



Visualizing Evolutionary Relationships of Multidomain Proteins: An Example from Receiver (REC) Domains of Sensor Histidine Kinases in the *Candidatus* Maribeggiatoa str. Orange Guaymas Draft Genome

Barbara J. MacGregor*

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

OPEN ACCESS

Edited by:

Kelly Wrighton, Ohio State University, USA

Reviewed by:

Brian P. Hedlund, University of Nevada, Las Vegas, USA Roland Hatzenpichler, Montana State University, 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: 16 May 2016 Accepted: 24 October 2016 Published: 14 November 2016

Citation:

MacGregor BJ (2016) Visualizing Evolutionary Relationships of Multidomain Proteins: An Example from Receiver (REC) Domains of Sensor Histidine Kinases in the Candidatus Maribeggiatoa str. Orange Guaymas Draft Genome. Front. Microbiol. 7:1780. doi: 10.3389/fmicb.2016.01780

For multidomain proteins, evolutionary changes may occur at the domain as well as the whole-protein level. An example is presented here, with suggestions for how such complicated relationships might be visualized. Earlier analysis of the Candidatus Maribeggiatoa str. Orange Guaymas (BOGUAY; Gammaproteobacteria) single-filament draft genome found evidence of gene exchange with the phylogenetically distant Cyanobacteria, particularly for sensory and signal transduction proteins. Because these are modular proteins, known to undergo frequent duplication, domain swapping, and horizontal gene transfer, a single domain was chosen for analysis, Recognition (REC) domains are short (~125 amino acids) and well conserved, simplifying sequence alignments and phylogenetic calculations. Over 100 of these were identified in the BOGUAY genome and found to have a wide range of inferred phylogenetic relationships. Two sets were chosen here for detailed study. One set of four BOGUAY ORFs has closest relatives among other Beggiatoaceae and Cyanobacteria. A second set of four has REC domains with more mixed affiliations, including other Beggiatoaceae, several sulfate-reducing Deltaproteobacteria and Firmicutes, magnetotactic Nitrospirae, one Shewanella and one Ferrimonas strain (both Gammaproteobacteria), and numerous Vibrio vulnificus and V. navarrensis strains (also Gammaproteobacteria). For an overview of the possible origins of the whole proteins and the surrounding genomic regions, colorcoded BLASTP results were produced and displayed against cartoons showing protein domain structure of predicted genes. This is suggested as a visualization method for investigation of possible horizontally transferred regions, giving more detail than scans of DNA composition and codon usage but much faster than carrying out full phylogenetic analyses for multiple proteins. As expected, most of the predicted sensor histidine kinases investigated have two or more segments with distinct BLASTP affiliations. For the first set of BOGUAY ORFs, the flanking regions were also examined, and the results suggest they are embedded in genomic stretches with complex histories. An automated method of creating such visualizations could be generally useful; a wish list for its features is given.

Keywords: Orange Guaymas Maribeggiatoa, Beggiatoaceae, Cyanobacteria, sensor histidine kinases, recognition (REC) domains, multidomain proteins, horizontal gene transfer

INTRODUCTION

The *Candidatus* Maribeggiatoa str. Orange Guaymas (BOGUAY) single-filament draft genome contains potential mobile genetic elements of several types (introns, inteins, and possible excision elements) with close relatives among the phylogenetically distant Cyanobacteria, suggesting a history of genetic exchange between these groups (MacGregor et al., 2013c). As identified by the top five BLASTP matches, the largest single category of potentially exchanged genes was for sensory and signal transduction proteins, raising the question of what environmental conditions these might respond to, what other genes they might interact with, and which lineages may have contributed (or received) which functions.

The multidomain nature of signal transduction proteins complicates phylogenetic inferences. Domains appear to be swapped at a high rate relative to overall genome evolution both within and between species (reviewed in Capra and Laub, 2012; Salazar and Laub, 2015), presumably allowing a range of regulatory adaptations to be tested within a population. The REC (recognition) domain was selected for this analysis, being short (\sim 125 amino acids) and easily aligned. In the studied cases, REC domains are phosphorylated by a histidine kinase upstream in a signaling chain, which may be on the same or a different protein. They may change conformation or dimerize, and interact with an element downstream in the chain (reviewed in Casino et al., 2010).

Understanding the evolution of multidomain signal transduction proteins is a daunting task, but could yield insights into the regulatory adaptations of bacterial species to their local environments. For the Beggiatoaceae, these include shallow hypersaline ponds, sulfidic seeps in freshwater lakes, and sulfidic deep sea sediments at vents, seeps, and the western continental margins of Africa and South America. These large vacuolated bacteria are sometimes found in close association with other species in microbial mats, and can be covered with epibionts (e.g., Fliss, 2014; Flood et al., 2016), suggesting opportunities for gene transfer. The individual cells or filaments may have higher than usual genome copy numbers (Angert, 2012), which could make them better able to carry out and tolerate genetic rearrangements of various sorts; these points are the topic of current research in several labs.

Analysis of such multidomain proteins would greatly benefit from phylogenetically nested domain-, protein-, local neighborhood-, and whole-genome visualization tools linked to the ever-expanding sequence database. Presented here is a detailed look at two sets of REC domains, illustrating both their possible history and how such visualizations might appear.

MATERIALS AND METHODS

Available Genomes of *Beggiatoaceae* and Related Large Sulfur Bacteria

Few *Beggiatoaceae* and related large sulfur gammaproteobacteria have as yet been sequenced, and fewer still are in cultivation; their classification is still in progress (Salman et al., 2011, 2013). There are complete or near-complete genome sequences for *Beggiatoa alba* B18LD (BegalDRAFT; Lucas et al., unpublished), *Beggiatoa leptomitiformis* D-420 (Fomenkov et al., 2015), *Thioploca ingrica* (THII; Kojima et al., 2015), and *Cand*. Maribeggiatoa "Orange Guaymas" (MacGregor et al., 2013a,b,c); a partial sequence for *Cand*. Thiomargarita nelsonii (Mußmann et al., unpublished); and one partial (*Cand*. Isobeggiatoa sp. PS; BGP) and one very partial (*Cand*. Parabeggiatoa sp. SS; BGS) genome for two filaments collected from Baltic Sea harbor sediment (Mußmann et al., 2007). By 16S rRNA phylogeny, *B. alba* is in a separate clade from the rest (Salman et al., 2013; *B. leptomitiformis* had not been sequenced when these names were proposed).

