Supplementary Material

Comparative analysis of tandem T7-like promoter containing regions in enterobacterial genomes reveals a novel group of genetic islands

Zehua Chen and Thomas D. Schneider [*]

Center for Cancer Research Nanobiology Program,
National Cancer Institute at Frederick, Frederick, MD 21702

[*] To whom correspondence should be addressed: Tel: 301-846-5581; Fax: 301-846-5598; Email: toms@ncifcrf.gov, http://www.ccrnp.ncifcrf.gov/~toms/

version = 1.37 of supp.tex 2006 Jan 30.

Table S1: List of enterobacterial genomes and T7-like phage hosts scanned in this study.

Bacteria genome scanned	Acc. No.	T7 island[a]
Escherichia coli K12*	NC_000913	–
Escherichia coli CFT073	NC_004431	–
Escherichia coli O157:H7	NC_002695	–
Escherichia coli O157:H7 EDL933	NC_002655	–
Escherichia coli strain 042	Sanger[b]	–
Escherichia coli E110019	NZ_AAJW00000000	–
Escherichia coli E22	NZ_AAJV00000000	E22
Shigella flexneri 2a strain 2457T	NC_004741	T-1, T-2, T-3
Shigella flexneri 2a strain 301	NC_004337	301-1, 301-3
Shigella boydii serotype 18 strain BS512	NZ_AAKA00000000	BS512
Shigella boydii serotype 4 strain 227	NC_007613	–
Shigella sonnei strain 53G	Sanger	–
Shigella sonnei strain Ss046	NC_007384	–
Shigella dysenteriae serotype 1 strain 197	NC_007606	–
Salmonella enterica serovar Typhi strain Ty2	NC_004631	Ty2
Salmonella enterica serovar Typhi strain CT18	NC_003198	CT18
Salmonella enterica serovar Typhimurium strain LT2	NC_003197	–
Salmonella enterica serovar Typhimurium DT104	Sanger	–
Salmonella enterica serovar Paratyphi A strain ATCC 9150	NC_006511	–
Salmonella enterica serovar Choleraesuis strain SC-B67	NC_006905	–
Salmonella enterica serovar Enteritidis PT4	Sanger	–
Salmonella bongori strain 12419	Sanger	–
Yersinia enterocolitica strain 8081	Sanger	Ye8081
Yersinia pestis CO92	NC_003143	–
Yersinia pestis KIM	NC_004088	–
Yersinia pestis strain 91001	NC_005810	–
Yersinia pseudotuberculosis IP 32953	NC_006155	–
Citrobacter rodentium strain ICC168	Sanger	CR
Erwinia carotovora strain SCRI1043	NC_004547	ECA
Buchnera aphidicola strain APS (Acyrthosiphon pisum)*	NC_002528	–
Buchnera aphidicola strain Bp (Baizongia pistaciae)*	NC_004545	–
Buchnera aphidicola strain Sg (Schizaphis graminum)*	NC_004061	–
*Blochmannia floridanus**	NC_005061	–
Blochmannia pennsylvanicus strain BPEN*	NC_007292	–
Photorhabdus luminescens subsp. laumondii TTO1	NC_005126	–
*Wigglesworthia glossinidia**	NC_004344	–
Pseudomonas putida KT2440 (gh-1 host)	NC_002947	–
Vibrio cholerae strain N16961 (VP4 host)	NC_002505/NC_002506	N16961

[a] – means no island, otherwise the island name is given.

[b] These are unfinished or finished genome sequences from the Sanger Institute. These sequences have not been deposited in the GenBank, so no accession numbers are available. These sequence data were produced by the Pathogen Sequencing Unit at the Sanger Institute and can be obtained from http://www.sanger.ac.uk/Projects/Microbes/.

Strains marked by * are non-pathogens, all others are pathogens.

Table S2: Similarity matches of the T7 island proteins.

T7 island protein[a]	Program	Significant database matches[b]	% Identity[c]	Expect	Function annotation
Int	BlastP	Possible integrase STY3193, NP_457435 (CT18)	96 (465/483)	0.0	Int_SG2
(BS512)		Putative integrase S3064, NP_838364 (T-3)	96 (455/473)	0.0	Site-specific
(ZP_00698814)		Putative integrase SF2866, NP_708645 (301-3)	96 (455/473)	0.0	integration
		Putative integrase S1981, NP_837478 (T-2)	75 (369/486)	0.0	
		Hypothetical protein t2953, NP_806646 (Ty2)	75 (367/487)	0.0	
		Integrase EcolE1_01003485, ZP_00718924 *E. coli* E110019	76 (333/435)	0.0	
		Integrase EcolE2_01002184, ZP_00729397 (E22)	30 (148/478)	2e-40	
		Probable phage integrase ECA2306, YP_050401 (ECA)	29 (146/488)	2e-38	
		Hypothetical protein VP0643, NP_797022 *Vibrio parahaemolyticus*	30 (123/404)	3e-33	
		Putative integrase SF1604, NP_707482 (301-1)	44 (75/170)	4e-27	
		Putative integrase SF1608, NP_707485 (301-1)	25 (75/293)	4e-11	
		Putative integrase S1739, NP_837274 (T-1)	25 (75/293)	4e-11	
		Hypothetical protein VCA0790, NP_233176 (N16961)	29 (111/374)	1e-19	
		Integrase VchoO_01003279, ZP_00755026 (O395)	29 (111/371)	1e-19	
	Blast2[d]	YE3373, (Ye8081)	71 (354/493)	0.0	
		Int, (CR)	68 (330/479)	0.0	
	CD-Blast[e]	cd01184, INT_SG2_C, DNA breaking-rejoining enzymes, 100% aligned		5e-46	
		cd01189, INT_phiLC3_C, phiLC3 phage integrases, 95.3% aligned		2e-12	
		cd01182, INT_REC_C, DNA breaking-rejoining enzymes, 94.4% aligned		2e-12	
		cd00397, DNA_BRE_C, DNA breaking-rejoining enzymes, 93.9% aligned		8e-12	
		cd00798, INT_XerDC, XerD and XerC integrases, 71.5% aligned		9e-11	
		cd00801, INT_P4, Bacteriophage P4 integrase, 70.3% aligned		2e-10	
Hyp1	BlastP	Hypothetical protein EcolE2_01002186, ZP_00729399 (E22)	65 (146/223)	2e-75	Putative
(BS512)		Hypothetical protein STY3192, NP_457434 (CT18)	77 (135/174)	1e-74	phage-related
(ZP_00698813)		Hypothetical protein, ZP_00669398 *N. eutropha* C71	36 (94/260)	9e-42	protein
		Hypothetical protein, ZP_00859983 *Bradyrhizobium*	38 (95/247)	2e-39	
		Hypothetical protein, ZP_00637571 *S. frigidimarina*	35 (85/239)	6e-34	
		Hypothetical protein, ZP_00875590 *Streptococcus suis*	34 (76/221)	1e-32	
		gp7, NP_862846 *Streptococcus mitis* phage SM1	25 (37/147)	6e-4	
	Blast2	Hyp1, (CR)	72 (150/207)	4e-84	

