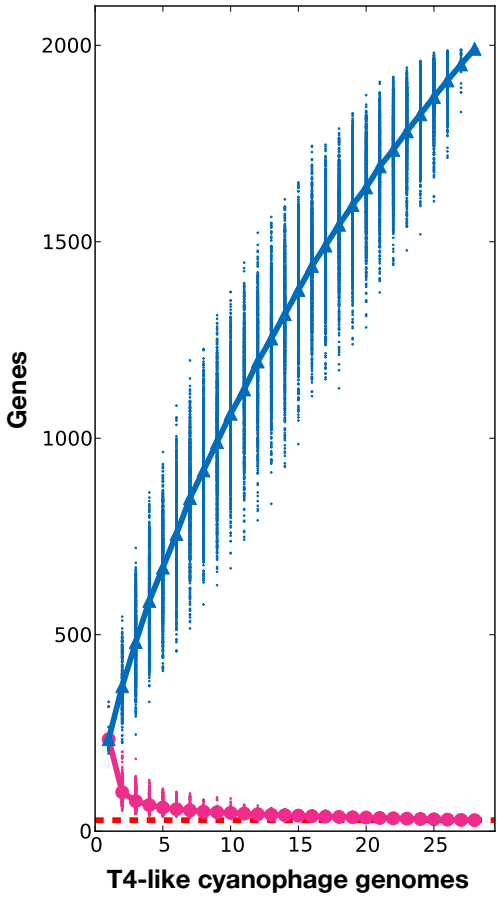
SUPPLEMENTARY MATERIALS AND METHODS

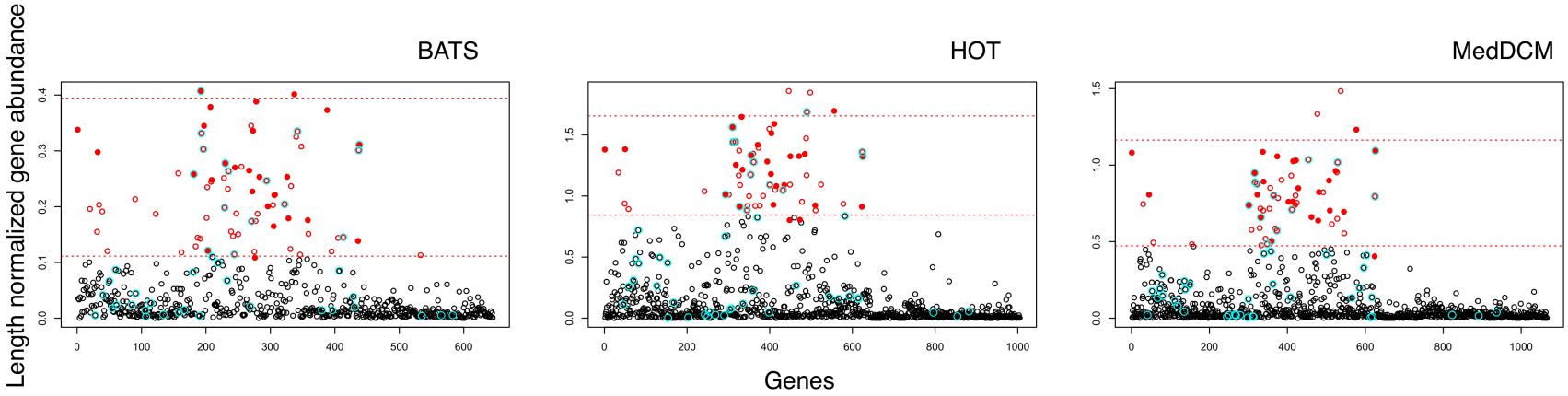
Identification of the PhoBR two component system in cyanobacterial hosts

To examine whether cyanophage genes containing *pho* box motifs were found in phage isolated from hosts with the PhoBR phosphate sensing and response system we systematically identified *phoR* and *phoB* genes in host genomes. Many sequenced *Prochlorococcus* and most *Synechococcus* genomes contain *phoR* and *phoB* genes (Kettler *et al.*, 2007). The recently sequenced and unpublished *Synechococcus* WH8109 does not have annotated *phoR* and *phoB* genes and so was probed for evidence of the PhoBR two-component system as follows. We used blastp to identify the reciprocal best BLAST hits between a custom database of 9,604 cyanobacterial orthologous groups from the ProPortal database and all genes in the *Synechococcus* WH8109 genome (Multi-FASTA file of peptides downloaded from Moore Foundation Microbial Genome Sequencing Project, http://moore.jcvi.org/, Genbank project accession: ACNY00000000) to identify *Synechococcus* WH8109 genes with reciprocal best hits to *Prochlorococcus* or *Synechococcus phoR* or *phoB* sequences.

SUPPLEMENTARY FIGURES



Supplementary Figure 1: Core and pan genomes for 28 cyanomyovirus genomes. Core genes, shared by all 28 genomes, are shown in pink; pan genome genes, found in one or more cyanomyovirus genomes, are shown in blue. Triangles and circles represent the average of the combination of values for each subset of genomes; each individual value is indicated by a small dot. The red dotted line hits the X-axis at 28, the total number of genes (26 single copy, and two multi-copy) shared by all 28 genomes.



Supplementary Figure 2: Length-normalized cyanomyovirus gene abundance in three environments. In all plots, red dashed lines indicate the first and third quartile cutoffs for core-like genes defined by the abundance of culture-defined core genes. Red filled circles are signature core genes; open red circles are metagenome-defined core genes, a few are present at levels higher than the culture defined core boundaries. A cyan outline indicates genes annotated with Pfam domains that are shared between phage and host genomes.

SUPPLEMENTARY TABLES

Supplementary Table 1: Cyanomyovirus genes found at signature core gene frequencies in one or two environmental metagenomic datasets.

