Supplementary Information

for

Comparison of large-insert, small-insert, and pyrosequencing libraries for metagenomic analysis

Thomas Danhorn Curtis R. Young Edward F. DeLong

Department of Civil and Environmental Engineering
Massachusetts Institute of Technology,
Cambridge, MA 02139

Effects of library choice on the perceived diversity of the sampled community

Supplementary Table 1 Diversity measures of libraries

Cruise	*Depth*	*Species Richness*			*Effective Species from Shannon Entropy*		
		Fosmid Ends	*Shotgun*	*454*	*Fosmid Ends*	*Shotgun*	*454*
HOT 179	25 m	729	1282	1240	121.6	31.4	12.2
	75 m	753	1127	1224	88.0	10.5	8.9
	125 m	930	1314	1293	150.3	40.0	25.0
	500 m	894	1365	1367	160.0	96.3	141.9
HOT 186	25 m	752	1011	1487	98.8	32.7	27.1
	75 m	702	895	1495	100.7	14.1	21.4
	110 m	643		1477	13.7		33.6
	500 m	914		1620	176.9		98.0
HOT 179[a]	125 m	1108	738	768	97.3	23.6	24.1

[a]Test for effects of read length and library size. Fosmid ends and shotgun reads were split to have the same average length as 454 reads. Shotgun and 454 data are random subsets with the same number of assigned reads as the fosmid library with split reads.

To assess if the choice of sequencing library affects the observed taxonomic diversity, we calculated both the species richness and the effective number of species using the Shannon entropy (for an overview of diversity measures and their relationship see Jost, 2006). The former is simply the number of different taxa, whereas the latter takes their relative abundance into account to calculate the diversity in equivalents of equally common species. Since MEGAN assigns BLAST hits not only to species but also to higher taxa, we chose to treat those equivalent to species, with the exception of the root of the taxonomic tree as well as the non-hits and unassigned reads, which were removed

before the calculation. The results are summarized in Supplementary Table 1. As a result of the smaller library sizes (Table 1 in main text), the fosmid ends data sets had somewhat lower species richness than shotgun and 454. For almost all samples, however, the fosmid libraries showed a much higher number of effective species. This is primarily a result of the reduction of highly dominant species such as the *Prochlorococcus* strains due to their GC content, which in turn increased the relative concentration of all other taxa, and a more even distribution resulted in a greater number of effective species, which is a measure of diversity. A notable exception was the fosmid library of HOT 186–110 m with only 13.4 effective species, which was consistent with the atypical taxonomic composition of this library—possibly the result of an aberration in its construction (see "Taxonomic composition of the data sets" in the main text, as well as Supplementary Table 8 and Supplementary Figure 2). While the higher diversity is a potentially beneficial side effect of the biased composition of fosmid libraries, their cost and comparatively small size limit their usefulness for capturing the largest possible cross section of a microbial community.

To show the influence of read length and library size on the calculation of the diversity measures, the last line in Supplementary Table 1 shows the effect of randomly splitting the fosmid and shotgun sequences of the HOT 179–125 m data set as described in section "GC content and bias" of the main text into chunks approximately equal to the length of the 454 reads and selecting random subsets from the 454 and the split shotgun library to achieve the same number of assigned BLAST hits as for the split fosmid sequences. The species richness is strongly correlated with the library size (compare the fosmid results, where splitting increases the number of reads, to the 454 numbers, where only a subset is used), while the read length has a smaller effect (the shotgun results are only slightly more reduced than 454 results). The Shannon Entropy-based measure is much less sensitive to the library size changes (compare the 454 results). Shorter read lengths do decrease the absolute numbers for fosmids and shotgun libraries, but the higher diversity in the fosmid library relative to the others when using Shannon Entropy is robust to changes in both library size and read length.

References

Jost L. (2006). Entropy and diversity. *Oikos* **113**: 363–375.

Tables

List of Tables

Supplementary Table 2 Archive information for sequence libraries

Cruise	*Depth*	*Library Type*	*Vector*	*Database*	*Accession or TI Numbers*
HOT 179	25 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		Shotgun	pUC18	NCBI Trace Archive	2242553305–2242566360, 2242586814–2242715268
		454	—	NCBI Seq. Read Arch.	SRX002155
	75 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		Shotgun	pUC18	NCBI Trace Archive	2242740565–2242890189
		454	—	NCBI Seq. Read Arch.	SRX000174
	125 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		Shotgun	pUC18	NCBI Trace Archive	2242890190–2243042992
		454	—	NCBI Seq. Read Arch.	SRX002157
	500 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		Shotgun	pUC18	NCBI Trace Archive	2243042993–2243052976, 2243078729–2243218158
		454	—	NCBI Seq. Read Arch.	SRX002159
HOT 186	25 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		Shotgun	pUC18	NCBI Trace Archive	2281908320–2281966591
		454	—	NCBI Seq. Read Arch.	SRX007372
	75 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		Shotgun	pUC18	NCBI Trace Archive	2281966592–2282003263, 2282006336–2282042335
		454	—	NCBI Seq. Read Arch.	SRX007369
	110 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		454	—	NCBI Seq. Read Arch.	SRX007370
	500 m	Fosmid Ends	pCC1FOS	CAMERA	CAM_P_0000828
		454	—	NCBI Seq. Read Arch.	SRX007371

Supplementary Table 3 Prevalence and GC content of taxonomic groups in HOT 179 sequencing libraries at 25–75 m using `blastn`

Depth:	*25 m*									*75 m*								
Library:	*Fosmid Ends*			*Shotgun*			*454*			*Fosmid Ends*			*Shotgun*			*454*		
Taxa	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]
Bacteria	85.4	46.8	13.0	93.5	34.9	9.7	97.2	32.7	8.5	91.0	43.7	12.0	97.8	32.8	6.7	98.4	32.2	7.7
Actinobacteria	3.7	61.7	7.4	0.7	61.4	8.2	0.3	64.3	10.2	2.3	60.2	8.1	0.2	59.1	9.9	0.1	62.2	10.6
Bacteroidetes	2.6	41.1	9.8	2.6	34.4	7.5	1.3	34.6	7.6	2.6	37.5	6.4	1.3	32.9	5.5	0.8	34.1	6.8
Flavobacteria	1.8	38.4	8.2	2.2	33.0	5.2	1.1	34.0	7.0	2.0	36.9	5.9	1.2	32.4	4.6	0.7	33.8	6.5
Cyanobacteria	30.1	39.8	12.9	53.8	32.8	7.0	77.3	31.9	7.3	38.9	37.3	11.1	78.5	32.1	5.5	83.9	31.8	7.1
Chroococcales	7.6	58.6	7.3	2.7	55.4	10.7	1.8	54.5	13.2	7.3	57.7	6.4	1.5	55.1	10.4	1.4	57.0	10.3
Prochlorales	22.3	32.5	4.7	50.8	31.7	4.2	75.2	31.3	6.2	31.1	32.4	4.6	76.8	31.7	4.3	82.3	31.4	6.3
Firmicutes	2.5	41.2	7.6	1.8	34.4	8.2	1.0	32.8	8.1	2.7	39.6	7.4	0.9	33.1	7.1	0.7	32.6	8.2
Proteobacteria	35.7	51.1	11.1	26.6	36.2	11.2	14.8	35.4	11.3	33.1	49.5	10.2	13.9	35.8	10.0	10.9	34.4	9.9
Alphaproteobacteria	15.5	50.5	11.5	20.0	33.5	9.1	11.7	33.4	9.5	13.7	49.6	10.6	9.0	33.1	8.3	8.8	32.8	8.5
Rhizobiales	2.9	56.7	9.9	0.7	50.0	12.9	0.3	51.1	14.2	2.7	54.3	8.3	0.3	46.8	12.5	0.2	49.5	13.7
Rhodobacterales	3.7	50.1	9.7	0.9	46.6	9.9	0.4	49.9	11.2	3.3	50.7	8.9	0.4	46.3	9.1	0.2	49.3	10.7
Rickettsiales	2.7	34.1	5.3	16.8	30.5	4.2	10.4	31.1	5.8	2.7	34.4	5.7	7.7	30.5	4.3	8.0	31.2	5.9
SAR11 cluster	2.3	33.4	5.0	16.3	30.4	4.1	10.0	31.0	5.8	2.1	33.7	5.0	7.3	30.4	4.2	7.7	31.1	5.9
Betaproteobacteria	3.9	59.3	9.3	0.9	55.8	12.5	0.4	57.2	13.7	2.6	56.4	9.0	1.7	48.1	8.4	0.2	53.2	14.5
Burkholderiales	3.1	60.0	9.0	0.7	58.7	10.3	0.3	61.1	11.0	2.1	57.4	8.5	1.6	48.3	8.0	0.1	58.6	12.3
Gammaproteobacteria	9.9	47.8	9.7	3.6	42.1	11.1	1.7	41.0	11.7	9.9	46.0	8.8	2.1	40.0	9.9	1.2	39.9	10.8
Alteromonadales	1.7	43.8	8.0	0.6	37.9	8.6	0.3	39.1	9.5	1.6	44.0	8.5	0.3	37.5	7.9	0.2	37.9	9.7
Enterobacteriales	1.0	45.8	8.5	0.8	42.8	12.0	0.2	35.6	11.2	1.3	46.5	10.2	0.3	35.2	8.1	0.2	35.8	10.4
Deltaproteobacteria	1.8	58.3	8.6	0.4	55.8	12.8	0.1	54.7	15.1	2.0	57.0	8.7	0.1	54.5	12.0	0.1	51.2	16.6
Archaea	1.3	51.2	9.6	0.4	39.3	12.2	0.2	36.0	14.9	1.0	47.0	10.4	0.2	37.3	11.9	0.2	33.5	13.5
Euryarchaeota	1.1	51.4	10.3	0.3	39.9	12.4	0.2	36.7	15.4	0.9	47.2	10.6	0.2	38.0	12.6	0.1	33.5	14.0
Thaumarchaeota	—	—	—	0.0	32.5	6.5	0.0	28.9	6.8	0.0	31.0	1.7	0.0	33.5	4.9	0.0	32.8	8.4
Eukaryota	7.2	50.8	11.7	3.0	40.3	13.4	1.1	32.9	14.1	3.6	48.6	12.4	1.0	36.2	12.7	0.7	29.3	12.2
Fungi	3.2	54.0	10.4	1.2	48.8	10.6	0.3	43.1	14.9	1.6	53.6	10.1	0.2	45.2	13.0	0.2	38.5	15.3
Viruses	2.3	41.0	9.2	1.9	36.1	5.0	1.1	36.6	6.7	2.4	39.4	7.1	0.4	36.7	5.8	0.4	36.4	6.3
Caudovirales	1.9	37.9	4.4	1.9	36.0	3.9	1.0	36.5	5.8	2.2	38.1	4.8	0.4	36.3	4.2	0.4	36.6	5.8

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 4 Prevalence and GC content of taxonomic groups in HOT 179 sequencing libraries at 125–500 m using `blastn`

Depth:	*125 m*									*500 m*								
Library:	*Fosmid Ends*			*Shotgun*			*454*			*Fosmid Ends*			*Shotgun*			*454*		
Taxa	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]
Bacteria	89.7	45.9	11.2	95.0	34.1	8.2	96.1	33.5	8.7	87.3	53.1	9.8	77.4	43.1	12.9	83.6	45.6	14.4
Actinobacteria	4.0	59.7	8.2	0.7	56.8	10.4	0.5	62.5	10.1	8.6	62.2	5.5	4.6	63.4	6.7	7.2	65.0	6.3
Bacteroidetes	2.7	40.6	7.9	3.0	33.0	5.8	1.7	33.9	7.4	2.9	47.0	9.3	2.5	39.0	9.3	2.6	39.8	10.7
Flavobacteria	1.9	38.1	6.3	2.6	32.5	5.0	1.4	33.2	6.5	1.1	40.8	6.7	1.6	36.4	6.9	1.7	36.9	8.7
Cyanobacteria	25.8	38.0	9.1	35.5	33.0	5.8	55.8	32.6	7.2	1.4	47.4	9.7	1.8	39.0	10.2	1.6	41.0	12.4
Chroococcales	3.4	55.5	7.9	0.8	50.4	12.2	0.8	53.6	12.1	0.7	49.9	9.6	0.5	44.7	12.3	0.6	47.3	12.8
Prochlorales	21.6	35.1	5.4	34.5	32.6	4.8	54.5	32.2	6.6	0.2	38.2	6.0	0.9	33.9	6.5	0.6	34.3	8.6
Firmicutes	4.4	42.6	8.4	2.8	33.9	7.3	2.1	32.8	8.4	4.1	46.9	9.2	4.0	39.8	9.4	4.8	39.9	11.3
Proteobacteria	34.8	49.7	10.5	44.4	34.0	9.0	30.5	34.1	9.7	42.6	52.4	9.6	49.3	40.7	12.0	51.0	42.9	13.6
Alphaproteobacteria	13.9	49.4	11.0	34.4	32.0	7.2	25.6	32.7	8.1	13.7	54.3	9.8	27.7	36.4	11.4	28.7	38.9	13.4
Rhizobiales	3.4	55.2	8.5	1.0	48.1	12.4	0.6	49.8	14.3	4.2	56.8	7.1	2.2	53.4	10.5	2.7	56.7	10.9
Rhodobacterales	2.8	50.4	7.9	0.9	45.6	10.0	0.5	48.4	10.9	2.7	54.2	7.8	1.7	51.7	9.3	2.1	55.8	9.5
Rickettsiales	3.0	33.9	5.4	31.1	30.4	4.2	23.6	31.4	6.0	1.6	35.7	4.1	21.2	31.1	4.4	20.4	32.2	6.1
SAR11 cluster	2.6	33.5	5.2	30.2	30.4	4.1	22.9	31.3	6.0	1.4	35.4	4.1	20.5	31.0	4.3	19.5	32.0	6.0
Betaproteobacteria	2.8	57.9	8.6	2.8	49.2	9.6	0.5	55.0	13.9	4.2	57.0	7.1	3.9	53.2	9.5	2.7	57.2	10.2
Burkholderiales	2.1	58.0	8.5	2.5	50.1	8.7	0.3	59.9	12.4	3.0	57.7	6.7	3.2	53.9	9.0	1.8	59.0	9.3
Gammaproteobacteria	10.7	46.5	8.8	4.4	40.1	9.7	2.6	39.8	11.0	13.7	47.2	8.7	11.2	43.1	8.5	12.2	44.7	10.0
Alteromonadales	1.7	43.4	8.1	0.8	38.9	8.3	0.5	38.6	9.3	2.6	45.4	6.9	2.1	43.2	6.3	3.7	44.4	7.5
Enterobacteriales	1.3	47.4	9.1	0.6	36.3	8.8	0.5	35.9	10.9	1.3	46.6	8.8	1.1	42.7	9.9	1.4	42.4	12.1
Deltaproteobacteria	2.2	58.3	8.6	0.4	51.8	11.9	0.2	54.0	14.4	4.0	56.4	8.2	1.8	54.0	9.9	2.2	57.0	10.3
Archaea	2.7	45.5	9.7	0.9	36.2	8.5	0.7	36.5	11.9	8.9	41.5	10.0	18.5	36.6	5.1	11.2	38.0	7.7
Euryarchaeota	1.8	48.0	8.9	0.5	38.2	10.7	0.4	37.3	13.9	2.5	54.8	8.6	1.1	45.2	12.0	1.2	43.8	14.1
Thaumarchaeota	0.5	34.4	5.3	0.4	33.8	3.3	0.2	34.8	5.9	6.1	36.0	3.1	17.0	36.0	3.5	9.6	37.1	5.8
Eukaryota	3.8	49.1	10.2	2.0	34.0	11.9	1.7	29.6	12.9	1.3	48.8	10.6	2.0	34.1	10.6	2.7	31.1	12.1
Fungi	2.1	50.7	9.4	0.5	42.4	11.9	0.4	39.8	15.2	0.7	50.6	9.1	0.6	41.8	10.5	0.6	40.1	12.9
Viruses	1.2	41.0	9.3	0.7	36.2	6.0	0.6	37.1	7.5	0.1	44.4	8.2	0.2	35.5	8.0	0.2	37.6	11.5
Caudovirales	1.1	38.7	5.6	0.6	36.3	4.2	0.6	37.5	6.1	0.0	35.4	0.0	0.1	35.2	5.3	0.1	41.6	10.5

