JOURNAL OF BACTERIOLOGY, June 2011, p. 2745–2755 Vol. 193, No. 11
0021-9193/11/$12.00 doi:10.1128/JB.00040-11

Biological Systems Discovery *In Silico*: Radical *S*-Adenosylmethionine Protein Families and Their Target Peptides for Posttranslational Modification[▽][†]

Daniel H. Haft* and Malay Kumar Basu

J. Craig Venter Institute, Rockville, Maryland

Received 9 January 2011/Accepted 25 March 2011

Data mining methods in bioinformatics and comparative genomics commonly rely on working definitions of protein families from prior computation. Partial phylogenetic profiling (PPP), by contrast, optimizes family sizes during its searches for the cooccurring protein families that serve different roles in the same biological system. In a large-scale investigation of the incredibly diverse radical *S*-adenosylmethionine (SAM) enzyme superfamily, PPP aided in building a collection of 68 TIGRFAMs hidden Markov models (HMMs) that define nonoverlapping and functionally distinct subfamilies. Many identify radical SAM enzymes as molecular markers for multicomponent biological systems; HMMs defining their partner proteins also were constructed. Newly found systems include five groupings of protein families in which at least one marker is a radical SAM enzyme while another, encoded by an adjacent gene, is a short peptide predicted to be its substrate for posttranslational modification. The most prevalent, in over 125 genomes, featuring a peptide that we designate SCIFF (*s*ix *c*ysteines *i*n *f*orty-*f*ive residues), is conserved throughout the class *Clostridia*, a distribution inconsistent with putative bacteriocin activity. A second novel system features a tandem pair of putative peptide-modifying radical SAM enzymes associated with a highly divergent family of peptides in which the only clearly conserved feature is a run of His-Xaa-Ser repeats. A third system pairs a radical SAM domain peptide maturase with selenocysteine-containing targets, suggesting a new biological role for selenium. These and several additional novel maturases that cooccur with predicted target peptides share a C-terminal additional 4Fe4S-binding domain with PqqE, the subtilosin A maturase AlbA, and the predicted mycofactocin and Nif11-class peptide maturases as well as with activators of anaerobic sulfatases and quinohemoprotein amine dehydrogenases. Radical SAM enzymes with this additional domain, as detected by TIGR04085, significantly outnumber lantibiotic synthases and cyclodehydratases combined in reference genomes while being highly enriched for members whose apparent targets are small peptides. Interpretation of comparative genomics evidence suggests unexpected (nonbacteriocin) roles for natural products from several of these systems.

Bioactive small molecules generated by biosynthetic pathways operating in bacteria can profoundly influence the population structures of mixed microbial communities by serving as enzymatic cofactors, signaling molecules, or toxins. Excepting the lantibiotics, however, natural products containing unusual amino acids through modification of ribosomally translated precursor peptides frequently have been overlooked (23).

The radical *S*-adenosylmethionine (rSAM) domain family (30) is a large superfamily of proteins with diverse members that generate a radical species by reductive cleavage of SAM. All radical SAM proteins discussed in this paper belong to Pfam (9) family PF04055. A few radical SAM enzymes have long been known to modify peptides or proteins; PqqE cross-links a tyrosine to a glutamate in an intramolecular cyclization of PqqA as the first step in pyrroloquinoline quinone (PQQ) biosynthesis (35), and AlbA performs three intramolecular cyclizations from cysteine side chains to synthesize the antilisterial bacteriocin subtilosin A from its precursor (18). The methylthiotransferase RimO modifies ribosomal protein S12 (20), and several families activate cognate enzymes dependent on a glycyl radical active site. However, functional diversity within the family is so great (10) that mere classification as a radical SAM enzyme says very little about its molecular target or biological process. An extensive list of 68 nonoverlapping subgroups resolved by TIGRFAMs hidden Markov models (HMMs) within the radical SAM domain superfamily is summarized in Table S1 in the supplemental material. Among these, RlmN is a methyltransferase acting on 23S RNA (33), MiaB is a methylthiotransferase acting on tRNA (20), and SplB is a DNA repair enzyme, spore photoproduct lyase, that directly repairs thiamine dimers (27). Radical SAM enzymes for cofactor biosynthesis include biotin synthase BioB, tyrosine lyase ThiH (thiamine pyrophosphate), lipoyl synthase LipA (6), CofG and CofH (coenzyme F_{420}) (11), the coproporphyrinogen dehydrogenase HemN (heme) and NirJ (heme d1) (4), and two enzymes of menaquinone biosynthesis via futalosine (16). HydG and HydE act in metal cluster assembly in iron-only hydrogenases (25), NifB acts in nitrogenase metal cluster assembly, and HmdB acts in 5,10-methenyltetrahydromethanopterin hydrogenase metallocofactor biosynthesis. Additional characterized radical SAM families have roles in lipid metabolism, small-molecule transformations, and natural product biosynthesis.

While much is known about individual radical SAM en-

* Corresponding author. Mailing address: J. Craig Venter Institute, 9704 Medical Center Dr., Rockville, MD 20850. Phone: (301) 795-7952. Fax: (301) 795-7060. E-mail: haft@jcvi.org.

† Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 8 April 2011.

zymes, the family on the whole exposes the limits of legacy annotation available from public archives and of the performance of automated annotation pipelines currently in use. Most annotations are overly generic ("radical SAM domain protein" or "FeS oxidoreductases"), while specific functional assignments such as "coenzyme PQQ synthesis protein E" or "arylsulfatase regulator" (a misnomer for an anaerobic sulfatase maturase) have propagated incorrectly to numbers of homologs whose function, on inspection, clearly must differ. Many radical SAM-containing biological systems are sufficiently rare and sparsely distributed that simple clustering methods such as the COG (clusters of orthologous groups) algorithm based on bidirectional best hit linkages (31) necessarily lump together proteins that differ in function, hindering inference about their roles in biological systems. Clearly, a large-scale reexamination of subfamilies within the radical SAM superfamily, with considerations of molecular phylogenetic trees, genome contexts, and system reconstructions performed during protein family construction (29), would serve the scientific community well. Findings from such efforts are discussed here. The work has resulted in many new protein family definitions, included in TIGRFAMs release 10.1, both for the radical SAM families themselves and for the additional protein families that are their partners in the same biological systems.

A number of radical SAM enzymes involved in peptide modification show mutual sequence similarity in the region C terminal to the region described by Pfam model PF04055 (2, 12). PqqE and AlbA were discussed above. In *Streptococcus thermophilus*, a radical SAM enzyme (family TIGR04080) introduces a cyclization between amino acids 2 (Lys) and 6 (Trp) at the KxxxW motif in the peptide AKGDGWVKM to create the possible quorum sensing system molecule Pep1357C (17). Recently, we described two new classes of peptide-modifying radical SAM enzymes. Family TIGR03962 is a radical SAM family putative maturase for mycofactocin, whose precursor peptide shows incredible sequence conservation across dozens of species throughout a taxonomic range that includes many actinobacteria and several *Chloroflexi*, *Clostridia*, *Deltaproteobacteria*, and *Archaea* (12). Family TIGR04064 contains putative maturases (12) for ribosomally translated natural product precursors of the Nif11-like and nitrile hydratase-like leader peptide families (14). These precursors appear to share a cleavage and export system but assort promiscuously with different classes of maturases, including lantibiotic synthases, cyclodehydratases, and radical SAM enzymes. All these radical SAM enzymes known or presumed to act on peptide targets carry additional 4Fe-4S cluster-binding motifs that they share with anaerobic sulfatase-maturating enzymes (2). The emerging picture suggests that additional close homologs may also act on peptide precursors. It should be noted, however, that several other members of this subgroup of radical SAM enzymes with extended C-terminal homology act on substrates that do not have ribosomal origin. Exceptions include BtrN, involved in synthesizing butirosin, an aminoglycoside antibiotic (36), and NirJ from heme d1 biosynthesis.

A prevailing notion is that most natural products made in bacteria by posttranslational modification from ribosomally translated peptides are bacteriocins, peptide antibiotics able to kill rival bacteria (23). This view, though well supported for lantibiotics, may have set too narrow a focus in experimental approaches to other ribosomal natural products; we will use the term ribosomally translated natural product (RTNP) rather than "putative bacteriocin" in the remainder of the discussion. Pyrroloquinoline quinone (PQQ) is a ribosomally derived RTNP but is a redox cofactor. *In silico* analysis of the mycofactocin system shows that its signatures in comparative genomics analyses follow the same "bioinformatics grammar" as do cofactors such as PQQ (invariant residues in the propeptide region, cooccurrence and coclustering with paralogous sets of cofactor-dependent enzymes, and no exporter), rather than the grammar of bacteriocins (conservation mostly in the leader peptide, cooccurrence and coclustering with export transporters, tandem paralogs commonly observed), and so mycofactocin is predicted to be a novel redox factor (12). Meanwhile, the first methanobactin (a copper-binding metallophore) for which the structure is known is now shown to derive from a translated peptide (19). These reports contribute to recent expansions in our recognition of RTNPs (28), the roles of radical SAM enzymes in their syntheses, and possible novel metabolic roles for their products.