Gene cluster	Specific to	Gene name	In host?	Cyanomyovirus genes in cluster	Pfam annotation	Pfam domain description	ProPortal gene cluster description
PhCOG204	BATS			7			Hypothetical protein
PhCOG703	BATS		Y	2			Hypothetical protein
PhCOG1424	BATS			3			Hypothetical protein
PhCOG71200	BATS			6			Hypothetical protein
PhCOG71673	BATS	Hli04	Y	3			High Light Inducible Proteins
PhCOG72544	BATS	2OG-Fe(II) oxygenase	Y	25	PF03171	2OG-Fe(II) oxygenase superfamily	2OG-Fe(II) oxygenase
PhCOG73152	BATS	pstS	Y	16	PF01547	Bacterial extracellular solute-binding protein	Phosphate transporter subunit
PhCOG71491	HOT	Dam	N	13	PF01555	DNA methylase	DNA adenine Methylase
PCOG72385	HOT			11			Hypothetical protein
PCOG71457	HOT	HN	Y	9			Hemagglutinin domain
PCOG71521	HOT			18			Hypothetical protein
PhCOG72994	HOT			15			Hypothetical protein
PhCOG258	MedDCM			16			Hypothetical protein
PhCOG1139	MedDCM			9			Hypothetical protein

Gene cluster	Specific to	Gene name	In host?	Cyanomyovirus genes in cluster	Pfam annotation	Pfam domain description	ProPortal gene cluster description
PhCOG71460	MedDCM			23			Hypothetical protein
PhCOG72482	MedDCM			19			Hypothetical protein
PhCOG224	BATS/HOT			11			Hypothetical protein
PhCOG3728	BATS			1			Hypothetical protein
PhCOG71554	BATS	PsbD	Y	20	PF00124	Photosynthetic reaction centre protein	photosystem II D2 protein
PhCOG71750	BATS	gp7		15			gp7
PhCOG72417	BATS/HOT	gp59		22	PF08994/PF08993	T4 gene 59 helicase, C terminal/T4 gene 59 helicase, N terminal	loader of gp41 DNA helicase
PhCOG73121	BATS		Y	27			Hypothetical protein
PhCOG198	BATS/HOT			16			Hypothetical protein
PhCOG71490	MedDCM		Y	24	PF04851/PF00270/PF00176/PF02562	Type III restriction enzyme, res subunit/DEAD/DEAH box helicase/SNF2 family N-terminal domain/PhoH	Helicase
PhCOG71360	HOT/MedDCM			26			Hypothetical protein
PhCOG71523	HOT/MedDCM	CP12	Y	24	PF02672	CP12	CP12
PhCOG72738	HOT/MedDCM			16			Hypothetical protein
PhCOG72130	BATS/MedDCM			16			Hypothetical protein
PCOG71682	BATS/MedDCM	talC		27	PF00923	Transaldolase	Transaldolase-like protein

Gene cluster	Specific to	Gene name	In host?	Cyanomyovirus genes in cluster	Pfam annotation	Pfam domain description	ProPortal gene cluster description
PhCOG71394	BATS/MedDCM	gp62		26			clamp loader subunit
PhCOG173	BATS/MedDCM			40			Hypothetical protein

Supplementary Table 2: Predicted pho boxes within 100bp upstream of genes in 28 cyanomyovirus genomes.

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
MED4-213	HQ634174	CPMG_00122	Orphan_1251	-	118107:118137	CTAAATAGGTTTTTTACCATTCTATGAACT
MED4-213	HQ634174	CPMG_00031	PhCOG71052	-	45027:45057	GTTAAGCTAAATATTATATGGAATTCTTAG
MED4-213	HQ634174	CPMG_00195	PhCOG71393	-	164324:164354	ATTTTTCTTGATTTCTATAAAATTTACCTA
MED4-213	HQ634174	CPMG_00061	PhCOG71488	+	77341:77371	TATAAAAACCAATTAAATTAGTGTCACACA
MED4-213	HQ634174	CPMG_00003	PhCOG72320	-	3222:3252	TATTATTAAATCTTAACGCAATCTAATTTT
MED4-213	HQ634174	CPMG_00003	PhCOG72320	-	3233:3263	TTTATAAATAATATTATTAAATCTTAACGC
MED4-213	HQ634174	CPMG_00144	PhCOG72398	-	128681:128711	AGTAATTTTTACTTTGTATAATGTTATACA
MED4-213	HQ634174	CPMG_00052	PhCOG72481	-	74197:74227	CTTACGTAATTGTTACGTATTGTTTCAACT
P-HM1	NC_015280	PHM1_050	PhCOG2046	+	34398:34428	ATTTGGTAGATTTTGGAAATATCTTAAAAA
P-HM1	NC_015280	PHM1_091	PhCOG71052	+	67888:67918	GTTAAGCTAAATATTATATGGAATTCTTAG
P-HM1	NC_015280	PHM1_119	PhCOG72320	+	110167:110197	TATTATTAAATCTTAACGCAATCTAATTTT
P-HM1	NC_015280	PHM1_199	PhCOG72398	+	165778:165808	CATAAAGAAAACTTAAGACTGTCACAAAAC
P-HM2	NC_015284	PHM2_075	Orphan_264	+	41500:41530	ATGCACTGCTCTTTAAGTCGAACTTAAGCA