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 5 Prevalence and GC content of taxonomic groups in HOT 179 sequencing libraries at 25–75 m using `blastx`

Depth:	*25 m*									*75 m*								
Library:	*Fosmid Ends*			*Shotgun*			*454*			*Fosmid Ends*			*Shotgun*			*454*		
Taxa	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]
Bacteria	68.5	48.2	12.3	86.0	36.3	10.9	94.8	33.8	8.4	80.8	45.1	11.2	94.6	33.6	7.8	96.8	33.4	7.8
Actinobacteria	3.2	54.4	12.1	1.2	49.4	13.6	0.3	45.3	12.4	3.5	51.1	11.0	0.6	48.2	12.3	0.3	47.2	11.9
Bacteroidetes	3.5	46.4	11.5	3.9	35.5	10.2	2.2	34.7	8.6	2.7	40.2	9.5	2.3	32.6	6.1	1.4	33.0	6.4
Flavobacteria	1.2	43.8	11.5	2.1	32.7	7.1	1.1	33.4	7.4	1.0	37.3	8.0	1.3	31.6	4.8	0.8	32.1	6.0
Cyanobacteria	12.1	41.2	13.1	35.4	32.9	7.1	59.3	32.3	7.0	16.6	37.9	11.4	63.3	32.1	5.5	67.9	32.3	6.8
Chroococcales	2.9	56.7	8.3	2.0	50.4	13.6	1.2	52.9	14.0	3.1	55.3	9.0	1.7	47.4	13.8	1.0	53.9	12.8
Prochlorales	7.4	32.5	5.6	31.3	31.4	4.1	50.1	31.0	5.6	11.8	32.2	4.7	58.5	31.4	4.1	57.9	31.1	5.6
Firmicutes	1.1	49.0	11.1	1.1	43.0	13.1	0.3	39.4	11.7	1.1	43.1	9.6	0.7	38.8	10.9	0.3	41.5	11.8
Proteobacteria	30.1	48.7	11.6	32.0	36.7	11.1	23.4	35.5	9.8	36.3	46.5	10.2	21.5	36.5	10.1	18.8	35.2	9.2
Alphaproteobacteria	9.9	45.5	10.5	20.1	33.1	8.2	15.6	33.1	8.5	11.5	44.8	9.5	11.8	33.1	7.9	12.5	33.0	8.0
Rhizobiales	1.8	48.2	10.7	2.0	37.3	10.6	0.6	39.1	10.7	2.7	46.7	9.7	1.4	36.8	9.5	0.5	39.4	10.3
Rhodobacterales	3.6	42.3	7.5	3.1	37.8	7.6	1.5	40.2	8.1	3.7	42.8	7.8	1.7	37.1	7.4	1.1	39.9	8.2
Rickettsiales	0.8	32.9	6.9	11.5	29.6	4.2	9.7	29.8	5.5	0.8	33.1	6.9	6.3	29.5	4.2	7.7	29.9	5.5
SAR11 cluster	0.8	32.2	5.7	11.4	29.6	4.1	9.6	29.8	5.5	0.7	32.9	6.5	6.2	29.5	4.1	7.6	29.9	5.5
Betaproteobacteria	3.1	60.0	11.0	1.4	52.1	14.6	0.4	47.6	14.1	2.3	54.1	12.2	1.8	45.8	12.1	0.3	41.7	11.4
Burkholderiales	2.6	61.2	10.3	1.2	54.0	14.2	0.3	51.2	13.4	1.9	55.6	12.0	1.6	47.7	11.7	0.2	43.9	11.9
Gammaproteobacteria	8.2	46.6	9.6	5.4	40.6	10.3	2.4	42.2	10.7	11.1	45.0	9.6	4.3	39.7	9.8	1.9	41.9	10.1
Alteromonadales	0.9	43.4	9.7	0.6	37.9	9.2	0.2	39.1	10.0	1.2	43.2	10.0	0.5	37.2	8.5	0.2	39.0	9.4
Enterobacteriales	0.2	46.0	10.6	0.3	47.0	12.1	0.1	36.9	9.2	0.3	46.2	11.1	0.1	38.0	11.3	0.0	38.2	9.3
Deltaproteobacteria	2.7	55.5	10.9	1.2	50.0	13.6	0.2	43.3	10.7	3.8	51.5	9.1	0.8	47.1	12.7	0.2	44.3	11.0
Archaea	2.9	50.4	7.0	0.8	45.6	10.8	0.2	45.2	11.0	2.3	48.8	6.6	0.3	42.7	10.2	0.1	41.0	10.2
Euryarchaeota	2.5	50.3	7.0	0.7	45.6	10.9	0.1	44.6	10.9	1.9	49.0	6.5	0.3	42.9	10.3	0.1	40.7	10.0
Thaumarchaeota	0.1	50.2	7.5	0.0	35.9	8.8	0.0	40.5	10.4	0.0	45.9	7.4	0.0	35.6	6.7	0.0	36.0	10.5
Eukaryota	12.2	54.7	9.1	4.5	47.6	13.0	0.5	47.4	13.7	5.8	54.5	9.6	2.0	43.9	13.1	0.2	49.1	13.1
Fungi	3.9	56.4	8.2	1.7	50.9	10.6	0.1	49.4	13.0	1.9	56.2	8.5	0.5	47.9	12.9	0.1	46.7	14.0
Viruses	2.2	44.4	13.0	3.7	36.0	7.2	2.2	35.5	5.9	1.9	41.5	10.9	0.8	36.8	8.1	0.9	35.7	5.6
Caudovirales	1.4	36.2	4.4	3.1	35.4	3.7	2.1	35.5	5.5	1.5	36.7	4.8	0.7	35.7	4.0	0.9	35.7	5.5

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 6 Prevalence and GC content of taxonomic groups in HOT 179 sequencing libraries at 125–500 m using `blastx`

Depth:	*125 m*									*500 m*								
Library:	*Fosmid Ends*			*Shotgun*			*454*			*Fosmid Ends*			*Shotgun*			*454*		
Taxa	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]	*%Rd.*[a]	*%GC*	*SD*[b]
Bacteria	78.8	47.1	10.0	90.5	35.8	9.7	93.9	35.5	8.9	81.7	51.6	9.4	77.2	43.9	11.8	77.3	46.0	12.5
Actinobacteria	5.6	49.5	8.5	1.9	46.8	10.4	0.9	45.5	9.7	4.4	59.7	8.6	3.6	59.9	10.4	3.2	62.7	9.2
Bacteroidetes	2.6	43.5	9.7	4.0	33.3	7.8	2.0	33.9	7.7	3.8	46.2	9.2	3.2	40.2	9.2	1.7	42.2	10.1
Flavobacteria	0.7	39.5	8.3	1.9	31.6	5.9	0.9	32.6	6.4	0.8	45.5	9.5	0.8	37.2	8.9	0.5	38.9	9.3
Cyanobacteria	9.4	39.0	9.6	21.6	33.1	6.1	35.7	33.5	7.0	1.3	51.6	9.0	1.3	41.2	11.0	0.7	45.6	12.5
Chroococcales	1.2	52.3	10.2	0.7	43.7	13.5	0.6	50.3	12.7	0.3	51.8	9.1	0.3	44.1	11.0	0.2	48.7	11.4
Prochlorales	6.0	34.6	5.5	18.7	32.2	4.6	28.6	32.0	5.9	0.1	47.8	8.5	0.4	33.4	8.3	0.2	36.2	10.2
Firmicutes	1.7	46.1	8.9	1.5	37.4	10.0	0.6	39.7	11.1	2.6	51.2	8.9	2.5	44.5	10.4	1.3	48.0	11.5
Proteobacteria	30.4	47.7	10.0	45.8	35.1	9.7	37.8	35.1	9.1	27.2	49.8	9.4	39.0	40.7	11.0	39.3	42.1	11.7
Alphaproteobacteria	9.4	45.7	9.9	30.9	32.3	7.4	26.6	33.0	7.9	6.7	51.0	9.5	17.9	36.5	11.1	18.2	38.2	12.1
Rhizobiales	2.2	47.9	10.2	3.0	37.0	10.2	1.0	40.3	11.1	2.0	51.9	8.2	2.2	45.9	11.3	1.6	49.4	11.5
Rhodobacterales	2.2	44.5	8.4	3.1	36.9	7.7	1.6	40.1	8.5	1.1	51.2	8.9	1.7	44.0	10.4	1.7	47.2	10.7
Rickettsiales	0.7	32.4	5.9	19.3	29.5	4.1	17.0	30.3	5.5	0.4	36.2	7.2	9.9	30.0	4.5	8.6	30.7	6.0
SAR11 cluster	0.7	32.3	5.2	19.1	29.5	4.1	16.8	30.3	5.4	0.4	35.9	7.2	9.8	30.0	4.4	8.4	30.7	5.9
Betaproteobacteria	2.3	54.9	11.4	2.3	46.5	12.0	0.6	44.1	11.4	1.8	53.9	8.6	2.1	48.1	10.0	1.2	50.0	10.9
Burkholderiales	1.9	56.1	11.3	2.0	48.1	11.8	0.4	45.7	11.3	1.3	55.0	8.1	1.6	49.3	9.9	0.8	51.6	11.0
Gammaproteobacteria	8.4	45.5	8.2	6.1	40.3	9.8	2.8	42.0	9.8	7.1	46.1	8.6	8.4	42.3	8.3	6.9	44.3	9.0
Alteromonadales	1.0	44.5	7.5	0.7	39.1	9.4	0.3	39.9	10.3	1.1	46.8	8.8	1.1	43.3	7.7	1.4	45.6	7.0
Enterobacteriales	0.2	49.4	7.2	0.2	37.5	9.8	0.1	37.7	9.4	0.2	49.9	7.7	0.2	40.9	10.0	0.1	45.4	12.3
Deltaproteobacteria	3.8	52.3	10.0	1.5	46.2	12.2	0.5	45.4	11.1	4.6	51.5	9.8	3.4	45.8	9.9	1.5	48.3	10.9
Archaea	5.2	47.4	6.1	1.4	40.8	9.2	0.8	40.3	9.3	7.5	46.2	11.3	14.5	37.0	6.5	14.6	37.3	6.4
Euryarchaeota	4.2	48.0	5.3	0.9	43.1	9.0	0.4	44.5	9.6	3.5	55.1	7.3	1.6	47.0	11.2	0.7	49.0	11.4
Thaumarchaeota	0.3	37.1	8.8	0.4	33.5	4.3	0.3	34.8	5.0	3.4	35.9	3.8	12.6	35.5	3.5	13.0	36.5	5.0
Eukaryota	5.4	54.4	9.0	2.8	45.4	12.7	0.5	51.1	11.9	1.3	53.7	9.3	1.5	40.9	12.2	0.3	45.6	10.8
Fungi	1.9	55.8	8.7	0.8	48.4	12.2	0.2	49.1	12.3	0.5	55.5	8.0	0.4	44.4	11.4	0.1	46.4	10.2
Viruses	1.1	41.2	10.3	1.2	36.8	8.1	1.1	36.4	6.1	0.4	52.1	10.8	0.9	36.7	8.5	0.4	36.3	6.9
Caudovirales	0.8	36.5	4.7	1.0	35.8	4.2	1.1	36.3	5.7	0.2	50.0	12.3	0.6	36.8	5.9	0.4	36.3	6.9

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 7 Prevalence and GC content of taxonomic groups in HOT 186 sequencing libraries at 25–75 m using `blastn`

