Supplemental material for:

Mobile elements in a single-filament orange Guaymas Basin *Maribeggiatoa* (*Beggiatoa*) sp. draft genome: Evidence for genetic exchange with cyanobacteria

Running title: Mobile elements in orange Guaymas *Maribeggiatoa*

Barbara J. MacGregor[1]#, Jennifer F. Biddle[2], and Andreas Teske[1]

[1]Department of Marine Sciences, University of North Carolina - Chapel Hill, Chapel Hill NC 27599

[2]College of Earth, Ocean, and the Environment, University of Delaware, Lewes DE 19958

#To whom correspondence should be addressed: bmacgreg@unc.edu

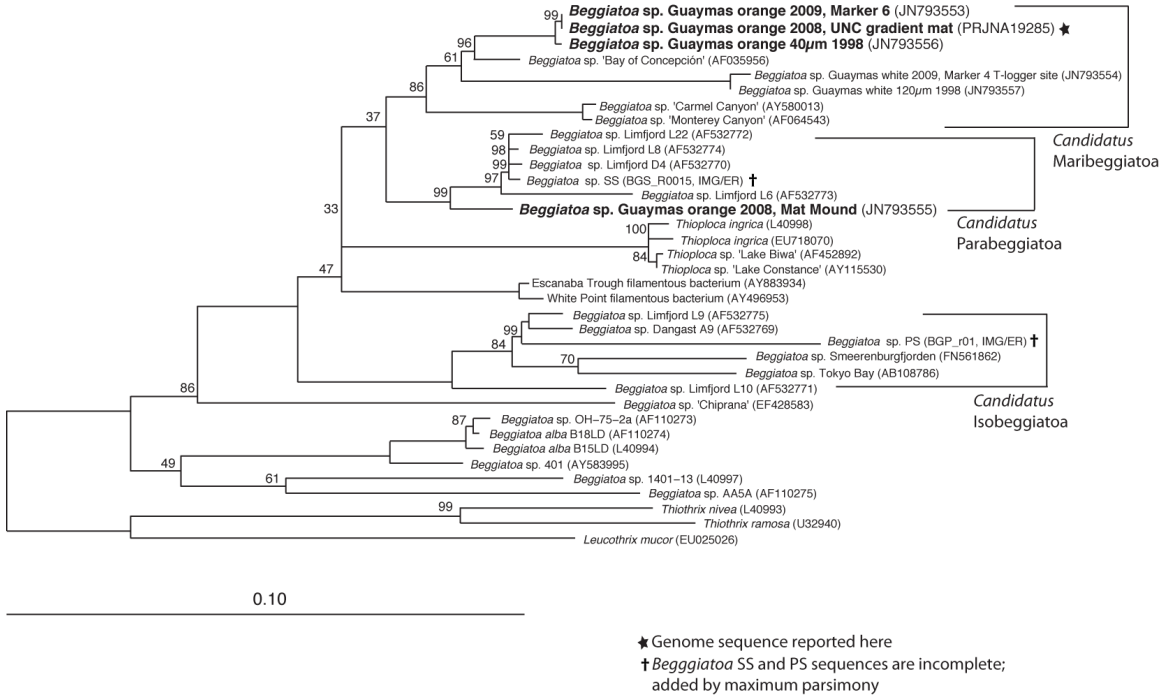

Figure S1. Phylogenetic tree of SSU rRNA genes from *Beggiatoaceae* and related species. Sequences obtained from orange Guaymas Basin filaments are highlighted, and collection sites indicated where known. The candidate species groupings are those proposed by Salman *et al*. (1), and the tree is adapted from McKay *et al*. (2). It was constructed using ARB's (3) neighbor-joining function with a Jukes-Cantor (4) distance correction, with two short sequences ("*Beggiatoa"* PS and SS, referred to in the text as BgP and BgS) added by maximum parsimony. The newly proposed species "Allobeggiatoa" (5), comprising narrow filaments from hypersaline environments, is not included here.

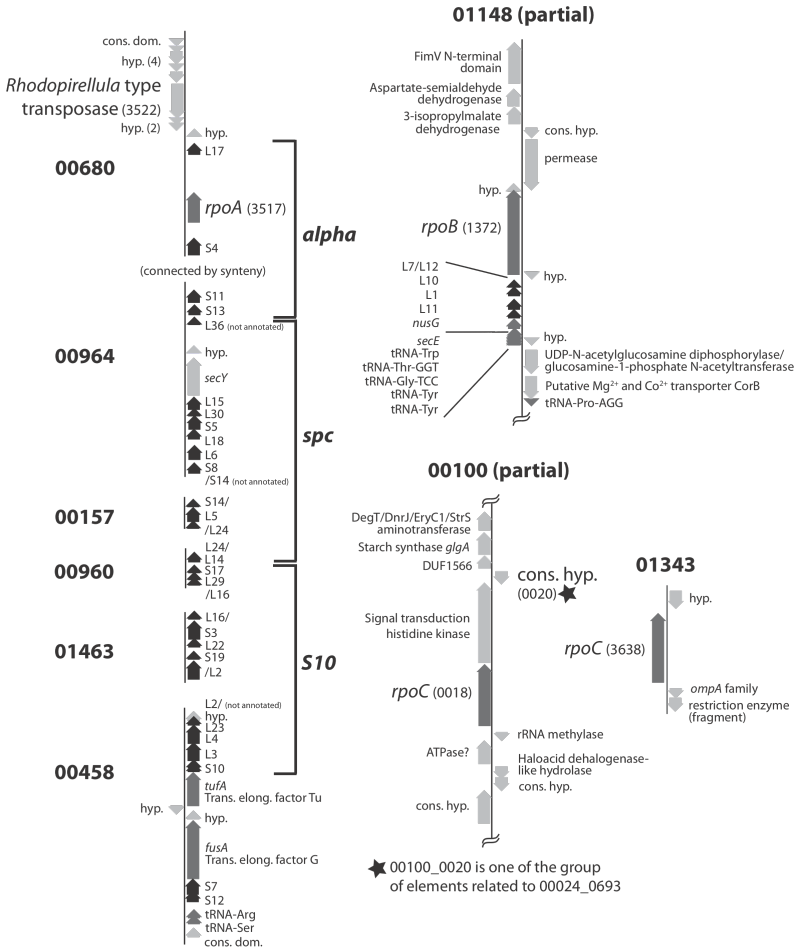

Figure S2. Organization of selected ribosomal protein and DNA-directed RNA polymerase genes. Contig numbers are in large bold type, and ORF numbers (where given) are in parentheses. Putative ribosomal protein genes are in black, RNA polymerase and other genes discussed in the text are in dark gray, and remaining ORFs are in light gray. The contigs that could be assembled are shown in order, with the common bacterial ribosomal protein operons bracketed. “hyp.”, hypothetical protein; “cons. hyp.”, conserved hypothetical protein; “cons. domain”, conserved domain protein.

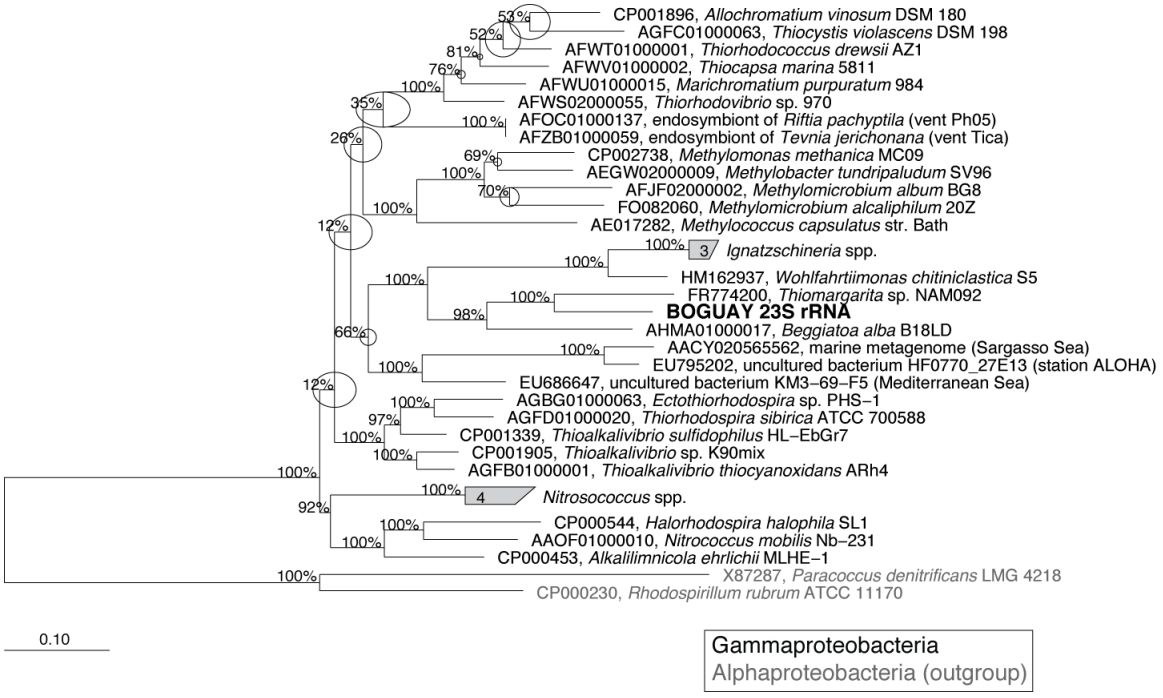

Figure S3. Maximum-likelihood phylogeny of orange Guaymas *Maribeggiatoa* 23S rRNA. Closest relatives of the BOGUAY sequence (with and without inserts) were identified by BLASTN searches of GenBank and IMG/ER, and aligned by the ARB (3) autoaligner with manual refinement. Maximum-likelihood phylogenies were inferred in ARB with RAxML rapid bootstrapping (6), using the GTRMIX nucleotide substitution model and 1000 bootstraps. Numbers represent bootstrap values for each node and circles indicate uncertain branchpoints. Most of the sequences shown do not contain intervening sequences, but filters to exclude these regions had little effect on tree topology (not shown). Two alphaproteobacterial sequences were used to root the tree. The scale bar represents base changes per position.

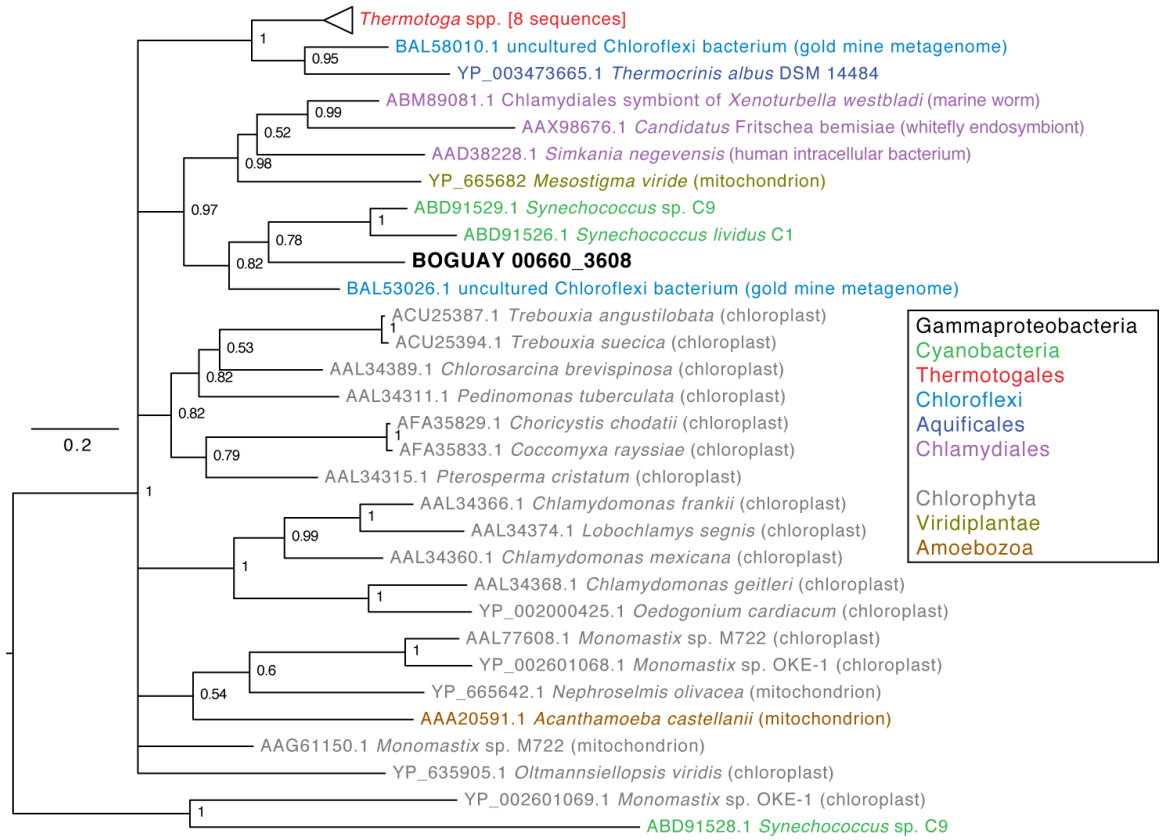

Figure S4. Phylogeny of putative homing endonucleases encoded within 23S rRNA genes by Bayesian inference. Relatives of BOGUAY 0660_3608 were identified by BLASTX searches of the GenBank nr and JCVI IMG/ER databases (August 2012). All sequences in the clades most similar to the BOGUAY sequence were included in the tree shown here. Sequences were aligned in MEGA5 (7) with MUSCLE (8), and phylogenetic analysis was carried out with MrBayes 3.2 (9) (two runs of four chains each, 2.5 million generations) with the gamma rate heterogeneity model (10). The temperature was dropped from 0.1 to 0.05 to increase the chain swap acceptance rate. A mixed prior amino acid substitution model was used and the cpREV model (11) had a posterior probability of 0.998, likely driven by the preponderance of chloroplast sequences. The final statistics suggested that the two runs had converged (not shown). The scale bar represents amino acid changes per position.

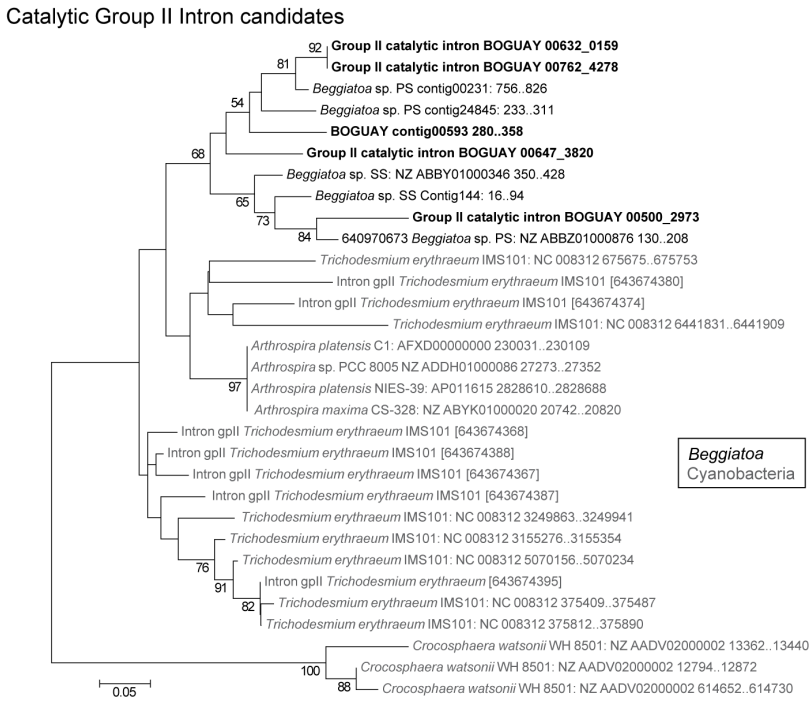

Figure S5. Inferred phylogeny of putative BOGUAY Group II catalytic introns and related sequences by neighbor joining. Four sequences annotated as BOGUAY Group II introns were used for BLASTN searches of the IMG/ER database (cutoff E value 9e-06, cutoff score 58 bits). For sequences not previously annotated (the majority), genome positions are indicated. Three *Crocosphaera watsonii* sequences were used to root the tree. Evolutionary analyses were conducted in MEGA5 (7), with sequences aligned by MUSCLE (8). Evolutionary histories were inferred by neighbor joining (12) with 500 bootstrap replicates (13). Branches corresponding to partitions reproduced in less than 50% of bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches for values ≥50%. Evolutionary distances were computed by the Poisson correction method (14) and are in units of amino acid substitutions per site. All ambiguous positions were removed for each sequence pair. Evolutionary distances are in units of base substitutions per site. The analysis involved 31 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 102 positions in the final dataset.