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-HM2	NC_015284	PHM2_075	Orphan_264	+	41511:41541	TTTAAGTCGAACTTAAGCAGCGATCCAGTC
P-HM2	NC_015284	PHM2_211	Orphan_282	+	171548:171578	ATTGACTTCTACGTAAAGTTTTGTTAATAT
P-HM2	NC_015284	PHM2_052	PhCOG2046	+	35576:35606	TTTCATAGTGATTTCATAGTCATTTAGAAA
P-HM2	NC_015284	PHM2_052	PhCOG2046	+	35587:35617	TTTCATAGTCATTTAGAAATGTATAAAAAA
P-HM2	NC_015284	PHM2_090	PhCOG71052	+	68113:68143	GTTAAGCTAAATATTATATGGAATTCTTAG
P-HM2	NC_015284	PHM2_072	PhCOG71068	+	40759:40789	TATAATAGTAGTATAAACAAACATTTACAT
P-HM2	NC_015284	PHM2_072	PhCOG71068	+	40770:40800	TATAAACAAACATTTACATCACATTAGATT
P-HM2	NC_015284	PHM2_238	PhCOG71329	+	181095:181125	ATTTAGACGACCTTGCTGATGAGTTATTCG
P-HM2	NC_015284	PHM2_121	PhCOG72320	+	114365:114395	TATTATTAAATCTTAACGCAATCTAATTTT
P-RSM1	HQ634175	CPPG_00012	Orphan_1256	-	11172:11202	ATTAAAATAAATGTAGTTATGTGTTTAATC
P-RSM1	HQ634175	CPPG_00017	PhCOG67	-	12637:12667	ATTGATTTTAATCTCATTACTCTTTATAAT
P-RSM1	HQ634175	CPPG_00026	PhCOG72994	-	14733:14763	ATTCAGAGTTATTTTACTCTATTATAATAA
P-RSM1	HQ634175	CPPG_00172	PhCOG73251	-	114014:114044	CTTACAGTGTTATTGTACATGTATTTGACA
P-RSM3	HQ634176	CPQG_00143	PhCOG186	-	114709:114739	TTTCAATAGTTCTTATAATTAGTATCGTAA
P-RSM3	HQ634176	CPQG_00041	PhCOG71490	-	42770:42800	TTTTATGCTATAATGATTATAAGTTAAATT
P-RSM3	HQ634176	CPQG_00119	PhCOG71682	-	98468:98498	TTTTTTTATGTCTTTATATACACATAGGTT
P-RSM3	HQ634176	CPQG_00026	PhCOG71906	-	38160:38190	TTTCAAATAATATTATACTATTATAATAGT
P-RSM3	HQ634176	CPQG_00026	PhCOG71906	-	38171:38201	TTTTACTACCATTTCAAATAATATTATACT
P-RSM3	HQ634176	CPQG_00122	PhCOG72385	-	99448:99478	CTCCCTCTTTTCTTAAGATTACATTAAGTG
P-RSM3	HQ634176	CPQG_00112	PhCOG72398	-	95861:95891	TATAATATGTACATAACTTTACATTAGTTA

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-RSM3	HQ634176	CPQG_00123	PhCOG73152	-	100668:100698	GTTAAATGTAGTTAAAGATACAGTTAAAAT
P-RSM3	HQ634176	CPQG_00036	PhCOG929	-	40576:40606	CTTCCATTTTAATTATTCATCCATTACAAT
P-RSM4	NC_015283	PRSM4_061	PhCOG1025	+	41560:41590	AGGAATACGTTTTTAAGTCGAACTTAAGTA
P-RSM4	NC_015283	PRSM4_079	PhCOG71068	+	49170:49200	TTTATAATAATAGTAATTAAACATTTTTAT
P-RSM4	NC_015283	PRSM4_141	PhCOG72130	+	121060:121090	ACTAAACTTTCCTTGTTATTCACTTCAAAA
P-RSM4	NC_015283	PRSM4_217	PhCOG72398	+	166652:166682	ATTATATATAATTGTAACAGTTCTTAATAA
P-RSM4	NC_015283	PRSM4_007	PhCOG72420	+	2191:2221	TTTCTAGTATAATTAAAACATACACAAGGA
P-RSM4	NC_015283	PRSM4_224	PhCOG72803	+	167815:167845	CTTTTTTTATAAATAATCTTATGATAACGT
P-RSM4	NC_015283	PRSM4_097	PhCOG72899	+	53702:53732	CTTGATATATTATTCATAATTTCATTTACA
P-RSM4	NC_015283	PRSM4_097	PhCOG72899	+	53726:53756	TTTACATCAGTATTAAAAAAATTTTGGAGA
P-RSM4	NC_015283	PRSM4_178	PhCOG73152	+	142812:142842	CTTTCGCAGTTCTTAACCTGTAGTTAAAGA
P-RSM4	NC_015283	PRSM4_178	PhCOG73152	+	142823:142853	CTTAACCTGTAGTTAAAGATACAAAACACT
P-RSM6	HQ634193	CPXG_00145	Orphan_1330	-	140661:140691	TCTCAAAATCGTTTAAACTACTATTATAGA
P-RSM6	HQ634193	CPXG_00145	Orphan_1330	-	140672:140702	ATTTACTTTAATCTCAAAATCGTTTAAACT
P-RSM6	HQ634193	CPXG_00145	Orphan_1330	-	140677:140707	TTTGTATTTACTTTAATCTCAAAATCGTTT
P-RSM6	HQ634193	CPXG_00142	PhCOG71252	-	139824:139854	GTTAATTAGGTCATTATTATTATTAAAGGT
P-RSM6	HQ634193	CPXG_00139	PhCOG71796	-	138784:138814	ATTATGTTATATTAAAAATATTTTTCGTAT
P-RSM6	HQ634193	CPXG_00139	PhCOG71796	-	138785:138815	TATTATGTTATATTAAAAATATTTTTCGTA
P-RSM6	HQ634193	CPXG_00139	PhCOG71796	-	138860:138890	TTTGTTCCAATTCTTATCATTAATTAAAAA
P-RSM6	HQ634193	CPXG_00096	PhCOG72321	-	63720:63750	TTTATATTATAAGTAAAACTATTTTTCCAA