Depth:	*25 m*									*75 m*								
Library:	*Fosmid Ends*			*Shotgun*			*454*			*Fosmid Ends*			*Shotgun*			*454*		
Taxa	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]
Bacteria	98.5	49.7	9.2	96.1	33.7	7.1	96.3	33.7	8.0	95.3	49.9	9.6	91.2	38.1	7.8	95.2	32.9	6.9
Actinobacteria	1.5	58.8	5.6	0.4	56.1	10.9	0.4	61.1	9.7	2.8	61.1	6.6	0.1	56.0	7.7	0.2	58.6	11.2
Bacteroidetes	3.0	42.1	9.1	4.8	33.2	5.1	3.0	34.2	6.1	3.0	42.3	7.3	2.4	34.8	4.6	2.5	33.5	5.6
Flavobacteria	2.2	39.6	7.7	4.3	32.9	4.6	2.6	33.8	5.6	1.6	39.4	7.1	2.1	34.5	4.1	2.2	33.2	5.2
Cyanobacteria	11.3	47.7	13.9	40.9	32.2	4.3	52.0	31.8	5.7	17.2	48.9	13.6	38.1	34.2	4.7	58.8	31.9	5.7
Chroococcales	6.6	58.9	6.3	0.4	43.0	13.1	0.6	49.8	13.3	10.9	58.1	6.8	0.6	53.4	9.2	0.4	50.1	13.3
Prochlorales	4.4	33.2	5.4	40.1	32.0	4.0	51.0	31.5	5.3	6.1	32.8	5.3	37.4	33.9	3.9	58.1	31.7	5.3
Firmicutes	1.6	42.9	8.4	2.3	32.9	5.5	1.9	34.0	7.0	1.5	43.7	8.1	3.0	46.3	7.6	1.8	33.7	6.9
Proteobacteria	69.0	50.3	8.3	39.4	34.8	8.5	33.3	35.7	9.6	59.9	50.5	8.3	12.9	40.6	8.4	27.0	34.3	8.3
Alphaproteobacteria	25.6	52.8	9.6	28.9	32.6	7.2	23.6	33.3	8.6	19.9	52.8	9.5	6.3	36.7	8.1	20.3	32.1	6.8
Rhizobiales	3.1	54.3	7.8	0.6	45.6	13.0	0.6	48.2	13.2	2.8	55.4	8.5	0.3	44.7	10.0	0.4	43.7	12.1
Rhodobacterales	10.9	54.8	8.1	1.1	47.2	10.2	1.1	51.6	10.7	9.2	53.6	8.3	0.8	46.6	7.9	0.5	46.8	10.4
Rickettsiales	2.3	33.6	5.2	25.8	30.8	3.9	20.8	31.1	4.9	1.5	34.1	6.2	4.7	33.2	4.0	18.9	31.1	5.1
SAR11 cluster	2.1	33.5	5.2	25.2	30.8	3.9	20.2	31.0	4.9	1.2	33.2	5.8	4.4	33.1	4.0	18.4	31.0	5.0
Betaproteobacteria	2.2	54.7	6.3	0.4	47.2	12.4	0.4	48.8	13.4	1.9	54.8	7.7	0.2	45.6	8.6	0.3	44.9	11.7
Burkholderiales	1.5	55.8	5.6	0.2	52.8	12.0	0.2	54.4	12.2	1.4	55.0	8.1	0.1	47.7	7.8	0.1	49.2	11.7
Gammaproteobacteria	34.6	47.7	6.5	7.8	41.1	7.9	7.2	42.0	8.0	31.7	48.1	6.7	5.4	44.6	6.2	4.8	41.5	7.9
Alteromonadales	20.7	45.4	4.9	3.4	42.5	5.4	3.6	43.3	5.9	18.7	45.6	5.0	2.7	45.0	4.2	2.3	43.4	6.2
Enterobacteriales	1.5	50.9	7.4	0.7	35.7	7.4	0.7	38.6	9.8	2.0	51.7	7.7	0.9	45.7	7.2	0.5	36.8	9.0
Deltaproteobacteria	0.8	54.7	6.7	0.2	48.4	12.5	0.2	49.8	13.9	1.4	55.7	7.8	0.1	48.8	9.2	0.2	47.4	13.3
Archaea	0.1	49.8	9.9	0.4	35.1	8.6	0.3	36.1	10.6	0.4	45.6	7.0	0.1	37.6	8.0	0.3	35.6	10.2
Euryarchaeota	0.1	49.4	10.6	0.3	35.3	9.1	0.3	36.2	11.1	0.4	45.8	7.6	0.1	37.6	8.2	0.3	35.6	10.5
Thaumarchaeota	—	—	—	0.0	37.4	0.7	0.0	34.4	5.2	—	—	—	0.0	35.0	2.4	0.0	31.4	6.7
Eukaryota	0.3	43.2	10.2	1.3	31.6	8.7	1.1	31.9	11.2	0.5	45.1	10.9	2.4	48.7	6.2	1.0	30.2	10.0
Fungi	0.2	45.5	10.2	0.3	36.6	8.9	0.3	39.9	12.6	0.3	49.3	8.1	0.1	39.2	9.4	0.2	37.1	11.5
Viruses	0.3	39.5	4.1	1.0	35.2	4.3	1.4	36.1	4.8	2.7	40.2	4.8	1.3	39.2	3.5	2.7	37.0	4.8
Caudovirales	0.3	39.4	4.0	0.9	35.4	3.5	1.4	36.2	4.4	2.6	40.1	4.4	1.3	39.2	3.4	2.6	37.1	4.7

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 8 Prevalence and GC content of taxonomic groups in HOT 186 sequencing libraries at 110–500 m using `blastn`

Depth:	*110 m*						*500 m*					
Library:	*Fosmid Ends*			*454*			*Fosmid Ends*			*454*		
Taxa	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]
Bacteria	96.0	39.9	9.0	91.1	34.2	7.6	90.8	49.9	9.5	82.5	41.6	13.0
Actinobacteria	1.2	58.0	7.8	0.5	58.2	10.9	7.1	59.9	6.9	4.5	64.4	6.4
Bacteroidetes	1.2	45.4	8.3	3.1	33.6	6.1	3.5	43.9	8.2	2.9	38.1	8.8
Flavobacteria	0.4	38.1	6.6	2.6	33.1	5.3	1.9	41.5	7.4	2.0	36.0	7.0
Cyanobacteria	69.3	36.2	5.3	42.4	34.1	5.8	3.2	41.7	9.6	6.9	35.0	7.8
Chroococcales	1.6	54.6	7.7	0.4	41.0	11.5	0.8	50.8	9.1	0.5	45.5	12.2
Prochlorales	67.3	35.7	4.4	41.6	34.0	5.7	1.8	35.9	5.4	6.0	33.6	5.9
Firmicutes	1.0	43.2	8.6	2.8	33.9	7.3	5.2	44.7	8.1	4.1	38.9	9.7
Proteobacteria	13.9	49.6	9.3	34.6	33.4	8.3	46.2	49.5	9.1	49.8	39.5	11.9
Alphaproteobacteria	6.0	48.8	9.7	28.3	32.0	6.9	15.3	49.2	10.6	32.9	35.9	10.6
Rhizobiales	1.1	52.0	8.4	0.6	45.2	12.6	4.0	54.2	7.5	2.0	53.3	11.6
Rhodobacterales	2.2	50.5	8.2	0.5	47.3	10.3	3.1	52.7	6.5	1.6	52.8	9.6
Rickettsiales	0.9	36.4	6.7	26.4	30.9	5.1	3.8	34.6	5.4	26.8	32.0	5.2
SAR11 cluster	0.6	34.8	5.8	25.6	30.9	5.1	3.5	33.9	4.4	25.8	31.9	5.1
Betaproteobacteria	1.0	55.8	6.6	0.4	49.5	13.0	3.5	54.9	6.7	1.9	54.6	10.2
Burkholderiales	0.6	57.2	5.8	0.3	54.4	11.7	2.3	56.2	6.1	1.2	56.8	9.1
Gammaproteobacteria	3.9	47.0	7.8	3.7	38.9	9.0	15.9	46.8	7.5	9.0	43.3	9.3
Alteromonadales	0.6	44.9	7.3	0.7	39.1	8.2	2.4	46.5	6.5	1.4	42.2	8.3
Enterobacteriales	0.4	49.4	8.0	0.6	35.9	9.0	1.7	47.8	6.7	1.1	41.4	10.4
Deltaproteobacteria	1.0	58.0	7.2	0.3	51.3	12.7	3.9	54.8	8.2	1.4	54.9	10.7
Archaea	0.3	48.2	11.8	0.6	38.3	11.0	4.8	41.7	9.2	13.3	37.5	5.8
Euryarchaeota	0.3	48.0	11.4	0.5	38.9	11.3	1.5	52.4	7.7	1.1	44.3	12.4
Thaumarchaeota	0.0	36.3	0.0	0.0	34.8	5.9	3.1	36.5	3.5	11.9	36.9	4.2
Eukaryota	0.6	45.9	12.1	1.9	34.0	12.4	1.2	44.1	10.6	1.7	32.5	11.4
Fungi	0.3	47.2	11.4	0.5	41.9	12.4	0.6	48.8	8.4	0.4	40.9	12.0
Viruses	2.3	40.9	4.0	5.1	36.3	4.8	0.3	41.1	8.0	0.6	36.9	5.7
Caudovirales	2.3	40.9	4.0	5.1	36.3	4.5	0.3	39.2	6.8	0.6	37.1	4.8

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 9 Prevalence and GC content of taxonomic groups in HOT 186 sequencing libraries at 25–75 m using `blastx`

Depth:	*25 m*									*75 m*								
Library:	*Fosmid Ends*			*Shotgun*			*454*			*Fosmid Ends*			*Shotgun*			*454*		
Taxa	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]
Bacteria	96.8	49.0	8.5	94.0	34.3	7.7	93.2	34.6	8.5	95.2	48.8	9.0	94.2	39.0	8.2	91.6	34.1	7.8
Actinobacteria	1.3	53.1	10.2	1.4	41.9	11.6	0.9	44.6	13.5	2.2	58.6	9.6	1.2	47.3	9.1	0.9	43.7	11.2
Bacteroidetes	4.9	46.6	10.3	6.8	33.4	6.8	4.8	33.6	7.3	3.4	44.6	10.7	5.4	34.8	5.6	4.0	32.3	5.9
Flavobacteria	2.3	43.7	10.5	4.1	32.4	5.6	2.6	32.5	6.2	1.5	39.8	10.1	3.7	34.0	4.6	2.1	31.7	5.3
Cyanobacteria	5.6	47.3	13.6	28.3	32.1	4.4	30.3	31.9	5.6	8.3	48.9	13.3	35.6	34.5	5.1	36.3	32.0	5.5
Chroococcales	3.1	57.5	8.0	0.6	37.5	11.4	0.5	42.6	13.7	4.7	57.5	7.2	1.0	45.9	11.4	0.4	42.1	13.3
Prochlorales	2.0	33.5	5.9	26.3	31.8	3.8	26.8	31.1	4.8	2.5	32.3	5.0	33.6	34.0	4.1	31.9	31.3	4.9
Firmicutes	0.8	48.1	8.0	1.3	37.9	10.1	0.7	37.9	10.2	0.7	47.9	8.2	0.8	42.1	8.5	0.7	37.7	9.8
Proteobacteria	60.4	48.9	8.2	45.4	34.8	8.2	41.7	35.5	9.1	56.5	48.3	8.4	38.6	42.0	8.4	35.2	34.9	8.5
Alphaproteobacteria	18.7	48.4	9.8	28.4	32.4	6.6	23.3	32.8	8.1	18.2	47.2	9.6	15.4	37.6	7.2	20.7	32.1	7.1
Rhizobiales	2.7	49.0	9.1	2.1	36.1	8.8	1.1	37.5	10.2	2.9	47.9	9.1	2.2	39.1	7.9	1.0	36.8	8.8
Rhodobacterales	9.2	49.0	9.6	3.5	38.2	7.9	2.6	40.7	9.5	9.0	47.1	9.3	5.1	39.7	6.2	1.9	38.7	7.8
Rickettsiales	1.0	32.6	5.4	18.5	30.1	3.8	13.8	29.5	4.7	0.5	31.9	6.3	4.2	32.6	4.3	12.8	29.4	4.9
SAR11 cluster	1.0	32.3	5.1	18.3	30.1	3.8	13.6	29.5	4.7	0.5	31.8	6.1	4.1	32.7	4.2	12.7	29.4	4.9
Betaproteobacteria	1.3	49.3	8.7	1.0	41.8	13.0	0.6	40.5	10.8	1.6	47.4	9.0	1.1	42.7	8.3	0.5	39.9	9.6
Burkholderiales	1.0	50.3	8.3	0.7	44.9	13.6	0.4	42.4	11.1	1.1	48.8	9.0	0.8	44.1	8.5	0.3	41.3	10.1
Gammaproteobacteria	29.0	48.8	7.0	10.7	39.1	8.0	9.1	40.5	8.4	25.7	48.6	7.1	15.9	46.1	7.3	6.5	40.5	8.2
Alteromonadales	11.4	45.0	4.7	3.1	41.3	6.1	2.8	43.1	6.0	10.3	45.3	5.1	3.5	44.2	5.0	1.8	43.4	6.3
Enterobacteriales	0.4	49.4	8.6	0.2	35.6	9.1	0.2	39.8	11.0	0.6	52.3	7.5	5.7	51.4	3.9	0.1	36.4	9.2
Deltaproteobacteria	2.0	53.4	7.0	1.1	43.5	12.6	0.6	44.0	12.1	2.4	52.7	8.4	1.2	46.8	9.3	0.7	42.7	11.0
Archaea	0.2	45.8	8.7	0.6	40.2	9.4	0.6	42.3	9.2	0.3	48.1	8.0	0.4	43.1	7.6	0.6	40.9	8.5
Euryarchaeota	0.1	46.0	8.1	0.5	40.6	8.8	0.5	42.4	9.2	0.2	48.2	8.4	0.3	43.5	7.5	0.5	41.0	8.4
Thaumarchaeota	0.0	43.2	0.0	0.0	33.8	10.8	0.0	36.6	9.4	0.0	43.3	10.1	0.0	39.8	8.0	0.0	35.6	7.2
Eukaryota	0.5	52.5	8.3	1.0	41.0	13.7	0.6	45.8	13.9	0.6	49.9	11.0	0.8	42.4	11.1	0.4	42.6	14.1
Fungi	0.2	50.2	8.5	0.3	44.2	11.7	0.2	47.0	14.4	0.2	51.5	10.0	0.3	44.5	10.4	0.1	44.0	14.2
Viruses	0.3	46.8	10.5	1.8	35.1	5.2	2.7	35.3	5.1	0.7	41.6	9.6	2.4	37.6	4.4	4.4	35.6	4.8
Caudovirales	0.2	41.7	9.0	1.6	35.1	3.4	2.6	35.3	4.5	0.5	38.6	5.1	2.2	37.5	3.3	4.3	35.6	4.4

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 10 Prevalence and GC content of taxonomic groups in HOT 186 sequencing libraries at 110–500 m using `blastx`

Depth:	*110 m*						*500 m*					
Library:	*Fosmid Ends*			*454*			*Fosmid Ends*			*454*		
Taxa	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]	*% Rd.*[a]	*% GC*	*SD*[b]
Bacteria	93.9	43.4	9.7	87.3	35.3	8.6	87.9	49.4	8.6	79.5	43.5	12.2
Actinobacteria	1.0	54.3	9.2	1.8	43.3	10.5	4.6	55.8	8.7	3.1	60.4	10.7
Bacteroidetes	1.4	47.0	9.7	3.9	32.7	6.7	3.8	45.5	8.8	2.7	39.6	9.6
Flavobacteria	0.4	43.5	10.8	1.8	31.4	5.5	0.8	43.3	9.4	0.8	36.6	9.2
Cyanobacteria	39.2	36.3	5.5	22.3	34.1	5.6	1.7	45.4	9.7	2.5	36.6	9.0
Chroococcales	1.4	47.4	10.3	0.3	39.2	10.2	0.3	48.8	8.6	0.2	45.0	11.7
Prochlorales	33.1	35.4	4.2	18.4	33.3	5.2	0.6	37.7	7.7	1.7	33.1	5.9
Firmicutes	0.7	51.0	8.2	1.3	37.9	10.3	2.6	49.3	8.6	1.5	44.5	10.9
Proteobacteria	24.8	45.8	8.4	37.0	34.0	8.5	33.9	47.9	8.4	39.1	40.0	11.1
Alphaproteobacteria	10.7	43.5	7.7	23.4	31.5	6.9	8.6	47.9	9.4	19.1	35.9	10.7
Rhizobiales	1.9	46.1	8.2	1.1	37.0	9.5	2.5	51.2	7.9	1.5	45.6	11.6
Rhodobacterales	4.0	42.4	7.8	1.5	37.6	7.8	1.5	48.7	8.6	1.4	44.6	10.6
Rickettsiales	0.3	31.4	6.1	15.1	29.1	4.8	0.9	33.0	5.2	10.0	30.1	5.0
SAR11 cluster	0.3	31.2	6.2	14.9	29.1	4.8	0.9	32.9	5.1	9.8	30.1	5.0
Betaproteobacteria	1.8	43.5	9.3	0.6	41.9	10.5	2.0	51.0	7.5	1.1	47.9	10.4
Burkholderiales	1.5	42.9	9.9	0.4	43.2	10.7	1.5	51.5	7.4	0.8	49.1	10.3
Gammaproteobacteria	4.7	47.0	7.2	4.5	38.5	8.2	10.6	46.0	7.6	6.5	42.7	8.9
Alteromonadales	0.6	47.1	8.0	0.4	38.1	8.3	1.2	45.3	7.8	0.6	43.3	9.6
Enterobacteriales	0.1	47.1	9.0	0.1	36.4	10.2	0.2	49.7	7.1	0.2	43.3	11.1
Deltaproteobacteria	2.8	53.1	8.3	1.1	44.2	11.6	3.9	50.3	8.5	1.9	46.8	10.5
Archaea	0.3	48.6	9.5	1.1	41.7	8.6	4.1	44.8	10.1	13.2	37.4	6.9
Euryarchaeota	0.3	47.5	9.7	0.9	41.7	8.4	1.9	51.6	8.0	1.3	48.3	11.1
Thaumarchaeota	0.0	51.7	0.0	0.0	36.3	10.5	1.8	36.4	4.3	11.3	35.9	4.2
Eukaryota	1.2	53.1	9.2	1.2	45.7	13.5	0.7	50.6	9.0	0.4	43.8	12.1
Fungi	0.5	52.5	9.5	0.4	46.9	14.0	0.2	52.0	8.6	0.1	46.5	13.3
Viruses	0.9	41.1	8.5	5.7	35.8	5.1	0.2	45.7	10.6	0.8	36.8	6.7
Caudovirales	0.8	39.1	5.5	5.5	35.6	4.4	0.2	42.1	9.3	0.8	36.4	5.2