Naming Conventions

Orange Guaymas Cand. Maribeggiatoa filament (abbreviated BOGUAY) draft genome open reading frames (ORFs) are referred to either by their IMG/ER locus tag (e.g., BOGUAY_1733) or by contig and locus tag (e.g., BOGUAY 00362_1733). Cand. Isobeggiatoa sp. PS, Cand. Parabeggiatoa sp. SS, and B. alba B18LD locus tags begin with BGP, BGS, or BegalDRAFT, respectively. ORFs from other species are referred to by GenBank accession number or, for those whose chromosomal location was checked, by IMG locus tags. In phylogenetic trees, numbers in parentheses (e.g., 5-119) indicate the amino acid residues used in the REC sequence alignment. At present, the same sequence may have two or three different accession numbers in GenBank and two different locus tags in IMG/ER; an attempt was made to be consistent, but because sequences were downloaded at different times, designations within a species may have more than one form.

Where multiple REC domains occur within a single putative BOGUAY protein, letter designations were used for the proteins and numbers for REC domains, in order of their position [e.g., BOGUAY 00362_1733 contains REC domains N1 (amino acids 8-120) and N2 (amino acids 148-260)].

Phylogenetic Reconstructions

Putative REC domains were identified in the BOGUAY genome within IMG/ER (www.jgi.doe.gov) by searching for relevant keywords and COG, KOG, and pfam numbers. The predicted amino acid sequences were then used to search the NCBI

ode used in					Domain l	used to buil	ullee
ther figures					•	*	
A WP_002684462.1 (706-821) Beggiatoa alba (BegalDRAFT_1092)	Cache_1	HAMP		ĸ	A		AS_9 PAS GGDE
B BOGUAY 00217_2058 (747_862) C WP_008477205.1 (564-679) <i>"Isobeggiatoa"</i> sp. PS (BGP_5448)	Cache_1	HAMP HAMP		к к	A A	R A R PAS_9	
D ↓ WP_045478691 (772-887) Thioploca ingrica (THIL 0443 = Gao060138_111079) E ↓ WP_008480732.1 (1-117) <i>"Isobeggiatoa"</i> sp. PS (BGP_0359)	Cache_1_sup	HAMP		к	А	R R PAS CI	HD R
F GOGUAY 00316_2961 (1148_1264)	COG3292	Y_Y_Y		к	А	R GGDEF	EAL
G	COG3292	Y_Y_Y		к	А	R PAS CI R PP2Cc	HD
WP_045480560 (1124-1239) Thioploca ingrica (THII_3682 = Ga0060138_111801)	0000202			ĸ	A	R SpollE	
J KHD11477.1 (126-241) Candidatus Thiomargarita nelsonii (OT06_10600) K YP_007120051.1 (602-717) Microcoleus sp. PCC 7113	CHASE			к	/A A	R PAS R PP2Cc	
L YP_007149058.1 (734-849) Cylindrospermum stagnale PCC 7417	Cache_1	HAMP		ĸ	A	R K A	
YP_007083863.1 (598-713) Oscillatoria acuminata PCC 6304 WP_006621627.1 (876-991) Arthrospira platensis	MASE1 Cache_1	HAMP PAS		к к	A A	R PP2Cc R CHD	
YP_005069038.1 (874-989) Arthrospira platensis NIES 39	Cache_1	HAMP PAS		к	A	R CHD	
WP_009782294.1 (893-1008) Lyngbya sp. PCC 8106 YP_722985.1 (742-857) Trichodesmium erythraeum IMS101	Cache_1 Cache_1	HAMP PAC HAMP		к к	A	R CHD R CHD	
WP_009785933.1 (640-755) Lyngbya sp. PCC 8106	CHASE_super			к	A	R SpollE/	
WP_017660438.1 (740-855) Geitlerinema sp. PCC 7105 WP_008478346.1 (1-104) "Isobeggiatoa" sp. PS (BGP_1186)	Cache_1	HAMP		к	A	R CHD TF R CHD	PR_2
WP_008477452.1 (727-842) "Isobeggiatoa" sp. PS (BGP_5629)	Cache_1	HAMP		к	A	R CHD	
BOGUAY 01308_0467 (707-822) WP_045475695.1 (712-827) Thioploca ingrica (THII_2325 = Ga0060138_113165)	Cache_1 Cache_1_sup	HAMP HAMP		к к	A A	R CHD R CHD	
WP_006619087.1 (722-837) Arthrospira platensis	Cache_1	HAMP		к	A	R PAS CI	HD
 YP_007084184.1 (636-751) Oscillatoria acuminata PCC 6304 WP_008187689.1 (732-847) Moorea producens 	CHASE			ĸ	A A	R CHD R CHD	
WP_006101034.1 (713-828) Coleofasciculus chthonoplastes	7TMR DISM_			к	A	R CHD	
YP_007122515.1 (696-811) Microcoleus sp. PCC 7113 YP_007088849.1 (1852-1967) Oscillatoria acuminata PCC 6304	7TMR DISM_ PKc AAA_1		3899 GAF	ĸ	A A	R CHD B CHD	
J WP_008310939.1 (713-828) Leptolyngbya sp. PCC 6406	Cache_1	HAMP		ĸ	A	R CHD	
↓ WP_007356084.1 (714-829) <i>Oscillatoria</i> YP_007072482.1 (719-834) <i>Leptolyngbya</i> sp. PCC 7376	COG4191 Cache_1	HAMP HAMP		к к	A A	R CHD R CHD	
WP_008317381.1 (788-903) Leptolyngbya sp. PCC 6406	Cache_1	HAMP		к	A/ /A	R CHD	
ACV42481.1 (748-863) Lyngbya majuscula 19L WP_006098419.1 (981-1095) Coleofasciculus chthonoplastes	Cache_1 Cache 1	HAMP HAMP PAS		к к	A A/ /A	R CHD R PAS CI	ЧD
WP_018398141.1 (688-802) filamentous cvanobacterium ESEC.1	7TMR DISM_	7TM		ĸ	A	R CHD	
WP_017718872.1 (505-620) Oscillatoria sp. PCC 10802 YP_007119567.1 (884-998) Microcoleus sp. PCC 7113		PAS HAMP PAS		к к	A A	R CHD R CHD	
YP_007070807.1 (734-849) Leptolyngbya sp. PCC 7376		HAMP		ĸ	A	R PAS CI	HD
WP_009786545.1 (719-834) Lyngbya sp. PCC 8106 YP_722386.1 (720-834) Trichodesmium erythraeum IMS101	Cache_1 Cache 1	HAMP HAMP		к к	A A	R CHD R CHD	
YP_007114488.1 (723-838) Oscillatoria nigro-viridis PCC 7112	Cache_1	HAMP		ĸ	Â	R CHD	
WP_007353506.1 (732-847) Oscillatoria	Cache_1 Cache_1	HAMP HAMP		к к	A A	R CHD R CHD	
YP_007114489.1 (683-798) Oscillatoria nigro-viridis PCC 7112 YP_007087963.1 (701-815) Oscillatoria acuminata PCC 6304	CHASE2	HAWF		ĸ	Â		R_superfamily
WP_018397756.1 (654-769) filamentous cyanobacterium ESFC.1 WP_017304826.1 (746-861) Spirulina subsalsa	Cache_1 Cache_1	HAMP HAMP		к к	A	R CHD R CHD	
WP_007304826.1 (746-661) Spirulina subsalsa	R PAS	HAWP		ĸ	Â	R CHD	
YP_007120333.1 (581-696) Microcoleus sp. PCC 7113	R PAS R PAS			к к	A A	R CHD R CHD	
WP_017305029.1 (560-675) Spirulina subsalsa WP_009592651.1 (666-780) Paenibacillus sp. HGF5	7TMR DISM_	7TM		ĸ	Â	R K_3	
YP_003243024.1 (666-780) Paenibacillus sp. Y412MC10	7TMR DISM			K	A	R K_3	
WP_006210711.1 (664-778) Paenibacillus vortex YP_004640948.1 (692-803) Paenibacillus mucilaginosus KNP414	7TMR DISM_ 7TMR DISM_			к к	A A	R K_3 R K_3 A	
WP_000488281.1 (725-839) Vibrio cholerae	_				A	R PAS	
WP_001907286.1 (215-329) Vibrio albensis WP_000488270.1 (725-839) Vibrio mimicus		HAMP		к	A	R PAS R PAS	
F WP_005474319.1 (708-822) Vibrio sp. 16		HAMP		K	A	R PAS/	
WP_005593238.1 (724-838) Vibrio		HAMP		к	A A	R PAS R PAS	
YP_001446594.1 (725-839) Vibrio campbellii ATCC BAA 1116					A	R PAS/	
VP_005024017.1 (725-839) Vibrio sp. EJY3 WP_004401931.1 (725-839) Vibrio nigripulchritudo		HAMP		к	A	R PAS/ R PAS/	
YP_002263922.1 (723-837) Aliivibrio salmonicida LFI1238		HAMP		К	A	R PAS/	
WP_006230602.1 (712-826) Photobacterium profundum YP_128751.1 (726-840) Photobacterium profundum SS9		HAMP HAMP		к К	A A	R PAS/ R PAS/	
WP_007470661.1 (723-837) Photobacterium sp. AK15		HAMP		К	A	R PAS/	
AEW28962.1 (724-838) Photobacterium damselae subsp. piscicida WP_006001166.1 (720-834) Desulfuromonas acetoxidans	7MTR DISM	HAMP 7TM		K K	A	R PAS/ R PAS/	
YP_006440036.1 (542-657) Turneriella parva DSM 21527		HAMP		ĸ	A	R SpollE	
YP_004120895.1 (716-827) Desulfovibrio aespoeensis Aspo_2 WP_019029155.1 (1103-1218) Colwellia piezophila	7TMR DISM_ COG3292	_7TM Y_Y_Y		к К	A A	R PAS R GGDEF	
WP_017463536.1 (1146-1260) Dyella ginsengisoli	/Vgb–SGL	Y_Y_Y		к	A	R GGDEF	
YP_526141.1 (1147-1262) Saccharophagus degradans 2_40 WP_008473219.1 (27-142) "Isobeggiatoa" sp. SS (BGS_0408)	COG3292	Y_Y_Y		к	A	R GGDEF R PP2Cc/	
YP_006762911.1 (724-838) Desulfobacula toluolica Tol2	7TMR DISM_	7TM		К/	Α	R PAS/	
WP_020286031.1 (710-825) Osedax symbiont Rs2		HAMP		к	А	R K	A
YP_004447710.1 (1194-1309) Haliscomenobacter hydrossis DSM 1100	Reg_prop	COG3292/	Y_Y_Y	ĸ	A	R CHD	
└── YP_004446809.1 (1201-1316) Haliscomenobacter hydrossis DSM 1100 └── WP_008476837.1 (253-368) "Isobeggiatoa" sp. PS (BGP_1126)		COG3292	Y_Y_Y	ĸ	A A	R CHD R	
WP_045478459.1 (352-467) Thioploca ingrica (THII_0070 = Ga0060138_111448)		HAMP		К	A	R CHD	
BOGUAY 01054_0100 (681-797) WP_002682771.1 (687-802) Beggiatoa alba (BegaIDRAFT_0207)		HAMP HAMP		к К	A A	R CHD R CHD	
YP_007058398.1 (738-853) Rivularia sp. PCC 7116 YP_007058396.1 (737-852) Rivularia sp. PCC 7116	Cache_1	HAMP		К	Α	R CHD	
YP_720573.1 (725-840) Trichodesmium erythraeum IMS101	Cache_1	HAMP HAMP		к К	A A	R CHD R CHD	
YP 720574.1 (722-837) Trichodesmium erythraeum IMS101	Cache_1	HAMP		к	A	R CHD	
WP_002694524.1 (888-1004) Microscilla marina WP_007129051.1 (724-836) Paenibacillus lactis	/CHASE3	GAF_3		к К	A A	R CHD R His_kinas	a A
WP_010651340.1 (685-798) <i>Oceanobacillus</i> sp. Ndiop	7TMR DISME	ED2 7TMI	R-DISM 7TM	ĸ	A	R K	A
	Ger	omonret	ophactoric			111.17.4	
0.10			eobacteria	4	K	HisKA	
		ochaete			A	HATPase	e_c
	Delt	aproteol	oacteria		R	REC	
		teroidete			L		
		nicutes	-				
			rio				
	Cya	nobacte	na				