Continued on Next Page…

Table S2 – Continued

T7 island protein[a]	Program	Significant database matches[b]	% Identity[c]	Expect	Function annotation
Hyp2 (BS512) (ZP_00698811)[f]	BlastP	Hypothetical protein S3062, NP_838362 (T-3)	85 (204/239)	4e-114	Putative phage anti-repressor
		Hypothetical protein SF2861, NP_708640 (301-3)	85 (204/239)	4e-114	
		Hypothetical protein t2951, NP_806644 (Ty2)	76 (177/232)	2e-93	
		putative phage-related protein ECA2309, YP_050404 (ECA)	61 (141/229)	2e-74	
		[g]Ribosome recycling factor EcolE2_01002192, ZP_00729405 (E22)	87 (130/148)	6e-70	
		[h]Anti-repressor protein, YP_033514 *Bartonella henselae*	33 (61/181)	4e-13	
		[h]DNA-binding protein Roi, YP_179421 *Campylobacter jejuni*	43 (42/97)	1e-11	
	Blast2	YE3371, (Ye8081)	60 (137/227)	2e-70	
	CD-Blast	Phage-encoded protein, COG3646, 44.3% aligned		2e-7	
Hyp3 (BS512) (ZP_00698810)	PSI-Blast	Hypothetical protein EcolE2_01002193, ZP_00729406 (E22)	87 (49/56)	4e-17	Hypothetical protein
		Hypothetical protein ECA2310, YP_050405 (ECA)	37 (20/53)	7e-18	
		Hypothetical protein t2950, NP_806643 (Ty2)	43 (22/53)	2e-14	
	Blast2	Hyp3, (301-3)	98 (53/54)	2e-24	
		Hyp3, (T-3)	98 (53/54)	2e-24	
		Hyp3, (301-1)	53 (32/60)	5e-7	
		Hyp3, (T-1)	53 (32/60)	5e-7	
		Hyp3, (CT18)	58 (29/50)	2e-6	
Hyp4 (BS512) (ZP_00698809)	PSI-Blast	Hypothetical protein S3060, NP_838361 (T-3)	92 (150/163)	8e-75	Hypothetical protein
		Hypothetical protein SF2860, NP_708639 (301-3)	92 (150/163)	8e-75	
		Hypothetical protein S1728, NP_837268 (T-1)	84 (138/163)	3e-75	
		Hypothetical protein SF1600, NP_707478 (301-1)	84 (138/163)	3e-75	
		Hypothetical protein STY3189, NP_457432 (CT18)	82 (135/163)	6e-74	
		Hypothetical protein EcolE2_01002194, ZP_00729407 (E22)	66 (109/163)	1e-68	
		Hypothetical protein ECA2311, YP_050406 (ECA)	23 (29/122)	7e-41	
		Hypothetical protein t2949, NP_806642 (Ty2)	21 (27/123)	2e-23	
	GAP[i]	Hypothetical protein VchoO_01003282, ZP_00755029 (O395)	30(40/135)		

Continued on Next Page...

Table S2 – Continued

T7 island protein[a]	Program	Significant database matches[b]	% Identity[c]	Expect	Function annotation
Hyp5 (BS512) (ZP_00698808)	BlastP	Hypothetical protein S3059, NP_838360 (T-3)	95 (314/328)	5e-176	Hypothetical protein (Other hits have an E-value > 0.5)
		Hypothetical protein SF2859, NP_838360 (301-3)	95 (314/328)	5e-176	
		Hypothetical protein STY3188, NP_457431 (CT18)	93 (302/324)	5e-170	
		Hypothetical protein S1727, NP_837267 (T-1)	91 (298/326)	7e-167	
		Hypothetical protein SF1599, NP_707477 (301-1)	91 (298/326)	7e-167	
		Hypothetical protein EcolE2_01002195, ZP_00729408 (E22)	81 (265/324)	8e-149	
		Hypothetical protein t2948, NP_806641 (Ty2)	29 (80/275)	6e-22	
		Hypothetical protein ECA2312, YP_050407 (ECA)	30 (91/294)	8e-22	
		Hypothetical protein VchoO_01003283, ZP_00755030 (O395)	24 (64/264)	1e-14	
		ORF27, AAF71189 *Vibrio cholerae*	24 (55/223)	9e-12	
		Hypothetical protein S0233, NP_835956 (T-2)	29 (36/124)	3e-8	
ECA2307 (ECA) (YP_050402)	BlastP	protein kinase, NP_523300 phage T3	35 (32/91)	0.001	Putative phage-related protein
		protein kinase, NP_041959 phage ϕYeO3-12	32 (29/88)	0.005	
		protein kinase, NP_041959 phage T7	35 (31/87)	0.013	
ECA2308 (ECA) (YP_050403)	BlastP	Hypothetical protein, NP_258393 *Spodoptera litura* NPV	28 (25/87)	2e-7	Prophage antirepressor
		Prophage antirepressor, YP_063130 *Leifsonia xyli* CTCB07	28 (47/164)	6e-7	
	CD-Blast	Prophage antirepressor, COG3617, 82.4% aligned		4e-15	
		pfam02498, Bro-N, BRO family, N-terminal domain, 92.8% aligned		3e-6	
CR3 (CR)	CD-Blast	pfam00239, Resolvase N terminal domain, 100% aligned		2e-27	Resolvase
		COG1961, PinR, Site-specific recombinases, 91.4% aligned		6e-27	

[a] All island proteins were used to Blast against GenBank (as of Dec, 2005). To avoid redundancy, only results for the six homologs (Int and Hyp1 to Hyp5) of the island BS512 are listed; the results are similar for other islands. Several non-homologous proteins that have significant database matches are also listed. The island name (in parenthesis) and GenBank accession number (when available, in parenthesis) are given underneath the protein name.

[b] For hits in T7 islands, the name of the island is given in parenthesis.

[c] Percent identity is given for each comparison; the number of identical amino acids and the total length of the alignment are given in parenthesis.

[d] The proteins of the islands Ye8081 and CR are not available in GenBank, so the program Blast2 was used for comparisons.

[e] Blast the Conserved Domain Database (CDD v2.05).

[f] The N-terminal region (1-62 AA) of this protein matches significantly with many IS629 ORF1 proteins, so only the C-terminal part (63-295 AA) was used for database search.

[g] No similarity between this protein and any ribosome recycling factor can be detected by BlastP, CD-Blast and PSI-Blast, so this may be an incorrect assignment.

[h] Using PSI-Blast, many significant hits ($E <$ 1e-20) of anti-repressor and Roi proteins were detected for Hyp2.

[i] The O395 Hyp4 (ZP_00755029, Figure S5) is only weakly similar to BS512 Hyp4 and not found by PSI-Blast, so the GCG program GAP (1) was used to compare these two proteins.

```
A         -----------
          21111111111--------- +++++
          09876543210987654321012345
K1F, NC_007456                       bits
p1   2491 atgacttaactatcactataggactc 32.9
p2   5569 taacctaaactatcactataggatta 31.5
p3   7200 gaccctaagctatcactataggacaa 27.6
p4   8516 agacctaaactatcactgtaggacag 30.8
p5  10883 aagactaaactatcaccttagggatg 29.5
p6  17194 aacactaaactatcactatagggact 34.4
p7  20015 aagcctaaactatcactataggaaca 36.6
p8  21014 aggactaaactatcactatagggaga 35.7
p9  25338 ttaccttaactatcactatagggact 30.8
p10 32614 aggactaaactatcactataggaaac 36.5
p11 38869 tggacttaactatcactataggaaga 33.9
```