[a]Percentage of reads assigned to a taxon

[b]Standard deviation of the GC content for each taxon (weighted by read length)

Supplementary Table 11 KEGG orthologs biased against Fosmid Ends

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{Fosmid\ Ends/Shotgun}$[b]	$Bias_{Fosmid\ Ends/454}$[b]
K11532	fructose-1,6-bisphosphatase II / sedoheptulose-1,7-bisphosphatase [EC:3.1.3.11 3.1.3.37]	*Prochlorococcus marinus*	$-\infty$	$-\infty$
K08262	delta12-fatty acid dehydrogenase [EC:1.14.99.33]	*Prochlorococcus marinus*	−4.2	−4.7
K05575	NADH dehydrogenase I subunit 4 [EC:1.6.5.3]	*Prochlorococcus marinus*	−3.7	−4.5
K07444	putative N6-adenine-specific DNA methylase [EC:2.1.1.-]	*Prochlorococcus marinus*	−3.4	−3.7
K01971	DNA ligase (ATP) [EC:6.5.1.1]	*Prochlorococcus marinus*	−2.9	−3.7
K04040	chlorophyll synthase [EC:2.5.1.62]	*Prochlorococcus marinus*	−3.1	−3.5
K00524	ribonucleotide reductase, class II [EC:1.17.4.1]	*Prochlorococcus marinus*	−2.7	−3.2
K03403	magnesium chelatase subunit H [EC:6.6.1.1]	*Prochlorococcus marinus*	−2.2	−3.0
K06876		Bacteria	−2.4	−2.6
K07042	probable rRNA maturation factor	Bacteria	−1.8	−1.6
K03798	cell division protease FtsH [EC:3.4.24.-]	*Prochlorococcus marinus*	−1.3	−2.0
K02551	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase [EC:2.2.1.9]	*Prochlorococcus marinus*	−1.1	−1.7
K01810	glucose-6-phosphate isomerase [EC:5.3.1.9]	Bacteria	−1.3	−1.4
K02316	DNA primase [EC:2.7.7.-]	Bacteria	−1.3	−1.3
K02337	DNA polymerase III subunit alpha [EC:2.7.7.7]	Bacteria	−1.1	−1.3
K01755	argininosuccinate lyase [EC:4.3.2.1]	Bacteria	−0.8	−1.2
K01000	phospho-N-acetylmuramoyl-pentapeptide-transferase [EC:2.7.8.13]	Bacteria	−0.7	−1.0

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Supplementary Table 12 KEGG orthologs biased in favor of Fosmid Ends

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{Fosmid\ Ends/Shotgun}$[b]	$Bias_{Fosmid\ Ends/454}$[b]
K10841	DNA excision repair protein ERCC-6	Bacteria	$+\infty$	$+\infty$
K14676	lysophospholipid hydrolase [EC:3.1.1.5]	cellular organisms	$+\infty$	$+\infty$
K08672	proprotein convertase subtilisin/kexin type 6 [EC:3.4.21.-]	*Synechococcus* sp. WH 7805	5.1	$+\infty$

[a]Taxon encompassing at least half the reads assigned to each K-number in Fosmid Ends (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 12 KEGG orthologs biased in favor of Fosmid Ends – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{Fosmid\ Ends/Shotgun}$[b]	$Bias_{Fosmid\ Ends/454}$[b]
K04955	hyperpolarization activated cyclic nucleotide-gated potassium channel 2	Bacteria	3.8	$+\infty$
K04344	voltage-dependent calcium channel P/Q type alpha-1A	Bacteria	3.5	$+\infty$
K04877	potassium voltage-gated channel Shaker-related subfamily A member 4	Bacteria	2.5	$+\infty$
K13172	serine/arginine repetitive matrix protein 2	Bacteria	1.9	$+\infty$
K10955	intestinal mucin-2	cellular organisms	1.9	$+\infty$
K13171	serine/arginine repetitive matrix protein 1	Bacteria	1.8	$+\infty$
K12311	lysosomal alpha-mannosidase [EC:3.2.1.24]	*Dictyostelium discoideum* AX4	5.8	6.2
K11367	chromodomain-helicase-DNA-binding protein 1 [EC:3.6.4.12]	root	5.4	5.7
K07129		Euryarchaeota	3.8	6.7
K11238	cytokinesis protein	root	2.5	6.6
K01115	phospholipase D [EC:3.1.4.4]	root	2.2	6.5
K00078	dihydrodiol dehydrogenase / D-xylose 1-dehydrogenase (NADP) [EC:1.3.1.20 1.1.1.179]	Bacteria	3.1	5.5
K07407	alpha-galactosidase [EC:3.2.1.22]	Bacteria	3.0	5.4
K05747	Wiskott-Aldrich syndrome protein	cellular organisms	2.1	6.2
K03646	colicin import membrane protein	cellular organisms	1.6	6.7
K02566	NagD protein	Bacteria	2.5	5.5
K01183	chitinase [EC:3.2.1.14]	cellular organisms	1.4	5.9
K03626	nascent polypeptide-associated complex subunit alpha	cellular organisms	2.1	4.6
K03832	periplasmic protein TonB	Bacteria	1.8	4.5
K09992	hypothetical protein	Bacteria	2.3	3.5
K01408	insulysin [EC:3.4.24.56]	cellular organisms	2.1	3.4
K01079	phosphoserine phosphatase [EC:3.1.3.3]	cellular organisms	1.9	3.2
K01134	arylsulfatase A [EC:3.1.6.8]	Bacteria	2.0	3.0
K06896		Enterobacteriaceae	1.6	3.3
K00054	hydroxymethylglutaryl-CoA reductase [EC:1.1.1.88]	cellular organisms	2.2	2.3
K01138	uncharacterized sulfatase [EC:3.1.6.-]	Bacteria	1.6	2.8
K13787	geranylgeranyl diphosphate synthase, type I [EC:2.5.1.1 2.5.1.10 2.5.1.29]	cellular organisms	1.8	2.6
K07126		Bacteria	1.4	2.9
K06944		Archaea	2.0	2.3
K10817	erythronolide synthase [EC:2.3.1.94]	Bacteria	1.4	2.6
K02519	translation initiation factor IF-2	Bacteria	1.0	1.5

[a]Taxon encompassing at least half the reads assigned to each K-number in Fosmid Ends (all depths combined).
[b]Bias measures are calculated combining all depths.

No KEGG orthologs were significantly biased against Shotgun.

Supplementary Table 13 KEGG orthologs biased in favor of Shotgun

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{Shotgun/Fosmid\ Ends}$[b]	$Bias_{Shotgun/454}$[b]
K00737	beta-1,4-mannosyl-glycoprotein beta-1,4-N-acetylglucosaminyltransferase [EC:2.4.1.144]	Bacteria	3.6	3.4
K07276	hypothetical protein	Candidatus *Pelagibacter ubique* HTCC1062	2.4	3.9
K13497	anthranilate synthase/phosphoribosyltransferase [EC:4.1.3.27 2.4.2.18]	Candidatus *Pelagibacter ubique*	1.8	3.4
K04744	LPS-assembly protein	Alphaproteobacteria	1.3	2.7

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Supplementary Table 14 KEGG orthologs biased against 454

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K03933	chitin-binding protein	Bacteria	$-\infty$	$-\infty$
K04459	dual specificity phosphatase [EC:3.1.3.16 3.1.3.48]	root	$-\infty$	$-\infty$
K04573	neurofilament medium polypeptide (neurofilament 3)	cellular organisms	$-\infty$	$-\infty$
K05673	ATP-binding cassette, subfamily C (CFTR/MRP), member 4	Eukaryota	$-\infty$	$-\infty$
K05746	enabled	cellular organisms	$-\infty$	$-\infty$
K05996	carboxypeptidase T [EC:3.4.17.18]	Bacteria	$-\infty$	$-\infty$
K06400	site-specific DNA recombinase	Bacteria	$-\infty$	$-\infty$
K06560	mannose receptor, C type	Bacteria	$-\infty$	$-\infty$
K06867		cellular organisms	$-\infty$	$-\infty$
K06977		Bacteria	$-\infty$	$-\infty$
K07000		Gammaproteobacteria	$-\infty$	$-\infty$
K07221	outer membrane porin	Proteobacteria	$-\infty$	$-\infty$
K07579	putative methylase	Euryarchaeota	$-\infty$	$-\infty$
K08177	MFS transporter, OFA family, oxalate/formate antiporter	Bacteria	$-\infty$	$-\infty$
K08217	MFS transporter, DHA3 family, macrolide efflux protein	Bacteria	$-\infty$	$-\infty$
K08287	dual-specificity kinase [EC:2.7.12.1]	Eukaryota	$-\infty$	$-\infty$
K08827	serine/threonine-protein kinase PRP4 [EC:2.7.11.1]	Bacteria	$-\infty$	$-\infty$
K09087	hairy and enhancer of split 2/6/7	Bacteria	$-\infty$	$-\infty$
K09291	nucleoprotein TPR	cellular organisms	$-\infty$	$-\infty$
K09506	DnaJ homolog subfamily A member 5	Eukaryota	$-\infty$	$-\infty$

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K09584	protein disulfide-isomerase A6 [EC:5.3.4.1]	Eukaryota	$-\infty$	$-\infty$
K09921	hypothetical protein	Proteobacteria	$-\infty$	$-\infty$
K10297	F-box protein 11	Bacteria	$-\infty$	$-\infty$
K10385	loricrin	Bacteria	$-\infty$	$-\infty$
K00851	gluconokinase [EC:2.7.1.12]	Bacteria	$-\infty$	$-\infty$
K10421	CAP-Gly domain-containing linker protein 1	cellular organisms	$-\infty$	$-\infty$
K10595	E3 ubiquitin-protein ligase HERC2 [EC:6.3.2.19]	cellular organisms	$-\infty$	$-\infty$
K10955	intestinal mucin-2	cellular organisms	$-\infty$	$-\infty$
K11093	U1 small nuclear ribonucleoprotein 70kDa	cellular organisms	$-\infty$	$-\infty$
K11244	cell wall integrity and stress response component	cellular organisms	$-\infty$	$-\infty$
K11429	histone-lysine N-methyltransferase SUV420H [EC:2.1.1.43]	cellular organisms	$-\infty$	$-\infty$
K11494	RCC1 and BTB domain-containing protein	cellular organisms	$-\infty$	$-\infty$
K11675	Ino eighty subunit 1	cellular organisms	$-\infty$	$-\infty$
K11855	ubiquitin carboxyl-terminal hydrolase 36/42 [EC:3.1.2.15]	cellular organisms	$-\infty$	$-\infty$
K12544	S-layer protein	Bacteria	$-\infty$	$-\infty$
K12571	PAB-dependent poly(A)-specific ribonuclease subunit 2 [EC:3.1.13.4]	Eukaryota	$-\infty$	$-\infty$
K12618	5'-3' exoribonuclease 1 [EC:3.1.13.-]	Eukaryota	$-\infty$	$-\infty$
K12811	ATP-dependent RNA helicase DDX46/PRP5 [EC:3.6.4.13]	Eukaryota	$-\infty$	$-\infty$
K12821	pre-mRNA-processing factor 40	Eukaryota	$-\infty$	$-\infty$
K12893	splicing factor, arginine/serine-rich 4/5/6	cellular organisms	$-\infty$	$-\infty$
K12896	splicing factor, arginine/serine-rich 7	Bacteria	$-\infty$	$-\infty$
K13091	RNA-binding protein 39	Eukaryota	$-\infty$	$-\infty$
K13106	pre-mRNA-splicing factor CWC26	cellular organisms	$-\infty$	$-\infty$
K13165	splicing factor, arginine/serine-rich 12	Bacteria	$-\infty$	$-\infty$
K13168	splicing factor, arginine/serine-rich 16	cellular organisms	$-\infty$	$-\infty$
K13170	splicing factor, arginine/serine-rich 18	root	$-\infty$	$-\infty$
K13171	serine/arginine repetitive matrix protein 1	Bacteria	$-\infty$	$-\infty$
K13172	serine/arginine repetitive matrix protein 2	Bacteria	$-\infty$	$-\infty$

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K13173	arginine and glutamate-rich protein 1	cellular organisms	$-\infty$	$-\infty$
K13276	bacillopeptidase F [EC:3.4.21.-]	Bacteria	$-\infty$	$-\infty$
K13448	calcium-binding protein CML	cellular organisms	$-\infty$	$-\infty$
K13507	glycerol-3-phosphate O-acyltransferase / dihydroxyacetone phosphate acyltransferase [EC:2.3.1.15 2.3.1.42]	Dikarya	$-\infty$	$-\infty$
K13908	mucin-5B	cellular organisms	$-\infty$	$-\infty$
K13909	mucin-7	cellular organisms	$-\infty$	$-\infty$
K13911	basic salivary proline-rich protein 1/2	cellular organisms	$-\infty$	$-\infty$
K14059	integrase	Bacteria	$-\infty$	$-\infty$
K14297	nuclear pore complex protein Nup98-Nup96	cellular organisms	$-\infty$	$-\infty$
K14306	nuclear pore complex protein Nup62	cellular organisms	$-\infty$	$-\infty$
K14392	sodium/pantothenate symporter	Bacteria	$-\infty$	$-\infty$
K14398	cleavage and polyadenylation specificity factor subunit 6/7	Bacteria	$-\infty$	$-\infty$
K14469	acrylyl-CoA reductase (NADPH) / 3-hydroxypropionyl-CoA dehydratase / 3-hydroxypropionyl-CoA synthetase [EC:1.3.1.84 4.2.1.116 6.2.1.36]	OM60 clade	$-\infty$	$-\infty$
K00073	ureidoglycolate dehydrogenase [EC:1.1.1.154]	Bacteria	$-\infty$	$-\infty$
K01099	phosphatidylinositol-bisphosphatase [EC:3.1.3.36]	cellular organisms	$-\infty$	$-\infty$
K01120	3',5'-cyclic-nucleotide phosphodiesterase [EC:3.1.4.17]	cellular organisms	$-\infty$	$-\infty$
K01178	glucoamylase [EC:3.2.1.3]	cellular organisms	$-\infty$	$-\infty$
K01317	acrosin [EC:3.4.21.10]	Bacteria	$-\infty$	$-\infty$
K01427	urease [EC:3.5.1.5]	*Prochlorococcus marinus*	$-\infty$	$-\infty$
K01514	exopolyphosphatase [EC:3.6.1.11]	Eukaryota	$-\infty$	$-\infty$
K02184	formin 2	cellular organisms	$-\infty$	$-\infty$
K02330	DNA polymerase beta subunit [EC:2.7.7.7 4.2.99.-]	cellular organisms	$-\infty$	$-\infty$
K02331	DNA polymerase phi subunit [EC:2.7.7.7]	cellular organisms	$-\infty$	$-\infty$
K02386	flagella basal body P-ring formation protein FlgA	Bacteria	$-\infty$	$-\infty$
K02397	flagellar hook-associated protein 3 FlgL	Proteobacteria	$-\infty$	$-\infty$
K02414	flagellar hook-length control protein FliK	Bacteria	$-\infty$	$-\infty$

