Supplementary Material for

Diversity and abundance of single-stranded DNA viruses in human faeces

Min-Soo Kim, Eun-Jin Park, Seong Woon Roh and Jin-Woo Bae*

*To whom correspondence should be addressed. E-mail: baejw@khu.ac.kr

This PDF file includes:

Figs. S1 to S3

Tables S1 to S4

Table S1. Features of participants and samples.

Individual	Sample ID	Birth Country	Age	Sex	Sample amount
Individual 1	F-A		28	Male	12.4g
Individual 2	F-B		29	Male	15.2g
Individual 3	F-C	South Korea	28	Male	14.4g
Individual 4	F-D		23	Female	14.1g
Individual 5	F-E		27	Female	12.4g

Table S2. The BLAST profile of five viromes comparing with three public databases, the SEED nr, CAMERA nr and CAMERA viral protein databases.

	No. of sequences (%)														
	SEED_nr BLASTx ($E<10^{-5}$)					CAMERA_v BLASTx ($E<10^{-5}$)					CAMERA_nr BLASTx ($E<10^{-3}$)				
	F-A	**F-B**	**F-C**	**F-D**	**F-E**	**F-A**	**F-B**	**F-C**	**F-D**	**F-E**	**F-A**	**F-B**	**F-C**	**F-D**	**F-E**
Raw reads	113,054	109,569	68,391	115,121	98,511	113,054	109,569	68,391	115,121	98,511	113,054	109,569	68,391	115,121	98,511
Quality-filtered	100,358	98,601	61,641	107,131	86,483	100,358	98,601	61,641	107,131	86,483	100,358	98,601	61,641	107,131	86,483
Non-redundant	90,273 (100%)	88,810 (100%)	53,342 (100%)	95,902 (100%)	71,806 (100%)	90,273 (100%)	88,810 (100%)	53,342 (100%)	95,902 (100%)	71,806 (100%)	90,273 (100%)	88,810 (100%)	53,342 (100%)	95,902 (100%)	71,806 (100%)
Unknown	80,523 (89.2%)	78,989 (88.94%)	47,246 (88.57%)	86,777 (90.49%)	62,534 (87.09%)	81,738 (90.55%)	83,190 (93.67%)	49,350 (92.5%)	81,694 (85.18%)	63,447 (88.36%)	78,234 (86.66%)	71,838 (80.89%)	43,535 (81.61%)	69,779 (72.76%)	54,422 (75.79%)
Known	9,750 (10.8%)	9,821 11.06%)	6,096 (11.43%)	9,125 (9.51%)	9,272 (12.91%)	8,535 (9.45%)	5620 (6.33%)	3,992 (7.5%)	14,208 (14.82%)	8,359 (11.64%)	12,039 (13.34%)	16,972 (19.11%)	9,807 (18.39%)	26,123 (27.24%)	17,384 (24.21%)
Virus	6,091 (6.75%)	2,125 (2.39%)	1,902 (3.57%)	2,976 (3.1%)	7,000 (9.75%)	8,467 (9.38%)	5,620 (6.33%)	3,952 (7.4%)	14,155 (14.76%)	8,348 (11.63%)	6,186 (6.85%)	2,381 (2.68%)	1,274 (2.39%)	10,255 (10.69%)	4,049 (5.64%)
Bacteria	3,599 (3.99%)	6,709 (7.55%)	4,103 (7.69%)	6,073 (6.33%)	2,180 (3.04%)	8 (0.01%)	0	35 (0.07%)	28 (0.03%)	0	5,661 (6.27%)	13,944 (15.7%)	8,263 (15.49%)	15,758 (16.43%)	13,198 (18.38%)
Archaea	13 (0.01%)	711 (0.8%)	13 (0.02%)	10 (0.01%)	19 (0.03%)	0	0	0	0	0	0	316 (0.36%)	0	0	0
Eukarya	47 (0.05%)	276 (0.31%)	78 (0.15%)	64 (0.07%)	73 (0.1%)	60 (0.07%)	0	5 (0.01%)	25 (0.03%)	11 (0.02%)	157 (0.14%)	133 (0.15%)	153 (0.29%)	86 (0.09%)	106 (0.15%)
Other	0	0	0	2 (0.002%)	0	0	0	0	0	0	35 (0.04%)	198 (0.22%)	117 (0.22%)	24 (0.03%)	31 (0.04%)

Table S3. The profiles of viral families of five viromes comparing with three public databases, the SEED nr, CAMERA nr and CAMERA viral protein databases.

