

Supplementary Material for “Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes”

Haitao Han^{1†}, Ziye Wang^{2†}, Shanfeng Zhu^{1*}

¹Institute of Science and Technology for Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, 200433, China.

²School of Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, China.

*Corresponding author(s). E-mail(s): zhuf@fudan.edu.cn;

†These authors contributed equally to this work.

Supplementary Note 1

High-performance binners reveals extensive microbial diversity

The microbial diversity of the MAGs recovered by the binners with the highest overall ranking scores was further analyzed. Firstly, the MQ MAGs recovered by MetaBAT 2, SemiBin 2, and the binners with the highest overall ranking scores were dereplicated to the species level for each data-binning combinations on the marine dataset (Supplementary Table 7).

Secondly, these species-level MAGs were classified using the Genome Taxonomy Database toolkit (GTDB-Tk, version 2.4.0) with the GTDB release r220. The results demonstrated that MAGs recovered from the binners with the highest overall ranking scores comprised the highest number of known families, genera, and species across the seven data-binning combinations (Supplementary Table 7). Additionally, these MAGs recovered from the binners with the highest overall ranking scores contained the highest number of unclassified species across the short_single, short_multi, long_single, long_multi, hybrid_single, and hybrid_multi data-binning combinations. This suggests a potential to expand our understanding of novel lineages.

Finally, all MAGs recovered by MetaBAT 2, SemiBin 2, and the binners with the highest overall ranking scores across the seven data-binning combinations were clustered into 5,282 species. These species were classified using GTDB-Tk and visualized in Supplementary Figs. 4 and 5. Supplementary Fig. 4 highlights five phyla that constitute more than 1% of the species within the bacterial domain. Among all the phyla, Pseudomonadota (52.61%), Bacteroidota (32.01%), Actinomycetota (3.31%) and Verrucomicrobiota (3.44%) were major components of the bacterial domain, while Thermoplasmata (75%) and Thermoproteota (10.71%) were major components of the archaeal domain.

Supplementary Table 1 Overview of short-read co-assemblies used for benchmarking.

	Human gut I	Human gut II	Cheese	Marine	Activated sludge
# contigs	133,003	535,433	66,255	1,203,821	6,297,919
N50	15,135	5,718	18,741	3,123	2,843
Total length	715,149,959	1,910,238,677	361,539,561	3,167,212,460	15,615,934,436

Supplementary Table 2 Overview of short-read single-sample assemblies used for benchmarking.

	Human gut I	Human gut II	Cheese	Marine	Activated sludge
# contigs (mean)	47,745	56,571	18,901	93,515	324,060
# contigs (max)	72,993	109,429	40,912	222,055	419,437
# contigs (min)	35,110	32,204	8,755	20,786	227,648
N50 (mean)	32,900	11,946	14,493	3,033	2,289
N50 (max)	43,440	20,881	24,128	3,738	3,958
N50 (min)	19,467	6,052	5,601	2,508	2,000
Total length (mean)	325,386,369	256,993,332	95,850,015	242,422,799	702,196,567
Total length (max)	413,344,356	401,812,298	185,792,967	579,027,752	900,384,805
Total length (min)	266,816,076	157,489,685	61,704,572	49,209,714	492,893,842

Supplementary Table 3 Overview of long-read single-sample assemblies used for benchmarking.

	Human gut I	Human gut II	Cheese	Marine	Activated sludge
# contigs (mean)	1,634	3,367	638	40,761	37,432
# contigs (max)	2,559	6,773	1,458	68,585	65,716
# contigs (min)	872	1,634	377	16,061	13,560
N50 (mean)	783,864	152,340	423,504	27,885	101,327
N50 (max)	1,113,454	265,218	1,442,587	38,185	191,848
N50 (min)	313,948	68,829	125,369	17,068	74,463
Total length (mean)	216,859,734	223,438,975	74,702,467	890,945,443	2,141,174,635
Total length (max)	274,513,079	455,827,593	157,884,716	1,463,531,463	3,566,969,444
Total length (min)	142,971,193	122,189,370	29,679,422	340,495,661	850,230,903

Supplementary Table 4 Overview of hybrid single-sample assemblies used for benchmarking.

	Human gut I	Human gut II	Cheese	Marine	Activated sludge
# contigs (mean)	37,103	34,584	14,076	53,379	179,580
# contigs (max)	59,218	66,221	33,023	134,030	219,349
# contigs (min)	24,192	16,450	6,729	11,787	122,382
N50 (mean)	81,858	47,665	57,450	31,420	29,783
N50 (max)	130,362	101,502	149,267	57,671	43,883
N50 (min)	46,811	16,605	9,518	8,411	16,782
Total length (mean)	343,885,590	299,011,137	104,509,978	396,047,279	1,664,070,909
Total length (max)	432,092,532	488,842,224	195,304,010	825,787,413	2,251,057,464
Total length (min)	284,087,294	177,322,146	67,076,014	120,920,638	989,372,543

Supplementary Table 5 Sample information of the marine, cheese, human gut I, human gut II and activated sludge datasets.

Dataset	Run accessions					
Marine (mNGS)	ERR9769234	ERR9769235	ERR9769236	ERR9769237	ERR9769238	ERR9769239
	ERR9769240	ERR9769241	ERR9769242	ERR9769243	ERR9769244	ERR9769245
	ERR9769246	ERR9769247	ERR9769248	ERR9769249	ERR9769250	ERR9769251
	ERR9769252	ERR9769253	ERR9769254	ERR9769255	ERR9769256	ERR9769257
Marine (HiFi)	ERR9769258	ERR9769259	ERR9769260	ERR9769261	ERR9769262	ERR9769263
	ERR9769275	ERR9769276	ERR9769277	ERR9769278	ERR9769279	ERR9769280
	ERR9769281	ERR9769282	ERR9769283	ERR9769284	ERR9769285	ERR9769286
	ERR9769287	ERR9769288	ERR9769289	ERR9769290	ERR9769291	ERR9769292
Cheese (mNGS)	ERR9769293	ERR9769294	ERR9769295	ERR9769296	ERR9769297	ERR9769298
	ERR9769299	ERR9769300	ERR9769301	ERR9769302	ERR9769303	ERR9769304
	SRR17227786	SRR17227787	SRR17227788	SRR17227790	SRR17227791	SRR17227792
	SRR17227793	SRR17227794	SRR17227795	SRR17227796	SRR17227797	SRR17227798
Cheese (HiFi)	SRR17227800	SRR17227801	SRR17227802			
	SRR18495437	SRR18495438	SRR18495439	SRR18495440	SRR18495441	SRR18495442
	SRR18495443	SRR18495444	SRR18495445	SRR18495446	SRR18495447	SRR18495448
	SRR18495449	SRR18495450	SRR18495451			
Human gut I (mNGS)	CRR344868	CRR344869	CRR344870			
Human gut I (HiFi)	CRR344871	CRR344872	CRR344873			
Human gut II (mNGS)	SRR18491179	SRR18491178	SRR18491210	SRR18490956	SRR18491125	SRR18491290
	SRR18491279	SRR18491268	SRR18491257	SRR18491246	SRR18491177	SRR18491110
	SRR18491099	SRR18491060	SRR18491049	SRR18491038	SRR18490999	SRR18490988
	SRR18491232	SRR18491221	SRR18491209	SRR18491198	SRR18491187	SRR18491089
Human gut II (Nanopore)	SRR18491078	SRR18491067	SRR18491168	SRR18491157	SRR18490978	SRR18490967
	SRR18491208	SRR18491207	SRR18491206	SRR18491205	SRR18491204	SRR18491203
	SRR18491202	SRR18491201	SRR18491200	SRR18491199	SRR18491197	SRR18491196
	SRR18491195	SRR18491194	SRR18491193	SRR18491192	SRR18491191	SRR18491190
Activated sludge (mNGS)	SRR18491189	SRR18491188	SRR18491186	SRR18491185	SRR18491184	SRR18491183
	SRR18491182	SRR18491181	SRR18491180	SRR18491092	SRR18491091	SRR18491090
	SRR11673989	SRR11673992	SRR11673995	SRR11673999	SRR11674002	SRR11674005
	SRR11674008	SRR11674012	SRR11674015	SRR11674018	SRR11674022	SRR11674025
Activated sludge (Nanopore)	SRR11674031	SRR11674035	SRR11674038	SRR11674041	SRR11674044	SRR11674048
	SRR11674051	SRR11674054	SRR11673976	SRR11674009	SRR11674045	
	SRR11673963	SRR11673964	SRR11673966	SRR11673967	SRR11673968	SRR11673969
	SRR11673970	SRR11673971	SRR11673972	SRR11673973	SRR11673974	SRR11673975
	SRR11673977	SRR11673978	SRR11673979	SRR11673980	SRR11673981	SRR11673982
	SRR11673983	SRR11673984	SRR11673985	SRR11673986	SRR11673988	
Dataset	Sample Description					
Marine	Surface seawater samples collected from the North Sea between March and May 2020 [1] (30 mNGS samples and 30 HiFi samples).					
Cheese	Rind samples of three washed-rind cheeses collected at five time points [2] (15 mNGS samples and 15 HiFi samples).					
Human gut I	Human fecal samples obtained from three healthy volunteers [3] (3 mNGS samples and 3 HiFi samples).					
Human gut II	Human fecal samples obtained from 30 healthy volunteers [4] (30 mNGS samples and 30 Nanopore samples).					
Activated sludge	Fresh activated sludge samples collected from 23 Danish wastewater treatment plants during August and September of 2018 [5] (23 mNGS samples and 23 Nanopore samples).					

For the human gut I, cheese, marine, and activated sludge datasets, we downloaded all the long-read samples along with their corresponding short-read samples from the original papers. In the human gut II dataset, considering the substantial computational resources and disk space required for multi-sample binning, we downloaded 30 Nanopore samples and their corresponding 30 mNGS samples.