FIGURE 1 | REC domain Tree 36. The tree was produced using RAxML rapid bootstrapping (Stamatakis, 2006) as implemented in ARB (Ludwig et al., 2004), using a random initial tree, the PROTGAMMA rate distribution and WAG amino acid substitution models, empirical amino acid frequencies, and branch optimization. Six *Clostridia* sequences were used to root the tree. Domain composition is from CDD (Marchler-Bauer et al., 2011). See Supplemental Table 1 for definition of domain abbreviations.



FIGURE 2 | Color-coded Tree 36 BLASTP results with cartoon representations of domain compositions. Each BLASTP hit is represented by a single bar, colored by phylum. Beggiatoaceae are indicated by wide black bars, other Gammaproteobacteria by normal ones. Black symbols at the top of some BLASTP results indicate domains with similar affiliations, identified by looking at the full species lists (not shown).

database with BLASTP, and the highest-scoring matching regions downloaded. One hundred hits per sequence were taken initially, but due to extensive overlap between sets, this was decreased to 50 for most searches. Because few genomic sequences are so far available for the *Beggiatoaceae*, there is no chance for search results to be dominated by very close relatives, as happens for example with pathogens. The final database included 4672 amino acid sequences, some of which derived from putative proteins with two or more separate REC domains. The REC domains were aligned in MEGA5 (Tamura et al., 2011) using MUSCLE (Edgar, 2004) and small adjustments made manually. Where only partial domains were retrieved by BLAST, upstream and/or downstream residues were added from the GenBank entries. The aligned sequences were exported to ARB (Ludwig et al., 2004), an initial guide tree computed by neighbor joining,

and subtrees centered on BOGUAY sequences computed first by neighbor joining and then with RAxML rapid bootstrapping (Stamatakis, 2006) using a random initial tree, the PROTMIX or more recently implemented PROTGAMMA rate distribution and WAG amino acid substitution models, empirical amino acid frequencies, and branch optimization. The final species composition of most trees is the result of several RAxML and neighbor-joining runs, with consistently outlying or unstable branches removed. The final set of 35 RAxML trees included 2254 REC domain sequences. The trees are numbered in order of the occurrence of the majority of their species in the initial neighborjoining guide tree. The three trees described here (9, 19, and 36) were updated by new BLASTP searches in Spring 2016. Sequences and alignments of the REC domains in each tree are provided in the Supplemental Files. The domain structure of all proteins



was predicted using the Conserved Domain Database (CDD, Marchler-Bauer et al., 2011). Protein domain abbreviations are those used by the CDD (summarized in Supplemental Table 1). Predicted chromosomal arrangements and ORF annotations were retrieved from IMG/ER.