		SEED_nr BLASTx (E <10^{-5})					CAMERA_v BLASTx (E <10^{-5})					CAMERA_nr BLASTx (E <10^{-3})				
		F-A	**F-B**	**F-C**	**F-D**	**F-E**	**F-A**	**F-B**	**F-C**	**F-D**	**F-E**	**F-A**	**F-B**	**F-C**	**F-D**	**F-E**
Viral sequences		6,090	2,125	1,902	2,976	7,000	8,467	5,620	3,952	14,155	8,348	6,186	2,381	1,274	10,255	4,049
dsDNA	Myo-	51	240	162	210	76	362	877	328	132	134	19	165	50	22	47
	Podo-	4,252	1,144	323	778	801	4,928	1,834	507	695	1,455	4,772	1,948	370	735	1,265
	Sipho-	121	224	289	278	47	702	1,291	1,043	10,381	203	42	41	413	8,809	15
	Unclassified Caudovirales	3	20	12	2	74	138	252	169	264	131	0	0	15	92	60
	Asfar-	1	0	1	1	0	6	0	5	0	0	7	0	6	0	0
	Irido-	0	11	1	0	0	27	44	0	0	0	13	6	0	2	0
	Herpes-	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0
	Adeno-	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
	Mimi-	4	29	5	5	2	14	28	29	20	0	3	3	6	7	0
	Phycodna-	11	22	2	0	3	23	24	6	0	6	0	5	0	0	0
	Pox-	5	55	13	0	0	8	19	20	0	0	0	0	0	0	0
	Tecti-	0	0	2	13	0	16	0	0	10	9	9	0	0	4	0
	Bicauda-	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
	Unclassified viruses	0	2	0	0	0	102	89	90	5	26	2	4	4	2	0
ssDNA	Ino-	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0
	Anello-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	Germini-	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
	Micro-	1,638	377	1,087	1,655	5,993	1,800	420	1,182	2,467	6,211	1,312	207	408	576	2,660
	Unclassified-	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0
Unclassified	Unclassified phages	0	1	3	26	1	339	723	566	181	173	5	2	2	0	0
RNA	Picorna-	4	0	2	4	3	0	0	0	0	0	2	0	0	2	0

Table S4. Comparison of viral and bacterial community structures and diversity.

	Sample	Richness[a]	Shannon-Wiener index (nats)
Viral assemblage	**F-A**	**34**	**3.04**
	F-B	**24**	**2.79**
	F-C	**26**	**2.84**
	F-D	**18**	**2.58**
	F-E	**401**	**4.19**
Bacterial community	**F-A**	**1636 / 2510**	**4.64**
	F-B	**1214 / 1632**	**4.25**
	F-C	**1607 / 2197**	**4.55**
	F-D	**914 / 1247**	**4.67**
	F-E	**1492 / 1953**	**4.59**

[a] In bacterial richness, Chao1 and Ace estimators are shown in order.