Supplementary Table 6 Average runtimes and memory usage of each metagenomic binning tool across seven data-binning combinations on the activated sludge dataset.

	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
Time (minutes)							
CONCOCT	nan	300.48 (± 174.31)	1119.7 (± 338.85)	440.0 (± 333.87)	646.48 (± 393.75)	642.39 (± 256.76)	877.0 (± 387.41)
MaxBin 2	nan	631.52 (± 173.76)	1222.52 (± 366.62)	123.74 (± 68.56)	208.91 (± 113.45)	554.57 (± 148.06)	1042.26 (± 293.1)
MetaBAT 2	463	6.83 (± 1.4)	50.48 (± 11.51)	6.48 (± 4.53)	22.74 (± 9.89)	8.09 (± 1.93)	52.7 (± 12.64)
VAMB	1971	29.09 (± 5.41)	123.35 (± 21.29)	15.78 (± 6.26)	54.91 (± 18.19)	35.35 (± 10.61)	137.0 (± 22.61)
CLMB	13198	391.3 (± 90.91)	671.13 (± 127.76)	286.04 (± 110.23)	344.17 (± 123.49)	611.87 (± 115.33)	697.26 (± 133.17)
MetaDecoder	8668	18.61 (± 0.95)	97.91 (± 30.03)	28.61 (± 17.25)	145.17 (± 85.78)	81.48 (± 33.43)	211.48 (± 65.02)
Binny	6976	475.04 (± 82.43)	673.13 (± 95.32)	1040.96 (± 343.48)	1038.26 (± 341.22)	1454.61 (± 303.03)	1585.61 (± 266.49)
MetaBinner	nan	355.52 (± 85.84)	345.65 (± 84.88)	313.09 (± 129.39)	296.48 (± 124.95)	394.35 (± 84.84)	508.57 (± 110.06)
SemiBin2	681	80.7 (± 7.32)	131.04 (± 42.88)	172.0 (± 36.98)	191.57 (± 43.57)	143.87 (± 19.2)	164.17 (± 20.58)
COMEBin	nan	nan	nan	1380.83 (± 501.65)	1434.52 (± 546.69)	nan	nan
VAMB (GPU)	322	14.65 (± 2.18)	57.39 (± 11.72)	11.52 (± 4.29)	27.17 (± 8.87)	15.35 (± 2.28)	61.17 (± 13.08)
SemiBin2 (GPU)	nan	44.83 (± 5.2)	64.57 (± 4.66)	107.35 (± 28.59)	125.83 (± 31.13)	88.09 (± 13.79)	106.87 (± 13.98)
COMEBin (GPU)	nan	698.09 (± 131.9)	670.13 (± 101.4)	251.26 (± 82.01)	302.96 (± 88.4)	614.57 (± 99.46)	733.48 (± 113.0)
DAS_tool		20.87 (± 5.12)	29.39 (± 5.46)	60.52 (± 26.23)	76.04 (± 35.57)	59.96 (± 15.72)	58.74 (± 15.38)
MAGScoT		14.96 (± 2.74)	16.04 (± 2.93)	45.78 (± 16.64)	46.3 (± 16.76)	37.61 (± 6.81)	37.3 (± 6.79)
Metawrap		297.13 (± 54.13)	400.22 (± 61.93)	882.87 (± 316.52)	988.87 (± 355.36)	841.96 (± 164.53)	567.87 (± 91.69)
Memory (GB)							
CONCOCT	nan	2.81 (± 0.43)	3.25 (± 0.5)	2.15 (± 0.71)	2.15 (± 0.71)	2.19 (± 0.31)	2.53 (± 0.37)
MaxBin 2	nan	1.71 (± 0.33)	1.81 (± 0.35)	2.57 (± 0.61)	2.57 (± 0.63)	3.53 (± 0.53)	3.53 (± 0.52)
MetaBAT 2	45.18	1.87 (± 0.31)	1.89 (± 0.32)	3.79 (± 1.1)	3.61 (± 1.15)	3.17 (± 0.55)	3.25 (± 0.57)
VAMB	8.34	0.36 (± 0.02)	0.56 (± 0.07)	1.96 (± 0.58)	2.09 (± 0.63)	1.1 (± 0.19)	1.55 (± 0.29)
CLMB	30	2.82 (± 0.44)	2.83 (± 0.41)	3.8 (± 0.99)	3.8 (± 0.99)	4.09 (± 0.62)	4.1 (± 0.62)
MetaDecoder	94.05	1.45 (± 0.23)	14.54 (± 1.63)	3.23 (± 1.17)	3.58 (± 1.27)	3.95 (± 0.73)	4.53 (± 1.95)
Binny	123.2	22.36 (± 3.36)	22.66 (± 3.55)	10.66 (± 3.8)	10.67 (± 3.81)	14.41 (± 2.2)	14.77 (± 2.16)
MetaBinner	nan	3.05 (± 0.55)	3.13 (± 0.57)	2.74 (± 0.72)	2.8 (± 1.06)	3.51 (± 0.49)	2.52 (± 0.33)
SemiBin2	73.86	16.54 (± 5.04)	4.16 (± 0.53)	4.93 (± 0.8)	5.1 (± 0.86)	39.46 (± 21.55)	6.32 (± 1.2)
COMEBin	nan	nan	nan	5.65 (± 0.58)	5.71 (± 0.56)	nan	nan
VAMB (GPU)	9.82	1.77 (± 0.03)	2.02 (± 0.07)	3.41 (± 0.58)	3.55 (± 0.63)	2.56 (± 0.2)	3.05 (± 0.29)
SemiBin2 (GPU)	nan	18.04 (± 5.1)	5.45 (± 0.54)	6.17 (± 0.8)	6.34 (± 0.86)	37.02 (± 11.7)	7.62 (± 1.21)
COMEBin (GPU)	nan	12.09 (± 1.79)	10.52 (± 1.58)	4.51 (± 1.38)	4.52 (± 1.37)	7.98 (± 1.21)	6.98 (± 0.99)
DAS_tool		0.84 (± 0.12)	0.84 (± 0.11)	1.53 (± 0.49)	1.53 (± 0.49)	1.33 (± 0.22)	1.33 (± 0.22)
MAGScoT		1.12 (± 0.16)	1.14 (± 0.16)	2.14 (± 0.71)	2.14 (± 0.71)	1.85 (± 0.32)	1.85 (± 0.32)
Metawrap		34.92 (± 0.13)	34.94 (± 0.11)	34.98 (± 0.08)	35.01 (± 0.07)	35.03 (± 0.05)	35.0 (± 0.07)

In the short_co data-binning combination, “nan” indicates that the corresponding binner failed to complete execution within two weeks based on our computational resources. In the other six data-binning combinations, “nan” signifies that the binner did not finish execution within one week based on our computational resources.

Supplementary Table 7 The number of MQ bins, species-level bins, identified lineage and unclassified species for the SemiBin 2, MetaBAT 2, and the binners with the highest overall ranking scores across seven data-binning combinations on the marine dataset.

	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
binner with the highest overall ranking score							
MQ bin	260	1153	1647	1515	2232	1844	2041
Species-level bin	260	542	866	1048	1706	648	737
Domain	2	2	2	2	2	2	2
Phylum	14	13	13	12	12	14	15
Class	22	18	19	17	17	20	24
Order	63	44	49	53	52	59	66
Family	98	70	77	82	82	92	98
Genus	162	109	127	139	148	153	165
Species	230	180	236	210	253	245	262
Unclassified species	30	47	39	433	394	105	128
SemiBin 2							
MQ bin	236	944	1105	869	1355	1408	1610
Species-level bin	236	328	384	439	730	508	607
Domain	2	2	2	2	2	2	2
Phylum	13	10	9	9	10	14	14
Class	21	14	13	15	16	20	20
Order	61	37	38	45	50	61	61
Family	93	62	64	69	78	90	94
Genus	146	96	101	124	136	151	155
Species	191	159	174	188	243	223	239
Unclassified species	45	14	16	43	59	84	106
MetaBAT 2							
MQ bin	235	682	1124	647	1313	1574	1896
Species-level bin	235	241	352	318	546	551	615
Domain	2	2	2	2	2	2	2
Phylum	15	9	9	8	10	13	14
Class	23	13	13	12	16	20	21
Order	64	35	35	30	47	56	62
Family	95	56	60	48	68	88	96
Genus	157	86	97	88	119	149	160
Species	208	134	165	132	214	227	255
Unclassified species	27	7	10	10	39	46	71

COMEBin is the binner with the highest overall ranking score for the hybrid_multi, hybrid_single, short_multi, and short_single data-binning combinations. MetaBinner is the binner with the highest overall ranking score for the long_multi and long_single data-binning combinations. Binny is the binner with the highest overall ranking score for the short_co data-binning combination.

Supplementary Table 8 MQ MAGs per minute and MQ MAGs per GB of memory for metagenomic binning tools across seven data-binning combinations in the activated sludge dataset.