RESULTS

Three REC-domain phylogenetic trees will be discussed here, numbered according to their position in the initial guide tree. Tree 36 includes primarily ORFs predicted to contain a single REC domain internal to a protein, closely related to Cyanobacterial sequences. Trees 19 and 9 include REC domains typically found together at the end of an ORF, often with a third partial or complete REC domain between them, and more diverse affiliations. Their inferred phylogenies are considered below, along with those of the proteins containing them and the surrounding genomic neighborhoods.

Tree 36: Cyanobacterial Affiliations

Inferred REC Domain Phylogeny

Tree 36 (Figure 1) includes three clusters of predicted Beggiatoaceae REC domains with cyanobacterial sequences as closest affiliates. More distantly related are several Spirochaete, Deltaproteobacteria, Bacteroidetes, and Firmicute sequences. The domain structures of the predicted proteins including them generally have one or two domains upstream of the HisKA-HATPase_c-REC (abbreviated KA-R) core, and one or occasionally two domains downstream of this, with a range of predicted sensory (e.g., Cache_1, PAS) and signal transduction (e.g., HAMP, CHD) roles.

Inferred Phylogeny of Other Domains

A full phylogenetic analysis of each domain of each protein would be very time-consuming, so a quicker method was sought for an overview, shown here for the first 12 sequences in the tree (coded A through L). These include 10 Beggiatoaceae and two Cyanobacteria sequences. First, BLASTP searches were



used to identify possible separately derived regions within each predicted protein (Supplemental Figure 1); all 12 appear to be mosaics. The matches were then illustrated as phylogenetically color-coded bars, displayed against the predicted domain structures (**Figure 2**). Wide black bars were used to distinguish Beggiatoaceae from other Gammaproteobacteria.

All but one REC domain is within a segment with primarily Cyanobacterial BLASTP hits, as expected. The exception is BOGUAY_2961 (Figure 2F): the REC domain is at the downstream end of a segment with a high proportion of Bacteroidetes matches (although Cyanobacteria are still among the highest-scoring), immediately upstream of a Cyanobacteria-related segment. All but the two shortest sequences (BGP_5448 and BGP_0359) also have at least one segment with substantially different associations. Consideration of the species lists themselves (not shown) identified similar segments in different predicted proteins, indicated by large black symbols; for example, BOGUAY_2058 (Figure 2B) and THII_0443 (Figure 2D) both have terminal HATPase_c domains

with similar mixed, particularly Deltaproteobacterial hits. This is an unusual position for this domain, more often found between a histidine kinase and REC domain. Only one set of matches (**Figure 2G**) is dominated by Gammaproteobacterial sequences, and these do not include any Beggiatoaceae except a "self" hit, suggesting that this segment too may have been exchanged horizontally.

Putting these cartoons in the context of the REC domain phylogenetic tree (**Figure 3**), hypotheses about how these predicted proteins may have been assembled can be suggested. For example, two *Beggiatoaceae* have a predicted terminal SpoIIE domain similar to that of the Cyanobacterial Mic7113_0733 (**Figures 3H,I,K**; pentagon symbol). All matches to all three segments are from Cyanobacteria, Beggiatoaceae, or (for THII_3682 only) unclassified bacteria. The Beggiatoaceae matches for this domain are dispersed rather than clustered (**Figures 2H,I,K**), suggesting that (alone or as part of some protein) it may have been introduced from Cyanobacteria to an ancestral Beggiatoaceae species, and then diverged within