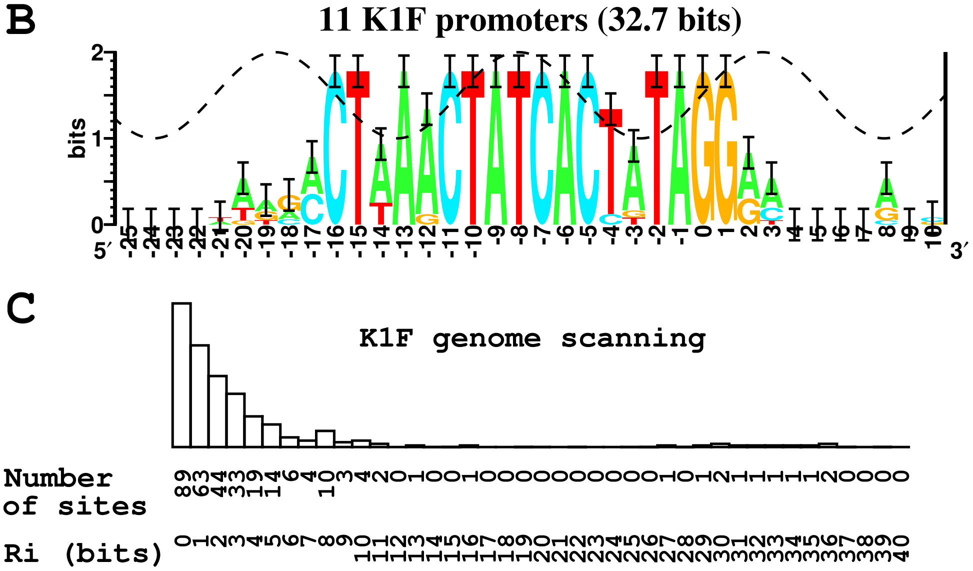
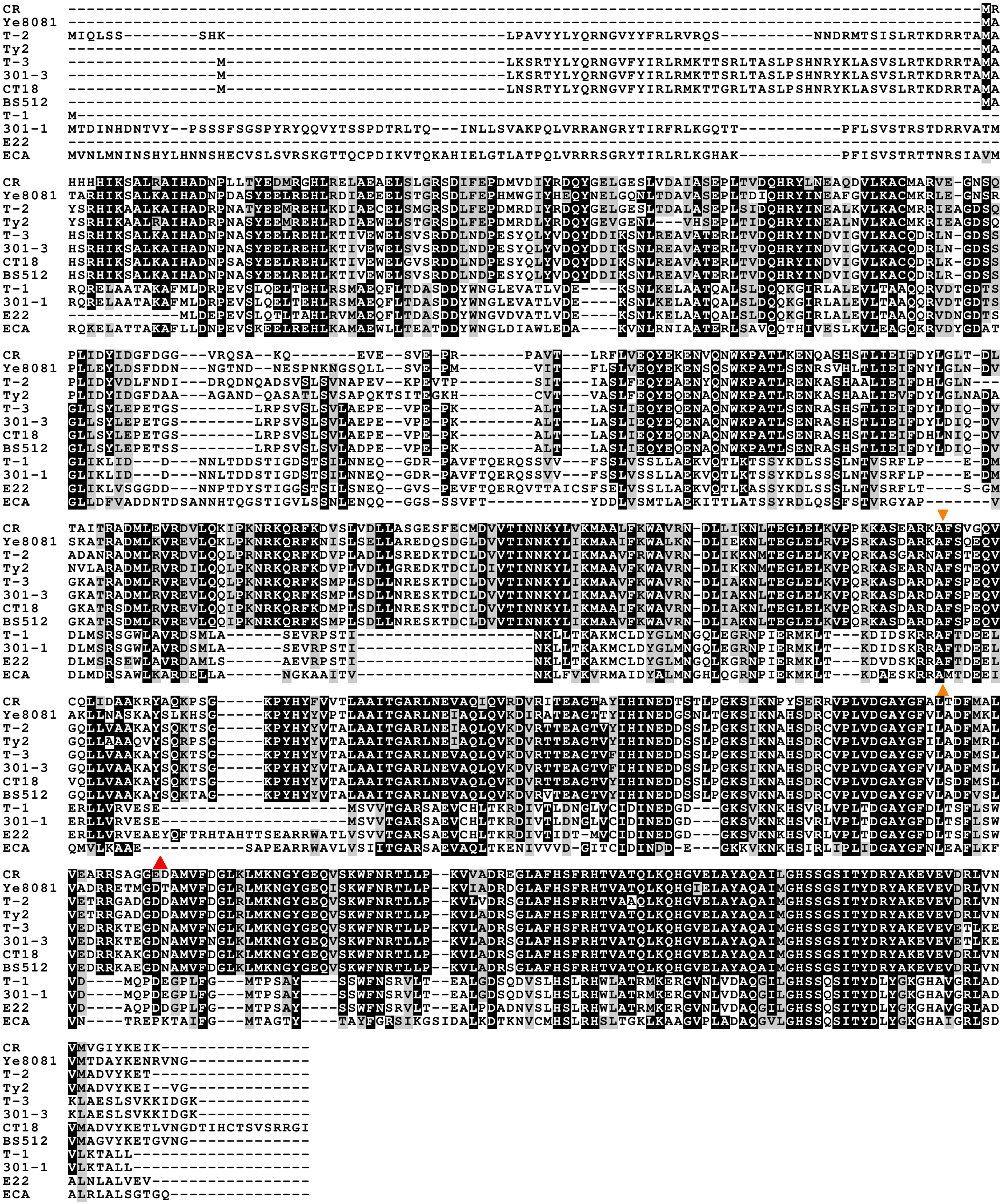


Figure S1: Prediction of phage K1F promoters.