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-RSM6	HQ634193	CPXG_00017	PhCOG72398	-	10236:10266	TTTGCTATATAATTATGTAACAGTTCTTTA
P-RSM6	HQ634193	CPXG_00017	PhCOG72398	-	10254:10284	CTTGACAGAAATTTAATCTTTGCTATATAA
P-SSM2	NC_006883	PSSM2_283	Orphan_1	+	208038:208068	CTTTCTTATTAAATAATAGTATGTTATTTC
P-SSM2	NC_006883	PSSM2_243	PhCOG2077	+	187040:187070	CTTTTCATATTTCTGAGTATTTCTTACTGT
P-SSM2	NC_006883	PSSM2_102	PhCOG611	+	69140:69170	ATTCCATCAAACTTCAAATTTTTTTTCGCT
P-SSM2	NC_006883	PSSM2_247	PhCOG173	+	189283:189313	TATAAACGGATACTAAACTTCCCTTAACCT
P-SSM2	NC_006883	PSSM2_247	PhCOG173	+	189294:189324	ACTAAACTTCCCTTAACCTTCTCTTAGTTT
P-SSM2	NC_006883	PSSM2_247	PhCOG173	+	189305:189335	CTTAACCTTCTCTTAGTTTACATATCAATT
P-SSM2	NC_006883	PSSM2_247	PhCOG173	+	189316:189346	CTTAGTTTACATATCAATTAAATTTGATAT
P-SSM2	NC_006883	PSSM2_247	PhCOG173	+	189321:189351	TTTACATATCAATTAAATTTGATATAATGT
P-SSM2	NC_006883	PSSM2_247	PhCOG173	+	189346:189376	AATGTATAGATATTACAGTTTCTTTAAAAA
P-SSM2	NC_006883	PSSM2_315.5	PhCOG72256	+	242844:242874	TATAATAACTACTAAATAATTTTTTAAAAC
P-SSM2	NC_006883	PSSM2_134	PhCOG72320	+	118001:118031	TTTAACAAATCTATAAATAAGTATAGATTC
P-SSM2	NC_006883	PSSM2_242	PhCOG72352	-	187105:187135	TTTATTATATATTATAACATACTTTTAGTA
P-SSM2	NC_006883	PSSM2_242	PhCOG72352	-	187106:187136	TTTTATTATATATTATAACATACTTTTAGT
P-SSM2	NC_006883	PSSM2_254	PhCOG72394	+	192699:192729	ATAAGAGTATTTTTTATTATTGTTTTATCA
P-SSM2	NC_006883	PSSM2_272	PhCOG72398	+	205579:205609	GTTAAGGTTTACTATATAATTATGTAAAGA
P-SSM2	NC_006883	PSSM2_054	PhCOG72455	+	48631:48661	TTTTATAATGTTTATATTATAACTTCGACT
P-SSM2	NC_006883	PSSM2_100	PhCOG72979	+	66957:66987	TTTTTGCTATACTTAATTTGTACACACGAA
P-SSM2	NC_006883	PSSM2_038	PhCOG73040	+	40786:40816	TGTAGGACAGGCTTTATAATTTTTTAAGAA

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241207:241237	ATATAATCTTTTTTAATTTATTATTTAATT
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241208:241238	TATAATCTTTTTTAATTTATTATTTAATTC
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241215:241245	TTTTTTAATTTATTATTTAATTCTTGATAA
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241218:241248	TTTAATTTATTATTTAATTCTTGATAATCA
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241267:241297	TTTCCGTTTCTGTTAATATAATCATTATAA
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241278:241308	GTTAATATAATCATTATAAATATTTAAAAT
P-SSM2	NC_006883	PSSM2_315	PhCOG73090	+	241289:241319	CATTATAAATATTTAAAATTATTGTGTGTT
P-SSM2	NC_006883	PSSM2_307	PhCOG73098	+	235236:235266	TATAAATCACCAATGATATAATATTAAACA
P-SSM2	NC_006883	PSSM2_200	PhCOG73202	+	162174:162204	ATTACTATATAGTTCAGTTGCGTTTAGCAA
P-SSM2	NC_006883	PSSM2_144	PhCOG73251	+	124862:124892	TATAAACAATTATTAATAAGGATATATTAA
P-SSM3	HQ337021	CYXG_00186	Orphan_1068	+	156076:156106	TTTATAAATAAAATCAGCAAGAGTTAACGT
P-SSM3	HQ337021	CYXG_00186	Orphan_1068	+	156087:156117	AATCAGCAAGAGTTAACGTGAATTTAGTAT
P-SSM3	HQ337021	PRAG_00035	Orphan_1216	+	28815:28845	TTTAAAGGTATTATAATATGGGTATAGACG
P-SSM3	HQ337021	PRAG_00066	PhCOG186	+	49853:49883	TTTCTATAGTTCTTATAATTAGTATTATAG
P-SSM3	HQ337021	CYXG_00010	PhCOG309	+	6976:7006	CCTAAAGAACAATTGATTATATAGTAATTT
P-SSM3	HQ337021	CYXG_00010	PhCOG309	+	6987:7017	ATTGATTATATAGTAATTTCTAAATAATAT
P-SSM3	HQ337021	CYXG_00034	PhCOG71272	+	21481:21511	CTTAATCCACTTTTACTCTTTTTTTACAAA
P-SSM3	HQ337021	CYXG_00075	PhCOG71460	+	82301:82331	CTTGCAATATTGTTAAATACTTTGTATAAT
P-SSM3	HQ337021	CYXG_00075	PhCOG71460	+	82309:82339	ATTGTTAAATACTTTGTATAATGTTGAACG
P-SSM3	HQ337021	PRAG_00179	PhCOG71490	+	124004:124034	TTTTATGCTATAATTAAATCAAGTTAATTT