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K03208	colanic acid biosynthesis glycosyl transferase WcaI	Bacteria	$-\infty$	$-\infty$
K03228	type III secretion protein SctT	Proteobacteria	$-\infty$	$-\infty$
K03260	translation initiation factor eIF-4F	cellular organisms	$-\infty$	$-\infty$
K03284	metal ion transporter, MIT family	Archaea	$-\infty$	$-\infty$
K03328	polysaccharide transporter, PST family	Bacteria	$-\infty$	$-\infty$
K00404	cb-type cytochrome c oxidase subunit I [EC:1.9.3.1]	Proteobacteria	$-\infty$	$-\infty$
K14325	RNA-binding protein with serine-rich domain 1	Bacteria	−9.3	−7.7
K13100	pre-mRNA-splicing factor CWC22	cellular organisms	−7.7	−6.3
K11294	nucleolin	Eukaryota	−7.6	−6.1
K02407	flagellar hook-associated protein 2	Proteobacteria	−7.2	−6.5
K13412	calcium-dependent protein kinase [EC:2.7.11.1]	Apicomplexa	−6.3	−6.4
K11901	type VI secretion system protein ImpB	Proteobacteria	−7.1	−5.5
K03932	polyhydroxybutyrate depolymerase	Bacteria	−6.8	−5.7
K09888	cell division protein ZapA	Candidatus *Pelagibacter ubique* HTCC1062	−5.7	−6.4
K12287	MSHA biogenesis protein MshQ	Bacteria	−6.0	−6.1
K03646	colicin import membrane protein	Bacteria	−6.7	−5.1
K07279	hypothetical protein	Proteobacteria	−6.4	−5.4
K13092	nuclear protein NHN1	cellular organisms	−6.6	−5.1
K06237	collagen, type IV, alpha	cellular organisms	−6.2	−5.4
K07288	uncharacterized membrane protein	Proteobacteria	−6.2	−5.3
K06907		unclassified T4-like viruses	−5.7	−5.7
K07029		Bacteria	−6.5	−4.9
K10380	ankyrin	cellular organisms	−6.7	−4.5
K04034	anaerobic magnesium-protoporphyrin IX monomethyl ester cyclase [EC:4.-.-.-]	Bacteria	−5.6	−5.4
K10352	myosin heavy chain	cellular organisms	−6.3	−4.6
K11941	glucans biosynthesis protein C [EC:2.1.-.-]	Bacteria	−6.2	−4.7
K05315	voltage-dependent calcium channel alpha 1, invertebrate	Bacteria	−6.5	−4.4
K11238	cytokinesis protein	cellular organisms	−6.6	−4.2
K08234	glyoxylase I family protein	Bacteria	−6.2	−4.5
K08819	Cdc2-related kinase, arginine/serine-rich [EC:2.7.11.22]	cellular organisms	−6.2	−4.4

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K07313	serine/threonine protein phosphatase 1 [EC:3.1.3.16]	Bacteria	−6.2	−4.4
K01975	2'-5' RNA ligase [EC:6.5.1.-]	cellular organisms	−5.2	−5.3
K01183	chitinase [EC:3.2.1.14]	Bacteria	−5.9	−4.4
K09607	immune inhibitor A [EC:3.4.24.-]	Bacteria	−6.2	−4.2
K01281	X-Pro dipeptidyl-peptidase [EC:3.4.14.11]	Actinomycetales	−6.4	−3.8
K05747	Wiskott-Aldrich syndrome protein	cellular organisms	−6.2	−4.1
K10908	DNA-directed RNA polymerase, mitochondrial [EC:2.7.7.6]	cellular organisms	−5.9	−4.4
K12567	titin [EC:2.7.11.1]	cellular organisms	−6.3	−3.9
K11653	AT-rich interactive domain-containing protein 1	cellular organisms	−5.7	−4.4
K03651	Icc protein	Gammaproteobacteria	−5.7	−4.3
K08296	phosphohistidine phosphatase [EC:3.1.3.-]	Bacteria	−5.5	−4.4
K14376	poly(A) polymerase [EC:2.7.7.19]	Eukaryota	−6.1	−3.8
K00689	dextransucrase [EC:2.4.1.5]	Bacteria	−5.4	−4.5
K08153	MFS transporter, DHA1 family, multidrug resistance protein	Bacteria	−6.2	−3.5
K01617	4-oxalocrotonate decarboxylase [EC:4.1.1.77]	Rhizobiales	−4.7	−4.9
K08720	outer membrane protein OmpU	Candidatus *Pelagibacter ubique*	−4.2	−5.5
K07141		Proteobacteria	−4.9	−4.6
K01781	mandelate racemase [EC:5.1.2.2]	Proteobacteria	−5.2	−4.3
K13668	phosphatidylinositol alpha-1,6-mannosyltransferase [EC:2.4.1.-]	Candidatus *Pelagibacter ubique*	−4.5	−4.9
K05303	macrocin O-methyltransferase [EC:2.1.1.101]	Bacteria	−4.7	−4.6
K09811	cell division transport system permease protein	Bacteria	−5.3	−4.0
K10799	tankyrase [EC:2.4.2.30]	cellular organisms	−5.2	−4.1
K09252	feruloyl esterase [EC:3.1.1.73]	Proteobacteria	−5.5	−3.7
K08589	gingipain R [EC:3.4.22.37]	*Porphyromonas gingivalis*	−4.9	−4.2
K02183	calmodulin	Eukaryota	−5.2	−3.9
K13735	adhesin/invasin	Bacteria	−4.2	−4.9
K01055	3-oxoadipate enol-lactonase [EC:3.1.1.24]	Proteobacteria	−4.9	−4.0
K12600	superkiller protein 3	Bacteria	−4.7	−4.1
K09774	lipopolysaccharide export system protein LptA	Proteobacteria	−5.0	−3.7
K06636	structural maintenance of chromosome 1	Eukaryota	−4.7	−3.9
K07243	high-affinity iron transporter	Bacteria	−4.7	−3.8
K07404	6-phosphogluconolactonase [EC:3.1.1.31]	Bacteria	−4.7	−3.5

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K01174	micrococcal nuclease [EC:3.1.31.1]	cellular organisms	−4.2	−4.1
K03571	rod shape-determining protein MreD	Alphaproteobacteria	−3.7	−4.5
K00949	thiamine pyrophosphokinase [EC:2.7.6.2]	Bacteria	−5.0	−3.2
K14572	midasin	cellular organisms	−4.4	−3.8
K02315	DNA replication protein DnaC	Bacteria	−3.7	−4.4
K14645	serine protease [EC:3.4.21.-]	Bacteria	−4.6	−3.4
K02422	flagellar protein FliS	Bacteria	−4.6	−3.5
K00441	coenzyme F420 hydrogenase beta subunit [EC:1.12.98.1]	Bacteria	−4.2	−3.9
K11647	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 2/4 [EC:3.6.4.-]	Eukaryota	−4.2	−3.9
K02843	heptosyltransferase II [EC:2.4.-.-]	Proteobacteria	−4.2	−3.9
K12603	CCR4-NOT transcription complex subunit 6 [EC:3.1.-.-]	cellular organisms	−4.5	−3.4
K01406	serralysin [EC:3.4.24.40]	Bacteria	−4.3	−3.6
K02396	flagellar hook-associated protein 1 FlgK	Proteobacteria	−4.3	−3.5
K03287	outer membrane factor, OMF family	*Prochlorococcus marinus*	−3.6	−4.2
K02406	flagellin	Bacteria	−4.5	−3.3
K06994	putative drug exporter of the RND superfamily	Bacteria	−5.0	−2.6
K07004		Bacteria	−4.7	−3.0
K09790	hypothetical protein	Bacteria	−4.4	−3.1
K03546	exonuclease SbcC	cellular organisms	−3.9	−3.5
K01090	protein phosphatase [EC:3.1.3.16]	cellular organisms	−4.6	−2.9
K10395	kinesin family member 4/7/21/27	Eukaryota	−4.5	−2.9
K01453		Proteobacteria	−4.0	−3.4
K01567		Bacteria	−4.2	−3.1
K00231	protoporphyrinogen oxidase [EC:1.3.3.4]	Bacteria	−4.2	−3.1
K03832	periplasmic protein TonB	Bacteria	−4.5	−2.7
K10592	E3 ubiquitin-protein ligase HUWE1 [EC:6.3.2.19]	Eukaryota	−3.9	−3.2
K14822	rRNA-processing protein CGR1	cellular organisms	−3.9	−3.2
K03626	nascent polypeptide-associated complex subunit alpha	cellular organisms	−4.6	−2.5
K03770	peptidyl-prolyl cis-trans isomerase D [EC:5.2.1.8]	Proteobacteria	−3.0	−4.1
K00908	Ca2+/calmodulin-dependent protein kinase [EC:2.7.11.17]	Eukaryota	−3.9	−3.0
K13730	internalin A	Bacteria	−3.8	−3.1

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K08151	MFS transporter, DHA1 family, tetracycline resistance protein	Bacteria	−4.2	−2.7
K00477	phytanoyl-CoA hydroxylase [EC:1.14.11.18]	Proteobacteria	−3.8	−2.9
K03726	helicase [EC:3.6.4.-]	*Nitrosopumilus maritimus* SCM1	−4.0	−2.8
K03454	mitochondrial carrier protein, MC family	Dikarya	−3.7	−3.0
K07403	membrane-bound serine protease (ClpP class)	Bacteria	−3.9	−2.8
K09228	KRAB domain-containing zinc finger protein	Dikarya	−3.3	−3.1
K13007	Fuc2NAc and GlcNAc transferase [EC:2.4.1.-]	Cyanobacteria	−3.2	−3.2
K07552	MFS transporter, DHA1 family, bicyclomycin/chloramphenicol resistance protein	Proteobacteria	−3.5	−2.8
K01684	galactonate dehydratase [EC:4.2.1.6]	Bacteria	−3.7	−2.5
K01004	phosphatidylcholine synthase [EC:2.7.8.24]	Proteobacteria	−3.7	−2.4
K03446	MFS transporter, DHA2 family, multidrug resistance protein B	Proteobacteria	−3.6	−2.6
K02841	heptosyltransferase I [EC:2.4.-.-]	Proteobacteria	−2.6	−3.6
K08857	NIMA (never in mitosis gene a)-related kinase [EC:2.7.11.1]	Eukaryota	−3.7	−2.4
K00087	xanthine dehydrogenase molybdenum-binding subunit [EC:1.17.1.4]	Bacteria	−3.6	−2.4
K06889		Proteobacteria	−2.9	−3.2
K06919	putative DNA primase/helicase	Bacteria	−3.5	−2.5
K03769	peptidyl-prolyl cis-trans isomerase C [EC:5.2.1.8]	Proteobacteria	−3.3	−2.6
K01238		Bacteroidetes	−3.0	−2.9
K08884	serine/threonine protein kinase, bacterial [EC:2.7.11.1]	Bacteria	−3.5	−2.4
K07068		Bacteria	−3.5	−2.3
K00982	glutamate-ammonia-ligase adenylyltransferase [EC:2.7.7.42]	Proteobacteria	−3.6	−2.1
K09800	hypothetical protein	*Prochlorococcus marinus*	−3.1	−2.5
K07001		Bacteria	−3.2	−2.4
K00924		Bacteria	−3.4	−2.1
K01467	beta-lactamase [EC:3.5.2.6]	Bacteria	−3.1	−2.4
K12960	5-methylthioadenosine/S-adenosylhomocysteine deaminase [EC:3.5.4.- 3.5.4.28]	Bacteria	−3.2	−2.3
K01133	choline-sulfatase [EC:3.1.6.6]	Proteobacteria	−3.5	−1.9
K05882	aryl-alcohol dehydrogenase (NADP+) [EC:1.1.1.91]	Proteobacteria	−3.0	−2.4

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K07263	zinc protease [EC:3.4.99.-]	Bacteria	−3.2	−2.1
K07152		Proteobacteria	−2.9	−2.4
K03006	DNA-directed RNA polymerase II subunit RPB1 [EC:2.7.7.6]	Eukaryota	−2.8	−2.4
K02014	iron complex outermembrane recepter protein	Proteobacteria	−2.8	−2.3
K03406	methyl-accepting chemotaxis protein	Proteobacteria	−2.6	−2.3
K03927	carboxylesterase type B [EC:3.1.1.1]	Proteobacteria	−2.9	−2.0
K07091	lipopolysaccharide export system permease protein	Candidatus *Pelagibacter ubique*	−1.9	−3.0
K01126	glycerophosphoryl diester phosphodiesterase [EC:3.1.4.46]	Bacteria	−2.5	−2.3
K06236	collagen, type I/II/III/V/XI, alpha	cellular organisms	−2.6	−2.2
K12823	ATP-dependent RNA helicase DDX5/DBP2 [EC:3.6.4.13]	Eukaryota	−3.4	−1.4
K08680	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase [EC:4.2.99.20]	Bacteria	−2.4	−2.3
K00983	N-acylneuraminate cytidylyltransferase [EC:2.7.7.43]	Bacteria	−2.0	−2.7
K10408	dynein heavy chain, axonemal	Trypanosomatidae	−2.9	−1.7
K06178	ribosomal large subunit pseudouridine synthase B [EC:5.4.99.12]	Proteobacteria	−2.8	−1.8
K00754		Bacteria	−1.9	−2.7
K08282	non-specific serine/threonine protein kinase [EC:2.7.11.1]	cellular organisms	−2.6	−1.9
K01113	alkaline phosphatase D [EC:3.1.3.1]	Bacteria	−2.6	−1.9
K07025	putative hydrolase of the HAD superfamily	Proteobacteria	−2.7	−1.7
K07126		Bacteria	−2.9	−1.5
K00077	2-dehydropantoate 2-reductase [EC:1.1.1.169]	Bacteria	−2.6	−1.8
K07011		Bacteria	−2.1	−2.2
K07024		Bacteria	−2.7	−1.6
K00849	galactokinase [EC:2.7.1.6]	Bacteria	−2.6	−1.7
K03771	peptidyl-prolyl cis-trans isomerase SurA [EC:5.2.1.8]	Proteobacteria	−2.0	−2.3
K01768	adenylate cyclase [EC:4.6.1.1]	Proteobacteria	−2.3	−1.9
K00320	coenzyme F420-dependent N5,N10-methenyltetrahydromethanopterin reductase [EC:1.5.99.11]	Bacteria	−2.5	−1.8
K01136	iduronate 2-sulfatase [EC:3.1.6.13]	Bacteria	−2.9	−1.3
K01179	endoglucanase [EC:3.2.1.4]	Bacteria	−2.5	−1.7