The functionally highly diverse radical SAM domain family represents approximately 0.5% of total proteins in anaerobic bacteria and many hundreds of different biological roles, and yet only a small number have been examined experimentally. Here, we have undertaken a broad study of the radical SAM family aimed at delineating subgroups where each approximates, as well as possible, the whole of a set of enzymes that share one particular function. If that enzyme happens to belong to a pathway that has multiple protein components, getting the granularity right for the radical SAM protein creates a mechanism through which phylogenetic profiling methods (15) help identify and build decision rules for recognizing those additional protein families that belong to the same system. Results from these analyses have identified a number of novel rSAM-containing genomic systems, described here, including a number with new RTNP precursor families. The interpretation of some of these radical SAM/RTNP systems suggests both new experimental strategies to look for microbial natural products and a broadened set of expectations for their possible biological roles.

MATERIALS AND METHODS

Phylogenetic profiling and protein family construction. Phylogenetic profiling studies were conducted using the ProPhylo system (M. J. Basu, J. D. Selengut, and D. H. Haft, 2010, available at ftp://ftp.jcvi.org/pub/data/ppp), with all-versus-all BLAST sequence comparison results based on a nonredundant collection of 1,466 complete and high-quality draft genomes. Phylogenetic profiles for existing hidden Markov models (HMMs), or candidate new families, were constructed by assigning the value 1 to species with a protein that scores above the TIGRFAMs (29) trusted cutoff (or estimated cutoff for a new candidate model) and the value 0 to species without such a protein. The query profile for the class *Clostridia* was created by assigning 1 to all children of node 186801 in the NCBI taxonomy tree and 0 to all other species in the reference genome set. Searches for protein families connected through biological processes were performed using partial phylogenetic profiling (PPP) (15) in ProPhylo, with manual inspection of PPP results.

Radical SAM proteins that did not score above the trusted cutoff value for any TIGRFAMs model were investigated in either of two ways. First, a collection of radical SAM domain-containing proteins was identified by match to Pfam model PF04055 and filtered of all members already covered by an existing TIGRFAMs HMM. A progressive alignment was constructed with Clustal W (32) and inspected manually to find sharply demarcated clusters with conserved protein architecture. Second, single radical SAM proteins were investigated across a range of depths with double-partial phylogenetic profiling (dPPP). dPPP is a special usage of PPP, provided by ProPhylo, that begins from a single protein and explores different depths in the list of best BLAST matches to that protein as a mechanism to produce a number of different query profiles. Each query profile is then used in turn. If a protein is one of several components in a subcellular system or pathway, an optimal depth may be found such that PPP based on that version of the query profile generates compelling statistical evidence for a correlated species distribution of the key proteins involved. Results produced in these analyses were inspected manually to find cohorts of proteins with PPP scores comparable to that of the query locus and yet well separated from the background. The background is taken to include most housekeeping proteins, as well as any large sets of proteins with very similar scores despite highly diverse functions and genomic locations. Once preliminary cutoffs defining sets of proteins for a novel protein family were determined, protein family construction and contribution to TIGRFAMs proceeded as previously described (29).

Validating conserved cohorts of protein families in novel systems *in silico*. Computational validation of suggested definitions for new biological systems was conducted by testing for the consistency of projections made for the systems across multiple genomes. Validation is considered successful if two or more different protein families (i) produce a sharp demarcation between the lowest scores for members and the highest scores for nonmembers, (ii) have almost perfectly matching sets of genomes with members of these families, and (iii) have largely different sets of genomes from each other in the set of the top-scoring sequences below the set cutoffs. That is, the putative system is not easily projected in its entirety onto additional species just by lowering cutoffs for all models. In most cases, the multiple genes encoding proteins from the same system tend to occur in clusters. Observing conserved gene neighborhoods provides additional validation for definitions of both whole systems and individual protein family definitions.

Because most of the systems described here feature at least one short peptide, often with an odd sequence composition, both gene-calling procedures that populate public databases with conceptual translations for predicted polypeptides and detection of such polypeptides by BLAST (1), as required in ProPhylo for PPP and dPPP, could miss important sequences. Correlations detected through PPP and dPPP were examined further manually, with downloading of genomic sequences and searching for missed gene calls performed as needed, by BLAST with translation of nucleic acid sequences or with HMM searches of six-frame translations.

Finding probable selenoproteins. Bacterial DNA regions of ~400 bp containing TGA sequences suspected to encode selenocysteine were searched for bacterial selenocysteine insertion sequences (SECIS) using the bSECISearch web server page http://genomics.unl.edu/bSECISearch (37), with default parameters. The regions searched always were large enough to include multiple additional candidate selenocysteine-encoding TGA sites, enough to show that bSECISearch rejects the vast majority of TGA sites that should not encode selenocysteine. Only one apparent false-positive SECIS was observed, for a TGA on the noncoding strand, among all sequences tested.

Hierarchical clustering. A consensus sequence for each of 54 nonoverlapping TIGRFAMs protein families was generated by the hmmemit program of HMMER 3 (8) from its hidden Markov model, which in turn was built by HMMER 3 from a curated seed alignment. The subtilosin A maturase AlbA sequence (which is unique, so that no TIGRFAMs HMM was constructed) was added to this collection, and all-versus-all sequence comparisons were performed with BLAST (1). Scores between sequences A and B were normalized to a value between 0 and 1 by dividing the log of the bit score by the log of the bit score of A versus itself or B versus itself, whichever was smaller. These scores were converted to distances by subtraction from 1 and then hierarchically clustered by the Ward method (34).

RESULTS

The SCIFF (*s*ix *C*ys *i*n *f*orty-*f*ive) system. Performing partial phylogenetic profiling to find which proteins best follow the property of belonging to the class *Clostridia* reveals that two of the best markers for the class are radical SAM proteins. This is not surprising, as many radical SAM enzymes are involved in modifications to tRNA or rRNA. The 16S RNA gene frequently is used as a basis of phylogenetic tree construction and taxonomic classification, and it would be reasonable to imagine an rRNA modification enzyme closely correlated with a lineage-specific character in the 16S RNA sequence. However, the radical SAM protein CPF_2198 from *Clostridium perfringens* ATCC 13124 is most similar to AlbA, PqqE, the Nif11 modification family TIGR04064, the CLI_3235 modification family TIGR04068 (see below), and the quinohemoprotein amine dehydrogenase maturation proteins of family TIGR03906. Family TIGR03974 was built to describe the distinctive clade of radical SAM proteins that includes CPF_2198.

Adjacent to CPF_2198 is CPF_2199, a 46-residue sequence with six cysteine residues in a very strongly conserved region of 23 amino acids, free of gaps, at positions 21 through 44. Partial phylogenetic profiling showed that every detectable homolog of CPF_2199 occurs next to a CPF_2198-related radical SAM protein. The relationship is reciprocal, except that this small protein was missed by gene calling software in 28 out of 128 genomes. Missed sequences, easily demonstrated by a tblastn search versus reference genomic sequences with the C-terminal consensus sequence GGCGECQTSCQSACKTSCTVGN QACE, showed no particular properties of sequence, species of origin, or local genomic context to explain why they were missed. Multiple sequence alignment for this family reveals an average length of about 45 residues, with the six cysteine residues each invariant or nearly so. The member from *Filifactor alocis* ATCC 35896 is unusual, with a nearly exact full-length tandem duplication. We designate the protein family SCIFF, for *s*ix *c*ysteines *i*n *f*orty-*f*ive residues. Figure 1 shows a multiple sequence alignment of the SCIFF protein family, which is described by TIGRFAMs model TIGR03973, and some genomic contexts that contain SCIFF system genes. The region N terminal to the first cysteine residue in SCIFF is much less strongly conserved (no invariant residues) than is the remainder (12 invariant residues), suggesting that the N-terminal region may be lost from the mature form. The small size and cysteine richness of the SCIFF protein suggest that it, like subtilosin A, is a target for posttranslational modification. Sequence similarities between the companion radical SAM family (TIGR03974) and previously known peptide-modifying radical SAM proteins support this claim.

The His-Xaa-Ser repeat/TIGR03977/TIGR03978 system. TIGRFAMs protein families TIGR03977 and TIGR03978 describe a tandem pair of radical SAM proteins that each show closer relationships of sequence similarity to peptide or protein modification enzymes than they show to any other characterized radical SAM proteins. TIGR03978 includes a region with the additional cysteine-rich motifs for 4Fe-4S cluster binding, resembling the motifs described by Benjdia (2). A majority of the radical SAM gene pairs from this system (25 of 36) are accompanied by a member of family TIGR03979, which is relatively poorly conserved except for an unusual region with five to seven tandem repeats of the motif His-Xaa-Ser, where Xaa usually is Arg, Ser, or Tyr. The C-terminal domain of many of these His-Xaa-Ser repeat proteins contains a peptidoglycan-binding domain that is shared by a number of enzymes involved in bacterial cell wall degradation, as described by PF01471. The His-Xaa-Ser motif is reminiscent of the variable putative bacteriocin widely distributed among strains of *Bacillus anthracis* and *Bacillus cereus* (13), with a Cys-Xaa-Xaa repeat. In thiazole/oxazole-modified microcin (TOMM) systems (21), multiple modifications to TOMM precursors with