	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
MQ MAGs per minute							
CONCOCT	nan	0.13 (± 0.04)	0.07 (± 0.01)	0.29 (± 0.38)	0.21 (± 0.02)	0.09 (± 0.06)	0.09 (± 0.06)
MaxBin 2	nan	0.05 (± 0.02)	0.05 (± 0.01)	0.63 (± 0.33)	0.54 (± 0.14)	0.11 (± 0.04)	0.08 (± 0.02)
MetaBAT 2	4.61	5.95 (± 1.54)	1.19 (± 0.27)	18.03 (± 5.7)	5.86 (± 2.76)	12.02 (± 3.32)	2.14 (± 1.09)
VAMB	0.67	0.61 (± 0.24)	0.52 (± 0.07)	5.72 (± 1.12)	1.9 (± 0.51)	2.9 (± 0.76)	0.67 (± 0.13)
CLMB	0.1	0.05 (± 0.02)	0.07 (± 0.01)	0.31 (± 0.06)	0.3 (± 0.05)	0.16 (± 0.04)	0.14 (± 0.04)
MetaDecoder	0.13	2.92 (± 0.78)	0.59 (± 0.14)	5.13 (± 1.3)	0.94 (± 0.34)	1.66 (± 0.54)	0.55 (± 0.2)
Binny	0.09	0.09 (± 0.02)	0.12 (± 0.02)	0.05 (± 0.02)	0.06 (± 0.02)	0.06 (± 0.02)	0.06 (± 0.01)
MetaBinner	nan	0.13 (± 0.04)	0.28 (± 0.04)	0.31 (± 0.06)	0.33 (± 0.05)	0.24 (± 0.07)	0.21 (± 0.05)
SemiBin2	1.9	0.64 (± 0.14)	0.48 (± 0.13)	0.71 (± 0.11)	0.67 (± 0.12)	0.59 (± 0.12)	0.55 (± 0.13)
COMEBin	nan	nan	nan	0.09 (± 0.02)	0.09 (± 0.03)	nan	nan
VAMB (GPU)	3.94	1.21 (± 0.49)	1.13 (± 0.22)	5.72 (± 1.12)	1.9 (± 0.51)	6.41 (± 1.3)	1.58 (± 0.57)
SemiBin2 (GPU)	nan	1.15 (± 0.22)	0.9 (± 0.16)	1.15 (± 0.11)	1.01 (± 0.14)	0.96 (± 0.22)	0.85 (± 0.22)
COMEBin (GPU)	nan	0.11 (± 0.02)	0.15 (± 0.01)	0.45 (± 0.08)	0.39 (± 0.07)	0.19 (± 0.04)	0.18 (± 0.04)
MQ MAGs per GB of memory							
CONCOCT	nan	11.29 (± 2.13)	22.44 (± 3.18)	25.14 (± 8.43)	40.6 (± 5.36)	20.23 (± 4.56)	23.04 (± 4.68)
MaxBin 2	nan	17.87 (± 4.7)	34.26 (± 4.01)	24.34 (± 3.9)	39.13 (± 6.58)	15.75 (± 3.57)	23.33 (± 5.64)
MetaBAT	28.35	21.35 (± 4.43)	30.91 (± 3.36)	27.7 (± 10.57)	33.78 (± 12.38)	30.09 (± 7.8)	32.02 (± 8.47)
VAMB	151.86	47.82 (± 16.71)	113.00 (± 13.39)	44.13 (± 7.45)	48.38 (± 6.03)	89.02 (± 13.06)	61.29 (± 19.34)
CLMB	42.33	6.43 (± 2.14)	17.54 (± 2.98)	22.58 (± 5.56)	26.64 (± 6.3)	24.17 (± 5.49)	23.12 (± 6.51)
MetaDecoder	12.16	54.58 (± 6.69)	52.28 (± 12.92)	40.87 (± 5.43)	32.34 (± 5.12)	31.21 (± 5.19)	25.93 (± 6.81)
Binny	5.24	2.00 (± 0.55)	3.71 (± 0.44)	5.24 (± 2.27)	6.10 (± 1.92)	6.13 (± 1.62)	6.84 (± 1.38)
MetaBinner	nan	14.86 (± 4.13)	29.81 (± 3)	33.94 (± 6.72)	34.83 (± 6.48)	26.16 (± 6.55)	40.74 (± 8.95)
SemiBin2	17.26	3.26 (± 0.69)	14.06 (± 2.07)	24.65 (± 4.09)	24.67 (± 4.25)	2.57 (± 1.35)	14.61 (± 4.1)
COMEBin	nan	nan	nan	19.47 (± 4.42)	20.17 (± 4.69)	nan	nan
VAMB (GPU)	129.12	9.8 (± 3.37)	31.56 (± 5.89)	44.13 (± 7.45)	48.38 (± 6.03)	37.92 (± 6.01)	30.2 (± 6.88)
SemiBin2 (GPU)	nan	2.96 (± 0.58)	10.73 (± 1.8)	19.68 (± 3.79)	19.81 (± 3.93)	2.48 (± 1.19)	11.97 (± 3.06)
COMEBin (GPU)	nan	6.47 (± 6.47)	9.8 (± 9.8)	25.1 (± 25.1)	26.08 (± 26.08)	14.86 (± 14.86)	18.47 (± 18.47)

In the short_co data-binning combination, “nan” indicates that the corresponding binner failed to complete execution within two weeks based on our computational resources. In the other six data-binning combinations, “nan” signifies that the binner did not finish execution within one week based on our computational resources.

Supplementary Table 9 NC MAGs per minute and NC MAGs per GB of memory for metagenomic binning tools across seven data-binning combinations in the activated sludge dataset.

	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
NC MAGs per minute							
CONCOCT	nan	0.03 (± 0.01)	0.01 (± 0.00)	0.07 (± 0.11)	0.04 (± 0.06)	0.02 (± 0.01)	0.02 (± 0.02)
MaxBin 2	nan	0.01 (± 0.01)	0.01 (± 0.00)	0.13 (± 0.13)	0.11 (± 0.05)	0.02 (± 0.01)	0.02 (± 0.01)
MetaBAT 2	0.51	1.66 (± 0.94)	0.31 (± 0.12)	3.83 (± 2.43)	1.15 (± 0.65)	3.38 (± 1.5)	0.45 (± 0.18)
VAMB	0.13	0.22 (± 0.15)	0.13 (± 0.04)	1.08 (± 0.5)	0.29 (± 0.09)	0.77 (± 0.34)	0.14 (± 0.05)
CLMB	0.02	0.02 (± 0.01)	0.02 (± 0.01)	0.06 (± 0.03)	0.05 (± 0.01)	0.04 (± 0.02)	0.03 (± 0.02)
MetaDecoder	0.02	0.62 (± 0.3)	0.14 (± 0.07)	0.89 (± 0.59)	0.17 (± 0.07)	0.39 (± 0.2)	0.11 (± 0.04)
Binny	0.02	0.03 (± 0.01)	0.03 (± 0.01)	0.02 (± 0.01)	0.02 (± 0.01)	0.02 (± 0.01)	0.02 (± 0.00)
MetaBinner	nan	0.02 (± 0.01)	0.05 (± 0.02)	0.07 (± 0.03)	0.07 (± 0.02)	0.05 (± 0.03)	0.04 (± 0.02)
SemiBin2	0.33	0.16 (± 0.07)	0.12 (± 0.05)	0.13 (± 0.05)	0.13 (± 0.06)	0.15 (± 0.06)	0.09 (± 0.03)
COMEBin	nan	nan	nan	0.02 (± 0.01)	0.02 (± 0.01)	nan	nan
VAMB (GPU)	0.81	0.44 (± 0.31)	0.29 (± 0.1)	1.44 (± 0.6)	0.62 (± 0.27)	1.68 (± 0.62)	0.31 (± 0.12)
SemiBin2 (GPU)	nan	0.28 (± 0.13)	0.23 (± 0.07)	0.22 (± 0.09)	0.2 (± 0.09)	0.25 (± 0.11)	0.14 (± 0.05)
COMEBin (GPU)	nan	0.02 (± 0.01)	0.03 (± 0.01)	0.09 (± 0.04)	0.08 (± 0.03)	0.05 (± 0.02)	0.05 (± 0.02)
NC MAGs per GB of memory							
CONCOCT	nan	2.4 (± 1.31)	3.87 (± 1.3)	5.93 (± 3.96)	6.1 (± 3.29)	4.2 (± 2.25)	3.82 (± 1.64)
MaxBin 2	nan	3.13 (± 2.00)	6.98 (± 2.38)	4.91 (± 1.88)	8.07 (± 2.73)	2.43 (± 0.99)	4.44 (± 1.47)
MetaBAT	5.2	5.89 (± 2.89)	8.13 (± 2.77)	5.81 (± 3.82)	6.94 (± 3.81)	8.42 (± 3.61)	7.08 (± 3.49)
VAMB	31.34	16.96 (± 10.27)	29.06 (± 8.71)	8.17 (± 3.3)	7.73 (± 2.88)	23.05 (± 6.18)	12.55 (± 6.36)
CLMB	8.93	2.27 (± 1.7)	5.72 (± 2.15)	3.89 (± 1.64)	4.09 (± 1.22)	6.49 (± 2.64)	4.98 (± 2.53)
MetaDecoder	2.13	11.5 (± 5.09)	11.83 (± 4.84)	6.82 (± 2.5)	5.67 (± 1.31)	7.3 (± 2.22)	4.98 (± 1.88)
Binny	1.26	0.55 (± 0.28)	0.83 (± 0.28)	1.7 (± 0.99)	2.15 (± 1.03)	1.99 (± 0.8)	2.1 (± 0.59)
MetaBinner	nan	2.5 (± 1.79)	5.13 (± 1.97)	7.06 (± 2.34)	6.92 (± 1.69)	5.27 (± 1.73)	7.57 (± 3.38)
SemiBin2	3.06	0.82 (± 0.4)	3.55 (± 1.11)	4.64 (± 1.9)	4.71 (± 1.85)	0.7 (± 0.56)	2.46 (± 0.95)
COMEBin	nan	nan	nan	3.98 (± 1.14)	4.19 (± 1.03)	nan	nan
VAMB (GPU)	26.68	3.5 (± 2.36)	8.05 (± 2.39)	4.5 (± 1.6)	4.33 (± 1.26)	9.8 (± 2.62)	6.1 (± 2.42)
SemiBin2 (GPU)	nan	0.74 (± 0.35)	2.7 (± 0.84)	3.69 (± 1.49)	3.76 (± 1.43)	0.67 (± 0.51)	2.01 (± 0.72)
COMEBin (GPU)	nan	1.32 (± 0.65)	1.92 (± 0.72)	5.27 (± 2.37)	5.55 (± 2.09)	4.12 (± 1.58)	4.89 (± 1.79)

In the short_co data-binning combination, “nan” indicates that the corresponding binner failed to complete execution within two weeks based on our computational resources. In the other six data-binning combinations, “nan” signifies that the binner did not finish execution within one week based on our computational resources.

