.39	10.77
	PhymmBL	27.88	22.98	19.40
	PhyloPythiaS-3fold	20.92	9.66	22.02
Class	PhyloPythiaS	34.31	10.77	14.66
	MEGAN	70.86	15.00	17.86
	Phymm	22.69	16.95	24.91
	PhymmBL	32.97	25.10	33.94
	PhyloPythiaS-3fold	23.28	14.27	20.39
Phylum	PhyloPythiaS	37.76	12.27	25.97
	MEGAN	75.17	10.65	18.54
	Phymm	26.95	21.77	41.69
	PhymmBL	36.72	29.93	50.72
	PhyloPythiaS-3fold	25.05	15.40	28.18
Domain	PhyloPythiaS	51.16	59.25	90.86
	MEGAN	99.47	34.01	67.33
	Phymm	51.13	55.37	91.01
	PhymmBL	52.07	57.64	93.54
	PhyloPythiaS-3fold	70.03	51.72	92.72

B. 300 bp

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	--	91.50	91.50
	MEGAN	--	79.62	79.62
	Phymm	--	2.74	2.74
	PhymmBL	--	2.19	2.19
	PhyloPythiaS-3fold	--	72.33	72.33
Family	PhyloPythiaS	32.12	6.31	47.93
	MEGAN	65.72	23.19	54.41
	Phymm	24.11	11.19	12.79
	PhymmBL	41.09	24.80	19.81
	PhyloPythiaS-3fold	38.02	19.30	49.68
Order	PhyloPythiaS	29.65	9.66	18.23
	MEGAN	71.65	23.98	29.05
	Phymm	24.76	20.06	17.71
	PhymmBL	38.27	33.18	28.25
	PhyloPythiaS-3fold	33.33	16.00	27.26
Class	PhyloPythiaS	44.76	15.39	22.09
	MEGAN	80.84	23.11	26.81
	Phymm	33.18	24.74	35.32
	PhymmBL	46.30	36.65	45.68
	PhyloPythiaS-3fold	33.50	20.69	28.89
Phylum	PhyloPythiaS	45.12	18.38	33.80
	MEGAN	88.09	21.29	28.48
	Phymm	36.38	30.29	50.82
	PhymmBL	52.89	41.78	59.33
	PhyloPythiaS-3fold	34.80	20.06	37.48
Domain	PhyloPythiaS	52.50	66.22	92.89
	MEGAN	99.86	40.34	63.21
	Phymm	53.58	62.29	92.96
	PhymmBL	57.20	72.46	94.46
	PhyloPythiaS-3fold	73.82	54.61	94.53

C. 500 bp

Rank	Method	Specificity	Sensitivity	Accuracy
Genus	PhyloPythiaS	--	90.53	90.53
	MEGAN	--	74.61	74.61
	Phymm	--	2.86	2.86
	PhymmBL	--	2.27	2.27
	PhyloPythiaS-3fold	--	69.90	69.90
Family	PhyloPythiaS	41.89	8.45	49.73
	MEGAN	67.88	26.60	54.66

	Phymm	33.48	14.85	15.56
	PhymmBL	50.46	29.04	22.75
	PhyloPythiaS-3fold	43.89	24.16	51.02
	PhyloPythiaS	35.25	13.13	21.35
	MEGAN	67.34	28.58	32.28
Order	Phymm	31.33	25.44	21.93
	PhymmBL	42.59	38.44	32.56
	PhyloPythiaS-3fold	38.76	20.28	31.21
	PhyloPythiaS	52.60	19.15	28.18
	MEGAN	81.59	28.25	31.74
Class	Phymm	38.96	28.77	40.35
	PhymmBL	53.33	41.67	50.70
	PhyloPythiaS-3fold	37.85	24.72	33.97
	PhyloPythiaS	51.87	23.06	39.15
	MEGAN	88.94	27.62	34.57
Phylum	Phymm	45.52	34.66	54.77
	PhymmBL	60.84	46.52	63.28
	PhyloPythiaS-3fold	38.78	22.74	42.49
	PhyloPythiaS	52.75	67.33	93.14
	MEGAN	99.85	50.73	71.04
Domain	Phymm	55.23	66.02	93.48
	PhymmBL	59.65	77.14	94.92
	PhyloPythiaS-3fold	76.30	56.04	95.29

D. 800 bp

Rank	Method	Specificity	Sensitivity	Accuracy
	PhyloPythiaS	--	90.21	90.21
	MEGAN	--	67.93	67.93
Genus	Phymm	--	3.49	3.49
	PhymmBL	--	2.64	2.64
	PhyloPythiaS-3fold	--	66.65	66.65
	PhyloPythiaS	48.36	10.76	52.13
	MEGAN	65.61	28.25	53.92
Family	Phymm	40.26	20.54	19.42
	PhymmBL	54.06	33.05	25.97
	PhyloPythiaS-3fold	48.18	30.10	52.21
	PhyloPythiaS	45.18	16.98	26.07
	MEGAN	64.05	32.01	36.07
Order	Phymm	33.30	30.79	27.52
	PhymmBL	45.17	43.20	37.88
	PhyloPythiaS-3fold	42.88	25.24	35.40
Class	PhyloPythiaS	59.58	22.13	33.99