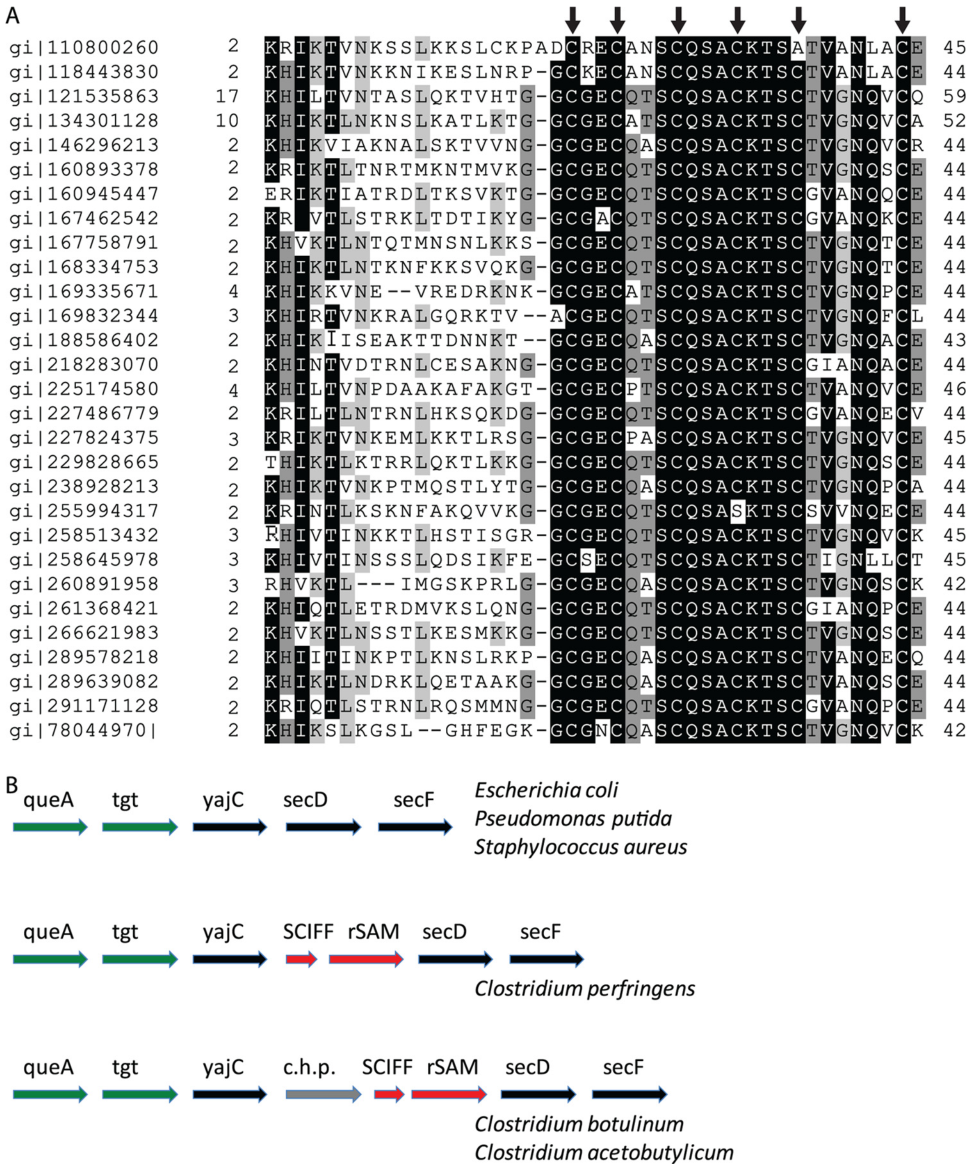
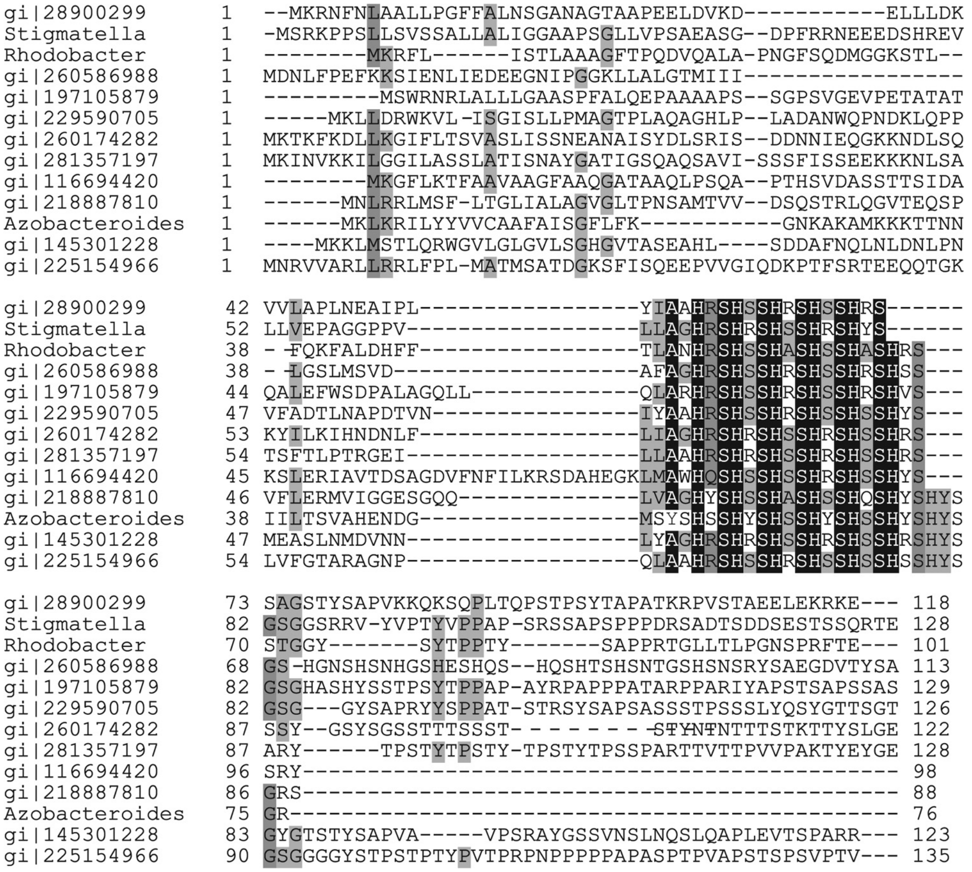

A
gi|110800260 2 KRIKTVNKSSLKKSLCKPADCRECANSCQSACKTSATVANLACE 45
gi|118443830 2 KHIKTVNKKNIKESLNRP-GCKECANSCQSACKTSCTVANLACE 44
gi|121535863 17 KHILTVNTASLQKTVHTG-GCGECQTSCQSACKTSCTVGNQVCQ 59
gi|134301128 10 KHIKTLNKNSLKATLKTG-GCGECATSCQSACKTSCTVGNQVCA 52
gi|146296213 2 KHIKVIAKNALSKTVVNG-GCGECQASCQSACKTSCTVGNQVCR 44
gi|160893378 2 KRIKTLTNRTMKNTMVKG-GCGECQTSCQSACKTSCTVGNQSCE 44
gi|160945447 2 ERIKTIATRDLTKSVKTG-GCGECQTSCQSACKTSCGVANQQCE 44
gi|167462542 2 KR VTLSTRKLTDTIKYG-GCGACQTSCQSACKTSCGVANQKCE 44
gi|167758791 2 KHVKTLNTQTMNSNLKKS-GCGECQTSCQSACKTSCTVGNQTCE 44
gi|168334753 2 KHIKTLNTKNFKKSVQKG-GCGECQTSCQSACKTSCTVGNQTCE 44
gi|169335671 4 KHIKKVNE--VREDRKNK-GCGECATSCQSACKTSCTVGNQPCE 44
gi|169832344 3 KHIRTVNKRALGQRKTV--ACGECQTSCQSACKTSCTVGNQFCL 44
gi|188586402 2 KHIKIISEAKTTDNNKT--GCGECQASCQSACKTSCTVGNQACE 43
gi|218283070 2 KHINTVDTRNLCESAKNG-GCGECQTSCQSACKTSCGIANQACE 44
gi|225174580 4 KHILTVNPDAAKAFAKGT-GCGECPTSCQSACKTSCTVANQVCE 46
gi|227486779 2 KRILTLNTRNLHKSQKDG-GCGECQTSCQSACKTSCGVANQECV 44
gi|227824375 3 KRIKTVNKEMLKKTLRSG-GCGECPASCQSACKTSCTVGNQVCE 45
gi|229828665 2 THIKTLKTRRLQKTLKKG-GCGECQTSCQSACKTSCTVGNQSCE 44
gi|238928213 2 KHIKTVNKPTMQSTLYTG-GCGECQASCQSACKTSCTVGNQPCA 44
gi|255994317 2 KRINTLKSKNFAKQVVKG-GCGECQTSCQSASKTSCSVVNQECE 44
gi|258513432 3 RHIVTINKKTLHSTISGR-GCGECQTSCQSACKTSCTVGNQVCK 45
gi|258645978 3 KHIVTINSSSLQDSIKFE-GCSECQTSCQSACKTSCTIGNLLCT 45
gi|260891958 3 RHVKTL---IMGSKPRLG-GCGECQASCQSACKTSCTVGNQSCK 42
gi|261368421 2 KHIQTLETRDMVKSLQNG-GCGECQTSCQSACKTSCGIANQPCE 44
gi|266621983 2 KHVKTLNSSTLKESMKKG-GCGECQTSCQSACKTSCTVGNQSCE 44
gi|289578218 2 KHIITINKPTLKNSLRKP-GCGECQASCQSACKTSCTVANQECQ 44
gi|289639082 2 KHIKTLNDRKLQETAAKG-GCGECQASCQSACKTSCTVANQSCE 44
gi|291171128 2 KRIQTLSTRNLRQSMMNG-GCGECQTSCQSACKTSCGVANQPCE 44
gi|78044970| 2 KHIKSLKGSL--GHFEGK-GCGNCQASCQSACKTSCTVGNQVCK 42
B
queA
tgt
yajC
secD
secF
Escherichia coli
Pseudomonas putida
Staphylococcus aureus
queA
tgt
yajC
SCIFF
rSAM
secD
secF
Clostridium perfringens
queA
tgt
yajC
c.h.p.
SCIFF
rSAM
secD
secF
Clostridium botulinum
Clostridium acetobutylicum