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-SSM3	HQ337021	PRAG_00179	PhCOG71490	+	124005:124035	TTTATGCTATAATTAAATCAAGTTAATTTT
P-SSM3	HQ337021	CYXG_00141	PhCOG173	+	128649:128679	GTTTAACTGTAGTTAAAGATATAGTTAAGG
P-SSM3	HQ337021	CYXG_00141	PhCOG173	+	128650:128680	TTTAACTGTAGTTAAAGATATAGTTAAGGT
P-SSM3	HQ337021	PRAG_00089	PhCOG173	+	64823:64853	GTTAAATGTAGTTAAAGATACAGTTAAGAT
P-SSM3	HQ337021	PRAG_00089	PhCOG173	+	64834:64864	TTAAAGATACAGTTAAGATGATCTAGATAT
P-SSM3	HQ337021	PRAG_00095	PhCOG71682	+	67395:67425	TTTATGTCACAATTAACACATACTTGACCC
P-SSM3	HQ337021	PRAG_00095	PhCOG71682	+	67406:67436	ATTAACACATACTTGACCCTACCCTAGCAT
P-SSM3	HQ337021	PRAG_00194	PhCOG71906	+	128630:128660	TTTTACTACCATTTCAAATAATATTATACT
P-SSM3	HQ337021	PRAG_00194	PhCOG71906	+	128641:128671	TTTCAAATAATATTATACTATTATAATAGT
P-SSM3	HQ337021	CYXG_00018	PhCOG72201	+	15627:15657	TTTAGTTCTATACTAAAGGGGTCTTAAATT
P-SSM3	HQ337021	CYXG_00087	PhCOG72321	+	92923:92953	ACTAACCACTTATTAAAAAACTTTTAGGAG
P-SSM3	HQ337021	PRAG_00091	PhCOG72385	+	66412:66442	CTCCCTCCTTTCTTAAGATTAAGTTAAGTG
P-SSM3	HQ337021	CYXG_00179	PhCOG72398	+	152420:152450	GTTGACAACCCCTTTATTTTTGCTATATAA
P-SSM3	HQ337021	PRAG_00102	PhCOG72398	+	70157:70187	TATAATATGTACATAACTTTACATTAGTTA
P-SSM3	HQ337021	CYXG_00081	PhCOG72672	+	86072:86102	TTTTATAAATATTTTTAGAAAACTGAAACA
P-SSM4	NC_006884	PSSM4_070.3	Orphan_36	+	51040:51070	TATTATAGCATTTTATTGTACTCTTAGCGT
P-SSM4	NC_006884	PSSM4_152	PhCOG186	+	149743:149773	TTTCAATAGTTCTTATAATTAGTATCGTAA
P-SSM4	NC_006884	PSSM4_049	PhCOG71490	+	43434:43464	TTTTATGCTATAATGATTATAAGTTAAATT
P-SSM4	NC_006884	PSSM4_171	PhCOG71682	+	165985:166015	TTTTTTTATGTCTTTATATACACATAGGTT
P-SSM4	NC_006884	PSSM4_064	PhCOG71906	+	48033:48063	TTTTACTACCATTTCAAATAATATTATACT

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-SSM4	NC_006884	PSSM4_064	PhCOG71906	+	48044:48074	TTTCAAATAATATTATACTATTATAATAGT
P-SSM4	NC_006884	PSSM4_177	PhCOG72398	+	168592:168622	TATAATATGTACATAACTTTACATTAGTTA
P-SSM4	NC_006884	PSSM4_112	PhCOG72724	+	121450:121480	GATAAGAAACTCATAAAAAAATTTCAAGAT
P-SSM4	NC_006884	PSSM4_169	PhCOG73152	+	163785:163815	GTTAAATGTAGTTAAAGATACAGTTAAAAT
P-SSM4	NC_006884	PSSM4_054	PhCOG929	+	45628:45658	CTTCCATTTTAATTATTCATCCATTACAAT
P-SSM5	HQ632825	PRTG_00044	Orphan_1229	-	29081:29111	TTTAAATACTTTCTTACCCAGACATCATAT
P-SSM5	HQ632825	PRTG_00125	Orphan_1230	-	94230:94260	CTTGACTAAATAATTAAAATGTGTTATTAT
P-SSM5	HQ632825	PRTG_00198	PhCOG2077	-	152304:152334	CTTTTCATATTTCTGAGTATTTCTTACTGT
P-SSM5	HQ632825	PRTG_00010	PhCOG611	-	18154:18184	ATTCCATCAAACTTCAAATTTTTTTTCGCT
P-SSM5	HQ632825	PRTG_00012	PhCOG71075	-	18790:18820	CTTAAACGATTATTAAATTGGCACTCACCC
P-SSM5	HQ632825	PRTG_00303	PhCOG71234	-	217296:217326	TTTAAAGATCTATTTCCAATTTCATTAACA
P-SSM5	HQ632825	PRTG_00194	PhCOG173	-	149999:150029	AATGTATAGATATTACAGTTTCTTTAAAAA
P-SSM5	HQ632825	PRTG_00194	PhCOG173	-	150024:150054	TTTACATATCAATTAAATTTGATATAATGT
P-SSM5	HQ632825	PRTG_00194	PhCOG173	-	150029:150059	CTTAGTTTACATATCAATTAAATTTGATAT
P-SSM5	HQ632825	PRTG_00194	PhCOG173	-	150040:150070	CTTAACCTTCTCTTAGTTTACATATCAATT
P-SSM5	HQ632825	PRTG_00194	PhCOG173	-	150051:150081	ACTAAACTTCCCTTAACCTTCTCTTAGTTT
P-SSM5	HQ632825	PRTG_00194	PhCOG173	-	150062:150092	TATAAACGGATACTAAACTTCCCTTAACCT
P-SSM5	HQ632825	PRTG_00158	PhCOG71852	-	126201:126231	GTTATTACTGCATTTCAATTTCGTTTATTT
P-SSM5	HQ632825	PRTG_00130	PhCOG72256	-	96504:96534	TATAATAACTACTAAATAATTTTTTAAAAC
P-SSM5	HQ632825	PRTG_00307	PhCOG72320	-	221029:221059	TTTAACAAATCTATAAATAAGTATAGATTC

