frontiers
in Microbiology

ORIGINAL RESEARCH
published: 16 December 2015
doi: 10.3389/fmicb.2015.01397

Abundant Intergenic TAACTGA Direct Repeats and Putative Alternate RNA Polymerase β′ Subunits in Marine *Beggiatoaceae* Genomes: Possible Regulatory Roles and Origins

Barbara J. MacGregor*

Department of Marine Sciences, University of North Carolina–Chapel Hill, Chapel Hill, NC, USA

OPEN ACCESS

Edited by:
Beth Orcutt,
Bigelow Laboratory for Ocean Sciences, USA

Reviewed by:
Jeremy Dodsworth,
California State University, San Bernardino, USA
Emily J. Fleming,
California State University, Chico, USA

***Correspondence:**
Barbara J. MacGregor
bmacgreg@unc.edu

Specialty section:
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology

Received: *28 August 2015*
Accepted: *23 November 2015*
Published: *16 December 2015*

Citation:
MacGregor BJ (2015) Abundant Intergenic TAACTGA Direct Repeats and Putative Alternate RNA Polymerase β′ Subunits in Marine Beggiatoaceae Genomes: Possible Regulatory Roles and Origins. Front. Microbiol. 6:1397. doi: 10.3389/fmicb.2015.01397

The genome sequences of several giant marine sulfur-oxidizing bacteria present evidence of a possible post-transcriptional regulatory network that may have been transmitted to or from two distantly related bacteria lineages. The draft genome of a *Cand.* "Maribeggiatoa" filament from the Guaymas Basin (Gulf of California, Mexico) seafloor contains 169 sets of TAACTGA direct repeats and one indirect repeat, with two to six copies per set. Related heptamers are rarely or never found as direct repeats. TAACTGA direct repeats are also found in some other *Beggiatoaceae*, *Thiocystis violascens*, a range of Cyanobacteria, and five Bacteroidetes. This phylogenetic distribution suggests they may have been transmitted horizontally, but no mechanism is evident. There is no correlation between total TAACTGA occurrences and repeats per genome. In most species the repeat units are relatively short, but longer arrays of up to 43 copies are found in several Bacteroidetes and Cyanobacteria. The majority of TAACTGA repeats in the *Cand.* "Maribeggiatoa" Orange Guaymas (BOGUAY) genome are within several nucleotides upstream of a putative start codon, suggesting they may be binding sites for a post-transcriptional regulator. Candidates include members of the ribosomal protein S1, Csp (cold shock protein), and Csr (carbon storage regulator) families. No pattern was evident in the predicted functions of the open reading frames (ORFs) downstream of repeats, but some encode presumably essential products such as ribosomal proteins. Among these is an ORF encoding a possible alternate or modified RNA polymerase beta prime subunit, predicted to have the expected subunit interaction domains but lacking most catalytic residues. A similar ORF was found in the *Thioploca ingrica* draft genome, but in no others. In both species they are immediately upstream of putative sensor kinase genes with nearly identical domain structures. In the marine *Beggiatoaceae*, a role for the TAACTGA repeats in translational regulation is suggested. More speculatively, the putative alternate RNA polymerase subunit could be a negative transcriptional regulator.

Keywords: heptamer repeats, DNA-directed RNA polymerase, beta prime subunit, *Beggiatoaceae*, Cyanobacteria, Bacteroidetes, orange Guaymas "Maribeggiatoa"

INTRODUCTION

Organic-rich sediments surrounding hydrothermal sites on the Guaymas Basin sea floor often host luxuriant microbial mats, visually dominated by large filamentous, vacuolated, orange-pigmented, and unpigmented *Beggiatoaceae* (Jannasch et al., 1989). From 16S rRNA data, these appear to belong to several distinct species. None of them are yet in culture, but physiological (McHatton et al., 1996) and genomic (MacGregor et al., 2013a) studies are consistent with a sulfur-oxidizing, nitrate-reducing metabolism. They are gradient dwellers, living between hot sulfidic fluids flowing up through the sediments below and cold, oxygenated overlying seawater. In general, the pigmented forms are found toward the center of mats, where flow rates (and temperature) are higher, while unpigmented forms are more concentrated at the periphery (McKay et al., 2012). The pigmentation is thought to be due to high concentrations of an octaheme cytochrome, possibly a nitrite reductase (MacGregor et al., 2013b). The Orange Guaymas *Cand.* "Maribeggiatoa" (BOGUAY) draft genome (MacGregor et al., 2013a) was obtained from a single orange filament cleaned of epibionts.

In the course of analyzing this genome, numerous short direct repeats of the heptanucleotide TAACTGA were noticed, particularly in intergenic regions directly upstream of translational start codons. The genomes of the marine *Beggiatoaceae Cand*. "Thiomargarita nelsonii" and *Thioploca ingrica,* and *Thiocystis violascens* (*Chromatiaceae*)—but not the freshwater *Beggiatoa alba*—also feature these repeats to varying degrees. Database searches further found TAACTGA direct repeats in some Cyanobacteria and a few Bacteroidetes, consistent with earlier evidence (MacGregor et al., 2013c) for genetic exchange between these groups and the *Beggiatoaceae*.

Tandem direct repeats of short nucleotide sequences have a very sporadic distribution in bacteria. In a comprehensive study, Mrázek et al. (2007) examined the distribution of what were termed long simple sequence repeats (LSSR) in prokaryotic genome sequences available at the time (2007). Repeat units of 1–11 nt were considered, and "long" was defined as series of repeats longer than statistically expected in a given genome. Species rich in LSSRs could be divided into those with repeat units primarily 1–4 or 5–11 nt long. They were phylogenetically scattered: for example, the 10 genomes identified with the most 5–11 nt repeats included four Betaproteobacteria (all *Burkholderia* spp.), two Cyanobacteria, three Actinobacteria, and one Gammaproteobacterium (*Xanthomonas campestris* ATCC 33913). Heptanucleotide repeats were the most abundant category in most genomes; it was proposed interaction of these with DNA polymerase might favor slippage and therefore duplications or deletions, and that 7 nt might be the length of sequence interacting with the polymerase. It was also noted that repeat units whose lengths are multiples of three were the most likely to be found within coding regions, presumably because series of them can be expanded and contracted without truncating a protein as long as they do not generate stop codons.

The same group went on to examine the genome-wide distribution of LSSR in several host-adapted pathogenic bacteria (Guo and Mrázek, 2008). Such repeats have been proposed and in some cases demonstrated to be involved in phase variation via slippage during DNA replication, turning on or off expression of virulence functions at either the transcriptional or the translational level. Some LSSR were in fact associated with antigenicity functions, such as envelope biogenesis genes, but COG classifications including these were not significantly overrepresented among the very diverse repeat-associated genes.

The genome-wide distribution of SSR (here abbreviating "simple satellite repeats") in *Escherichia coli* has also been examined (Gur-Arie et al., 2000), considering only 1–6 nt units. For tetranucleotides, the longest unit reported in this regard, 78.9% of repeats were found in coding regions—very nearly the same proportion of the whole genome that is coding (79.5%). The repeats in intergenic regions did not show any particular concentration near translational start sites.

The two experimentally studied examples of bacterial tandem repeats between a promoter and a start codon are both upstream of surface proteins involved in phase variation in the respiratory pathogen *Moraxella catarrhalis*. A tract of either 9 or 10 G residues occurs 30 nt upstream of the translational start for the UspA1 gene (Lafontaine et al., 2001), which allows adhesion of the bacterium to human epithelial cells. Nine-residue G tracts were associated with high expression and 10-residue tracts with low expression. The tetranucleotide AGAT is found in strain-dependent copy numbers (from 6 to 23) in the 5′ untranslated regions of mRNAs for UspA2 (Attia and Hansen, 2006), a surface protein conferring resistance to human serum. Mutational studies in one strain found highest UspA2 expression with 18 copies.

This study describes the distribution of TAACTGA heptamer repeats in the BOGUAY genome, and the limited number of other species in which they have been found. Possible roles in translational regulation and genome rearrangement will be considered, depending on the length and position of the different repeat arrays. A possible alternate or derived RNA polymerase beta prime subunit gene identified in the Orange Guaymas "Maribeggiatoa" and *Thioploca ingrica* genomes is also discussed.

MATERIALS AND METHODS

An orange tuft retrieved from core 4489-10 from RV *Atlantis*/HOV *Alvin* cruise AT15-40 (13 December 2008) at the UNC Gradient Mat site in Guaymas Basin, Gulf of California, Mexico (latitude 27° 0.450300′ N, longitude 111° 24.532320′ W, depth 2001 m) was cleaned of epibionts; its DNA amplified, tested for genetic purity, sequenced, assembled, and annotated; and the genome sequence checked for completeness, as previously described (MacGregor et al., 2013a,c). A total of 99.3% of the sequence was assembled into 822 contigs, suggesting good coverage was achieved. 4.7 Mb of sequence was recovered, with 80% of it forming large (≥15 kb) contigs. Throughout this paper, the genome is referred to as BOGUAY (from "*Beggiatoa* orange Guaymas") and annotated sequences are referred to by 5-digit contig and 4-digit open reading frame (ORF) numbers (e.g., 00024_0691) or by ORF number alone (e.g., BOGUAY_0691). Additional sequence analysis was carried out using a combination of the JCVI-supplied annotation,

TABLE 1 | Orientation of TAACTGA repeats in the BOGUAY genome.

Repeats	Orientation																Total
	Toward start codon, no RBS				Toward start codon, with RBS			Toward stop codon				Contig end	Other				
	Intergenic	Overlap	In ORF	Total	Intergenic	Overlap	Total	Intergenic	Overlap	In ORF	Total	Total	Intergenic	In ORF	Total		
2	40	3	3	46	10	1	11	8	–	–	8	2	–	1	1		68
2, split	15	–	–	15	1	–	1	1	–	–	1	–	1	–	1		18
3	27	4	1	32	2	1	3	1	1	1	3	2	–	–	–		40
3, split	11	1	1	13	–	–	–	–	–	–	–	–	–	1	1		14
4	15	3	1	19	–	–	–	–	–	–	–	–	–	–	–		19
4, split	4	–	–	4	–	–	–	–	–	–	–	2	–	–	–		6
5	2	–	–	2	–	–	–	–	–	–	–	–	–	–	–		2
5, split	–	–	–	–	–	–	–	–	–	–	–	–	–	1	1		1
6	1	–	–	1	–	–	–	–	–	–	–	–	–	–	–		1
Inverted repeat	–	–	–	–	–	–	–	–	–	–	–	–	–	1	1		1
Total	115	11	6	132	13	2	15	10	1	1	12	6	1	4	5		170

"Split" sets have a different but related 7-mer between two TAACTGA sequences.

the IMG/ER (Markowitz et al., 2009) and RAST (Aziz et al., 2008) platforms, and BLASTN, BLASTX, and BLASTP and PSIBLAST searches of the GenBank nr databases. Nucleic acid and amino acid sequence alignments were performed in MEGA5 (Tamura et al., 2011) using MUSCLE (Edgar, 2004) and small adjustments made manually. For identification of other TAACTGA-containing genomes, the GenBank nr database was searched with seven direct repeats of the TAACTGA sequence, using the default "short query" settings. For each strain with a sequence identified by this search, the genome sequence was searched for all TAACTGA direct repeats (in both orientations). RNA structure predictions are the first results from a minimum free energy calculation using the default settings of the MaxExpect algorithm from the RNAstructure Web Server (http://rna.urmc.rochester.edu/RNAstructureWeb/, Reuter and Mathews, 2010). Translations were done via the ExPASy portal of the Swiss Institute of Bioinformatics (Artimo et al., 2012). Protein domains were identified in CDD (Marchler-Bauer et al., 2011).

RESULTS AND DISCUSSION

Overview of Sequenced *Beggiatoaceae*

The *Beggiatoaceae* family of giant sulfur bacteria includes species with a range of morphologies and habitats, very few of which have as yet been cultivated. Their classification is still in progress (Salman et al., 2011, 2013), but it is clear that many strains formerly designated *Beggiatoa* should be reclassified. Genomic sequence data are currently available for a small but diverse selection of these: complete or near-complete genome sequences for *B. alba* B18LD (Lucas et al. unpublished), *Thioploca ingrica* (Kojima et al., 2015), and Orange Guaymas "Maribeggiatoa" (MacGregor et al., 2013a,b,c); a partial sequence for *Cand.* "Thiomargarita nelsonii" (Mußmann et al., unpublished); and very partial sequences for two single filaments from the Baltic Sea, designated *Cand.* "Isobeggiatoa" PS and SS (Mussmann et al., 2007). By 16S rRNA gene sequence analysis, *B. alba* is in a separate clade from the rest of these (Salman et al., 2013).

Abundance and Distribution of TAACTGA Repeats in the BOGUAY and Other *Beggiatoaceae* Genomes

The Orange Guaymas "Maribeggiatoa" (BOGUAY) genome, with ∼5330 annotated genes, contains some 169 sets of direct TAACTGA repeats and one indirect repeat, with between two and six copies per set (**Table 1**). Thirty-six of the sets are split by one or two different but related 7 bp sequences. Their distribution is not random: most are in a "forward" orientation upstream of a putative start codon, with the largest single category ending 1 nt upstream (**Figure 1**). All but 25 sets are completely intergenic. Of the rest, 14 overlap the end of an upstream ORF, with 13 in forward orientation to a downstream ORF; 10 are interior to ORFs in reverse orientation (Supplemental Table 1); and one is an inverted repeat near the end of an ORF, with the repeat units separated by one base pair (**Table 2**). There are an additional 819 singletons, whose distribution was not examined, for a total of 1357. TAACTGA repeats are also found in the "Isobeggiatoa"

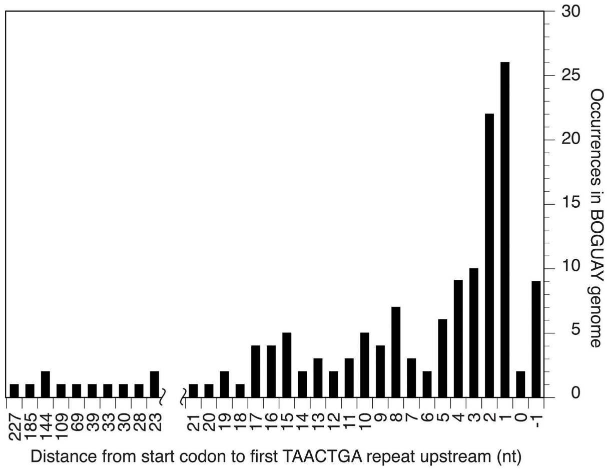


FIGURE 1 | Distance between start codon and first repeat for BOGUAY intergenic TAACTGA repeats.

sp. PS and SS genomes, but these are too incomplete for thorough comparison. Of other sequenced *Beggiatoaceae*, Cand. "Thiomargarita nelsonii" has a similar number of repeats, and a higher proportion of doublets and triplets, but fewer longer sets; *T. ingrica* has a similar number of TAACTGA copies, but very few as direct repeats; and *B. alba* has less than half as many total copies and no direct repeats (**Figures 2A,B**, Supplemental Table 2).