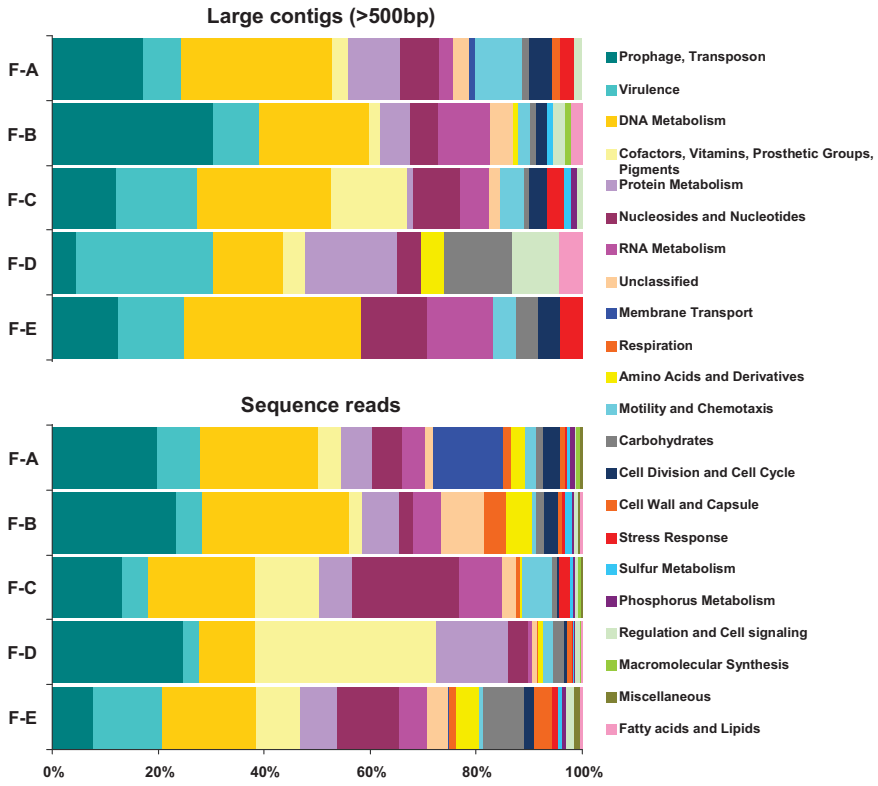
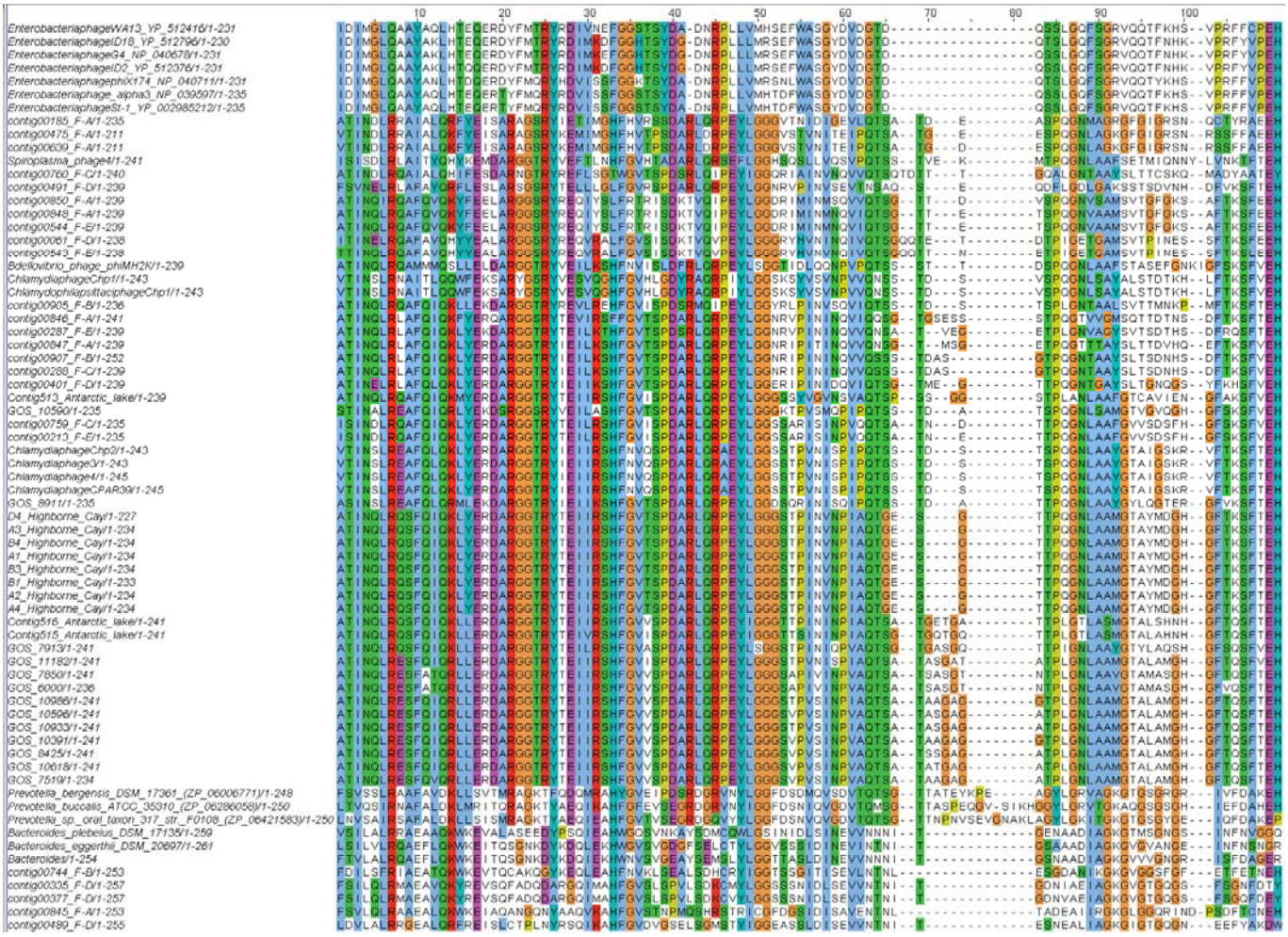