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-SSM5	HQ632825	PRTG_00199	PhCOG72352	+	152238:152268	TTTTATTATATATTATAACATACTTTTAGT
P-SSM5	HQ632825	PRTG_00199	PhCOG72352	+	152239:152269	TTTATTATATATTATAACATACTTTTAGTA
P-SSM5	HQ632825	PRTG_00187	PhCOG72394	-	146646:146676	ATAAGAGTATTTTTTATTATTGTTTTATCA
P-SSM5	HQ632825	PRTG_00169	PhCOG72398	-	133766:133796	GTTAAGGTTTACTATATAATTATGTAAAGA
P-SSM5	HQ632825	PRTG_00016	PhCOG72979	-	20336:20366	TTTTTGCTATACTTAATTTGTACACACGAA
P-SSM5	HQ632825	PRTG_00077	PhCOG73040	-	46167:46197	TGTAGGACAGGCTTTATAATTTTTTAAGAA
P-SSM5	HQ632825	PRTG_00093	PhCOG73056	-	69018:69048	AATTAACACCACTTATATTAGCGTTAAGAT
P-SSM5	HQ632825	PRTG_00096	PhCOG73058	-	75208:75238	CATTAACACTAATTAAAGATACATTCGGAA
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98059:98089	CATTATAAATATTTAAAATTATTGTGTGTT
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98070:98100	GTTAATATAATCATTATAAATATTTAAAAT
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98081:98111	TTTCCGTTTCTGTTAATATAATCATTATAA
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98130:98160	TTTAATTTATTATTTAATTCTTGATAATCA
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98133:98163	TTTTTTAATTTATTATTTAATTCTTGATAA
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98140:98170	TATAATCTTTTTTAATTTATTATTTAATTC
P-SSM5	HQ632825	PRTG_00131	PhCOG73090	-	98141:98171	ATATAATCTTTTTTAATTTATTATTTAATT
P-SSM5	HQ632825	PRTG_00139	PhCOG73098	-	104112:104142	TATAAATCACCAATGATATAATATTAAACA
P-SSM5	HQ632825	PRTG_00162	PhCOG73121	-	131307:131337	CTTTCTTATTAAATAATAGTATGTTATTTC
P-SSM5	HQ632825	PRTG_00297	PhCOG73251	-	214170:214200	TATAAACAATTATTAATAAGGATATATTAA
P-SSM7	NC_015290	PSSM7_129	PhCOG129	+	117269:117299	ATTAACACATACTTTTAACTTGTATAAATA
P-SSM7	NC_015290	PSSM7_219	PhCOG217	+	168070:168100	CTTAGGAATAGTTTCAGCAACAGTTGTATA

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
P-SSM7	NC_015290	PSSM7_226	PhCOG224	+	170723:170753	TTTGACACATAAATAAATCAACCGTATAAT
P-SSM7	NC_015290	PSSM7_174	PhCOG173	+	142552:142582	TATAAACAAATACTAAACTTCCCTTAACCT
P-SSM7	NC_015290	PSSM7_174	PhCOG173	+	142563:142593	ACTAAACTTCCCTTAACCTTCTCTTAGTTT
P-SSM7	NC_015290	PSSM7_174	PhCOG173	+	142574:142604	CTTAACCTTCTCTTAGTTTACATATCAATT
P-SSM7	NC_015290	PSSM7_174	PhCOG173	+	142585:142615	CTTAGTTTACATATCAATTAAATTTGATGT
P-SSM7	NC_015290	PSSM7_174	PhCOG173	+	142590:142620	TTTACATATCAATTAAATTTGATGTAATGT
P-SSM7	NC_015290	PSSM7_174	PhCOG173	+	142601:142631	ATTAAATTTGATGTAATGTATACATAATAC
P-SSM7	NC_015290	PSSM7_207	PhCOG71682	+	165038:165068	GTTGTACCCCTCTTTTTTATGTGTTAATAT
P-SSM7	NC_015290	PSSM7_207	PhCOG71682	+	165049:165079	CTTTTTTATGTGTTAATATATACTATAACC
P-SSM7	NC_015290	PSSM7_173	PhCOG71713	+	141972:142002	TTTAAGATTTTCATAGTGTGTGATTAAATA
P-SSM7	NC_015290	PSSM7_236	PhCOG71968	+	178970:179000	ATTATGATCTTTATAAAAAGAACTTTGTAA
P-SSM7	NC_015290	PSSM7_213	PhCOG72398	+	167104:167134	GTTGACAAAACTTTACAATTGCTATATAAT
S-PM2	AJ630128	S-PM2p177	Orphan_44	+	136624:136654	GTTAAGGAACTCTTCGGAGTTCCTTTTTTT
S-PM2	AJ630128	S-PM2p102	Orphan_90	+	72766:72796	AGAAAATAATAAATAATTTTAGATTAAATT
S-PM2	AJ630128	S-PM2p204	PhCOG1328	+	149619:149649	GTTTATACACCCCTGAATATGCCATAAATA
S-PM2	AJ630128	S-PM2p063	PhCOG1370	+	19563:19593	GATTTCTTAAGATTTAAGATTTCTTAAGAT
S-PM2	AJ630128	S-PM2p063	PhCOG1370	+	19564:19594	ATTTCTTAAGATTTAAGATTTCTTAAGATT
S-PM2	AJ630128	S-PM2p063	PhCOG1370	+	19574:19604	ATTTAAGATTTCTTAAGATTTAAGATTTCT
S-PM2	AJ630128	S-PM2p063	PhCOG1370	+	19580:19610	GATTTCTTAAGATTTAAGATTTCTTAAGAT
S-PM2	AJ630128	S-PM2p063	PhCOG1370	+	19581:19611	ATTTCTTAAGATTTAAGATTTCTTAAGATT