Direct Repeats of Sequences Similar to TAACTGA are Rare in the BOGUAY Genome

A survey of the BOGUAY genome for heptamers with a single-base difference to TAACTGA (**Table 3**) showed that while some of these are in similar or greater abundance than TAACTGA as singletons, the maximum number of doublets for any of them was six, and only two had any longer sets of direct repeats (one of four units, one of six). Several scrambled versions of TAACTGA were also searched; all are at lower to considerably lower abundance as singletons, and none is found as even a single direct repeat. Factors such as coding potential likely influence the distribution of each of these, and some permutations may be selected against as interfering with whatever function(s) TAACTGA repeats may have, but TAACTGA does appear to be a favored sequence.

Predicted Characteristics of RNA and Amino Acid Sequences that Might be Produced from TAACTGA Repeats

If the BOGUAY TAACTGA repeats have common function(s), these could be at the DNA, RNA, or in a few cases protein level. At the DNA level, repeat sequences can serve as recombinational and mutational hot spots (reviewed in Lovett, 2004; Zhou et al., 2014), or as binding sites for regulatory proteins. They could conceivably also mark the site of transposon excisions; some transposon insertions can generate 7 nt direct repeats (Sallam et al., 2006), although in the studied cases they seem usually to resolve to singletons upon excision (Foster et al., 1981).

At the RNA level, the repeats may again be protein-binding sites (or interrupt existing ones), and/or impart secondary structure. As direct repeats in up to six copies, however, TAACTGA is not predicted to generate any particular RNA secondary structure in either orientation (**Table 3**), unless by interaction with surrounding sequences.

At the protein level, translation of TAACTGA and its reverse complement (TCAGTTA) reveals what is probably a major factor controlling genomic distribution of these sequences. In the "forward" orientation, translation of TAACTGA repeats yields the repeating amino acid sequence LITDN–, where dashes represent stop codons. These can therefore overlap the end of coding sequences by no more than 18 nt, or two full repeats plus four nucleotides. If repeats are carried by mobile elements, their introduction into coding sequences in forward orientation will terminate the gene, and usually be deleterious. In some locations it might be tolerated however, for example between the subunits of modular proteins, or at the beginning or end of a protein. Possible examples will be discussed below.

Translation of repeats in the "reverse" orientation yields the repeating sequence LSVISYQ. At first glance, this suggests a leucine zipper dimerization domain (reviewed in Parry et al., 2008), with nonpolar residues in the first (L) and fourth (I) positions, but there are no charged amino acids for interactions on the other face of the predicted helix, and the nonpolar third position (V) is unusual. According to the algorithm of Bornberg-Bauer et al. (1998), this sequence does not have the requisite leucine zipper coiled-coil structure even when 20 or more amino acid repeats are included. *Ab initio* structure predictions (Xu and Zhang, 2012) for a peptide composed of seven LSVISYQ repeats (and several variants) suggest a structure dominated by antiparallel beta sheets (not shown), but structure in a real protein would depend on the number of repeats and on interactions with the rest of the protein.

Compared to other similar heptamers, TAACTGA has no obvious special features (**Table 3**): several have similar genomic abundances, many yield apparently similar local RNA conformations, a majority can be translated in "reverse" orientation, and all single-base mutants yield one or more stop codons in "forward" orientation. None of these properties shows a strong correlation with chromosomal abundance, or with occurrence as direct repeats. Assuming all relevant properties have been considered, this is consistent with TAACTGA repeats arising in one lineage and being horizontally transferred to others. The alternatives that this particular sequence became repeated independently in multiple isolated lineages, or was preserved as such in only a few, seem less likely.

Abundance and Distribution of TAACTGA Repeats in the Cyanobacteria and Bacteroidetes

A GenBank search for TAACTGA direct repeats found a very limited phylogenetic distribution (**Figure 2**). Outside of the *Beggiatoaceae*, considering only complete or near-complete

TABLE 2 | TAACTGA repeats within or overlapping BOGUAY ORFs.

A) Forward orientation

Upstream ORF with repeats		Repeat region sequence	Downstream ORF		
Description	COG group		Locus tag	Description	COG group
ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein	T	AAGGATG**AACTGATAACTGATAACTGATAACTGATAACTGA**AAAAaATG	00163_1011	response regulator receiver domain protein	T
adenylate and guanylate cyclase catalytic domain protein	T	aAAA**AACTGATAACTGATAACTGATAACTGATAACTGATAA**ACATG	00391_1539	hemerythrin HHE cation binding domain protein	T
PAS domain S-box	T	GGCTAAAAATTTTGAAATGTCTAAACTTAAAAAAT**TGATAACTGATAACTGATAACTGA**TTATG	01232_0432	response regulator receiver domain protein	
Tol-Pal system-associated acyl-CoA thioesterase	R	**TAACTGATAACTGATAACTGATAACTGA**AAAATG	00647_3823	TolQ	U
glycine cleavage system T protein	E	TTTT**TAACTGATAACTGATAACTGATAA**TG	00883_3308	glycine cleavage system H protein	E
COG3222, DUF2064	S	AGATGAC**AACTGATAACTGATAACTGATAACTGATA**TG	00162_0503	homoserine kinase, ThrB	R
CheR methyltransferase, SAM binding domain protein	N, T	AAGAATTTATCGCCGATTAAAC**ACTGATAACTGATAACTGATAACTGA**AAATG	00806_2998	protein-glutamate methylesterase CheB	
ABC transporter, ATP-binding protein	G, M	AAT**TGATAACTGATAACTGATAACT**A**ATAACTGATAA**AATG	00127_3139	methyltransferase domain protein	H
electron transport complex, RnfABCDGE type, E subunit	C	GCCCTAAAAAATGCCATAGATAAACGTTT**ATAACTGATAACTGATAA**TTGAACAATG	01035_2232	Uncharacterized protein conserved in bacteria COG3122	S
regulatory protein, FmdB family	S	ATGCCTATT~273nt~AGTTCAAGCA**CTGATAACTGATAACTGA**ATA**AGGA**AAATTATG	00106_0256	aspartyl-tRNA synthetase	J
hypothetical protein		ATGCCAGAT~111nt~AAAGTCA**TGATAACTGATAACTGATAACTGATA**C**AGGAG**TACACAAAAATG	00614_2860	roadblock/LC7 domain protein	R
hypothetical protein		ATGAGTTATCAG~90nt~TGGATGAAAA**ATAACTGATAACTG**AATATCTAATTA~181nt~GGATATTTTGATGGTGAGCAATTAGATAGCTGTGATG	01005_2993	hypothetical protein	
hypothetical protein		CAAAATTGT**TAACTGATAACTGATAA**TTGAAAAAAATGGCTATTCTGCGTTTAATG	00155_2445	hypothetical protein	
hypothetical protein		ATGGCAGGTTTAG~101nt~TTAAGCCATAATCATTTTCCAGAA**ATAACTGATAACTGATAACTGAT**TA	00472_0533	putative potassium uptake protein TrkA (on opposite strand, coming in from end of contig)	P

B) Reverse orientation

Repeat region sequence	ORF downstream of repeats (if any)		
	Locus tag	Description	COG group
CATGTATAAATTCG~60nt~AAACTAA**TAACTGATAACTGA**ATCCACTACTTTTTGTAA UDP-muramoylpentapeptide beta-N-acetylglucosaminyltransferase (MurG) (00938_0721)			
TTAAAGATTAAGCATTAATTATAA~50nt~ACCAA**TAACTGA**TAACTGG**TAACTGATAACTGA**AAATGAGTATTCACAATAAATTTA hypothetical protein (00199_2704)	00199_2702	methyl-accepting chemotaxis protein signaling domain protein	N, T
TTAAAAGAATA~60nt~CAACGGA**TAACTGATAACTGA**ATGaAAGAAAACT~24nt~CAT hypothetical protein (01017_0756)	01017_0755	Uncharacterized conserved protein	
TAAATTTGGC~124nt~ACCAGTTGAAA**TAACTGATAACTGA**TAATGGGCAACCACAAgGGAtTGCCCCTACTACTGATAACATAAGGAGTTT~124nt~TGATGAAGACTTG hypothetical protein (01192_0163)	01192_0161	glutamine synthetase, type I (>100 nt downstream)	E
TCAGATCACATTTTGATAATGAACTGTAGTCAGTTTTCCTAACGGG**TAACTGATAACTGATAACTGATAACTGA**TTATGAAAAG~25nt~ATCAT conserved domain protein (01153_1105)	01153_1104	receptor family ligand-binding protein	E
TTAAAAATC~40nt~AGAT**TAACTGATAACTGATAACTGA**TTAATTTAAACTTT~65nt~CGCCAAAAACAAAACATCCACCCCC~195nt~TATGAGATTTTA hypothetical protein (01318_2105)	01318_2106	conserved hypothetical protein (>200 nt downstream)	
TAATTTTAAT~96nt~TAATACCCtTTTTTCaAAAAAGGGTATTTTTATTGATTGA**TAACTGATAACTGATAACTGA**AAATATGGATTC~42nt~GGGCAT hypothetical protein (00794_2084)	00794_2083	putative proton/sodium-glutamate symport protein GltT	C
GTTAGCTGA**TAACTGATAACTGA**AAACTGAAAATGAAACCTGCACCTTATG putative corrinoid ABC transporter permease (00106_0223)	00106_0222	lipofamily protein	
CATCAATTCA~54nt~TTTTCC**TAACTGA**GCATTGA**TAACTGATAACTGATAACTGATAACTGA**TCAT 3048 3047 glycosyl hydrolase (01182_3048) conserved domain protein (01182_3047)			
DNA-directed RNA polymerase, beta' subunit (00100_0018) CATG~371nt~T**TAACTGA**TTACTGA**TAACTGA**TTACTGA**TAACTGA**TATG~1641nt~AT**TCAGTTATCAGTTA**TCAATTA**TCAGTTA**TCAGTGAATA~1991nt~TAG			

C) Inverted repeat

Upstream ORF			Downstream ORF		
Description	COG group		Locus tag	Description	COG group
segregation and condensation protein B		AAGTTG**TAACTGA**T**TCAGTTA**AGTAGGGTGGGTAGA~45nt~ATTCAA**AGGA**AACCGCAATGGCTTATGGTACATTTAAT**TCAGTTA**	00780_1610	conserved hypothetical protein	

➡ TAACTGA repeat	—● stop codon
→ start codon	□ Possible RBS

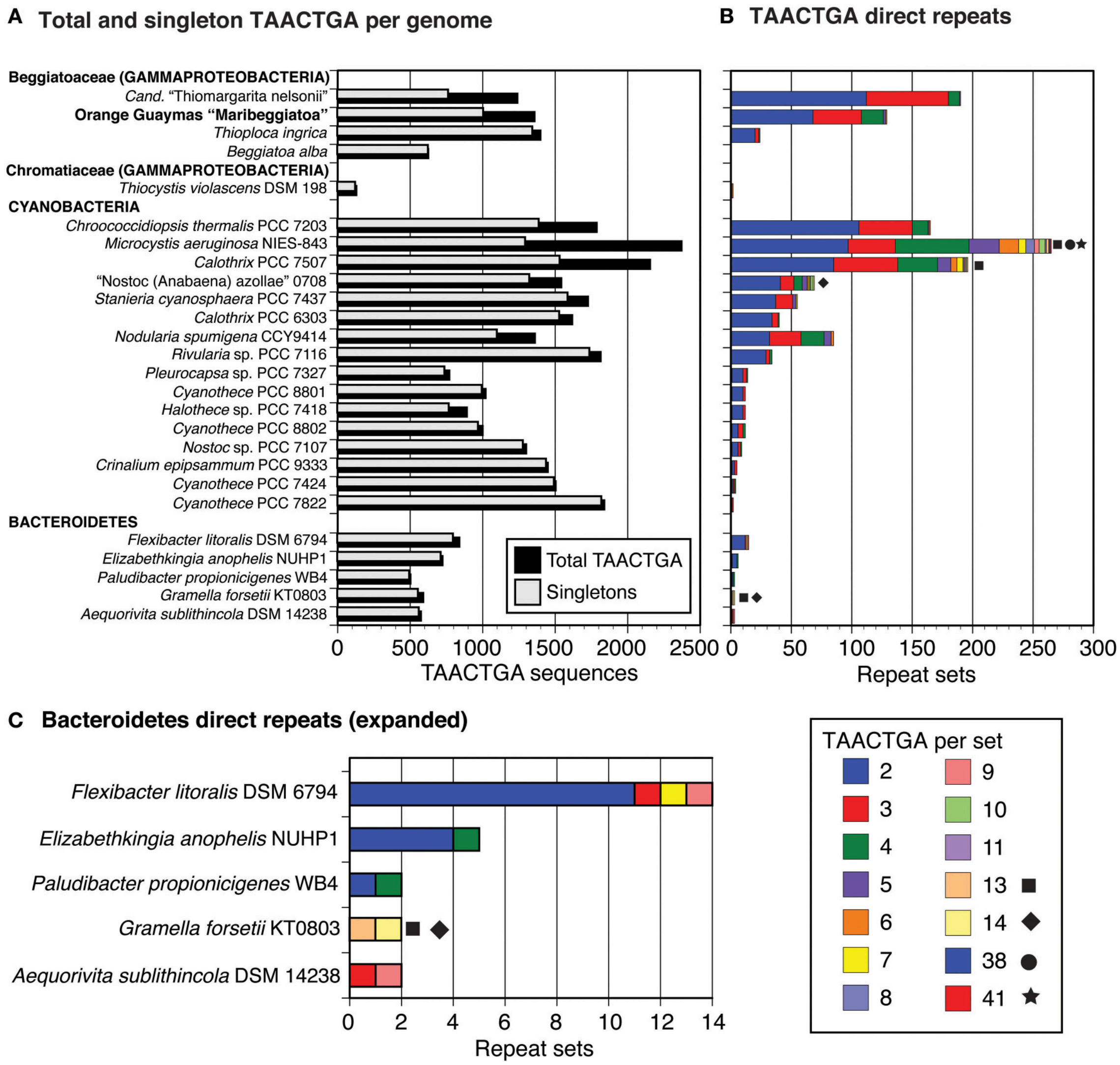


FIGURE 2 | Number and length of TAACTGA repeat sets in different species. The GenBank nr database was searched with seven direct repeats of the TAACTGA sequence, using the default "short query" settings. For each strain with a sequence identified by this search, the genome sequence was searched for all TAACTGA direct repeats (in both orientations), and these were classified by the number of repeats they contain. The strains were sorted in order of number of two-repeat copies within each phylogenetic group. *Beggiatoa alba* contains no sets of repeats, but was included to present a complete set of available *Beggiatoaceae* genomes. **(A)** Total and singleton repeats in all species. **(B)** Repeat sets, classified by length. Especially long sets are highlighted by symbols. **(C)** Expanded view for Bacteroidetes repeat sets.