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

Supplementary Table 14 KEGG orthologs biased against 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K08776	puromycin-sensitive aminopeptidase [EC:3.4.11.-]	*Nitrosopumilus maritimus* SCM1	−2.7	−1.5
K03833	selenocysteine-specific elongation factor	Bacteria	−2.2	−1.9
K04744	LPS-assembly protein	Alphaproteobacteria	−1.4	−2.7
K04763	integrase/recombinase XerD	Bacteria	−2.6	−1.4
K07031		Bacteria	−2.4	−1.5
K02674	type IV pilus assembly protein PilY1	Candidatus *Pelagibacter ubique*	−1.9	−1.9
K07749	formyl-CoA transferase [EC:2.8.3.16]	Proteobacteria	−2.4	−1.4
K00067	dTDP-4-dehydrorhamnose reductase [EC:1.1.1.133]	Bacteria	−1.7	−1.9
K00517		Proteobacteria	−2.1	−1.5
K11720	lipopolysaccharide export system permease protein	Candidatus *Pelagibacter ubique* HTCC1062	−1.0	−2.5
K00837		Bacteria	−2.2	−1.4
K03307	solute:Na+ symporter, SSS family	Bacteria	−2.2	−1.3
K00599		Bacteria	−1.9	−1.6
K01130	arylsulfatase [EC:3.1.6.1]	Bacteria	−2.5	−0.8
K06148	ATP-binding cassette, subfamily C, bacterial	Bacteria	−1.6	−1.7
K00912	tetraacyldisaccharide 4'-kinase [EC:2.7.1.130]	Alphaproteobacteria	−1.5	−1.8
K07124		Bacteria	−1.5	−1.7
K00100		Proteobacteria	−1.9	−1.3
K09815	zinc transport system substrate-binding protein	Proteobacteria	−1.5	−1.6
K02319	DNA polymerase I [EC:2.7.7.7]	T4-like viruses	−1.7	−1.2
K01692	enoyl-CoA hydratase [EC:4.2.1.17]	Proteobacteria	−1.9	−1.0
K02517	lipid A biosynthesis lauroyl acyltransferase [EC:2.3.1.-]	Proteobacteria	−1.0	−1.8
K03750	molybdopterin biosynthesis protein MoeA	Proteobacteria	−1.5	−1.2
K00979	3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) [EC:2.7.7.38]	Proteobacteria	−1.1	−1.5
K06902	MFS transporter, UMF1 family	Bacteria	−1.6	−1.0
K01426	amidase [EC:3.5.1.4]	Proteobacteria	−1.6	−0.7
K07003		Proteobacteria	−1.4	−0.8
K02519	translation initiation factor IF-2	Bacteria	−1.5	−0.5
K00059	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	Bacteria	−1.1	−0.7
K00540		Bacteria	−0.5	−0.7

[a]Taxon encompassing at least half the reads assigned to each K-number in Shotgun (all depths combined).
[b]Bias measures are calculated combining all depths.

Supplementary Table 15 KEGG orthologs biased in favor of 454

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K01429	urease subunit beta [EC:3.5.1.5]	*Prochlorococcus marinus*	$+\infty$	3.2
K00777		Candidatus *Pelagibacter ubique*	3.5	2.3
K08641	D-alanyl-D-alanine dipeptidase [EC:3.4.13.-]	*Prochlorococcus marinus*	3.1	1.2
K05578	NADH dehydrogenase I subunit 6 [EC:1.6.5.3]	*Prochlorococcus marinus*	2.9	1.3
K02705	photosystem II CP43 chlorophyll apoprotein	*Prochlorococcus marinus*	2.4	1.5
K07769	two-component system, OmpR family, sensor histidine kinase NblS [EC:2.7.13.3]	*Prochlorococcus marinus*	2.9	1.0
K00304	sarcosine oxidase, subunit delta [EC:1.5.3.1]	Candidatus *Pelagibacter ubique* HTCC1062	1.6	2.1
K01428	urease subunit alpha [EC:3.5.1.5]	Bacteria	2.3	0.9
K00394	adenylylsulfate reductase, subunit A [EC:1.8.99.2]	Candidatus *Pelagibacter ubique* HTCC1062	2.1	1.0
K00412	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]	Alphaproteobacteria	2.0	0.8
K03798	cell division protease FtsH [EC:3.4.24.-]	Cyanobacteria	2.0	0.7
K01696	tryptophan synthase beta chain [EC:4.2.1.20]	Bacteria	1.8	0.8
K01662	1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]	Bacteria	1.9	0.5
K02690	photosystem I P700 chlorophyll a apoprotein A2	Cyanobacteria	1.3	1.0
K02111	F-type H+-transporting ATPase subunit alpha [EC:3.6.3.14]	Bacteria	1.3	0.8
K04043	molecular chaperone DnaK	Bacteria	1.5	0.6
K04077	chaperonin GroEL	Bacteria	1.5	0.5
K02274	cytochrome c oxidase subunit I [EC:1.9.3.1]	Bacteria	0.9	1.0
K00962	polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	Bacteria	1.3	0.6
K06207	GTP-binding protein	Bacteria	1.3	0.6
K03046	DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	Bacteria	1.2	0.7
K02469	DNA gyrase subunit A [EC:5.99.1.3]	Cyanobacteria	1.2	0.6
K03070	preprotein translocase subunit SecA	Bacteria	1.3	0.5
K01937	CTP synthase [EC:6.3.4.2]	Bacteria	1.2	0.6
K01740	O-acetylhomoserine (thiol)-lyase [EC:2.5.1.49]	Bacteria	1.2	0.6
K03569	rod shape-determining protein MreB and related proteins	Bacteria	1.0	0.7

[a]Taxon encompassing at least half the reads assigned to each K-number in 454 (all depths combined).
[b]Bias measures are calculated combining all depths.

Continued on next page

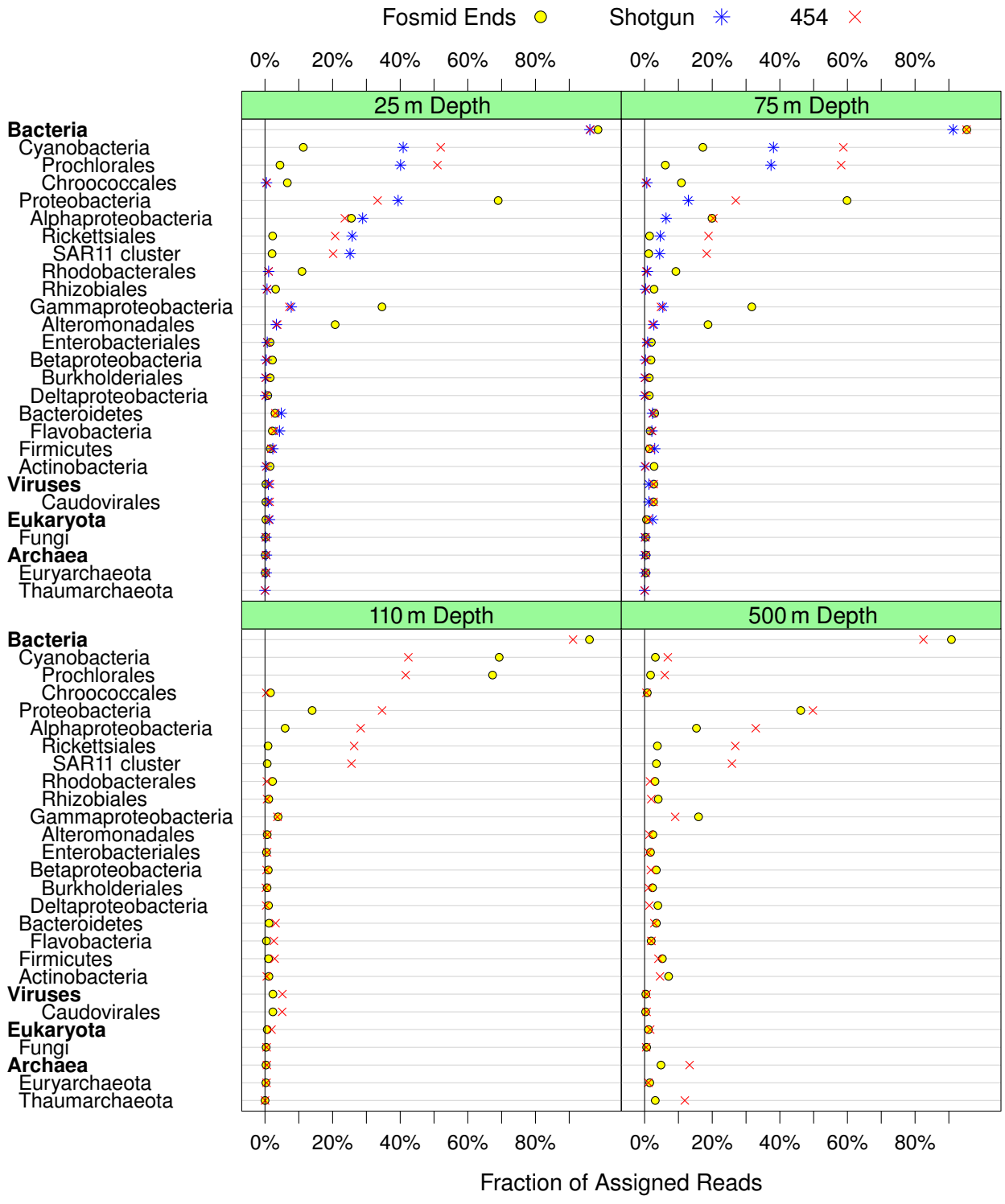
Supplementary Table 15 KEGG orthologs biased in favor of 454 – continued

K-Number	*Function*	*Consensus Taxon*[a]	$Bias_{454/Fosmid\ Ends}$[b]	$Bias_{454/Shotgun}$[b]
K01955	carbamoyl-phosphate synthase large subunit [EC:6.3.5.5]	Bacteria	1.1	0.6
K01649	2-isopropylmalate synthase [EC:2.3.3.13]	Bacteria	1.1	0.5
K01895	acetyl-CoA synthetase [EC:6.2.1.1]	Bacteria	1.1	0.5
K03703	excinuclease ABC subunit C	Cyanobacteria	1.1	0.5
K00284	glutamate synthase (ferredoxin) [EC:1.4.7.1]	*Prochlorococcus marinus*	0.9	0.7
K02358	elongation factor EF-Tu [EC:3.6.5.3]	Bacteria	0.9	0.7
K03043	DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	Bacteria	1.0	0.6
K00281	glycine dehydrogenase [EC:1.4.4.2]	Bacteria	0.9	0.7
K03702	excinuclease ABC subunit B	Bacteria	0.9	0.6
K02112	F-type H+-transporting ATPase subunit beta [EC:3.6.3.14]	Bacteria	0.8	0.7
K03701	excinuclease ABC subunit A	Bacteria	0.8	0.6
K02355	elongation factor EF-G [EC:3.6.5.3]	Bacteria	0.6	0.8
K01687	dihydroxy-acid dehydratase [EC:4.2.1.9]	Bacteria	0.8	0.5
K01873	valyl-tRNA synthetase [EC:6.1.1.9]	Bacteria	0.7	0.6
K01338	ATP-dependent Lon protease [EC:3.4.21.53]	Proteobacteria	0.7	0.6

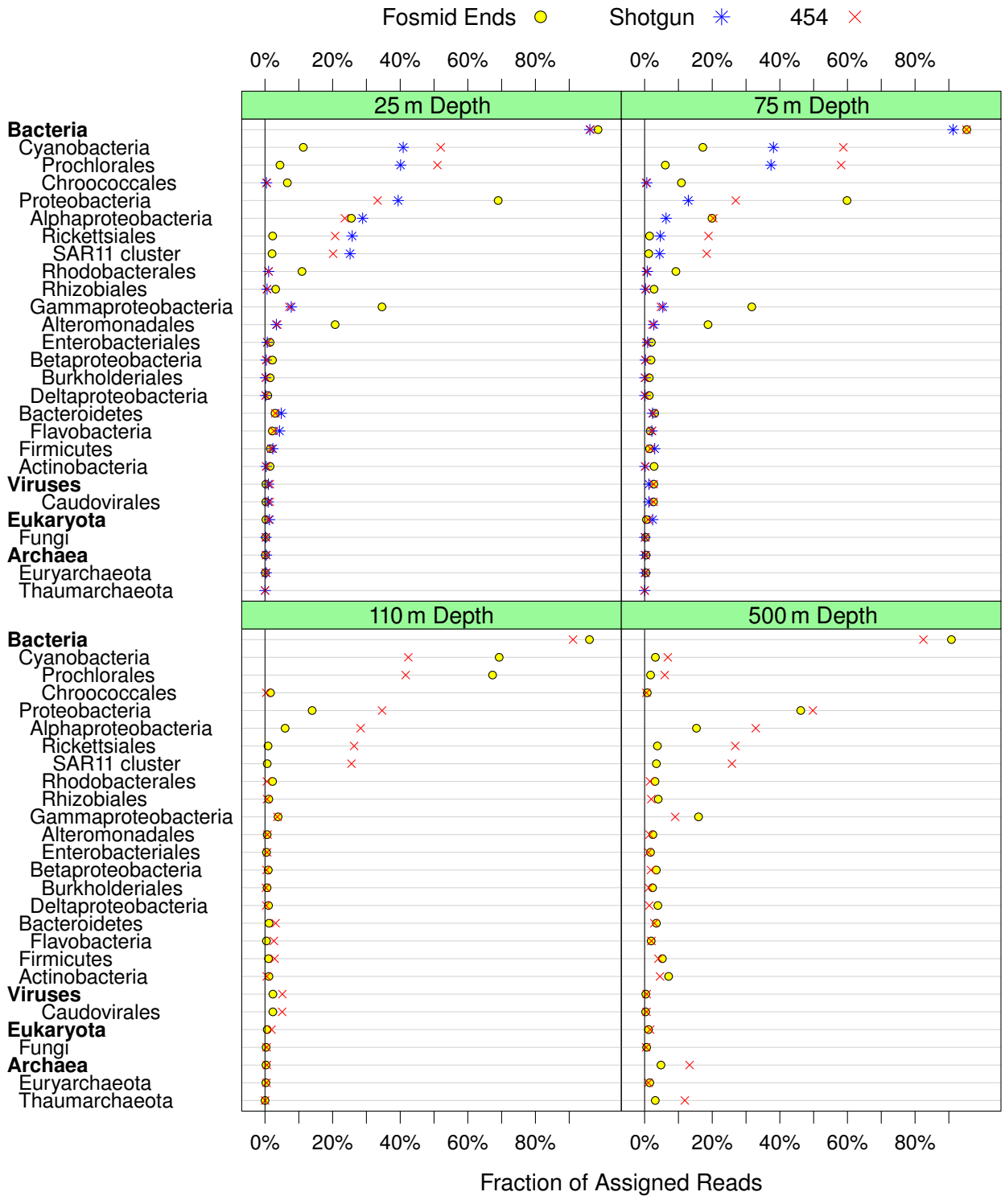
[a]Taxon encompassing at least half the reads assigned to each K-number in 454 (all depths combined).
[b]Bias measures are calculated combining all depths.