FIG. 1. The SCIFF system multiple sequence alignment and genomic regions. (A) The TIGRFAMs seed alignment TIGR03973 for the SCIFF (*s*ix *c*ysteines *i*n *f*orty-*f*ive residues) protein is shown shaded according to the degree of sequence identity in each column. Sequences more than 80% identical were removed. The six cysteines that are universal or nearly so are indicated with arrows. A run of 10 residues, SCQSACKTSC, is invariant except for two sequences with one conservative substitution each. The first, third, fourth, and fifth cysteines are flanked on one or both sides by amino acids with small side chains (Gly, Ser, or Ala), as is common for posttranslational modifications that cross-link cysteines to other residues during peptide maturation. The species of origin for the sequences shown, in order from top to bottom, are *Clostridium perfringens* ATCC 13124, *Clostridium novyi* NT, *Thermosinus carboxydivorans* Nor1, *Desulfotomaculum reducens* MI-1, *Caldicellulosiruptor saccharolyticus* DSM 8903, *Clostridium* sp. strain L2-50, *Faecalibacterium prausnitzii* M21/2, *Paenibacillus larvae* subsp. *larvae* BRL-230010, *Clostridium scindens* ATCC 35704, *Epulopiscium* sp. 'N.t. morphotype B', *Anaerofustis stercorihominis* DSM 17244, "*Candidatus* Desulforudis audaxviator" MP104C, *Natranaerobius thermophilus* JW/NM-WN-LF, *Eubacterium biforme* DSM 3989, *Dethiobacter alkaliphilus* AHT 1, *Anaerococcus lactolyticus* ATCC 51172, *Acidaminococcus* sp. strain D21, *Shuttleworthia satelles* DSM 14600, *Selenomonas flueggei* ATCC 43531, *Eubacterium saphenum* ATCC 49989, *Desulfotomaculum acetoxidans* DSM 771, *Dialister invisus* DSM 15470, *Ammonifex degensii* KC4, *Subdoligranulum variabile* DSM 15176, *Clostridium hathewayi* DSM 13479, *Thermoanaerobacter italicus* Ab9, *Ethanoligenens harbinense* YUAN-3, *Filifactor alocis* ATCC 35896, and *Carboxydothermus hydrogenoformans* Z-2901. (B) Genome region figure showing the SCIFF precursor and its maturase (red) appearing in a housekeeping gene context with the queuosine tRNA modification genes *queA* and *tgt* (green) and the Sec system subunit genes *yajC*, *secD*, and *secF* (black). In some species, an additional conserved hypothetical protein (c.h.p.) is also present (gray).

repetitive, low-complexity sequence are performed by a single cyclodehydratase. The radical SAM enzymes in the His-Xaa-Ser repeat system might behave similarly. Strikingly, an HMM built from the repeating His-Xaa-Ser region from the seed alignment of TIGR03979, flanked by only five residues on one side, three on the other, always detects a His-Xaa-Ser repeat-containing protein that lies next to the radical SAM proteins from this system, and these proteins include examples lacking



```
gi|28900299     1   --MKRNFNLAALLPGFFALNSGANAGTAAPEELDVKD-----------ELLLDK
Stigmatella     1   -MSRKPPSLLSVSSALLALIGGAAPSGLLVPSAEASG--DPFRRNEEEDSHREV
Rhodobacter     1   --------MKRFL------ISTLAAAGFTPQDVQALA-PNGFSQDMGGKSTL--
gi|260586988    1   MDNLFPEFKKSIENLIEDEEGNIPGGKLLALGTMIII-----------------
gi|197105879    1   ---------MSWRNRLALLLGAASPFALQEPAAAAPS--SGPSVGEVPETATAT
gi|229590705    1   -----MKLLDRWKVL-ISGISLLPMAGTPLAQAGHLP--LADANWQPNDKLQPP
gi|260174282    1   MKTKFKDLLKGIFLTSVASLISSNEANAISYDLSRIS--DDNNIEQGKKNDLSQ
gi|281357197    1   MKINVKKILGGILASSLATISNAYGATIGSQAQSAVI-SSSFISSEEKKKNLSA
gi|116694420    1   --------MKGFLKTFAAVAAGFAAQGATAAQLPSQA--PTHSVDASSTTSIDA
gi|218887810    1   ------MNLRRLMSF-LTGLIALAGVGLTPNSAMTVV--DSQSTRLQGVTEQSP
Azobacteroides  1   ------MKLKRILYYVVCAAFAISGFLFK-----------GNKAKAMKKKTTNN
gi|145301228    1   ----MKKLMSTLQRWGVLGLGVLSGHGVTASEAHL----SDDAFNQLNLDNLPN
gi|225154966    1   MNRVVARLLRRLFPL-MATMSATDGKSFISQEEPVVGIQDKPTFSRTEEQQTGK

gi|28900299    42   VVLAPLNEAIPL-----------------YIAAHRSHSSHRSHSSHRS------
Stigmatella    52   LLVEPAGGPPV------------------LLAGHRSHRSHSSHRSHYS------
Rhodobacter    38   --FQKFALDHFF-----------------TLANHRSHSSHASHSSHASHRS---
gi|260586988   38   --LGSLMSVD-------------------AFAGHRSHSSHRSHSSHRSHSS---
gi|197105879   44   QALEFWSDPALAGQLL-------------QLARHRSHSSHRSHSSHRSHVS---
gi|229590705   47   VFADTLNAPDTVN----------------IYAAHRSHSSHRSHSSHSSHYS---
gi|260174282   53   KYILKIHNDNLF-----------------LIAGHRSHRSHSSHRSHSSHRS---
gi|281357197   54   TSFTLPTRGEI------------------LLAAHRSHRSHSSHRSHSSHRS---
gi|116694420   45   KSLERIAVTDSAGDVFNFILKRSDAHEGKLMAWHQSHSSHSSHRSHSSHYS---
gi|218887810   46   VFLERMVIGGESGQQ--------------LVAGHYSHSSHASHSSHQSHYSHYS
Azobacteroides 38   IILTSVAHENDG-----------------MSYSHSSHYSHSSHYSHSSHYSHYS
gi|145301228   47   MEASLNMDVNN------------------LYAGHRSHRSHSSHRSHSSHRSHYS
gi|225154966   54   LVFGTARAGNP------------------QLAAHRSHSSHRSHSSHSSHSSHYS

gi|28900299    73   SAGSTYSAPVKKQKSQPLTQPSTPSYTAPATKRPVSTAEELEKRKE--- 118
Stigmatella    82   GSGGSRRV-YVPTYVPPAP-SRSSAPSPPPDRSADTSDDSESTSSQRTE 128
Rhodobacter    70   STGGY-------SYTPPTY-------SAPPRTGLLTLPGNSPRFTE--- 101
gi|260586988   68   GS-HGNSHSNHGSHESHQS--HQSHTSHSNTGSHSNSRYSAEGDVTYSA 113
gi|197105879   82   GSGHASHYSSTPSYTPPAP-AYRPAPPPATARPPARIYAPSTSAPSSAS 129
gi|229590705   82   GSG---GYSAPRYYSPPAT-STRSYSAPSASSSTPSSSLYQSYGTTSGT 126
gi|260174282   87   SSY--GSYSGSSTTTSSST-----------STYNTNTTTSTKTTYSLGE 122
gi|281357197   87   ARY------TPSTYTPSTY-TPSTYTPSSPARTTVTTPVVPAKTYEYGE 128
gi|116694420   96   SRY---------------------------------------------- 98
gi|218887810   86   GRS---------------------------------------------- 88
Azobacteroides 75   GR----------------------------------------------- 76
gi|145301228   83   GYGTSTYSAPVA-----VPSRAYGSSVNSLNQSLQAPLEVTSPARR--- 123
gi|225154966   90   GSGGGGYSTPSTPTYPVTPRPNPPPPPAPASPTPVAPSTSPSVPTV--- 135
```

FIG. 2. Multiple sequence alignment of His-Xaa-Ser proteins. Sequences were aligned by MUSCLE and minimally hand edited at sites from the first His-Xaa-Ser repeat to the C terminus. The three shortest sequences are shown at their full lengths, although others have additional C-terminal sequence not shown. Three sequences, identified by genus names, were not previously identified as protein-coding features. The sequences shown, in order from top to bottom, are from *Vibrio parahaemolyticus* RIMD 2210633, *Stigmatella aurantiaca* DW4/3-1, *Rhodobacter* sp. strain SW2, *Blautia hansenii* DSM 20583, *Phenylobacterium zucineum* HLK1, *Pseudomonas fluorescens* SBW25, *Bacteroides* sp. strain D2, *Victivallis vadensis* ATCC BAA-548, *Ralstonia eutropha* H16, *Desulfovibrio vulgaris* strain Miyazaki F, "*Candidatus* Azobacteroides pseudotrichonymphae" genomovar CFP2, *Aeromonas salmonicida* subsp. *salmonicida* A449, and *Opitutaceae* bacterium TAV2. Member sequences occur in close proximity to paired radical SAM enzymes, one each from families TIGR03977 and TIGR03978.