genomes, TAACTGA repeats were identified in one other sulfur-oxidizing Gammaproteobacterium (*Thiocystis violascens* DSM 198), 15 Cyanobacteria, and 5 Bacteroidetes. This distribution is similar to that previously noted for the *fdxN* element excision-controlling factor proteins XisH and XisI (MacGregor et al., 2013c). An updated (May 2015) database search found that at least one of these was annotated in all cyanobacterial genomes with TAACTGA repeats except *Stanieria cyanosphaera* PCC 7437, but not in the Bacteroidetes represented (although they are found in some other genera in this group) and not in *T. ingricans* or *T. violascens* (Supplemental Table 7). The hypothetical protein BOGUAY_0693, which has 29 close matches in the BOGUAY genome, has matches in some but not all of the same cyanobacteria, the other *Beggiatoaceae*, and *Flexibacter litoralis*, but not in the remaining Bacteroidetes or *T. violascens* (Supplemental Table 7). Whether or not a common transfer mechanism is involved, this is consistent with a history of genetic exchange among some Cyanobacteria and *Beggiatoaceae*.

As in the *Beggiatoaceae*, there is no necessary correlation between number of singletons and number of repeats (**Figure 2**, Supplemental Table 2); for example, *Cyanothece* PCC 7424 has more singleton and nearly as many total copies as "Nostoc azollae" 0708, but 3 vs. 69 sets of repeats. There are no obvious morphologies, metabolic types, or habitats common to all the species found: for example, *Microcystis aeruginosa* NIES-843

TABLE 3 | TAACTGA-like sequences in the BOGUAY genome.

DNA sequence (forward)	Total and direct-repeat occurrences in BOGUAY genome						Predicted RNA minimum free energy structure for six direct repeats				Amino acid repeat unit	
	Total copies	Repeats in set					Forward		Reverse complement		Forward	Reverse complement
		2	3	4	5	6	Type	kcal mol^{-1}	Type	kcal mol^{-1}		
TAACTGA AND SINGLE-BASE MUTATIONS												
TAA***T***TGA	1908	6					Stem-loop	−4.3	One pair	3.5	LIIDN–	SIINYQL
TAA***A***TGA	1416	1					One pair	3.6	One pair	3.5	MINDK–	SFIIYHL
TAACTGA	1357	68	40	19	1	1	One pair	3.9	One pair	3.9	LITDN–	SVISYQL
AAACTGA	1354	4					One pair	3.9	One pair	3.9	KLKTEN-	SVFSFQF
TAACT***C***A	1206						One pair	3.9	Stem-loop	−10.2	LITHNS-	VMSYEL-
TAACT***T***A	1018						Stem-loop	−6.1	Stem-loop	−9.7	LITYNL-	VISYKL-
TA***T***CTGA	943						Stem-loop	−14.6	Stem-loop	−13.4	YLISDI-	SDIRYQI
CAACTGA	829	3					One pair	3.1	Stem-loop	−6.3	QLTTDN-	SVVSCQL
T***T***ACTGA	810					1	Stem-loop	−4.4	One pair	3.9	LLITDY-	SVISNQ-
TAACT***A***A	786	4		1			One pair	3.9	Stem-loop	−4.3	LITNN–	LLVISY-
T***C***ACTGA	786	1					Stem-loop	−15.1	Stem-loop	−21.2	SLITDH-	SVISDQ-
T***G***ACTGA	778						One pair	3.9	One pair	3.9	LMTDD–	SVISHQS
TAAC***G***GA	742	1					Stem-loop	−4.6	One pair	3.4	RITDNG-	SVIRYPL
TA***C***CTGA	701						One pair	3.9	Stem-loop	−11.0	YLIPDT-	SGIRYQV
TAA***G***TGA	632						One pair	3.2	One pair	3.9	VISDK–	SLITYHL
TAACTG***T***	625	1					Stem-loop	−12.8	Stem-loop	−11.8	LLTVNC-	QLTVNS-
GAACTGA	606						Stem-loop	−4.1	Stem-loop	−2.2	ELRTEN-	SVLSSQF
TA***G***CTGA	598						Stem-loop	−18.4	Stem-loop	−10.4	LIADS–	SAISYQL
TAAC***A***GA	544						One pair	3.9	One pair	3.8	QITDNR-	SVICYLL
TAACTG***G***	476						Stem-loop	−11.8	Stem-loop	−14.4	LVTGNW-	PVTSYQL
TAAC***C***GA	470	1					One pair	3.4	Stem-loop	−12.8	PITDNR-	SVIGYRL
TAACTG***C***	351						Stem-loop	−4.0	Stem-loop	−11.8	LLTANC-	QLAVSS-
SHUFFLED TAACTGA (SELECTION)												
ATATCAG	1030						Stem-loop	−13.4	Stem-loop	−14.0	ISDIRYQ	YLISDI-
ATAATCG	867						Stem-loop	−14.0	Stem-loop	−15.4	SIIDNR-	RLSIIDY
CTAAGTA	322						Stem-loop	−13.8	Stem-loop	−14.5	VLSTKY-	YLVLST-
TCGAATA	319						Stem-loop	−9.4	Stem-loop	−11.0	SNIEYRI	YSIFDIR
TAACTAG	160						Stem-loop	−16.6	Stem-loop	−15.6	LVTSN–	LLVTSY-

*DNA sequences are arranged by number of occurrences. The TAACTGA sequence itself is outlined. Single-base differences to it are in bold italics. For each DNA sequence, an RNA structure was predicted for six direct repeats. Amino acid sequences were predicted for 7 direct repeats, but only a single repeat unit is shown. Shaded boxes indicate amino acid sequences containing stop codons. RNA structure predictions are the first results from a minimum free energy calculation using the default settings of the MaxExpect algorithm from the RNAstructure Web Server [*http://rna.urmc.rochester.edu/RNAstructureWeb/*, (Reuter and Mathews, 2010)]. Translations were done via the ExPASy portal of the Swiss Institute of Bioinformatics (Artimo et al., 2012).*

(NC7) is a colonial freshwater cyanobacterium isolated in Japan (Otsuka et al., 2000); *Elizabethkingia anophelis* NUHP1 is a Gram negative rod from a mosquito midgut collected in The Gambia (Kämpfer et al., 2011); and *Aequorivita sublithincola* DSM 14238 is an endolithic Gram negative bacterium found as rods or filaments, isolated from within a quartz rock in Antarctica (Bowman and Nichols, 2002). This complicates the argument just made for horizontal transfer; characterization of other heptamer repeats and additional genomic sequencing may clarify this issue.

Cyanobacteria

Among the Cyanobacteria, the sequenced genomes of the freshwater, bloom-forming *M. aeruginosa*, particularly strains NIES-843 (Kaneko et al., 2007) and PCC 7806 (Frangeul et al., 2008), have high proportions of repeated sequences. This has been proposed to be part of an evolutionary strategy relying on genome plasticity, with a comparatively high number of horizontally acquired genes and repeated genes and sequences (Humbert et al., 2013). These include a range of repeating heptamers, with TAACTGA repeats often mixed with others. A complete analysis was not carried out here, but a small random sample of the 265 sets of *M. aeruginosa* NIES-843 TAACTGA repeats suggests that they may play more or different roles than in BOGUAY. Of 24 sets of repeats mapped in detail (Supplemental Table 3), 22 were intergenic and two in "reverse" orientation within ORFs encoding small hypothetical proteins. Of the intergenic sets, just six were in "forward" orientation relative to a downstream start codon, and at a range of distances (from 1 to 214 nt). Eight sets were in reverse origin relative to a start codon and eight were between stop codons. All of the latter are in the same orientation on the chromosome; it would be interesting to see whether this pattern holds throughout the genome. If this is a representative sample, it is a clear contrast to the BOGUAY genome, where most sets of repeats are intergenic and in "forward" orientation to a relatively nearby start codon. The chromosomal arrangement is not known because the genome is not closed.

Repeat distributions in four *Cyanothece* strains with relatively few TAACTGA copies were also compared (Supplemental Table 4). *Cyanothece* PCC 8801 and 8802 are very similar, with nine sets of repeats in matching positions in terms of flanking ORFs and only small intergenic sequence differences, mostly indels in 7 nt increments. Seven of these repeat sets are just upstream of a start codon, one just upstream of a putative Shine-Dalgarno (SD) sequence, and one in reverse orientation near the upstream ORF. PCC 8802 has an additional intergenic set relatively far upstream from a start codon; each strain has an intergenic plasmid-borne set, but between different ORFs; and PCC 8801 has one set in reverse orientation internal to an ORF. In PCC 7424, there are only three sets of repeats, none in positions matching the other two strains. All are intergenic and in "forward" orientation, at varying distances from the nearest start codon. The closest relatives of the flanking ORFs are all from strain PCC 7822, including those flanking its only set of repeats. Overall, whether TAACTGA and related repeats derive from a common cyanobacterial ancestor or are transmitted by some mobile element, they appear to have followed strain-specific paths here as in other lineages.

Bacteroidetes

The distribution of TAACTGA repeats in the Bacteroidetes (**Figure 2C**) suggests they could also have more than one role in this group. *F. litoralis* DSM 6794 is similar to BOGUAY, on a more limited scale. Of 14 repeat sets, 12 are intergenic and in the "forward" orientation relative to a start codon between 1 and 43 nt downstream (Supplemental Table 5). One set of seven repeats is located immediately downstream of a stop codon, in reverse orientation, and a set of two is located within a putative PurC (SAICAR synthase) gene, near its end. In *Paludibacter propionicigenes* WB4 there are just two sets of direct repeats, one close to a start codon and the other toward the center of a long intergenic region (Supplemental Table 5).

The remaining three Bacteroidetes strains have different distributions. *Gramella forsetii* KT0803 and *A. sublithincola* DSM 14238 have only two sets of TAACTGA direct repeats each, but three of these are quite long (**Figure 2C**). All are intergenic and in "forward" orientation relative to the downstream ORF, but only one is immediately upstream of a start codon, and the intergenic regions contain other heptamer direct repeats as well (Supplemental Figure 3). For both *A. sublithincola* sets and one of the *G. forsetii* ones, the closest matches to the upstream and downstream ORFs are found in the same close relative (*A. capsosiphonis* DSM 23843 and *G. echinicola* DSM 19838, respectively), which have shorter intergenic regions without obvious sets of repeats, although the immediate gene neighborhoods appear the same (Supplemental Figure 3A). In the second *G. forsetii* example (Supplemental Figure 3B), at least the downstream ORF may have been acquired by horizontal transfer. The closest match to the upstream ORF is from the Bacteroidetes strain *Gillisia limnaea* DSM 15749, which has a similar local gene neighborhood, except that instead of a homolog of the downstream ORF there is a short hypothetical protein encoded on the opposite strand. No sets of direct repeats are evident in this intergenic region. Downstream, the closest match to the *G. forsetii* ORF is from *Bacillus azotoformans* LMG9581, which has no other apparent local similarity to *G. forsetii*. A phylogenetic reconstruction for this ORF and a comparison of intergenic regions in other *Gramella, Gillisia*, and *Bacillus* strains would be needed to propose a history for this small region, but the pattern so far suggests a role in gene rearrangement for these intergenic repeats.

E. anophelis NUHP1 has sets of TAACTGA repeats between only three pairs of ORFs, which are not very long (four sets of two, one set of four), but in two cases they are part of nearly identical intergenic regions containing larger assemblages of heptamer repeats and flanked by ORFs encoding putative proteins with stretches of high identity (Supplemental Table 6). Comparisons with closest neighbors (all *Elizabethkingia* strains) were difficult because the contigs identified often end partway through the repeat region, likely because of assembly difficulties. The third repeat set is a single pair, found toward the center of a relatively long (295 bp) intergenic region with no other obvious repeats.

Canonical Ribosome Binding Sites are Rare in Repeat-Containing BOGUAY Intergenic Regions

The TAACTGA repeats in the BOGUAY genome are generally positioned close to start codons (**Figure 1**), overlapping the expected ribosome binding site. The Shine-Dalgarno (SD) sequence predicted from the 16S rRNA genes of BOGUAY and other sequenced *Beggiatoaceae* is the same as that of *E. coli* (AGGAGGU). With only one G residue per heptamer in either orientation, the repeat sequence itself has little SD character, so most of the ORFs downstream of them have no obvious ribosome binding site. For an overview of the genome, any four consecutive bases from the AGGAGGU sequence ending 4–13 nt upstream of a start codon was considered an SD, recognizing that this may lead to over- or undercounting. The number of such sequences was estimated at 1346 (Supplemental Table 8), accounting for 25% of the 5272 predicted protein-coding genes. This is toward the low side for bacteria overall, but by no means unmatched (Ma et al., 2002). Of intergenic regions with repeats, just 15 (∼10%) also include SD sequences (**Table 4**), with the repeats ending between 2 and 25 bp upstream of them.