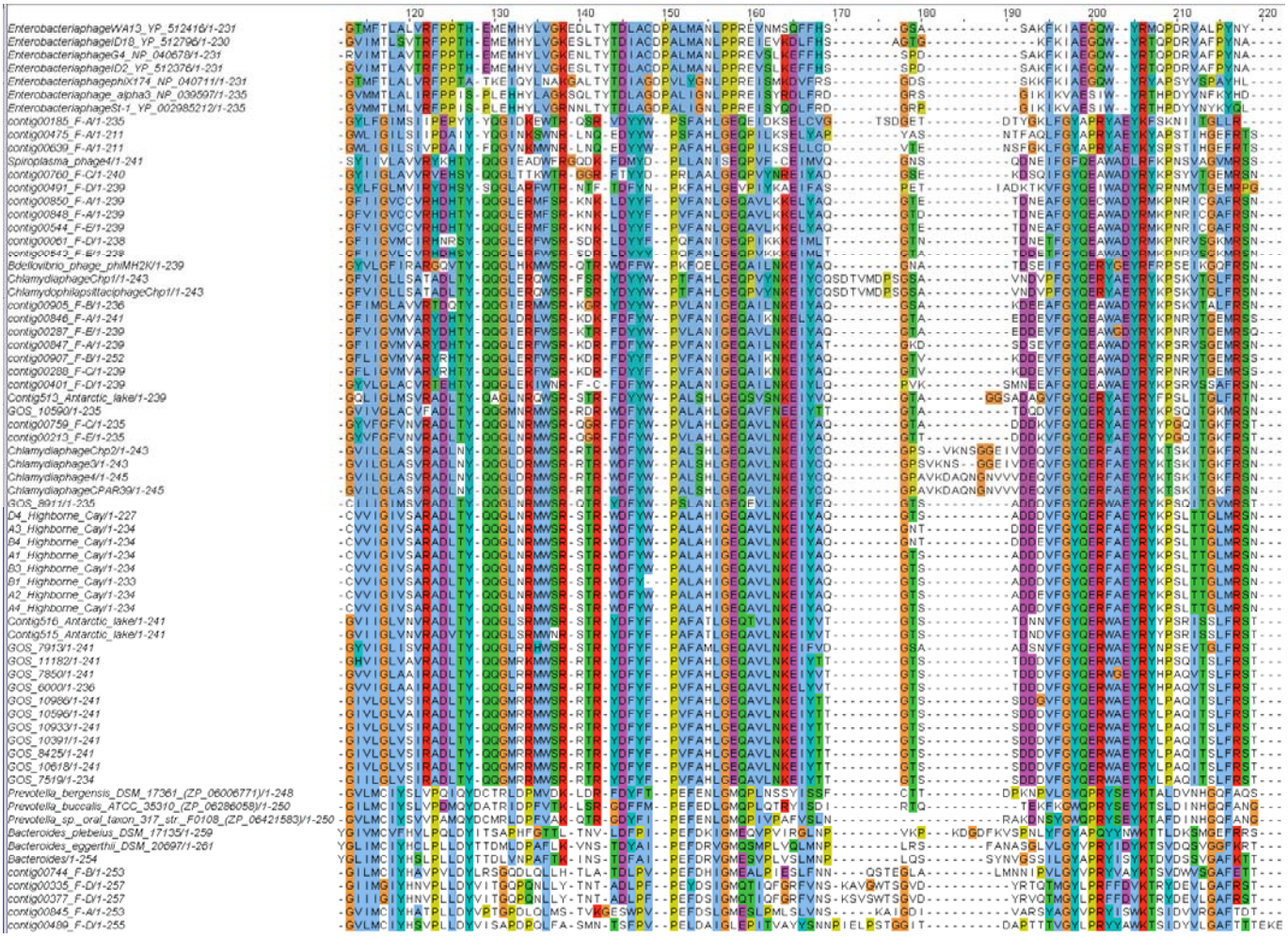