significant sequence similarity to family TIGR03979 anywhere outside the repeat region. An alignment of this expanded collection of His-Xaa-Ser proteins, after removal of redundant sequences, is shown in Fig. 2. In some cases, the corresponding open reading frame never was called as a gene in the deposited reference genome (e.g., *Stigmatella aurantiaca* DW4/3-1, *Bacteroides caccae* ATCC 43185, and *Rhodobacter* sp. strain SW2). Sequence similarity clearly is minimal outside the repeat region. The combination of a His-Xaa-Ser repeat peptide with the TIGR03977/TIGR03978 radical SAM gene pair occurs across an extremely broad taxonomic range, including *Bacteroidetes*; *Firmicutes*; the alpha, beta, gamma, and delta divisions of the *Proteobacteria*; and *Verrucomicrobia*. An additional protein family that occurs only in this context is TIGR03976. In one branch of this family, proteins are ~90 residues in length, share a near-invariant LLNDYxLRE motif, and with default parameters in PSI-BLAST converge after one round to find only members of this branch. However, PPP finds additional family members from other branches, also coclustered with the His-Xaa-Ser proteins and radical SAM pair. PSI-BLAST originating from these alternate starting points converges eventually to include the full set of proteins now modeled in TIGR03976. This family completes the definition of a conserved four-gene cassette.

A selenocysteine-containing radical SAM-modified peptide. For the radical SAM protein family TIGR04082, the most similar characterized protein once again is the bacteriocin-maturing radical SAM enzyme AlbA, followed by PqqE. This family includes GSU_1560 from *Geobacter sulfurreducens* PCA, next to tandem very short open reading frames translated as GSU_1558 and GSU_1559. Examination of several homologs to GSU_1558 showed examples of somewhat longer peptides with a Cys residue aligned with the predicted GSU_1558 stop codon. This finding suggests selenocysteine incorporation, since *G. sulfurreducens* contains a selenocysteine incorporation system. A search for selenocysteine incorporation (SECIS) elements in the GSU_1558 region showed that the immediate UGA stop codon and the next in-frame stop codon to follow, also UGA, were followed by SECIS elements homologous to each other. Interpretation of these two UGA codons as selenocysteines extends GSU_1558 through GSU_1559, which is in the same frame. TIGRFAMs

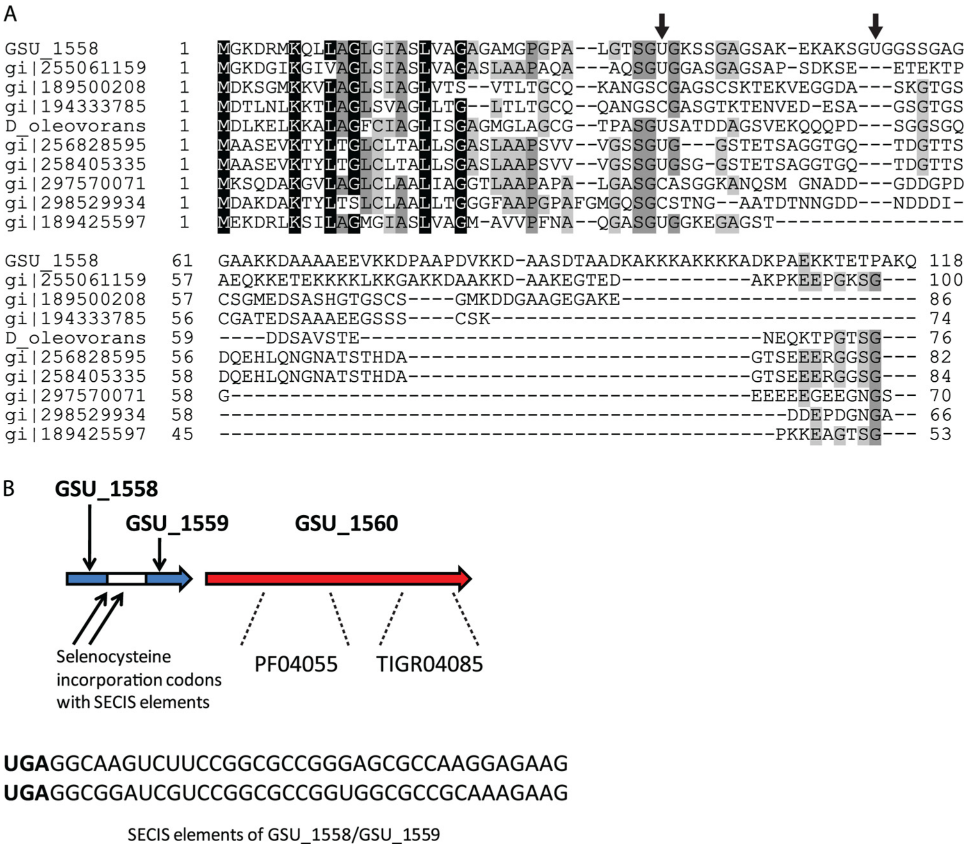

A
GSU_1558 1 MGKDRMKQLLAGLGIASLVAGAGAMGPGPA--LGTSGUGKSSGAGSAK-EKAKSGUGGSSGAG
gi|255061159 1 MGKDGIKGIVAGLSIASLVAGASLAAPAQA---AQSGUGGASGAGSAP-SDKSE---ETEKTP
gi|189500208 1 MDKSGMKKVLAGLSIAGLVTS--VTLTGCQ--KANGSCGAGSCSKTEKVEGGDA---SKGTGS
gi|194333785 1 MDTLNLKKTLAGLSVAGLLTG--LTLTGCQ--QANGSCGASGTKTENVED-ESA---GSGTGS
D_oleovorans 1 MDLKELKKALAGFCIAGLISGAGMGLAGCG--TPASGUSATDDAGSVEKQQQPD---SGGSGQ
gi|256828595 1 MAASEVKTYLTGLCLTALLSGASLAAPSVV--VGSSGUG---GSTETSAGGTGQ---TDGTTS
gi|258405335 1 MAASEVKTYLTGLCLTALLSGASLAAPSVV--VGSSGUGSG-GSTETSAGGTGQ---TDGTTS
gi|297570071 1 MKSQDAKGVLAGLCLAALIAGGTLAAPAPA--LGASGCASGGKANQSM GNADD---GDDGPD
gi|298529934 1 MDAKDAKTYLTSLCLAALLTGGGFAAPGPAFGMGQSGCSTNG--AATDTNNGDD---NDDDI-
gi|189425597 1 MEKDRLKSILAGMGIASLVAGM-AVVPFNA--QGASGUGGKEGAGST---------------
GSU_1558 61 GAAKKDAAAAEEVKKDPAAPDVKKD-AASDTAADKAKKKAKKKKADKPAEKKTETPAKQ 118
gi|255061159 57 AEQKKETEKKKKLKKGAKKDAAKKD-AAKEGTED-----------AKPKEEPGKSG--- 100
gi|189500208 57 CSGMEDSASHGTGSCS----GMKDDGAAGEGAKE------------------------- 86
gi|194333785 56 CGATEDSAAAEEGSSS----CSK------------------------------------ 74
D_oleovorans 59 ----DDSAVSTE--------------------------------NEQKTPGTSG--- 76
gi|256828595 56 DQEHLQNGNATSTHDA------------------------------GTSEEERGGSG--- 82
gi|258405335 58 DQEHLQNGNATSTHDA------------------------------GTSEEERGGSG--- 84
gi|297570071 58 G---EEEEEGEEGNGS-- 70
gi|298529934 58 --DDEPDGNGA-- 66
gi|189425597 45 ---PKKEAGTSG--- 53
B
GSU_1558
GSU_1559
GSU_1560
Selenocysteine incorporation codons with SECIS elements
PF04055
TIGR04085
UGAGGCAAGUCUUCCGGCGCCGGGAGCGCCAAGGAGAAG
UGAGGCGGAUCGUCCGGCGCCGGUGGCGCCGCAAAGAAG
SECIS elements of GSU_1558/GSU_1559

FIG. 3. Multiple sequence alignment and genomic region view of selenobacteriocin precursor peptides. (A) Multiple alignment. The letter U represents UGA (normally a stop) codon at the start of a bacterial selenocysteine insertion element (SECIS) translated as selenocysteine (SeCys), the 21st amino acid. The two alignment columns that contain at least one U are indicated with arrows; all non-SeCys residues in those columns are Cys. Model TIGR04081 describes sequences up to the column immediately past the first selenocysteine-containing column. Sequences, in order from top to bottom, include putative (seleno)bacteriocins from *Geobacter sulfurreducens* PCA (extended), *Geobacter* sp. strain M18 (extended), *Chlorobium phaeobacteroides* BS1, *Prosthecochloris aestuarii* DSM 271, *Desulfococcus oleovorans* Hxd3 (no gene shown in GenBank), *Desulfomicrobium baculatum* DSM 4028 (extended), *Desulfohalobium retbaense* DSM 5692 (extended), *Desulfurivibrio alkaliphilus* AHT2, *Desulfonatronospira thiodismutans* ASO3-1, and *Geobacter lovleyi* SZ (extended). (B) Corrected genomic region for the GSU_1558/GSU_1559 and GSU_1560 genes from *Geobacter sulfurreducens* PCA. Diagonal arrows indicate the positions of the two predicted SeCys residues. Underneath the arrow diagram are the identified selenocysteine insertion elements, or SECIS. The SECIS elements begin with UGA codons that are translated as SeCys and are 80% identical through their first 30 bases.