	MEGAN	78.37	31.96	37.71
	Phymm	44.69	32.67	45.32
	PhymmBL	55.34	45.53	55.54
	PhyloPythiaS-3fold	42.56	28.88	39.32
	PhyloPythiaS	58.70	29.27	44.46
	MEGAN	88.53	33.05	43.20
Phylum	Phymm	52.97	39.03	57.68
	PhymmBL	66.04	50.47	66.04
	PhyloPythiaS-3fold	42.41	25.82	47.71
	PhyloPythiaS	55.45	81.52	94.24
	MEGAN	99.91	54.84	73.81
Domain	Phymm	58.00	72.34	93.44
	PhymmBL	62.10	81.09	94.96
	PhyloPythiaS-3fold	78.53	57.23	95.88

E. **1000 bp**

Rank	Method	Specificity	Sensitivity	Accuracy
	PhyloPythiaS	--	89.44	89.44
	MEGAN	--	66.32	66.32
Genus	Phymm	--	3.21	3.21
	PhymmBL	--	2.58	2.58
	PhyloPythiaS-3fold	--	65.40	65.40
	PhyloPythiaS	49.04	11.67	53.00
	MEGAN	65.48	29.02	53.79
Family	Phymm	45.68	21.81	20.35
	PhymmBL	57.44	34.58	27.04
	PhyloPythiaS-3fold	50.63	33.08	53.10
	PhyloPythiaS	46.58	19.29	27.71
	MEGAN	65.02	32.64	36.95
Order	Phymm	38.01	32.25	29.11
	PhymmBL	47.24	44.65	39.58
	PhyloPythiaS-3fold	45.23	27.51	37.62
	PhyloPythiaS	64.88	24.66	36.67
	MEGAN	78.57	33.22	39.31
Class	Phymm	46.31	33.94	46.86
	PhymmBL	57.99	46.90	57.01
	PhyloPythiaS-3fold	43.35	31.01	41.94
	PhyloPythiaS	63.71	31.75	47.03
	MEGAN	89.64	34.66	45.18
Phylum	Phymm	54.95	39.97	58.97
	PhymmBL	67.20	51.17	67.18
	PhyloPythiaS-3fold	44.54	27.40	50.23

Domain	PhyloPythiaS	56.08	81.58	94.89
	MEGAN	99.95	55.10	73.82
	Phymm	57.40	70.10	93.71
	PhymmBL	61.11	78.62	94.91
	PhyloPythiaS-3fold	79.55	57.76	96.21

Supplementary Table 14: Execution time comparison for different methods for characterization of the three real metagenome samples. The sample sizes are approximately 16 Mb, 113 Mb and 72 Mb for TW, TS28 and TS29 respectively.

Method	Time (DD:HH:MM:SS)		
	TW	TS28	TS29
PhyloPythiaS	00:00:08:36	00:01:13:43	00:00:46:28
PhyloPythia	00:03:12:43	01:08:04:25	00:21:18:27
PhymmBL	00:15:09:51	07:13:54:01	04:15:53:44
MEGAN	00:12:10:14	--	--

Supplementary Note

Genomes and Draft assemblies

As reference data for the different experiments we use the sequences of complete genomes and whole genome shotgun projects from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genbank/), state of May 2009. This corresponds to 851 complete genomes and 917 whole genome shotgun assemblies. The use of this data is described more in detail for the individual experiments.

Performance measures

For evaluation on the simulated data sets we compute the sensitivity and specificity[2] of assignments, averaged over all n clades at a respective taxonomic rank[3]. Thus, the average sensitivity, or macro-accuracy, and specificity are defined as follows:

$$specificity = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i}$$

$$sensitivity = \frac{1}{n+1}\left(\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} + \frac{TP_{-1}}{TP_{-1} + FN_{-1}}\right)$$

The index -1 denotes items that do not belong to any of the n modeled clades for a given rank. Furthermore, we compute the classification accuracy, which corresponds to the overall number of correctly classified items at a given taxonomic rank. Note that while the macro-accuracy measures the classification accuracy averaged over all classes

represented in a test data set, the accuracy measures classification performance for a given data set in a way that every input item contributes equally. This distinction becomes important if clades are represented unequally in a given data set, such as is often the case for metagenomic data. In this case, the overall classification accuracy becomes a more relevant performance measure than the macro-accuracy of assignments.

$$accuracy = \frac{TP}{TP + FN}$$

Ideally, a method should score well in terms of all measures.

As for real metagenome samples the correct taxonomic assignment of the fragments is not known, measuring the binning performance on real metagenome samples cannot be done with traditional measures such as the accuracy, sensitivity and specificity cannot be calculated. We use here an intuitive and informative measure for assessing the binning performance of a method[3].