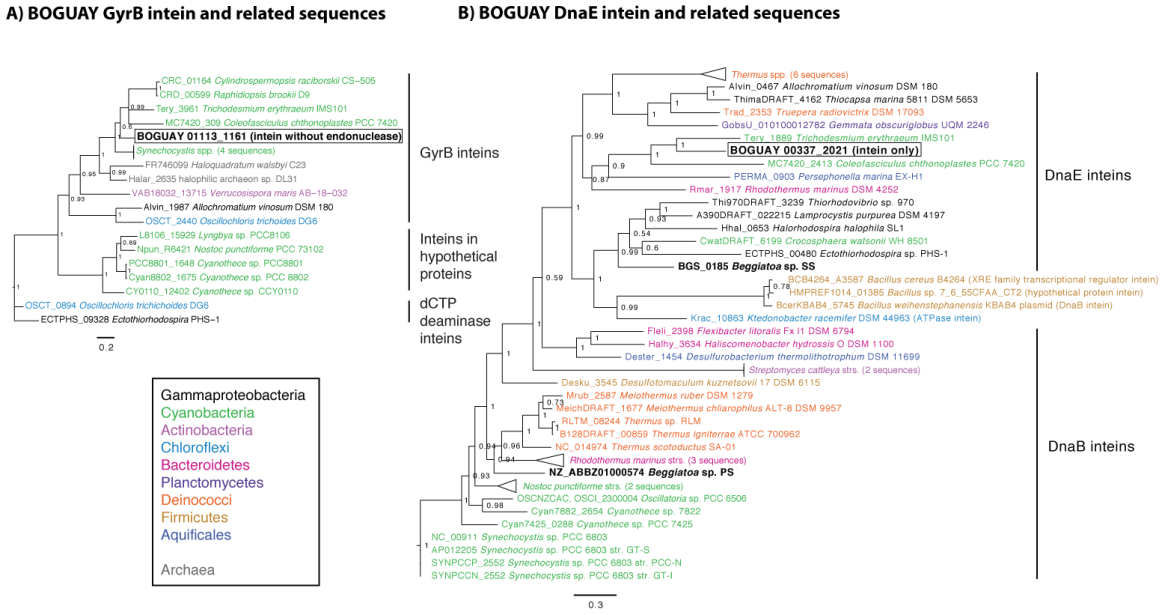

Figure S6. Phylogeny of the putative BOGUAY GyrB and DnaE inteins by Bayesian inference. Relatives of the BOGUAY sequences (highlighted by boxes) were identified by BLASTX searches of the GenBank nr and JCVI IMG/ER databases. Sequences were aligned in MEGA5 (7) using MUSCLE (8). Phylogenetic analysis was carried out with MrBayes 3.2 (9). Two runs of four chains each were carried out for both datasets. The scale bars represent amino acid changes per position. **A)** The analysis was run for 350,000 generations, and final statistics indicated convergence (not shown). The prior amino acid substitution model was mixed, and BLOSUM 62 (15) had a posterior probability of 1.000. **B)** The analysis was run for 500,000 generations, and final statistics indicated convergence (not shown). The prior amino acid substitution model was mixed, and the WAG model (16) had a posterior probability of 1.000.

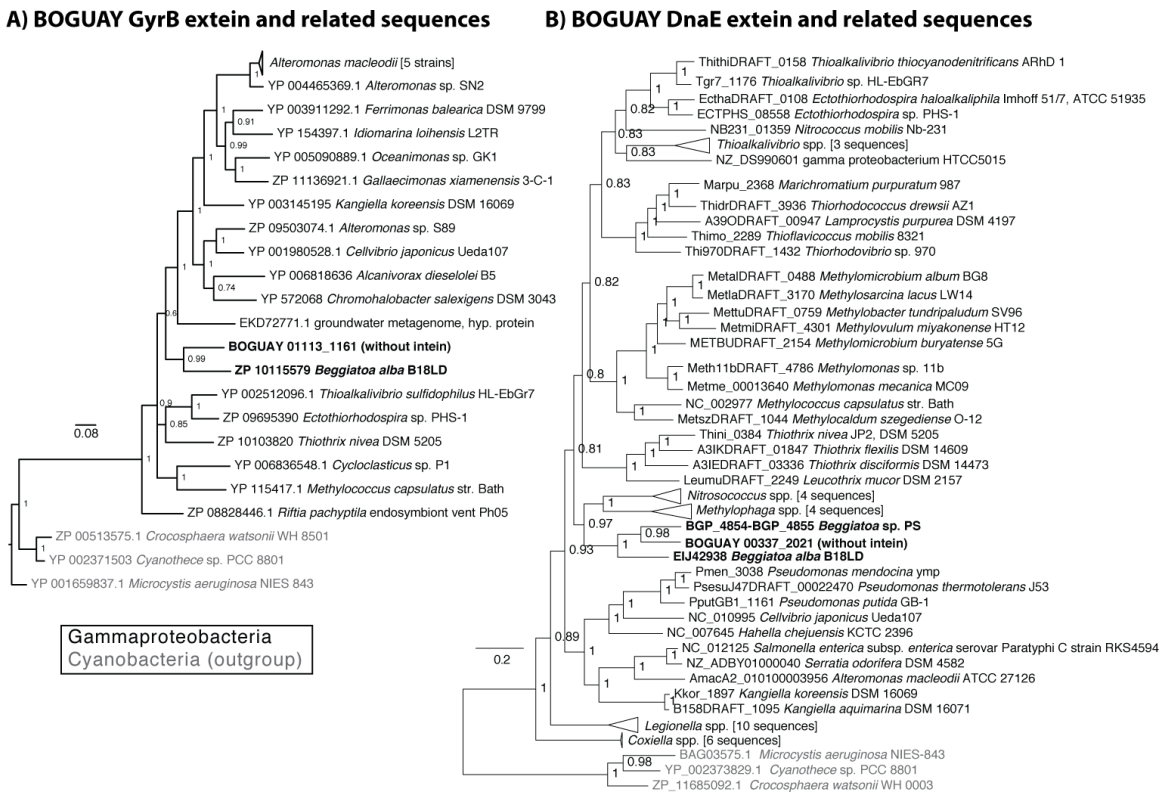

Figure S7. Phylogeny of the putative BOGUAY GyrB and DnaE exteins by Bayesian inference. Relatives of the BOGUAY 01113_1161 and 00337_2021 inferred extein amino acid sequences were identified by BLASTX searches of the GenBank nr and JCVI IMG/ER databases. *Beggiatoaceae* are highlighted in boldface. Sequences are identified by GenBank accession numbers or IMG/ER locus tags and phylogenetic analysis was carried out with MrBayes 3.2 (9). Three cyanobacterial sequences were used to root the trees. For both genes, the prior amino acid model was mixed and the WAG amino acid substitution model (16) had a posterior probability of 1.000. The final statistics indicated convergence was reached between the two runs in both cases (not shown). The scale bars represent amino acid changes per position. **A)** Sequence alignments were done in COBALT (17). The analysis was run for 400,000 generations (2 runs of 4 chains each). **B)** Sequence alignments were done in MEGA5 (7) using MUSCLE (8). MrBayes was run for 4 million generations (2 runs of 4 chains each). The temperature was dropped from 0.1 to 0.03 to increase the chain swap acceptance rate.

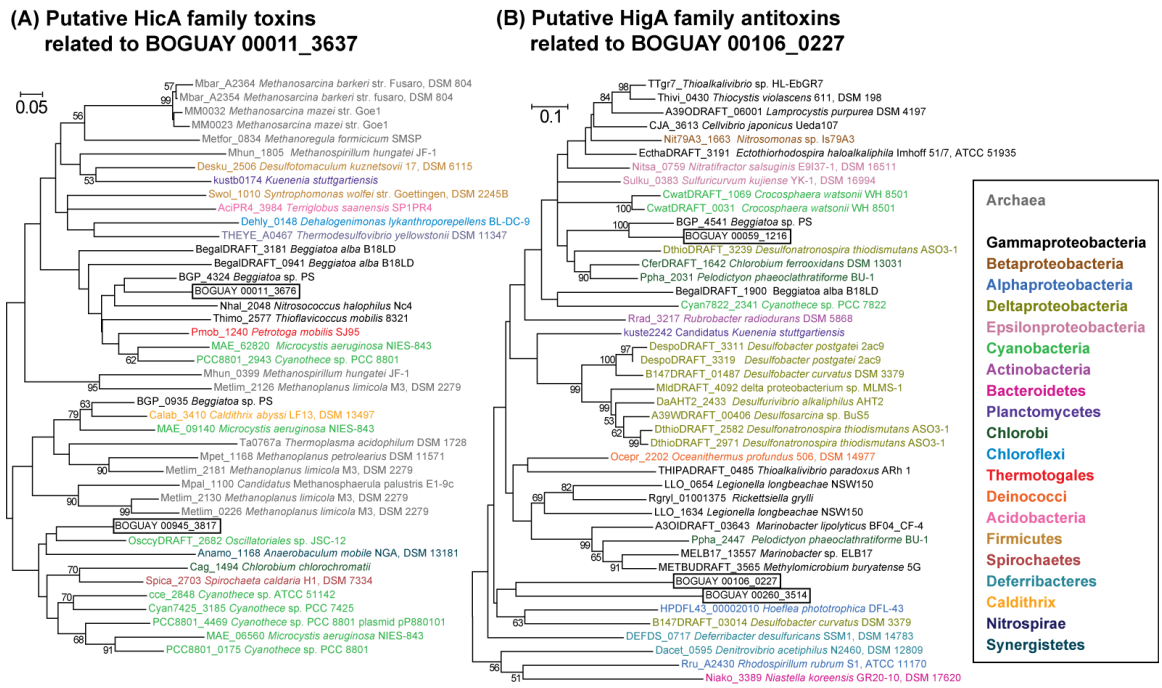

Figure S8. Examples of inferred toxin and antitoxin phylogenies. Evolutionary analyses were conducted in MEGA5 (7), with sequences aligned by MUSCLE (8). Evolutionary histories were inferred by neighbor joining (12) with 500 bootstrap replicates (13). Branches corresponding to partitions reproduced in less than 50% of bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches for values ≥50%. Evolutionary distances were computed by the Poisson correction method (14) and are in units of amino acid substitutions per site. All ambiguous positions were removed for each sequence pair. BOGUAY sequences are highlighted by boxes. **A)** For HicA, the analysis involved 42 amino acid sequences, with 127 positions in the final dataset. **B)** For HigA, the analysis involved 44 amino acid sequences, with 167 positions in the final dataset.

Table S1. Predicted ribosomal protein genes.

Ribosomal protein[a]	Gene	Contig_ORF	*E. coli* operon (18)	Notes
S1	*rpsA*	00726_1454		Separated from other ribosomal protein genes, in conserved neighborhood (includes *ihfB*, *cmk*, *aroA*)
S2	*rpsB*	01035_2226		Separated from other ribosomal protein genes, in conserved neighborhood (includes *tsf*, *pyrH*, *frr*)
S3	*rpsC*	01463_4525	S10	
S4	*rpsD*	00680_3516	alpha	By synteny, may be downstream of S11
S5	*rpsE*	00964_3699	*spc*	
S6	*rpsF*	00936_0911	S6	S6, S18, and L9 together in somewhat conserved cluster with *dnaB*
S7	*rpsG*	00458_3456	*str*	
S8	*rpsH*	00964_3696	*spc*	
S9	*rpsI*	01192_0185		S9 and L13 together, which is seen in other bacteria, but surrounding genes seem unusual; S21 separate, in conserved neighborhood (includes *dnaG*, *rpoD*)
S10	*rpsJ*	00458_3451, 00458_3450	S10	Run of 7 T's in sequence where closest matches have 3 probable sequencing errors, accounting for split gene
S11	*rpsK*	00964_3705	alpha	Probably follows S4, by synteny
S12	*rpsL*	00458_3457	*str*	
S13	*rpsM*	00964_3704	alpha	
S14	*rpsN*	00157_4992 (upstream portion)	*spc*	Downstream portion found upstream of S8 on contig 00964
S15	*rpsO*	00969_3087		Separate from other ribosomal protein genes, in conserved cluster (includes *truB*, *pnp*)
S16	*rpsP*	01318_2101		S16 and L19 together in short conserved cluster (includes *trmD*, *rimM*)
S17	*rpsQ*	00960_4877	S10	
S18	*rpsR*	00936_0910	S6	S6, S18, and L9 together in somewhat conserved cluster with *dnaB*
S19	*rpsS*	01463_4527	S10	
S20	*rpsT*	00665_1281		Separate from other ribosomal proteins, which seems usual; no obvious conserved neighborhood
S21	*rpsU*	01192_0212		S9 and L13 together, which is seen in other bacteria, but surrounding genes seem unusual; S21 separate, in conserved neighborhood (includes *dnaG*, *rpoD*)
L1	*rplA*	01148_1368	L11	
L2	*rplB*	01463_4528 (downstream portion)	S10	Upstream portion found on contig 00458, including 00458_3446 and the DNA from there to the end of the contig
L3	*rplC*	00458_3449	S10	
L4	*rplD*	00458_3448	S10	
L5	*rplE*	00157_4991	*spc*	
L6	*rplF*	00964_3697	*spc*	
L7/L12	*rplL*	01148_1370	L10	
L9	*rplI*	00936_0908	S6	S6, S18, and L9 together in somewhat conserved cluster with *dnaB*
L10	*rplJ*	01148_1369	L10	
L11	*rplK*	01148_1367	L11	
L13	*rplM*	01192_0186		S9 and L13 together, which is seen in other bacteria, but surrounding genes seem unusual; S21 separate, in conserved neighborhood (includes *dnaG*, *rpoD*).
L14	*rplN*	00960_4876	*spc*	

L15	*rplO*	00964_3701	*spc*	
L16	*rplP*	01463_4524 (upstream); 00960_4879 (downstream)	S10	
L17	*rplQ*	00680_3518	alpha	
L18	*rplR*	00964_3698	*spc*	
L19	*rplS*	01318_2098		S16 and L19 together in short conserved cluster (includes *trmD*, *rimM*)
L20	*rplT*	00883_3299		L20 and L35 together in conserved cluster (includes *thrS* and *infC*)
L21	*rplU*	00832_2867		L21 and L27 found with *cgtA*, which seems somewhat conserved
L22	*rplV*	01463_4526	S10	
L23	*rplW*	00458_3447	S10	
L24	*rplX*	00157_4990 (downstream portion)	*spc*	Upstream portion found on contig 00960, downstream of *rplN*
L25 (CTC form)		00369_1647		Where studied, this form of L25 is produced under stress conditions (19). Found in conserved neighborhood (with *ispE*, tRNA-Gln-TTG, *prs*, *pth*).
L27	*rpmA*	00832_2868		L21 and L27 found with *cgtA*, which seems somewhat conserved
L28	*rpmB*	00660_3605		Found immediately downstream of 5S rRNA gene, which does not seem a conserved position. Most orthologs found upstream of L33, but this one is not.
L29	*rpmC*	00960_4878	S10	
L30	*rpmD*	00964_3700	*spc*	
L31	*rpmE*	01153_1097		Neighborhood not obviously conserved; just upstream of heme export protein genes *ccmB* and *ccmC*.
L32	*rpmF*	00396_2184		Found in conserved neighborhood, just upstream of some phospholipid synthesis genes (includes *plsX*, *fabH*, *fabD*)
L33	*rpmG*	00163_0997		Just upstream of *pgk*, position does not seem conserved
L34	*rpmH*	00696_0322	L34	Found in somewhat conserved cluster (includes *rnpA* and *yidC*)
L35	*rpmI*	00883_3300		L20 and L35 together in conserved cluster (includes *thrS* and *infC*)
L36	*rpmJ*	Contig 00964 (positions 5778-5891)	*spc*	

[a] Historically, "L26" was found to be identical to L20 (20), and "L8" was shown to be a complex of the identical proteins L7 and L12 with L10 (21).