Figures

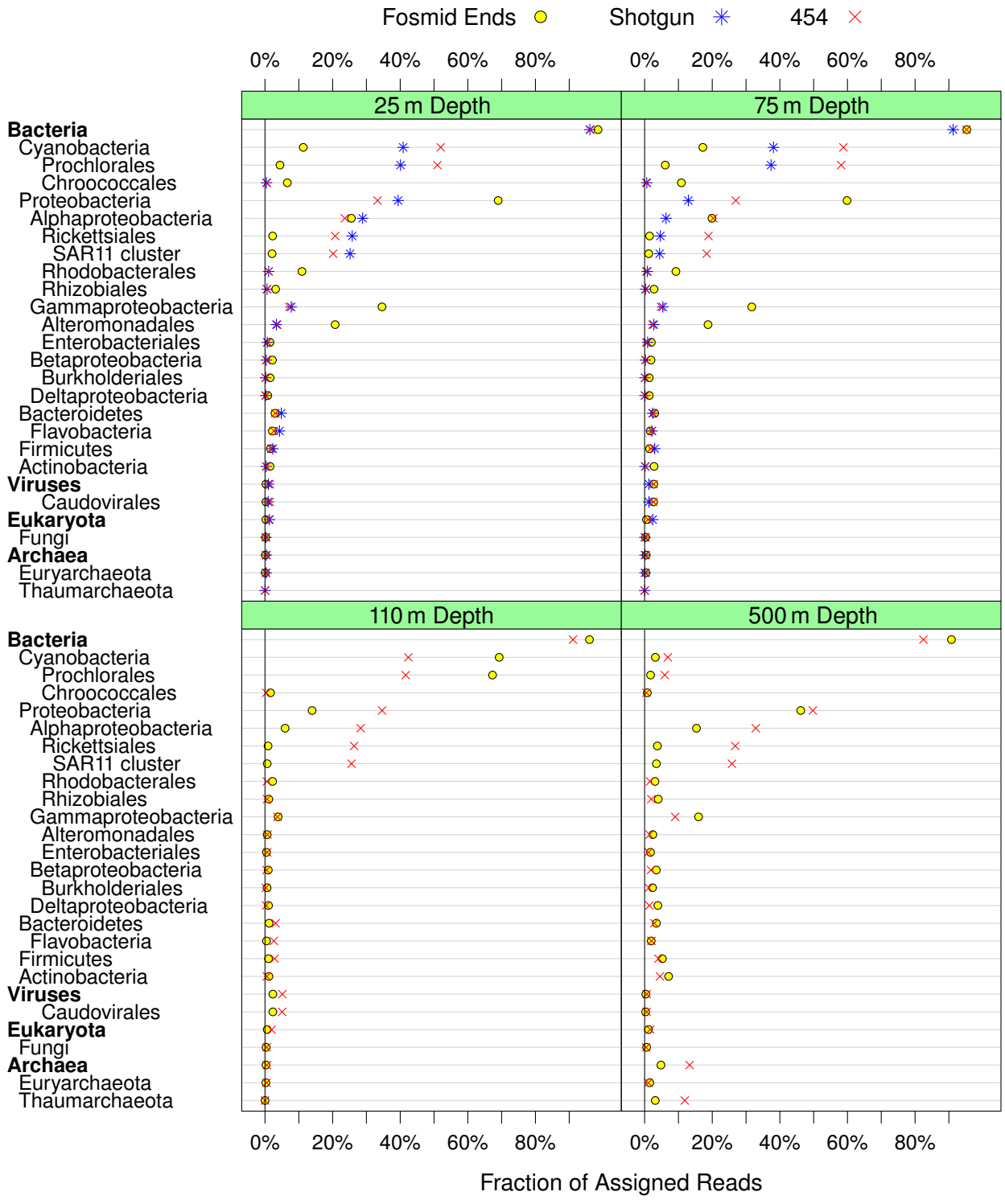
List of Figures



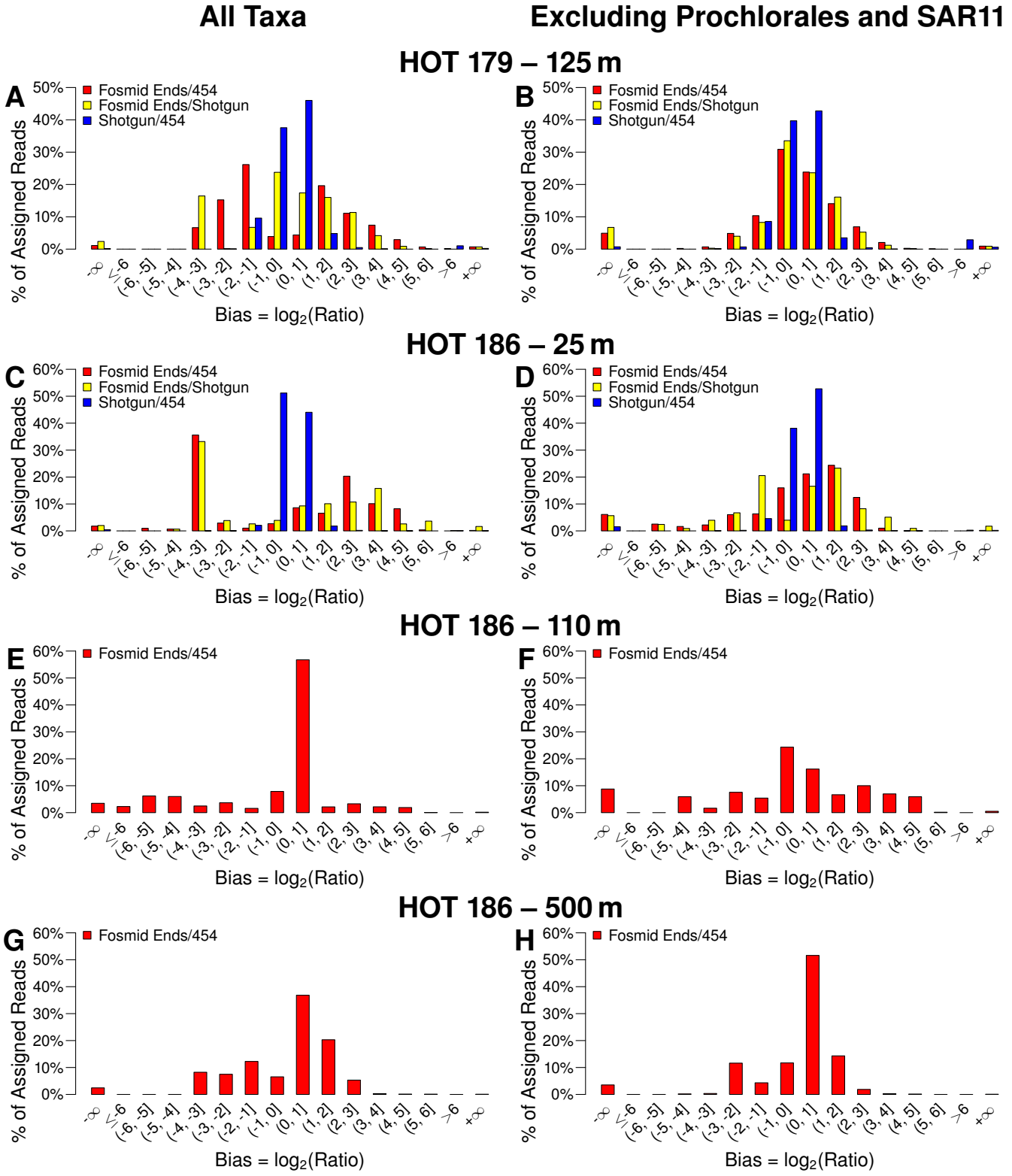
Supplementary Figure 1 Taxonomic comparison of libraries in HOT 179, analyzed using `blastx`. A representative subset of taxa is shown and their indentations reflect the hierarchical rank in the NCBI taxonomy. The number of reads assigned to each taxon, including its child taxa, is displayed as percentage of all successfully assigned reads in the library.



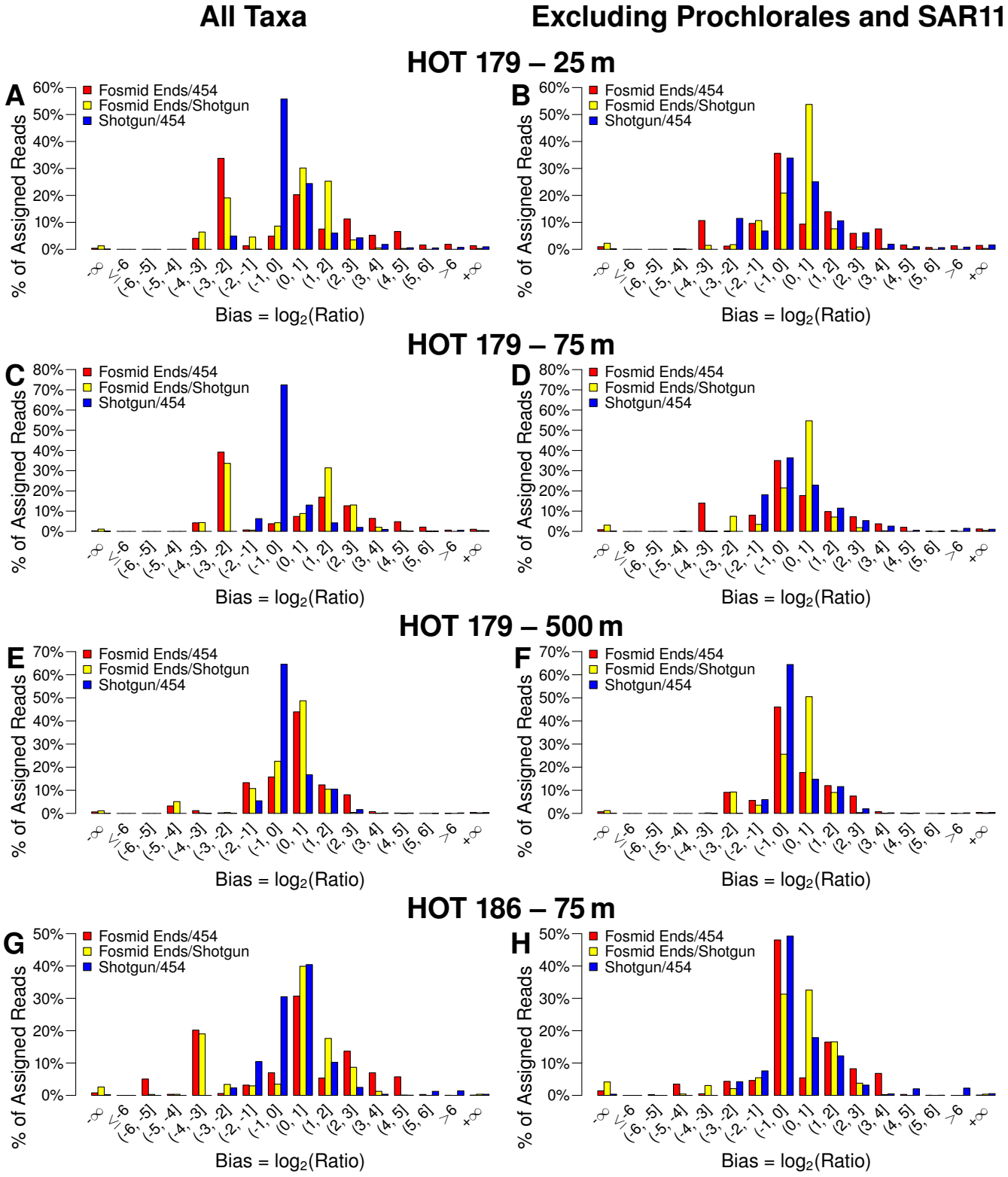
Supplementary Figure 2 Taxonomic comparison of libraries in HOT 186, analyzed using `blastn`. A representative subset of taxa is shown and their indentations reflect the hierarchical rank in the NCBI taxonomy. The number of reads assigned to each taxon, including its child taxa, is displayed as percentage of all successfully assigned reads in the library.



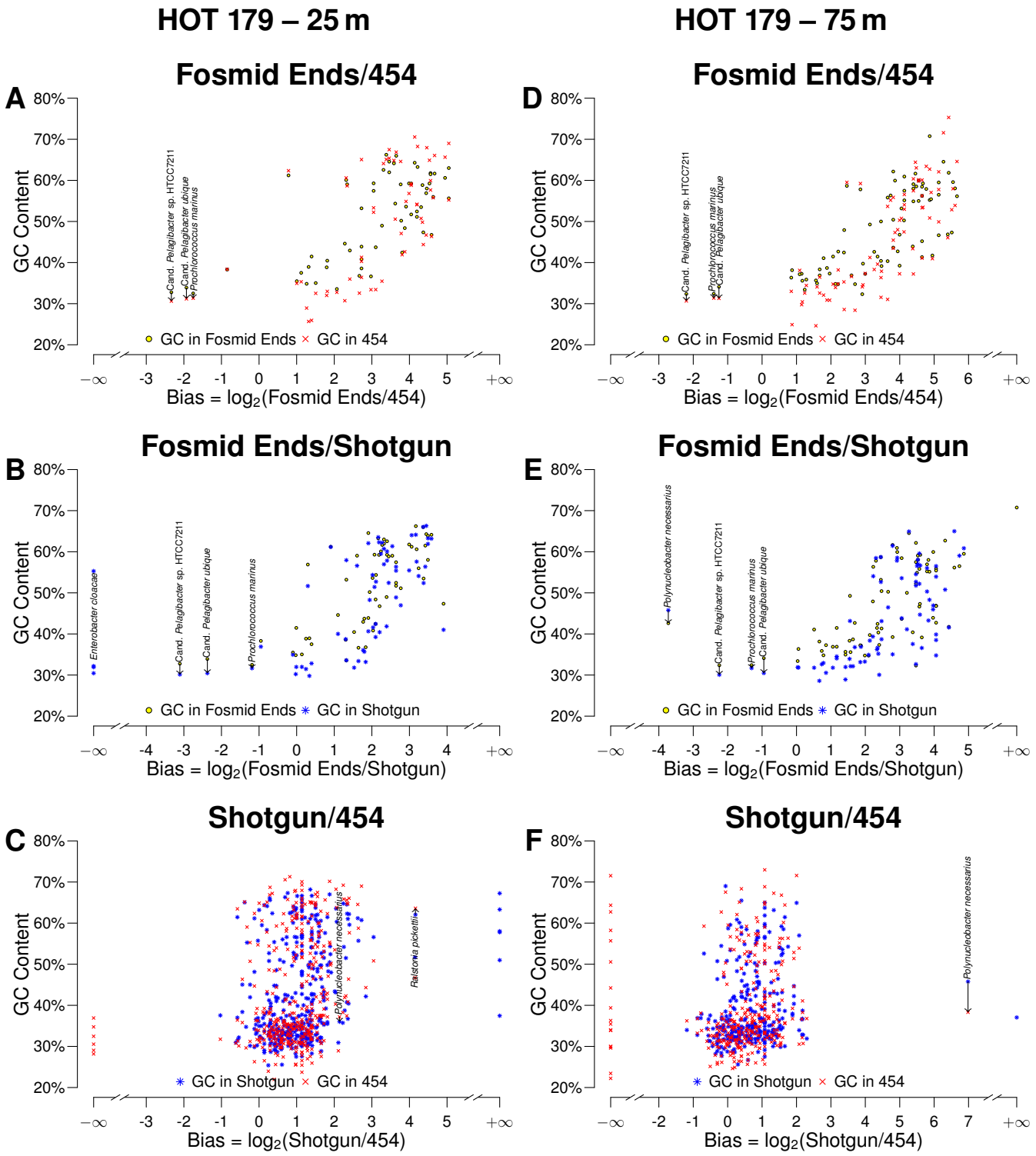
Supplementary Figure 3 Taxonomic comparison of libraries in HOT 186, analyzed using `blastx`. A representative subset of taxa is shown and their indentations reflect the hierarchical rank in the NCBI taxonomy. The number of reads assigned to each taxon, including its child taxa, is displayed as percentage of all successfully assigned reads in the library.