Fig. S1. Functional gene profiles of five viromes comparing to the SEED database. Both the sequence reads and large contigs (>500 bp) of five viromes were compared in BLAST searches (BLASTx, *E*-value < 10^{-2}). The functional categories including higher than 1% of the identified sequences were shown.


10
20
30
40
50
60
70
80
90
100
EnterobacteriaphageWA13_YP_512416/1-231
EnterobacteriaphageID18_NP_512796/1-230
EnterobacteriaphageG4_NP_040678/1-231
EnterobacteriaphageID2_YP_512376/1-231
EnterobacteriaphagephiX174_NP_040711/1-231
Enterobacteriaphage_alpha3_NP_039597/1-235
EnterobacteriaphageSt-1_YP_002985212/1-235
contig00185_F-A/1-235
contig00475_F-A/1-211
contig00639_F-A/1-211
Spiroplasma_phage4/1-241
contig00760_F-C/1-240
contig00491_F-D/1-239
contig00850_F-A/1-239
contig00848_F-A/1-239
contig00544_F-E/1-239
contig00061_F-D/1-238
contig00543_F-E/1-238
Bdellovibrio_phage_phiMH2K/1-239
ChlamydiaphageChp1/1-243
ChlamydophilapsittaciphageChp1/1-243
contig00905_F-B/1-236
contig00846_F-A/1-241
contig00287_F-E/1-239
contig00847_F-A/1-239
contig00907_F-D/1-252
contig00288_F-C/1-239
contig00401_F-D/1-239
Contig513_Antarctic_lake/1-239
GOS_10590/1-235
contig00759_F-C/1-235
contig00213_F-E/1-235
ChlamydiaphageChp2/1-243
Chlamydiaphage3/1-243
Chlamydiaphage4/1-245
ChlamydiaphageCPAR39/1-245
GOS_8911/1-235
D4_Highborne_Cay/1-227
A3_Highborne_Cay/1-234
B4_Highborne_Cay/1-234
A1_Highborne_Cay/1-234
B3_Highborne_Cay/1-234
B1_Highborne_Cay/1-233
A2_Highborne_Cay/1-234
A4_Highborne_Cay/1-234
Contig516_Antarctic_lake/1-241
Contig515_Antarctic_lake/1-241
GOS_7913/1-241
GOS_11182/1-241
GOS_7850/1-241
GOS_6000/1-236
GOS_10986/1-241
GOS_10596/1-241
GOS_10933/1-241
GOS_10391/1-241
GOS_8425/1-241
GOS_10618/1-241
GOS_7519/1-234
Prevotella_bergensis_DSM_17361_(ZP_06006771)/1-248
Prevotella_buccalis_ATCC_35310_(ZP_06286058)/1-250
Prevotella_sp._oral_taxon_317_str._F0108_(ZP_06421583)/1-250
Bacteroides_plebeius_DSM_17135/1-259
Bacteroides_eggerthii_DSM_20697/1-261
Bacteroides/1-254
contig00744_F-B/1-253
contig00335_F-D/1-257
contig00377_F-D/1-257
contig00845_F-A/1-253
contig00489_F-D/1-255