Assume that for a metagenome data set the reads are assembled into contigs and that a set of contigs are known to jointly originate from a given genome, based on mate pair information, which is denoted by their grouping into a scaffold. A binning method is used for taxonomic assignment of the contigs. The scaffold-contig consistency measures the consistency of the taxonomic assignments for a scaffold in terms of its constituent contig assignments. For this purpose, each scaffold is first labeled with the assignment of

one of its constituent contigs with the lowest taxonomic rank. In case there are multiple lowest rank assignments, then the assignment with longest collective contig length is used. The consistency of scaffold assignments is then measured with respect to this taxonomic label. For each contig of a scaffold, the taxonomic assignment is considered to be consistent if it is either the same or a more general taxonomic assignment with respect to the true taxonomic origin of the scaffold; otherwise it is considered an inconsistent assignment. The percentage of consistently assigned contig base-pairs is the scaffold-contig consistency. The scaffold-contig consistency is then averaged over all the scaffolds with the same assignment, to measure the assignment consistency of a clade. Furthermore, we also calculate average taxonomic distance of contig assignments in terms of the path distance to the scaffold label as a more fine grained consistency measure. High scaffold-contig consistency is a desirable property for a binning method. For a given data set we use the same reference taxonomy for all the methods for calculating scaffold-contig consistency.

METHODS

Large margin structured output learning

Following standard nomenclature, we denote the input and output spaces by X and Y, respectively. A supervised method learns a function $f : X \to Y$, which assigns an output y to a given input x. In other words, a supervised learning algorithm learns a function f given a set of input-output pairs, $S = \{(x_i, y_i) \in X \times Y : i = 1,2,\ldots,N\}$ drawn from an unknown joint distribution $p(X,Y)$. This function can then be used to assign output to

new input items. A function is said to have a good generalization capacity if it performs well on unseen data. Vapnik[4] showed that a maximum margin classification technique, called support vector machine (SVM)[5], has a very good generalization ability. SVMs have become state-of-the-art methods for classification tasks and have shown excellent empirical performance. Recently, methods have been proposed[6-8] which extend the large margin framework to structured output problems. These methods exploit the known inter-dependencies between the elements in the output space to learn large margin functions.

We model the taxonomic classification problem as a structured output prediction problem where the inputs are the genomic sequence fragments and the outputs are the paths in a known taxonomy. The association between the input and output is encoded with a joint feature space. We introduce these concepts first before describing the learning and inference processes.

Input space

Various studies[3,9-11] have shown that the composition of a genomic sequence in terms of its constituent sub-strings carries a phylogenetic signal which can be used for phylogenetic classification. As before, we use sequence composition-derived features to represent the input space X. Counts of oligonucleotides of length 4, 5 and 6 represent a sequence fragments in this input space, which are normalized by the sequence fragment length. Each sequence is thus represented by a vector of length 5376, we denote this

transformation with φ. A suffix tree-based algorithm is used to compute sequence composition in linear time with respect to sequence length. Each input feature (oligonucleotide counts) is normalized to mean zero and standard deviation one. See supplementary methods for details.

Output space

Our output structure is a hierarchy of the evolutionary clades. This hierarchical representation encodes the evolutionary relationships between the clades. In particular, we use the relationships defined by the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/) as the reference, unless specified otherwise. In this structured representation, each output y corresponds to a valid path in the hierarchy. Each path can be encoded as a binary vector of length n, where n is the number of nodes. In this vector, the elements corresponding to the nodes in the path are set to one. We denote this binary encoding by $\wp$. A hidden terminal child node is added to an internal node if some training examples are assigned to it.

Joint feature space

Joint kernels are used to explicitly specify known input-output and output-output correlations resulting in an extended feature space. We encode the relationship between an input vector x and an output path y via a joint feature map Ψ. The joint feature map is defined as the Kronecker (tensor) product between the input sequence x and the binary vector representation of output path y.

$$\Psi(x,y) = \varphi(x) \otimes \wp(y)$$

The Kronecker product kernel can be decomposed into element-wise product of input and output kernels:

$$K_{XY}\left(\Psi(x,y),\Psi(x',y')\right) = K_X(x,x')^{\mathrm{T}} K_Y(y,y').$$

Loss function

The discrepancy between two outputs is measured using a loss function $\Delta(y,\hat{y})$. For structured output problems the traditional zero-one loss, used in flat classification, is not appropriate, because the distance between true and predicted output in the taxonomic tree has to be taken into account. In this case, problem-specific loss functions can be used, which satisfies two conditions: 1) the loss is 0 if two outputs are the same and 2) the loss function is monotonic with respect to the discrepancy to the true output. In the case of PhyloPythiaS the output structure is taxonomy, and an output is a path in the hierarchy. We measure the discrepancy between a pair of paths using the shortest path distance between the terminal nodes of the corresponding paths. This is accomplished by finding the lowest common ancestor of the terminal nodes. Various other loss functions can be implemented[12,13] but our experiments (data not shown) showed that they have little or no effect on the predictive performance.