Functional Classification of BOGUAY ORFs Downstream of TAACTGA Repeats

The COG (Clusters of Orthologous Groups; Tatusov et al., 1997) classifications of ORFs with and without upstream repeats were compared (**Table 5**). Categories F, D, Q, E, and J were particularly overrepresented among those with repeats, while only category A was as strongly underrepresented. Note however that 63% of all ORFs and 29% of those with repeats have not been classified at all, and some 8% more of each are in categories R (general function prediction only) and S (function unknown). No clear picture of a possible transcriptional or translational regulatory role for TAACTGA repeats is apparent at this level, particularly since it is not known whether regulation is positive or negative. Several concentrations of repeat sequences will be considered in more detail below.

TAACTGA Repeats within Open Reading Frames

While most of the TAACTGA repeats in the BOGUAY genome are intergenic, suggesting a regulatory role, there are exceptions. The coding regions of 25 putative BOGUAY proteins contain or overlap 24 sets of direct repeats, with one set found in overlapping ORFs (BOGUAY_3048 and _3047). In 13 of these, between one partial and two complete repeats overlap the stop codon of an upstream gene in "forward" orientation relative to a downstream gene (**Table 2A**); as mentioned above, forward repeats generate stop codons in all three reading frames, so these are necessarily at the end of ORFs. In only two of these was a recognizable SD sequence found between the end of the repeats and the start codon of the downstream ORF. In three more ORFs, sets of repeats were found within or overlapping one end or the other of the putative coding sequence, but not directly upstream of another (**Table 2B**). One example was also found of an indirect repeat near the end of an ORF, with one base pair separating the two copies (**Table 2C**).

The 11 ORFs containing "reverse" repeats (**Table 2B**) have no apparent amino acid sequence similarity outside the repeat-encoded region (Supplemental Figure 1). Seven are short hypothetical or conserved-domain proteins with no assigned functions. One of these overlaps a putative glycosyl hydrolase (BOGUAY 01182_3048); the repeat-encoded amino acids in the latter are near the C-terminal end of the predicted protein, with little homology to otherwise close database relatives and outside the CDD-defined glycosyl hydrolase domain that includes most of the rest of the ORF (not shown). The repeat-encoded amino acids of two of the other ORFs with assigned functions are likewise outside regions of assigned function, either toward the very beginning (corrinoid ABC permease, BOGUAY 00106_0223) or very end (MurG, BOGUAY 00938_0721) of their respective amino acid sequences. The exception is BOGUAY 00100_0018, an ORF encoding a putative protein similar to an RNA polymerase beta prime subunit, discussed below.

If the repeats were or are mobile within the genome, their insertion within coding sequences seems to have been successful primarily at the periphery of at least the primary structure of proteins. For repeats in "forward" orientation, this is a necessary consequence of their sequence, which encodes stop codons in all three reading frames. "Reverse" repeats could in principle occur anywhere, but most insertions are likely deleterious. Those at the end of proteins, or perhaps splitting a protein into two new functional proteins, are probably more likely to become fixed.

Direct TAACTGA repeats are also found within hypothetical proteins in *Beggiatoa* sp. PS and some cyanobacteria, particularly *M. aeruginosa* strains. A BLASTP search of the GenBank protein database with 7, 14, or 21 LSVISYQ repeats yielded mostly predicted amino acid sequences annotated as hypothetical proteins. The shorter variant yielded the most perfect matches (Supplemental Table 9). The phylogenetic distribution of at least the top hits was quite restricted: 61 cyanobacterial sequences, of which 25 were from *M. aeruginosa* and 25 from *Moorea producens*; 17 Gammaproteobacterial sequences, of which 12 were from *Pseudoalteromonas* spp. and 2 from *Beggiatoaceae*; 9 from the Betaproteobacterium *Burkholderia pseudomallei*; 6 from Alphaproteobacteria, of which 4 were from *Ehrlichia ruminantium*; and one reportedly from a bird. Interestingly, one of these was annotated as an FdxN element excision controlling factor protein-like protein (BAG05441.1 from *M. aeruginosa* NIES-843). However, given the large number of these in the database, and the fact that it has no BLASTP matches from this group, this is suspected to be a misannotation. Similarly, the *B. pseudomallei* predicted protein (KGC53376) described as a putative 60S ribosomal protein L19 does not seem to actually belong to this group.

TAACTGA Repeats in Putative Ribosomal Protein Operons

One COG category overrepresented among BOGUAY ORFs preceded by repeats is J (Translation) with 13 examples, including four upstream of putative genes for ribosomal proteins (S1, L3, S4, and S21; **Figure 3**) and five others within putative

TABLE 4 | BOGUAY ORFs preceded by both Shine-Dalgarno sequences and TAACTGA repeats.

Upstream ORF	Repeat region sequence	Downstream ORF		Downstream COG category
hypothetical protein	(>100 bp to next ORF) TTTGTGATTAGTGATAAAGCAACCTTATTTTTAACTGATAACTGACAAAGGAACAATTTATATG	00031_3793	pfam14072 (DndB - DNA sulfur modification-associated)	none
conserved domain protein, ABC transporter type 1, transmembrane domain	(>100 bp to next ORF) TCGTTTTTCAGATTGTTGGTTTTAGACTCTGATAACTGATAACTGAACTAAGGAGAATTAAATTATTATG	00848_4300	V-type H(+)-translocating pyrophosphatase	C
beta propeller domain	(>100 bp to next ORF) TTTATGTATACCTAATTAAAACAACACTGATAACTGATAACTGAACAATAAGGTAAAATTTCCCAATATG	01124_0963	succinate dehydrogenase flavoprotein subunit	C
Inner membrane protein CreD	CATGTTAAATATTAAGTGAA~66nt~ACTGATAATTAATAACTGATAACTGATAAGGAGATTACACCTTCAATG	01192_0212	ribosomal protein S21	J
hypothetical protein	(>100 bp to next ORF) TATAAGGTAACTCAAATTTGACTGACAAATATTTTAACTGATAACTGATAATTCAAAAGGAATATAGGTG	00666_1840	amino acid carrier protein	E
phosphoglycerate kinase family protein	TAATTTAAAGATG~70nt~AAAGTTGACTGATAACTGATAACTGATAACTTAATAGGAAAATTTTATG	00163_1000	fructose-bisphosphate aldolase, class II, Calvin cycle subtype	G
3-oxoacyl-(acyl-carrier-protein) reductase	(>100 bp to next ORF) TTATTATGAATGAATAACTGATAACTGACAACTGATAAATGATAATTGAAAAGAGGACTAATTTGTTATG	00396_2179	acyl carrier protein	I, Q
regulatory protein, FmdB family	ATGCCTATT~273nt~AGTTCAAGCACTGATAACTGATAACTGAATAAGGAAAATTATG	00106_0256	aspartyl-tRNA synthetase	J
sulfite reductase, dissimilatory-type alpha subunit	TGAATTAACTGATAACTAATAACTAATAACTGATAACTGAATTAGGAGGTTATTTATG	01191_1510	sulfite reductase, dissimilatory-type beta subunit	C
response regulator receiver domain protein	TAATTATTGGCATTAATTAACTGATAACTGACAACACTTAACTGATAACTTAAAAGGATATGATG	00362_1734	response regulator receiver domain protein	none
septum site-determining protein MinC	TGATTAATGGTTAATAGTTAACTGATAACTGATAACTTAAAACAGAATAACTGATACAGGAAAGTATTTTG	00127_3142	septum site-determining protein MinD	D
glycogen~starch synthases, ADP-glucose type	TAGTTAGTGTTCAGTTATCAGTTAACTGATAACCGATAACTGAACACTGATAACAGAGGTTATTTTATG	00241_1023	glycosyl hydrolase, family 57	none
hypothetical protein	TAGTCATCAGTCTTGTTAAATGACACTGATAACTGATAACTGATAACTGATAAAGGAAAACATACGTATG	01295_4360	hypothetical protein	none
acetyl-CoA carboxylase, biotin carboxyl carrier protein	TAATAATTTATTAGTTATCAGTTATCAGTCAAAAAAACCTAACTGATAACTGATAACTGATATTGAGGTTAAAGTACATG	00861_2920	acetyl-CoA carboxylase, biotin carboxylase subunit	I
hypothetical protein	ATGCCAGAT~111nt~AAAGTCATGATAACTGATAACTGATAACTGATACAGGAGTACACAAAAATG	00614_2860	roadblock/LC7 domain protein	R

⟶ TAACTGA repeat
→ start codon
—● stop codon
□ Shine-Dalgarno sequence

Shine-Dalgarno sequences are defined, here and throughout, as having 4 or more consecutive matches to the consensus sequence AGGAGGT. Entries are organized to highlight similar arrangements of SDs and repeats.

r-protein operons (*pnp*, *fusA*) or nearby (COG2976, *pheS*, BOGUAY_0218). Only one of these (S21, **Figure 3H**) also has a ribosome-binding site by the criteria used above. As described previously (MacGregor et al., 2013a), BOGUAY ribosomal protein genes are organized similarly to those in *E. coli* (see e.g., Fu et al., 2013 for an illustration) and many other bacteria. Where studied, these are transcribed as multigene operons, with translation generally regulated by a negative feedback loop involving one of the proteins encoded by the operon. Short noncoding RNAs transcribed within these operons may also play a role (Khayrullina et al., 2012). There is no experimental evidence regarding transcription of any BOGUAY genes, but it is worth noting that all TAACTGA repeats within BOGUAY r-protein operons are internal to the standard operons, suggesting a role in translational rather than transcriptional regulation. Insertion of a mobile element at such an internal site might also be favorable compared to insertion in a promoter region for these highly expressed operons, although given the essential role of ribosomes any insertion at all seems potentially disruptive.

This distribution has some overlap with the other *Beggiatoaceae* "T. nelsonii" and *T. ingrica*. In particular, all three species have TAACTGA repeats upstream of their putative S1 subunit genes (Supplemental Figure 2): BOGUAY has 5, beginning 1 nt upstream; "T. nelsonii" has three copies but with gaps between them, also beginning 1 nt upstream; and *T. ingrica* has 2 copies, beginning 21 nt upstream. The sequence of this gap is nearly identical (18 of 21 nt) to the *B. alba* sequence over this stretch; *B. alba* of course has no repeats. This shared sequence does not include a ribosome-binding site, by the definition used here, but does have an AGGG and an AGGGG run.

Three of the four putative BOGUAY r-protein genes preceded by repeats (S1, S4, and L3) are also among those with proposed extraribosomal functions in *E. coli* (reviewed in Aseev and Boni, 2011). During translation, S1 is involved in ribosome docking and in unfolding of structured mRNAs (Duval et al., 2013), interacting with AT-rich regions upstream of the SD sequence (if there is one), as well as with downstream sequences (Tzareva et al., 1994). In *E. coli*, S1 is required for translation of all mRNAs with leader sequences (reviewed in Hajnsdorf and Boni, 2012), while leaderless mRNAs can be translated by ribosomes lacking it (reviewed in Byrgazov et al., 2013). Like several other ribosomal proteins, it inhibits translation of its own operon: at least *in vitro*, free S1 competes with ribosome-bound S1 for mRNA binding upstream of the start codon (Boni et al., 2001). Again *in vitro*,

TABLE 5 | COG classification of BOGUAY ORFs downstream of TAACTGA repeats compared to whole genome.

COG category		Number of ORFs downstream of repeats	% of ORFs downstream of repeats	% of all ORFs	Fold difference (%repeats/%total)
OVERREPRESENTED DOWNSTREAM OF TAACTGA REPEATS					
F	Nucleotide metabolism and transport	6.5	4.55	0.82	5.6
D	Cell cycle control and mitosis	3	2.60	0.62	4.2
Q	Secondary metabolites	1.8	1.95	0.47	4.1
E	Amino acid metabolism and transport	12.5	9.09	2.36	3.9
J	Translation	13	8.44	2.39	3.5
H	Coenzyme metabolism	7.5	5.19	1.78	2.9
I	Lipid metabolism	2.8	1.95	0.80	2.4
O	Post-translational modification, protein turnover, chaperone functions	7	4.55	1.94	2.3
G	Carbohydrate metabolism and transport	4	2.60	1.16	2.2
M	Cell wall/membrane/envelope biogenesis	7	5.84	3.12	1.9
T	Signal transduction	7	4.55	2.48	1.8
C	Energy production and conversion	6.5	4.55	2.85	1.6
V	Defense mechanisms	1	0.65	0.45	1.4
K	Transcription	1.5	1.30	1.23	1.1
UNDERREPRESENTED					
U	Intracellular trafficking and secretion	2	1.30	1.49	0.9
L	Replication and repair	1.5	1.30	1.69	0.8
N	Cell motility	0.5	0.65	1.16	0.6
P	Inorganic ion transport and metabolism	1	0.65	1.83	0.4
A	RNA processing and modification	0	0.00	0.04	0.0
UNCATEGORIZED					
R	General functional prediction only	6.8	4.55	4.28	1.1
S	Function unknown	8	4.55	3.91	1.2
None assigned		46	29.22	63.16	0.5
TOTAL		147	100.00	100.00	

Fractional occurrences were used for ORFs assigned to more than one category.

it is reported to have a transcriptional role as well: *E. coli* S1 co-purifies with RNAP and stimulates transcriptional cycling (Sukhodolets et al., 2006).

The *E. coli* S4 ribosomal protein, in addition to negatively regulating translation of its own operon, is proposed to form part of transcriptional antitermination complexes that may also include L1, L3, and L4 (Torres et al., 2001), with S4 binding RNAP directly.

Candidate Repeat-Binding Proteins

The frequent position of the TAACTGA repeats upstream of and apparently replacing SD sequences, including five direct repeats directly upstream of the S1 gene (**Figure 3**), suggests that they might play a role in translation. Several categories of known translational regulatory proteins have properties that suggest them as candidates.