model TIGR04081 describes the conserved region shared by all homologs of GSU_1558 up to the first stop codon now reinterpreted as selenocysteine. For every additional homolog detected that stopped short like the originally reported GSU_1558, rather than continuing through 40 or so residues of additional low-complexity sequence, the genome of the species of origin encoded a selenocysteine incorporation system, the stop codon was UGA, and we detected a SECIS (selenocysteine insertion) element (37) specifically at that UGA site. These additional species included *Desulfococcus oleovorans* Hxd3, *Desulfohalobium retbaense* DSM 5692, and *Geobacter* sp. strain M18. The multiple alignment presented in Fig. 3 shows that predicted selenocysteine residues from previously truncated sequences align with cysteine residues, just as occurs when selenoproteins belong to families of redox enzymes. This strict concurrence of predicted selenocysteine incorporation sites with cysteine residue positions in other homologs means that all requirements for identifying a new family of selenoproteins are satisfied. This family is odd among selenocysteine-containing protein families in that it is likely not an enzyme. Beyond the cysteine/selenocysteine position, sequences continue but become low complexity and hypervariable, as often observed for the propeptide regions of lantibiotic precursors and thiazole/oxazole-modified microcin precursors. All findings are consistent with the interpretation that TIGR04081 describes the conserved N-terminal region of a novel, selenocysteine-containing family of RTNP precursors. There appears to be no previous report of a putative selenobacteriocin, nor of any ribosomal peptide natural product precursor translated with a noncanonical amino acid (selenocysteine or pyrrolysine) before it undergoes additional posttranslational modification (23, 38).

The CLI_3235-type system. TIGR04068 describes another radical SAM protein family. Among all radical SAM enzymes with experimental demonstrations of function or biological process, TIGR04068 is again most similar to known peptide-modifying radical SAM enzymes. It occurs regularly in combination with small peptides in the family of CLI_3235 from

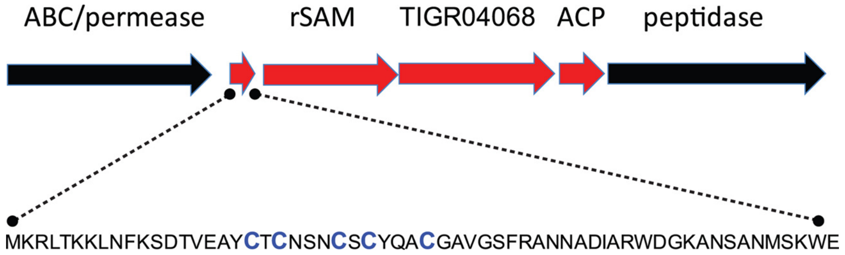

ABC/permease
rSAM
TIGR04068
ACP
peptidase
MKRLTKKLNFKSDTVEAYCTCNSNCSCYQACGAVGSFRANNADIARWDGKANSANMSKWE

FIG. 4. A ribosomal peptide natural product cassette in *Clostridium botulinum* A2 Kyoto. This six-gene cluster for a CLI_3235-type system shows two genes for which models were not constructed at the left and right (black). At the left is a transporter, CLM_3254, with both the ATP-binding and permease domains of ABC transporters. At the right is a peptidase, CLM_3249. The central four genes (red) are a locally cysteine-rich putative RTNP precursor of family TIGR04065, a radical SAM enzyme of family TIGR04068, a conserved hypothetical protein described by family TIGR04066, and an acyl carrier protein homolog described by TIGR04069. The related cassette in *Clostridium botulinum* F Langeland contains the same genes in the same order (plus one additional gene). Cassettes are flanked by unrelated genes in the genomes of these two *C. botulinum* strains.

Clostridium botulinum F Langeland (Fig. 4). Exploring sequence relationships among these small peptides revealed a family whose members are about 60 amino acids in length that is fairly well conserved in the N-terminal region while Cys rich and otherwise divergent in both length and sequence in the C-terminal region. Two rather divergent branches of the family of CLI_3235-like small, Cys-rich precursor peptides are described by models TIGR04065 and TIGR04067. Many of these radical SAM/RTNP gene pairs occur along with a gene for an additional protein that belongs to the acyl carrier protein (ACP) family and yet, strikingly, lacks the conserved Ser residue that serves as an acyl group attachment site. This novel ACP-like family found in putative RTNP maturation cassettes is described by TIGR04069. These regions also include predicted export ABC transporters such as CLI_3236.

The TIGR03913/Y_X(10)_GDL system. Yet another member of the PqqE-like radical SAM protein families is TIGR03913. This member, too, appears to be a candidate peptide-modifying enzyme. It tends to occur next to trios of tandem genes, members of family PF08898, that share a conserved C-terminal region of about 60 residues. The unusual arrangement of this trio of small proteins, mutually closely homologous to each other but not to any known enzyme or structural protein, next to a PqqE-like radical SAM protein, suggests again a cognate relationship between a radical SAM enzyme and its substrate(s). The occurrence of short homologous polypeptides next to a probable maturase cassette recalls examples of two-chain bacteriocins such as thuricin CD (26). Although the evidence from genome context and C-terminal sequence similarity to PqqE and the C-2x-C-5x-C-3x-C motif (2) is relatively weak evidence for peptide modification, the prediction becomes somewhat stronger in the context of similarity to the other PqqE-like radical SAM proteins described by us recently (12) and in this paper.

A set of 68 subfamilies defined in TIGRFAMs, within the radical SAM domain family defined by Pfam model PF04055, is presented in Table S1 in the supplemental material. The families described in detail above differ from most of the remaining families in Table S1, involved in processes such as RNA modification, cofactor biosynthesis (other than PQQ), lipid metabolism, enzyme activation, etc., in that those described in this article have an identified, family-specific cognate small peptide encoded nearby. Most of these new families display greater sequence similarities to each other and to experimentally characterized enzymes of subtilosin A biosynthesis and PQQ biosynthesis than to the majority of radical SAM enzymes known to act on nonproteinaceous substrates, as for RNA modification or biotin biosynthesis biosynthesis. In Table S1, models designated equivalog (or hypothetical-equivalog) describe families in which all members should be conserved in one specific function since their most recent common ancestor, while subfamily models are broader and may fully contain one or more equivalog families.

Figure S1 in the supplemental material shows the results of a hierarchical clustering of consensus sequences from over 50 nonoverlapping protein families with particularly good assignments to distinct biological processes. The results show considerable similarities within several functional subsets of radical SAM enzymes. One of these is RNA modification. Another, however, is peptide-modifying radical SAM enzymes together with anaerobic sulfatase activators and NirJ family proteins, and this block marks the best opportunity for prospecting for new peptide modification systems.

Comparative abundance of peptide modification enzymes. Table 1 shows the number of known or putative peptide modification enzymes identified by the HMMs described in this paper, found in a collection of 1,466 prokaryotic reference genomes. The genome set was filtered to reduce the numbers of near-identical genomes. In this set of genomes, 1,892 radical SAM enzymes were found with C-terminal extended sequence matched by TIGR04085. Of these, 269 are assigned as anaerobic sulfatase activators and 136 as NirJ-like proteins for heme d1 biosynthesis, meaning that they act on globular proteins, or in cofactor biosynthesis, rather than in peptide modification, while 413 enzymes are assigned to peptide maturase processes.

These classifications leave over 1,000 TIGR04085 domain-containing radical SAM enzymes from the set of reference genomes unassigned. Among these, additional peptide-modifying radical SAM enzymes are likely to be found. *Faecalibacterium prausnitzii* is a human gut bacterium whose numbers are significantly reduced in Crohn's disease (24). Only two proteins with domain TIGR04085 occur in human microbiome reference strain M21/2, representing 2 of its 16 total radical SAM enzymes. One is the SCIFF system putative maturase. The other is gi|160944046. At a distance of just 94 bp on the same strand is a missed reading frame for a 267-amino-acid protein,

TABLE 1. Counts of peptide modification enzymes for various classes of target peptide

Precursor family	Enzyme class	HMM(s)	Count (in 1,466 reference genomes)	Reference
Subtilosin A	rSAM + TIGR04085	None	1	18
YydG	rSAM	TIGR04078	1	5
KxxxW	rSAM + TIGR04085	TIGR04080	1	17
Nif11 leader	rSAM + TIGR04085	TIGR04064	5	12
CLI_3235	rSAM + TIGR04085	TIGR04068	11	This work
Y_X(10)_GDL	rSAM + TIGR04085	TIGR03913	12	This work
Selenopeptides	rSAM + TIGR04085	TIGR04082	13	This work
HxS repeats	rSAM + TIGR04085	TIGR03978	36	This work
Mycofactocin	rSAM + TIGR04085	TIGR03962	42	12
SCIFF	rSAM + TIGR04085	TIGR03974	130	This work
All thiazole/oxazole	Cyclodehydratase	TIGR03603, TIGR03882	146	21
PqqA	rSAM + TIGR04085	TIGR02109	161	35
All lantibiotics	LanB/C or LanM	PF04738, PF05147	478	3

in which the last 154 amino acids contain 38 evenly spaced cysteines, including 30 pairs spaced CxxxC. The cysteine richness resembles those of the SCIFF and CLI_3235-type system target peptides, while the periodicity recalls the His-Xaa-Ser system. Similar arrangements of PF04055/TIGR04085 radical SAM enzymes next to CxxxC-repeat peptides exist in other bacterial reference genomes, such as *Ruminococcus* sp. strain 5_1_39BFAA. Because the identification so far of radical SAM enzymes that act on peptide targets clearly has not been exhaustive, it seems likely that peptide targets of radical SAM enzymes in aggregate are at least as abundant as lantibiotic synthase or cyclodehydratase targets.