Learning and Inference

The large-margin structured-output formulation learns a continuously-valued scoring function such that for a particular input the correct output ranks higher than any other output. In this case, the input space X corresponds to the genome sequence fragment derived features, while the output space Y corresponds to the taxonomic hierarchy. The learned function takes the form,

$$f(x;w) = argmax_{y \in Y} F(x,y;w) .$$

Here the scoring function F is a linear function;

$$F(x,y;w) = w^T \Psi(x,y).$$

where Ψ maps the input output pair (x,y) to a joint feature space. We use the maximum margin structural support vector machine framework[7] to find the parameter vector w. Specifically, we use a linear penalty term (1-slack) with the slack-rescaling formulation of the structural SVM[6] for learning the optimal w given a training set of size N. The primal formulation is given below;

$$\min_{w,\xi \geq 0} \frac{1}{2} w^T w + \frac{C}{N} \sum_{i=1}^{N} \xi_i$$

$$s.t.\ w^T\ \left[\Psi(x_i, y_i) - \Psi(x_i, \hat{y}_i)\right] \geq 1 - \frac{\xi_i}{\Delta(y_i, \hat{y}_i)} \quad \forall i,\ \forall \hat{y}_i \in Y$$

This formulation has only one slack variable per example that is shared by all constraints of the example. Joachims et al.[6] show that dual form of this problem has a sparse solution (in number of non-zero dual variables) independent of the size of the training set. Also a tractable cutting plane algorithm is proposed which iteratively adds the most strongly violated constraint to the working set. The reader is referred to Joachims et al.[6] for the dual formulation and more details on the algorithm.

The training procedure gives the optimal linear discriminant w for a given training set and the parameter C. At inference time, for classification of a new example, the score for the new input x is computed for all possible outputs in the joint feature space and the new example is assigned to the highest scoring output path.

$$argmax_{y \in Y} w^T \Psi(x, y)$$

Ensemble of classifiers

Individual classifiers can be inaccurate, especially if the data is noisy or heterogeneous. For this reason we use an ensemble of classifiers instead of a single classifier. The idea behind an ensemble strategy is that if one classifier makes a mistake it can be corrected

by other classifiers, which in effect improves specificity. We designed a simple ensemble strategy to combine multiple classifier outputs, where each output is a path in the hierarchy. We define an ensemble output as the majority vote lowest node of different output paths. In other words, we take the longest path in which majority of the classifiers are in agreement. Note that this ensemble supports the prediction of partial paths.

Sequence composition space

Sequence composition allows each sequence fragment to be represented as a fixed length vector and similarity between a pair of sequences can be computed using this vectorized representation. Let $f(\alpha_1,\alpha_2..\alpha_k)$ be frequency of an oligonucleotide of length k then for all possible oligonucleotides of length k with 4-letter DNA alphabet {A, C, G, T}. A sequence thus can be represented as a vector $\lfloor f_1, f_2, f_{4^k} \rfloor$ of length 4^k. In our experiments we use a concatenation of oligonucleotides of lengths 4, 5 and 6. Thus, each sequence fragment is represented as a vector of length 5376.

Suffix trees were used for efficient enumeration of the oligonucleotides in a string. A suffix tree can be constructed in linear time and space. Once a suffix tree is available, sub-strings can be enumerated in time linearly proportional to the sub-string length. We used the gsuffix library (http://gsuffix.sourceforge.net), which provides an implementation of generalized suffix tree algorithm in C language.

Pre-processing and models

As the sequence fragments can be of different lengths, we need to account for the lengths, so that they can be compared. Various normalization factors can be used, like string length, number of k-mers or markov-chains[14]. We use sequence length as normalization factor for PhyloPythiaS.

Once all the feature vectors for the sequence fragments in the training set are computed, the data set is represented as a matrix in the SVM-light sparse matrix format. The columns of this matrix represent the oligonucleotide features and the rows represent individual training sequence fragments. All the columns of this matrix (the oligonucleotide features) are normalized to have zero mean and standard deviation of one. The mean and standard deviation of every feature (or any other pre-processing information) is stored in the model files for later use in the prediction.

For efficiency purpose instead of storing the support vectors, which is a common practice, we directly store the optimal weight vector in the model file. As we strictly use a linear kernel this does not result in any penalties.

We use sequence fragments of length 1000, 3000, 5000, 10000, 15000 and 50000 to create six structural SVM models. At prediction time three models closest to the length of the test sequence are used in an ensemble. We used approximately 10,000 examples to train each model. All models were trained with a parameter setting of C=1000,

determined as setting with best-accuracy in 3-fold cross-validation on data used in[3]. When sample-specific data was available we use a sliding window of 10% of the fragment length to create the sample-specific training examples. The sample-specific examples were used in addition to the 10,000 examples generated using publically available sequences.