Ribosomal Protein S1

Interaction with the S1 subunit is one possibility. S1 has a relatively weak and reversible association with the ribosome, and is added last in assembly (Subramanian and Vanduin, 1977). In *E. coli* and many other Gram negative bacteria, it is composed of six linked oligonucleotide/oligosaccharide binding (OB)-fold domains; where studied, the four C-terminal domains are RNA-binding, while the two N-terminal domains make protein-protein contacts with ribosomal, and other proteins (reviewed in Hajnsdorf and Boni, 2012). The BOGUAY S1 protein is predicted to have a typical Gram negative S1 structure (not shown).

The *E. coli* S1 gene itself (*rpsA*) lacks a strong SD sequence and does not require one for expression (Boni et al., 2001). The upstream region forms three hairpins, which contribute to its translational efficiency (Boni et al., 2001; Skorski et al., 2006). Different secondary structures can be predicted for the intergenic region upstream of the BOGUAY S1 gene, depending how much of this and the coding sequence are included in the calculation (not shown), but they have no obvious similarity to those in *E. coli*. Without experimental evidence, or knowledge of the transcriptional start site, they cannot be assigned a function. One argument against a TAACTGA-binding role for S1 is the reported non-specificity of S1 RNA recognition, limited to a preference for AT-rich sequences (reviewed in Aseev and Boni, 2011). TAACTGA repeats are somewhat AT rich, but do not produce long polypyrimidine tracts.

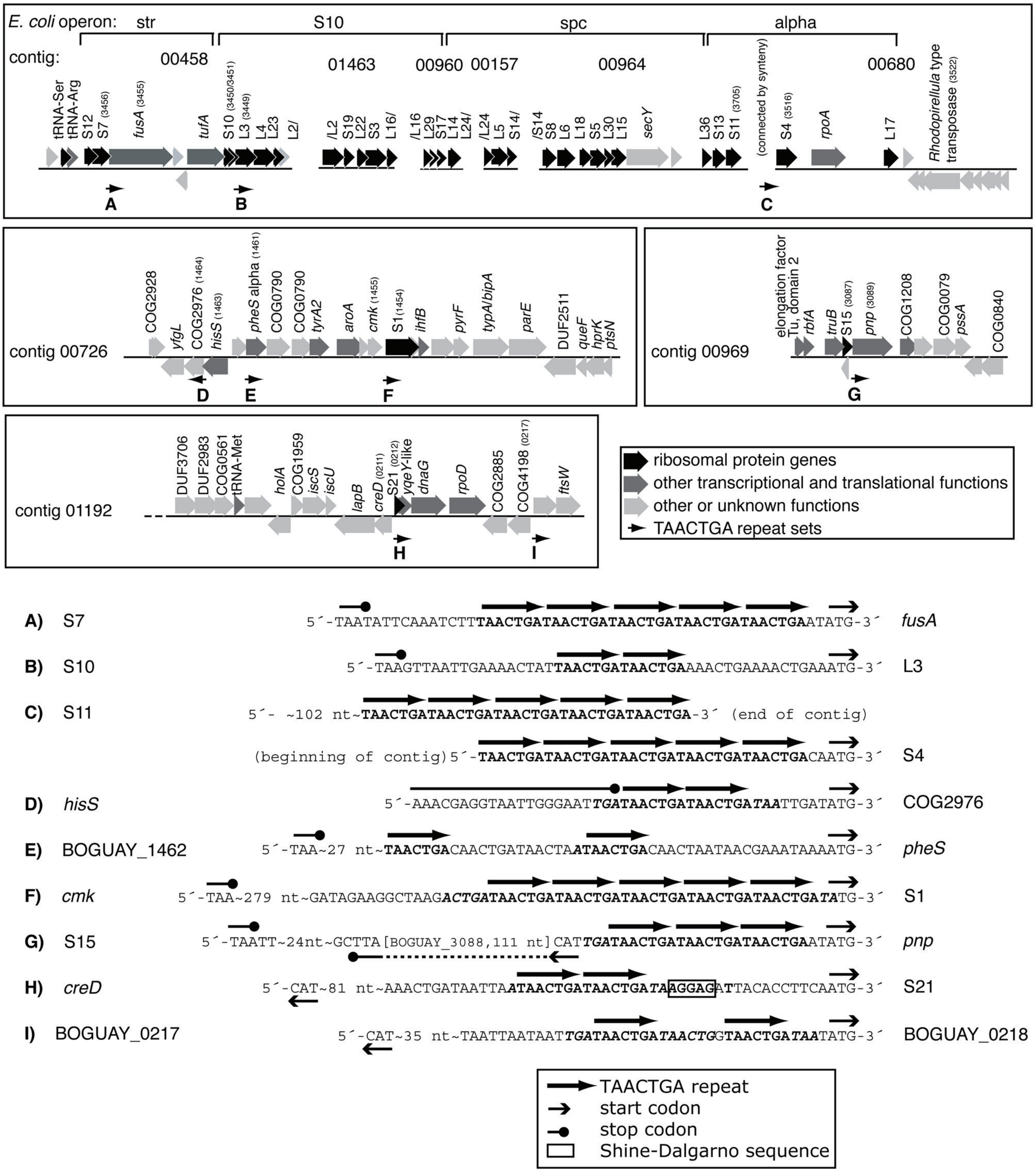

E. coli operon: str S10 spc alpha
contig: 00458 01463 00960 00157 00964 00680
tRNA-Ser
tRNA-Arg
S12
S7 (3456)
fusA (3455)
tufA
S10 (3450/3451)
L3 (3449)
L4
L23
L2/
/L2
S19
L22
S3
L16/
/L16
L29
S17
L14
L24/
/L24
L5
S14/
/S14
S8
L6
L18
S5
L30
L15
secY
L36
S13
S11 (3705)
(connected by synteny)
S4 (3516)
rpoA
L17
Rhodopirellula type transposase (3522)
A
B
C
contig 00726
COG2928
yfgL
COG2976 (1464)
hisS (1463)
pheS alpha (1461)
COG0790
COG0790
tyrA2
aroA
cmk (1455)
S1 (1454)
ihfB
pyrF
typA/bipA
parE
DUF2511
queF
hprK
ptsN
D
E
F
contig 00969
elongation factor Tu, domain 2
rbfA
truB
S15 (3087)
pnp (3089)
COG1208
COG0079
pssA
COG0840
G
contig 01192
DUF3706
DUF2983
COG0561
tRNA-Met
holA
COG1959
iscS
iscU
lapB
creD (0211)
S21 (0212)
yqeY-like
dnaG
rpoD
COG2885
COG4198 (0217)
ftsW
H
I
ribosomal protein genes
other transcriptional and translational functions
other or unknown functions
TAACTGA repeat sets
A) S7 5´-TAATATTCAAATCTTTAACTGATAACTGATAACTGATAACTGATAACTGAATATG-3´ fusA
B) S10 5´-TAAGTTAATTGAAAACTATTAACTGATAACTGAAAACTGAAAACTGAAATG-3´ L3
C) S11 5´- ~102 nt~TAACTGATAACTGATAACTGATAACTGATAACTGA-3´ (end of contig)
(beginning of contig)5´-TAACTGATAACTGATAACTGATAACTGATAACTGACAATG-3´ S4
D) hisS 5´-AAACGAGGTAATTGGGAATTGATAACTGATAACTGATAATTGATATG-3´ COG2976
E) BOGUAY_1462 5´-TAA~27 nt~TAACTGACAACTGATAACTAATAACTGACAACTAATAACGAAATAAAATG-3´ pheS
F) cmk 5´-TAA~279 nt~GATAGAAGGCTAAGACTGATAACTGATAACTGATAACTGATAACTGATAACTGATATG-3´ S1
G) S15 5´-TAATT~24nt~GCTTA[BOGUAY_3088,111 nt]CATTGATAACTGATAACTGATAACTGAATATG-3´ pnp
H) creD 5´-CAT~81 nt~AAACTGATAATTAATAACTGATAACTGATAAGGAGATTACACCTTCAATG-3´ S21
I) BOGUAY_0217 5´-CAT~35 nt~TAATTAATAATTGATAACTGATAACTGGTAACTGATAATATG-3´ BOGUAY_0218
TAACTGA repeat
start codon
stop codon
Shine-Dalgarno sequence

FIGURE 3 | TAACTGA repeats in and near putative BOGUAY ribosomal protein operons. Repeats are found upstream of putative genes for **(A)** *fusA* (elongation factor G); **(B)** ribosomal protein L3; **(C)** ribosomal protein S4; **(D)** a COG2976 protein; **(E)** *pheS* (phenylalanine-tRNA ligase, alpha subunit); **(F)** ribosomal protein S1; **(G)** *pnp* (polynucleotide phosphorylase); **(H)** ribosomal protein S1; and **(I)** ORF BOGUAY_0218.

Cold Shock Proteins

As a second possibility, the cold shock proteins (CSPs; since shown to include proteins with other roles) are OB-fold proteins with a single S1-like domain that can bind single-stranded RNA or DNA. Intriguingly, X-ray crystallography (Sachs et al., 2012) and microarray binding (Morgan et al., 2007) studies of *Bacillus subtilis* CspB have shown that it can bind heptamer direct repeats (reviewed in Horn et al., 2007), with one protein per heptamer, although only weak sequence specificity (e.g., stronger binding to TTCTTTT than TTTTTT) has been demonstrated. During cold shock, CSPs bind both non-specifically to general RNA and specifically to the 5′ untranslated region of selected mRNAs; this selection has been proposed to rely more on secondary structure than primary sequence (Giuliodori et al., 2004), but limited work has been done on this question. It seems conceivable that some Csp-like proteins might bind in a sequence-specific manner.

There are several putative proteins with cold shock domains in the BOGUAY genome (Supplemental Table 10). Two include just a single cold shock domain, and are annotated as CspA and CspE; two have a downstream Excalibur calcium-binding domain; and one has a downstream DUF1264 domain. According to a CDD (Marchler-Bauer et al., 2011) search, the CSP-Excalibur architecture is found in 301 other proteins in the GenBank nr protein database, of which 298 are Proteobacterial; 256 of these are Gammaproteobacterial. Similarly, the CDS-DUF1264 architecture is found in 801 nr sequences, of which 766 are Proteobacterial and 614 Gammaproteobacterial. Cyanobacteria were the next largest group, but with just 13 examples. It is not uncommon for a single Gammaproteobacterial genome to encode more than one CSP domain protein (not shown).

PSORTb 3.0 (Yu et al., 2010) predicts the putative BOGUAY CspA and CspE to be cytoplasmic, by similarity to known proteins (Supplemental Table 10). The CSP-DUF1264 protein is predicted to possess four internal helices and be a cytoplasmic membrane protein, making it an unlikely translational regulatory protein. No prediction could be made for the two CSP-Excalibur putative proteins (while the name stands for "extracellular calcium-binding region," this is due to the proteins the domain was originally identified in Rigden et al. (2003); other proteins containing it may or may not be extracellular). At least two (the putative CspA and CspE) and possibly four of these CSP-like proteins are therefore candidates for TAACTGA binding. While so-called cold-shock domain proteins need not respond to temperature, temperature is likely an important environmental clue in the Guaymas Basin microbial mats, signaling the intensity of the hydrothermal flow that supplies sulfide to the sulfide-oxidizing BOGUAY strain and its relatives.

CsrA-Like Proteins

As a third possibility, CsrA (*E. coli* carbon storage regulatory protein) and related proteins bind to single-stranded RNA, in some cases inhibiting translation by competing with ribosomes for binding to Shine-Dalgarno sequences. They play a role in processes including motility, biofilm formation, quorum sensing, and virulence in a wide range of bacteria (reviewed in Romeo et al., 2013; Van Assche et al., 2015). The BOGUAY genome contains a *csrA* candidate (BOGUAY 00153_2343) with a strong possible SD site (AGGAG, 7 nt from the start codon), consistent with the autoregulation often found for these genes (Romeo et al., 2013). However, the known RNA binding sites for CsrA proteins, whether on target RNAs or on regulatory small RNAs, are centered on SD-like GGA motifs with more than 7 nt spacing (reviewed in Duss et al., 2014). These are not found in TAACTGA repeats in either orientation, making these unlikely to be recognized by a canonical CsrA.

Possible Secondary or Repurposed RNA Polymerase Beta Prime Subunits in BOGUAY and *Thioploca Ingrica*

The BOGUAY genome encodes two putative RNAP β′ subunits (MacGregor et al., 2013c), an unusual feature also found in the recently sequenced *T. ingrica* genome, but not in *B. alba*. The partial "Isobeggiatoa" PS sequence includes only one. The *Cand.* "Thiomargarita nelsonii" genome is annotated with two (OT06_22820, OT06_51635), each on short contigs with no surrounding ORFs, but their sequences are identical except that OT06_51635 is missing 214 C-terminal amino acids where its contig ends; if this is in fact a duplication, it would seem to be a fairly recent one. Of 100 top BLASTP hits to BOGUAY 00100_0018, the alternate beta prime gene, only one intraspecific pair of beta prime genes was found. *Nitrosococcus watsonii* C-113 has an apparent tandem duplication of its beta (Nwat_2177, Nwat_2165) and beta prime (Nwat_2176, Nwat_2164) genes along with surrounding ribosomal protein and other translation-related genes. The two putative beta prime subunit genes are 100% identical at the nucleotide level; again, if this is a duplication, it appears recent.

BOGUAY and *T. ingrica*, by contrast, each have two different beta prime-like genes (**Figure 4**). One of these (BOGUAY_3638, THII_2732) appears to include all the expected catalytic and subunit-interaction sites of a bacterial beta prime subunit, and is very similar to the single "Isobeggiatoa" PS sequence (BGP_5131). The other (BOGUAY_0018, THII_0330) has the N-terminal subunit interaction and DNA-binding sites expected for an RNA polymerase beta prime subunit, but the *T. ingrica* sequence has several active-site substitutions, and neither has a complete catalytic site D–D–D sequence. The N-terminal domains resemble other beta prime sequences, but the C-terminal domains differ from each other and their genomic partners in the variable S13 region. The BOGUAY ORF has three TAACTGA units in forward orientation just upstream of its start codon, interleaved with two TTACTGA sequences, and three in reverse orientation within the ORF, one direct TCAGTTA repeat and a third unit separated by the related 7-mer TCAATTA (**Figure 6A**, below). These encode the amino acid sequence LSVINYQLS and fall within a variable region of the predicted beta prime N-terminal domain (**Figure 4**), which in the *E. coli* crystal structure is in a surface loop near the alpha II subunit (Murakami, 2013).