Supplementary Table 10 HQ MAGs per minute and HQ MAGs per GB of memory for metagenomic binning tools across seven data-binning combinations in the activated sludge dataset.

	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
HQ MAGs per minute							
CONCOCT	nan	0.01 (± 0.00)	0.00 (± 0.00)	0.06 (± 0.09)	0.03 (± 0.05)	0.01 (± 0.01)	0.01 (± 0.01)
MaxBin 2	nan	0.00 (± 0.00)	0.00 (± 0.00)	0.12 (± 0.12)	0.10 (± 0.05)	0.01 (± 0.01)	0.01 (± 0.01)
MetaBAT 2	0.08	0.34 (± 0.3)	0.06 (± 0.03)	18.03 (± 5.7)	0.91 (± 0.53)	2.45 (± 1.09)	0.32 (± 0.13)
VAMB	0.01	0.06 (± 0.06)	0.02 (± 0.01)	0.86 (± 0.35)	0.26 (± 0.08)	0.58 (± 0.24)	0.09 (± 0.03)
CLMB	0.00	0.00 (± 0.00)	0.00 (± 0.00)	0.05 (± 0.02)	0.04 (± 0.01)	0.03 (± 0.02)	0.02 (± 0.01)
MetaDecoder	0.00	0.10 (± 0.08)	0.02 (± 0.02)	0.73 (± 0.47)	0.14 (± 0.05)	0.28 (± 0.14)	0.08 (± 0.03)
Binny	0.00	0.01 (± 0.00)	0.01 (± 0.00)	0.01 (± 0.01)	0.02 (± 0.01)	0.02 (± 0.01)	0.01 (± 0.00)
MetaBinner	nan	0.00 (± 0.00)	0.01 (± 0.00)	0.06 (± 0.03)	0.06 (± 0.02)	0.03 (± 0.02)	0.03 (± 0.01)
SemiBin2	0.04	0.03 (± 0.02)	0.02 (± 0.01)	0.10 (± 0.03)	0.09 (± 0.03)	0.18 (± 0.08)	0.07 (± 0.02)
COMEBin	nan	nan	nan	0.01 (± 0.01)	0.01 (± 0.01)	nan	nan
VAMB (GPU)	0.07	0.12 (± 0.11)	0.05 (± 0.03)	0.86 (± 0.35)	0.54 (± 0.23)	1.27 (± 0.45)	0.21 (± 0.07)
SemiBin2 (GPU)	nan	0.05 (± 0.04)	0.04 (± 0.02)	0.16 (± 0.05)	0.14 (± 0.05)	0.18 (± 0.08)	0.10 (± 0.03)
COMEBin (GPU)	nan	0.00 (± 0.00)	0.01 (± 0.00)	0.07 (± 0.03)	0.07 (± 0.02)	0.04 (± 0.01)	0.03 (± 0.01)
HQ MAGs per GB of memory							
CONCOCT	nan	0.56 (± 0.46)	0.47 (± 0.3)	5.19 (± 3.42)	5.23 (± 2.88)	3.22 (± 1.91)	2.9 (± 1.17)
MaxBin 2	nan	1.14 (± 0.84)	2.23 (± 0.87)	4.42 (± 1.69)	7.08 (± 2.37)	1.89 (± 0.86)	3.36 (± 1.23)
MetaBAT	0.49	1.17 (± 0.97)	1.49 (± 0.83)	4.46 (± 2.41)	5.41 (± 2.56)	6.09 (± 2.57)	5.09 (± 2.21)
VAMB	2.87	4.63 (± 4.09)	4.75 (± 2.84)	6.59 (± 2.48)	6.76 (± 2.7)	17.43 (± 4.61)	8.51 (± 3.81)
CLMB	0.9	0.55 (± 0.53)	0.91 (± 0.54)	3.25 (± 1.4)	3.61 (± 1.08)	4.85 (± 1.85)	3.33 (± 1.53)
MetaDecoder	0.13	1.85 (± 1.62)	1.63 (± 1.06)	5.56 (± 1.97)	4.66 (± 1.14)	5.32 (± 1.76)	3.59 (± 1.41)
Binny	0.18	0.15 (± 0.09)	0.18 (± 0.09)	1.27 (± 0.62)	1.62 (± 0.73)	1.54 (± 0.66)	1.6 (± 0.43)
MetaBinner	nan	0.39 (± 0.39)	0.78 (± 0.39)	6.14 (± 1.93)	6.01 (± 1.64)	3.8 (± 1.36)	5.23 (± 1.96)
SemiBin2	0.33	0.15 (± 0.11)	0.67 (± 0.34)	3.33 (± 1.02)	3.39 (± 1.05)	0.49 (± 0.37)	1.74 (± 0.57)
COMEBin	nan	nan	nan	3.07 (± 0.95)	3.34 (± 1.04)	nan	nan
VAMB (GPU)	2.44	0.95 (± 0.85)	1.33 (± 0.83)	6.59 (± 2.48)	3.78 (± 1.15)	7.46 (± 2.12)	4.15 (± 1.43)
SemiBin2 (GPU)	nan	0.14 (± 0.1)	0.51 (± 0.26)	2.65 (± 0.8)	2.71 (± 0.84)	0.49 (± 0.37)	1.43 (± 0.44)
COMEBin (GPU)	nan	0.19 (± 0.17)	0.38 (± 0.21)	4.05 (± 1.46)	4.4 (± 1.42)	2.94 (± 1.17)	3.38 (± 1.33)

In the short_co data-binning combination, “nan” indicates that the corresponding binner failed to complete execution within two weeks based on our computational resources. In the other six data-binning combinations, “nan” signifies that the binner did not finish execution within one week based on our computational resources.

Supplementary Table 11 Overall ranking scores and mean of overall ranking scores for seven data-binning combinations based on five real datasets. A lower average of ranks for a algorithm signifies its superior performance.

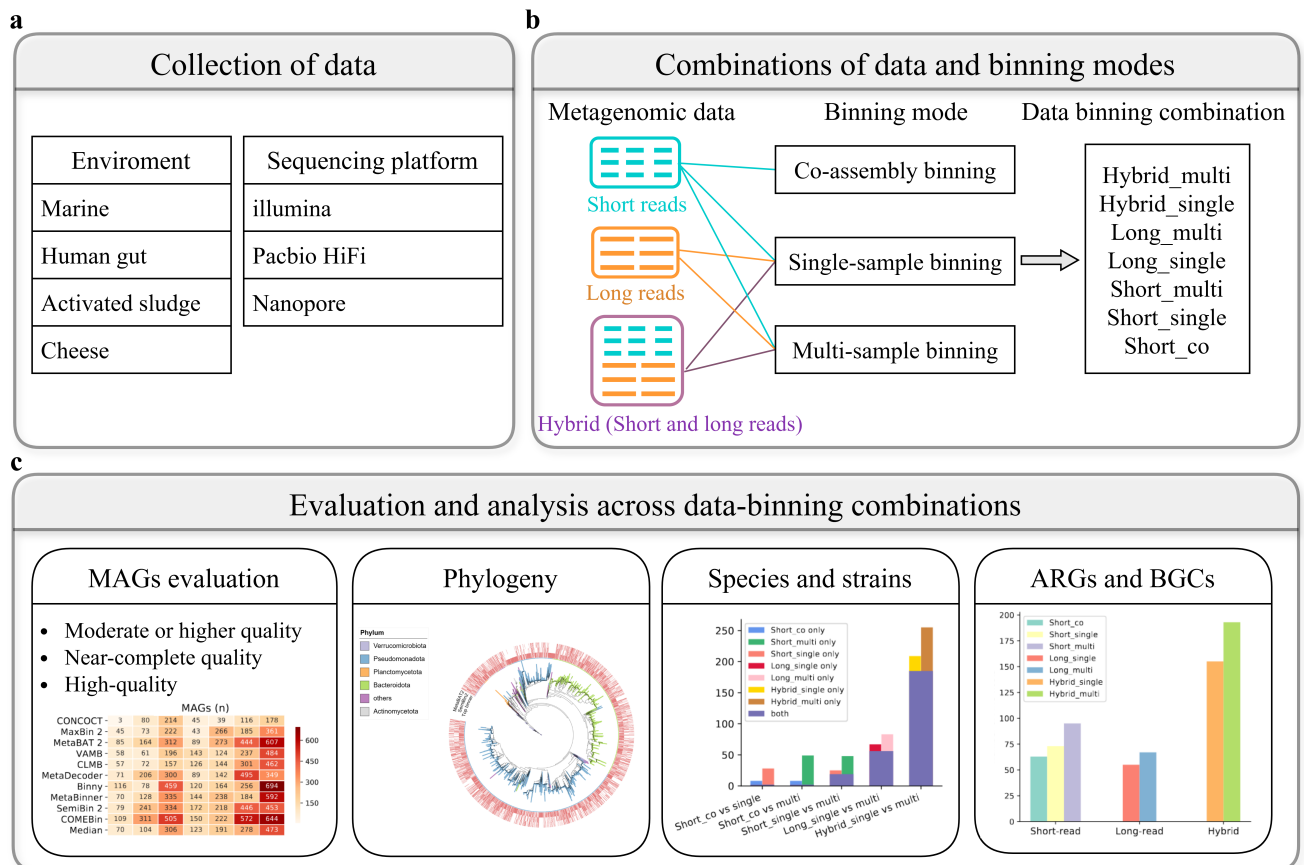
	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi	Mean
MAGScoT	1.63	1.72	1.85	1.92	2.06	1.86	1.77	1.83
DASTool	2.33	2.52	2.34	2.15	2.03	2.62	2.59	2.37
Metawrap	2.04	1.76	1.81	1.94	1.91	1.52	1.64	1.80

Supplementary Table 12 The number of NC MAGs and corresponding unique species and strain for each data-binning combination.