At prediction time three models closest to the length of the test sequence are used is an ensemble, as discussed above.

Use of dynamic programming

Each output in our structured output prediction problem is a path in taxonomy. Both learning and inference processes depend on a compatibility score, which measures the strength of association between an input-output pair. Let's assume two paths P1 and P2 in the hierarchy are represented by nodes $\{n_1, n_2, n_3\}$ and $\{n_1, n_2\}$, respectively. Please note that this not the binary representation of the paths, but just enumeration of constituent nodes. Furthermore assume the dependency relationship $n_1 \succ n_2 \succ n_3$, saying that n1 is parent of n2 and n2 is parent of n3. The compatibility score of these paths for a given input vector x are given by;

$$F(x, P_1) = F(x, n_1) + F(x, n_2) + F(x, n_3)$$

$$F(x, P_2) = F(x, n_1) + F(x, n_2)$$

Thus calculation of the compatibility score of the whole path, needs compatibility score of its constituent nodes. Using this knowledge, the hierarchy can be traversed in top-down fashion enumerating compatibility of all possible paths (outputs), which is necessary for both learning and inference. As a concrete example, the compatibility score for path P1 can be rewritten as follows;

$$F(x, P_1) = F(x, P_2) + F(x, n_3)$$

Same data for all methods

A fair comparison between different methods tailored for same task requires that they are provided with the same information. In the binning task we use different sources of reference sequences; NCBI complete genomes, NCBI draft assembles and sample specific data (when available). In the following section we describe the steps taken in order to allow the different taxonomic classification methods to make use of the same information.

Phymm and PhymmBL

The PhymmBL package was downloaded from the Nature Methods website (http://www.nature.com/nmeth/journal/v6/n9/extref/nmeth.1358-S2.zip). This software by default downloads the NCBI RefSeq data and builds IMM on the corresponding sequences. Phymm does not allow training on arbitrary sequences (unless

some specific conditions on the fasta headers and folder names are met). We changed the perl scripts to allow use of arbitrary training data, so that NCBI draft assemblies and sample specific data could be used.

MEGAN

MEGAN (version 3.9) was downloaded from the website (http://www-ab.informatik.uni-tuebingen.de/software/megan). MEGAN can detect standard NCBI names in the BLAST output, so including sample specific data was straight forward. We created various BLAST databases; NCBI complete genomes, NCBI draft assemblies and sample specific data (when available) using the "formatdb" program (available with blast at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/). For sample specific data, care was taken to include the organism names in the fasta headers before formatting them as a BLAST database, so that MEGAN could detect their taxonomic position. Default MEGAN parameters for LCA were used. Database searches were performed using blastn to appropriate databases using blast alias files. Complexity filter was turned off with option –F "m D" when performing blast searches.

PhyloPythiaS and PhyloPythia

Both PhyloPythia and PhyloPythiaS can directly incorporate arbitrary training data, given appropriate class labels. So using appropriate data for different experiments was straightforward.

TEST AND REFERENCE DATA SETS

All the experiments were performed in hold-out fashion. That is, if a genome sequence or a part of it was present in a test data set then the corresponding reference sequence was removed from the reference sequence set. A 3-fold validation experiment was performed on the simulated short fragment data sets (p. 43). Additional details, if any, are mentioned in the respective context.

Simulated acid mine drainage data set (simMC)

We analyzed the simulated acid mine drainage data set (simMC)[15] to evaluate the performance of the different binning methods. We used the data set of contigs assembled with the Arachne assembler, which consist of 7307 contigs of which ~99% come from six strains of three species (two strains each); *Rhodopseudomonas palustris, Bradyrhizobium sp. BTAi1* and *Xylella fastidiosa*. The average contig length is 2332 bp.

We used the NCBI complete genomes as reference/training of models for the taxonomic assignment of this data set. Controlled sets of genomes were excluded as described in the evaluation section below.

Simulated short fragment data sets (simSF)

The benchmark data sets were constructed with two constraints: First, the fragments to be characterized should not belong to any of the organisms represented among the reference sequences, as metagenome sample populations are rarely among the available sequenced isolate genomes. Secondly, they should be chosen such that the

closest reference genomes are found at different taxonomic ranks, to model different degrees of evolutionary relatedness of metagenome sample populations to available reference sequences. To simulate this set-up, sequences from the NCBI genomes database were used as reference data for model construction. One hundred isolate sequences from the NCBI whole genome shotgun database with no mapping to any of the genera of the reference data were used for testing. Of the latter, 48 belong to a family, 39 to an order and 13 to a class of the reference taxonomy (Supplementary table 3 online). Approximately 10,000 non-overlapping fragments of 100, 300, 500, 800 and 1000 bp in length were randomly sampled from the selected isolate sequences to create fragment test sets of varying lengths.