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
S-PM2	AJ630128	S-PM2p063	PhCOG1370	+	19591:19621	ATTTAAGATTTCTTAAGATTTAAACTTTAA
S-PM2	AJ630128	S-PM2p064	PhCOG1371	+	19907:19937	ATTAATATTACTATTTGATTATTTTCACCC
S-PM2	AJ630128	S-PM2p208	PhCOG4725	+	154886:154916	ATTGATATTCGATTACCTAAATTTTAAAAA
S-PM2	AJ630128	S-PM2p208	PhCOG4725	+	154896:154926	GATTACCTAAATTTTAAAAAATTTTTCCCG
S-PM2	AJ630128	S-PM2p208	PhCOG4725	+	154897:154927	ATTACCTAAATTTTAAAAAATTTTTCCCGC
S-PM2	AJ630128	S-PM2p179	PhCOG71554	+	137765:137795	ATTGACCTTTATGTTAAGGTATGTTAAAAT
S-PM2	AJ630128	S-PM2p176	PhCOG71555	+	135237:135267	TATAATAAATAGGTAAACAAATGTTAAGGA
S-PM2	AJ630128	S-PM2p251	PhCOG72199	+	193756:193786	CTTAAGATTTGCTTAAGATTAGATTTTTTG
S-PM2	AJ630128	S-PM2p107	PhCOG72320	+	78276:78306	CTAAATTAATAAATAAACTATAGATAAAAT
S-RSM4	NC_013085	SRSM4_170	Orphan_1019	-	139936:139966	TTTAACAAAAACCTCAAAATGTCTAAAAAG
S-RSM4	NC_013085	SRSM4_215	Orphan_1050	-	176037:176067	ATTCTACTACATTTACATCTGTATTAAGGA
S-RSM4	NC_013085	SRSM4_012	Orphan_966	-	4201:4231	GTTATCATTGGCATTAATCTTTTTTAGTCA
S-RSM4	NC_013085	SRSM4_111	Orphan_994	-	75627:75657	AATCTAAATAGATTTAGATAAATTTGATAT
S-RSM4	NC_013085	SRSM4_106	PhCOG72320	-	70458:70488	AATAAACTTAGATTAATTATACGGAAATTA
S-RSM4	NC_013085	SRSM4_028	PhCOG72398	-	10818:10848	CTTGACACATCTTTACAATTGCTATATACT
S-RSM4	NC_013085	SRSM4_104	PhCOG73121	-	67783:67813	TGTAAATAGTATATAATGATTTGTAAAGTT
S-RSM4	NC_013085	SRSM4_234	PhCOG761	+	190956:190986	GATTACTATTATTTAAACGGAAAATATATT
S-RSM4	NC_013085	SRSM4_234	PhCOG761	+	190967:190997	TTTAAACGGAAAATATATTTTTTTTAATGG
S-ShM2	NC_015281	SShM2_117	Orphan_493	+	110243:110273	TTTTATAAATATTTCTAGAAACTGTAAAGA
S-ShM2	NC_015281	SShM2_124	PhCOG72321	+	117109:117139	TGTAATTATTTCGTAAATAACATTTTTTCT

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
S-SM1	NC_015282	SSM1_117	PhCOG71433	+	100379:100409	ATTAAATAATACTTAAGGTAAAAACTTCAT
S-SM1	NC_015282	SSM1_172	PhCOG173	+	135311:135341	TTTATTGATACCTTAAACTGCCCTTACCCA
S-SM1	NC_015282	SSM1_063	PhCOG72528	+	40031:40061	ATACTACAGTAGTTTATCTTTACTTCAACA
S-SM1	NC_015282	SSM1_063	PhCOG72528	+	40042:40072	GTTTATCTTTACTTCAACATCTTTAACAAT
S-SM2	NC_015279	SSM2_091	Orphan_544	+	59197:59227	TTATAATTGATCTTGAAGGTTAATTCATCT
S-SM2	NC_015279	SSM2_218	PhCOG173	+	163970:164000	CTTAACCTTTTCTTGGTGGACTATTTTTTT
S-SM2	NC_015279	SSM2_118	PhCOG72320	+	101677:101707	TTTAATTTATAAATAAATATAGATTTTATA
S-SM2	NC_015279	SSM2_008	PhCOG73063	-	4889:4919	TATAAATAATTTTTAACCTTATATATTATG
S-SSM2	JF974292	CPLG_00179	PhCOG72321	-	128325:128355	TGTAATTATTTCGTAAATAACATTTTTTCT
S-SSM2	JF974292	CPLG_00144	PhCOG72352	+	105160:105190	AATGATAAATACTTAGACATGAATCCATAT
S-SSM2	JF974292	CPLG_00185	PhCOG72672	-	135192:135222	TTTTATAAATATTTCTAGAAACTGTAAAGA
S-SSM5	NC_015289	SSSM5_127	Orphan_751	+	116354:116384	CTTGACACCCTCTTTTTTTATGCTATAATT
S-SSM5	NC_015289	SSSM5_044	PhCOG1016	+	35326:35356	ACTAAAGATCATTTAATTTTGATTTAAAAG
S-SSM5	NC_015289	SSSM5_041	PhCOG73097	-	34237:34267	CTTTCAATGAGACTAACTAGTTCTTCAGGA
S-SSM7	NC_015287	SSSM7_237	Orphan_620	+	193797:193827	AGTTAGAATAGTTTATGTAAGTTTTAATAA
S-SSM7	NC_015287	SSSM7_084	Orphan_674	+	50686:50716	ATTAAGTAACATTTAAAAAAGGTAAAATAA
S-SSM7	NC_015287	SSSM7_084	Orphan_674	+	50726:50756	GTTTATTATCTCTTTAAACATTTTATGATT
S-SSM7	NC_015287	SSSM7_084	Orphan_674	+	50727:50757	TTTATTATCTCTTTAAACATTTTATGATTA
S-SSM7	NC_015287	SSSM7_080	Orphan_677	+	49791:49821	AATTAATCGTCTTTAATTGTTACTTAGTGG
S-SSM7	NC_015287	SSSM7_070	Orphan_684	+	47427:47457	TATAATGAGTACATAACAAACATTTAAGGA