Table S2. Predicted tRNA and tRNA synthetase genes. Annotation by JCVI, IMG/ER, and RAST unless otherwise indicated.

tRNA	Contig_ORF	Cognate tRNA synthetase (contig_ORF)	tRNAs spared by bacterial wobble-base rules (22)
tRNA-Ala-GGC tRNA-Ala-TGC tRNA-Ala-CGC	00059_1200 00660_3610 00170_4111	alanyl-tRNA synthetase (00632_0118)	tRNA-Ala-AGC
tRNA-Arg-ACG tRNA-Arg-CCG tRNA-Arg-CCT tRNA-Arg-TCT[4]	00458_3458 00550_5173 00170_4110 00822 (34896-34858...34588-34541)	arginyl-tRNA synthetase (00429_4146)	tRNA-Arg-TCG tRNA-Arg-GCG
tRNA-Asn-GTT	00696_0310	aspartyl/glutamyl-tRNA amidotransferase subunit A (00150_0829, 0828)[3]; subunit B (00338_4326); subunit C (00150_0830)	tRNA-Asn-ATT
tRNA-Asp-GTC	00214_3254	aspartyl-tRNA synthetase (00106_0256)	tRNA-Asp-ATC
tRNA-Cys-GCA	00394_1822	cysteinyl-tRNA synthetase (00362_1727)	tRNA-Cys-ACA
tRNA-Gln-CTG tRNA-Gln-TTG	01192_0184 00369_1645	glutaminyl-tRNA synthetase (00824_3103)	
tRNA-Glu-CTC tRNA-Glu-TTC	00938_0746 00059_1201	glutamyl-tRNA synthetase (00822_0376)	
tRNA-Gly-CCC tRNA-Gly-GCC tRNA-Gly-TCC	00397_1685 00394_1823 01148_1362	glycyl-tRNA synthetase, alpha subunit (01308_0438); beta subunit (01308_0437)	tRNA-Gly-ACC
tRNA-His-GTG	00822_0367	histidyl-tRNA synthetase (00726_1463)	tRNA-His-ATG
tRNA-Ile-GAT tRNA-kIle-TAT	00660_3611 00369_1667[1]	isoleucine-tRNA synthetase (00043_2361) Specificity changed from Met (CAT) to Ile (TAT) by tRNA(Ile)-lysidine synthase (BOGUAY 00794_2073)	tRNA-Ile-AAT
tRNA-Leu-CAA tRNA-Leu-CAG tRNA-Leu-GAG tRNA-Leu-TAG tRNA-Leu-TAA[4]	00153_2340 00059_1208 00149_4873 01068_5251 00214_3263 00394 (7562 - 7524...7223 - 7175)	leucyl-tRNA synthetase (00356_1908)	tRNA-Leu-AAG
tRNA-Lys-CTT tRNA-Lys-TTT	00214_3264 00478_0883	lysyl-tRNA synthetase (01090_3690)	
tRNA-eMet-CAT tRNA-iMet-CAT	01341_2394 01192_0203[2]	methionyl-tRNA synthetase (00356_1919)	

tRNA-Phe-GAA	00241_1022	phenylalanyl-tRNA synthetase, alpha subunit (00726_1461); beta subunit (00794_2075, 2076)3	tRNA-Phe-AAA
tRNA-Pro-CGG tRNA-Pro-GGG tRNA-Pro-TGG	01232_0410 01148_1356 00822_0369	prolyl-tRNA synthetase (00194_2279)	tRNA-Pro-AGG
tRNA-Ser-CGA tRNA-Ser-GCT tRNA-Ser-GGA tRNA-Ser-TGA	00031_3796 00458_3459 00967_1430 00906_2616	seryl-tRNA synthetase (00226_3292)	tRNA-Ser-ACT, tRNA-Ser-AGA
tRNA-Thr-CGT tRNA-Thr-GGT tRNA-Thr-TGT	00901_3209 01148_1363 00513_4070	threonyl-tRNA synthetase (00883_3302)	tRNA-Thr-AGT
tRNA-Trp-CCA	01148_1364	tryptophanyl-tRNA synthetase (00780_1607)	
tRNA-Tyr-GTA	01148_1359 01148_1361	tyrosyl-tRNA synthetase (01113_1172)	tRNA-Tyr-ATA
tRNA-Val-CAC tRNA-Val-GAC tRNA-Val-TAC	00285_1246 00883_3303 00214_3255 00214_3256	valyl-tRNA synthetase (00696_0295)	tRNA-Val-AAC

[1] Predicted by TFAM (23).

[2] Predicted by TFAM (23) and tRNdb (24).

[3] Phenylalanyl-tRNA synthetase and aspartyl/glutamyl-tRNA amidotransferase subunit A each span two predicted ORFs, possibly due to amplification or sequencing errors.

[4] tRNA-Arg-TCT and tRNA-Leu-TAA genes contain intervening sequences, which are possibly group I introns. tRNA domains are annotated below according to the Transfer RNA Database (tRNAdb (24)). tRNA sequences are underlined. Orange, acceptor stem; green, D-stem; blue, anticodon stem; red, anticodon; pink, TΨC-stem. Note that tRNA-Arg-TCT ends in CAA rather than CCA.

tRNA-Arg-TCT (reverse complement of IMG/ER 2502791056 Beggiatoa sp. "Orange Guaymas" : BOGUA_contig00822 34541..34896)

GCGCTCGTAGCTCATCTGGATAGAGCATCGGCCTTCTTTTTTTGTGTTAAGCAAATTGGACCCATAACGGAAAACAGGA
ACTGTTATGGTGGATGGTGCTATATCGGTGAAACCTGTCAAATGGCAATACCGAGGAAACCGAGGGGATGGTGAAAGGG
CAAATCTTTATAAAAATTTTGCAGGTAACCATTCGGTCAACTTAATTAAGTTGCCGGGATCCGTAGAGGCCATACGCAC
CATGCCTGAGATGGCAAAGAGATGGTCCAGACCGCAAACATTAATTTGGTAGCGAAAGCTATAGTGGTATGAAGCCGAG
GGTAGCAGGTTCGAGTCCTGCCGAGCGCAAAATTTTTCAA

tRNA-Leu-TAA (reverse complement of IMG/ER 2502790796 Beggiatoa sp. "Orange Guaymas" : BOGUA_contig00394 7175..7562)

GCCCGGGTGGCGAAATTTGGTAGACGCAAGGGACTTAAACTAATTTGAGCACCTAAGGCGAAATCTTTAGGTGAATGGA
GTCTAATTCGGTGAAACCTTTAAACTGGCAACGTCGAGCTAAGCTTAAGCTTATGTCAAGGATTATGGGCAAAAATTTT
CCCCTTCCCCTTATCTTCTTGGTAGAGCATCTGAAAGTGTAGAGGCCATACGGCTCCTGCCTGACTAAATTCCCCTATC
CCCATTGGGTTAGAGAAAATTTAAAGGCAAAGAAATGGTCCAGACCACAAACATATTTGATTAATTATGAGTATGGCAG
TGAAAACTGTAGTTGGTAAGAAAATCCCTCGGTGGAAACACCGTACCGGTTCGAGTCCGGTCCCGGGTACCA

162 **Table S3. Putative group II catalytic introns and adjacent ORFs.**

Contig	ORF	Assignment	Position on contig (strand)	Closest BLASTX or BLASTN matches	Comments
00500	(not assigned)	Possible transposase	ca. 6098-6340 (+)	First BLASTX/nr hit is putative transposase from gamma proteobacterium IMCC1989 (EGG94688.1)	Identified by BLASTX/nr in GenBank
	2969	KilA-N domain (DNA-binding) protein?	6758-7240 (+)	First BLASTX/nr hit is hypothetical protein ThvES_00001210 [*Thiovulum* sp. ES (EJF07782.1)]	
	2970	Transposase, IS4-like	7525-7941 (+)	First BLASTX/nr hit is BgP; second is transposase IS4 family protein [*Chlorobium limicola* DSM 245 (ACD89665.1)]	
	2971	Retron-type reverse transcriptase	8161-9495 (+)	First BLASTX/nr hit is BgP; second is RNA-directed DNA polymerase [*Moorea producta* 3L (EGJ28487.1)]	
	2972	HNH endonuclease domain protein	9511-9990 (+)	First BLASTX/nr hit is PS; second is hypothetical protein L8106_29020 [*Lyngbya* sp. PCC 8106 (EAW34308.1)]	
	2973	**Group II catalytic intron**	10054-10133 (+)	Blastn/IMG gives one BgP and one BgS match; first BLASTN/nr hit is from *Trichodesmium erythraeum* IMS101, complete genome (CP000393.1)	
	2974	Transposase	10169-10684 (+)	First BLASTX/nr hit is BgS; second is transposase, IS605 OrfB family [*Thiorhodospira sibirica* ATCC 700588 (EGZ44075.1)]	
00647	3820	**Group II catalytic intron**	325-403 (+)	First BLASTN/IMG match is to BgP second is to *Trichodesmium erythraeum* IMS101, complete genome (CP000393.1)	First ORF on contig, coming in from end
	3821	KilA-N domain (DNA-binding)	432-1097 (+)	First BLASTX/nr match is KilA-N domain-containing protein, partial [*Thiovulum* sp. ES (EJF05835.1)]	
	3822	Conserved hypothetical protein	1189-1776 (+)	Several BLASTX/IMG hits in BgP, one in BgS; first non-*Beggiatoaceae* hit by BLASTX/nr is hypothetical protein CRC_01820 [*Cylindrospermopsis raciborskii* CS-505 (EFA69593.1)]	

00632	0159	**Group II catalytic intron**	52147-52225 (-)	Identical to 00762_4278; next two BLASTN/nr hits are BgP and then the other two BOGUAY sequences; first non-*Beggiatoaceae* match is from *Cyanothece* sp. PCC 7822, complete genome (CP002198.1)	Coming in from end of contig
	(not assigned)	Possible fragment of transposase	ca. 51554-51477 (-)	First BLASTN/nr hit is transposase (fragment) [*Microcystis aeruginosa* PCC 9806 (CCI14123.1)]	Identified by BLASTX/nr in GenBank
00762	4278	**Group II catalytic intron**	2845-2923 (+)	Identical to 00362_0159; next two BLASTN/nr hits are BgP and then other two BOGUAY; first non-*Beggiatoaceae* match is from *Cyanothece* sp. PCC 7822, complete genome (CP002198.1)	Second ORF on contig
	4279	Retron-type reverse transcriptase	3026-3544 (+)	First BLASTX/nr match is BgP; second is RNA-directed DNA polymerase [*Moorea producta* 3L (EGJ28487.1)]	First ORF on contig
00593	(not assigned)	**Group II catalytic intron**	280-350 (+)	Identified by BLASTN in IMG/ER with other Group II sequences	On short (1600 bp) contig
	4776	HicA toxin domain protein	861-1013 (+)	First BLASTX hit in IMG/ER is HicA-related protein [BGP_0189]; second is toxin-antitoxin systems (TAS) HicA [*Acetobacter pasteurianus* IFO 3283-32, APA32_08280]	
	4777	HicB family antitoxin	1114-1446 (+)	First BLASTX hit in IMG/ER is HicB-related protein [BGP_0190]; second is HicB-related protein [*Synechococcus* sp. PCC 7002, SYNPCC7002_A1564]	

163

Table S4. Putative XisHI genes. Shorter fragments are indicated by "frag".

Code (Fig. 6)	XisH		XisI		Gene order and comments
	ORF	length (aa)	ORF	length (aa)	
1	00997_3078	139	00997_3079	111	3078 (H) > 3079 (I)
2	00342_3380	138	00342_3379	110	3380 (H) > 3379 (I)
3	00362_1738	138	00362_1737	112	1738 (H) > 1737 (I)
4	00833_4598	138	00833_4597	111	RecB fam endo > 4598 (H) > 4596 (I) > CRISPR endo
5	00946_4151	138	00946_4150	116	4151 (H) > 4150 (I)
6	01026_3196	138	01026_3195	111	3196 (H) > 3195 (I)
7	00948_1636	123	00948_1637	110	1636 (H) > 1637 (I)
8	01341_2380	103	01341_2379	111	2380 (H) > 2379 (I)
9	00553_4207	103	00553_4206	93	4208 (H frag) >4207 (H frag) > 4206 (short I)
---	00553_4208	38			
10	00494_3982	42	00494_3984	111	3982 (H frag) > 3983 (H frag) > 3984 (I)
---	00494_3983	93			
11	00754_1884	45	00754_1883	111	1885 (H frag) > 1884 (H frag) > 1883 (I)
---	00754_1885	53			
12	00155_2449	56	00155_2450	37	2449 (H frag) > 2450 (I frag)
13			01159_4377	117	4377 (At beginning of contig, probably missing upstream XisH)

Table S5. Phylogenetic distribution of XisHI- and "BOGUAY 00024_0693"-like sequences. PCC, Pasteur Culture Collection cyanobacterial database (http://www.pasteur.fr/).

IMG/ER Code	Species	Morphology and physiology	Differentiation	Motility	Habitat or source	Gene copies (-, not found)			References
						XisH	XisI	"0693"	
Gammaproteobacteria									
BOGUAY	Orange Guaymas *Beggiatoa* (*Cand.* "Maribeggiatoa")	Multicellular filaments; sulfur oxidizer		Gliding	Microbial mat near deep-sea hydrothermal vent	11	12	29	This paper
BGP	*Beggiatoa* (*Cand.* "Isobeggiatoa") sp. PS*	Multicellular filaments; sulfur oxidizer		Gliding	Intertidal sediments	4	5	27	(25)
BGS	*Beggiatoa* (*Cand.* "Parabeggiatoa") sp. SS*	Multicellular filaments; sulfur oxidizer		Gliding	Intertidal sediments	-	-	1	(25)
BegalDRAFT	*Beggiatoa alba* B18LD	Multicellular filaments; sulfur oxidizer		Gliding	Rice field ditch sediment	-	1	1	(26)
A3IKDRAFT	*Thiothrix flexilis* DSM 14609	Variable morphology; forms slightly bent filaments, which may include elongated or swollen cells; some strains form rosettes or holdfasts; sulfur oxidizer	Gonidia	Rosette-forming strains produce gliding gonidia	Activated sludge suffering from bulking	2	3	-	(27)
Thini	*Thiothrix nivea* DSM 5205	Single cells, sheathed filaments, or rosettes; sulfur oxidizer		Gliding	Sulfide-containing well water; some *Thiothrix* found in activated sludge	-	2	1	(28, 29)

169

Chloroflexi									
CCHmeta	*Candidatus* "Chlorothrix halophila"	Filamentous; anoxygenic phototroph		Gliding	Hypersaline microbial mat (Guerrero Negro)	3	3	5	(30)
Haur	*Herpetosiphon aurantiacus* ATCC 23779 (DSM 785)	Sheathed filaments; unusual cell envelope, not yet completely defined		Gliding	Freshwater lake; some found in activated sludge	4	5	-	(28, 31, 32)
OSCT	*Oscillochloris trichoides* DG6	Multicellular filaments		Gliding	Sulfide spring	1	-	-	(33)
Rcas	*Roseiflexus* sp. RS-1	Multicellular filaments; thermophilic, anoxygenic phototroph, aerobic chemotroph		Gliding	Found in microbial mats	-	-	1	(34)
RoseRS	*Roseiflexus castenholzii* DSM 13941	Multicellular filaments; thermophilic, anoxygenic phototroph aerobic chemotroph		Gliding	Found in microbial mats	-	-	1	(35)
Bacteroidetes									
Fleli	*Flexibacter litoralis* DSM 6794	Filamentous		Gliding	Seawater aquarium outflow	-	-	1	(36)
Halhy	*Haliscomenobacter hydrossis* DSM 1100 (ATCC 27775)	Sheathed, needle-like filaments; chemoorganotroph		No	Activated sludge; "probably" found in freshwater environments (37)	8	13	-	(28, 37)
Runsl	*Runella slithyformis* DSM 19594 (ATCC 29530)	Curved rods; occasional filaments and coils		No	Freshwater	1	2	-	(38)
Slin	*Spirosoma linguale* DSM 74 (ATCC 33905)	Pleiomorphic: vibroid, horseshoe, ring, and spiral forms		No	Laboratory water bath	2	2	-	(39)
B153DRAFT	*Spirosoma panaciterrae* DSM 21099	Rod-shaped		Gliding	Ginseng field	1	2	-	(40)