The genomic context of the BOGUAY beta prime is unusual. Of the four *Beggiatoaceae* beta prime genes for which some surrounding sequence is available (**Figure 5A**), all but BOGUAY have beta and beta prime genes immediately adjacent, as do many if not most other bacteria (Dandekar et al., 1998). Upstream of the putative beta subunit gene, BOGUAY, *T. ingrica*, and *B. alba* each have a NusG and four ribosomal protein genes; the "Isobeggiatoa" PS contig does not extend upstream. Downstream, the BOGUAY beta and beta prime subunit genes have apparently become separated, being internal to separate contigs. Comparing the beta/beta prime intergenic regions in the other three species,

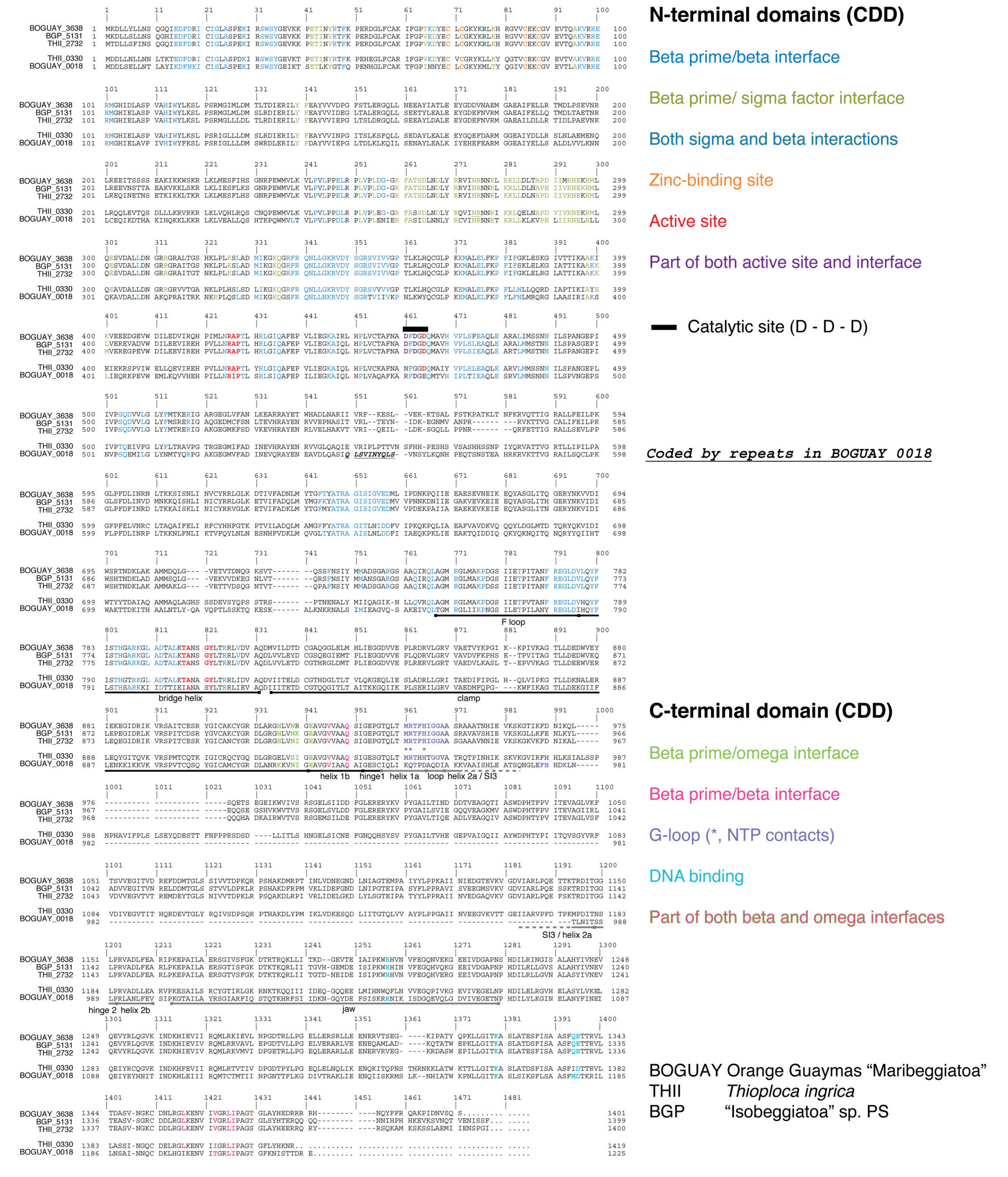

N-terminal domains (CDD)
Beta prime/beta interface
Beta prime/ sigma factor interface
Both sigma and beta interactions
Zinc-binding site
Active site
Part of both active site and interface
Catalytic site (D - D - D)
Coded by repeats in BOGUAY 0018
C-terminal domain (CDD)
Beta prime/omega interface
Beta prime/beta interface
G-loop (*, NTP contacts)
DNA binding
Part of both beta and omega interfaces
BOGUAY Orange Guaymas "Maribeggiatoa"
THII Thioploca ingrica
BGP "Isobeggiatoa" sp. PS
BOGUAY_3638
BGP_5131
THII_2732
THII_0330
BOGUAY_0018
F loop
bridge helix
clamp
helix 1b
hinge1
helix 1a
loop
helix 2a / SI3
SI3 / helix 2a
hinge 2
helix 2b
jaw

FIGURE 4 | Alignment of RNA polymerase beta prime and beta-prime like sequences from the BOGUAY, "Isobeggiatoa" PS, and *Thioploca ingrica* genomes. Sequences were aligned in MEGA5.2.2 (Tamura et al., 2011) using Muscle (Edgar, 2004). Trigger loop and SI3 annotation are after Windgassen et al. (2014), F loop and bridge loop annotation after Miropolskaya et al. (2014), jaw annotation after Opalka et al. (2010), and clamp annotation after Davis et al. (2007). Other putative domains were identified in CDD (Conserved Domain Database; Marchler-Bauer et al., 2011). Active-site and G-loop regions are boxed, and details of these shown to the right of the complete alignment.

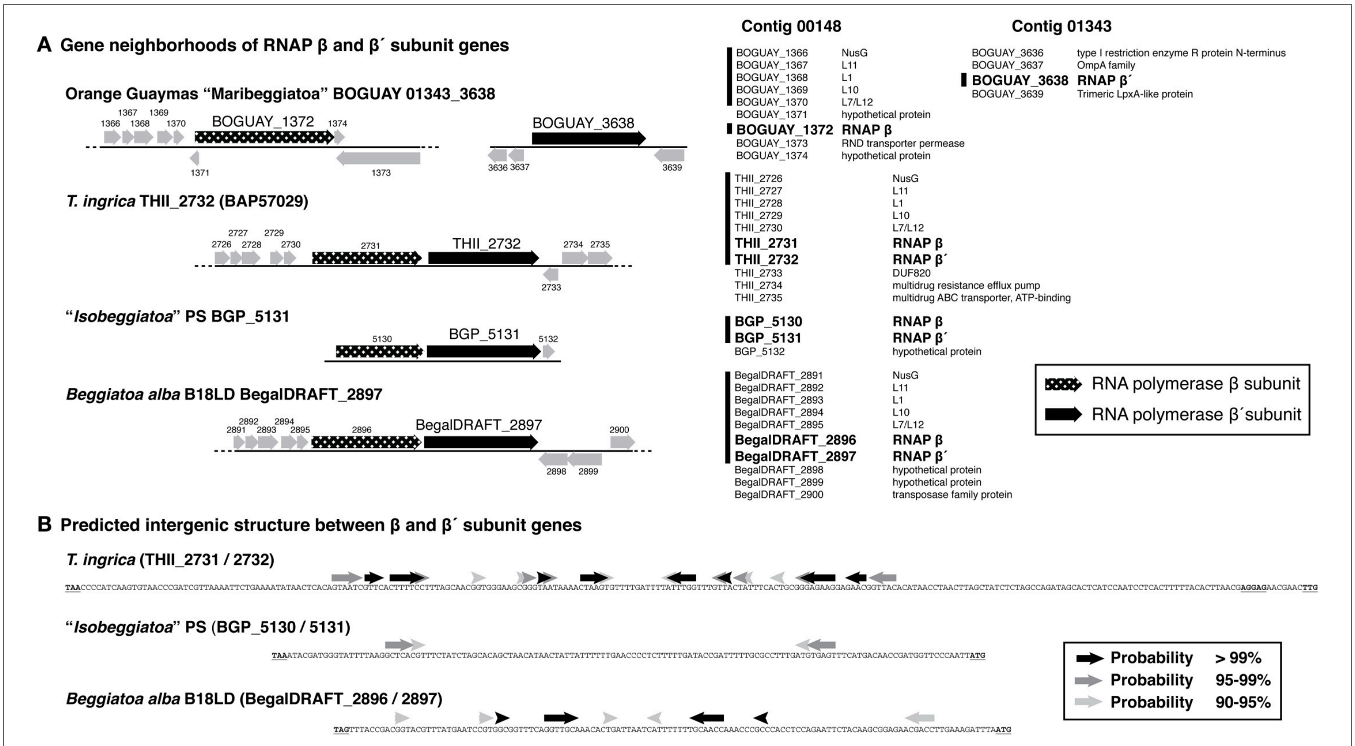
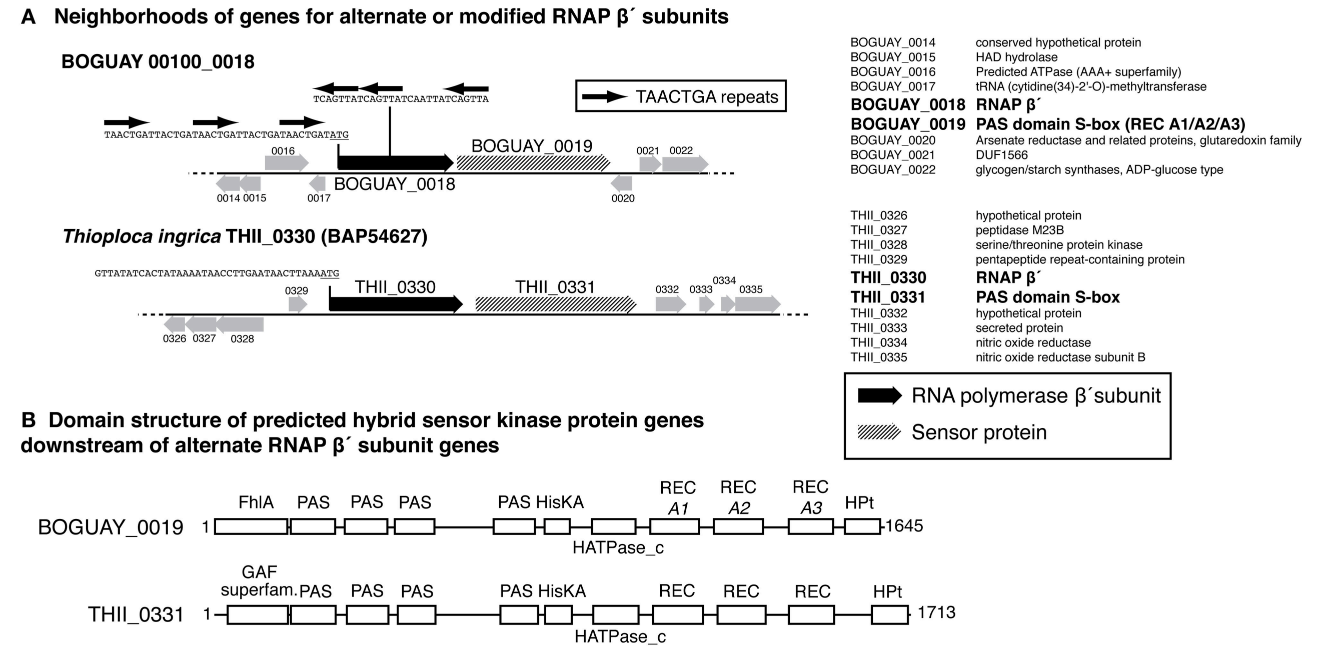

A Gene neighborhoods of RNAP β and β´ subunit genes
Orange Guaymas "Maribeggiatoa" BOGUAY 01343_3638
1366 1367 1368 1369 1370 BOGUAY_1372 1374
1371 1373
BOGUAY_3638
3636 3637 3639
T. ingrica THII_2732 (BAP57029)
2726 2727 2728 2729 2730 2731 THII_2732 2734 2735
2733
"Isobeggiatoa" PS BGP_5131
5130 BGP_5131 5132
Beggiatoa alba B18LD BegalDRAFT_2897
2891 2892 2893 2894 2895 2896 BegalDRAFT_2897 2900
2898 2899
Contig 00148
BOGUAY_1366 NusG
BOGUAY_1367 L11
BOGUAY_1368 L1
BOGUAY_1369 L10
BOGUAY_1370 L7/L12
BOGUAY_1371 hypothetical protein
BOGUAY_1372 RNAP β
BOGUAY_1373 RND transporter permease
BOGUAY_1374 hypothetical protein
Contig 01343
BOGUAY_3636 type I restriction enzyme R protein N-terminus
BOGUAY_3637 OmpA family
BOGUAY_3638 RNAP β´
BOGUAY_3639 Trimeric LpxA-like protein
THII_2726 NusG
THII_2727 L11
THII_2728 L1
THII_2729 L10
THII_2730 L7/L12
THII_2731 RNAP β
THII_2732 RNAP β´
THII_2733 DUF820
THII_2734 multidrug resistance efflux pump
THII_2735 multidrug ABC transporter, ATP-binding
BGP_5130 RNAP β
BGP_5131 RNAP β´
BGP_5132 hypothetical protein
BegalDRAFT_2891 NusG
BegalDRAFT_2892 L11
BegalDRAFT_2893 L1
BegalDRAFT_2894 L10
BegalDRAFT_2895 L7/L12
BegalDRAFT_2896 RNAP β
BegalDRAFT_2897 RNAP β´
BegalDRAFT_2898 hypothetical protein
BegalDRAFT_2899 hypothetical protein
BegalDRAFT_2900 transposase family protein
RNA polymerase β subunit
RNA polymerase β´subunit
B Predicted intergenic structure between β and β´ subunit genes
T. ingrica (THII_2731 / 2732)
TAACCCCATCAAGTGTAACCCGATCGTTAAAATTCTGAAAATATAACTCACAGTAATCGTTCACTTTTCCTTTAGCAACGGTGGGAAGCGGGTAATAAAACTAAGTGTTTTGATTTTATTTGGTTTGTTACTATTTCACTGCGGGAGAAGGAGAACGGTTACACATAACCTAACTTAGCTATCTCTAGCCAGATAGCACTCATCCAATCCTCACTTTTTACACTTAACGAGGAGAACGAACTTG
"Isobeggiatoa" PS (BGP_5130 / 5131)
TAAATACGATGGGTATTTTAAGGCTCACGTTTCTATCTAGCACAGCTAACATAACTATTATTTTTTGAACCCCTCTTTTTGATACCGATTTTTGCGCCTTTGATGTGAGTTTCATGACAACCGATGGTTCCCAATTATG
Beggiatoa alba B18LD (BegalDRAFT_2896 / 2897)
TAGTTTACCGACGGTACGTTTATGAATCCGTGGCGGTTTCAGGTTGCAAACACTGATTAATCATTTTTTTGCAACCAAACCCGCCCACCTCCAGAATTCTACAAGCGGAGAACGACCTTGAAAGATTTAATG
Probability > 99%
Probability 95-99%
Probability 90-95%