DISCUSSION

We undertook an analysis of the radical SAM family by using phylogenetic profiling approaches, in which codistributed protein families each provide information to guide the proper construction of the other. The results of this analysis included several apparent discoveries of new peptide modification systems. Because the comparative genomics methods require multiple copies of a system to exist among the collected 1,466 reference genomes analyzed, we did not attempt to study systems with rarities comparable to those of the subtilosin A, YydG, and KxxxW systems. Therefore, there may be many additional undiscovered radical SAM-mediated peptide modification systems. Those that we did identify, however, featured radical SAM enzymes with considerable mutual sequence similarity C terminal to the region covered by Pfam model PF04055. This additional region always contained a Cys-rich motif for an additional 4Fe-4S binding site, either as described previously (2) or in modified form. We constructed an HMM, TIGR04085, that readily identifies a branch of the radical SAM domain superfamily that appears highly enriched in peptide- and protein-modifying enzymes. The combined signature PF04055 plus TIGR04085, for a protein that is neither NirJ nor an anaerobic sulfatase maturase, marks a protein as a candidate peptide maturase. The often very small genes that encode ribosomal natural product precursors are easily overlooked; the identification of a new good marker for modified peptide precursors will aid in the detection of additional natural product biosynthesis systems.

Several of our newly described peptide modification systems show very different kinds of signatures in comparative genomics analyses than are typical among known bacteriocin production systems. To provide an interpretation of the SCIFF system, we examined apparent features from its “bioinformatics grammar.” We introduced this approach previously when we showed that the mycofactocin system exhibited a number of signatures more consistent with a role as a molecular cofactor or redox carrier than a role as a bacteriocin (12). Biological systems that differ entirely in their makeup, such that no protein from the first system shows any sequence similarity to any component of the second, may obey similar sets of constraints if they perform similar roles, such as both producing bacteriocins or both producing a cofactor. For the different types of systems in which the core feature marking the system is a peptide maturase next to a target peptide, it is possible to identify additional aspects of its bioinformatics grammar as an aid to making well-formed hypotheses about possible biological roles.

Features in the grammars that distinguish bacteriocins from cofactor biosynthesis systems are summarized in Table 2. Bacteriocins must be exported, while cofactors usually remain inside the cell. Consequently, bacteriocin biosynthesis loci typically are flanked by transporter genes. Bacteriocin families evolve, presumably, under strong positive selection, such that the propeptide region typically shows greater sequence divergence than the leader peptide. There may be several paralogous target peptides in the genome together with a single maturase. A polypeptide serving as a cofactor precursor, by contrast, will be encoded by a single-copy gene and will exhibit several invariant residues, such as the Glu and Tyr that are cross-linked in the first step in the biosynthesis of PQQ. Patterns of phylogenetic distribution in collections of hundreds to thousands of reference genomes clearly encode key clues to a system’s role and contrast sharply between the SCIFF system (virtually universal in *Clostridia*) and the His-Xaa-Ser system (sporadically distributed and regularly surrounded by markers of transposition and DNA integration). The SCIFF system is missing from only one complete genome classified as *Clostridia* (*Halothermothrix orenii* H 168) and two draft genomes, which makes it substantially better conserved than endospore formation, for example. It occurs in just three other *Firmicutes*, plus two species classified outside the *Firmicutes* by the current NCBI taxonomy tree: *Bacteroides capillosus* ATCC 29799 and

Bacteroides pectinophilus ATCC 43243. But these two genomes have numerous markers for endospore formation shared with low-GC Gram-positive sporeformers, have no markers of outer membranes, have closest matches of housekeeping proteins to other *Clostridia*, and clearly are misnamed and misclassified. Unlike typical cofactor biosynthesis systems, the SCIFF gene pair does not show coclustering or cooccurrence with paralogous families of cofactor-dependent enzymes, nor with any transcription factor. Instead, it has a tendency to appear in genomes next to the universal *secD* gene. In fact, while the genes for the YajC and SecD subunits of the Sec complex are adjacent in species as widely separated as *Escherichia coli*, *Mycobacterium tuberculosis*, and *Staphylococcus aureus*, the SCIFF gene pair occurs near *secD* in more than 50 species and in several species occurs between *yajC* and *secD* with no other intervening genes. This unusual location suggests constitutive expression. The near-perfect correspondence between the gene pair and classification within the *Clostridia*, the hint of constitutive expression, and its colocalization with housekeeping genes would seem to argue against the hypothesis that the SCIFF system makes an episodically produced metabolite such as a pheromone or a bacteriocin.

The His-Xaa-Ser repeat peptide system suggests novel chemistries for peptide modification. The precursor peptides are accompanied by not one but two radical SAM enzymes. In this system, comparative genomics identifies a presumptive peptide target that is longer than most bacteriocin precursors. However, only one small region of that protein family, the His-rich tripeptide repeat region, shows notable conservation, and we propose that region as the peptide modification target. A single enzyme creates three different cross-links through cysteine side chains in subtilosin A (18), while cyclodehydratases are shown to act at multiple sites, and on heterologous targets, in thiazole/oxazole-modified microcin precursors (21). It is likely that multiple His-Xaa-Ser sites are modified and that for each repeat the two enzymes act sequentially, although it is unclear if the target would be the histidine residue of each repeat, the serine, or both.

A surprising feature of the His-Xaa-Ser system, although one fully consistent with its highly sporadic species distribution, is that genes immediately neighboring its four-gene cassette are enriched in mobility markers: transposases, integrases, plasmid partitioning proteins, mobilization proteins, primases, restriction system proteins, toxin-antitoxin system (addiction module) proteins, and various phage protein homologs. At least 16 of 36 His-Xaa-Ser systems have such markers identifiable within three genes on one (10 cases) or both (6 cases) sides of the His-Xaa-Ser four-gene cassette. This exceeds the 25% rate that we observe for neighborhoods of Pfam model PF04013 family restriction systems and the 10 to 15% rates that we observe for one arsenite and one tellurite resistance marker. In general, proteins flanking His-Xaa-Ser systems and associated with mobility lack pairwise homology to each other, suggesting that the His-Xaa-Ser system does indicate the presence of any one specific type of mobile element. The sporadic distribution and coclustering with mobility markers cannot be consistent with housekeeping functions but could be consistent either with participation in mechanisms of lateral transfer or in providing a rapidly selectable trait such as immunity to an antibiotic, a toxic metal, or phage infection. In-

TABLE 2. Features of RTNP systems

Feature	Bacteriocins	Redox factors (PQQ/mycofactocin)	SCIFF	His-Xaa-Ser
Overall conservation	Weak	Strong	Strong	Weak
Most conserved region	Leader peptide	Mature region	Mature region	Mature region
Taxonomic range	Each bacteriocin system individually rare	Abundant in certain lineages, sporadic elsewhere	Near universal in *Clostridia*, rare elsewhere	Common but sporadically distributed
Additional paralogs of target peptide	Often	No	No	No
Typical no. of components	Variable	5–6 (PQQ), 5–7 (mycofactocin)	2	≥4
Regulated expression	Yes	Yes	Constitutive?	Unknown
Common neighbors	Transporters	Redox enzymes	Housekeeping genes	Mobility genes

terestingly, histidine often serves to provide a metal-binding ligand (22). Furthermore, a recent study proved the ribosomal origin of the methanobactin-OB3b (19), a highly modified peptide natural product used by methanotrophs to acquire copper and change its redox state from Cu(II) to Cu(I). Natural products made from His-Xaa-Ser repeat-containing precursors conceivably could resemble siderophores and methanobactins more closely than bacteriocins, binding to and perhaps conferring resistance to one or more toxic heavy metals. The actual function, however, is unknown.

We have presented evidence that family TIGR04081 contains naturally occurring selenoproteins and inferred that the family undergoes radical SAM-mediated posttranslational modifications. We hypothesize that the mature form may function as a selenium-containing bacteriocin, that is, a selenobacteriocin. According to this hypothesis, the Se atom remains in the mature form of the natural product derived from the precursor peptide and may form a part of a novel peptide modification. In Fig. 3, a column that is always Cys or SeCys occurs near the boundary that separates the consistently homologous N-terminal domain from a region that tends to be repetitive and low in complexity even in species with Cys instead of SeCys. All predicted SeCys residues are flanked on at least one side, and usually both, by glycine residues, as often observed for Cys residues subject to cyclization-forming modifications in other RTNP precursors.