170

170

Planctomycetes									
GobsU	*Gemmata obscuriglobus* UQM 2246	Unicellular; budding		Multitrichous flagella during swarmer stage	Freshwater	-	1	-	(41)
Cyanobacteria									
AM1	*Acaryochloris marina* MBIC11017	Unicellular		No	Symbiont of colonial ascidian (non-obligate)	1	1	-	(42)
FJ461733	*Anabaena* sp. 90				Lake Vesijärvi, Finland	-	1	1	(43)
Ava	*Anabaena variabilis* ATCC 29413 (PCC 7937)	Filamentous	Heterocysts, akinetes	(no flagella)	Freshwater (PCC)	4	5	1	(44)
APCC8	*Arthrospira* sp. PCC 8005	Filamentous; non-heterocystous				2	3	-	(45)
AmaxDRAFT	*Arthrospira maxima* CS-328	Filamentous; non-diazotrophic	(no heterocysts)	Gliding	Freshwater (but salt-tolerant)	2	3	-	(46)
SPLC1	*Arthrospira platensis* C1 (PCC 9438)	Spiral-shaped; filamentous		Non-gliding	Freshwater	1	2	-	(47)
NIES39	*Arthrospira platensis* NIES-39	Filamentous; non-diazotrophic	(no heterocysts)	Gliding	Freshwater (but salt-tolerant)	5	5	-	(48)
AplaP	*Arthrospira platensis* str. Paraca	Spiral-shaped			Freshwater	5	5	-	
CwatDRAFT	*Crocosphaera watsonii* WH 8501	Unicellular; aerobic nitrogen fixation			South Atlantic	3	7	-	(49-51)
Cy51472	*Cyanothece* sp. ATCC 51472 (BH63E)	Unicellular; aerobic diazotroph (internal membranes and granules)			Seawater, Port Aransas TX	4	5	-	(52)
cce	*Cyanothece* sp. ATCC 51142 (BH68)	Unicellular; aerobic diazotroph (internal membranes and granules)			Intertidal area	4	5	-	(52)
CY0110	*Cyanothece* sp. CCY0110	Unicellular; aerobic diazotroph		No (IMG/ER)	Coastal marine	2	2	-	

171

171

PCC7424	*Cyanothece* sp. PCC 7424	Unicellular; anaerobic diazotroph		No (PCC)	Rice field soil	2	4	1	(53)
Cyan7425	*Cyanothece* sp. PCC 7425	Unicellular; anaerobic diazotroph		No (PCC)	Rice field soil	2	2	4	(53)
Cyan7822	*Cyanothece* sp. PCC 7822	Unicellular; aerobic diazotroph	(no heterocysts)	No (PCC)	Rice field soil	4	5	2	(54)
PCC8801	*Cyanothece* sp. PCC 8801	Unicellular; aerobic diazotroph (internal granules...)		No (PCC)	Rice field soil	2	4	2	(55)
Cyan8802	*Cyanothece* sp. PCC 8802	Unicellular; aerobic diazotroph		No (PCC)	Rice field soil	3	4	2	(55)
CRC	*Cylindrospermopsis raciborskii* CS-505 (CR1/SDS)	Straight trichomes (when isolated); diazotrophic	Heterocysts, tapered end cells		Fresh water	-	-	2	(56, 57)
L8106	*Lyngbya* sp. PCC 8106	Anaerobic diazotroph (originally reported as aerobic (str. Osc. 23), but PCC says not in pure culture)		Gliding	Intertidal cyanobacterial mat	3	6	1	(58)
MC7420	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420	Non-diazotrophic; straight trichomes of cylindrical cells		Yes	Salt marsh	6	17	12	(53)
---	*Microcystis aeruginosa* NIES-843	Unicellular, colony-forming			Freshwater lake	6	8	-	(59)
N9414	*Nodularia spumigena* CCY9414	Filamentous	Heterocysts, akinetes		Baltic Sea surface waters	3	7	1	(60)
alr	*Nostoc* sp. PCC 7120	Aerobic nitrogen fixation; filamentous	Heterocysts	No		1	1	-	(53, 61)
Aazo	*'Nostoc azollae'* 0708	Filamentous (in one life stage), diazotrophic	Motile hormogonia, akinetes	Yes	Symbiont of water fern *Azolla filiculoides*	1	2	1	(62, 63)
Npun	*Nostoc punctiforme* PCC 73102 (ATCC 29133)	Aerobic nitrogen fixation; filamentous	Heterocysts, akinetes	Transient (PCC)	Root section, *Macrozamia* sp.	5	8	3	(53)
OsccyDRAFT	Oscillatoriales sp. JSC-12				Freshwater	1	1	2	(IMG/ER)

172

OSCI	*Oscillatoria* sp. PCC 6506	Anaerobic diazotroph; filamentous		Yes		4	3	2	(53)
CRD	*Raphidiopsis brookii* D9	Straight filaments ; non-diazotrophic	Akinetes; no heterocysts		Freshwater	-	-	2	(57, 64)
CYB	*Synechococcus* sp. JA-2-3B	Coccoid, unicellular		Yes (IMG)	Octopus Spring microbial mat	-	-	1	(65)
SYNPCC7002	*Synechococcus* sp. PCC 7002	Non-diazotrophic		No (PCC)	Fish pen mud	1	1	-	(53)
Syn60A14	*Synechococcus* sp. PE A1-1 60AY4M2				Mushroom Spring microbial mat	-	-	1	(66)
Syn63A14	*Synechococcus* sp. PE A1-1 63AY4M1				Mushroom Spring microbial mat	-	-	1	(66)
Syn65A14	*Synechococcus* sp. PE A1-1 65AY640				Mushroom Spring microbial mat	-	-	1	(66)
Tery	*Trichodesmium erythraeum* IMS101	Filamentous, diazotrophic	Diazocytes	Yes (IMG)	Coastal marine waters	-	2	7	

* Partial genomes.

Table S6. Summary of BOGUAY restriction enzyme phylogeny. ORFs encoding putative restriction enzymes and associated DNA modification enzymes were classified by comparison with the REBASE database (67); this does not agree in all details with the automated annotations. Closest relatives were then identified by BLASTX searches of the IMG/ER or GenBank nr databases, which are more complete. See Tables S5-S7 for information about individual ORFs. RE, restriction enzyme; MT, DNA methyltransferase.

Classification of closest relative	Type II		Type IIG	Type IIS	Type I		Type III		Type IV
	RE	MT			RE	MT	RE	MT	RE
Gammaproteobacteria	3	4							1
Betaproteobacteria	3	1	1						
Deltaproteobacteria			1	1		1			
Epsilonproteobacteria		1	1						
Cyanobacteria	1	4	4	2	3				
Bacteroidetes	2	3	4						
Chloroflexi			1					1	
OP1	1								
Firmicutes		2	1						
Mycoplasma		1							
Chlorobi			1				1		
Archaea		1	13						

TABLE S7. Candidate Type II restriction/modification genes. Methyltransferases (MTs) and restriction enzymes (REs) are grouped by the putative or experimentally verified recognition site of the closest relative found in REBASE ((67); not shown), with groups separated by heavy lines. The upstream ORF of a pair or group is listed first, where relevant.

			First non-*Beggiatoaceae* BLASTX hit in IMG/ER or GenBank* (September 2012)			
Role	**IMG/ER locus tag**	**Notes**	**IMG/ER or GenBank ID**	**Phylogenetic group**	**Species**	**Expec-tation**
RE	00065_4458 (plus sequence to beginning of contig)	Conserved domain protein; at beginning of contig, probably missing some upstream sequence	Thini_1894	Gamma-proteobacteria	*Thiothrix nivea* JP2, DSM 5205	8e-20
MT	00065_4457	DNA (cytosine-5-)-methyltransferase	Thini_1895	Gamma-proteobacteria	*Thiothrix nivea* JP2, DSM 5205	0e+00
MT	00125_3719	DNA (cytosine-5-)-methyltransferase; ORFs are head-to-head; no other evident RE or MTase nearby	NMBNZ0533_0312	Beta-proteobacteria	*Neisseria meningitidis* NZ-05/33	0e+00
RE	00125_3720	ScaI restriction endonuclease	NMBNZ0533_0311	Beta-proteobacteria	*Neisseria meningitidis* NZ-05/33	7e-48
RE	00161_4393	hindVP restriction endonuclease	NIES39_A03530	Cyanobacteria	*Arthrospira platensis* NIES-39	0e+00
MT	00161_4392	DNA (cytosine-5-)-methyltransferase	cce_4745	Cyanobacteria	*Cyanothece* sp. BH68, ATCC 51142	0e+00
RE?	01171_2312	conserved hypothetical protein	AAR23810.1	Beta-proteobacteria	*Neisseria mucosa* subsp. *heidelbergiensis*	2e-81
MT	01171_2313	DNA (cytosine-5-)-methyltransferase	ZP_03273423.1	Cyanobacteria	*Arthrospira maxima* CS-328	6e-174
MT	00198_4532	Site-specific DNA methylase	ZP_10103377.1	Gamma-proteobacteria	*Thiothrix nivea* DSM 5205	3e-81
RE?	00198_4533	conserved hypothetical protein	ZP_10103376.1	Gamma-proteobacteria	*Thiothrix nivea* DSM 5205	7e-89
RE	00286_0609	ApaLI-like restriction endonuclease	ZP_01688676.1	Bacteroidetes	*Microscilla marina* ATCC 23134	2e-105
MT	00286_0610	DNA (cytosine-5-)-methyltransferase; frame shift and change in quality of match at ORF boundary	YP_004870322.1	Gamma-proteobacteria	*Glaciecola nitratireducens* FR1064	5e-130
	00286_0611	DNA (cytosine-5-)-methyltransferase				
MT	00935_1705	DNA (cytosine-5-)-methyltransferase; No evident RE nearby	ref|ZP_03272602.1	Cyanobacteria	*Arthrospira maxima* CS-328	8e-103
MT	00632_0145	DNA (cytosine-5-)-methyltransferase; no evident RE nearby	CCI06041.1	Cyanobacteria	*Microcystis aeruginosa* PCC 7941	3e-109
MT	00594_4464	DNA methylase	YP_003443578.1	Gamma-proteobacteria	*Allochromatium vinosum* DSM 180	3e-93
RE	00594_4463	Bpu10I restriction endonuclease	YP_003443577.1	Gamma-proteobacteria	*Allochromatium vinosum* DSM 180	3e-122

191

MT	00830_3287	modification methylase BsoBI family protein	P70986.2	Firmicutes	*Geobacillus stearothermophilus*	1e-178
RE	00830_3286	XcyI restriction endonuclease	BAL54860.1	OP1	uncultured candidate division OP1 bacterium	2e-124
MT	00873_2163	DNA adenine methylase	ZP_03391304.1	Bacteroidetes	*Capnocytophaga sputigena* Capno	3e-134
RE	00873_2164	DpmII restriction endonuclease	EJF45781.1	Bacteroidetes	*Capnocytophaga ochracea* str. Holt 25	8e-115
MT	00076_3240	DNA N-6-adenine-methyltransferase (Dam); no evident RE nearby	YP_875554.1	Archaea	*Cenarchaeum symbiosum* A	9e-73
RE	01168_5011	Restriction endonuclease BglII. Only ORF annotated on short contig; seems to be missing a little of beginning of protein. Downstream part of contig gave nothing obvious.	YP_002463562.1	Chloroflexi	*Chloroflexus aggregans* DSM 9485	1e-101
MT	00623_4986	DNA (cytosine-5-)-methyltransferase; only ORF annotated on short contig	AAO48713.1	Firmicutes	*Geobacillus stearothermophilus*	3e-144
RE?	00390_4915	Conserved domain protein; BLASTP with cognate methyltransferase (MSsp6803I, REBASE) found no new methyltransferases in BOGUAY genome. Possibly belongs with part of downstream ORF 00390_4913.	ZP_08466487.1	Beta-proteobacteria	*Kingella kingae* ATCC 23330	1e-30
MT	00632_0150	DNA (cytosine-5-)-methyltransferase. Internal to contig, no REs evident nearby.	YP_002960718.1	Mycoplasmas	*Mycoplasma conjunctivae* HRC/581	2e-137
MT	00260_3510 00260_3511	DNA (cytosine-5-)-methyltransferase; Internal to contig, no REs evident nearby; ORFs appear to be fragments of a single gene; BLASTX result is for two combined.	EIA42719.1	Epsilon-proteobacteria	*Campylobacter coli* 90-3	4e-166
MT	01017_0763	type II DNA modification methyltransferase M.TdeIII; internal to contig, no REs evident nearby	YP_003140803.1	Bacteroidetes	*Capnocytophaga ochracea* DSM 7271	0.0
MT	00696_0279 00696_0278	Internal to contig, no REs evident nearby; ORFs appear to be fragments of a single gene; BLASTX result is for two combined.	ZP_07628911.1	Bacteroidetes	*Prevotella amnii* CRIS 21A-A	1e-56

192 * IMG/ER had occasional capacity issues, so GenBank was used. Where tested, the
193 two databases gave similar results.
194

Table S8. Possible type IIG restriction/methylation enzymes.

		First non-*Beggiatoa* BLASTX hit in IMG/ER or GenBank* (September 2012)			
IMG/ER locus tag	**Notes**	**IMG/ER or GenBank ID**	**Phylogenetic group**	**Species**	**Expec-tation**
01346_4733	Conserved domain protein. At end of small contig; gene probably continues past it. See also 01346_4732 below. 01346_4733, 00815_4735, and 01192_0160 nearly identical.	YP_005921158.1	Archaea	*Methanosaeta harundinacea* 6Ac	2e-14
00815_4735	Hypothetical protein. At end of contig; possibly should be fused upstream with 00815_4734 (below); gene probably continues downstream. 01346_4733, 00815_4735, and 01192_0160 nearly identical.	YP_005921158.1	Archaea	*Methanosaeta harundinacea* 6Ac	2e-18
01192_0160	Conserved domain protein. At end of contig, gene probably continues downstream. 01346_4733, 00815_4735, and 01192_0160 nearly identical.	YP_005921158.1	Archaea	*Methanosaeta harundinacea* 6Ac	2e-15
01171_2301	Type I restriction enzyme R protein N-terminus (HSDR_N). Match covers only a small part of BmuSORF1564P (REBASE).	YP_005920925.1	Archaea	*Methanosaeta harundinacea* 6Ac	1e-141
00513_4066	Conserved hypothetical protein. At beginning of short contig, gene probably continues upstream; should probably be joined with 01097_5208	YP_005921158.1	Archaea	*Methanosaeta harundinacea* 6Ac	2e-94
01097_5208	Conserved domain protein. Alone on contig; should probably be joined with 00513_4066	YP_001047072.1	Archaea	*Methanoculleus marisnigri* JR1	3e-76
00205_2665	Conserved domain protein. At beginning of contig; matches second part of McoGP6ORF996P. Probably not directly linked to 01307_4975 (gap in match).	YP_004383567.1	Archaea	*Methanosaeta concilii* GP6	2e-127
01434_4930	Conserved domain protein. Alone on contig, gene probably continues to either side; similar but not identical to 01293_5122	YP_006545502	Archaea	*Methanoculleus bourgensis* MS2	1e-132

196

01293_5122	Conserved hypothetical protein. Alone on contig, gene probably continues to either side; similar but not identical to 01434_4930.	Aboo_1333	Archaea	*Aciduliprofundum boonei* T469	1e-68
01051_2212	N-6 DNA methylase.	ThePKDRAFT_1685	Archaea	*Thermococcus* sp. PK	0e+00
00462_4797	Eco57I restriction endonuclease. Only ORF annotated on contig; sequence probably continues to either side.	AAU84371.1	Archaea	uncultured archaeon GZfos9D8	1e-86
00686_4838	Conserved domain protein. Alone on contig, gene probably continues upstream	AAU84371.1	Archaea	uncultured archaeon GZfos9D8	1e-68
00796_5264	Type I restriction enzyme R protein N-terminus (HSDR_N). Alone on contig, gene may continue upstream	AAU82382.1	Archaea	Uncultured archaeon GZfos17C7	2e-66
00921_4204	Hypothetical protein. At end of contig, gene probably continues downstream.	CBH38221.1	Archaea	Uncultured archaeon	1e-07
01346_4732	Eco57I restriction endonuclease. At beginning of small contig, gene probably continues upstream. See also 01346_4733 above.	ZP_01628230.1	Cyanobacteria	*Nodularia spumigena* CCY9414	3e-50
00202_4928	Conserved domain protein. Alone on contig, gene probably continues upstream	ZP_01630033.1	Cyanobacteria	*Nodularia spumigena* CCY9414	9e-52
00814_5069	Conserved domain protein. Alone on contig, gene probably continues to either side; match is to more upstream part of target than for 01293_5122 or 01434_4930, with gap between	ZP_01628230.1	Cyanobacteria	*Nodularia spumigena* CCY9414	4e-101
00014_2805	ORFs seem to be fragments of a single gene	CCI20701.1	Cyanobacteria	*Microcystis aeruginosa* PCC 9807	0e+00
00014_2806					
01343_3636	Type I restriction enzyme R protein N-terminus (HSDR_N). At end of contig, gene probably continues downstream.	ZP_10245112.1	Chloroflexi	*Nitrolancetus hollandicus* Lb	4e-87
01307_4975	Restriction enzyme Eco57I domain protein. Alone on contig, gene probably continues in both directions.				7e-157
00163_1009	DNA methylase.	ZP_08504525.1	Beta-proteobacteria	*Methyloversatilis universalis* FAM5	0e+00
00815_4734	Conserved domain protein. One of only two ORFs on contig - see 00815_4735 above. Gene may continue upstream.	ZP_07203318.1	Delta-proteobacteria	Delta proteobacterium NaphS2	9e-100

00414_4874	Eco57I restriction endonuclease. Alone on contig; gene probably continues downstream.	ZP_10404968.1	Epsilon-proteobacteria	*Thiovulum* sp. ES	2e-169
00574_4989	Eco57I restriction endonuclease. Alone on contig.	YP_003996560.1	Bacteroidetes	*Leadbetterella byssophila* DSM 17132	2e-48
01256_5046	Type I restriction enzyme R protein N-terminus (HSDR_N). Alone on contig; gene probably continues on either end.	MED217_09797	Bacteroidetes	*Leeuwenhoekiella blandensis* MED217	0e+00
01133_5261	Hypothetical protein. Alone on contig, gene probably continues on either end; nearly identical to 01355_5253.	MED217_09797	Bacteroidetes	*Leeuwenhoekiella blandensis* MED217	3e-73
01355_5253	Hypothetical protein. Alone on contig, gene probably continues on either end; nearly identical to 01355_5261.	MED217_09797	Bacteroidetes	*Leeuwenhoekiella blandensis* MED217	8e-76
01223_4511	Eco57I restriction endonuclease. ORFs seem to be fragments of same gene	ZP_04060130.1	Firmicutes	*Staphylococcus hominis* SK119	5e-70
01223_4512					
00541_3766	BseRI endonuclease family protein	Ctha_1068	Chlorobi	*Chloroherpeton thalassium* ATCC 35110	2e-30

* IMG/ER had occasional capacity issues, so GenBank was used. Where tested, the two databases gave similar results.