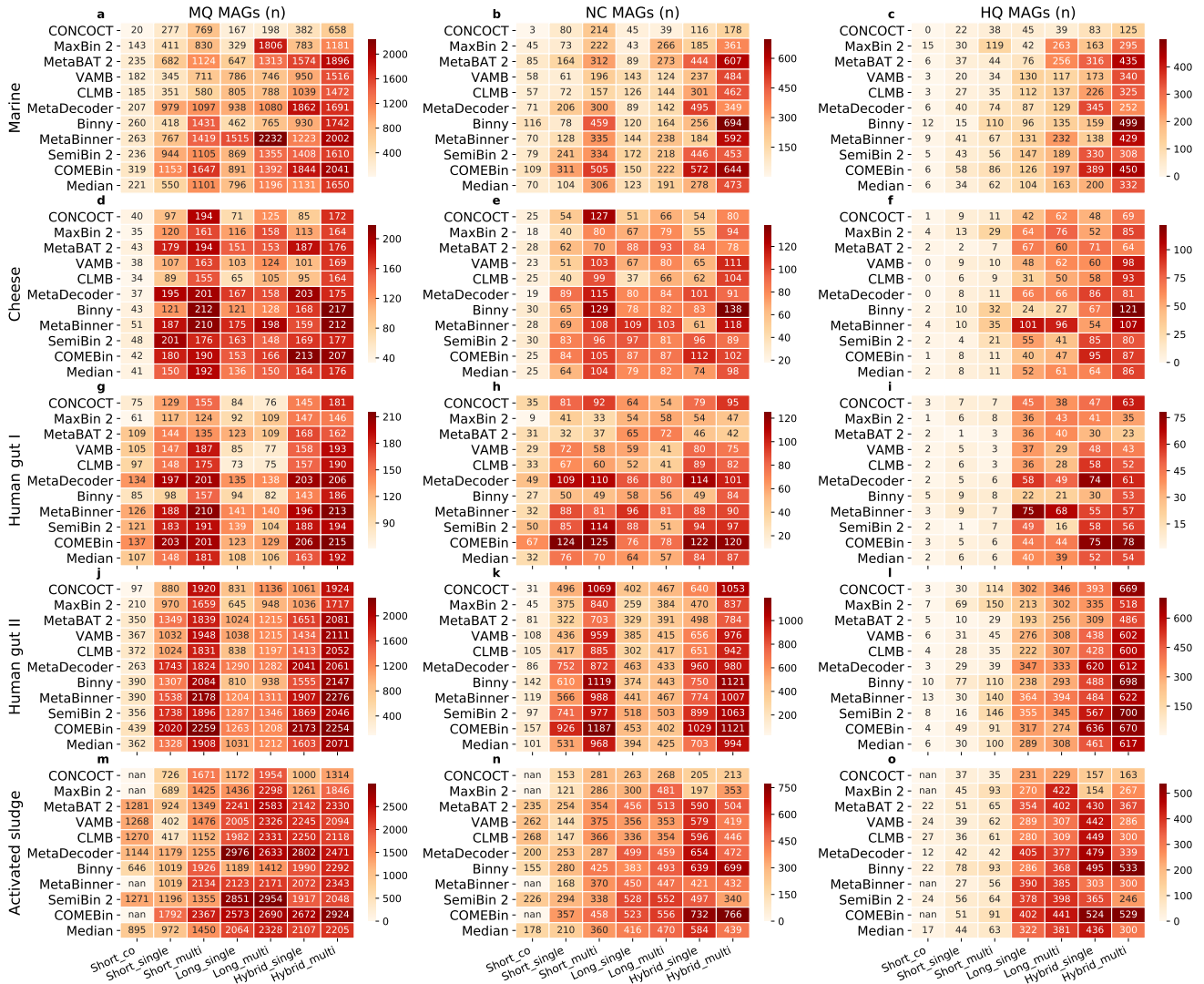
	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
#NC MAG	111	335	513	208	310	634	767
#Species	111	96	135	70	86	136	157
#Strain	111	152	229	70	87	297	374

Supplementary Table 13 The number of HQ MAGs and corresponding unique species and strain for each data-binning combination.

	Short_co	Short_single	Short_multi	Long_single	Long_multi	Hybrid_single	Hybrid_multi
#HQ MAG	8	64	115	187	282	447	552
#Species	8	21	30	67	82	109	115
#Strain	8	28	48	67	83	208	257



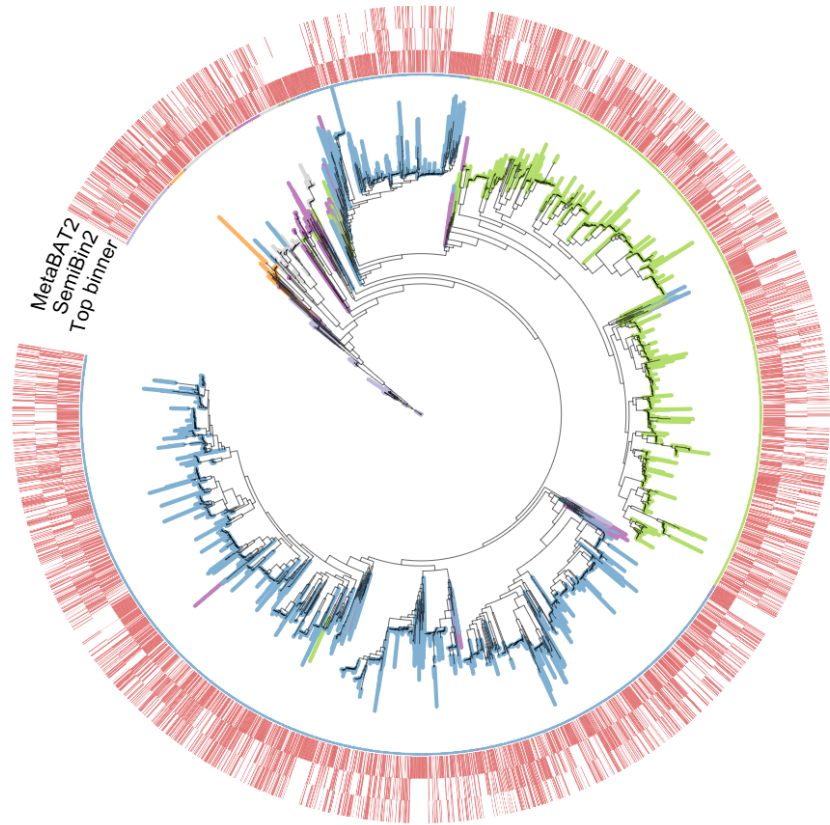
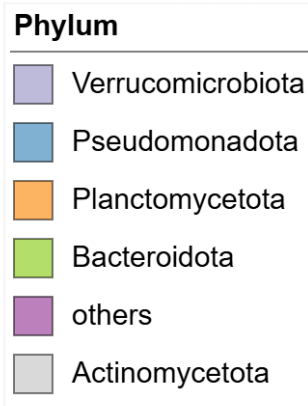
Supplementary Figure 1 Overview of the benchmarking workflow. **a**, A total of 202 sequencing samples from five datasets, covering four environments and three sequencing platforms, were used in this study. **b**, Seven data-binning combinations were introduced by integrating short-read, long-read, and hybrid data with co-assembly, single-sample, and multi-sample binning, respectively. **c**, Benchmarking of 13 metagenomic binning tools across seven data-binning combinations, followed by further analysis of the MAGs recovered from these combinations.



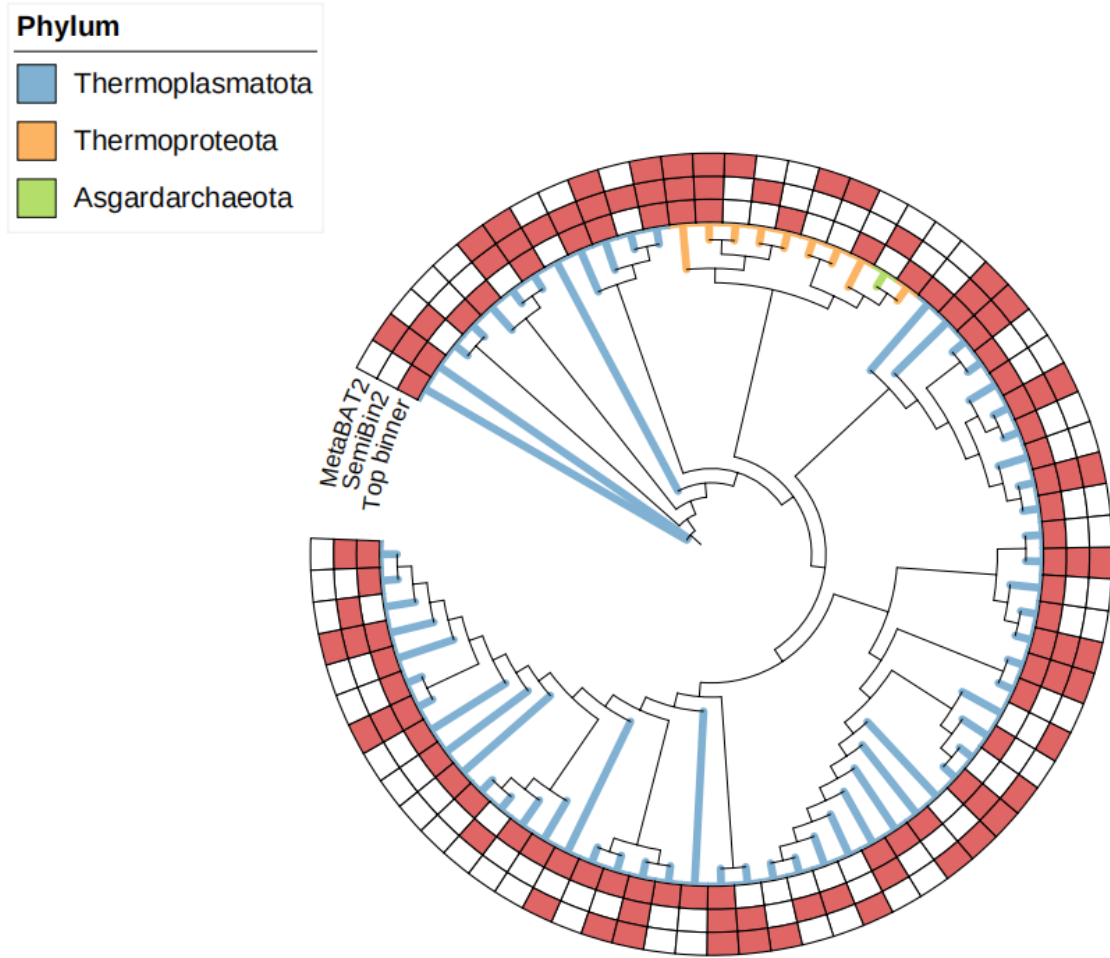
Supplementary Figure 2 Number of MQ, NC and HQ MAGs recovered from the real datasets. **a, d, g, j, m** The number of MQ MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets, respectively. **b, e, h, k, n** The number of NC MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets, respectively. **c, f, i, l, o** The number of HQ MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets, respectively. “nan” denotes that the corresponding binner failed to complete execution within two weeks based on the computational resources we used.