The simulated short fragment test set is sampled equally from a large number of organisms with varying degrees of relationships to the reference data. Fragment lengths vary from 100bp to 1000bp and the test sequence fragments belong to 'unknown genomes', i.e. they were not used for training or reference. This is one of the most complex tasks in metagenome sample classification; for a real sample corresponding to the task of assigning individual unassembled reads of rare organisms without reference sequences available to correct higher-level clades. The test fragments do not map to any genus in the reference taxonomy (or available reference sequences). The lowest clades that the fragments map to in the reference taxonomy are at varying taxonomic ranks above the rank of genus (Supplementary Table 3). Thus, no assignment to a genus-level clade is the optimal result for fragments of this data set; meaning that genus-level

assignment specificity can be computed, while sensitivity of assignments, indicates the portion of correctly 'not assigned' test fragments.

Besides the hold-out experiment, we furthermore performed 3-fold cross validation for PhyloPythiaS on the pooled data of complete genome sequences and whole genome assemblies. The data were split into three random sets according to their genus affiliations. Genome sequences belonging to one of these sets were used to generate short fragment test data, while the sequences of other two sets were used for training. This procedure was repeated for each of the three sets and assignment accuracy determined. The averaged sensitivity, specificity and accuracy values obtained are reported in Supplementary Table 2.

We used the genome sequences from NCBI complete genomes as reference data for model construction. For PhyloPythiaS, we built different length models (1000, 3000, 5000, 10000, 15000 and 50000 base-pairs) with approximately 10,000 input examples and the output hierarchy restricted to at least three genomes per leaf node. This resulted in a hierarchy of 192 nodes, out of which 101 were leaf nodes. 178 nodes had sequences assigned to them in training, as the lowest level representative in the hierarchy for a given organism.

Tammar wallaby gut microbiome (TW)

Microbial communities from the gut of the Australian Tammar wallaby *(Macropus*

eugenii) were sequenced by Sanger sequencing[16] (GenBank accession number ADGC00000000). This sample consists of approximately 13.572 Mb of assembled DNA sequence, with contig lengths varying in length from 438 bp to 27,865 bp (average length 2,276.38 bp) (Supplementary Fig. 1). 16S rRNA analysis determined that organisms from the phyla *Firmicutes* and *Bacteroidetes* and the gamma-subdivision of *Proteobacteria* are abundant. This sample contains three abundant microbial populations, namely Wallaby gut 1 (WG-1 - a population of an uncultured *Succinivibrionaceae* bacterium), WG-2 (of a novel deep branching lineage within the *Lachnospiraceae*) and WG-3 (a novel bacterium of the *Erysipelotrichaceae*).

For taxonomic sample characterization, sample-specific models were constructed by combining publicly available sequences from NBCI (complete genomes and draft assemblies) with sample-specific data identified based on taxonomic marker genes and sequencing of a scaffold metagenome library.

Human gut metagenome samples (HG-TS28 and HG-TS29)

Two metagenome sequence samples from the gut of two human monozygotic, female twins were obtained by 454 deep sequencing of the total fecal community DNA with 454 Titanium single- and paired-end protocols[17] (referred to as TS28 and TS29). We analyzed approximately 113 Mb and 72 Mb of assembled contig sequences for TS28 and TS29, respectively. Sample-specific training data was obtained with BLASTN homology searches versus a reference database of 118 sequenced gut genomes. Training data was

identified based on the following criteria; e-value<10-5, bitscore>50, percent identity>90, percent sequence aligned>90, and total contig length >2 kb. Furthermore, all significant matches were required to originate from the same reference genome.

PhyloPythiaS and PhyloPythia models were constructed for 29 (14+15) genus- and family-level clades abundant in the sample and relevant higher-level taxonomic clades (Supplementary Table 10) using data from 5,548 and 3,391 sample-specific contigs and 1,775 microbial complete and draft microbial genomes. For PhyloPythiaS, sample-specific data was selected with active sampling for training, while for PhyloPythia, a subset was taken. For the training of PhymmBL, only assembled and draft genome sequences were used. Due to excessive computational requirements of homology searches on this data set, we did not perform binning with MEGAN.

Evaluation on simulated short fragment data

Supplementary Fig. 2 and Table 2 summarize the performance of the different binning methods on the simulated short fragment data sets. None of the tested methods show acceptable performance on these data sets. As expected, all methods show improving performance with increasing fragment length and a trade-off between sensitivity and specificity. Overall, MEGAN shows the superior specificity compared to the other methods. MEGAN is conservative due to its LCA algorithm, in the sense that it makes very specific assignments at the cost of sensitivity. Of the sequence composition-based methods, PhyloPythiaS and Phymm, PhyloPythiaS shows better specificity with compromised sensitivity.