Table S9. Potential Type I, IIS, III, and IV restriction/modification systems. Put., putative.

			First non-*Beggiatoa* BLASTX hit in IMG/ER or GenBank* (September 2012)			
Role	**IMG/ER locus tag**	**Notes**	**IMG/ER or GenBank ID**	**Phylogenetic group**	**Species**	**Expec-tation**
TYPE I						
RE	00287_3058	Type I restriction enzyme R protein N-terminus (HSDR_N)	OSCI_870027	Cyanobacteria	*Oscillatoria* sp. PCC 6506	5e-63
RE	00871_2430	Type I restriction enzyme R protein N-terminus (HSDR_N)	OSCI_870027	Cyanobacteria	*Oscillatoria* sp. PCC 6506	2e-55
RE	01192_0187	Type I restriction enzyme R protein N-terminus (HSDR_N)	CRC_02234	Cyanobacteria	*Cylindrospermopsis raciborskii* CS-505	2e-09
MT	01248_2637	N-6 DNA Methylase	DND132_0555	Delta-proteobacteria	*Desulfovibrio* sp. ND132	6e-86
TYPE IIS						
MT	00948_1621	N-6 DNA Methylase	SYNPCC7002_F0089	Cyanobacteria	*Synechococcus* sp. PCC 7002	0e+00
MT	00948_1622	Type I restriction modification DNA specificity domain protein	Gura_3599	Delta-proteobacteria	*Geobacter uraniireducens* Rf4	2e-75
MT	00375_5197	Conserved domain protein; on small contig with two other short ORFs	OSCI_870027	Cyanobacteria	*Oscillatoria* sp. PCC 6506	1e-10
TYPE III						
RE	00264_4522	hypothetical protein	Cpha266_0934	Chlorobi	*Chlorobium phaeobacteroides* DSM 266	1e-20
MT	00264_4523	DNA methylase	Rcas_4397	Chloroflexi	*Roseiflexus castenholzii* HLO8, DSM 13941	0e+00
TYPE IV						
RE	00282_5161	Conserved domain protein	A3G3DRAFT_00738	Gamma-proteobacteria	*Moraxella boevrei* DSM 14165	1e-30

* IMG/ER had occasional capacity issues, so GenBank was used. Where tested, the two databases gave similar results.

Table S10. Summary of putative BOGUAY toxin and antitoxin genes. See Table S9 for details of location and arrangement.

Toxins		Antitoxins	
Category	**Number**	**Category**	**Number**
HicA	14	HicB	16
MazF	8	MazE	1
RelE	23	RelB/DinJ	2
HigB	1	HigA	7
Txe/YoeB	5	Axe	1
PemK/Doc	1	PHD	11
HipA	2		
PIN domain family	33		
Fic domain family	5		
		AbrB	8
		Xre	14
Unclassified	21	Unclassified	43
TOTAL	**113**	**TOTAL**	**103**

Table S11. Putative toxin-antitoxin system operons. Genes were annotated by JCVI, JGI, and/or RAST unless otherwise indicated.

Toxin and/or antitoxin *(contigs)*	**Occur-rences**	**Notes**
Toxin directly upstream of antitoxin (27)		
HicA>HicB> *(162, 593, 1124)*	3	
HicA>AbrB> *(832)*	1	
HicA>antitoxin> *(472)*	1	
RelE>HigA> *(106, 260)*	2	
RelE(JGI)/HigB(RAST)>HigA> *(1092)*	1	
RelE>Xre> *(253, 804, 1014, 1090, 1124)*	5	
RelE>HicB> *(261)*	1	
HigB>HigA> *(59)*	1	Possible HigB identified by BLASTX upstream of HigA (00059_1216)
toxin>Xre antitoxin> *(64, 763)*	2	One possible antitoxin gene (00763_4162) found by PSI-BLAST
toxin>antitoxin> *(100, 100, 105, 404, 472, 759, 775, 779, 997, 1102)*	10	Two occurrences on contig 0100 are in tandem; one possible antitoxin gene (00100_0028) found by PSI-BLAST
Antitoxin directly upstream of toxin (36)		
HicB>HicA> *(11, 297, 348, 665, 855, 945)*	6	
(RelB)>RelE(JGI)/YafQ(RAST)> *(1185)*	1	Possible RelB gene found by BLASTX upstream of 01185_3401
PHD>RelE> *(1069)*	1	
antitoxin>RelE> *(852, 938, 938, 1124)*	4	Two possible antitoxin genes (00825_3969 and intergenic region downstream of 00938_0720) and one possible RelE gene (00938_0740) identified by BLASTX searches upstream of annotated RelE genes
(MazE)>MazF> *(1160)*	1	01160_3501 identified as possible MazE gene by PSI-BLAST
AbrB>MazF> *(945)*	1	
antitoxin>MazF> *(472)*	1	
PHD(JGI)/YefM(RAST)>Txe/YoeB> *(935)*	1	
Xre>HipA> *(469)*	1	
PHD>PIN family toxin> *(666, 1142, 1223)*	3	
AbrB>Pin family toxin> *(822, 1171, 1279)*	3	
antitoxin>PIN family toxin> *(25, 205, 359, 692, 696, 791, 1113, 1236)*	8	Two possible antitoxin genes (00696_0277 and intergenic region upstream of 00205_2659) and one possible PIN family toxin gene (00692_3338) identified by BLASTX or PSI-BLAST
antitoxin>Fic domain toxin> *(162)*	1	Possible toxin gene (00162_0498) identified by PSI-BLAST
antitoxin>toxin> *(241, 322, 606, 1308)*	4	One possible toxin gene identified by BLASTX downstream of an annotated toxin gene (00606_4541)

216

Tandem repeats of MazF toxin (2)		
AbrB(JGI)/MazE(RAST)>MazF>MazF> *(1220)*	1	At end of very short contig
AbrB>MazF>MazF> *(1124)*	1	
Toxin>antitoxin>toxin (1)		
RelE>Xre>PIN family toxin> *(1302)*	1	No obvious additional antitoxin nearby
Toxin>antitoxin>antitoxin (1)		
RelE>HigA>RelB/DinJ> *(130)*	1	At end of contig
Antitoxin>Toxin>Toxin> (1)		
Axe>Txe/YoeB>RelE> *(1192)*	1	Internal to contig
Toxins with no adjacent antitoxin (41)		
HicA> *(95, 971, 1026)*	3	One is on very short contig (00095_5170); one (01026_3179) has ORF with weak PHD similarity upstream
RelE> *(139, 523, 696, 1142, 1279)*	5	One (00523_4677) is at the end of a contig
MazF> *(871)*	1	At beginning of short contig, no obvious antitoxin nearby
Txe/YoeB> (*92, 362, 1308)*	3	One at beginning of contig, two internal to contigs with no obvious antitoxin genes nearby
HipA> *(614)*	1	Internal to contig, no obvious antitoxin gene nearby
PIN family toxin> *(168, 318, 364, 369, 696, 835, 865, 883, 938, 948, 1026, 1051, 1056, 1155, 1159, 1171, 1171, 1197)*	18	One at beginning of contig, three near ends of contigs, rest internal to contigs
Fic domain toxin> *(239, 692, 804, 1142)*	4	One at beginning of a contig, one at end, two internal
PemK(JGI)/Doc(RAST)> *(822)*	1	Internal to contig, no obvious antitoxin gene nearby
toxin> *(97, 138, 445, 1026, 1142)*	5	One internal to contig, three at end of contigs, one at beginning
Antitoxins with no adjacent toxin (34)		
HicB> *(311, 337, 664, 666, 855, 1051)*	6	All are internal to contigs
HigA> *(244, 1070)*	2	One internal to a contig, one at beginning; no apparent toxin genes near either one
Xre> *(191, 267, 318, 694, 746)*	5	One is at beginning of contig, others are internal
AbrB> *(585)*	1	Internal to contig, no obvious toxin gene nearby
PHD> *(64, 79, 264, 316, 597, 1301)*	6	Two at beginning of contigs, one at end of contig, three internal to contigs
antitoxin> *(39, 91, 121, 138, 285, 523, 835, 848, 851, 948, 991, 1159, 1179, 1232)*	14	Eight internal to contigs, three at beginning, three at end

217

Table S12. Unidentified or partially identified BOGUAY ORFs with high similarity to cyanobacterial sequences. The cyanobacterial match from a BLASTP search of UniProt (68) with the lowest e-value is shown for each ORF. ORFs putatively encoding mobile elements or genes commonly transferred by them (restriction enzymes, transposases, XisH elements, toxin/antitoxin systems, etc.) are not included; some are shown elsewhere. UniProt reports the five best BLASTP matches above a cutoff value (E-value ≤ 1.00E-05); many of these proteins also had high-scoring non-cyanobacterial matches.

IMG/ER identification (number of proteins)	BOGUAY ORF	Length	%id	Score	e-value	UniProt ID of match	Protein name of match	Organism name
conserved domain proteins (16)	00065_4458	327	62.79	55.1	2.00E-06	A0ZGZ3_NODSP	Putative uncharacterized protein	*Nodularia spumigena* CCY9414
	00162_0494	1015	50.86	113	5.00E-24	B4VH37_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00282_5161	462	63.74	112	1.00E-23	P72665_SYNY3	LlaI.2 protein	*Synechocystis* sp. PCC 6803
	00316_2957	204	58.16	242	1.00E-62	B4W0S7_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00390_4915	205	58.33	117	5.00E-25	Q4BYE4_CROWT	Putative uncharacterized protein	*Crocosphaera watsonii* WH 8501
	00392_3127	205	55	70.5	5.00E-11	C7QWM8_CYAP0	Putative uncharacterized protein	*Cyanothece* sp. PCC 8802
	00467_2253	859	64.76	148	2.00E-34	B8HKK8_CYAP4	Metallophosphoesterase	*Cyanothece* sp. PCC 7425
	00472_0541	498	51.95	82.8	1.00E-14	B2IVA1_NOSP7	Putative uncharacterized protein	*Nostoc punctiforme* PCC 73102
	00521_4854	368	54.97	172	1.00E-41	B1WNM1_CYAA5	Probable glycosyl transferase	*Cyanothece* sp. ATCC 51142
	00696_0331	684	51.61	59.3	1.00E-07	C7QYC7_CYAP0	GUN4 domain protein	*Cyanothece* sp. PCC 8802
	00787_5175	603	64.36	275	1.00E-72	Q112E2_TRIEI	Putative uncharacterized protein	*Trichodesmium erythraeum* IMS101
	00970_5280	2361	54.14	185	2.00E-45	Q3M1D8_ANAVT	Serine/threonine protein kinase and signal transduction histidine kinase with GAF and PAS/PAC sensor	*Anabaena variabilis* ATCC 29413

	01049_5275	2361	52.69	184	4.00E-45	Q3M1D8_ANAVT	Serine/threonine protein kinase and signal transduction histidine kinase with GAF and PAS/PAC sensor	*Anabaena variabilis* ATCC 29413
	01113_1154	259	62.34	98.2	2.00E-19	Q3M8J7_ANAVT	Putative uncharacterized protein	*Anabaena variabilis* ATCC 29413
	01183_4449	292	58.93	65.5	2.00E-09	B4VMX7_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	01285_3954	503	52.17	149	1.00E-34	A3IV21_9CHRO	Sensor protein	*Cyanothece* sp. CCY0110
Conserved hypothetical proteins (36)	00085_4727	600	56.1	242	3.00E-62	B7K1N7_CYAP8	Putative uncharacterized protein	*Cyanothece* sp. PCC 8801
	00106_0239	106	56	118	2.00E-25	B4AUQ8_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. PCC 7822
	00107_4619	133	50.78	120	6.00E-26	C7QYD7_CYAP0	Putative uncharacterized protein	*Cyanothece* sp. PCC 8802
	00168_4764	121	52.94	62.4	1.00E-08	Q8YZW8_ANASP	All0337 protein	*Nostoc* sp. PCC 7120
	00168_4765	61	57.38	82	2.00E-14	Q8YZW7_ANASP	Asl0338 protein	*Nostoc* sp. PCC 7120
	00198_4533	98	68.83	117	7.00E-25	B4VT91_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00241_1042	200	52	224	4.00E-57	B4VJL2_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00267_1484	290	54.12	266	1.00E-69	A0ZCM7_NODSP	Putative uncharacterized protein	*Nodularia spumigena* CCY9414
	00318_3317	170	61.4	148	2.00E-34	B4VRV8_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00369_1660	75	57.75	84	5.00E-15	B1XNX7_SYNP2	Putative uncharacterized protein	*Synechococcus* sp. PCC 7002
	00426_4242	287	65.22	379	2.00E-103	B4W504_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00463_3783	206	50.62	93.2	8.00E-18	Q55573_SYNY3	Slr0184 protein	*Synechocystis* sp. PCC 6803
	00463_3787	202	53.03	224	5.00E-57	Q110I8_TRIEI	Putative uncharacterized protein	*Trichodesmium erythraeum* IMS101
	00465_4452	170	60.22	115	2.00E-24	B4VRV8_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00536_3249	135	58.02	167	4.00E-40	B0CB22_ACAM1	Putative uncharacterized protein	*Acaryochloris marina* MBIC11017
	00632_0148	84	60.24	120	4.00E-26	A0YPF0_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
	00664_4932	76	61.84	100	4.00E-20	A8YBV5_MICAE	Genome sequencing data, C278	*Microcystis aeruginosa* PCC 7806
	00696_0320	89	62.86	107	5.00E-22	Y560_SYNJB	UPF0161 protein CYB_0560	*Synechococcus* sp. JA-2-3B'a(2-13)
	00701_1076	115	67.29	174	2.00E-42	Q3M559_ANAVT	Putative uncharacterized protein	*Anabaena variabilis* ATCC 29413
	00883_3306	170	71.86	242	1.00E-62	B4VRV8_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420