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
S-SSM7	NC_015287	SSSM7_046	Orphan_701	+	40640:40670	GATAATTCAATCTTATAGTATAATCAAACT
S-SSM7	NC_015287	SSSM7_009	Orphan_728	+	5326:5356	TTTTAGTATGTTTTGGTTTAAATTTTCCCT
S-SSM7	NC_015287	SSSM7_135	PhCOG410	+	113817:113847	TTTCTGACTCGTTTATAAAATACTTTTTCA
S-SSM7	NC_015287	SSSM7_056	PhCOG71272	+	43856:43886	TTTCACCATACTATAATAAGTACATAACAA
S-SSM7	NC_015287	SSSM7_147	PhCOG72320	+	126705:126735	ATTATACGATATCTAAACACGTTTTACGAC
S-SSM7	NC_015287	SSSM7_148	PhCOG72321	+	127842:127872	GTTAATTTTTAACTAACAACTCTTAGACCG
S-SSM7	NC_015287	SSSM7_269	PhCOG72398	+	213022:213052	CTTGACAAAATTTTATATTTCCTATATAAT
S-SSM7	NC_015287	SSSM7_003	PhCOG72419	+	644:674	ATTGACTATTACATAATTATCTGGTATAAT
Syn1	NC_015288	Syn1_051	PhCOG1310	+	51958:51988	TTTTAGTGACACTTGATGAAGTGTCTACTA
Syn1	NC_015288	Syn1_076	PhCOG1371	+	59314:59344	ATTGATGTATTTTTGATACTAATTGATTAA
Syn1	NC_015288	Syn1_131	PhCOG73251	+	123003:123033	CTTGACATGTTCTTGATCATCGACTATAAT
Syn10	HQ634191	CPUG_00151	PhCOG71524	-	135732:135762	AATTAAATTCCGTTAATAACATTTTAGGAT
Syn10	HQ634191	CPUG_00151	PhCOG71524	-	135743:135773	TTTAGTTGATCAATTAAATTCCGTTAATAA
Syn10	HQ634191	CPUG_00188	PhCOG72398	-	163468:163498	CTTGACAAAAGATTAAGTTTGCTATATAAT
Syn10	HQ634191	CPUG_00148	PhCOG72551	-	133547:133577	CTTCGCCCATATATAATTCAACCTTTAATT
Syn19	NC_015286	Syn19_155	PhCOG72355	+	137340:137370	CTTTAGACTCAATTCAACCTACTATTATAG
Syn19	NC_015286	Syn19_009	PhCOG72425	-	5140:5170	AATAAATAATGATTACACCTTATTTAGGCA
Syn19	NC_015286	Syn19_009	PhCOG72425	-	5144:5174	CATAAATAAATAATGATTACACCTTATTTA
Syn19	NC_015286	Syn19_010	PhCOG73063	+	5076:5106	CTTCATCATCTCTAAAACCAGCATTAGGGT
Syn2	HQ634190	CPTG_00065	PhCOG72355	+	71451:71481	CTTTAGACTCAATTCAACCTACTATTATAG

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Predicted pho box
Syn2	HQ634190	CPTG_00128	PhCOG72416	+	109020:109050	ATTTACAACTACTTAAGTCCTATATTGAAG
Syn2	HQ634190	CPTG_00196	PhCOG72484	+	159262:159292	GTTTGTCCGCTATTAACAACATTTCAAATA
Syn2	HQ634190	CPTG_00138	PhCOG73063	+	114415:114445	CTTCATCATCTCTAAAACCAGCATTAGGGT
Syn30	HQ634189	CPRG_00093	PhCOG1168	-	66544:66574	TATTAAAGGAAACTCACTAAACCTTAAATA
Syn30	HQ634189	CPRG_00138	PhCOG173	-	97688:97718	TTTATTGATATGTTAAACTGCCCTTACCCT
Syn30	HQ634189	CPRG_00133	PhCOG72099	-	94591:94621	CATAATACATAATTAAAGAGACCTTCCAAA
Syn30	HQ634189	CPRG_00010	PhCOG72501	-	13352:13382	GTTGAGGGGTATATTACTTTTATTTCACAC
Syn30	HQ634189	CPRG_00187	PhCOG72677	-	131817:131847	ATTAGCAGTCATTTACATTATACTAAAACT
Syn33	NC_015285	Syn33_120	Orphan_396	+	108556:108586	ATGGAAGACTATTTTATCATAAAGTAATTG
Syn33	NC_015285	Syn33_014	Orphan_413	+	7089:7119	ATACTGAGAAGATTAAATTATTCTTAGATA
Syn33	NC_015285	Syn33_014	Orphan_413	+	7100:7130	ATTAAATTATTCTTAGATAAATCACTCTTT
Syn33	NC_015285	Syn33_226	PhCOG71968	+	171103:171133	ATTGGGACTTATATAAAAAAGACCTAATTA
Syn33	NC_015285	Syn33_187	PhCOG72003	+	150087:150117	TTTTATAAATAAATATATTGGAATTAATAC
Syn33	NC_015285	Syn33_050	PhCOG72528	+	36202:36232	TGTAATTTATCATTGATCATTTCATTACAA
Syn33	NC_015285	Syn33_050	PhCOG72528	+	36213:36243	ATTGATCATTTCATTACAATGGGTAAAACT
Syn33	NC_015285	Syn33_224	PhCOG72578	+	170197:170227	GGTTCCATATTATTACACAAGACTTAAAAG
Syn33	NC_015285	Syn33_084	PhCOG72899	+	47313:47343	CTTGAGATATTATTCAATAATACTTTAGTG
Syn33	NC_015285	Syn33_084	PhCOG72899	+	47324:47354	ATTCAATAATACTTTAGTGTCAATCTAAAA
Syn9	NC_008296	BSV9_gp209	PhCOG72398	+	163317:163347	CTTGACAAAAGATTAAGTTTGCTATATAAT
Syn9	NC_008296	BSV9_gp20	PhCOG72551	+	16036:16066	CTTCGCCCATATATAATTCAACCTTTAATT

Supplementary Table 3: Predicted pho boxes more than 100bp upstream of genes in the *PhCOG173*/*pstS*/*phoA* region.

Genome	Accession	Gene Id	Cluster Id	Strand	Genomic location	Distance upstream of gene	Predicted pho box
S-SSM7	NC_015287	SSSM7_291	PhCOG173	+	220904:220934	121	GTTGACAAATACATACTTGTCGTTTAACCT
S-SSM5	NC_015289	SSSM5_163	PhCOG71713	+	137087:137117	142	AGAGTTTGAGCTTACAAGATTAACCAAGTG
P-RSM1	HQ634175	CPPG_00148	PhCOG71713	-	97759:97789	135	GGTAAAAACTTGTTCATTTTTTCTCACATG