FIGURE 5 | Gene neighborhoods and intergenic sequence for putative RNA polymerase beta and beta prime subunit genes from *Beggiatoaceae* genome sequences. (A) Gene neighborhoods are shown as cartoons (left) and lists (right). Vertical lines alongside lists indicate ORF sequences common to at least three species (the "Isobeggiatoa" PS genome is very partial). RNA polymerase subunit genes are in bold. **(B)** Predicted stem-loop structures in the beta/beta prime intergenic region are the first results from a minimum free energy calculation using the default settings of the MaxExpect algorithm from the RNAstructure Web Server (http://rna.urmc.rochester.edu/RNAstructureWeb/, Reuter and Mathews, 2010), with arrows shaded by the probability of the interaction.

that in *T. ingrica* is longer (338 nt) than those in *B. alba* and BGP (126 and 133 nt, respectively). It also includes a stronger potential stem-loop structure (**Figure 5B**), possibly a transcriptional terminator. One scenario is that the two genes became transcriptionally uncoupled in a common ancestor of *T. ingrica* and the BOGUAY strain, making the intergenic region a viable site for genomic rearrangements and introduction (by whatever mechanism) of TAACTGA repeats. If the putative beta prime variants are in fact expressed, perhaps the separation of beta and beta prime allows the levels of the three proteins to be separately regulated.

Predicted Sensor Proteins are Immediately Downstream of the Putative Secondary Beta Prime Subunit Genes

Each of the variant beta prime genes is immediately followed by a predicted hybrid sensor kinase gene (**Figure 6**). These have nearly identical structures according to the Conserved Domain Database (CDD; Marchler-Bauer et al., 2011): a GAF-superfamily domain, four PAS domains, a histidine kinase, three REC domains, and an HPT domain. GAF domains, which include those in FhlA (formate hydrogen lyase transcriptional activator)-family proteins, bind and respond to cyclic-nucleotide second messengers (Aravind and Ponting, 1997). PAS domains are intracellular or periplasmic redox sensors responsive to various stimuli, including light and oxygen, with specificity determined partly by small-molecule cofactors such as a heme or flavin (Taylor and Zhulin, 1999; Kneuper et al., 2010). HisKA-HATPase_c (histidine kinase A—histidine-kinase-like ATPase) domains respond to sensor inputs by autophosphorylating on a histidine residue, which in turn typically phosphorylates a response regulator (REC) domain aspartate residue (Stock et al., 2000), changing its conformation and, for example, promoting dimerization and DNA binding. HPt (histidine-containing phosphotransfer) domains transfer phosphate groups to other proteins along phosphorylation cascades (Matsushika and Mizuno, 1998). Both the BOGUAY and *T. ingrica* putative sensor proteins are strongly predicted by PSORTb (Yu et al., 2010) to be inner-membrane proteins, by comparison with *E. coli* BarA, which was localized in a membrane proteomic survey (Daley et al., 2005). As is usual with the highly modular sensor proteins, neither has any other full-length matches in current databases, although each of the subdomains does. There is not yet enough known about sensor proteins to predict what stimuli these might respond to, or what their upstream and downstream interaction partners might be, but it can be hypothesized that they sense a condition in the periplasm and transmit that information to cytoplasmic elements via a phosphorylation cascade, which may directly or indirectly contact the variant beta prime.


A Neighborhoods of genes for alternate or modified RNAP β´ subunits
BOGUAY 00100_0018
TCAGTTATCAGTTATCAATTATCAGTTA
TAACTGA repeats
TAACTGATTACTGATAACTGATTACTGATAACTGATATG
0016
BOGUAY_0019
0021 0022
0014 0015
0017
BOGUAY_0018
0020
Thioploca ingrica THII_0330 (BAP54627)
GTTATATCACTATAAAATAACCTTGAATAACTTAAAATG
0329
THII_0330
THII_0331
0332 0333 0334 0335
0326 0327 0328
RNA polymerase β´subunit
Sensor protein
B Domain structure of predicted hybrid sensor kinase protein genes downstream of alternate RNAP β´ subunit genes
BOGUAY_0019 1
FhlA PAS PAS PAS
PAS HisKA
HATPase_c
REC A1
REC A2
REC A3
HPt
1645
THII_0331 1
GAF superfam.
PAS PAS PAS
PAS HisKA
HATPase_c
REC REC REC
HPt
1713

BOGUAY_0014	conserved hypothetical protein
BOGUAY_0015	HAD hydrolase
BOGUAY_0016	Predicted ATPase (AAA+ superfamily)
BOGUAY_0017	tRNA (cytidine(34)-2'-O)-methyltransferase
BOGUAY_0018	**RNAP β´**
BOGUAY_0019	**PAS domain S-box (REC A1/A2/A3)**
BOGUAY_0020	Arsenate reductase and related proteins, glutaredoxin family
BOGUAY_0021	DUF1566
BOGUAY_0022	glycogen/starch synthases, ADP-glucose type

THII_0326	hypothetical protein
THII_0327	peptidase M23B
THII_0328	serine/threonine protein kinase
THII_0329	pentapeptide repeat-containing protein
THII_0330	**RNAP β´**
THII_0331	**PAS domain S-box**
THII_0332	hypothetical protein
THII_0333	secreted protein
THII_0334	nitric oxide reductase
THII_0335	nitric oxide reductase subunit B

FIGURE 6 | (A) Gene neighborhoods for putative alternate RNA polymerase beta prime subunit genes from the BOGUAY and *Thioploca ingrica* genome sequences. Gene neighborhoods are shown as cartoons (left) and lists (right). Positions of TAACTGA repeats within and upstream of BOGUAY 00100_0018 are indicated; the corresponding upstream sequence from *T. ingrica,* which has no repeats in this region, is included for comparison. **(B)** Predicted domain structures of putative downstream sensor proteins. Domains were identified in CDD (Marchler-Bauer et al., 2011).

SUMMARY AND PERSPECTIVES

TAACTGA Repeats May Play Different Roles in Different Species

The draft genomes of Orange Guaymas "Maribeggiatoa" (BOGUAY) and *Cand*. "Thiomargarita nelsonii," and to a lesser extent *T. ingrica*, contain an unusually high number of TAACTGA direct repeats, while close relative *B. alba* and apparently all but one other sequenced Gammaproteobacterium (*T. violascens*, also a sulfur oxidizer) have none at all. TAACTGA direct repeats were also found in Cyanobacteria, especially in species known for harboring long repetitive arrays, and in a few Bacteroidetes. This is consistent with earlier evidence for genetic exchange among these groups (MacGregor et al., 2013c), particularly the Cyanobacteria and some *Beggiatoaceae*, although no exchange mechanism is obvious as yet. Once introduced into a genome, whether by exchange or mutation, the tolerated sites and orientations for repeats will be determined by sequence characteristics such as length, coding potential, and propensity to form secondary structures, and by their interaction with existing cellular machinery. For the BOGUAY intergenic TAACTGA repeats, a plausible scenario is that they were recognized by an existing nucleic acid-binding protein—perhaps a ribosomal subunit, perhaps a protein that interacts with these—and over time a regulatory network evolved by selection for individuals with favorable protein interaction(s) and combinations of insertions. The original introduction may have happened in the common ancestor of a branch of the *Beggiatoaceae*, with different networks evolving (or not) in each subsequent lineage. The very long arrays in species such as *M. aeruginosa* and *G. forsetii* suggest a role in genome rearrangement may have evolved in these. Acquisition of additional genome sequences for the *Beggiatoaceae* may help illuminate this history.

Another possibility is that a TAACTGA-binding protein is the mobile element. On entering a new species, it could interact with pre-existing "good-enough" RNA or DNA sequences, with closer matches and useful locations evolving over time. Identification of repeat-binding protein(s) in the BOGUAY genome and evaluation of their inferred phylogeny and gene neighborhoods in other species could help in evaluating this model.

TAACTGA Repeats May Play a Role in Translational Regulation in the BOGUAY Strain

In the BOGUAY genome, most of the TAACTGA repeats are in "forward" orientation immediately upstream of putative start codons and overlapping the expected ribosome-binding site, suggesting that they may have taken on a role in translational regulation in this species. Genes and ORFs lacking recognizable Shine-Dalgarno sequences are prevalent in BOGUAY and many other bacterial genomes (Ma et al., 2002), including such highly expressed genes as the *E. coli* ribosomal protein S1 gene (*rpsA*; Aseev and Boni, 2011); in BOGUAY, only a small proportion of these are preceded by TAACTGA repeats. Possibilities for the translational role of the BOGUAY repeats, not all mutually exclusive, include:

a) Canonical BOGUAY ribosomes are able to bind efficiently enough to the repeats for production of even highly translated proteins, despite the absence of sequence complementary to the 16S rRNA.
b) Ribosomes with different subunit compositions—in particular, those lacking S1—may have different binding sites, as already recognized for leaderless mRNAs; this could include TAACTGA repeats.
c) Repeats may be recognized by some other RNA-binding protein (e.g., a Csp-like one), which then recruits ribosomes.
d) Repeats are irrelevant, these genes are translated like leaderless mRNAs by ribosomes lacking S1.

Possible Function of Second RNA Polymerase Beta Prime Subunit-Like Proteins in BOGUAY and *Thioploca Ingrica*

Another unusual feature of the BOGUAY genome is a second RNA polymerase beta prime-like ORF, also found in *T. ingrica*, and immediately upstream of multisensor kinases in both. In BOGUAY, this putative alternate or modified gene is both preceded by and contains TAACTGA repeats. The BOGUAY genome has the additional peculiarity that the beta and "normal" beta prime genes are not adjacent, but rather internal to separate contigs. Assuming the beta prime-like gene is expressed, one possibility is that it associates with other RNA polymerase subunits, forming either a functional or a non-functional complex: the absence of key catalytic residues suggests it would be non-functional, but this would need experimental testing. This is somewhat supported by the physical separation of the beta and beta prime genes in BOGUAY, and their possible transcriptional separation in *T. ingrica*: if two proteins are competing for the beta prime role, it may be beneficial to regulate their production separately from that of their common partners. In BOGUAY, the TAACTGA repeats upstream of the beta prime-like ORF suggest that it may be part of their putative global regulatory network.

Perspectives

Experimental tests of these ideas will be challenging in an uncultivated, difficult to collect species. Some basic questions may be answerable by transcriptomic analysis of samples collected from different Guaymas Basin sites and/or preincubated under different conditions (temperature, oxygen, sulfide). Are ORFs preceded by repeats up- and downregulated in concert? Is the second beta-prime like ORF transcribed, and if so, under what conditions? How does its expression pattern compare with that of other RNA polymerase subunit genes? There are also indications from the partial "Isobeggiatoa" genome sequences that the more accessible Baltic Sea *Beggiatoaceae* may have similar repeat distributions. *In vitro* identification of repeat-binding proteins might be possible from total mat protein preparations, or by heterologous expression and isolation of cloned (or synthesized) genes for candidate proteins.

FUNDING

The Guaymas Basin project was funded by NSF OCE 0647633. Genome sequencing was performed by the J. Craig Venter Institute, with funding from The Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project. The use of RAST was supported in part by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services (NIAD) under contract HHSN266200400042C.