	00898_4415	209	55.28	221	4.00E-56	B4VMF9_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00917_2249	167	51.23	162	1.00E-38	B0JQS5_MICAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *aeruginosa* NIES-843
	00919_4129	178	50.86	179	1.00E-43	P74585_SYNY3	Slr0668 protein	*Synechocystis* sp. PCC 6803
	00925_4058	375	54.08	108	1.00E-22	A0ZIG6_NODSP	Putative uncharacterized protein	*Nodularia spumigena* CCY9414
	01051_2217	202	57.29	233	1.00E-59	B2IZ66_NOSP7	Putative uncharacterized protein	*Nostoc punctiforme* PCC 73102
	01054_0106	200	67.51	269	1.00E-70	B4VJL2_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	01092_1328	80	53.12	74.3	4.00E-12	B7K457_CYAP8	Putative uncharacterized protein	*Cyanothece* sp. PCC 8801
	01101_2519	53	77.55	86.7	7.00E-16	B7K0D8_CYAP8	Putative uncharacterized protein	*Cyanothece* sp. PCC 8801
	01101_2520	136	51.47	137	3.00E-31	A0YJ86_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
	01124_0955	110	57.27	135	1.00E-30	A3IXS9_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. CCY0110
	01124_0956	72	73.24	102	2.00E-20	B2IUZ8_NOSP7	Putative uncharacterized protein	*Nostoc punctiforme* PCC 73102
	01155_3625	64	51.11	57	6.00E-07	B1WV08_CYAA5	Putative uncharacterized protein	*Cyanothece* sp. ATCC 51142
	01171_2312	133	77.59	193	2.00E-47	B5W0E8_SPIMA	Putative uncharacterized protein	*Arthrospira maxima* CS-328
	01171_2317	112	64.86	153	6.00E-36	B0C9W1_ACAM1	Putative uncharacterized protein	*Acaryochloris marina* MBIC11017
	01269_2033	460	60.57	557	8.00E-157	B7K3K4_CYAP8	Putative uncharacterized protein	*Cyanothece* sp. PCC 8801
	01318_2108	463	57.69	65.5	2.00E-09	Q8YRN8_ANASP	Alr3406 protein	*Nostoc* sp. PCC 7120
Domain of unknown function (DUF) proteins (15)								
DUF29	00097_4303	147	53.06	160	4.00E-38	B4VZY0_9CYAN	Conserved domain protein, putative	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
DUF29	00128_4517	155	54.68	150	5.00E-35	B4VTS1_9CYAN	Conserved domain protein, putative	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
DUF29	00136_0657	147	55.24	156	6.00E-37	B4VZY0_9CYAN	Conserved domain protein, putative	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
DUF29	00542_5239	140	54.35	147	5.00E-34	B2IU74_NOSP7	Putative uncharacterized protein	*Nostoc punctiforme* PCC 73102
DUF29	00556_2295	150	51.06	145	1.00E-33	B4B870_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. PCC 7822
DUF29	00791_1925	150	51.82	130	3.00E-29	B0JSS6_MICAN	Putative uncharacterized protein	*Microcystis aeruginosa* NIES-843
DUF99	00337_2025	194	62.5	248	2.00E-64	A0YLC1_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
DUF262	00285_1239	372	70.43	534	7.00E-150	B4VSD6_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
DUF488	00072_1969	205	50.53	199	5.00E-49	Q2JPA8_SYNJB	Putative uncharacterized protein	*Synechococcus* sp. JA-2-3B'a(2-13)

DUF488	00072_1970	150	71.79	117	4.00E-25	A0YKB1_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
DUF488	00072_1971	148	60.78	71.2	3.00E-11	B5W5J5_SPIMA	Putative uncharacterized protein	*Arthrospira maxima* CS-328
DUF2283	01103_3563	70	52.83	65.1	2.00E-09	Q7NN51_GLOVI	Gsr0563 protein	*Gloeobacter violaceus*
DUF3696	00285_1238	449	66.02	337	2.00E-90	B4VSD5_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
DUF4065	00358_4902	142	58.42	125	1.00E-27	Q7NEI2_GLOVI	Glr3897 protein	*Gloeobacter violaceus*
DUF4291	00344_4267	218	58.78	184	3.00E-45	B0BYR1_ACAM1	Putative uncharacterized protein	*Acaryochloris marina* MBIC11017
Hypothetical proteins (15)								
	00016_1750	170	78.92	262	1.00E-68	B4VRV8_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00100_0035	95	54.74	123	6.00E-26	Q117F5_TRIEI	Putative uncharacterized protein	*Trichodesmium erythraeum* IMS101
	00209_4817	715	56.1	50.1	7.00E-05	B8HUI7_CYAP4	Oxidoreductase domain protein	*Cyanothece* sp. PCC 7425
	00356_1923	92	63.89	50.8	4.00E-05	B2IW04_NOSP7	Putative uncharacterized protein	*Nostoc punctiforme* PCC 73102
	00516_5007	1728	53.85	51.2	3.00E-05	B4B098_9CHRO	Putative phytochrome sensor protein	*Cyanothece* sp. PCC 7822
	00822_0359	76	70.21	64.3	4.00E-09	B2IW71_NOSP7	Putative uncharacterized protein	*Nostoc punctiforme* PCC 73102
	00830_3281	162	69.77	62.8	1.00E-08	A0YJ12_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
	00830_3283	162	67.44	61.6	3.00E-08	A0YJ12_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
	00871_2432	204	50.36	142	8.00E-33	B4W0S7_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	00961_3361	170	67.86	235	2.00E-60	B4VRV8_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
	01017_0788	103	52.17	56.6	8.00E-07	B4B4G3_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. PCC 7822
	01109_3959	279	61.02	87.4	4.00E-16	B4B2R3_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. PCC 7822
	01226_4080	99	59.46	53.5	7.00E-06	Q116W9_TRIEI	Putative uncharacterized protein	*Trichodesmium erythraeum* IMS101
	01283_5029	719	60.78	65.5	2.00E-09	A8YKD4_MICAE	Genome sequencing data, C323	*Microcystis aeruginosa* PCC 7806
	01308_0455	1878	57.39	134	2.00E-30	Q2JN13_SYNJB	Putative helicase	*Synechococcus* sp. JA-2-3B'a(2-13)
Other proteins (39)								
Excalibur calcium-binding domain.	00322_3106	111	53.62	91.3	3.00E-17	A8YLM3_MICAE	Genome sequencing data, C326	*Microcystis aeruginosa* PCC 7806
FGE-sulfatase	01179_4746	90	53.49	92.8	1.00E-17	B4VQ60_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420

GAF domain protein	00488_5109	1837	53.64	250	1.00E-64	B0C2A3_ACAM1	Sensor protein	*Acaryochloris marina* MBIC11017
GTP-binding protein SEC4, small G protein superfamily, and related Ras family GTP-binding proteins	01232_0381	165	50.91	175	1.00E-42	Q8YSQ4_ANASP	All3030 protein	*Nostoc* sp. PCC 7120
Leucine-rich repeat (LRR) protein	00697_1991	1109	54.29	163	1.00E-38	B2IUT6_NOSP7	Miro domain protein	*Nostoc punctiforme* PCC 73102
Leucine-rich repeat (LRR) protein, contains calponin homology domain	00128_4520	1041	50.45	98.6	2.00E-19	Q10Y31_TRIEI	Small GTP-binding protein	*Trichodesmium erythraeum* IMS101
methyltransferase, FkbM family	01087_4784	175	51.74	194	6.00E-48	A0YVC6_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
PemK-like protein	00822_0360	119	70.79	139	1.00E-31	Q7NI95_GLOVI	Glr2288 protein	*Gloeobacter violaceus*
pentapeptide repeat protein	00162_0528	1033	50.91	132	2.00E-29	Q119S0_TRIEI	Pentapeptide repeat	*Trichodesmium erythraeum* IMS101
pentapeptide repeat protein	00697_1984	506	53.19	85.1	5.00E-15	B4VMN7_9CYAN	Pentapeptide repeat protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
pentapeptide repeat protein	01248_2638	740	56.48	101	2.00E-20	B5W655_SPIMA	Pentapeptide repeat protein	*Arthrospira maxima* CS-328
pentapeptide repeat protein	01248_2639	740	59.74	87.4	4.00E-16	B5W655_SPIMA	Pentapeptide repeat protein	*Arthrospira maxima* CS-328
phage uncharacterized family protein	00050_4778	339	58.93	432	2.00E-119	Q3M5G8_ANAVT	Putative uncharacterized protein	*Anabaena variabilis* ATCC 29413
Predicted ATP-binding protein involved in virulence	00019_4994	333	56.52	83.2	8.00E-15	B4VLH3_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Predicted ATPase	00436_5127	1827	56.25	111	3.00E-23	B2J5V9_NOSP7	Sensor protein	*Nostoc punctiforme* PCC 73102
Predicted ATPase	00436_5128	1942	54.55	78.2	3.00E-13	B9YPA1_ANAAZ	Sensor protein	'*Nostoc azollae*' 0708
Predicted ATPase	00516_5008	2001	53.88	236	2.00E-60	A3IU74_9CHRO	Sensor protein	*Cyanothece* sp. CCY0110
Predicted ATPase	00666_1851	394	54.06	415	5.00E-114	A8YMG4_MICAE	Similar to tr|Q8YZF3|Q8YZF3	*Microcystis aeruginosa* PCC 7806
Predicted ATPase	01333_3876	500	50.5	446	3.00E-123	Q7NEM8_GLOVI	Gll3851 protein	*Gloeobacter violaceus*
Predicted carbamoyl transferase, NodU family	00337_2029	615	67.92	78.2	3.00E-13	Y1178_SYNY3	Uncharacterized protein sll1178	*Synechocystis* sp. PCC 6803
Predicted carbamoyl transferase, NodU family	01101_2521	612	74.18	759	0.00E+00	B4YB60_APHFL	O-carbamoyltransferase	*Aphanizomenon flos-aquae* NH-5

Predicted haloacid-halidohydrolase and related hydrolases	00024_0694	256	64.63	327	7.00E-88	B8HRS0_CYAP4	HAD-superfamily hydrolase, subfamily IA, variant 3	*Cyanothece* sp. PCC 7425
Predicted integral membrane protein, DUF2282	00917_2238	104	56.57	109	8.00E-23	B0CCJ4_ACAM1	Putative uncharacterized protein	*Acaryochloris marina* MBIC11017
Predicted O-methyltransferase	00639_4881	306	54.34	328	3.00E-88	B7JYZ4_CYAP8	Putative uncharacterized protein	*Cyanothece* sp. PCC 8801
putative K+-dependent Na+/Ca+ exchanger	00575_1873	358	56.18	393	1.00E-107	B4VUK0_9CYAN	Putative K+-dependent Na+/Ca+ exchanger	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
putative membrane protein, uncharacterized protein conserved in archaea	01113_1168	813	52.19	643	0.00E+00	B4W3A2_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
regulatory protein, FmdB family	00726_1456	72	57.14	52	2.00E-05	A0YKC6_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
Uncharacterized conserved protein	00394_1826	147	59.59	179	8.00E-44	B4WJA5_9SYNE	Conserved domain protein	*Synechococcus* sp. PCC 7335
Uncharacterized conserved protein	00394_1827	107	73.68	142	9.00E-33	B4WJA6_9SYNE	Conserved domain protein, putative	*Synechococcus* sp. PCC 7335
Uncharacterized conserved protein	01408_4978	610	53.31	333	1.00E-89	Q8YS65_ANASP	All3226 protein	*Nostoc* sp. PCC 7120
Uncharacterized protein conserved in cyanobacteria	00203_2410	212	58.49	266	1.00E-69	A0YSQ3_9CYAN	Putative uncharacterized protein	*Lyngbya* sp. PCC 8106
Uncharacterized protein conserved in cyanobacteria	00665_1267	260	56	247	8.00E-64	B7KE70_CYAP7	Putative uncharacterized protein	*Cyanothece* sp. PCC 7424
Uncharacterized protein conserved in cyanobacteria	01113_1176	295	60	283	2.00E-74	B7K3N9_CYAP8	Putative uncharacterized protein	*Cyanothece* sp. PCC 8801
Uncharacterized protein conserved in cyanobacteria	01204_2884	162	60.51	192	9.00E-48	B4AVB9_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. PCC 7822
Uncharacterized protein conserved in cyanobacteria	01308_0482	247	56.56	283	1.00E-74	B4VR64_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
WD domain, G-beta repeat protein	00155_2444	1270	53.58	312	3.00E-83	B4AXY5_9CHRO	WD-40 repeat protein	*Cyanothece* sp. PCC 7822

	WD domain, G-beta repeat protein	00194_2270	1526	52.13	300	2.00E-79	YY46_ANASP	Uncharacterized WD repeat-containing protein alr3466	*Nostoc* sp. PCC 7120
	WD domain, G-beta repeat protein	00394_1818	706	51.41	320	1.00E-85	B4W111_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420

223

224

Table S13. Identified BOGUAY ORFs with high similarity to cyanobacterial sequences. The cyanobacterial match from UniProt (68) with the lowest e-value is shown for each ORF. ORFs putatively encoding mobile elements or genes commonly transferred by them (restriction enzymes, transposases, XisH elements, toxin/antitoxin systems, etc.) are not included; some are shown elsewhere. UniProt reports the five best BLASTP matches above a cutoff value (E-value ≤ 1.00E-05); many of these proteins also had high-scoring non-cyanobacterial matches.

IMG/ER identification	BOGUAY ORF	Length	%id	Score	e-value	UniProt ID of match	Protein Name of match	Organism name
Sensory and signal transduction proteins								
adenylate and Guanylate cyclase catalytic domain protein	00024_0700	586	62.61	282	2.00E-74	B4VJK7_9CYAN	Adenylate and Guanylate cyclase catalytic domain protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
adenylate and Guanylate cyclase catalytic domain protein	00391_1540	1805	50.22	223	3.00E-56	Q110P3_TRIEI	Adenylate/guanylate cyclase	*Trichodesmium erythraeum* IMS101
adenylate and Guanylate cyclase catalytic domain protein	00962_4708	1805	55.06	341	5.00E-92	Q110P3_TRIEI	Adenylate/guanylate cyclase	*Trichodesmium erythraeum* IMS101
adenylate and Guanylate cyclase catalytic domain protein	01034_3367	352	63.08	429	6.00E-118	B5WA99_SPIMA	Adenylate/guanylate cyclase	*Arthrospira maxima* CS-328
Adenylate cyclase, family 3 (some proteins contain HAMP domain)	00231_3849	1119	55.17	85.1	2.00E-15	Q111B1_TRIEI	Adenylate/guanylate cyclase	*Trichodesmium erythraeum* IMS101
Adenylate/guanylate kinase	01074_4351	744	53.12	224	1.00E-56	Q10W13_TRIEI	Adenylate/guanylate cyclase	*Trichodesmium erythraeum* IMS101
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	00948_1635	1875	56.11	386	4.00E-105	B4VQ57_9CYAN	Sensor protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	00961_3363	424	57.04	306	2.00E-81	B0C2A5_ACAM1	Sensor protein	*Acaryochloris marina* MBIC11017
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	01017_0784	1660	51.02	457	1.00E-126	B0C8M6_ACAM1	Sensor protein	*Acaryochloris marina* MBIC11017

ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	01054_0100	974	57.26	666	0.00E+00	B4W013_9CYAN	PAS fold family	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	01074_4353	280	56.36	307	8.00E-82	B5W6J4_SPIMA	Sensor protein	*Arthrospira maxima* CS-328
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	01245_5106	421	55.34	272	3.00E-71	B0C8M5_ACAM1	Sensor protein	*Acaryochloris marina* MBIC11017
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	01285_3956	421	53.94	276	1.00E-72	B0C8M5_ACAM1	Sensor protein	*Acaryochloris marina* MBIC11017
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	01308_0467	1119	51.62	1077	0.00E+00	Q111B1_TRIEI	Adenylate/guanylate cyclase	*Trichodesmium erythraeum* IMS101
histidine kinase A domain protein	00301_1142	750	50.39	120	4.00E-26	Q10X63_TRIEI	Sensor protein	*Trichodesmium erythraeum* IMS101
PAS domain S-box	00470_1304	878	50.43	331	2.00E-88	B4W0L0_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Response regulator containing a CheY-like receiver domain and a GGDEF domain	00697_1997	437	71.01	108	2.00E-22	B9YKK5_ANAAZ	Sensor protein	'*Nostoc azollae*' 0708
response regulator receiver domain protein	00024_0680	145	61.72	168	1.00E-40	A0YM57_9CYAN	Two-component response regulator	*Lyngbya* sp. PCC 8106
response regulator receiver domain protein	00024_0682	888	57.93	184	3.00E-45	B0C2H4_ACAM1	Sensor protein	*Acaryochloris marina* MBIC11017
response regulator receiver domain protein	00478_0855	129	57.36	171	2.00E-41	A3IHA5_9CHRO	Two-component response regulator	*Cyanothece* sp. CCY0110
response regulator receiver domain protein	00948_1638	131	63.41	173	4.00E-42	B4VQ56_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
response regulator receiver domain protein	01034_3366	125	70.25	179	8.00E-44	A0YPA4_9CYAN	Two-component response regulator	*Lyngbya* sp. PCC 8106
Sensory transduction histidine kinase	00470_1295	459	55.47	448	1.00E-123	B4W249_9CYAN	PAS fold family	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420