				Dom		_	build tre	-	
EDN71132.1 (749-863) Cand. Isobeggiatoa sp. PS			CAE 2	Tropomyosin 1	KA	R*	R sup	R* —	-
EDN68471.1 (461–575) Cand. Isobeggiatoa sp. PS			GAF_2	OmpH sup/	KA	R*	R/	n	
EDN68655.1 (437–552) Cand. Isobeggiatoa sp. PS				/PRK13923_sup	KA	R*	R_sup/	R*	
EDN72941.1 (355–470) Cand. Parabeggiatoa sp. SS				111(10520_30p	KA	R*	R R		
EDN71014.1 (1223–1337) Cand. Isobeggiatoa sp. PS		HAMP	GAE 2	Tropomyosin 1	KA	B*	CheY	B*	
EDN71011.1 (929-1048) Cand. Isobeggiatoa sp. PS		HAMP	GAF 2	/PRK13923/	KA	B*	CheY	B*	
WP_052491805.1 (1086-1200) Thioploca ingrica	Tar Tsr LBD/	HAMP	GAF 2	/Nup54	KA	R*		R*	
EDN68110.1 (331-445) Cand. Isobeggiatoa sp. PS					KA	R*	R	R*	
BOGUAY 00806_2995 (1111-1226) (L1)		HAMP	GAF_2	/Nup54	KA	R*	R	R*	
WP_062153829.1 (992–1106) Beggiatoa leptomitiformis D-402			GAF_2	/Nup54	KA	R*	R/	R*	
WP_002684997.1 (1002–1116) Beggiatoa alba B18LD			GAF_2	/Nup54	KA	R*	CheY	R*	
WP_052492084.1 (880-994) Thioploca ingrica	Tar_Tsr_LBD_sup	HAMP	GAF_2	Tropomyosin_1/	KA	R*	R/	R*	
EDN68203.1 (358-477) Cand. Isobeggiatoa sp. PS					KA	R*	R_sup	R*	BEGGIATOACEAE
EDN70752.1 (796-910) Cand. Isobeggiatoa sp. PS			GAF_2	PAS	KA	R*	CheY	R*	
BOGUAY 00153_2324 (70-186) (K1)						R*	R_sup/	R*	
WP_052492165.1 (1070-1184) Thioploca ingrica		HAMP	GAF_2	Mod_r	KA	R*	CheY	R* B*	
WP_062149449.1 (1025-1139) Beggiatoa leptomitiformis D-402		HAMP HAMP	GAF_2	PAS /PRK13923_sup	KA	R* B*	CheY	R* B*	
WP_002690178.1 (1048–1162) Beggiatoa alba B18LD WP_062154998.1 (864–983) Beggiatoa leptomitiformis D-402	Tar Tsr LBD sup		GAF_2 GAF_2	PAS /PRK13923_sup	KA KA	R*	CheY CheY	R*	
WP_002154996.1 (864–983) Beggiatoa leptomitionnis D=402 WP_002690554.1 (866–980) Beggiatoa alba B18LD	Tar Tsr LBD sup		GAF_2 GAF 2	ds DNAbind/	KA	R*	B/	R*	
BOGUAY 00286_0624 (70–193) (J1)	Tal_Tsl_LDD_sup	TIAWF	GAF_2	us_DNAbinu/	NA.	R*	n/	B*1	
WP_052491912.1 (1316–1430) Thioploca ingrica	Tar Tsr LBD	HAMP	GAF_2	/PRK13923	KA	R*	R	R*	
WP_052491710.1 (798–912) Thioploca ingrica	141_131_600	1 IZAIWI	GAF 2	HemX/	KA	R*	R/	R*	
WP 052491909.1 (872–986) Thioploca ingrica	Tar Tsr LBD sup	HAMP	GAF 2	/PRK13923 sup	KA	R*	Rsup	R*	
WP 052491910.1 (1199–1313) Thioploca ingrica	Fic_sup/	HAMP	FhIA	OmpH/	KA	B*	B/	R*	
BOGUAY 01204 2878 (135-250) (M1)					/A	R*	R sup/	R*	
WP_045434835.1 (816-930) bacterium UASB270, wastewater treatment sludge	3	HAMP	GAF_2	/Tropomyosin_1	KA	R*	R	R*	
ABD75785.1 (939-1053) uncultured bacterium, tidal flat		HAMP	GAF_2		KA	R*	R	R*	
ABD75783.1 (578-692) uncultured bacterium, tidal flat				/TMF_TATA_bd	KA	R*	R	R*	
WP_015759177.1 (770-884) Desulfotomaculum acetoxidans DSM 771		HAMP	GAF_3	/PRK13923	KA	R*	R	R*	
WP_014956971.1 (991–1105) Desulfobacula toluolica Tol2		HAMP		/Tropomyosin_1	KA	R*	R	R	lot in Tree 9
WP_035223370.1 (991–1105) Desulfobacula sp. TS		HAMP	GAF_2	/Tropomyosin_1	KA	R*	R/		lot in Tree 9
WP_028585323.1 (1008–1122) Desulfobulbus mediterraneus DS	SM 13871	HAMP	GAF_2	Tropomyosin_1/	KA KA	R* B*	R/	R	
WP_061058081.1 (986–1100) Vibrio vulnificus ATL 6–1306 WP_017422776.1 (986–1223) Vibrio vulnificus ATCC 27562		/HAMP /HAMP		RILP_like_sup RILP like sup	KA	R*	R R	R R	
WP_017422776.1 (986–1223) Vibrio Vulnincus ATCC 27562 WP_045609177.1 (989–1195) Vibrio vulnificus SC9740		/HAMP		RILP_like_sup	KA	R*	R	R	
WP_045569434.1 (981–1095) Vibrio sp. S234–5		HAMP		/Tropomyosin 1	KA	R*	B	R	
WP_039442982.1 (981–1182) Vibrio navarrensis 08–2462		HAMP		/Tropomyosin_1	KA	B*	R	R	
WP_039463312.1 (981–1095) Vibrio navarrensis 0053–83		HAMP	GAF 3	/Tropomyosin 1	KA	R*	R	R	
GAK87599.1 (999–1113) Vibrio ponticus JCM 19238		/HAMP	GAF 2	/Tropomyosin 1	KA	B*	B		
WP_012141995.1 (990-1104) Shewanella sediminis HAW-EB3		/HAMP	GAF_2	/Tropomyosin_1	KA	R*	R	R	
WP_028109261.1 (970-1084) Ferrimonas futtsuensis DSM 18154		HAMP	GAF_2	ApoLp_III	KA	R*	R	R	
WP_020000022.1 (1005–1119) Desulfovibrio desulfuricans ATCC 29578		HAMP	GAF_2	Cortex_I_coil	KA	R*	R	R	
WP_011189034.1 (974–1088) Desulfotalea psychrophila LSv54		HAMP	GAF_2	HemX/	KA	R*	R	R*	
EKD40401.1 (238-352) uncultured bacterium, subsurface aquifer sediment					A	R*	R/	R*	
KJR99450.1 (989–1103) Desulfobulbaceae bacterium BRH_c16a		HAMP	GAF_3	OmpH	KA	R*	R/	R*	
WP_015415991.1 (1009-1123) Desulfovibrio piezophilus C1TLV30		HAMP	0.45	PTZ00423	KA	R*	R	R*	
WP_034640925.1 (996-1110) Desulfovibrio inopinatus DSM 10711		HAMP	GAF	/Tropomyosin_1	KA	R*	R	R*	
WP_028586737.1 (1029–1143) Desulfocurvus vexinensis DSM 17965	Tor Tor LPD	HAMP	GAF GAF 2	Tropomyosin_1	KA KA	R* B*	R	R* R*	
WP_052567468.1 (809–928) Cand. Magnetobacterium casensis KJU86825.1 (806–919) Cand. Magnetobacterium bavaricum TM–1	Tar_Tsr_LBD_sup Tar-Tsr_LBD_sup		GAF_2 GAF_2	HemX/ OmpH/	KA	н [.] В*		н [.] R*_	
KUU86825.1 (806–919) Cand. Magnetobacterium bavaricum TM-1 KWT91955.1 (805–919) Nitrospirae bacterium HCH-1	Tar Tsr LBD sup		GAF_2 GAF_2	/Tropomyosin 1	KA	R*		R	
WP_051328394.1 (1058–1172) Desulfatirhabdium butyrativorans DSM 18734		HAMP	GAF_2 GAF_2	Tropomyosin 1/	KA	B*	B/	R*	
CBX27662.1 (632–747) uncultured Desulfobacterium sp., enrichment culture N47		HAMP	GAF_2		KA	R*	R	R*	Gammaproteobacte
WP_012805237.1 (970–1084) Desulfomicrobium baculatum DSM 4028		HAMP	GAF 2	DivIVA_sup	KA	R*	R	R*	Betaproteobacteria
WP_013163556.1 (909-1023) Desulfurivibrio alkaliphilus AHT 2		PAS_9	GAF 2	/Tropomyosin_1	KA	R*	R	R I	Deltaproteobacteria
EAT02265.1 (23–137) delta proteobacterium MLMS–1				· · · · · · · · · · · · · · · · · · ·		R*	R	R	Nitrospirae
AGX87357.1 (995–1109) Cand. Symbiobacter mobilis CR		HAMP	GAF	Tropomyosin_1	KA	R*	R	R*	Nitrospinae
WP_005009692.1 (876–990) Nitrospina gracilis 3/211			GAF_2	/Tropomyosin_1	KA	R*	R/	R	
WP_018085118.1 (1028-1142) Desulfurispora thermophila DSM 16022	HAMP	HAMP	GAF_2	Tropomyosin_1	KA	R*	R	R*	Firmicutes
									Unclassified bacteri

FIGURE 5 | REC domain Tree 19. See legend to Figure 1 for methods. BOGUAY domain compositions are highlighted by boxes. Three Paenibacillus sequences were used to root the tree.

these. Alternatively, there could have been several transfers from different Cyanobacteria to different Beggiatoaceae, but this seems less likely.