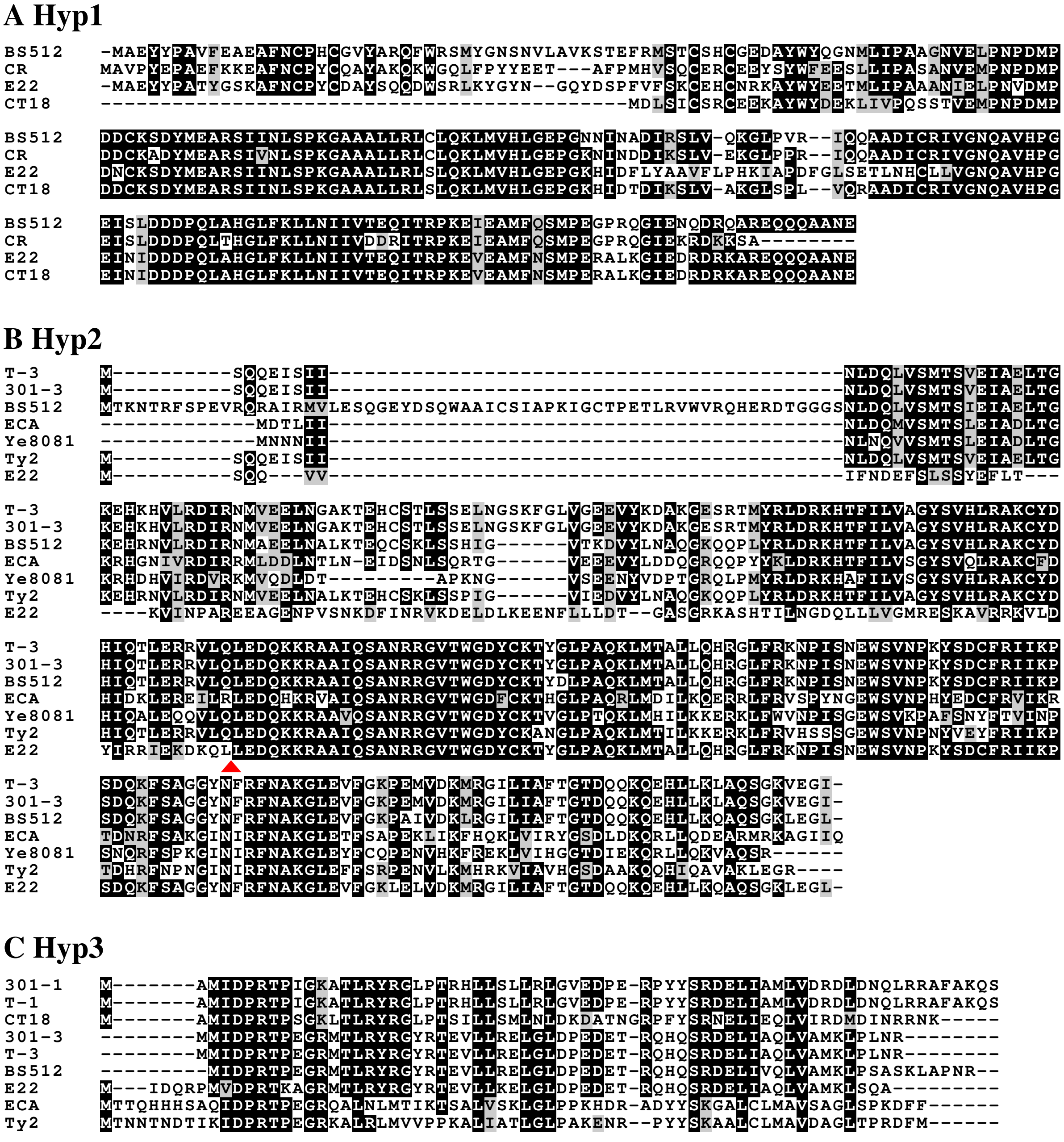
(A) As in previous work (2), the promoters from NC_007456 (3) were aligned from −20 to +5, relative to the transcription start (at 0). Individual information (bits) is given for each promoter. (B) The sequence logo was made as previously described (2). The height of each letter is proportional to the frequency of that base at each position, and the height of the letter stack is the conservation in bits (4). (C) The 11-site K1F promoter model was used to scan the K1F genome; all 11 promoters gave greater than 27 bits of information, while the background was lower than 17 bits.



```
CR     --------------------------------------------------------------------------------------------------MR
Ye8081 --------------------------------------------------------------------------------------------------MA
T-2    MIQLSS--------SHK------------------------------LPAVYYLYQRNGVYYFRLRVRQS----------NNDRMTSISLRTKDRRTAMA
Ty2    --------------------------------------------------------------------------------------------------MA
T-3    ----------------M------------------------------LKSRTYLYQRNGVFYIRLRMKTTSRLTASLPSHNRYKLASVSLRTKDRRTAMA
301-3  ----------------M------------------------------LKSRTYLYQRNGVFYIRLRMKTTSRLTASLPSHNRYKLASVSLRTKDRRTAMA
CT18   ----------------M------------------------------LNSRTYLYQRNGVFYIRLRMKTTGRLTASLPSHNRYKLASVSLRTKDRRTAMA
BS512  --------------------------------------------------------------------------------------------------MA
T-1    M---------------------------------------------------------------------------------------------------
301-1  MTDINHDNTVY--PSSSFSGSPYRYQQVYTSSPDTRLTQ---INLLSVAKPQLVRRANGRYTIRFRLKGQTT-----------PFLSVSTRSTDRRVATM
E22    ----------------------------------------------------------------------------------------------------
ECA    MVNLMNINSHYLHNNSHECVSLSVRSKGTTQCPDIKVTQKAHIELGTLATPQLVRRRSGRYTIRLRLKGHAK-----------PFISVSTRTTNRSIAVM

CR     HHHHIKSALRAIHADNPLLTYEDMRGHLRELAEAELSLGRSDIFEPDMVDIYRDQYGELGESLVDAIASEPLTVDQHRYLNEAQDVLKACMARVE-GNSQ
Ye8081 TARHIKSALKAIHADNPDASYEELREHLRDIAEEELSTGRSDLFEPHMWGIYHEQYNELGQNLTDAVASEPLTDIQHRYINEAFGVLKACMKRLE-GNSR
T-2    YSRHIKAALKAIHADRPNATYEEMREHLKDIAECELSMGRSDLFEPDMRDIYRDQYGELGESLTDALASEPLSIDQHRYINEALKVLKACMRRIEAGDSQ
Ty2    YSRHIKAALRAIHADRPNASYEEMREHLRDIAEWELSTGRSDLFEPDMRDLYRDQYGEVGENL---VHSEPLTIDQHRYINEALNVLKACMKRIEAGDSQ
T-3    HSRHIKSALKAIHADNPNASYEELREHLKTIVEWELSVSRDDLNDPESYQLYVDQYDDIKSNLREAVATERLTVDQHRYINDVIGVLKACQDRLN-GDSS
301-3  HSRHIKSALKAIHADNPNASYEELREHLKTIVEWELSVSRDDLNDPESYQLYVDQYDDIKSNLREAVATERLTVDQHRYINDVIGVLKACQDRLN-GDSS
CT18   HSRHIKSALKAIHADNPSASYEELREHLKTIVEWELGVSRDDLNDPESYQLYVDQYDDIKSNLREAVATERLTVDQHRYINDVIGVLKACQDRLK-GDSS
BS512  HSRHIKSALKAIHADNPNASYEELREHLKTIVEWELSVSRDDLNDPESYQLYVDQYDDIKSNLREAVATERLTVDQHRYINDVIGVLKACQDRLR-GDSS
T-1    RQRELAATAKAFMLDRPEVSLQELTEHLRSMAEQFLTDASDDYWNGLEVATLVDE----KSNLKELAATQALSLDQQKGIRLALEVLTAAQQRVDTGDTS
301-1  RQRELAATAKAFMLDRPEVSLQELTEHLRSMAEQFLTDASDDYWNGLEVATLVDE----KSNLKELAATQALSLDQQKGIRLALEVLTAAQQRVDTGDTS
E22    ------------MLDEPEVSLQTLTAHLRVMAEQFLTDASDDYWNGVDVATLVDE----KSNLKELAATQALSLDQQKGIRLALEVLTAAQQRVDNGDTS
ECA    RQKELATTAKAFLLDNPEVSKEELREHLKAMAEWLLTEATDDYWNGLDIAWLEDA----KVNLRNIAATERLSAVQQTHIVESLKVLEAGQKRVDYGDAT

CR     PLIDYIDGFDGG---VRQSA--KQ-------EVE--SVE-PR-------PAVT---LRFLVEQYEKENVQNWKPATLKENQASHSTLIEIFDYLGLT-DL
Ye8081 PLLEYLDSFDDN---NGTND--NESPNKNGSQLL--SVE-PM--------VIT---FLSLVEQYEKENSQSWKPATLSENRSVHLTLIEIFNYLGLN-DV
T-2    PLIDYVDLFNDI---DRQDNQADSVSLSVNAPEV--KPEVTP--------SIT---IASLFEQYEAENYQNWKPATLRENKASHAALIEIFDHLGLN---
Ty2    PLIDYIDGFDAA---AGAND-QASATLSVSAPQKTSITEGKH--------CVT---VASLFEQYEQENAQNWKPATLRENKASHAALIEVFDYLGLNADA
T-3    GLLSYLEPETGS--------LRPSVSLSVLAEPE--VPE-PK--------ALT---LASLIEQYEQENAQNWKPATLSENRASHSTLIEIFDYLDIQ-DV
301-3  GLLSYLEPETGS--------LRPSVSLSVLAEPE--VPE-PK--------ALT---LASLIEQYEQENAQNWKPATLSENRASHSTLIEIFDYLDIQ-DV
CT18   GLLSYLEPETGS--------LRPSVSLSVLAEPE--VPE-PK--------ALT---LASLIEQYEQENAQNWKPATLSENRASHSTLIEIFDHLNIQ-DV
BS512  GLLSYLEPETSS--------LRPSVSLSVLADPE--VPE-PK--------ALT---LASLIEQYEQENAQNWKPATLSENRASHSTLIEIFDYLDIQ-DV
T-1    GLIKLID---D---NNLTDDSTIGDSTSILNNEQ--GDR-PAVFTQERQSSVV---FSSLVSSLLAEKVQTLKTSSYKDLSSSLNTVSRFLP----E-DM
301-1  GLIKLID---D---NNLTDDSTIGDSTSILNNEQ--GDR-PAVFTQERQSSVV---FSSLVSSLLAEKVQTLKTSSYKDLSSSLNTVSRFLP----E-DM
E22    GLIKLVSGGDD---NNPTDYSTIGGSTSILSNEQ--GVS-PEVFTQERQVTTAICSFSELVSSLLAEKVQTLKASSYKDLSSSLNTVSRFLT----S-GM
ECA    GLLDFVADDNTDSANNHTQGSTIGVLSSNLENQQ--GGS-SSVFT-----------YDDLVSMTLAEKITTLATSSYRDLQSSFSTVRGYAP-------V

CR     TAITRADMLEVRDVLQKIPKNRKQRFKDVSLVDLLASGESFECMDVVTINNKYLVKMAALFKWAVRN-DLLIKNLTEGLELKVPPKKASEARKAFSVGQV
Ye8081 SKATRADMLKVREVLQKLPKNRKQRFKNISLSELLAREDQSDGLDVVTINNKYLIKMAAIFKWALKN-DLIEKNLTEGLELRVPSRKASDARKAFSQEQV
T-2    ADANRADMLRVRDVLQQLPRNRKQRFKDVPLADLLSREDKTDCLDVVTINNKYLIKMAAVFRWAVRN-DLIKKNMTEGLELKVPQRKASGARNAFSTEQV
Ty2    NVLARADMLRVRDILQQLPKNRKQRFKDVPLVDLLGREDKTDCLDIVTINNKYLIKMAAVFKWAVRN-DLIKKNMTEGLELKVPQRKASEARNAFSTEQV
T-3    GKATRADMLRVREVLQQLPKNRKQRFKSMPLSDLLNRESKTDCLDVVTINNKYLIKMAAVFKWAVRN-DLIAKNLTEGLELKVPQRKASDARDAFSPEQV
301-3  GKATRADMLRVREVLQQLPKNRKQRFKSMPLSDLLNRESKTDCLDVVTINNKYLIKMAAVFKWAVRN-DLIAKNLTEGLELKVPQRKASDARDAFSPEQV
CT18   GKATRSDMLRVREVLQQIPKNRKQRFKDMPLSDLLNRESKTDCLDVVTINNKYLIKMAAIFKWAVRN-DLIAKNLTEGLELKVPQRKASDARDAFSPEQV
BS512  GKATRSDMLRVREVLQQIPKNRKQRFKSMPLSDLLNRESKTDCLDVVTINNKYLIKMAAVFKWAVRN-DLIAKNLTEGLELKVPQRKASDARDAFSPEQV
T-1    DLMSRSGWLAVRDSMLA------SEVRPSTI-------------------NKLLTKAKMCLDYGLMNGQLEGRNPIERMKLT---KDIDSKRRAFTDEEL
301-1  DLMSRSGWLAVRDSMLA------SEVRPSTI-------------------NKLLTKAKMCLDYGLMNGQLEGRNPIERMKLT---KDIDSKRRAFTDEEL
E22    DLMSRSEWLAVRDAMLS------AEVRPSTI-------------------NKLLTKAKMCLDYGLMNGQLKGRNPIERMKLT---KDVDSKRRAFTDEEL
ECA    DLMDRSAWLKARDELLA------NGKAAITV-------------------NKLFVKVRMAIDYALMNGHLQGRNPIEKMKLT---KDAESKRRAMTDEEI

CR     CQLIDAAKRYAQKPSG-----KPYHYFVVTLAAITGARLNEVAQIQVRDVRITEAGTAYIHINEDTSTLPGKSIKNPYSERRVPLVDGAYGFALTDFMAL
Ye8081 AKLLNASKAYSLKHSG-----KPYHYYVPTLAAITGARLNEIAQLQVKDIRATEAGTTYIHINEDGSNLTGKSIKNAHSDRCVPLVDGAYGFVLADFMKL
T-2    GQLLVAAKAYSQKTSG-----KPYHYYVTALAAITGARLNEIAQLQVKDVRTTEAGTVYIHINEDDSSLPGKSIKNAHSDRCVPLVDGAYGFILADFMAL
Ty2    GQLLAAAQVYSQRPSG-----KPYHYYVTALAAITGARLNEIAQLQVKDVRTTEAGTVYIHINEDDSSLPGKSIKNAHSDRCVPLVDGAYGFILADFMRL
T-3    GQLLVAAKAYSQKTSG-----KPYHYYVTALAAITGARLNEVAQLQVKDVRTTEAGTVFIHINEDDSSLPGKSIKNAHSDRCVPLVDGAYGFVLADFMSL
301-3  GQLLVAAKAYSQKTSG-----KPYHYYVTALAAITGARLNEVAQLQVKDVRTTEAGTVFIHINEDDSSLPGKSIKNAHSDRCVPLVDGAYGFVLADFMSL
CT18   VQLLVAAKAYSQKTSG-----KPYHYYVTALAAITGARLNEVAQLQVKDVRTTEAGTVYIHINEDDSSLPGKSIKNAHSDRCVPLVDGAYGFVLSDFMSL
BS512  GQLLVAAKAYSQKTAG-----KPYHYYVTALAAITGARLNEVAQLQVKDVRVTEAGTVYIHINEDDSSLPGKSVKNAHSDRCVPLVDGAYGFVLADFVSL
T-1    ERLLVRVESE--------------------MSVVTGARSAEVCHLTKRDIVTLDNGLVCIDINEDGD---GKSVKNKHSVRLVPLTDGAYGFDLTSFLSW
301-1  ERLLVRVESE--------------------MSVVTGARSAEVCHLTKRDIVTLDNGLVCIDINEDGD---GKSVKNKHSVRLVPLTDGAYGFDLTSFLSW
E22    ERLLVRVEAEYQFTRHTAHTTSEARRWATLVSVVTGARSAEVCHLTKRDIVTIDT-MVCIDINEDGD---GKSVKNKHSVRLVPLTDGAYGFDLTSFLSW
ECA    QMVLKAAE-----------SAPEARRWAVLVSIITGARSAEVAQLTKENIVVVD-GITCIDINDD-E---GKKVKNKHSIRLIPLIDGAYGFNLEAFLKF

CR     VEARRSAGGEDAMVFDGLKLMKNGYGEQVSKWFNRTLLP--KVVADREGLAFHSFRHTVATQLKQHGVELAYAQAILGHSSGSITYDRYAKEVEVDRLVN
Ye8081 VADRRETMGDTAMVFDGLRLMKNGYGEQISKWFNRTLLP--KVIADRDGLAFHSFRHTVATQLKQHGIELAYAQAIMGHSSGSITYDRYAKEVEVDRLVN
T-2    VETRRGADGDDAMVFDGLRLMKNGYGEQVSKWFNRTLLP--KVLVDRSGLAFHSFRHTVAAQLKQHGVELAYAQAIMGHSSGSITYDRYAKEVEVDRLVN
Ty2    VETRRGADGDDAMVFDGLRLMKNGYGEQVSKWFNRTLLP--KVLADRSGLAFHSFRHTVATQLKQHGVELAYAQAIMGHSSGSITYDRYAKEVEVDRLVN
T-3    VEDRRKTEGDNAMVFNGLKLMKNGYGEQVSKWFNRTLLP--KVLADRSGLAFHSFRHTVATQLKQHGVELAYAQAIMGHSSGSITYDRYAKEVEVETLKE
301-3  VEDRRKTEGDNAMVFNGLKLMKNGYGEQVSKWFNRTLLP--KVLADRSGLAFHSFRHTVATQLKQHGVELAYAQAIMGHSSGSITYDRYAKEVEVETLKE
CT18   VEDRRKAKGDNAMVFDGLKLMKNGYGEQVSKWFNRTLLP--KVLADRSGLAFHSFRHTVATQLKQHGVELAYAQAIMGHSSGSITYDRYAKEVEVDRLVN
BS512  VEDRRKAEGDNAMVFDGLKLMKNGYGEQVSKWFNRTLLP--KVLADRSGLAFHSFRHTVATQLKQHGVELAYAQAIMGHSSGSITYDRYAKEVEVDRLVN
T-1    VD----MQPDEGPLFG---MTPSAY----SSWFNSRVLT--EALGDSQDVSLHSLRHWLATRMKERGVNLVDAQGILGHSSQSITYDLYGKGHAVGRLAD
301-1  VD----MQPDEGPLFG---MTPSAY----SSWFNSRVLT--EALGDSQDVSLHSLRHWLATRMKERGVNLVDAQGILGHSSQSITYDLYGKGHAVGRLAD
E22    VD----AQPDDGPLFG---MTPSAY----SSWFNSRVLT--EALPDADNVSLHSLRHWLATRMKERGVNLVDAQGILGHSSQSITYDLYGKGHAVGRLAD
ECA    VN----TREPKTAIFG---MTAGTY----TAYFGRSIKGSIDALKDTKNVCMHSLRHSLTGKLKAAGVPLADAQGVLGHSSQSITYDLYGKGHAIGRLSD

CR     VMVGIYKEIK----------------
Ye8081 VMTDAYKENRVNG-------------
T-2    VMADVYKET-----------------
Ty2    VMADVYKEI--VG-------------
T-3    KLAESLSVKKIDGK------------
301-3  KLAESLSVKKIDGK------------
CT18   VMADVYKETLVNGDTIHCTSVSRRGI
BS512  VMAGVYKETGVNG-------------
T-1    VLKTALL-------------------
301-1  VLKTALL-------------------
E22    ALNLALVEV-----------------
ECA    ALRLALSGTGQ---------------
```