Serine/threonine protein kinase	00291_5252	1830	63.78	239	7.00E-62	B4VIF3_9CYAN	Adenylate and Guanylate cyclase catalytic domain protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Serine/threonine protein kinase	00676_3755	2361	52.48	989	0.00E+00	Q3M1D8_ANAVT	Serine/threonine protein kinase and signal transduction histidine kinase with GAF and PAS/PAC sensor	*Anabaena variabilis* ATCC 29413
Serine/threonine protein kinase	00962_4709	1830	64.13	285	3.00E-75	B4VIF3_9CYAN	Adenylate and Guanylate cyclase catalytic domain protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Serine/threonine protein kinase	01127_5053	1830	63.04	318	4.00E-85	B4VIF3_9CYAN	Adenylate and Guanylate cyclase catalytic domain protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Serine/threonine specific protein phosphatase involved in glycogen accumulation, PP2A-related	00406_3932	872	55.72	363	2.00E-98	B4WKD5_9SYNE	Ser/Thr protein phosphatase family protein	*Synechococcus* sp. PCC 7335
Signal transduction histidine kinase	00024_0678	1875	58.77	437	3.00E-120	B4VQ57_9CYAN	Sensor protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Signal transduction histidine kinase	00024_0681	1875	59.59	383	3.00E-104	B4VQ57_9CYAN	Sensor protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Signal transduction histidine kinase	00093_4605	857	51.93	432	7.00E-119	B4AZE6_9CHRO	Sensor protein	*Cyanothece* sp. PCC 7822
Signal transduction histidine kinase	00697_1995	2724	66.04	142	3.00E-32	B4VXY3_9CYAN	Sensor protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Predicted ATPase	00837_5334	1728	59.39	203	5.00E-51	B4B098_9CHRO	Putative phytochrome sensor protein (ser-thr kinase by PSI-BLAST - BM)	*Cyanothece* sp. PCC 7822
Cell wall biogenesis, membrane production, and possible chromosome partitioning proteins								
Glycosyltransferases involved in cell wall biogenesis	00034_4585	325	63.38	445	4.00E-123	A8YNB5_MICAE	Genome sequencing data, C328	*Microcystis aeruginosa* PCC 7806
Glycosyltransferases involved in cell wall biogenesis	00359_2116	385	53.64	122	9.00E-27	B8HUG1_CYAP4	Glycosyl transferase family 2	*Cyanothece* sp. PCC 7425
Glycosyltransferases involved in cell wall biogenesis	00388_3714	160	56.67	155	1.00E-36	B1WVH2_CYAA5	Glycosyl transferase, group 2	*Cyanothece* sp. ATCC 51142
Predicted polypeptide N-acetylgalactosaminyltransferase	00138_2549	703	53.78	286	5.00E-75	B7K809_CYAP7	Glycosyl transferase family 2	*Cyanothece* sp. PCC 7424

Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	00267_1479	383	68.94	172	8.00E-42	B2IYK0_NOSP7	DegT/DnrJ/EryC1/StrS aminotransferase	*Nostoc punctiforme* PCC 73102
Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	00267_1480	383	56.91	282	2.00E-74	B2IYK0_NOSP7	DegT/DnrJ/EryC1/StrS aminotransferase	*Nostoc punctiforme* PCC 73102
peptidase, M23 family	00938_0730	338	54.17	131	2.00E-28	Q05SU0_9SYNE	Peptidoglycan-binding LysM	*Synechococcus* sp. RS9916
glycosyltransferase, group 1 family protein, phosphatidylinositol biosynthesis/Sulfolipid synthase	00904_5273	362	51.83	196	1.00E-48	Q6ZET2_SYNY3	Sll5048 protein	*Synechocystis* sp. PCC 6803
N-acetylglucosaminyltransferase complex, subunit PIG-A/SPT14, required for phosphatidylinositol biosynthesis/Sulfolipid synthase	00905_5269	362	50.51	104	3.00E-21	Q6ZET2_SYNY3	Sll5048 protein	*Synechocystis* sp. PCC 6803
oxidoreductase, 2-nitropropane dioxygenase family protein	00286_0602	554	68.92	729	0.00E+00	A0ZCD7_NODSP	2-nitropropane dioxygenase, NPD (according to pfam, now called nitronate monooxygenase; in SEED however it is Enoyl-[acyl-carrier-protein] reductase [FMN] (EC 1.3.1.9), inferred for PFA pathway	*Nodularia spumigena* CCY9414
phosphopantetheine attachment domain protein	00430_5228	3242	59.09	114	2.00E-24	B2IXJ9_NOSP7	Amino acid adenylation domain protein	*Nostoc punctiforme* PCC 73102
Myo-inositol-1-phosphate synthase	01069_2501	367	75.55	585	2.00E-165	B0C6C6_ACAM1	Myo-inositol-1-phosphate synthase	*Acaryochloris marina* MBIC11017
CobQ/CobB/MinD/ParA nucleotide binding domain protein	01124_0945	449	52.24	448	5.00E-124	B1XNZ9_SYNP2	Chromosome partitioning protein, ParA ATPase family	*Synechococcus* sp. PCC 7002
CobQ/CobB/MinD/ParA nucleotide binding domain protein (ATPases involved in chromosome partitioning)	00037_3952	354	50.71	349	3.00E-94	B5W8M8_SPIMA	Cobyrinic acid ac-diamide synthase	*Arthrospira maxima* CS-328
Transport proteins								
ABC-type Mn/Zn transport systems, ATPase component	00016_1762	289	54.47	279	2.00E-73	P73086_SYNY3	ABC transporter	Synechocystis sp. PCC 6803
ABC-type Mn2+/Zn2+ transport systems, permease components	00016_1763	276	63.1	314	7.00E-84	B0JY65_MICAN	ABC-3 transport family protein	*Microcystis aeruginosa* NIES-843

putative ferrous iron transport protein A	01185_3408	211	60.81	99.4	1.00E-19	B0JLG8_MICAN	Ferrous iron transport protein A	*Microcystis aeruginosa* NIES-843
ferrous iron transport protein B	01185_3407	774	59.06	909	0.00E+00	B0JLG7_MICAN	Ferrous iron transport protein B	*Microcystis aeruginosa* NIES-843
putative K+-dependent Na+/Ca+ exchanger	00575_1873	367	54.27	388	5.00E-106	B7KDB9_CYAP7	Na+/Ca+ antiporter, CaCA family	*Cyanothece* sp. PCC 7424
putative K+-dependent Na+/Ca+ exchanger	00632_0121	374	50.96	105	1.00E-21	C7QRB7_CYAP0	Sensor protein	*Cyanothece* sp. PCC 8802
transporter, divalent anion:Na+ symporter (DASS) family protein	00344_4265	461	50.42	470	2.00E-130	B4W1Y8_9CYAN	Transporter, DASS family	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Predicted membrane protein involved in D-alanine export; Acyltransferase required for palmitoylation of Hedgehog (Hh) family of secreted signaling proteins	00300_4486	495	51.44	382	4.00E-104	Q7NM22_GLOVI	Alginate O-acetylation protein	*Gloeobacter violaceus*
multidrug resistance protein, SMR family	00103_3279	106	54.9	122	2.00E-26	Q8Z018_ANASP	Multidrug exporter	*Nostoc* sp. PCC 7120
Possible secondary metabolite production proteins								
AMP-binding enzyme; Non-ribosomal peptide synthetase modules and related proteins; Acyl-CoA synthetase	00450_4350	1098	60.71	564	2.00E-158	B0JJX1_MICAN	McnA protein	*Microcystis aeruginosa* NIES-843
beta-ketoacyl synthase, N-terminal domain protein	01230_4663	2206	57.02	409	2.00E-112	D2JNV9_9NOSO	Cis-AT polyketide synthase	*Nostoc* sp. 'Peltigera membranacea cyanobiont'
glycosyl hydrolase family 3 N-terminal domain protein	00984_4621	140	69.66	141	2.00E-32	B1X1W0_CYAA5	Putative uncharacterized protein	*Cyanothece* sp. ATCC 51142
hydantoinase B/oxoprolinase	00342_3374	1215	56.89	528	5.00E-148	B4WH03_9SYNE	Hydantoinase/oxoprolinase domain family protein	*Synechococcus* sp. PCC 7335
N-methylhydantoinase B/acetone carboxylase, alpha subunit	00342_3373	519	64.71	75.5	2.00E-12	B4B8A0_9CHRO	5-oxoprolinase (ATP-hydrolyzing)	*Cyanothece* sp. PCC 7822
PfaB family protein	00286_0603	1556	53.52	1491	0.00E+00	A0ZFX9_NODSP	Beta-ketoacyl synthase	*Nodularia spumigena* CCY9414
Polyketide synthase modules and related proteins	00241_1056	2206	54.33	632	6.00E-179	D2JNV9_9NOSO	Cis-AT polyketide synthase	*Nostoc* sp. 'Peltigera membranacea cyanobiont'
Polyketide synthase modules and related proteins	00971_2005	575	52.93	600	3.00E-169	B2J2F2_NOSP7	KR	*Nostoc punctiforme* PCC 73102

polyketide-type polyunsaturated fatty acid synthase PfaA	00971_2006	1735	54.61	1374	0.00E+00	A0ZFX6_NODSP	Heterocyst glycolipid synthase	*Nodularia spumigena* CCY9414
putative linear gramicidin synthetase LgrC	00172_4829	4458	51.1	265	4.00E-69	B2IXJ7_NOSP7	Amino acid adenylation domain protein	*Nostoc punctiforme* PCC 73102
putative tyrocidine synthetase TycC	00780_1591	1298	50.93	1150	0.00E+00	B5STA6_OSCAG	OciC (Fragment)	Planktothrix agardhii NIES-205
ornithine cyclodeaminase/mu-crystallin family protein	00100_0036	350	50.74	358	6.00E-97	B4VZM8_9CYAN	Ornithine cyclodeaminase/mu-crystallin family	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420

229

Other proteins

4-hydroxybenzoate synthetase (chorismate lyase)	01073_4016	110	62.62	143	6.00E-33	B4B8N0_9CHRO	Putative uncharacterized protein	*Cyanothece* sp. PCC 7822
Adenylylsulfate kinase and related kinases	01034_3365	205	65.33	277	5.00E-73	CYSC_ACAM1	Adenylyl-sulfate kinase	*Acaryochloris marina* MBIC11017
ADP-ribosylglycohydrolase family protein	00150_0826	268	66.44	197	2.00E-49	B9YPM5_ANAAZ	ADP-ribosylation/Crystallin J1	'*Nostoc azollae*' 0708
Amidohydrolase	00199_2709	390	62.07	454	7.00E-126	B8HUF9_CYAP4	Amidohydrolase 2	*Cyanothece* sp. PCC 7425
aspartate racemase	00162_0519	265	55.42	280	1.00E-73	B5VZU3_SPIMA	Aspartate racemase	*Arthrospira maxima* CS-328
Cysteine synthase	00100_0038	340	56.01	369	3.00E-100	B4VZM9_9CYAN	Cysteine synthase	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Dimethylglycine N-methyltransferase	00162_0520	553	53.51	301	7.00E-80	A0YIB3_9CYAN	Putative glycine-sarcosine methyltransferase	*Lyngbya* sp. PCC 8106
Dimethylglycine N-methyltransferase	00360_3387	277	60.53	340	7.00E-92	Q83WC3_APHHA	Dimethylglycine N-methyltransferase	*Aphanothece halophytica*
ATPase family associated with various cellular activities (AAA)	00338_4328	238	70.21	139	2.00E-31	B4W316_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
ATPase family associated with various cellular activities (AAA)	00356_1916	320	62.3	369	2.00E-100	B0C9T9_ACAM1	Methanol dehydrogenase regulatory protein, putative	*Acaryochloris marina* MBIC11017
Predicted ATPase	00959_5086	382	63.47	283	9.00E-75	C7QR95_CYAP0	ATPase-like protein	*Cyanothece* sp. PCC 8802
dihydroorotase, homodimeric type	01054_0074	345	64.5	448	5.00E-124	PYRC_ACAM1	Dihydroorotase	*Acaryochloris marina* MBIC11017
Fe-S oxidoreductase	00162_0524	516	51.92	333	1.00E-89	Q10WJ6_TRIEI	Radical SAM	*Trichodesmium erythraeum* IMS101
conserved domain protein	00162_0525	516	57.32	207	4.00E-52	Q10WJ6_TRIEI	Radical SAM	*Trichodesmium erythraeum* IMS101
Fe-S oxidoreductase	00162_0526	516	53.69	554	8.00E-156	Q10WJ6_TRIEI	Radical SAM	*Trichodesmium erythraeum* IMS101
GMC oxidoreductase (Glucose-methanol-choline oxidoreductase family)	01079_5157	472	58.97	50.1	8.00E-05	B5W9M6_SPIMA	Glucose-methanol-choline oxidoreductase	*Arthrospira maxima* CS-328
oxidoreductase, zinc-binding dehydrogenase family protein	01124_0941	319	56.88	337	1.00E-90	B4W230_9CYAN	Alcohol dehydrogenase GroES-like domain family	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420

Flagellar motor protein OmpA	01009_5001	215	50.31	125	3.00E-27	Q3MC14_ANAVT	Putative uncharacterized protein	*Anabaena variabilis* ATCC 29413
gamma-glutamyltranspeptidase	00696_0298	472	57.5	263	2.00E-68	Q4BU94_CROWT	Gamma-glutamyltranspeptidase (Fragment)	*Crocosphaera watsonii* WH 8501
GDP-mannose 4,6-dehydratase	01226_4081	363	79.06	600	1.00E-169	Q5N5L7_SYNP6	GDP-mannose 46-dehydratase	*Synechococcus elongatus* PCC 6301
glycogen/starch synthases, ADP-glucose type	00241_1024	491	65.44	667	0.00E+00	B4AUH6_9CHRO	Glycogen/starch synthase, ADP-glucose type	*Cyanothece* sp. PCC 7822
Heparan sulfate D-glucosaminyl 3-O-sulfotransferase	00397_1672	309	51.13	303	2.00E-80	B7KBX2_CYAP7	Sulfotransferase	*Cyanothece* sp. PCC 7424
Heparan sulfate D-glucosaminyl 3-O-sulfotransferase	00397_1674	298	51.85	310	1.00E-82	B4B416_9CHRO	Sulfotransferase	*Cyanothece* sp. PCC 7822
Protein-tyrosine sulfotransferase TPST1/TPST2	00847_4028	320	51.05	293	1.00E-77	B7KGL5_CYAP7	Sulfotransferase	*Cyanothece* sp. PCC 7424
Uncharacterized conserved protein, Nitro_FeMo-Co	00301_1138	130	53.6	139	7.00E-32	B4VK95_9CYAN	Dinitrogenase iron-molybdenum cofactor, putative	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
molybdenum cofactor biosynthesis protein B	00478_0861	195	65.12	238	2.00E-61	Q4C4T4_CROWT	Molybdenum cofactor biosynthesis protein	*Crocosphaera watsonii* WH 8501
NAD(P)(+) transhydrogenase (AB-specific), alpha subunit	00327_3230	538	66.15	694	0.00E+00	B0CDX4_ACAM1	NAD(P) transhydrogenase, alpha subunit	*Acaryochloris marina* MBIC11017
peptide deformylase	01181_3853	179	52.3	192	1.00E-47	Q3M2X2_ANAVT	Peptide deformylase	*Anabaena variabilis* ATCC 29413
Phosphoenolpyruvate carboxykinase (ATP)	00100_0013	633	60.16	790	0.00E+00	B0JW12_MICAN	Putative uncharacterized protein	*Microcystis aeruginosa* NIES-843
phosphoglucomutase	00138_2548	544	68.75	783	0.00E+00	B7JZ50_CYAP8	Phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain I	*Cyanothece* sp. PCC 8801
phosphopantetheine attachment domain protein	00430_5228	3242	59.09	114	2.00E-24	B2IXJ9_NOSP7	Amino acid adenylation domain protein	*Nostoc punctiforme* PCC 73102
Predicted transcriptional regulator	00696_0316	127	57.89	131	2.00E-29	B4VNB4_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
putative transcriptional regulator, PvuIIC	00125_3718	78	52.94	76.6	8.00E-13	B4AWI0_9CHRO	Transcriptional regulator, XRE family	*Cyanothece* sp. PCC 7822
short chain dehydrogenase	00906_2610	657	52.02	664	0.00E+00	A3IUF6_9CHRO	Short chain dehydrogenase	*Cyanothece* sp. CCY0110
retroviral aspartyl protease	00266_3919	55	57.14	60.1	1.00E-07	B4VXN4_9CYAN	Putative uncharacterized protein	*Coleofasciculus* (*Microcoleus*) *chthonoplastes* PCC 7420
Subtilisin-related protease/Vacuolar protease B	00384_3830	1344	53.15	390	2.00E-106	B4VZU0_9CYAN	Cna protein B-type domain	*Coleofasciculus* (*Microcoleus*)

								chthonoplastes PCC 7420
Subtilisin-related protease/Vacuolar protease B	00384_3831	2534	52.33	389	9.00E-106	B0JT18_MICAN	Putative peptidase	*Microcystis aeruginosa* NIES-843
UDP-glucose 4-epimerase/UDP-sulfoquinovose synthase	00665_1257	400	72.73	601	4.00E-170	B1XII0_SYNP2	NAD dependent epimerase/dehydratase family protein	*Synechococcus* sp. PCC 7002