BLASTP Relatives of Neighboring Predicted Proteins

To search for the boundaries of possible mobile elements that could have introduced sensor proteins to the Beggiatoaceae genomes, the BLASTP visualizations were continued to either side of the ORFs encoding the putative REC domain-containing proteins until either sequences with only Gammaproteobacterial matches or the end of a contig was reached (Figure 4; see Supplemental Table 2 for ORF descriptions). Nearly all of the Beggiatoaceae ORFs are part of regions with apparently complex histories. The illustration for a vertically transmitted gene is expected to resemble that of T. ingrica Ga0060138_111077 (Figure 4D), with Beggiatoaceae sequences first (wide black bars) and then strictly Gammaproteobacteria, for which there are many more than 100 (the number of bars) sequenced genomes available. This is only suggestive evidence; proof would require a closer look at the species identities and sequences.

The *B. alba* region in **Figure 4A** appears to have the simplest exchange history. On the bottom strand are four consecutive ORFs (BegalDRAFT 1093-1096) whose closest affiliations are to just one other Beggiatoaceae sequence, followed by a mixture of species with Alphaproteobacteria predominating. These are predicted to encode a cobalt-nickel transporter (Supplemental Table 2A), a likely candidate for gene transfer, since heavy-metal resistance is often carried on mobile elements (reviewed in

Bouzat and Hoostal, 2013); the simplest interpretation would be that it was acquired by the *B. alba* lineage after its divergence from the other Beggiatoaceae, although differential retention is also a possibility. Immediately upstream of this, on the opposite strand, is the predicted REC domain protein gene (BegalDRAFT_1092), with primarily Cyanobacterial affiliations. All four regions of this ORF have sporadically distributed Beggiatoaceae matches (wide black bars), suggesting that they may have diverged since their acquisition by a common ancestor. These two possibly transferred segments are flanked by at least several ORFs (as far as checked) with primarily Gammaproteobacterial BLASTP hits.

Most of the other Beggiatoaceae regions illustrated appear to have more complex histories. A few include predicted genes that may record at least part of the transfer mechanism: toxins and an antitoxin (Figures 4F,H), a transposase (Figure 4I), and two restriction endonucleases and a possible associated methylase (Figures 4D,F,H). BOGUAY_1145 and T. ingrica Ga0060138_111801 (Figures 4H,I) appear to be related by rearrangement, with homologous ORFs (Ga0060138_111802, BOGUAY_1144) encoding putative hypothetical proteins immediately upstream of very similar REC domain protein genes. The ORF just downstream of this one in T. ingrica (Ga0060138_111803) also has a BOGUAY homolog (BOGUAY_1128), but it is further from the REC domain protein, and this gene pair has less similar affiliations, although both are annotated as CHAT domain ("Caspase HetF Associated with Tprs"; possible peptidases) proteins with N-terminal tetratricopeptide repeats (groups TPR_16



FIGURE 6 | REC domain Tree 9. See legend to Figure 1 for methods. BOGUAY domain compositions are highlighted by boxes. Three Paenibacillus sequences were used to root the tree.

and TPR_12, respectively). Two of the Cyanobacterial REC domain proteins in Tree 36 (**Figure 1**: WP_0176660438, TPR_2; YP_007087963, TPR_16) are themselves annotated with C-terminal tetratricopeptide domains. Such repeats suggest a possible mechanism for recombination, but if there was a repeat-mediated event, these sequences appear to have diverged considerably since then (not shown). No other potential homolog pairs were identified.

As a crude measure of the likelihood of finding transposases on a given genome segment, the assembled genome length was divided by the number of annotated transposases and putative transposases annotated in IMG. Leaving aside the very incomplete *Cand*. Parabeggiatoa sp. SS genome, estimates range from one "transposase" every ~24,000 bp in *B. leptomitiformis* (178 in 4.3 mbp), through one every ~67–69,000 bp in *B. alba* and *T. ingrica*, to one every ~92,000 bp in BOGUAY (52 in 4.8 mbp). It is therefore not unusual to find one on the genome fragments illustrated in **Figure 4**, but neither can a contribution to transfer or rearrangement of the sensor proteins be ruled out. The annotation of toxin/antitoxin and restriction/methylation genes is more difficult because of the many classes of each (see e.g., MacGregor et al., 2013c, for BOGUAY), but they are not especially rare. It could be informative to investigate how many of these possible mobility functions are found in regions with phylogenetically mixed vs. homogeneous blastp affiliations.

By contrast to the wide phylogenetic range of blastp hits in the Beggiatoaceae genome segments, both of the Cyanobacterial predicted REC domain protein genes (**Figures 4K,L**) appear to be in much more stable regions. The *Microcoleus* sp. PCC 7113 one is upstream of three ORFs with predominately Cyanobacterial and Beggiatoaceae affiliations, then flanked by ORFs with only Cyanobacterial hits. The *Cylindrospermum stagnale* PCC 7417 one is flanked by ORFs with nearly or entirely Cyanobacterial affiliations. This suggests the Cyanobacteria as the immediate



FIGURE 7 | Domain structure and color-coded BLASTP matches for selected Tree 19 sequences. Vertical lines in domain cartoons indicate possible phylogenetic break points as identified by BLASTP, used to produce the color-coded diagrams. Open black (Gammaproteobacteria) and maroon (Firmicutes) bars were used for *Vibrio* and *Paenibacillus* spp., respectively, since these were dominant matches for the downstream segments.

donor of these REC domain-containing gene segments to the Beggiatoaceae, rather than the other way around.

Trees 19 and 9: Paired REC Domains Inferred REC Domain Phylogeny

The REC domains shown in Trees 19 and 9 (Figures 5, 6) are found as the first and last, respectively, of series of two or three REC domains at the C-terminal end of predicted proteins. In the Beggiatoaceae and most of the other species shown they occur together, and have similar inferred phylogenies. This suggests they may have undergone lateral transfer primarily as a unit, while upstream domains are more variable. The main exceptions are several predicted Deltaproteobacterial proteins found only in Tree 19, and a cluster of Sphingobacterales (Bacteroidetes; Pedobacter spp. in particular) found only in Tree 9. Because these are RAxML trees, identical sequences could not be included; Pedobacter is a well-studied genus with many published genomes, so the species shown are only a subset. Similarly, while the lists of Vibrio species do not completely overlap between the two trees, each one shown is only an example of a much larger group of primarily Vibrio vulnificus and V. navarrensis strains. For purposes of this paper it was not considered necessary to make a full concordance. An impression of the number of identical sequences can be gained from Figure 7 (open black bars).