ACKNOWLEDGMENTS

Thanks to the captain and crews of R/V *Atlantis* and HOV *Alvin* for two enjoyable cruises to Guaymas Basin, and to Drs. Alecia Septer and Lisa Nigro for helpful comments on the manuscript. Genome comparisons were carried out in IMG/ER (https://img.jgi.doe.gov/cgi-bin/er/main.cgi).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01397

REFERENCES

Aravind, L., and Ponting, C. P. (1997). The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.* 22, 458–459. doi: 10.1016/S0968-0004(97)01148-1

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603. doi: 10.1093/nar/gks400

Aseev, L. V., and Boni, I. V. (2011). Extraribosomal functions of bacterial ribosomal proteins. *Mol. Biol.* 45, 739–750. doi: 10.1134/S0026893311050025

Attia, A. S., and Hansen, E. J. (2006). A conserved tetranucleotide repeat is necessary for wild-type expression of the *Moraxella catarrhalis* UspA2 protein. *J. Bacteriol.* 188, 7840–7852. doi: 10.1128/JB.01204-06

Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Boni, I. V., Artamonova, V. S., Tzareva, N. V., and Dreyfus, M. (2001). Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1. *EMBO J.* 20, 4222–4232. doi: 10.1093/emboj/20.15.4222

Bornberg-Bauer, E., Rivals, E., and Vingron, M. (1998). Computational approaches to identify leucine zippers. *Nucleic Acids Res.* 26, 2740–2746. doi: 10.1093/nar/26.11.2740

Bowman, J. P., and Nichols, D. S. (2002). Aequorivita gen. nov., a member of the family Flavobacteriaceae isolated from terrestrial and marine Antarctic habitats. *Int. J. Syst. Evol. Microbiol.* 52, 1533–1541. doi: 10.1099/00207713-52-5-1533

Byrgazov, K., Vesper, O., and Moll, I. (2013). Ribosome heterogeneity: another level of complexity in bacterial translation regulation. *Curr. Opin. Microbiol.* 16, 133–139. doi: 10.1016/j.mib.2013.01.009

Daley, D. O., Rapp, M., Granseth, E., Melén, K., Drew, D., and Von Heijne, G. (2005). Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science* 308, 1321–1323. doi: 10.1126/science.1109730

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328. doi: 10.1016/S0968-0004(98)01274-2

Davis, C. A., Bingman, C. A., Landick, R., Record, M. T., and Saecker, R. M. (2007). Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by *Escherichia coli* RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7833–7838. doi: 10.1073/pnas.0609888104

Duss, O., Michel, E., Konté, N. D. D., Schubert, M., and Allain, F. H. T. (2014). Molecular basis for the wide range of affinity found in Csr/Rsm protein-RNA recognition. *Nucleic Acids Res.* 42, 5332–5346. doi: 10.1093/nar/gku141

Duval, M., Korepanov, A., Fuchsbauer, O., Fechter, P., Haller, A., Fabbretti, A., et al. (2013). *Escherichia coli* ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol.* 11:e1001731. doi: 10.1371/journal.pbio.1001731

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Foster, T. J., Lundblad, V., Hanley-Way, S., Halling, S. M., and Kleckner, N. (1981). Three Tn10-associated excision events: relationship to transposition and role of direct and inverted repeats. *Cell* 23, 215–227. doi: 10.1016/0092-8674(81)90286-5

Frangeul, L., Quillardet, P., Castets, A. M., Humbert, J. F., Matthijs, H. C. P., Cortez, D., et al. (2008). Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* 9:274. doi: 10.1186/1471-2164-9-274

Fu, Y., Deiorio-Haggar, K., Anthony, J., and Meyer, M. M. (2013). Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria. *Nucleic Acids Res.* 41, 3491–3503. doi: 10.1093/nar/gkt055

Giuliodori, A. M., Brandi, A., Gualerzi, C. O., and Pon, C. L. (2004). Preferential translation of cold-shock mRNAs during cold adaptation. *RNA* 10, 265–276. doi: 10.1261/rna.5164904

Guo, X., and Mrázek, J. (2008). Long simple sequence repeats in host-adapted pathogens localize near genes encoding antigens, housekeeping genes, and pseudogenes. *J. Mol. Evol.* 67, 497–509. doi: 10.1007/s00239-008-9166-5

Gur-Arie, R., Cohen, C. J., Eitan, Y., Shelef, L., Hallerman, E. M., and Kashi, Y. (2000). Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome Res.* 10, 62–71. doi: 10.1101/gr.10.1.62

Hajnsdorf, E., and Boni, I. V. (2012). Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie* 94, 1544–1553. doi: 10.1016/j.biochi.2012.02.010

Horn, G., Hofweber, R., Kremer, W., and Kalbitzer, H. R. (2007). Structure and function of bacterial cold shock proteins. *Cell. Mol. Life Sci.* 64, 1457–1470. doi: 10.1007/s00018-007-6388-4

Humbert, J. F., Barbe, V., Latifi, A., Gugger, M., Calteau, A., Coursin, T., et al. (2013). A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLoS ONE* 8:e70747. doi: 10.1371/journal.pone.0070747

Jannasch, H. W., Nelson, D. C., and Wirsen, C. O. (1989). Massive natural occurrence of unusually large bacteria (*Beggiatoa* sp.) at a hydrothermal deep-sea vent site. *Nature* 342, 834–836. doi: 10.1038/342834a0

Kämpfer, P., Matthews, H., Glaeser, S. P., Martin, K., Lodders, N., and Faye, I. (2011). *Elizabethkingia anophelis* sp. nov., isolated from the midgut of the mosquito Anopheles gambiae. *Int. J. Syst. Evol. Microbiol.* 61, 2670–2675. doi: 10.1099/ijs.0.026393-0

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., et al. (2007). Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res.* 14, 247–256. doi: 10.1093/dnares/dsm026

Khayrullina, G. A., Raabe, C. A., Hoe, C. H., Becker, K., Reinhardt, R., Tang, T. H., et al. (2012). Transcription analysis and small non-protein coding RNAs associated with bacterial ribosomal protein operons. *Curr. Med. Chem.* 19, 5187–5198. doi: 10.2174/092986712803530485

Kneuper, H., Scheu, P. D., Etzkorn, M., Sevvana, M., Dünnwald, P., Becker, S., et al. (2010). "Sensing ligands by periplasmic sensing histidine kinases with sensory PAS domains," in *Sensory Mechanisms in Bacteria: Molecular Aspects of Signal Recognition,* eds S. Spiro and R. Dixon. (Norfolk, UK: Caister Academic Press), 39–59.

Kojima, H., Ogura, Y., Yamamoto, N., Togashi, T., Mori, H., Watanabe, T., et al. (2015). Ecophysiology of Thioploca ingrica as revealed by the complete genome sequence supplemented with proteomic evidence. *ISME J.* 9, 1166–1176. doi: 10.1038/ismej.2014.209

Lafontaine, E. R., Wagner, N. J., and Hansen, E. J. (2001). Expression of the *Moraxella catarrhalis* UspA1 protein undergoes phase variation and is regulated at the transcriptional level. *J. Bacteriol.* 183, 1540–1551. doi: 10.1128/JB.183.5.1540-1551.2001

Lovett, S. T. (2004). Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* 52, 1243–1253. doi: 10.1111/j.1365-2958.2004.04076.x

Ma, J., Campbell, A., and Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.* 184, 5733–5745. doi: 10.1128/JB.184.20.5733-5745.2002

MacGregor, B. J., Biddle, J. F., Harbort, C., Matthysse, A. G., and Teske, A. (2013a). Sulfide oxidation, nitrate respiration, carbon acquisition, and electron transport pathways suggested by the draft genome of a single orange Guaymas Basin *Beggiatoa* (*Cand. Maribeggiatoa)* sp. filament. *Mar. Genomics* 11, 53–65. doi: 10.1016/j.margen.2013.08.001

MacGregor, B. J., Biddle, J. F., Siebert, J. R., Staunton, E., Hegg, E. L., Matthysse, A. G., et al. (2013b). Why orange Guaymas Basin *Beggiatoa* (*Maribeggiatoa*) spp. are orange: single-filament genome-enabled identification of an abundant octaheme cytochrome with hydroxylamine oxidase, hydrazine oxidase, and nitrite reductase activities. *Appl. Environ. Microbiol.* 79, 1183–1190. doi: 10.1128/AEM.02538-12

MacGregor, B. J., Biddle, J. F., and Teske, A. (2013c). Mobile elements in a single-filament orange Guaymas Basin *Beggiatoa* (*Maribeggiatoa*) sp. draft genome: evidence for genetic exchange with cyanobacteria. *Appl. Environ. Microbiol.* 79, 3974–3985. doi: 10.1128/AEM.03821-12

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189

Markowitz, V. M., Mavromatis, K., Ivanova, N. N., Chen, I.-M. A., Chu, K., and Kyrpides, N. C. (2009). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25, 2271–2278. doi: 10.1093/bioinformatics/btp393

Matsushika, A., and Mizuno, T. (1998). The structure and function of the histidine-containing phosphotransfer (HPt) signaling domain of the *Escherichia coli* ArcB sensor. *J. Biochem.* 124, 440–445. doi: 10.1093/oxfordjournals.jbchem.a022132

McHatton, S. C., Barry, J. P., Jannasch, H. W., and Nelson, D. C. (1996). High nitrate concentrations in vacuolate, autotrophic marine *Beggiatoa* spp. *Appl. Environ. Microbiol.* 62, 954–958.

McKay, L. J., MacGregor, B. J., Biddle, J. F., Albert, D. B., Mendlovitz, H. P., Hoer, D. R., et al. (2012). Spatial heterogeneity and underlying geochemistry of phylogenetically diverse orange and white *Beggiatoa* mats in Guaymas Basin hydrothermal sediments. *Deep-Sea Res. I* 67, 21–31. doi: 10.1016/j.dsr.2012.04.011

Miropolskaya, N., Esyunina, D., Klimasauskas, S., Nikiforov, V., Artsimovitch, I., and Kulbachinskiy, A. (2014). Interplay between the trigger loop and the F loop during RNA polymerase catalysis. *Nucleic Acids Res.* 42, 544–552. doi: 10.1093/nar/gkt877

Morgan, H. P., Estibeiro, P., Wear, M. A., Max, K. E. A., Heinemann, U., Cubeddu, L., et al. (2007). Sequence specificity of single-stranded DNA-binding proteins: a novel DNA microarray approach. *Nucleic Acids Res.* 35:e75. doi: 10.1093/nar/gkm040

Mrázek, J., Guo, X., and Shah, A. (2007). Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8472–8477. doi: 10.1073/pnas.0702412104

Murakami, K. S. (2013). X-ray crystal structure of *Escherichia coli* RNA polymerase σ^{70} holoenzyme. *J. Biol. Chem.* 288, 9126–9134. doi: 10.1074/jbc.M112.430900

Mussmann, M., Hu, F. Z., Richter, M., de Beer, D., Preisler, A., Jørgensen, B. B., et al. (2007). Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol.* 5:e230. doi: 10.1371/journal.pbio.0050230

Opalka, N., Brown, J., Lane, W. J., Twist, K. A., Landick, R., Asturias, F. J., et al. (2010). Complete structural model of *Escherichia coli* RNA polymerase from a hybrid approach. *PLoS Biol.* 8:e1000483. doi: 10.1371/journal.pbio.1000483

Otsuka, S., Suda, S., Li, R. H., Matsumoto, S., and Watanabe, M. M. (2000). Morphological variability of colonies of *Microcystis* morphospecies in culture. *J. Gen. Appl. Microbiol.* 46, 39–50. doi: 10.2323/jgam.46.39

Parry, D. A., Fraser, R. D., and Squire, J. M. (2008). Fifty years of coiled-coils and α-helical bundles: a close relationship between sequence and structure. *J. Struct. Biol.* 163, 258–269. doi: 10.1016/j.jsb.2008.01.016

Reuter, J. S., and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129. doi: 10.1186/1471-2105-11-129

Rigden, D. J., Jedrzejas, M. J., and Galperin, M. Y. (2003). An extracellular calcium-binding domain in bacteria with a distant relationship to EF-hands. *FEMS Microbiol. Lett.* 221, 103–110. doi: 10.1016/S0378-1097(03)00160-5

Romeo, T., Vakulskas, C. A., and Babitzke, P. (2013). Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environ. Microbiol.* 15, 313–324. doi: 10.1111/j.1462-2920.2012.02794.x

Sachs, R., Max, K. E. A., Heinemann, U., and Balbach, J. (2012). RNA single strands bind to a conserved surface of the major cold shock protein in crystals and solution. *RNA* 18, 65–76. doi: 10.1261/rna.02809212

Sallam, K. I., Mitani, Y., and Tamura, T. (2006). Construction of random transposition mutagenesis system in *Rhodococcus erythropolis* using IS1415. *J. Biotechnol.* 121, 13–22. doi: 10.1016/j.jbiotec.2005.07.007

Salman, V., Amann, R., Girnth, A. C., Polerecky, L., Bailey, J. V., Høgslund, S., et al. (2011). A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. *Syst. Appl. Microbiol.* 34, 243–259. doi: 10.1016/j.syapm.2011.02.001

Salman, V., Bailey, J. V., and Teske, A. (2013). Phylogenetic and morphologic complexity of giant sulphur bacteria. *Antonie Van Leeuwenhoek* 104, 169–186. doi: 10.1007/s10482-013-9952-y

Skorski, P., Leroy, P., Fayet, O., Dreyfus, M., and Hermann-Le Denmat, S. (2006). The highly efficient translation initiation region from the *Escherichia coli rpsA* gene lacks a Shine-Dalgarno element. *J. Bacteriol.* 188, 6277–6285. doi: 10.1128/JB.00591-06

Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). Two-component signal transduction. *Annu. Rev. Biochem.* 69, 183–215. doi: 10.1146/annurev.biochem.69.1.183

Subramanian, A. R., and Vanduin, J. (1977). Exchange of individual ribosomal proteins between ribosomes as studied by heavy-isotope transfer experiments. *Mol. Gen. Genet.* 158, 1–9. doi: 10.1007/BF00455113

Sukhodolets, M. V., Garges, S., and Adhya, S. (2006). Ribosomal protein S1 promotes transcriptional cycling. *RNA* 12, 1505–1513. doi: 10.1261/rna.2321606

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631–637. doi: 10.1126/science.278.5338.631

Taylor, B. L., and Zhulin, I. B. (1999). PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* 63, 479–506.

Torres, M., Condon, C., Balada, J. M., Squires, C., and Squires, C. L. (2001). Ribosomal protein S4 is a transcription factor with properties remarkably similar to NusA, a protein involved in both non-ribosomal and ribosomal RNA antitermination. *EMBO J.* 20, 3811–3820. doi: 10.1093/emboj/20.14.3811

Tzareva, N. V., Makhno, V. I., and Boni, I. V. (1994). Ribosome-messenger recognition in the absence of the Shine-Dalgarno interactions. *FEBS Lett.* 337, 189–194. doi: 10.1016/0014-5793(94)80271-8

Van Assche, E., Van Puyvelde, S., Vanderleyden, J., and Steenackers, H. P. (2015). RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Front. Microbiol.* 6:141. doi: 10.3389/fmicb.2015.00141

Windgassen, T. A., Mooney, R. A., Nayak, D., Palangat, M., Zhang, J. W., and Landick, R. (2014). Trigger-helix folding pathway and SI3 mediate catalysis and hairpin-stabilized pausing by *Escherichia coli* RNA polymerase. *Nucleic Acids Res.* 42, 12707–12721. doi: 10.1093/nar/gku997

Xu, D., and Zhang, Y. (2012). *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 80, 1715–1735. doi: 10.1002/prot.24065

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615. doi: 10.1093/bioinformatics/btq249

Zhou, K., Aertsen, A., and Michiels, C. W. (2014). The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol. Rev.* 38, 119–141. doi: 10.1111/1574-6976.12036

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.