Figure S2: Alignment of the integrases.

The integrases from the 12 T7 islands were aligned using the program T_Coffee (5), and the alignment was reformatted using the program Boxshade. Different integrases are represented by the corresponding island names. The integrases of the islands T-1 and 301-1 have been split into two parts by an insertion sequence (Figure 4). These parts were merged and used for this alignment. The red triangle indicates the break point, which is close to the junction (indicated by two orange triangles) of the N-terminal and C-terminal domains.



A Hyp1

```
BS512    -MAEYYPAVFEAEAFNCPHCGVYARQFWRSMYGNSNVLAVKSTEFRMSTCSHCGEDAYWYQGNMLIPAAGNVELPNPDMP
CR       MAVPYEPAEFKKEAFNCPYCQAYAKQKWGQLFPYYEET---AFPMHVSQCERCEEYSYWFEESLLIPASANVEMPNPDMP
E22      -MAEYYPATYGSKAFNCPYCDAYSQQDWSRLKYGYN--GQYDSPFVFSKCEHCNRKAYWYEETMLIPAAANIELPNVDMP
CT18     --------------------------------------------MDLSICSRCEEKAYWYDEKLIVPQSSTVEMPNPDMP

BS512    DDCKSDYMEARSIINLSPKGAAALLRLCLQKLMVHLGEPGNNINADIRSLV-QKGLPVR--IQQAADICRIVGNQAVHPG
CR       DDCKADYMEARSIVNLSPKGAAALLRLCLQKLMVHLGEPGKNINDDIKSLV-EKGLPPR--IQQAADICRIVGNQAVHPG
E22      DNCKSDYMEARSIINLSPKGAAALLRLSLQKLMVHLGEPGKHIDFLYAAVFLPHKIAPDFGLSETLNHCLLVGNQAVHPG
CT18     DDCKSDYMEARSIINLSPKGAAALLRLSLQKLMVHLGEPGKHIDTDIKSLV-AKGLSPL--VQRAADICRIVGNQAVHPG

BS512    EISLDDDPQLAHGLFKLLNIIVTEQITRPKEIEAMFQSMPEGPRQGIENQDRQAREQQQAANE
CR       EISLDDDPQLTHGLFKLLNIIVDDRITRPKEIEAMFQSMPEGPRQGIEKRDKKSA--------
E22      EINIDDDPQLAHGLFKLLNIIVTEQITRPKEVEAMFNSMPERALKGIEDRDRKAREQQQAANE
CT18     EINIDDDPQLAHGLFKLLNIIVTEQITRPKEVEAMFNSMPERALKGIEDRDRKAREQQQAANE
```