BLASTP Affiliations of Complete Proteins

For an overview of the affiliations of the complete predicted proteins, BLASTP searches (not shown) were used as above to identify possible boundaries in a subset of the Tree 19/Tree 9 sequences (**Figure 7**). First, this illustrates a problem with

the method in its current state. Deeply sequenced genera such as *Vibrio* and *Paenibacillus* can dominate visually even if they represent only a small part of the phylogenetic range; they were assigned special colors to distinguish them from other Gammaproteobacteria and Firmicutes.

The downstream segment of this predicted protein group appears to have undergone duplication (or more) within the *Beggiatoaceae*, as evidenced by the clusters of wide black bands and similar banding patterns (e.g., compare BOGUAY "L1" and the *B. leptomitiformis*, *B. alba*, and *T. ingrica* predicted proteins grouped with it).

Two of the predicted BOGUAY proteins (coded "K" and "L"), several of the *Isobeggiatoa* ones, and the *Parabeggiatoa* one begin at position 1 of their respective contigs, and therefore may be missing upstream domains. This can affect the BLASTP results, particularly for modular proteins. For example, BOGUAY 00153_2324 ("K") appears distinct from the downstream segments of the other four predicted proteins shown for its REC domain clade (**Figure 7**, top right). However, when equivalent segments of each ORF are used, results for all five are similar (**Figure 8**, compare "A to end" and "B to end"). In particular, there are fewer Cyanobacterial and more *Paenibacillus* matches at the lower end of the scale, especially for the marine strains BOGUAY and *T. ingrica*.

Identification of Possible Transfer Mechanisms

A simpler look was taken at gene neighborhoods for these trees than for Tree 36. Possible indicators of genome rearrangements and horizontal gene transfer were highlighted in the immediate neighborhoods (arbitrarily defined as the 50





	[3785] CheR [3784] CheB	(Dtox_3786)
[14630] [14640] [100237] [14800]	[14470] ParC topoisomerase-4 subunit A [14480] ParE topoisomerase-4 subunit B [14680] CheR [14640] CheR [14600] RecJ single-stranded-DNA-specific exonuclease	Desulfobacula toluolica Tol2 (TOL2_14620)
[0023] [00245]	[00237] CheR [00238] CheB [00245] Small Multidrug Resistance protein	Desulfobulbus mediterraneus (G494DRAFT_00236)
[03781] [03780] (03789] [03770]	[03789] assimilatory nitrite reductase (NAD(P)H) small subunit [03788] assimilatory nitrite reductase (NAD(P)H) large subunit precursor [03781] (DABA [03770] assimilatory nitrite reductase (NAD(P)H) small subunit [03770] assimilatory nitrite reductase (NAD(P)H) small subunit [03769] assimilatory nitrite reductase (NAD(P)H) large subunit precursor	Vibrio vulnificus ATCC 27562 (B736DRAFT_03782)
	[05162, 05163] CheR [05164] CheB [05171] DNA segregation ATPase FtsK/SpolIIE, S-DNA-T family [1658] DNA topoisomerase	Vibrio ponticus JCM19238 (Ga0058818_05160)
[02120]	[1647] CheR [1646] CheB [02134] Transposase and inactivated derivatives	Shewanella sediminis HAW-EB3 (Ssed_1648)
[02119]	02120] CheR (02119] CheB (021018] Wuldfrug resistance efflux pump [021018] Muldfrug resistance efflux pump [02101] Superinfection exclusion protein B	Ferrimonas futtsuensis DSM 18154 (G505DRAFT_02121)
[1792] [1791]	[01948] CheR [01947] CheB [1792] CheR	Desulfovibrio desulfuricans ATCC 29578 (EK04DRAFT_01949)
[12717] [12718] [12735]	[1791] CheB [12717] CheR [12718] CheB	Desulfotalea psychrophila LSv54 (DP1793) Desulfovibrio piezophilus C1TLV30
[01578] [01580]	[12728] RecA [12735] Phage repressor protein C [01579] CheR [01580] CheB	(BN4_12716) Desulfovibrio inopinatus (G451DRAFT_01578)
[00805] [00804] [00796] [00780] [00790]	[00805] CheR [00704] CheB [00796] SSB [00796] MolB [00789] MotA	Desulfocurvus vexinensis DSM 17965 (G495DRAFT_00806)
[02703] [02702] [02700] [104411]	[02706] zinc-finger of transposase IS204/IS1001/IS1096/IS1165 [02703] CheR [02702] CheB [02700] Predicted nuclease, HicB family	Cand. Magnetobacterium casensis (McasDRAFT 02704)
(1044410) (1044409) (1044408)	[1044410] CheB [1044409] HicB-like [1044408] YgIT-type zinc finger domain-containing protein	Cand. Magnetobacterium bavaricum TM- (Ga0078519_104441)
	[00744] transposase, IS605 OrfB family, central region [00739] Xerb Integrase [00729] zinc transport system substrate-binding protein	Desulfatirhabdium butyrivorans DSM 1873 (G492DRAFT_00730)
	(0021) CheR (0020) CheB	Desulfomicrobium baculatum DSM 4028 (Dbac_0022)
	[1332] CheR [1331] CheB [1278] Abortive infection bacteriophage resistance protein	Desulfurivibrio alkaliphilus AHT2 (DaAHT2_1333)
[1249]	[1275] Type III RE [1273] adenine-specific DNA-methyltransferase [1268] transposase	Cand. Symbiobacter mobilis CR (Cenrod_1265)
	[2534344977] CheR [2534344968] MotA [2534344966] MotB	Nitrospina gracilis (NITGR_600027)
[00893] [0889]	[00873] DNA replication and repair protein RecN [00892] CheR [00893] CheB [00899] XerD recombinase	Desulfurispora thermophila DSM 16022 (B064DRAFT_00891)
	 CheR CheB transposon functions toxins and antitoxins restriction enzymes and ass 	ociated DNA methylases
	[1440] [00237] [00238] [00238] [00238] [00238] [00238] [00238] [002370] [003700] [0037	[1440] [1440]



kb displayed by IMG) of the Tree 19 putative REC domain proteins (**Figures 9–11**). These include homologous ORFs, duplicated genes, transposons, toxin and antitoxin genes, XerD integrase/recombinase, and restriction enzyme and associated methylase genes. The predicted transposons, toxin/antitoxin, and restriction enzymes are sporadically distributed and no clearly related sets were found, nor do the two XerD-like sequences (Tree 10D, F) have any significant similarity (not shown). If any of these do have common ancestors, they have diverged or decayed considerably.