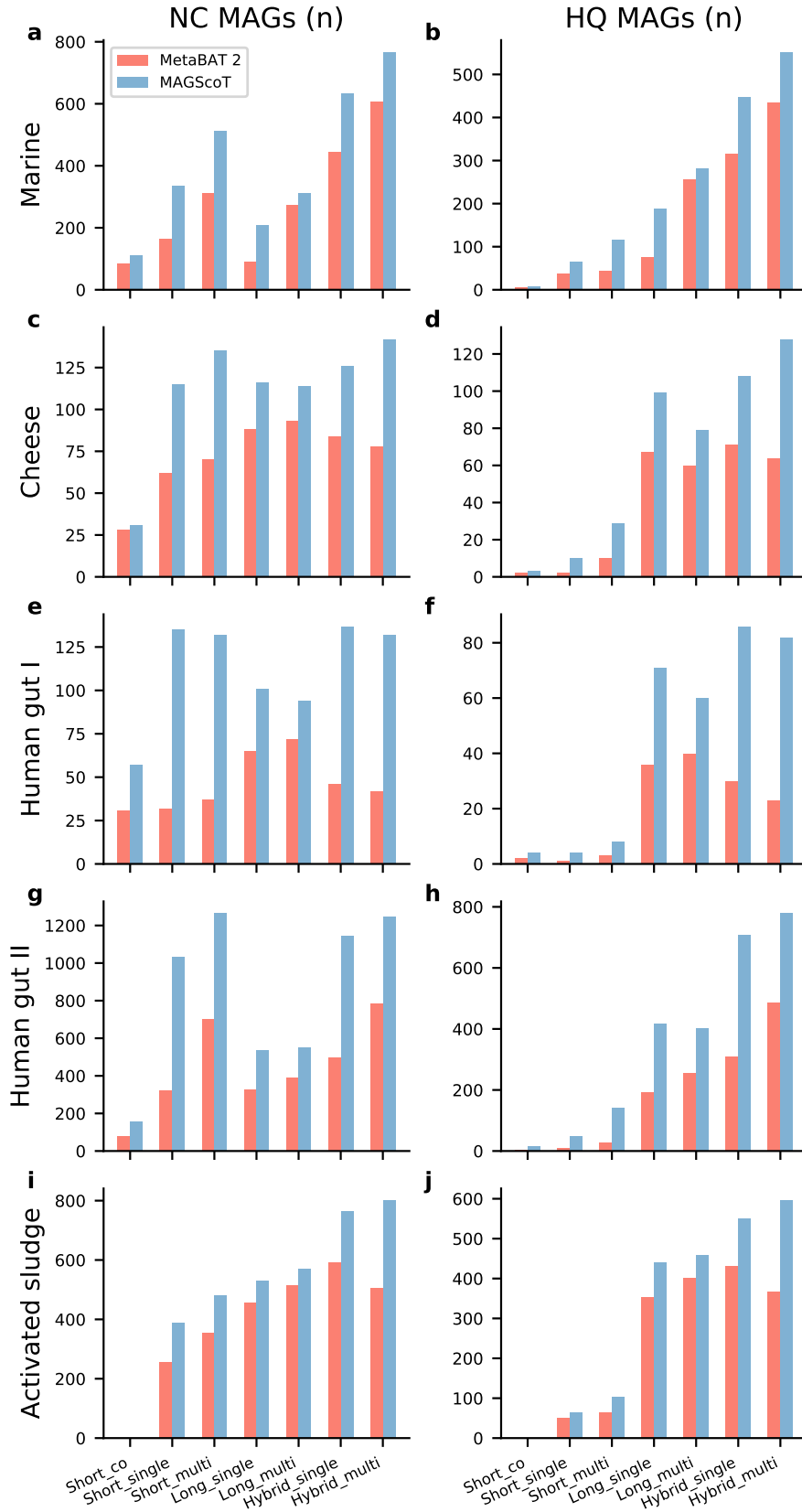
Supplementary Figure 3 Ranking scores (see Methods) of ten binners for each dataset across seven data-binning combinations. “nan” denotes that CONCOCT, MaxBin 2, MetaBinner, and COMEBin (GPU) cannot complete execution within two weeks on the activated sludge short-read co-assembly data (see Methods for computational resources). Consequently, these binners are ranked last on the activated sludge short-read co-assembly data.



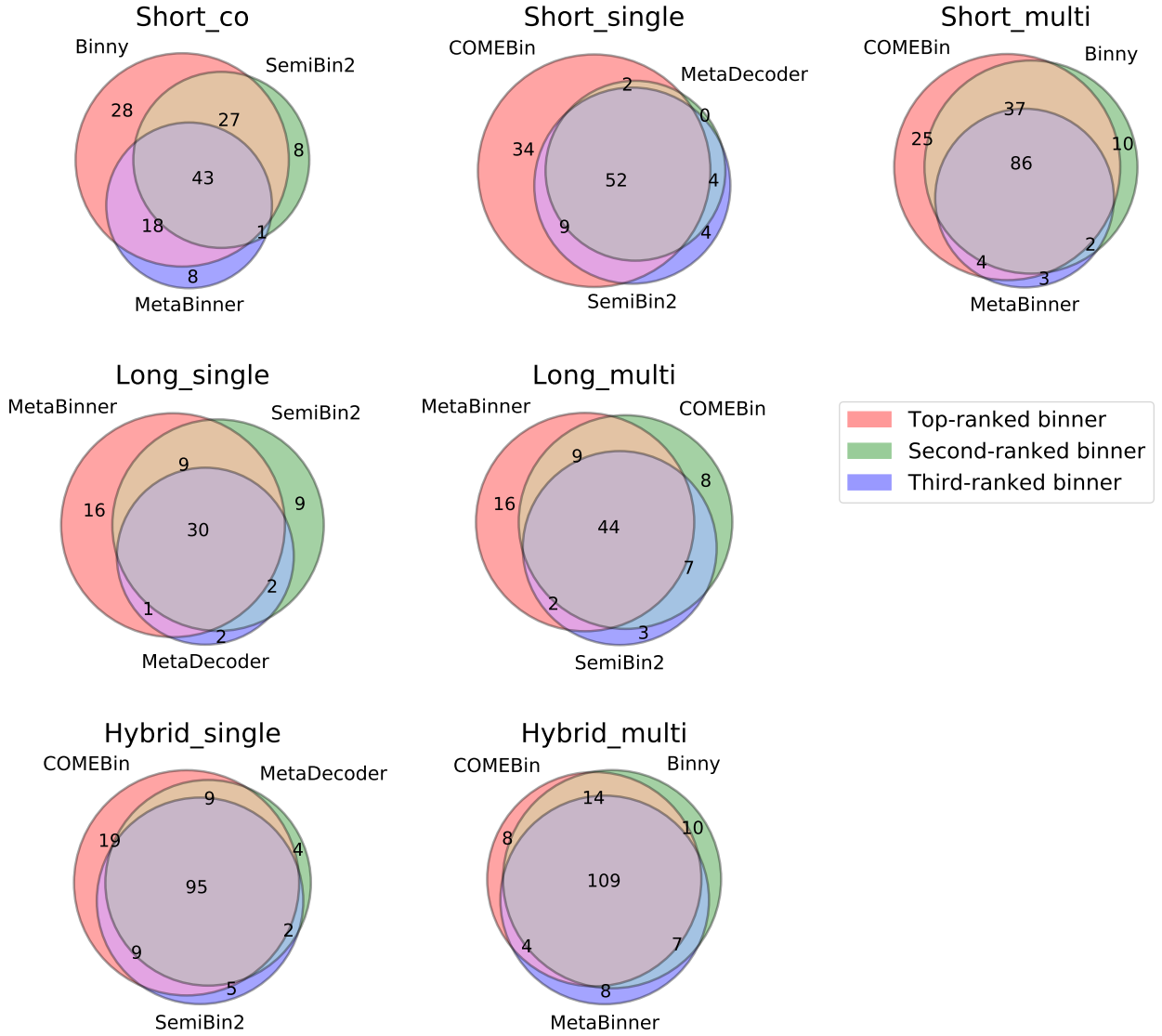
Supplementary Figure 4 Distribution of bacterial phyla for MAGs recovered by MetaBAT 2, SemiBin 2, and the binners with the highest overall ranking scores across seven data-binning combinations on the marine dataset. The “Top binner” refers to the binners with the highest overall ranking scores. Specifically, it denotes COMEBin for hybrid_multi, hybrid_single, short_multi, and short_single data-binning combinations; MetaBinner for long_multi and long_single data-binning combinations; and Binny for short_co data-binning combination. This figure was drawn using iTOL v7 [6].