B Hyp2

```
T-3      M----------SQQEISII-------------------------------------------NLDQLVSMTSVEIAELTG
301-3    M----------SQQEISII-------------------------------------------NLDQLVSMTSVEIAELTG
BS512    MTKNTRFSPEVRQRAIRMVLESQGEYDSQWAAICSIAPKIGCTPETLRVWVRQHERDTGGGSNLDQLVSMTSIEIAELTG
ECA      -------------MDTLII-------------------------------------------NLDQMVSMTSLEIADLTG
Ye8081   -------------MNNNII-------------------------------------------NLNQVVSMTSLEIADLTG
Ty2      M----------SQQEISII-------------------------------------------NLDQLVSMTSVEIAELTG
E22      M----------SQQ---VV-------------------------------------------IFNDEFSLSSYEFLT---

T-3      KEHKHVLRDIRNMVEELNGAKTEHCSTLSSELNGSKFGLVGEEVYKDAKGESRTMYRLDRKHTFILVAGYSVHLRAKCYD
301-3    KEHKHVLRDIRNMVEELNGAKTEHCSTLSSELNGSKFGLVGEEVYKDAKGESRTMYRLDRKHTFILVAGYSVHLRAKCYD
BS512    KEHRNVLRDIRNMAEELNALKTEQCSKLSSHIG------VTKDVYLNAQGKQQPLYRLDRKHTFILVAGYSVHLRAKCYD
ECA      KRHGNIVRDIRRMLDDLNTLN-EIDSNLSQRTG------VEEEVYLDDQGRQQPYYKLDRKHTFILVSGYSVQLRAKCFD
Ye8081   KRHDHVIRDVRKMVQDLDT---------APKNG------VSEENYVDPTGRQLPMYRLDRKHAFILVSGYSVHLRAKCYD
Ty2      KEHRNVLRDIRNMVEELNALKTEHCSKLSSPIG------VIEDVYLNAQGKQQPLYRLDRKHTFILVAGYSVHLRAKCYD
E22      ----KVINPAREEAGENPVSNKDFINRVKDELDLKEENFLLLDT--GASGRKASHTILNGDQLLLVGMRESKAVRRKVLD

T-3      HIQTLERRVLQLEDQKKRAAIQSANRRGVTWGDYCKTYGLPAQKLMTALLQHRGLFRKNPISNEWSVNPKYSDCFRIIKP
301-3    HIQTLERRVLQLEDQKKRAAIQSANRRGVTWGDYCKTYGLPAQKLMTALLQHRGLFRKNPISNEWSVNPKYSDCFRIIKP
BS512    HIQTLERRVLQLEDQKKRAAIQSANRRGVTWGDYCKTYDLPAQKLMTALLQHRGLFRKNPISNEWSVNPKYSDCFRIIKP
ECA      HIDKLEREILRLEDQHKRVAIQSANRRGVTWGDFCKTHGLPAQRLMDILKQERRLFRVSPYNGEWSVNPHYEDCFRVIKR
Ye8081   HIQALEQQVLQLEDQKKRAAVQSANRRGVTWGDYCKTVGLPTQKLMHILKKERKLFWVNPISGEWSVKPAFSNYFTVINP
Ty2      HIQTLERRVLQLEDQKKRAAIQSANRRGVTWGDYCKANGLPAQKLMTILKKERKLFRVHSSSGEWSVNPNYVEYFRIIKP
E22      YIRRIEKDKQLLEDQKKRAAIQSANRRGVTWGDYCKTYGLPAQKLMTALLQHRGLFRKNPISNEWSVNPKYSDCFRIIKP

T-3      SDQKFSAGGYNFRFNAKGLEVFGKPEMVDKMRGILIAFTGTDQQKQEHLLKLAQSGKVEGI-
301-3    SDQKFSAGGYNFRFNAKGLEVFGKPEMVDKMRGILIAFTGTDQQKQEHLLKLAQSGKVEGI-
BS512    SDQKFSAGGYNFRFNAKGLEVFGKPAIVDKLRGILIAFTGTDQQKQEHLLKQAQSGKLEGL-
ECA      TDNRFSAKGINIRFNAKGLETFSAPEKLIKFHQKLVIRYGSDLDKQRLLQDEARMRKAGIIQ
Ye8081   SNQRFSPKGINIRFNAKGLEYFCQPENVHKFREKLVIHGGTDIEKQRLLQKVAQSR------
Ty2      TDHRFNPNGINIRFNAKGLEFFSRPENVLKMHRKVIAVHGSDAAKQQHIQAVAKLEGR----
E22      SDQKFSAGGYNFRFNAKGLEVFGKLELVDKMRGILIAFTGTDQQKQEHLLKQAQSGKLEGL-
```

C Hyp3

```
301-1    M-------AMIDPRTPIGKATLRYRGLPTRHLLSLLRLGVEDPE-RPYYSRDELIAMLVDRDLDNQLRRAFAKQS
T-1      M-------AMIDPRTPIGKATLRYRGLPTRHLLSLLRLGVEDPE-RPYYSRDELIAMLVDRDLDNQLRRAFAKQS
CT18     M-------AMIDPRTPSGKLTLRYRGLPTSILLSMLNLDKDATNGRPFYSRNELIEQLVIRDMDINRRNK-----
301-3    --------MMIDPRTPEGRMTLRYRGYRTEVLLRELGLDPEDET-RQHQSRDELIAQLVAMKLPLNR--------
T-3      --------MMIDPRTPEGRMTLRYRGYRTEVLLRELGLDPEDET-RQHQSRDELIAQLVAMKLPLNR--------
BS512    ---------MIDPRTPEGRMTLRYRGYRTEVLLRELGLDPEDET-RQHQSRDELIVQLVAMKLPSASKLAPNR--
E22      M---IDQRPMVDPRTKAGRMTLRYRGYRTEVLLKELGLDPEDET-RQHQSRDELIAQLVAMKLSQA---------
ECA      MTTQHHHSAQIDPRTPEGRQALNLMTIKTSALVSKLGLPPKHDR-ADYYSKGALCLMAVSAGLSPKDFF------
Ty2      MTNNTNDTIKIDPRTPEGRKALRLMVVPPKALIATLGLPAKENR--PYYSKAALCLMAVDAGLTPRDFM------
```

Figure S3: Alignments of Hyp1, Hyp2 and Hyp3 proteins.

The C-terminal region of E22-Hyp2 aligns well with the other Hyp2 proteins, while the N-terminal region aligns poorly, suggesting that E22-Hyp2 is a recombinant protein. Only the C-terminal region (starting after the red triangle) of this alignment was used to infer a tree for Hyp2 proteins (Figure 9A).