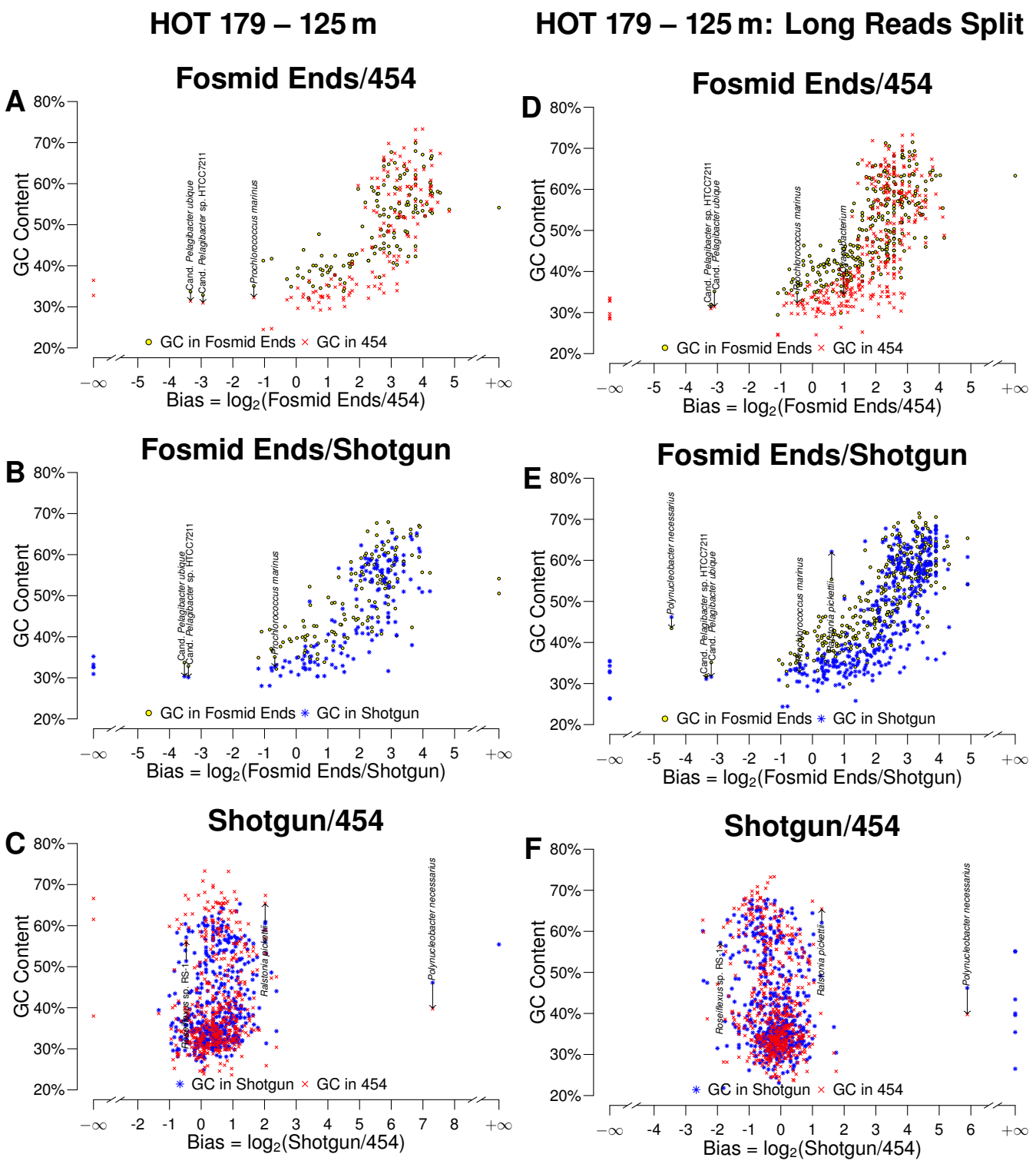
Supplementary Figure 4 Distribution of the bias measure in pairwise comparisons of library types, using `blastn` analysis results. Positive and negative infinity of the bias reflect the absence of some taxa in one of the libraries. The colors of the histogram bars are indicative of the library types being compared. The y-axis shows the percentage of reads assigned to taxa falling into the respective bias interval, averaged over both libraries.



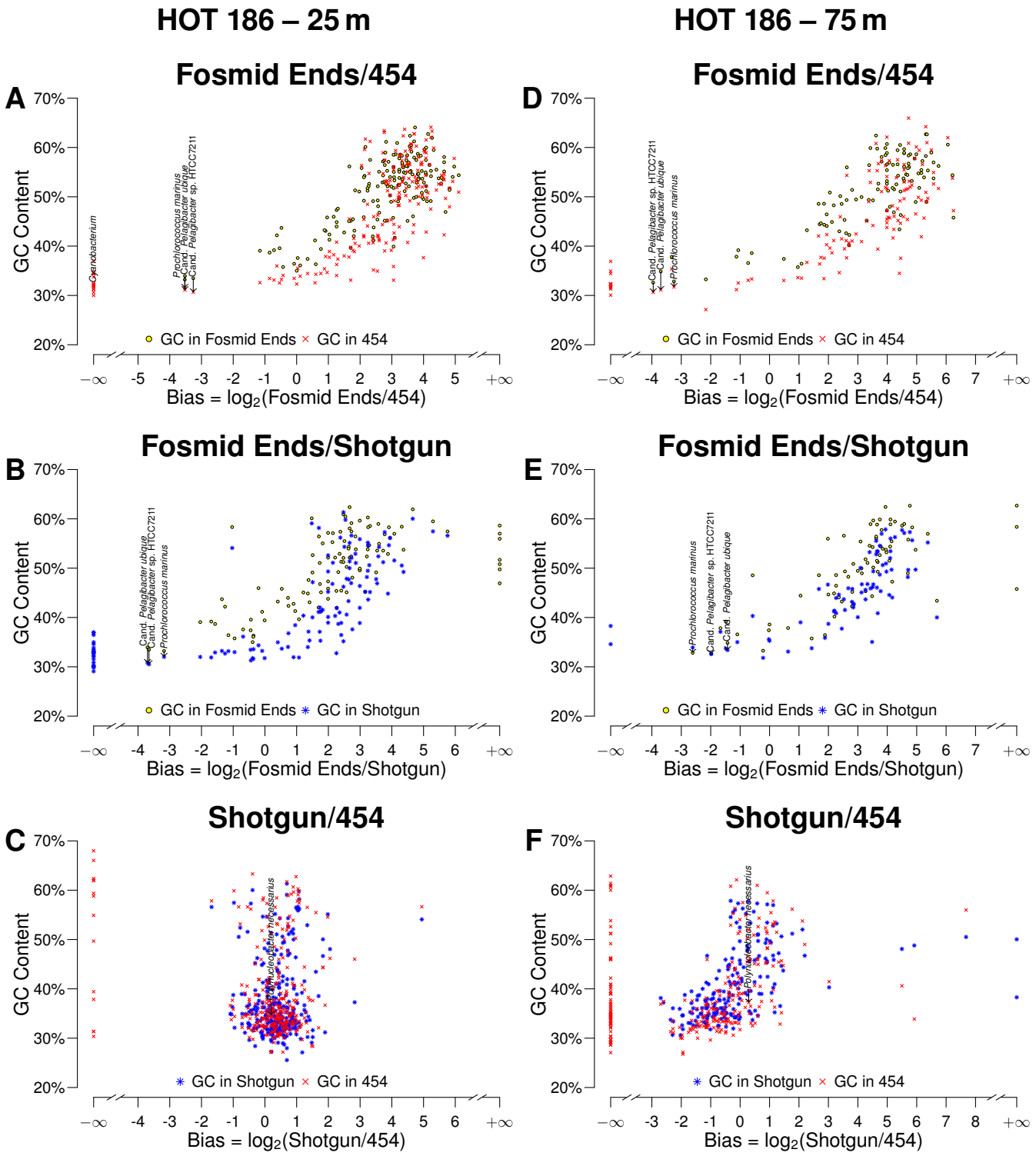
Supplementary Figure 5 Distribution of the bias measure in pairwise comparisons of library types, using `blastx` analysis results. Positive and negative infinity of the bias reflect the absence of some taxa in one of the libraries. The colors of the histogram bars are indicative of the library types being compared. The y-axis shows the percentage of reads assigned to taxa falling into the respective bias interval, averaged over both libraries.



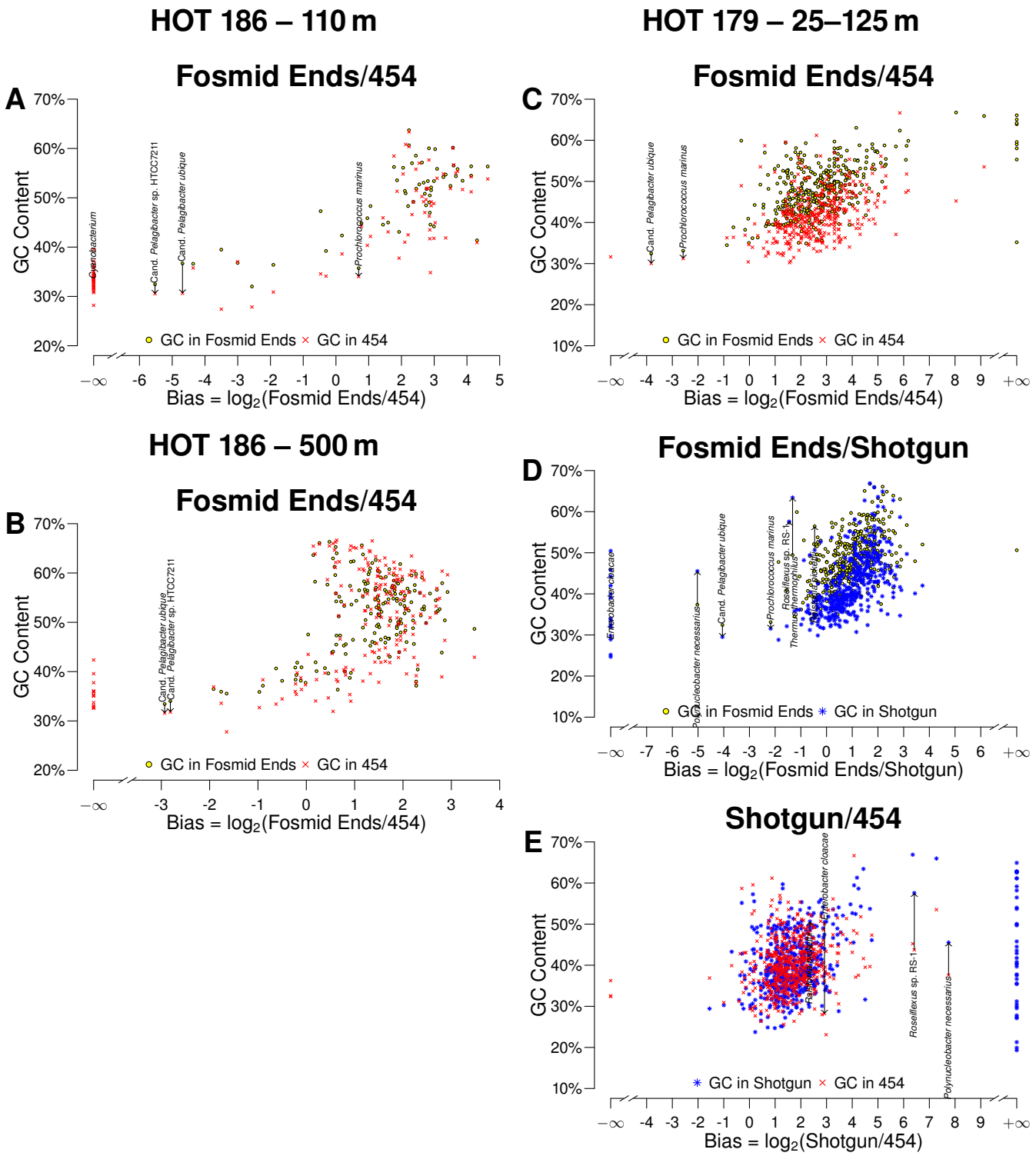
Supplementary Figure 6 GC contents plotted versus bias for pairwise library comparisons of HOT 179 25 m and 75 m `blastn` results. Each taxon is represented by two symbols, one for the GC content in each library. Taxa with less than 5 reads in each library are only shown if they are not represented at all in one library (infinite bias measure), where their expected representation based on their percentage in the other library would be at least 2 reads.



Supplementary Figure 7 GC contents plotted versus bias for pairwise library comparisons of HOT 179–125 m `blastn` results. Panels A–C show the original data. In D–F the fosmid and shotgun sequences were randomly split before the analysis, to yield a length distribution of with a mean of 108.8 nt (equivalent to 454 sequences). Each taxon is represented by two symbols, one for the GC content in each library. Taxa with less than 5 reads in each library are only shown if they are not represented at all in one library (infinite bias measure), where their expected representation based on their percentage in the other library would be at least 2 reads. In D–F the split yielded multiple BLAST hits for many original reads, resulting in more taxa with read numbers above the cutoff.



Supplementary Figure 8 GC contents plotted versus bias for pairwise library comparisons, for HOT 186–25 m and 75 m `blastn` results. Each taxon is represented by two symbols, one for the GC content in each library. Taxa with less than 5 reads in each library are only shown if they are not represented at all in one library (infinite bias measure), where their expected representation based on their percentage in the other library would be at least 2 reads.



Supplementary Figure 9 GC contents plotted versus bias for pairwise library comparisons. Panels A and B show the `blastn` results for HOT 186, 110 m and 500 m, respectively. C–E show the combined photic zone of HOT 179, using `blastx` data. Each taxon is represented by two symbols, one for the GC content in each library. Taxa with less than 5 reads in each library are only shown if they are not represented at all in one library (infinite bias measure), where their expected representation based on their percentage in the other library would be at least 2 reads.