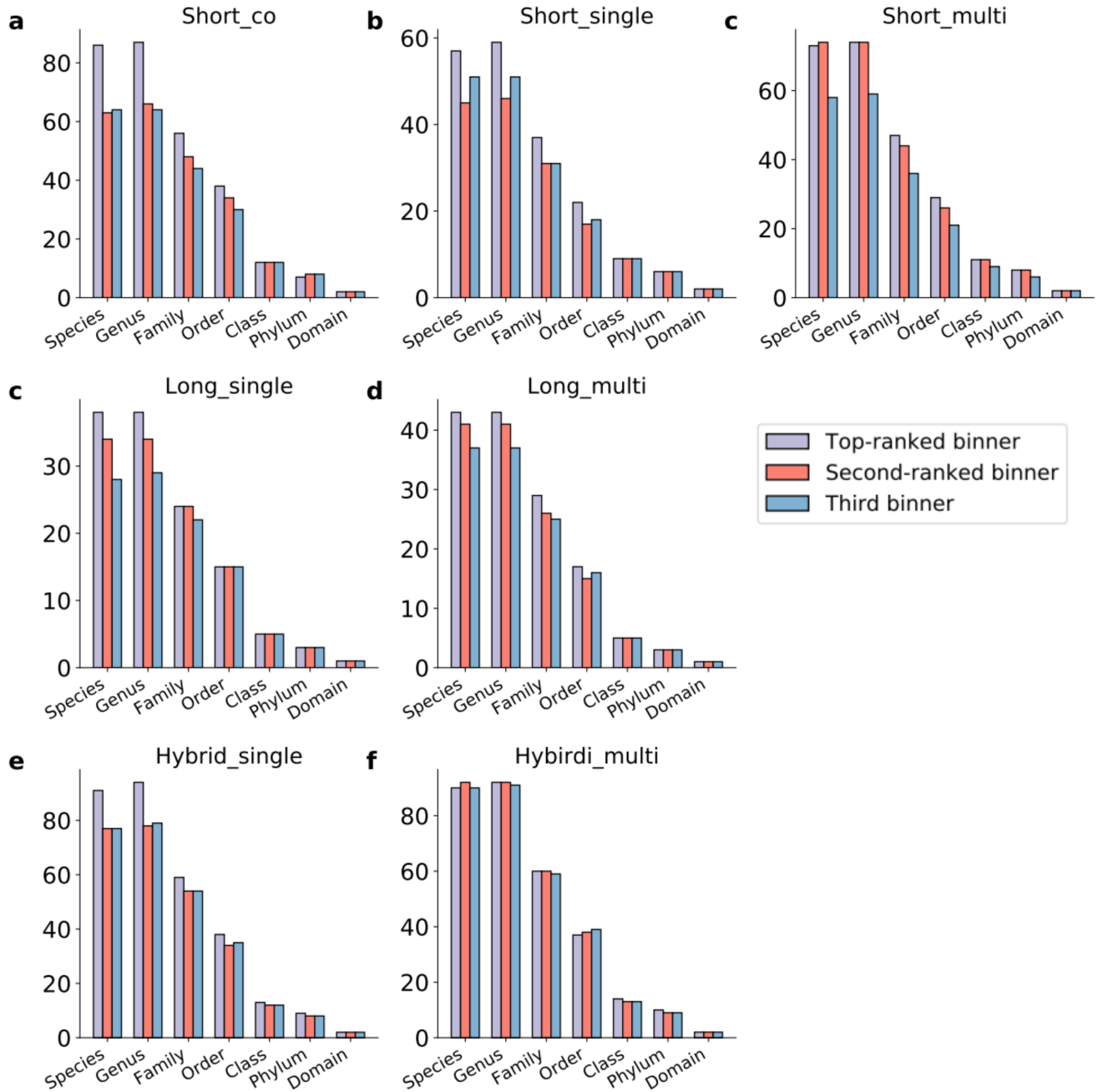
Supplementary Figure 5 Distribution of archaeal phyla for MAGs recovered by MetaBAT 2, SemiBin 2, and the binners with the highest overall ranking scores across seven data-binning combinations on the marine dataset. The “Top binner” in the figure refers to the binners with the highest overall ranking scores. Specifically, COMEBin is the binner with the highest overall ranking score for the hybrid_multi, hybrid_single, short_multi, and short_single data-binning combinations. MetaBinner is the binner with the highest overall ranking score for the long_multi and long_single data-binning combinations. Binny is the binner with the highest overall ranking score for the short_co data-binning combination. This figure was drawn using iTOL v7 [6].



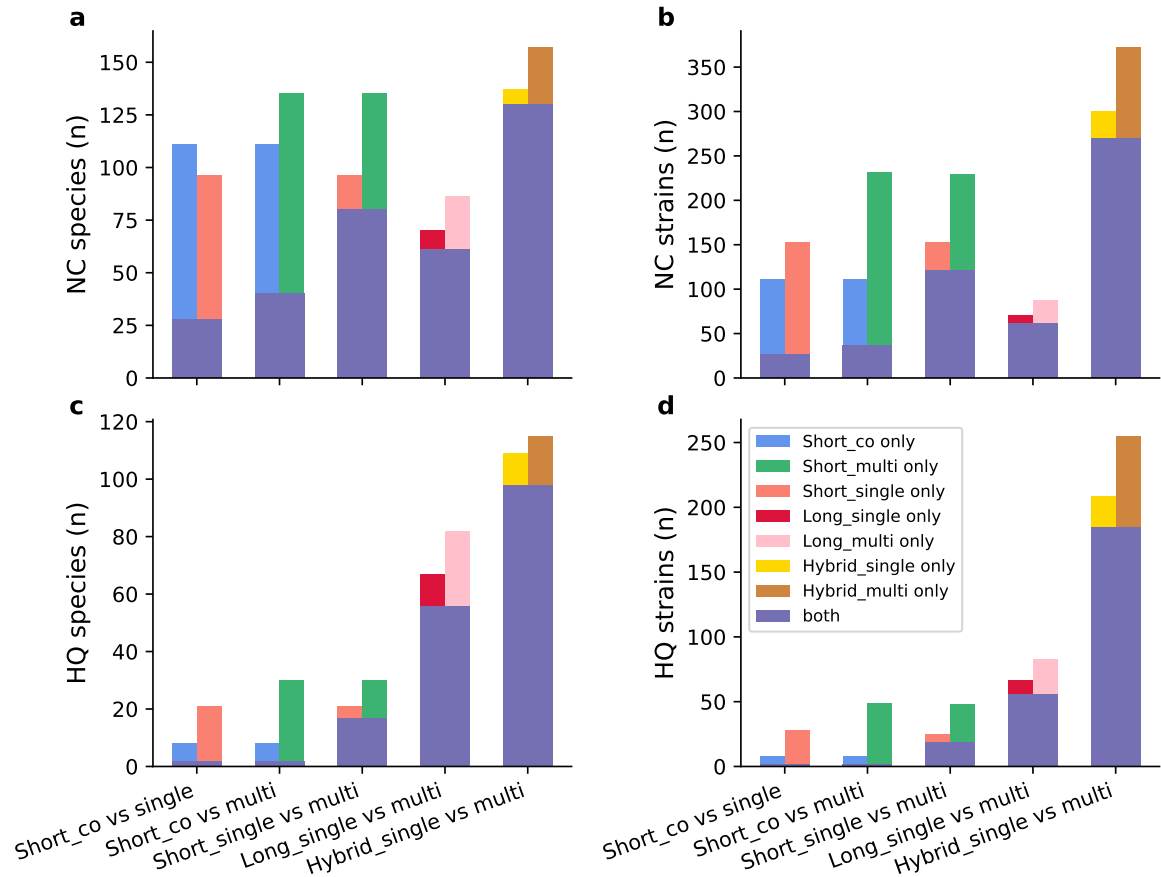
Supplementary Figure 6 Number of NC and HQ MAGs recovered by MetaBAT 2 and MAGScoT. **a, b** The number of NC and HQ MAGs recovered from marine dataset. **c, d** The number of NC and HQ MAGs recovered from the cheese dataset. **e, f** The number of NC and HQ MAGs recovered from the human gut I dataset. **g, h** The number of NC and HQ MAGs recovered from the human gut II dataset. **i, j** The number of NC and HQ MAGs recovered from the activated sludge dataset.