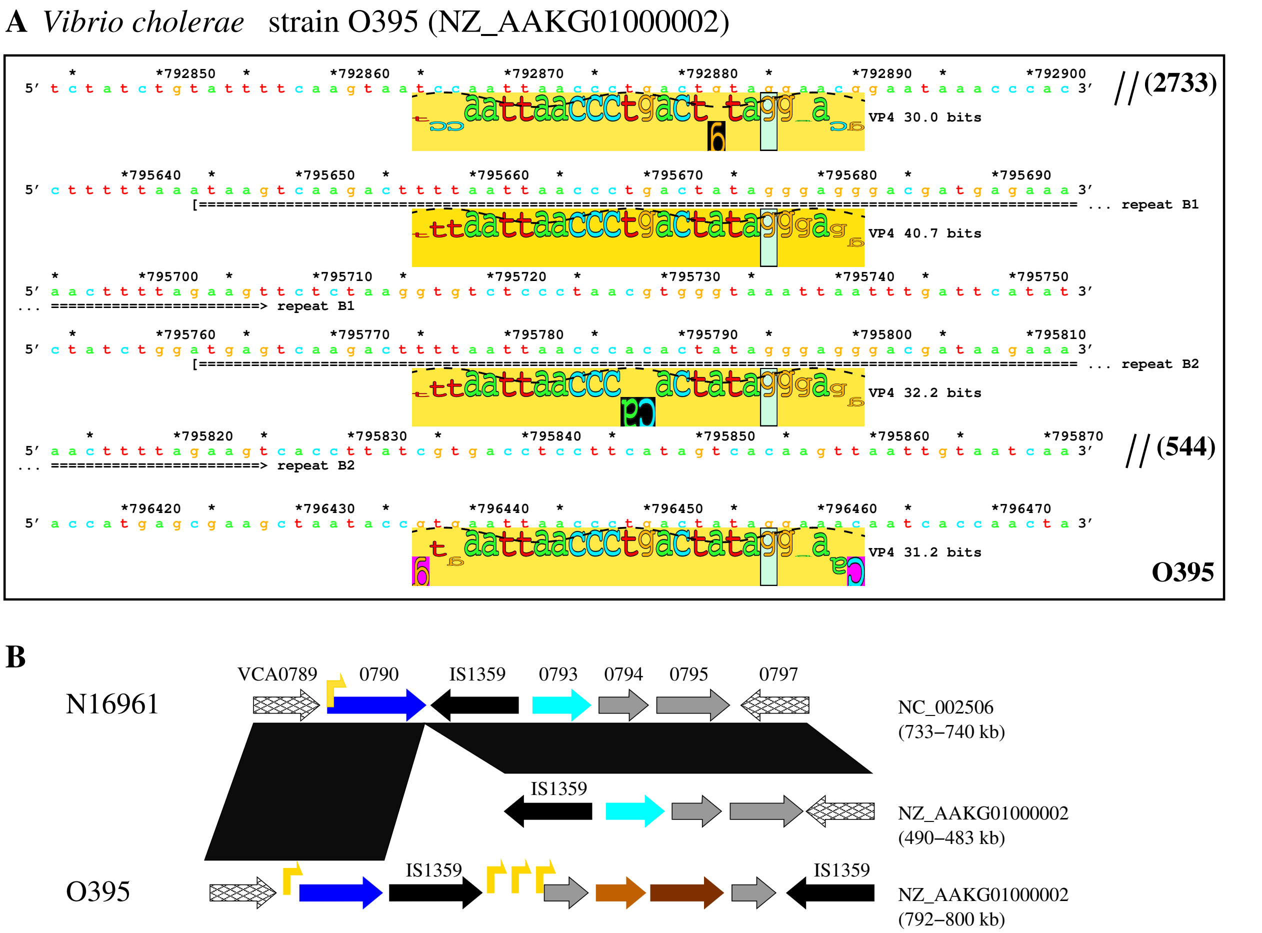
A Hyp4

```
T-1      MMFNNNNWKLSVTDINLYENTVSLDGQPYPLSLAIKTLIPGYLSGLPSTSREAMELLEALAEAGVTIGNFFSNDLMTAYGRR
301-1    MMFNNNNWKLSVTDINLYENTVSLDGQPYPLSLAIKTLIPGYLSGLPSTSREAMELLEALAEAGVTIGNFFSNDLMTAYGRR
CT18     MLFNNNDWKLSVTDINPYENTVSLDGQSYPLSLAIKTLIPGYLSGLPSTSRAAMELLEALAEAGVSIGNFFSNDLMTAYGRR
T-3      MMFNNNNWKLSVTDINLYENTVSLDGQSYPLSLAIKTLIPGYLSGLPSTSREAMELLEALAEAGVTIGNFFSNDLMTAYQRR
301-3    MMFNNNNWKLSVTDINLYENTVSLDGQSYPLSLAIKTLIPGYLSGLPSTSREAMELLEALAEAGVTIGNFFSNDLMTAYQRR
BS512    MLFNNNEWKLSVTDIDLYANTCKLDGQSYPLSLAIKTLIPGYLSGLPSTSREAMELLEALAEAGVAIGNFFSNDLLTAYQRR
E22      MLFDQNDWKLQVTDIDLYANTCKLDGESYPLSLALKTLIPGYLSGLSPTSSASMELLEALAEAGVTISNFFSNDLLTSYQRR
ECA      --MSIAYRKLDIT-LSADKETVLVFGQELSTKYFCEVVIPTMLNGCGNDAGKTNSILNDVHAAGLNAGDYTTFSRWWSESNA
Ty2      --MSISYRKLDIA-LSADKETVLVFGQELSTKYFTEIVVTTMLNSTGSDMANSNRILNDIHAAGLDAGDYGKYSRWWAQSNA

T-1      QMNKRAEAERIAKEQRLQAERMREENMTDAEWQKEL-QRREQVKAERRTYGESLRSATHSAGRSRAAIVADLESGGNWMDSL
301-1    QMNKRAEAERIAKEQRLQAERMREENMTDAEWQKEL-QRREQVKAERRTYGESLRSATHSAGRSRAAIVADLESGGNWMDSL
CT18     QQNKRAEAERIAKEQRIQAERMREENMTDAEWQKEL-QRREQVKAERRTYGEHLRSATHSAGRSRASMVADIESTDNWMDSL
T-3      QMNKRAEAERIAKELASQKERTREMFMTEDEWQKEL-QRREQVKAERRTYGENLRSATHSAGRSRAAIVADLESGGNWMDSL
301-3    QMNKRAEAERIAKELASQKERTREMFMTEDEWQKEL-QRREQVKAERRTYGENLRSATHSAGRSRAAIVADLESGGNWMDSL
BS512    QMNKRVEAERIAKELASQKERTREMFMTEDEWQKEL-QRREQVKAERRTYGENLRSATHSAGRSRASIMADLDSGANWMDSL
E22      QVNKQREAERIEREQRVLAERMAELHMTEEERIKAN-QKRDQQKAERHAYGDSIRNAMSSTGRSRAAKLVEIDGMDNWMDSL
ECA      Q--ARQEAERKRIEAEQHRERMAAMHATPAEIAAERAEKARRVEDAQRKFG-----------HKGAAFGL------------
Ty2      Q--ERQEAERRRKEAKAHQERMAAIHATPEEIAKAVAERKAREEALIKRFG-----------NKGAAFGL------------
```