Both Phymm and PhymmBL show comparably low sensitivity at the genus level. This is caused by the composition of the test data, for which none of the test fragments belong to any of the genus-level clades that are part of the models. Both methods 'over-bin' by assigning a substantial fraction of sequences to genus-level clades that should be left unassigned. It is interesting to note the drastic performance improvement of PhymmBL compared to Phymm for all fragment lengths and at all taxonomic ranks. At family level, which is the lowest taxonomic rank with valid assignments to clades included in the model, the improvement in specificity is approximately 12-18%, with a bigger effect on shorter fragments, and around 13% improvements in sensitivity. This, together with high specificity observed for MEGAN indicates that sequence homology information is beneficial for short fragment assignment, in the absence of sample-specific training data and when closely related genomes are available. For sequence composition, we attribute part of the degraded performance to the comparatively weak and noisy compositional signal of short fragments.

A "dip" is observed in the specificity at the order level for PhyloPythiaS and other methods (Supplementary Table 2). This is due to the construction of the data set. More specifically, the test fragments have varying degree of evolutionary relationship with the reference sequences (Supplementary Table 3). This is the reason for non-monotonous behavior of the performance measures on this data set.

The high sensitivity observed for PhyloPythiaS at genus level is due to the fact that no assignments were made at this level. As none of the test fragments can be mapped to any reference genus, this is the correct behavior.

Evaluation for dominant populations from novel species, genera or order-level clades

As most of the microorganism diversity is still unknown[18] it is very unlikely that complete genome sequences of the dominant populations are available as a reference for the taxonomic assignment of a metagenome sample. In some cases it is possible to obtain limited amounts of sequence data for the dominant populations by phylogenetic analysis of conserved marker-genes for the sample or sequencing of additional fosmid libraries (sample-specific data). Thus, it is crucial to assess the performance of the binning methods when limited amounts or no reference data from closely related organisms are available.

On the simulated 'simMC' data set, we evaluated performance of the different taxonomic classification methods by retaining 100 kb randomly selected contiguous fragments from the three dominant strains each as reference data and removing all genomes of the (1) same genus, (2) same order and (3) same class for the dominant strains. These different experiments are referred to as 'New genus', 'New order' and 'New class' respectively. This allows us to examine the performance in more realistic settings. Supplementary Fig. 1 and Table 1 summarize the results. A drastic drop in the

sensitivity and accuracy of the alignment-based methods (MEGAN and PhymmBL) can be seen in the absence of the closely related genomes. This is due to the lack of homologous regions, as only 100 kb of sequence are available for the dominant populations. On the other hand, composition-based methods (PhyloPythiaS and Phymm) show better sensitivity and accuracy, of which PhyloPythiaS shows superior performance. This demonstrates strength of composition-based methods and the ability of PhyloPythiaS to learn accurate models from limited amounts of reference data.

For the assembled metagenome from the Tammar wallaby gut we evaluated the performance of PhyloPythiaS and PhymmBL in the presence and absence of the sample specific data. The results (Supplementary Tables 5 and 8) indicate PhymmBL's over-binning tendency of assigning most sequences to genus-level clades. These assignments can be misleading if genera of the dominant sample populations are not included in the reference model. For PhymmBL, out of 530 contigs that were assigned to WG-1, when sample-specific data was included, only 33 contigs were assigned to the consistent parental clade *Gammaproteobacteria* without sample-specific data, accompanied by a large number of inconsistent assignments in comparison to assignments of the sample-specific model. In contrast, for the same population, PhyloPythiaS assigned 243 out of 477 contigs to the consistent general clade *Bacteria*, in the absence of sample-specific data, thus avoiding false positive assignments. Similar observations were made for other populations (data not shown). This suggests that PhyloPythiaS is better at assigning fragments of the 'known unknowns' in metagenome data sets and is robust with respect to the reference data.

Evaluation for sample populations with closely related genomes available

When closely related complete genomes sequences are available for the populations in the metagenome sample, alignment-based methods are at an advantage as the sample fragments can be aligned to the respective reference genomes with high confidence.

For the simMC data set, in the 'known species' experiment, complete genome sequecnes from NCBI were used as reference data for model training with exception of those genomes used to create the simMC data set. No training data of the genomes of the respective populations included in simMC were used. Though the exact genomes were removed, the reference data includes genomes of either same species (for *Rhodopseudomonas palustris* and *Xylella fastidiosa*) or same genus (for *Bradyrhizobium sp. BTAi1*). The results are shown in Supplementary Fig. 1 and Table 1.A. At lower taxonomic ranks (genus and family), the alignment-based methods show higher sensitivity and accuracy compared to the composition-based methods. At higher taxonomic ranks sensitivity and accuracy of all methods become more similar. PhyloPythiaS maintains high specificity at all taxonomic ranks, while other methods show lower specificity at lower taxonomic ranks.

For the human gut metagenomes (HG-TS28 and HG-TS28) PhyloPythiaS and PhyloPythia consistently show a similar performance across all taxonomic ranks (Supplementary Tables 11-13). PhymmBL also shows a high scaffold-contig consistency, but, in

comparison, lesser amounts of sequence are characterized. As no sample-specific data was included for the training of PhymmBL for assignment of these samples, the high consistency observed in the absence of sample-specific training data is likely due to the large number of gut genome sequences from related taxa (122) in reference data which contribute to model quality.