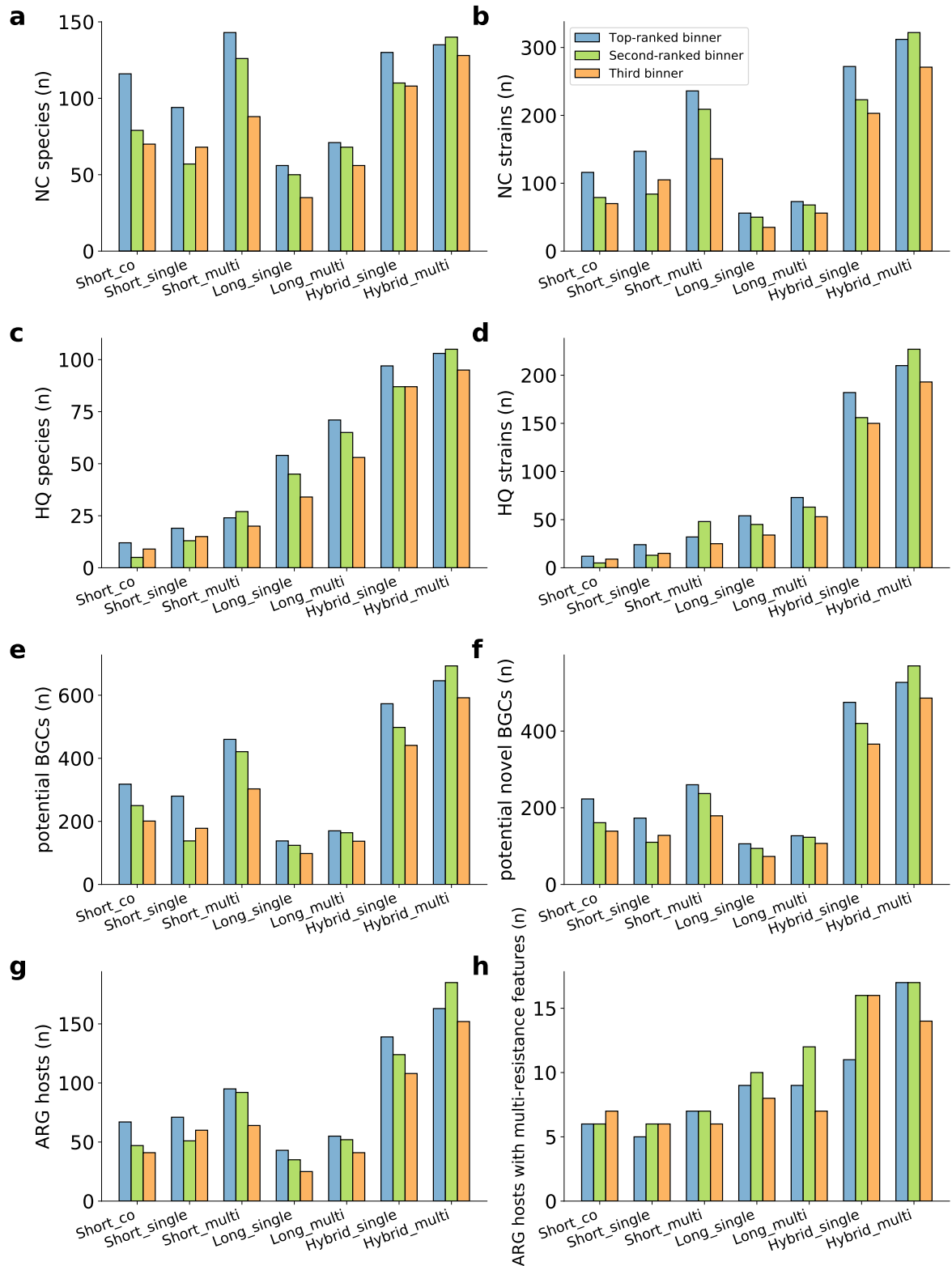
Supplementary Figure 7 Number of NC species recovered by the top three high-performance binners across seven data-binning combinations in the marine dataset. Venn diagrams illustrate the overlap of NC species recovered by these binners. The “Top-ranked binner” (red), “Second-ranked binner” (green), and “Third-ranked binner” (purple) refer to the binners with the highest, second-highest, and third-highest overall ranking scores for each data-binning combination, respectively.



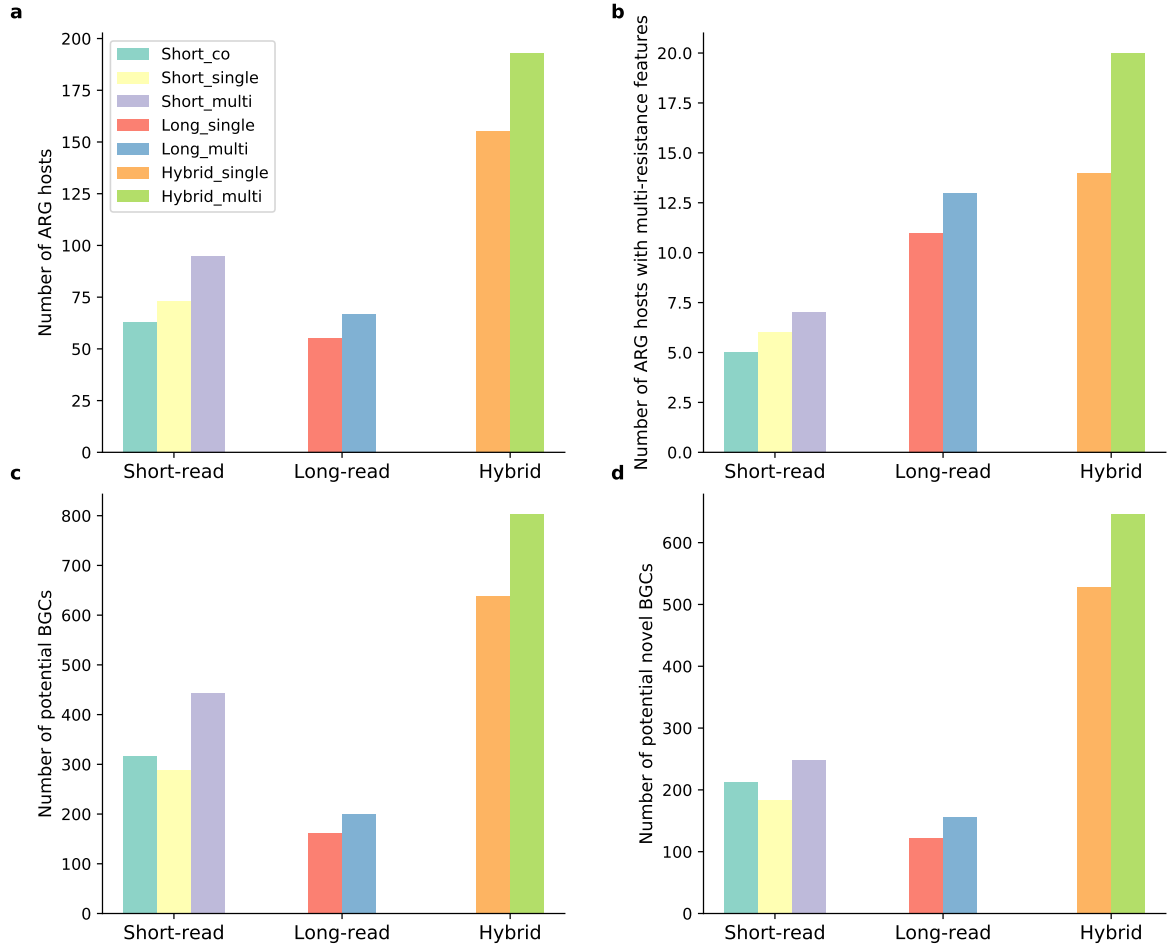
Supplementary Figure 8 The number of distinct taxa at the species, genus, family, order, class, phylum, and domain level. GTDB-Tk was used to annotate the NC species recovered by the top three high-performance binners across seven data-binning combinations in the marine dataset.



Supplementary Figure 9 Comparison of binning modes across diverse data types on the marine dataset. Dereplication was performed on the MAGs recovered from distinct binning modes within specific types of data. **a, b** Overlap and unique species or strain among the NC MAGs recovered by different binning modes. **c, d** Overlap and unique species or strain among the HQ MAGs recovered by different binning modes.



Supplementary Figure 10 Comparison of the top three high-performance binners on the marine dataset. **a, b** Number of species or strains after dereplication of NC MAGs across the seven data-binning combinations. **c, d** Number of species or strains after dereplication of HQ MAGs across the seven data-binning combinations. **e, f** The number of potential BGCs or novel BGCs in NC strains across the seven data-binning combinations. **g, h** The number of potential ARG hosts or ARG hosts with multi-resistance features across the seven data-binning combinations. The “Top-ranked binner”, “Second-ranked binner”, and “Third-ranked binner” refer to the binners with the highest, second-highest, and third-highest overall ranking scores for each data-binning combination, respectively.



Supplementary Figure 11 Potential ARG Hosts and BGCs Identified in refined NC Strains from the Marine Dataset. **a**, The number of potential ARG hosts across seven data-binning combinations. **b**, The number of potential ARG hosts with multi-resistance features across seven data-binning combinations. **c**, The number of potential BGCs in NC strains across seven data-binning combinations. **d**, The number of potential novel BGCs in NC strains across seven data-binning combinations.

References

- [1] Orellana, L.H., Krüger, K., Sidhu, C., Amann, R.: Comparing genomes recovered from time-series metagenomes using long-and short-read sequencing technologies. *Microbiome* **11**(1), 105 (2023)
- [2] Saak, C.C., Pierce, E.C., Dinh, C.B., Portik, D., Hall, R., Ashby, M., Dutton, R.J.: Longitudinal, multi-platform metagenomics yields a high-quality genomic catalog and guides an in vitro model for cheese communities. *Msystems* **8**(1), 00701–22 (2023)
- [3] Jia, L., Wu, Y., Dong, Y., Chen, J., Chen, W.-H., Zhao, X.-M.: A survey on computational strategies for genome-resolved gut metagenomics. *Briefings in Bioinformatics*, 162 (2023)
- [4] Chen, L., Zhao, N., Cao, J., Liu, X., Xu, J., Ma, Y., Yu, Y., Zhang, X., Zhang, W., Guan, X., *et al.*: Short-and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nature Communications* **13**(1), 3175 (2022)
- [5] Singleton, C.M., Petriglieri, F., Kristensen, J.M., Kirkegaard, R.H., Michaelsen, T.Y., Andersen, M.H., Kondrotaitė, Z., Karst, S.M., Dueholm, M.S., Nielsen, P.H., *et al.*: Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature Communications* **12**(1), 2009 (2021)
- [6] Letunic, I., Bork, P.: Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**(W1), 256–259 (2019)