230

SUPPLEMENTAL REFERENCES

1. **Salman V, Amann R, Girnth AC, Polerecky L, Bailey JV, Hogslund S, Jessen G, Pantoja S, Schulz-Vogt HN.** 2011. A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. Systematic and Applied Microbiology **34:**243-259.
2. **McKay LJ, MacGregor BJ, Biddle JF, Albert DB, Mendlovitz HP, Hoer DR, Lipp JS, Lloyd KG, Teske AP.** 2012. Spatial heterogeneity and underlying geochemistry of phylogenetically diverse orange and white *Beggiatoa* mats in Guaymas Basin hydrothermal sediments. Deep-Sea Research I **67:**21-31.
3. **Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüßmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H.** 2004. ARB: a software environment for sequence data. Nucleic Acids Research **32:**1363-1371.
4. **Jukes TH, Cantor CR.** 1969. Evolution of protein molecules, p. 21-132. *In* Munro HN (ed.), Mammalian Protein Metabolism, vol. III. Academic Press, New York.
5. **Hinck S, Mußmann M, Salman V, Neu TR, Lenk S, de Beer D, Jonkers HM.** 2011. Vacuolated *Beggiatoa*-like filaments from different hypersaline environments form a novel genus. Environmental Microbiology **13:**3194-3205.
6. **Stamatakis A.** 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688-2690.
7. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution **28:**2731-2739.
8. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32:**1792-1797.
9. **Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP.** 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology **61:**539-542.
10. **Yang ZH.** 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution **11:**367-372.
11. **Adachi J, Waddell PJ, Martin W, Hasegawa M.** 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. Journal of Molecular Evolution **50:**348-358.
12. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4:**406-425.

13. **Felsenstein J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783-791.
14. **Zuckerkandl E, Pauling L.** 1965. Evolutionary convergence and divergence in proteins, p. 97-166. *In* Bryson V, Vogel HJ (ed.), Evolving Genes and Proteins. Academic Press, New York.
15. **Henikoff S, Henikoff JG.** 1992. Amino-acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America **89:**10915-10919.
16. **Whelan S, Goldman N.** 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution **18:**691-699.
17. **Papadopoulos JS, Agarwala R.** 2007. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics **23:**1073-1079.
18. **Zengel JM, Lindahl L.** 1994. Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*, p. 331-370. *In* Cohn WE, Moldave K (ed.), Progress in Nucleic Acid Research and Molecular Biology, vol. 47.
19. **Korobeinikova AV, Gongadze GM, Korepanov AP, Eliseev BD, Bazhenova MV, Garber MB.** 2008. 5S rRNA-recognition module of CTC family proteins and its evolution. Biochemistry-Moscow **73:**156-163.
20. **Stöffler G.** 1974. Structure and function of the *Escherichia coli* ribosome: Immunochemical analysis, p. 615-667. *In* Nomura M, Tissiéres A, Lengyel P (ed.), Ribosomes. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
21. **Pettersson I, Hardy SJS, Liljas A.** 1976. The ribosomal protein L8 is a complex of L7/L12 and L10. FEBS Letters **64:**135-138.
22. **Grosjean H, de Crécy-Lagard V, Marck C.** 2010. Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. FEBS Letters **584:**252-264.
23. **Tåquist H, Cui Y, Ardell DH.** 2007. TFAM 1.0: an online tRNA function classifier. Nucleic Acids Research **35:**W350-W353.
24. **Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J.** 2009. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Research **37:**D159-D162.
25. **Mußmann M, Hu FZ, Richter M, de Beer D, Preisler A, Jørgensen BB, Huntemann M, Glöckner FO, Amann R, Koopman WJH, Lasken RS, Janto B, Hogg J, Stoodley P, Boissy R, Ehrlich GD.** 2007. Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. PLoS Biology **5:**1923-1937.
26. **Mezzino MJ, Strohl WR, Larkin JM.** 1984. Characterization of *Beggiatoa alba*. Archives of Microbiology **137:**139-144.
27. **Aruga S, Kamagata Y, Kohno T, Hanada S, Nakamura K, Kanagawa T.** 2002. Characterization of filamentous Eikelboom type 021N bacteria and description of *Thiothrix disciformis* sp. nov. and *Thiothrix flexilis* sp. nov. International Journal of Systematic and Evolutionary Microbiology **52:**1309-1316.

28. **Nielsen PH, Kragelund C, Seviour RJ, Nielsen JL.** 2009. Identity and ecophysiology of filamentous bacteria in activated sludge. FEMS Microbiology Reviews **33:**969-998.
29. **Larkin JM, Shinabarger DL.** 1983. Characterization of *Thiothrix nivea*. International Journal of Systematic Bacteriology **33:**841-846.
30. **Klappenbach JA, Pierson BK.** 2004. Phylogenetic and physiological characterization of a filamentous anoxygenic photoautotrophic bacterium '*Candidatus* Chlorothrix halophila' gen. nov., sp nov., recovered from hypersaline microbial mats. Archives of Microbiology **181:**17-25.
31. **Sutcliffe IC.** 2010. A phylum level perspective on bacterial cell envelope architecture. Trends in Microbiology **18:**464-470.
32. **Holt JG, Lewin RA.** 1968. *Herpetosiphon aurantiacus* gen. et sp. n., a new filamentous gliding organism. J. Bacteriol. **95:**2407-2408.
33. **Keppen OI, Baulina OI, Lysenko AM, Kondrateva EN.** 1993. A new green bacterium belonging to the Chloroflexaceae family. Microbiology **62:**179-185.
34. **Klatt CG, Bryant DA, Ward DM.** 2007. Comparative genomics provides evidence for the 3-hydroxypropionate autotrophic pathway in filamentous anoxygenic phototrophic bacteria and in hot spring microbial mats. Environmental Microbiology **9:**2067-2078.
35. **Hanada S, Takaichi S, Matsuura K, Nakamura K.** 2002. *Roseiflexus castenholzii* gen. nov., sp nov., a thermophilic, filamentous, photosynthetic bacterium that lacks chlorosomes. International Journal of Systematic and Evolutionary Microbiology **52:**187-193.
36. **Lewin RA, Lounsbery DM.** 1969. Isolation, cultivation and characterization of flexibacteria. Journal of General Microbiology **58:**145-170.
37. **van Veen WL, van der Kooij D, Geuze ECWA, van der Vlies AW.** 1973. Investigations on the sheathed bacterium *Haliscomenobacter hydrossis* gen. n., sp. n., isolated from activated sludge. Antonie van Leeuwenhoek **39:**207-216.
38. **Larkin JM, Williams PM.** 1978. *Runella slithyformis* gen. nov., sp. nov., a curved, nonflexible, pink bacterium. International Journal of Systematic Bacteriology **28:**32-36.
39. **Lail K, Sikorski J, Saunders E, Lapidus A, Del Rio TG, Copeland A, Tice H, Cheng JF, Lucas S, Nolan M, Bruce D, Goodwin L, Pitluck S, Ivanova N, Mavromatis K, Ovchinnikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Chain P, Brettin T, Detter JC, Schütze A, Rohde M, Tindall BJ, Goker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klen HP, Chen F.** 2010. Complete genome sequence of *Spirosoma linguale* type strain (1^T). Standards in Genomic Sciences **2:**176-185.
40. **Ten LN, Xu JL, Jin FX, Im WT, Oh HM, Lee ST.** 2009. *Spirosoma panaciterrae* sp nov., isolated from soil. International Journal of Systematic and Evolutionary Microbiology **59:**331-335.

41. **Franzmann PD, Skerman VBD.** 1984. *Gemmata obscuriglobus*, a new genus and species of the budding bacteria. Antonie van Leeuwenhoek Journal of Microbiology **50:**261-268.
42. **Miyashita H, Ikemoto H, Kurano N, Miyachi S, Chihara M.** 2003. *Acaryochloris marina* gen. et sp. nov. (Cyanobacteria), an oxygenic photosynthetic prokaryote containing chl *d* as a major pigment. J. Phycol. **39:**1247-1253.
43. **Leikoski N, Fewer DP, Jokela J, Wahlsten M, Rouhiainen L, Sivonen K.** 2010. Highly diverse cyanobactins in strains of the genus *Anabaena*. Applied and Environmental Microbiology **76:**701-709.
44. **Skill SC, Smith RJ.** 1987. Synchronous akinete germination and heterocyst differentiation in *Anabaena* PCC 7937 and *Nostoc* PCC 6720. Journal of General Microbiology **133:**299-303.
45. **Janssen PJ, Morin N, Mergeay M, Leroy B, Wattiez R, Vallaeys T, Waleron K, Waleron M, Wilmotte A, Quillardet P, de Marsac NT, Talla E, Zhang C-C, Leys N.** 2010. Genome sequence of the edible cyanobacterium *Arthrospira* sp. PCC 8005. Journal of Bacteriology **192:**2465-2466.
46. **Carrieri D, Momot D, Brasg IA, Ananyev G, Lenz O, Bryant DA, Dismukes GC.** 2010. Boosting autofermentation rates and product yields with sodium stress cycling: application to production of renewable fuels by cyanobacteria. Applied and Environmental Microbiology **76:**6455-6462.
47. **Cheevadhanarak S, Paithoonrangsarid K, Prommeenate P, Kaewngam W, Musigkain A, Tragoonrung S, Tabata S, Kaneko T, Chaijaruwanich J, Sangsrakru D, Tangphatsornruang S, Chanprasert J, Tongsima S, Kusonmano K, Jeamton W, Dulsawat S, Klanchui A, Vorapreeda T, Chumchua V, Khannapho C, Thammarongtham C, Plengvidhya V, Subudhi S, Hongsthong A, Ruengjitchatchawalya M, Meechai A, Senachak J, Tanticharoen M.** 2012. Draft genome sequence of *Arthrospira platensis* C1 (PCC9438).
48. **Fujisawa T, Narikawa R, Okamoto S, Ehira S, Yoshimura H, Suzuki I, Masuda T, Mochimaru M, Takaichi S, Awai K, Sekine M, Horikawa H, Yashiro I, Omata S, Takarada H, Katano Y, Kosugi H, Tanikawa S, Ohmori K, Sato N, Ikeuchi M, Fujita N, Ohmori M.** 2010. Genomic structure of an economically important cyanobacterium, *Arthrospira* (*Spirulina*) *platensis* NIES-39. DNA Research **17:**85-103.
49. **Webb EA, Ehrenreich IM, Brown SL, Valois FW, Waterbury JB.** 2009. Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, *Crocosphaera watsonii*, isolated from the open ocean. Environmental Microbiology **11:**338-348.
50. **Waterbury JB, Willey JM.** 1988. Isolation and growth of marine planktonic cyanobacteria. Methods in Enzymology **167:**100-105.
51. **Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, Steward GF, Hansen A, Karl DM.** 2001. Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. Nature **412:**635-638.

52. **Reddy KJ, Haskell JB, Sherman DM, Sherman LA.** 1993. Unicellular, aerobic nitrogen-fixing cyanobacteria of the genus *Cyanothece*. J. Bacteriol. **175:**1284-1292.
53. **Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY.** 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. Journal of General Microbiology **111:**1-61.
54. **Singh PK.** 1973. Nitrogen fixation by the unicellular blue-green alga *Aphanothece*. Arch. Mikrobiol. **92:**59-62.
55. **Huang T-C, Chow T-J.** 1988. Comparative studies of some nitrogen-fixing unicellular cyanobacteria isolated from rice fields. Journal of General Microbiology **134:**3089-3097.
56. **Saker ML, Neilan BA.** 2001. Varied diazotrophies, morphologies, and toxicities of genetically similar isolates of *Cylindrospermopsis raciborskii* (Nostocales, Cyanophyceae) from northern Australia. Applied and Environmental Microbiology **67:**1839-1845.
57. **Stucken K, John U, Cembella A, Murillo AA, Soto-Liebe K, Fuentes-Valdes JJ, Friedel M, Plominsky AM, Vasquez M, Glockner G.** 2010. The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. PLoS One **5:**e9235.
58. **Stal LJ, Krumbein WE.** 1981. Aerobic nitrogen fixation in pure cultures of a benthic marine *Oscillatoria* (cyanobacteria). FEMS Microbiology Letters **11:**295-298.
59. **Kaneko T, Nakajima N, Okamoto S, Suzuki I, Tanabe Y, Tamaoki M, Nakamura Y, Kasai F, Watanabe A, Kawashima K, Kishida Y, Ono A, Shimizu Y, Takahashi C, Minami C, Fujishiro T, Kohara M, Katoh M, Nakazaki N, Nakayama S, Yamada M, Tabatai S, Watanabe MM.** 2007. Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. DNA Research **14:**247-256.
60. **Staal M, Stal LJ, Hekkert ST, Harren FJM.** 2003. Light action spectra of N_2 fixation by heterocystous cyanobacteria from the Baltic Sea. Journal of Phycology **39:**668-677.
61. **Adolph KW, Haselkorn R.** 1971. Isolation and characterization of a virus infecting the blue-green alga *Nostoc muscorum*. Virology **46:**200-208.
62. **Ran LA, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng WW, Lapidus A, Lowry S, Haselkorn R, Bergman B.** 2010. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One **5:**e11486.
63. **Zheng WW, Bergman B, Chen B, Zheng SP, Xiang G, Rasmussen U.** 2009. Cellular responses in the cyanobacterial symbiont during its vertical transfer between plant generations in the *Azolla microphylla* symbiosis. New Phytologist **181:**53-61.
64. **Stucken K, Murillo AA, Soto-Liebe K, Fuentes-Valdes JJ, Mendez MA, Vasquez M.** 2009. Toxicity phenotype does not correlate with phylogeny of *Cylindrospermopsis raciborskii* strains. Systematic and Applied Microbiology **32:**37-48.

65. **Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF.** 2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME Journal **1:**703-713.
66. **Becraft ED, Cohan FM, Kuhl M, Jensen SI, Ward DM.** 2011. Fine-scale distribution patterns of *Synechococcus* ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. Applied and Environmental Microbiology **77:**7689-7697.
67. **Roberts RJ, Vincze T, Posfai J, Macelis D.** 2010. REBASE - a database for DNA restriction and modification: enzymes, genes, and genomes. Nucleic Acids Research **38:**D234-D236.
68. **Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fazzini F, Fedotov A, Foulger R, Garavelli J, Castro LG, Huntley R, Jacobsen J, Kleen M, Laiho K, Legge D, Lin QA, Liu WD, Luo J, Orchard S, Patient S, Pichler K, Poggioli D, Pontikos N, Pruess M, Rosanoff S, Sawford T, Sehra H, Turner E, Corbett M, Donnelly M, van Rensburg P, Xenarios I, Bougueleret L, Auchincloss A, Argoud-Puy G, Axelsen K, Bairoch A, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, deCastro E, Coudert E, Cusin I, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehant S, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Kappler T, Keller G, Lara V, Lemereier P, Lieberherr D, Martin X, Masson P, Moinat M, Morgat A, Paesano S, Pedruzzi I, Pilbout S, Poux S, Pozzato M, Redaschi N, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stanley E, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Barker WC, Chen CM, Chen YX, Dubey P, Huang HZ, Mazumder R, McGarvey P, Natale DA, Natarajan TG, Nchoutmboube J, Roberts NV, Suzek BE, Ugochukwu U, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Zhang JA.** 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Research **39:**D214-D219.