B Hyp5

```
T-1      MT------------------TIKDA-----------FQFGI--------EPVRITDTDNI-QVNEGL--------PTNAD
301-1    MT------------------TIKDA-----------FQFGI--------EPVRITDTDNI-QVNEGL--------PTNAD
T-3      MN------------------MTKNA-----------FQFGI--------EPVRITDTDNI-QVNEGL--------PTNAD
301-3    MN------------------MTKNA-----------FQFGI--------EPVRITDTDNI-QVNEGL--------PTNAD
BS512    MT------------------MTKDA-----------LQFGI--------EPVRITDTDNM-QVNEGL--------PTNAD
ECA      MNTRSNNQIKRIDTTQPKVKTLDDLKTHMVDAKGQKVPNSVWGTDATRTEIVSMPDSTGG-ILNVGYTDRSGRPQPFGND
Ty2      M--SITNQIN-------KASSLASLRQPQRDKDGQIIKGSIWGTDISRTDYVQMTNGQDAQVLNVGFTDRSGRPMAFGND
E22      MT------------------TIKDA-----------FQFGI--------EPARITNTDNI-QVSES---------GAATE
CT18     -M------------------TIKDA-----------FQFGI--------EPVRITDTDNI-QINEGL--------PTNAD
T-2      M-TT----------------------------------------------------------------------------

T-1      PQVYALQLAKTVKAMLN-GVLKDAQENIPFPVEVLPTRNSLPTPIIAHTLADRSVVVPVRGGK-RP-EVVTVPS--GQEI
301-1    PQVYALQLAKTVKAMLN-GVLKDAQENIPFPVEVLPTRNSLPTPIIAHTLADRSVVVPVRGGK-RP-EVVTVPS--GQEI
T-3      PQVYALQLAKTVKAMLN-GVLKDAQDNIPFPVEVLPTRNSLPTPIIAHTLADRSVLVPVRGGK-RP-EVVTAPS--GTEI
301-3    PQVYALQLAKTVKAMLN-GVLKDAQDNIPFPVEVLPTRNSLPTPIIAHTLADRSVLVPVRGGK-RP-EVVTAPS--GTEI
BS512    PQVYAFQLAKLVKTMLN-SVLKDAQDNIPFPVEVLPTRNSLPTPIIAHTLADRSVLVPVRGGK-RP-EVVTAPS--GTEI
ECA      VHDYIQALEASLEATFNDGDFADAMQQHIFPSEFRPF-AALPTPLNVQLLEDRTVTFNPSGKRDRS-NLPTAKAIAGSAV
Ty2      SHDFVQALEDSLDETFNDGDFRTAVMENVFPCEFRPY-GTDPTPIRRQALQNRTVRFKHNGAFNPAKDAPSTTAIKGASV
E22      PQVYALTLAKTVKSMLN-SVLKDAQDNIPFPVELLPTRNSLPTPIVSHLLADRSVVVPVRGGK-AP-TVVTATS--GTEI
CT18     PQVYAFELAKLVKTMLN-GVLKSAQENIPFPVEVLPTRNSLPTPIIAHTLADRSVVVPVRGGK-RP-EVVTAPS--GTEI
T-2      ----------------------------------------------SNT-------------------------------

T-1      VVEPI-------EQAILISEQTKLWDAKSSTGFTQGTLQQDAMNICENVVRTINARMVDVLESSKLLKTVELPVLTGSLT
301-1    VVEPI-------EQAILISEQTKLWDAKSSTGFTQGTLQQDAMNICENVVRTINARMVDVLESSKLLKTVELPVLTGSLT
T-3      TVEPI-------EQAILVSHQTKLWDQKSTTGFTQGTLQQDALNICDNVIRTINSKMVDVLESSKLLKTVELPALTGSLT
301-3    TVEPI-------EQAILVSHQTKLWDQKSTTGFTQGTLQQDALNICDNVIRTINSKMVDVLESSKLLKTVELPALTGSLT
BS512    TVEPI-------EQAILVSHQTKLWDQKSTTGFTQGTLQQDALNICDNVVRTINSKMVDVLESSKLLKTVELPALTGSLT
ECA      MVAPVLHGDDNTQTGILCSTATHVSDFDMMGGWTDGSLHQTVANMQLQFCRIMAGNVAKVLSKTPDLVTIECDALSSKPK
Ty2      TVTPVLNPETN-KTGILCSAATHVSDFDMMGGWTEGALIQTVGDMQVQYSRQMFAAVVDVLKDTPDLQIIEAAPLSGKPS
E22      TVEPI-------EKAILVSHQTKLWDAKSTTGFTQGTLQQDALNICENVVRTINSKMVEVLESSKLLKSVDVSVLTGTLT
CT18     TVEPI-------EQAILVSHQTKLWDAKSSTGFTQGTLQQDALNICENVVRTINSKMVDVLESSKLLKTVELPVLTGSLT
T-2      --------------------------------------------------------------------------------

T-1      AKADAIMDALYENTESSFGSEVSDYGIIAHESQLKALSRLAAKQGFSGEDAIVDMLGTDIAYYNGEDKGVFMLAKRFTAL
301-1    AKADAIMDALYENTESSFGSEVSDYGIIAHESQLKALSRLAAKQGFSGEDAIVDMLGTDIAYYNGEDKGVFMLAKRFTAL
T-3      AKADAIMDALYENTESSFGSEVSDYGIIAHESHLKALSRLAAKQGFGGEDAIVDMLGTDVAYYNGEDRGVFMMAKRFTAL
301-3    AKADAIMDALYENTESSFGSEVSDYGIIAHESHLKALSRLAAKQGFGGEDAIVDMLGTDVAYYNGEDRGVFMMAKRFTAL
BS512    AKADAIMDALYENTESSFGSEVSDYGIIAHESHLKALSRLAAKQGFGGEDAIVDMLGTDIAYYNGEDRGVFMMAKRFTAL
ECA      DAAEDLLDYLAINLPVHLGATLDAYALMVPEKLEAVLERAAQRA--GHED-ASELFGCTIMGYLGEDTGVYLLPKGFAML
Ty2      DQAEDLLDTLALNLPVELGNTLSDYAVLVPERLEAILDRAAQRA--GHED-ISELLGCTVCSYAGDDTGVYLLPKRFASI
E22      ERAEHILDALYENTESAYGSEVTDYGVIVHESHLKALSRLAAKQGFGGEDAITDMLGTDLAYYNGTDRGIFMLAKRFTCL
CT18     AKADTIMDALYDNTESSFGSEVADYGIIAHESHLKALSRLAAKQGFGGEDAIVDMLGTDIAYYNGADRGVFMMAKRFTAL
T-2      ---------------------LSDYAVLVPERLEAILDRAAQRA--GHED-ISELLGCTVCSYAGDDTGIYLLPKRFASI

T-1      SFGCFRHDGENITVVLSRDGDSQSHDLEILGKVFVVAEAATTIKMGTGS--ATAVLPVVKRLKFTKTEA-----
301-1    SFGCFRHDGENITVVLSRDGDSQSHDLEILGKVFVVAEAATTIKMGTGS--ATAVLPVVKRLKFTKTEA-----
T-3      SFGCFRHDGESITVVLSRDGDSQSHDLEILGKVFVVAEAATTIKMGTGS--ATAVLPVVKRLSFTKEAN-----
301-3    SFGCFRHDGESITVVLSRDGDSQSHDLEILGKVFVVAEAATTIKMGTGS--ATAVLPVVKRLSFTKEAN-----
BS512    SFGCFRHDGENVTVVLSRDGDSQSHDLEILGKVFVVAEAATTIKMGTGS--ATAVLPVVKRLSFTKDS------
ECA      SFRSTKED-DTVKVIVTRDPNRAGYDVELITVIDVMATGSVKVKQGEFNVEATAEFPVVHRLTFKSA-------
Ty2      SFRSTKDA-KTVDVKVTRNSNTAGYDLELISVVDVLATGSVKVKAGEFDVEKDASFPLIHVIRFTTPE-------
E22      SFGCFRHDGEHITVVLSRDGDSQSHDLEIFGKIFVVAEAATTIKMTSGS--ATAVLPVVKRLKFATK-------
CT18     SFGCFRHDGENITVVLSRDGNTQSHDLEILGKVFVVAEAATTVKMGTGS--AAAVLPVVKRLSFTKTS------
T-2      SFRSTK-DAKTVDVKVTRNSNTAGYDLELISVVDVLATGSVKVKAGEFDVEKDASFPLIHVIRFTTPRVTINPE
```

Figure S4: Alignments of Hyp4 and Hyp5 proteins.



A *Vibrio cholerae* strain O395 (NZ_AAKG01000002)

* *792850 * *792860 * *792870 * *792880 * *792890 * *792900
5′ t c t a t c t g t a t t t t c a a g t a a t c c a a t t a a c c c t g a c t g t a g g a a c g g a a t a a a c c c a c 3′
// (2733)
VP4 30.0 bits
*795640 * *795650 * *795660 * *795670 * *795680 * *795690
5′ c t t t t t a a a t a a g t c a a g a c t t t t a a t t a a c c c t g a c t a t a g g g a g g g a c g a t g a g a a a 3′
... repeat B1
VP4 40.7 bits
* *795700 * *795710 * *795720 * *795730 * *795740 * *795750
5′ a a c t t t t a g a a g t t c t c t a a g g t g t c t c c c t a a c g t g g g t a a a t t a a t t t g a t t c a t a t 3′
... ========================> repeat B1
* *795760 * *795770 * *795780 * *795790 * *795800 * *795810
5′ c t a t c t g g a t g a g t c a a g a c t t t t a a t t a a c c c a c a c t a t a g g g a g g g a c g a t a a g a a a 3′
... repeat B2
VP4 32.2 bits
* *795820 * *795830 * *795840 * *795850 * *795860 * *795870
5′ a a c t t t t a g a a g t c a c c t t a t c g t g a c c t c c t t c a t a g t c a c a a g t t a a t t g t a a t c a a 3′
... ========================> repeat B2
// (544)
*796420 * *796430 * *796440 * *796450 * *796460 * *796470
5′ a c c a t g a g c g a a g c t a a t a c c g t g a a t t a a c c c t g a c t a t a g g a a a c a a t c a c c a a c t a 3′
VP4 31.2 bits
O395

B

N16961
VCA0789
0790
IS1359
0793
0794
0795
0797
NC_002506
(733–740 kb)
IS1359
NZ_AAKG01000002
(490–483 kb)
O395
IS1359
IS1359
NZ_AAKG01000002
(792–800 kb)

Figure S5: Putative T7 islands in *Vibrio cholerae* genomes.
(A) Sequence walkers of tandem VP4 promoters in the island O395. (B) Genome organization of the putative T7 islands N16961 and O395. Symbol key is given in Figure 4.

REFERENCES

1. Womble, D. D. (2000) GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol Biol,* **132**, 3–22.

2. Chen, Z. and Schneider, T. D. (2005) Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. *Nucleic Acids Res,* **33**, 6172–6187.

3. Scholl, D. and Merril, C. (2005) The genome of bacteriophage K1F, a T7-like phage that has acquired the ability to replicate on K1 strains of *Escherichia coli*. *J Bacteriol,* **187**, 8499–8503.

4. Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.,* **18**, 6097–6100 http://www.ccrnp.ncifcrf.gov/˜toms/paper/logopaper/.

5. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.,* **302**, 205–217.