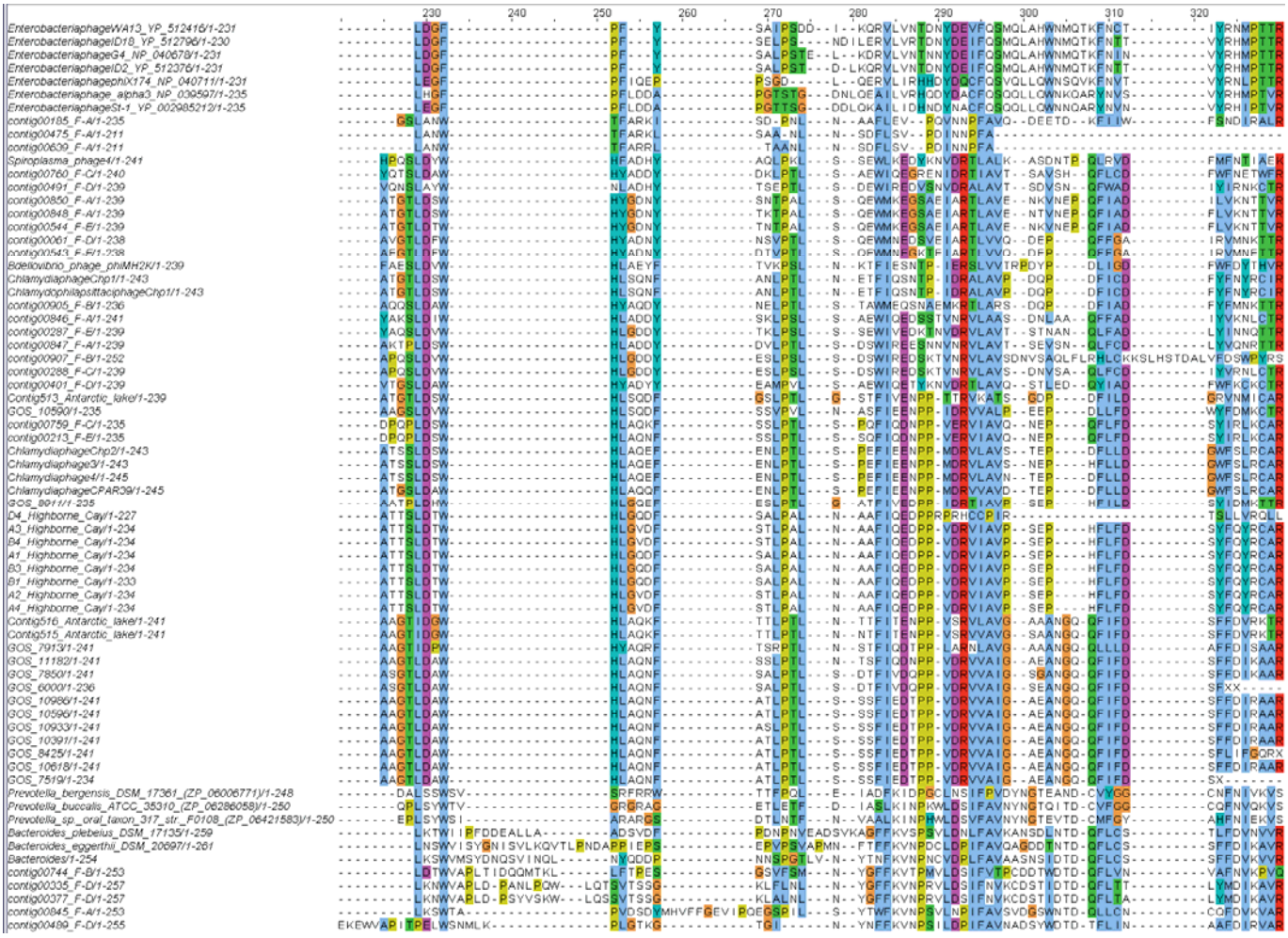


Fig. S2. Multiple sequence alignment of the capsid protein sequences of the microphages from human faeces, cultured isolates and environmental samples. MUSCLE program (http://www.ebi.ac.uk/Tools/muscle) was used for sequence alignment and aligned partial capsid protein sequences of human faeces, Sagarsso Sea, Highborne Cay, Antarctic lake, *Bacteroides* and *Prevotella* and cultured isolations (Chlamydia-, Bdellovibrio- and Spiroplasma-, Enterobacteria phages) were visualized using Jalview (http://www.jalview.org/).