There are three examples of similar gene neighborhoods. First, the two Sphingobacterium regions from Tree 9 are very similar (Figure 11). Second, nearly all non-Beggiatoaceae have predicted genes for chemotaxis methyltransferase CheR/CheB directly downstream of the predicted REC domain protein gene; the exceptions are Nitrospina gracilis and two Pedobacter spp. (CheR only; Figures 10E, 11) and Desulfatirhabdium butyrivorans (neither; Figure 10D). By contrast, CheR/CheB are found in only one of the Beggiatoaceae neighborhoods (BOGUAY "L1"; Figure 9), where they are separated from the Tree 19/Tree 9 ORF BOGUAY_2995 by a second multi-REC domain signal transduction histidine kinase (BOGUAY_2996; CDD-predicted domains REC-HisKA-HATPase_c-RE C/-REC/).

Third, NitT/TauT family transport system substrate-binding proteins are annotated directly upstream and downstream of one T. ingrica ORF (Figure 9, Ga0060138_112336); directly upstream of one BOGUAY (Figure 9, BOGUAY 00286_0624); between two of a set of three T. ingrica ORFs (Figure 9, Ga0060138_113542, 113544); directly upstream of the Symbiobacterium ORF (Figure 10E); and farther downstream of one B. alba ORF. This is suggestive of a pair of co-transferred genes, and a duplication event in T. ingrica. The upstream ORF of the T. ingrica pair is highly similar to the one from BOGUAY (Evalue 8e-156, the closest current database match) and somewhat less so to the one from Symbiobacter (2e-15), suggesting these may be related by inheritance (within the Beggiatoaceae) or transfer. None of the remaining putative NitT/TauT sequences are significantly similar, however (not shown), including the two flanking the T. ingrica ORF; this REC domain protein may have taken on a role in more than one transport system.

There are also two examples of tandem repeats of highly similar ORFs that could record a recombination event: putative genes for large and small assimilatory nitrate reductase subunits in *V. vulnificus* ATCC 27562 (**Figure 10C**) and numerous other *V. vulnificus* strains (not shown), and two putative dihydrolipoamide dehydrogenase genes just downstream of the

REC domain protein genes in the two *Sphingobacterium* strains shown (**Figure 11**) and several others (not shown).

DISCUSSION

In attempting to gain an overview of the sensor complement of the single-filament "Maribeggiatoa" Orange Guaymas (BOGUAY) genome, and its possible evolutionary origins, several obstacles were encountered. Their modular structure means that domains or domain clusters, rather than whole proteins, are the relevant evolutionary units in many cases. Ideally, phylogenetic reconstructions could be carried out for each individual domain or coherent group of domains, and the individual analyses recombined for visualization at different levels of detail.

From the work presented here, some of the desirable features of such a bioinformatic tool for modular proteins became clear. Many of the pieces are already available, but linkages are not seamless. In particular, the ability to reorder maps and diagrams with reference to user-generated phylogenetic trees (or, failing that, in user-selected order) would greatly ease the production of figures such as those shown.

The wish list also includes the following: (1) The data should remain easily updateable. (2) Segments defined by BLASTP scores (with default settings) may correctly segment some proteins, but well-defined domains are likely a better choice. It would be useful to be able to group or ungroup these; for example, HisKA-HATPase-REC seems to be a generally conserved unit. (3) Significance cutoffs should be adjustable, and large sets of near-identical sequences collapsible. (4) The level of phylogenetic resolution used for illustration should be customizable and flexible; in some cases phylum-level resolution is sufficient, in other cases—even in the same tree—species or even strain level may be needed (e.g., the many *V. vulnificus* strains). (5) Visual display of BLASTP results would be improved by including some representation of scores, perhaps as a bar graph parallel to

REFERENCES

- Angert, E. R. (2012). DNA replication and genomic architecture of very large bacteria. Annu. Rev. Microbiol. 66, 197–212. doi: 10.1146/annurev-micro-090110-102827
- Bouzat, J. L., and Hoostal, M. J. (2013). Evolutionary analysis and lateral gene transfer of two-component regulatory systems associated with heavy-metal tolerance in bacteria. J. Mol. Evol. 76, 267–279. doi: 10.1007/s00239-013-9558-z
- Capra, E. J., and Laub, M. T. (2012). Evolution of two-component signal transduction systems. Annu. Rev. Microbiol. 66, 325–347. doi: 10.1146/annurev-micro-092611-150039
- Casino, P., Rubio, V., and Marina, A. (2010). The mechanism of signal transduction by two-component systems. *Curr. Opin. Struct. Biol.* 20, 763–771. doi: 10.1016/j.sbi.2010.09.010
- 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
- Fliss, P. S. (2014). Pearl in the Mud: Genome Assembly and Binning of a Cold Seep Thiomargarita nelsonii Cell and Associated Epibionts from an Environmental Metagenome. M.S. MSc, University of Minnesota.

the phylogenetically color-coded bars. (6) User-controlled color choices for phylogenetic and functional groups at different levels of specificity would aid understanding and presentation of results.

ETHICS STATEMENT

No human or animal subjects were involved in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

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. The Guaymas Basin project was funded by NSF OCE 0647633.

ACKNOWLEDGMENTS

Thanks to the captain and crews of R/V Atlantis and HOV Alvin for two enjoyable cruises to Guaymas Basin. Jennifer Biddle isolated the "Maribeggiatoa" filament sequenced and prepared DNA for sequencing. 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. 2016.01780/full#supplementary-material

- Flood, B. E., Fliss, P. S., Jones, D. S., Dick, G. J., Jain, S., Kaster, A. K. et al. (2016). Single-cell (meta-)genomics of a dimorphic Candidatus *Thiomargarita nelsonii* reveals genomic plasticity. *Front. Microbiol.* 7:603. doi: 10.3389/fmicb.2016.00603
- Fomenkov, A., Vincze, T., Grabovich, M. Y., Dubinina, G., Orlova, M., Belousova, E., et al. (2015). Complete genome sequence of the freshwater colorless sulfur bacterium *Beggiatoa leptomitiformis* neotype strain D-402(T). *Genome Announc.* 3, e01436–e01415. doi: 10.1128/genomeA.01436-15
- 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
- Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, et al. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res.* 32, 1363–1371. doi: 10.1093/nar/gkh293
- 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 singlefilament 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. N., 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
- 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, 1923–1937. doi: 10.1371/journal.pbio.00 50230
- Salazar, M. E., and Laub, M. T. (2015). Temporal and evolutionary dynamics of two-component signaling pathways. *Curr. Opin. Microbiol.* 24, 7–14. doi: 10.1016/j.mib.2014.12.003
- 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
- Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446
- 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

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.

Copyright © 2016 MacGregor. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.