NUCMER ANALYSIS OF PHYLOPYTHIAS AND PHYLOPYTHIA PREDICTIONS FOR WG-1 POPULATION OF THE TAMMAR WALLABY GUT MICROBIOME

NUCmer[19] was used to align the contigs predicted as WG-1 by PhyloPythiaS and PhyloPythia, respectively, to the 43 scaffolds obtained for the WG-1 genome. The alignment coordinate output (Supplementary material online) shows the contigs aligned to the genome of WG-1 for both filtered and unfiltered alignments. In filtering the NUCmer output has additionally been run through the delta-filter (settings -q and -r), which leave only the contig-genome alignments that form the longest consistent set for the query and reference. This reduces the coverage but gives a better mapping.

MARKER GENE BASED BIN VALIDATION OF THE HUMAN GUT FAECAL METAGENOMES

All genes from the microbiome bins were assigned to STRING orthologous[20] groups. A neighbor-joining tree was built using clustalw[21] version 2.0.12 for each set of marker genes after aligning the translated gene sequences from 122 gut genomes and the binned scaffolds[17]. Individual sequences were assigned to taxa based on the consensus

taxonomy of all sequences found at the first node. Additionally, the frequency of consistent taxonomy between database marker genes and nearest neighbor sequences was tallied, and used as a control for the frequency of mis-assignment due to alignment errors, improper clustering, and/or disagreement with the marker genes and NCBI taxonomy. We furthermore used cd-hit[22] to cluster the protein sequences of the gut samples and 122 gut genomes at 60% identity. The taxonomic consistency of genes within these clusters and the respective bin assignments was then analyzed. Both PhyloPythia and PhyloPythiaS show a high consistency of taxonomic bin assignments within protein clusters (supplementary Table 13).

Availability

The PhyloPythiaS implementation is freely available for academic use at

http://binning.bioinf.mpi-inf.mpg.de/download/

AUTHOR CONTRIBUTIONS

K.R.P. implemented the method and performed the experiments, A.C.M., K.R.P and P.H. designed the experimental set-up, P.P., M.M., P.J.T. contributed data sets, P.P. and P.J.T. performed data analysis and evaluation, K.R.P. and A.C.M wrote the paper, T.S. and A.C.M. devised the project and A.C.M. supervised the work.

References

1. Kestler, H.A. *et al.*, *BMC Bioinformatics* **9**, 67 (2008).
2. Baldi, P. & Brunak, S., *Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)*. (The MIT Press, 2001).
3. McHardy, A.C., Garcia-Martin, H., Tsirigos, A., Hugenholtz, P., & I., R., *Nat. Methods* **4** (1), 63-72 (2007).
4. Vapnik, V.N., *The Nature of Statistical Learning Theory*. (Springer, 1995).
5. Boser, B., Guyon, I., & Vapnik, V.N., in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (eds. Haussler, D.), 144-152 (ACM Press, 1992).
6. Joachims, T., Finley, T., & Yu, C., *Machine Learning* **77** (1), 27-59 (2009).
7. Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y., *J. Mach. Learn. Res.* **6**, 1453-1484 (2005).
8. Taskar, B., Guestrin, C., & Koller, D., *Advances in Neural Information Processing Systems 16* **16**, 25-32 (2004).
9. Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I., & Coster, J., *Genome Res.* **11** (8), 1404-1409 (2001).
10. Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., & Ikemura, T., *Genome Inform Ser Workshop Genome Inform* **13**, 12-20 (2002).
11. Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., & Fertil, B., *Mol. Biol. Evol.* **16** (10), 1391-1399 (1999).
12. Rousu, J., Saunders, C., Szedmak, S., & Shawe-Taylor, J., *Journal of Machine Learning Research* **7**, 1601-1626 (2006).
13. Cesa-Bianchi, N., Gentile, C., & Zaniboni, L., *Journal of Machine Learning Research* **7**, 31-54 (2006).
14. Mrazek, J., *Mol. Biol. Evol.* **26** (5), 1163-1169 (2009).
15. Mavromatis, K. *et al.*, *Nat. Methods* **4** (6), 495-500 (2007).
16. Pope, P.B. *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **107** (33), 14793-14798 (2010).
17. Turnbaugh, P.J. *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **107** (16), 7503-7508 (2010).
18. Hugenholtz, P., *Genome Biol* **3** (2), (2002).
19. Delcher, A.L., Phillippy, A., Carlton, J., & Salzberg, S.L., *Nucleic Acids Res* **30** (11), 2478-2483 (2002).
20. von Mering, C. *et al.*, *Nucleic Acids Res* **35** (Database issue), D358-362 (2007).
21. Larkin, M.A. *et al.*, *Bioinformatics* **23** (21), 2947-2948 (2007).
22. Li, W. & Godzik, A., *Bioinformatics* **22** (13), 1658-1659 (2006).