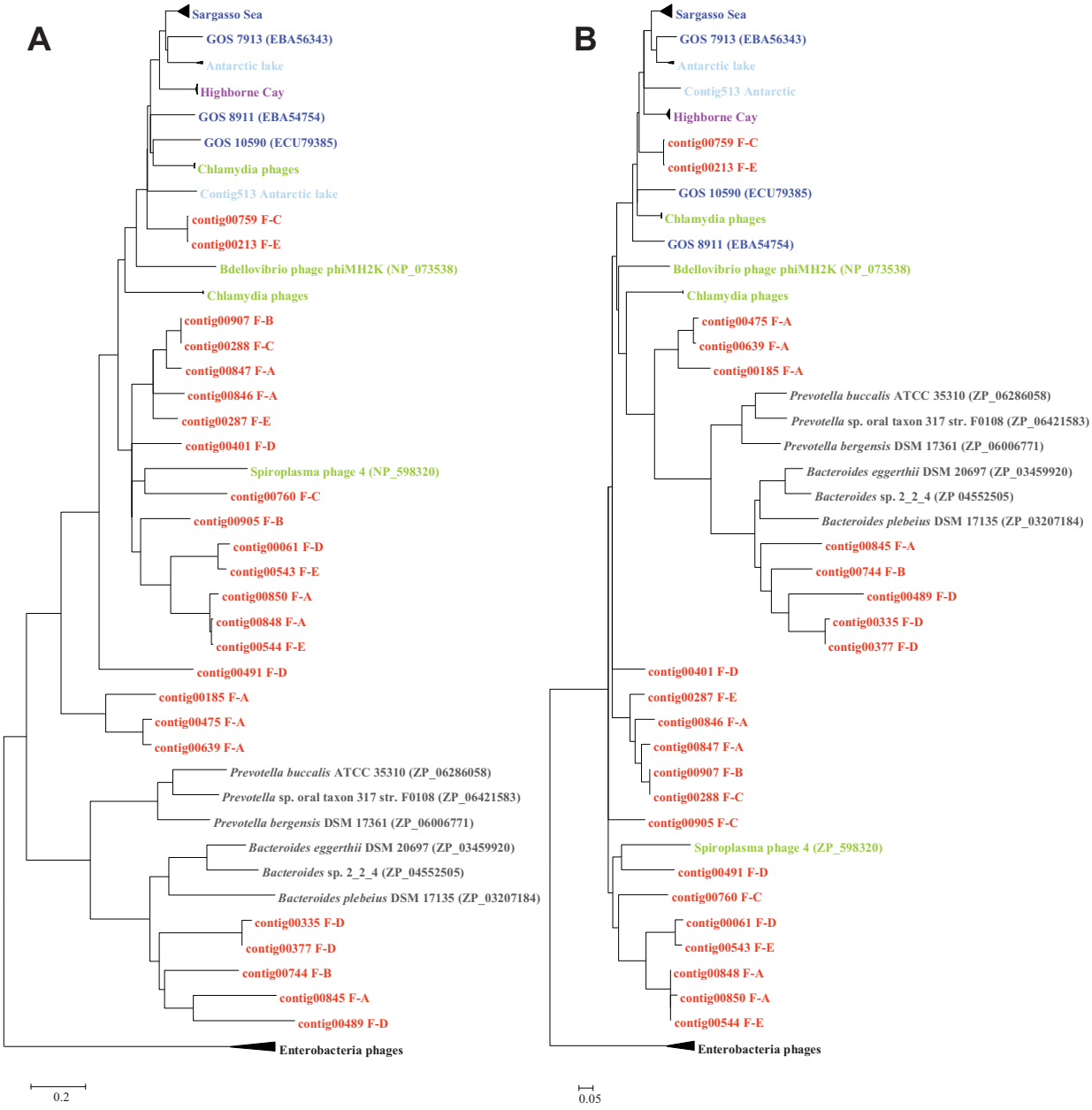


Fig. S3. Phylogenetic trees of the capsid genes of the microphages from human faeces, environmental samples, *Bacteroides* and *Prevotella* and cultured isolations at the protein (A) and nucleic (B) levels. The sequences were shown in colour as follows: human faeces (red), Sagarsso Sea (deep blue), Highborne Cay (purple), Antarctic lake (light blue), *Bacteroides* and *Prevotella* (gray) and cultured isolations (Chlamydia-, Bdellovibrio- and Spiroplasma phages, green; Enterobacteria phages, black). The phylogenetic trees were constructed based on neighbor-joining algorithm using MEGA 4. Enterobacteria phages were used as outgroups. The scale bars represents 0.2 amino-acid and 0.05 nucleotide substitutions per site, respectively.