An alternative hypothesis, however, is that the selenium atom, or sulfur atom in those species with Cys instead of SeCys at the equivalent sequence position, contributes to the molecular mechanism of modifications that occur elsewhere on the peptide, as a new example of substrate-assisted catalysis (7). Substrate-assisted catalysis would help to guarantee that the peptide-modifying action of the radical SAM enzyme works only on appropriate targets while still allowing great plasticity in the local sequence environment around the sites to be modified. The principle of substrate-assisted catalysis may help explain the occurrence of somewhat longer leader peptides in recently discovered RTNP precursor families that show large paralogous family expansions (14). It would therefore be of great interest to learn experimentally whether selenium remains part of the mature product or is removed with the leader peptide.

Nearly all known families in the selenoproteome are enzymes, most resembling thiol oxidoreductases with CxxU, UxxC, or related motifs. A recent study of the selenoproteome boosted the number of prokaryotic selenocysteine-containing protein families to over 50, with as few as two members found *in silico* taken as sufficient to identify a new family (38). Selenocysteine-containing RTNP families, however, present a special challenge because, in contrast to enzymes, RTNP families tend to be individually rare and tend to show weak if any sequence conservation in their propeptide regions. These traits interfere both with BLAST-based search sensitivity and with ratification of the family during manual review. The recognition here of a first selenocysteine-containing RTNP precursor family may now assist in the discovery of additional families.

The work presented here describes a contribution of many new protein family definitions for use in automated annotation pipelines that will produce more accurate genomic annotation of radical SAM enzymes in the future. In particular, it presents bioinformatics-based evidence for new families of modified peptides and new families of peptide modification enzymes. Recent bioinformatics and experimental work is expanding our catalog of, and appreciation for, natural products made from ribosomally produced peptides (28). The new systems that we have described are considerably more widespread than previously studied model systems with radical SAM involvement in peptide modification other than PQQ biosynthesis (Table 1). This work defines a domain (TIGR04085), found in the C-terminal region of a subset of radical SAM enzymes, for which member proteins show several variants of a previously noted iron-sulfur cluster-binding motif (see Fig. S1 in the supplemental material). This domain defines a molecular marker for probable peptide modification systems, perhaps the most abundant of any type. Several of these new biosynthetic systems are widely distributed in bacteria and should be investigated for biological roles other than bacteriocin-like antimicrobial activity.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01 HG004881 from the National Human Genome Research Institute.

Hierarchical clustering (Fig. S1 in the supplemental material) was performed by Doug Rusch.

REFERENCES

1. **Altschul, S. F., et al.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Benjdia, A., et al.** 2010. Anaerobic sulfatase-maturating enzyme—a mechanistic link with glycyl radical-activating enzymes? FEBS J. **277:**1906–1920.
3. **Bierbaum, G., and H. G. Sahl.** 2009. Lantibiotics: mode of action, biosynthesis and bioengineering. Curr. Pharm. Biotechnol. **10:**2–18.
4. **Brindley, A. A., R. Zajicek, M. J. Warren, S. J. Ferguson, and S. E. Rigby.** 2010. NirJ, a radical SAM family member of the d1 heme biogenesis cluster. FEBS Lett. **584:**2461–2466.
5. **Butcher, B. G., Y. P. Lin, and J. D. Helmann.** 2007. The yydFGHIJ operon of Bacillus subtilis encodes a peptide that induces the LiaRS two-component system. J. Bacteriol. **189:**8616–8625.
6. **Challand, M. R., et al.** 2009. Product inhibition in the radical S-adenosylmethionine family. FEBS Lett. **583:**1358–1362.
7. **Dall'Acqua, W., and P. Carter.** 2000. Substrate-assisted catalysis: molecular basis and biological significance. Protein Sci. **9:**1–9.
8. **Eddy, S. R.** 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. **23:**205–211.
9. **Finn, R. D., et al.** 2008. The Pfam protein families database. Nucleic Acids Res. **36:**D281–D288.
10. **Frey, P. A., A. D. Hegeman, and F. J. Ruzicka.** 2008. The radical SAM superfamily. Crit. Rev. Biochem. Mol. Biol. **43:**63–88.
11. **Graham, D. E., H. Xu, and R. H. White.** 2003. Identification of the 7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase required for coenzyme F(420) biosynthesis. Arch. Microbiol. **180:**455–464.
12. **Haft, D.** 2011. Bioinformatic evidence for a widely distributed, ribosomally produced electron carrier precursor, its maturation proteins, and its nicotinoprotein redox partners. BMC Genomics **12:**21.
13. **Haft, D. H.** 2009. A strain-variable bacteriocin in Bacillus anthracis and Bacillus cereus with repeated Cys-Xaa-Xaa motifs. Biol. Direct **4:**15.
14. **Haft, D. H., M. K. Basu, and D. A. Mitchell.** 2010. Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. BMC Biol. **8:**70.
15. **Haft, D. H., I. T. Paulsen, N. Ward, and J. D. Selengut.** 2006. Exopolysaccharide-associated protein sorting in environmental organisms: the PEP-CTERM/EpsH system. Application of a novel phylogenetic profiling heuristic. BMC Biol. **4:**29.
16. **Hiratsuka, T., et al.** 2008. An alternative menaquinone biosynthetic pathway operating in microorganisms. Science **321:**1670–1673.
17. **Ibrahim, M., et al.** 2007. Control of the transcription of a short gene encoding a cyclic peptide in Streptococcus thermophilus: a new quorum-sensing system? J. Bacteriol. **189:**8844–8854.
18. **Kawulka, K. E., et al.** 2004. Structure of subtilosin A, a cyclic antimicrobial peptide from Bacillus subtilis with unusual sulfur to alpha-carbon cross-links: formation and reduction of alpha-thio-alpha-amino acid derivatives. Biochemistry **43:**3385–3395.
19. **Krentz, B. D., et al.** 2010. A comparison of methanobactins from Methylo-

sinus trichosporium OB3b and Methylocystis strain Sb2 predicts methanobactins are synthesized from diverse peptide precursors modified to create a common core for binding and reducing copper ions. Biochemistry **49:**10117–10130.

20. **Lee, K. H., et al.** 2009. Characterization of RimO, a new member of the methylthiotransferase subclass of the radical SAM superfamily. Biochemistry **48:**10162–10174.
21. **Lee, S. W., et al.** 2008. Discovery of a widely distributed toxin biosynthetic gene cluster. Proc. Natl. Acad. Sci. U. S. A. **105:**5879–5884.
22. **Lippi, M., A. Passerini, M. Punta, B. Rost, and P. Frasconi.** 2008. MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. Bioinformatics **24:**2094–2095.
23. **McIntosh, J. A., M. S. Donia, and E. W. Schmidt.** 2009. Ribosomal peptide natural products: bridging the ribosomal and nonribosomal worlds. Nat. Prod. Rep. **26:**537–559.
24. **Mondot, S., et al.** 2011. Highlighting new phylogenetic specificities of Crohn's disease microbiota. Inflamm. Bowel Dis. **17:**185–192.
25. **Posewitz, M. C., et al.** 2004. Discovery of two novel radical S-adenosylmethionine proteins required for the assembly of an active [Fe] hydrogenase. J. Biol. Chem. **279:**25711–25720.
26. **Rea, M. C., et al.** 2010. Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against Clostridium difficile. Proc. Natl. Acad. Sci. U. S. A. **107:**9352–9357.
27. **Rebeil, R., and W. L. Nicholson.** 2001. The subunit structure and catalytic mechanism of the Bacillus subtilis DNA repair enzyme spore photoproduct lyase. Proc. Natl. Acad. Sci. U. S. A. **98:**9038–9043.
28. **Schmidt, E. W.** 2010. The hidden diversity of ribosomal peptide natural products. BMC Biol. **8:**83.
29. **Selengut, J. D., et al.** 2007. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. **35:**D260–D264.
30. **Sofia, H. J., G. Chen, B. G. Hetzler, J. F. Reyes-Spindola, and N. E. Miller.** 2001. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. Nucleic Acids Res. **29:**1097–1106.
31. **Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin.** 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:**33–36.
32. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
33. **Toh, S. M., L. Xiong, T. Bae, and A. S. Mankin.** 2008. The methyltransferase YfgB/RlmN is responsible for modification of adenosine 2503 in 23S rRNA. RNA **14:**98–106.
34. **Ward, J. H.** 1963. Hierachical grouping to optimize an objective function. J. Am. Stat. Assoc. **58:**236–244.
35. **Wecksler, S. R., et al.** 2009. Pyrroloquinoline quinone biogenesis: demonstration that PqqE from Klebsiella pneumoniae is a radical S-adenosyl-L-methionine enzyme. Biochemistry **48:**10151–10161.
36. **Yokoyama, K., M. Numakura, F. Kudo, D. Ohmori, and T. Eguchi.** 2007. Characterization and mechanistic study of a radical SAM dehydrogenase in the biosynthesis of butirosin. J. Am. Chem. Soc. **129:**15147–15155.
37. **Zhang, Y., and V. N. Gladyshev.** 2005. An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. Bioinformatics **21:**2580–2589.
38. **Zhang, Y., and V. N. Gladyshev.** 2008. Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. PLoS Genet. **4:**e1000095.